Recurrent Spatial-Temporal Attention Network for Action Recognition in Videos

Wenbin Du, Yali Wang, and Yu Qiao, *Senior Member, IEEE*

***Abstract*— Recent years have witnessed the popularity of using recurrent neural network (RNN) for action recognition in videos. However, videos are of high dimensionality and contain rich human dynamics with various motion scales, which makes the traditional RNNs difficult to capture complex action information. In this paper, we propose a novel recurrent spatial-temporal attention network (RSTAN) to address this challenge, where we introduce a spatial-temporal attention mechanism to adaptively identify key features from the global video context for every time-step prediction of RNN. More specifically, we make three main contributions from the following aspects. First, we reinforce the classical long short-term memory (LSTM) with a novel spatial-temporal attention module. At each time step, our module can automatically learn a spatial-temporal action representation from all sampled video frames, which is compact and highly relevant to the prediction at the current step. Second, we design an attention-driven appearance-motion fusion strategy to integrate appearance and motion LSTMs into a unified framework, where LSTMs with their spatial-temporal attention modules in two streams can be jointly trained in an end-to-end fashion. Third, we develop actor-attention regularization for RSTAN, which can guide our attention mechanism to focus on the important action regions around actors. We evaluate the proposed RSTAN on the benchmark UCF101, HMDB51 and JHMDB data sets. The experimental results show that, our RSTAN outperforms other recent RNN-based approaches on UCF101 and HMDB51 as well as achieves the state-of-the-art on JHMDB.**

***Index Terms*— Action recognition, RSTAN, spatial-temporal attention, attention-driven fusion, actor-attention regularization.**

I. Introduction

ACTION recognition in videos has been highlighted in computer vision research, due to its wide applications in video retrieval, surveillance, human-computer interaction, etc [1]. However, it is challenging to classify actions in the real-world videos, because complex background clutters and motion styles may cause various confusions, and high-dimensionality of videos may also restrict the recognition performance.

Recently, there is a growing interest in developing deep neural networks for action recognition [2]–[8]. Several early approaches exploited convolutional neural networks (CNNs) to learn deep representations from action videos [3], [4]. Karpathy et al. introduced a large scale Sports-1M video data set to train deep CNNs [4]. Simonyan et al. proposed two-stream CNNs, which achieved a competitive recognition performance by handling RGB images and optical flows separately with appearance and motion CNNs [9]. However, CNN-based architectures only captured the temporal motions in a short scale. This problem can be alleviated via recurrent neural networks (RNNs), especially LSTM [10] which has proved effective for modeling video sequences [2], [5], [6]. In most of these works, the inputs to LSTM are the high-level features extracted from the fully-connected layer of CNNs, but these features lack the fine details about action. Attention-based RNNs [11], [12] have been proposed to address this issue. However, most existing approaches only applied attention on the convolutional cube at the current step, which may be insufficient to make a reasonable prediction without taking into account the spatial-temporal context among video frames. As shown in Fig. 1, the 80th frame itself may not be sufficient to recognize *ThrowDiscus*, since the actor looks like standing still. On the other hand, this prediction can be largely enhanced with the 6th and 30th frames which contain the typical motions of *ThrowDiscus*. Similarly, the 55th and 6th frames yield highly complementary cues for the 30th frame.

Motivated by these facts, this paper proposes a novel recurrent spatial-temporal attention network (RSTAN) for action recognition in videos, which can adaptively learn a compact and detailed spatial-temporal feature to enhance action recognition at each time step of LSTM. Specifically, we make the main contributions from the following perspectives.

- We propose a spatial-temporal (S-T) attention module for LSTM. Instead of only attending to the salient regions of the current frame, our attention module exploits the global context for identifying spatial-temporal cues which are strongly-relevant to the prediction at the current step

Manuscript received May 27, 2017; revised October 5, 2017; accepted November 18, 2017. Date of publication November 29, 2017; date of current version December 27, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant U1613211, Grant 61633021, and Grant 61502470, in part by the Shenzhen Basic Research Program under Grant JCYJ20150925163005055 and Grant JCYJ20160229193541167, and in part by the External Cooperation Program of BIC Chinese Academy of Sciences under Grant 172644KYSB20150019 and Grant 172644KYSB20160033. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Pierre-Marc Jodoin. *(Wenbin Du and Yali Wang contributed equally to this work.) (Corresponding author: Yu Qiao.)*

W. Du is with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, China and also with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China (e-mail: wb.du@siat.ac.cn).

Y. Wang is with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China (e-mail: yl.wang@siat.ac.cn).

Y. Qiao is with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China, and also with The Chinese University of Hong Kong, Hong Kong (e-mail: yu.qiao@siat.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2778563

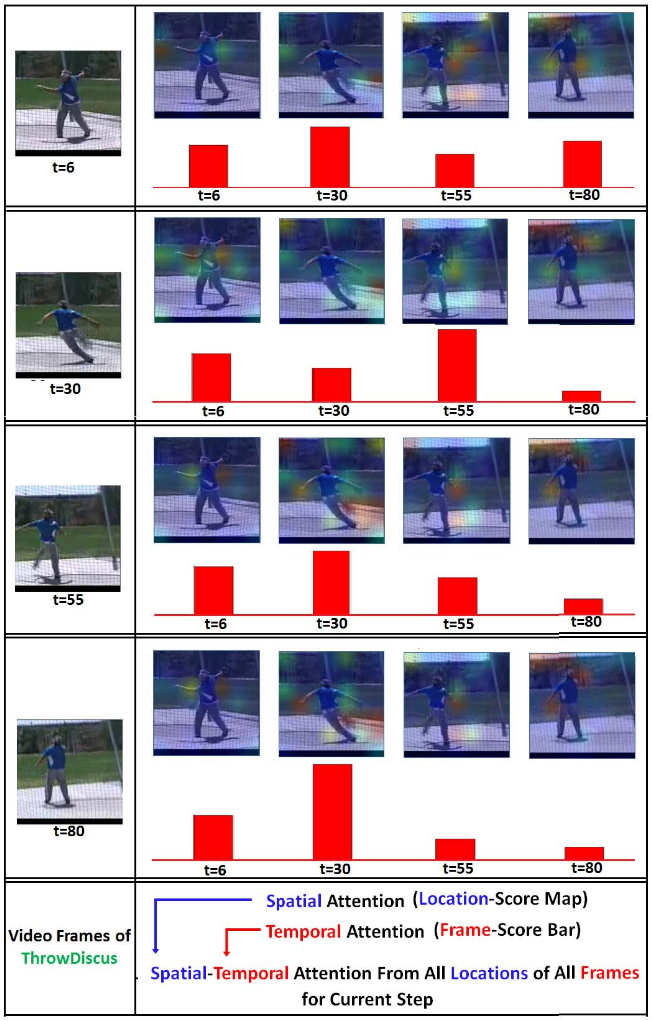

t=6
t=30
t=55
t=80
t=6
t=30
t=55
t=80
Video Frames of
ThrowDiscus
Spatial Attention (Location-Score Map)
Temporal Attention (Frame-Score Bar)
Spatial-Temporal Attention From All Locations of All Frames
for Current Step

Fig. 1. Illustration of our spatial-temporal (S-T) attention for *ThrowDiscus* in UCF101. At each time step, our S-T attention exploits video-level awareness for action recognition, where it not only attends to the salient regions of the current frame, but also leverages the global context for identifying spatial-temporal cues. Since these spatial-temporal cues are strongly-relevant to the prediction at the current step, our S-T attention can effectively enhance action representation at every step of LSTM.

of LSTM. The resulting spatial-temporal feature is complementary to the high-level fully-connected feature of CNNs, and thus their cooperation can effectively enhance action representation.

- We design an attention-driven appearance-motion (A-M) fusion strategy. By fusing S-T attention, we can integrate appearance and motion LSTMs into a unified framework, where LSTMs with their spatial-temporal attention modules in two streams can be jointly trained in an end-to-end fashion. Consequently, one LSTM (appearance or motion) can exploit the complementary action characteristics from the other LSTM to improve its discriminative power.
- We develop an actor-attention (A-A) regularization for the proposed RSTAN. It can further guide the learning procedure of our spatial-temporal attention, with more focus on the important regions where actions are most likely to occur.
- To show the effectiveness of our approach, we conduct extensive experiments on the benchmark UCF101, HMDB51 and JHMDB data sets. The empirical results demonstrate that, the proposed RSTAN outperforms other recent RNN-based approaches on UCF101 and HMDB51, and achieves the state-of-the-art on JHMDB.

The reminder of this paper is organized as follows. We first review related works for action recognition in Section II. After this, we introduce our recurrent spatial-temporal attention network (RSTAN) in detail in Section III. Then, we conduct a number of experiments, analyze the obtained results and show the effectiveness of our RSTAN in Section IV. Finally, we conclude this work in Section V.

II. Related Work

Action recognition has received intensive research works in recent years. In this section, we review the previous works which are related to our approach.

A. Hand-Crafted Features for Action Recognition

Early approaches for action recognition mainly rely on hand-crafted features [13]–[18], which represented videos by using a number of local descriptors. One popular approach is the improved dense trajectory (iDT) [17], consisting of Histogram of Oriented Gradients (HOG) [19], Histogram of Optical Flow (HOF), Motion Boundary Histogram (MBH) extracted along the spatial-temporal trajectories. By using bag of visual words (BovW) or its variants, these descriptors can be encoded into a high-dimension feature vector for action recognition [20], [21]. However, hand-crafted approaches may only capture the local contents, and thus lack the discriminative capacity to classify complex actions [22]. Additionally, several 3D action recognition approaches have been proposed by using human skeletons in a multi-task learning framework [23], [24].

B. CNNs for Action Recognition

Inspired by the remarkable successes of deep CNNs in image recognition [25]–[28], several works aimed at designing effective CNNs for action recognition in videos [3], [4], [7]–[9], [29]. Karpathy et al. introduced a large scale Sports-1M video data set to train deep CNNs [4], while their fusion strategies may be limited for temporal modeling. Another well-known approach is 3D CNN [3], [7], [8], which is a 3D extension of the standard 2D CNN by treating the time-domain as the third dimension. However, the high training complexity may require massive data sets [8] or 3D convolution kernel factorization [7]. Alternatively, two-stream CNNs [9], [29] can avoid this problem by training appearance and motion CNNs with RGB images and optical flows separately. One limitation in this approach is that the stacked optical flows can only capture motion information in short temporal scale. To improve the performance of this architecture, several extensions have been proposed by deep descriptors [22], [30]–[32], key volume mining [33], multiple streams fusions [34], [35], different pooling strategies [36]–[41], and temporal modeling [42]–[45].

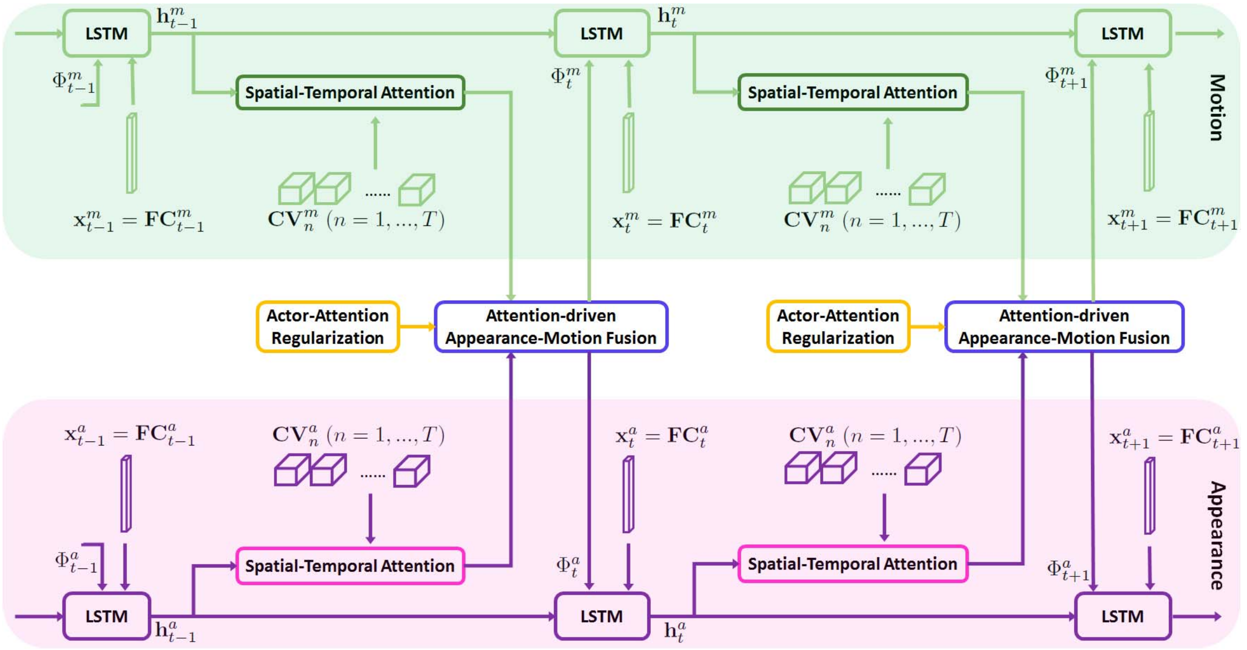

LSTM
$\mathbf{h}^m_{t-1}$
LSTM
$\mathbf{h}^m_t$
LSTM
Φ^m_{t-1}
Spatial-Temporal Attention
Φ^m_t
Spatial-Temporal Attention
Φ^m_{t+1}
Motion
$\mathbf{x}^m_{t-1} = \mathbf{FC}^m_{t-1}$
$\mathbf{CV}^m_n \ (n = 1, ..., T)$
$\mathbf{x}^m_t = \mathbf{FC}^m_t$
$\mathbf{CV}^m_n \ (n = 1, ..., T)$
$\mathbf{x}^m_{t+1} = \mathbf{FC}^m_{t+1}$
Actor-Attention Regularization
Attention-driven Appearance-Motion Fusion
Actor-Attention Regularization
Attention-driven Appearance-Motion Fusion
$\mathbf{x}^a_{t-1} = \mathbf{FC}^a_{t-1}$
$\mathbf{CV}^a_n \ (n = 1, ..., T)$
$\mathbf{x}^a_t = \mathbf{FC}^a_t$
$\mathbf{CV}^a_n \ (n = 1, ..., T)$
$\mathbf{x}^a_{t+1} = \mathbf{FC}^a_{t+1}$
Appearance
Φ^a_{t-1}
Spatial-Temporal Attention
Φ^a_t
Spatial-Temporal Attention
Φ^a_{t+1}
LSTM
$\mathbf{h}^a_{t-1}$
LSTM
$\mathbf{h}^a_t$
LSTM

Fig. 2. Our Recurrent Spatial-Temporal Attention Network (RSTAN, better viewed in color). At the t-th step, the previous hidden state $\mathbf{h}^{\star}_{t-1}$ guides our spatial-temporal attention to use the convolutional cubes of CNNs, i.e., $\mathbf{CV}^{\star}_1, \ldots, \mathbf{CV}^{\star}_T$, to automatically learn a spatial-temporal feature $\Phi^{\star}_t$ at the current step ($\star$ indicates a or m for appearance or motion). This contextual feature $\Phi^{\star}_t$ is complementary to the high-level feature of CNNs, i.e., $\mathbf{FC}^{\star}_t$, and thus their cooperation can enhance action representation at each step of LSTM. Furthermore, we integrate appearance and motion streams into a unified framework by using our attention-driven fusion, where LSTMs with their spatial-temporal attention modules can be jointly trained in an end-to-end fashion. In this case, one stream can take advantage of the complementary action characteristics in the other stream to improve its discriminative capacity. Finally, we design an actor-attention regularization to guide our attention to the important action regions around actors.

C. RNNs for Action Recognition

The sequential nature of video inspires researchers to learn video representations by recurrent neural networks, especially LSTM which has proven successful in modeling different types of sequences [10], [46]–[50]. In [2], [5], and [51], LSTMs were trained on top of two-stream CNNs for action recognition, while an unsupervised learning framework was explored by training LSTM with self-prediction in [6]. However, the inputs to these LSTMs are the high-level features obtained from the fully-connected (FC) layer of CNNs, which may lack fine action details in video frames [52]. Ballas *et al.* [52] and Gammulle *et al.* [53] applied features from different layers of CNNs as the inputs to RNNs, by designing multi-layer gated recurrent unit network or various feature fusion strategies. Recently, attention has been incorporated into LSTMs for video recognition [11], [12], [54]–[56], inspired by its efficiency for image understanding [48], [49]. Yeung *et al.* [12] introduced temporal attention into LSTM for emphasizing the key temporal segments. But their attention mechanism ignored the spatial (location) cues of action. Sharma *et al.* [54] designed an attention-driven LSTM by highlighting important spatial locations at each step of LSTM. However, the convolutional cubes in this work are extracted from RGB images, which largely ignored motion cues [11]. Attention in [11] and [55] integrated appearance and motion into a unified framework. However, similar to [54], the attention factors are determined by the convolutional features at the current step, which lack rich spatial-temporal contexts among video frames.

Different from all previous works, we propose a recurrent spatial-temporal attention network (RSTAN) in this paper, which integrates appearance and motion LSTMs using a novel spatial-temporal attention mechanism. At each time step, our attention can identify the relevant spatial-temporal regions within the global context, and the resulting contextual feature is compact and complementary to the current input of LSTM for action representation enhancement.

III. Recurrent Spatial-Temporal Attention Network (RSTAN)

In this section, we describe our Recurrent Spatial-Temporal Attention Network (RSTAN) for action recognition in videos. Specifically, we first extract features from two-stream CNNs. Then, we design a novel spatial-temporal attention module for LSTM, where we leverage the convolutional feature cubes at all time steps to automatically learn a spatial-temporal feature vector at the current step. Next, we develop an attention-driven appearance-motion fusion strategy to integrate appearance and motion streams into a unified framework. Finally, we propose an actor-attention regularization, which can guide our attention to the important action regions around actors. The whole framework of RSTAN is shown in Fig. 2.

A. Feature Extraction From Two-Stream CNNs

First, we extract appearance and motion features of video frames for action representation. To achieve it, we feed RGB images and stacked optical flows of video frames into the

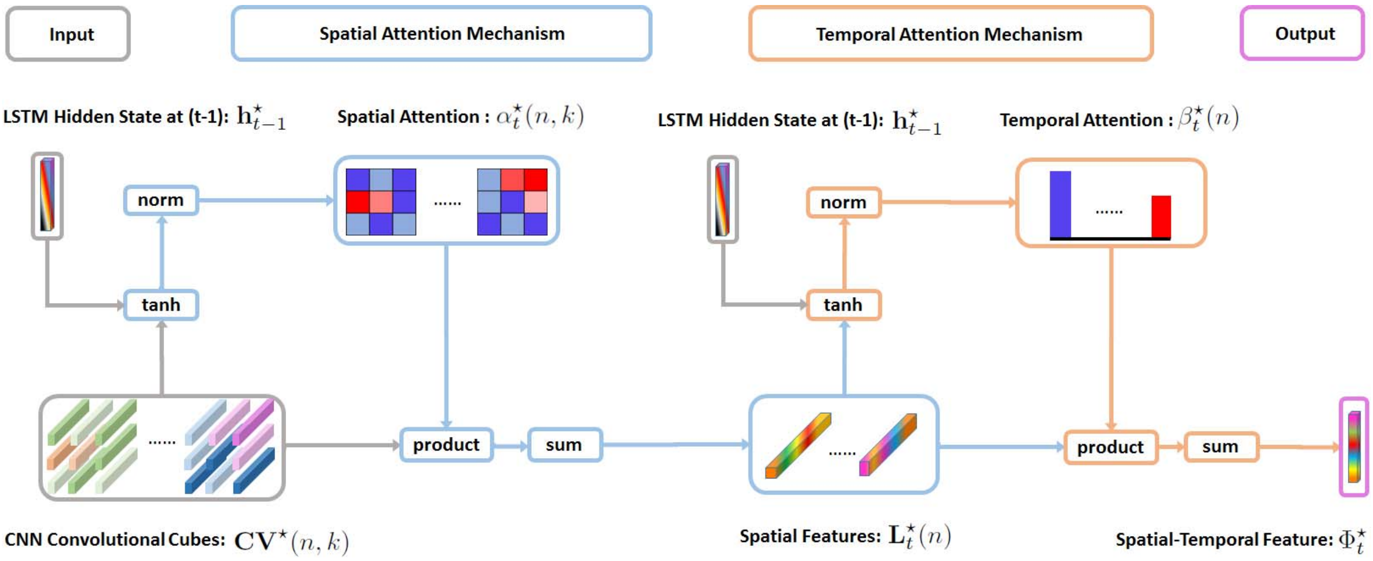

Input
Spatial Attention Mechanism
Temporal Attention Mechanism
Output
LSTM Hidden State at (t-1): $\mathbf{h}^\star_{t-1}$
Spatial Attention : $\alpha^\star_t(n,k)$
LSTM Hidden State at (t-1): $\mathbf{h}^\star_{t-1}$
Temporal Attention : $\beta^\star_t(n)$
norm
tanh
norm
tanh
......
product
sum
product
sum
CNN Convolutional Cubes: $\mathbf{CV}^\star(n,k)$
Spatial Features: $\mathbf{L}^\star_t(n)$
Spatial-Temporal Feature: $\Phi^\star_t$

Fig. 3. The proposed spatial-temporal (S-T) attention for LSTM (better viewed in color). At the t-th step, the S-T attention uses the previous hidden state $\mathbf{h}^\star_{t-1}$ as guidance to automatically learn a spatial-temporal feature $\Phi^\star_t$ from all convolutional cubes, i.e., $\mathbf{CV}^\star_1, \ldots, \mathbf{CV}^\star_T$ ($\star$ denotes a or m for appearance or motion). **Spatial Attention**: we first use the previous hidden state, $\mathbf{h}^\star_{t-1}$, and the feature vector at the k-th location of the n-th convolutional cube, $\mathbf{CV}^\star(n,k)$, to compute the normalized location-score, $\alpha^\star_t(n,k)$, where $n = 1, \ldots, T$ and $k = 1, \ldots, K^2$. Then we use $\mathbf{CV}^\star(n,k)$ and $\alpha^\star_t(n,k)$ to summarize T spatial features at the t-th step, $\mathbf{L}^\star_t(n)$, where $n = 1, \ldots, T$. **Temporal Attention**: we use $\mathbf{h}^\star_{t-1}$ and $\mathbf{L}^\star_t(n)$ to compute the normalized frame-score, $\beta^\star_t(n)$, where $n = 1, \ldots, T$. Then we use $\mathbf{L}^\star_t(n)$ and $\beta^\star_t(n)$ to obtain a compact spatial-temporal feature $\Phi^\star_t$ at the t-th step. More details can be found in the text.

widely-used two-stream CNNs architecture [9], [29], [42]. Note that, we here describe feature extraction in general for notation simplicity. More implementation details can be found in our experiments.

For the t-th video frame, we extract the convolutional feature cubes, $\mathbf{CV}^\star_t \in \mathbb{R}^{K \times K \times d_{cv}}$, from the **convolutional** layer of two-stream CNNs,

$$\mathbf{CV}^\star_t = \{\mathbf{CV}^\star(t,1), \ldots, \mathbf{CV}^\star(t,K^2)\}, \tag{1}$$

where $\star$ represents a for appearance-stream CNN or m for motion-steam CNN, $\mathbf{CV}^\star_t$ consists of d_{cv} feature maps with spatial size of $K \times K$, and we denote this cube as a set of feature vectors at different spatial locations, i.e., $\mathbf{CV}^\star(t,k) \in \mathbb{R}^{d_{cv}}$ where $k = 1, \ldots, K^2$. In addition, we extract the d_{fc} dimension feature vector, $\mathbf{FC}^\star_t \in \mathbb{R}^{d_{fc}}$, from the **fully-connected** layer of two-stream CNNs.

As a result, each video can be represented as a sequence of extracted features of sampled frames, i.e., $\{\mathbf{CV}^\star_t, \mathbf{FC}^\star_t\}^T_{t=1}$. In the next, we design a novel spatial-temporal attention for LSTM to model this sequence.

B. Spatial-Temporal (S-T) Attention for LSTM

After extracting the sequences of deep features from two-stream CNNs, it is natural to use LSTM to encode the temporal structure of these sequentially-ordered features. A straightforward way is to use $\mathbf{FC}^a_t$ and $\mathbf{FC}^m_t$ as the t-th input to train appearance and motion LSTMs respectively, as in previous works [2], [5]. But the performance of this approach is often restricted due to the fact that the high-level $\mathbf{FC}^\star_t$ often lacks detailed location information of action. Alternatively, attention to the convolutional cubes can capture fine spatial information [11], [54]. However, only attention to the convolutional cube at the current step may not be sufficient to capture complex spatial-temporal cues of different actions. Hence, we introduce a novel spatial-temporal attention module to identify the key contextual information from all convolutional cubes at each-step prediction in LSTM. An illustration is shown in Fig. 3.

1) Spatial Attention: Given the previous hidden state $\mathbf{h}^\star_{t-1}$ of LSTM, we first estimate which locations of convolutional cubes among all frames are important to the t-th step. To achieve it, we design a spatial attention mechanism with the guidance of $\mathbf{h}^\star_{t-1}$,

$$\tilde{\alpha}^\star_t(n,k) = \mathbf{v}^\star_\alpha \tanh(\mathbf{A}^\star_h \mathbf{h}^\star_{t-1} + \mathbf{A}^\star_{cv} \mathbf{CV}^\star(n,k) + \mathbf{b}^\star_\alpha) \tag{2}$$

where $\tilde{\alpha}^\star_t(n,k) \in \mathbb{R}$ is the un-normalized location score of $\mathbf{CV}^\star(n,k)$ at the t-th step, $\mathbf{CV}^\star(n,k)$ is the feature vector at the k-th location of the n-th convolutional cube ($n = 1, \ldots T$; $k = 1, \ldots, K^2$), and $\{\mathbf{v}^\star_\alpha, \mathbf{A}^\star_h, \mathbf{A}^\star_{cv}, \mathbf{b}^\star_\alpha\}$ are the spatial attention model parameters. For the n-th convolutional cube, we normalize $\tilde{\alpha}^\star_t(n,k)$ to the spatial location score $\alpha^\star_t(n,k)$,

$$\alpha^\star_t(n,k) = \frac{(\exp\{\tilde{\alpha}^\star_t(n,k)\})^{\gamma_\alpha}}{\Sigma^{K^2}_{j=1}(\exp\{\tilde{\alpha}^\star_t(n,j)\})^{\gamma_\alpha}}, \tag{3}$$

where γ_α is introduced to control the sharpness of the location-score map.

As the location-score $\alpha^\star_t(n,k)$ reflects the spatial importance of the k-th feature vector in the n-th convolutional cube, we propose to use $\alpha^\star_t(n,k)$ as weights to summarize the n-th convolutional cube as a spatial feature $\mathbf{L}^\star_t(n)$,

$$\mathbf{L}^\star_t(n) = \Sigma^{K^2}_{k=1} \alpha^\star_t(n,k)\mathbf{CV}^\star(n,k). \tag{4}$$

where $n = 1, \ldots, T$. Each of $\{\mathbf{L}^\star_t(n)\}^T_{n=1}$ emphasizes the important spatial details of action, with regard to the t-th step.

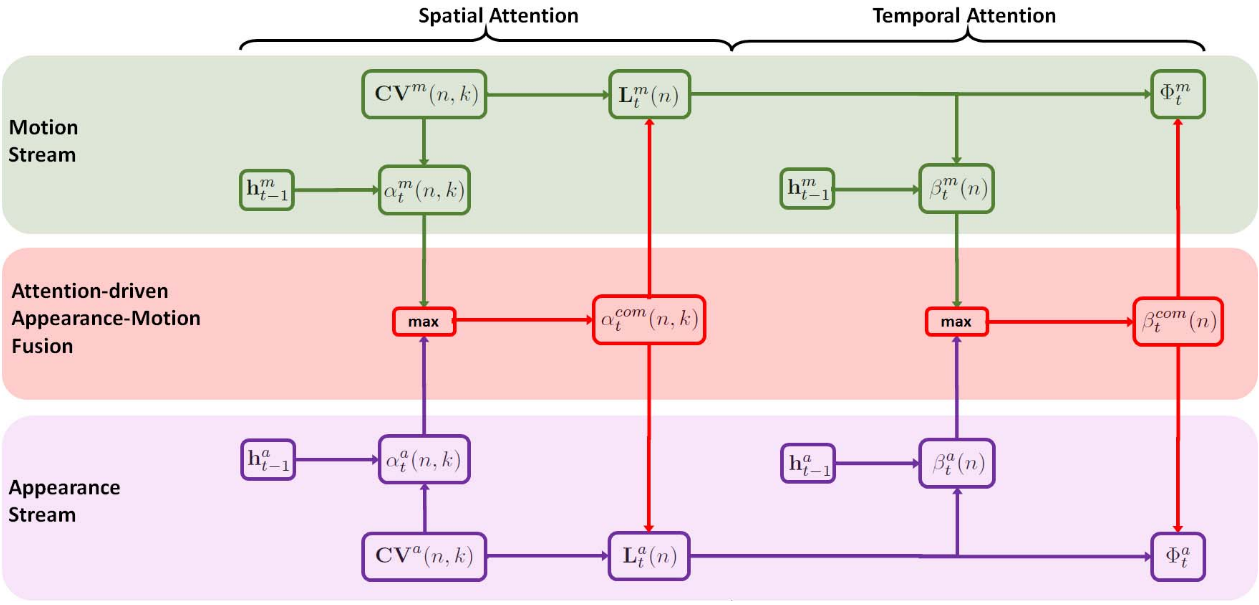


Fig. 4. Our attention-driven appearance-motion fusion (better viewed in color). By fusing S-T attention modules in the appearance and motion streams, we integrate two streams into a unified framework. **First**, we operate max-fusion on the location-scores of spatial attention, to highlight the important spatial regions in both streams. Then we use $\mathbf{CV}^\star(n,k)$ and the fused score $\alpha_t^{com}(n,k)$ to compute the spatial feature $\mathbf{L}_t^\star(n)$ respectively for appearance and motion streams. **Next**, we use the obtained $\mathbf{L}_t^\star(n)$ to calculate the frame-score $\beta_t^\star(n)$ for different streams. Then we operate max-fusion on the frame-scores of temporal attention, to highlight the important temporal frames in both streams. Subsequently, we use $\mathbf{L}_t^\star(n)$ and the fused score $\beta_t^{com}(n)$ to obtain $\Phi_t^\star$ for two streams. **Finally**, we apply Φ_t^a and Φ_t^m to appearance and motion LSTMs in Eq. (8)-(13). The obtained hidden states $\mathbf{h}_t^a$ and $\mathbf{h}_t^m$ are used in Eq. (16)-(17) to jointly train LSTMs and their S-T attention modules in an end-to-end way. More details can be found in the text.

2) Temporal Attention: After obtaining T spatial features $\{\mathbf{L}_t^\star(n)\}_{n=1}^T$, we estimate which ones are important to the t-th step. To achieve it, we propose a temporal attention mechanism to estimate the frame-score of each spatial feature as follows,

$$\tilde{\beta}_t^\star(n) = \mathbf{v}_\beta^\star \tanh(\mathbf{B}_h^\star \mathbf{h}_{t-1}^\star + \mathbf{B}_L^\star \mathbf{L}_t^\star(n) + \mathbf{b}_\beta^\star), \tag{5}$$

$$\beta_t^\star(n) = \frac{(\exp\{\tilde{\beta}_t^\star(n)\})^{\gamma_\beta}}{\Sigma_{j=1}^T (\exp\{\tilde{\beta}_t^\star(j)\})^{\gamma_\beta}}, \tag{6}$$

where $\tilde{\beta}_t^\star(n)$, $\beta_t^\star(n) \in \mathbb{R}$ are respectively the un-normalized and normalized frame score of the n-th spatial feature $\mathbf{L}_t^\star(n)$ ($n = 1, \ldots, T$), $\{\mathbf{v}_\beta^\star, \mathbf{B}_h^\star, \mathbf{B}_L^\star, \mathbf{b}_\beta^\star\}$ are temporal attention model parameters, and γ_β is the sharpness parameter for normalization.

As the frame-score $\beta_t^\star(n)$ reflects the temporal importance of the n-th spatial feature for the t-th step, we propose to use $\beta_t^\star(n)$ to summarize all spatial features as a spatial-temporal feature $\Phi_t^\star$,

$$\Phi_t^\star = \Sigma_{n=1}^T \beta_t^\star(n) \mathbf{L}_t^\star(n). \tag{7}$$

The main novelty of our spatial-temporal attention module is that, instead of only focusing on the spatial locations in the current frame, *it can automatically learn a spatial-temporal feature $\Phi_t^\star$. This feature can leverage the global context to capture the important spatial-temporal cues, which are strongly relevant to the prediction at the current step*. In addition, the dimensionality of $\Phi_t^\star$ is the same as the one of $\mathbf{CV}^\star(n,k) \in \mathbb{R}^{d_{cv}}$. Hence, $\Phi_t^\star$ is a compact feature without suffering from high-dimensionality, which can alleviate the training difficulty.

3) Our Spatial-Temporal Attention for LSTM: After obtaining the spatial-temporal feature $\Phi_t^\star$, we feed it into LSTM as an extra input at the t-th step,

$$\begin{aligned}
\mathbf{i}_t^\star &= \sigma(\mathbf{U}_{i_\star}^x \mathbf{x}_t^\star + \mathbf{U}_{i_\star}^\phi \Phi_t^\star + \mathbf{U}_{i_\star}^h \mathbf{h}_{t-1}^\star + \mathbf{b}_{i_\star}), && (8)\\
\mathbf{f}_t^\star &= \sigma(\mathbf{U}_{f_\star}^x \mathbf{x}_t^\star + \mathbf{U}_{f_\star}^\phi \Phi_t^\star + \mathbf{U}_{f_\star}^h \mathbf{h}_{t-1}^\star + \mathbf{b}_{f_\star}), && (9)\\
\mathbf{o}_t^\star &= \sigma(\mathbf{U}_{o_\star}^x \mathbf{x}_t^\star + \mathbf{U}_{o_\star}^\phi \Phi_t^\star + \mathbf{U}_{o_\star}^h \mathbf{h}_{t-1}^\star + \mathbf{b}_{o_\star}), && (10)\\
\mathbf{g}_t^\star &= \tanh(\mathbf{U}_{g_\star}^x \mathbf{x}_t^\star + \mathbf{U}_{g_\star}^\phi \Phi_t^\star + \mathbf{U}_{g_\star}^h \mathbf{h}_{t-1}^\star + \mathbf{b}_{g_\star}), && (11)\\
\mathbf{c}_t^\star &= \mathbf{f}_t^\star \odot \mathbf{c}_{t-1}^\star + \mathbf{i}_t^\star \odot \mathbf{g}_t^\star, && (12)\\
\mathbf{h}_t^\star &= \mathbf{o}_t^\star \odot \tanh(\mathbf{c}_t^\star), && (13)
\end{aligned}$$

where $\star$ represents either a (appearance) or m (motion), the sets of $\mathbf{U}$ and $\mathbf{b}$ are the parameters of LSTM, $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and tanh functions, $\odot$ is the element-wise multiplication, $\mathbf{i}_t^\star$, $\mathbf{f}_t^\star$ and $\mathbf{o}_t^\star$ are the input, forget and output gates, $\mathbf{g}_t^\star$, $\mathbf{c}_t^\star$ and $\mathbf{h}_t^\star$ are the candidate memory, the memory state and the hidden state, $\mathbf{x}_t^\star$ is the input to LSTM, which is the feature vector obtained from the fully-connected layer of CNN, i.e., $\mathbf{FC}_t^\star$.

Since our contextual feature $\Phi_t^\star$ contains the detailed spatial-temporal cues at the current step, *$\Phi_t^\star$ is complimentary to the high-level input feature of LSTM, i.e., $\mathbf{x}_t^\star = \mathbf{FC}_t^\star$. As a result, the cooperation of both features allows LSTM to learn a discriminative action representation at each step*. In the following, we propose an attention-driven appearance-motion fusion strategy to integrate appearance and motion LSTMs within a unified framework.

C. Attention-Driven Appearance-Motion (A-M) Fusion

Our attention module in Section III-B is designed separately for appearance or motion streams. Motivated by the fact that

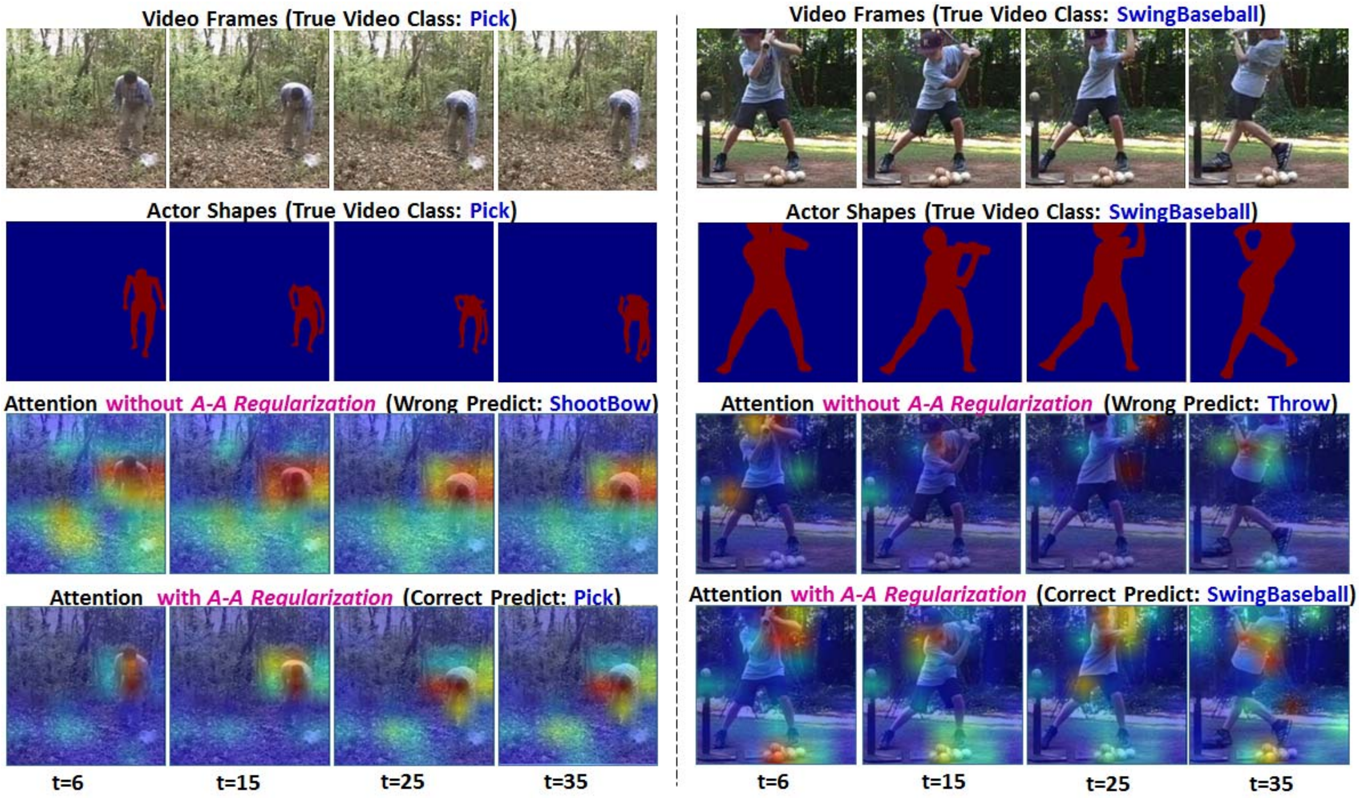

Video Frames (True Video Class: Pick)
Actor Shapes (True Video Class: Pick)
Attention without A-A Regularization (Wrong Predict: ShootBow)
Attention with A-A Regularization (Correct Predict: Pick)
Video Frames (True Video Class: SwingBaseball)
Actor Shapes (True Video Class: SwingBaseball)
Attention without A-A Regularization (Wrong Predict: Throw)
Attention with A-A Regularization (Correct Predict: SwingBaseball)
t=6
t=15
t=25
t=35

Fig. 5. Illustration of our actor-attention (A-A) regularization (*Pick* and *SwingBaseball* in JHMDB). One can see that, attention without A-A regularization may either overestimate the action-irrelevant regions (for the *Pick* action) or underestimate the action-relevant regions (for the *SwingBaseball* action), which could lead to the wrong predictions. On the contrary, attention with A-A regularization can reasonably highlight the important regions around actors for action recognition. Furthermore, it can enhance the relevant contextual information in these regions (such as baseballs in *Swing Baseball*) as well as reduce the influence of irrelevant regions (such as trees and grasses in *Pick*), in order to assist our network to make the correct predictions.

appearance and motion streams are highly complementary to describe different aspects of action in videos [34], we propose an attention-driven appearance-motion (A-M) fusion strategy to incorporate two streams into a unified framework. An illustration is shown in Fig. 4.

First, we fuse the location-scores of spatial attention to highlight the important spatial regions in both appearance and motion streams. This can be achieved by taking the maximum of $\alpha_t^a(n,k)$ (appearance) and $\alpha_t^m(n,k)$ (motion),

$$\alpha_t^{com}(n,k) = \max\{\alpha_t^a(n,k), \alpha_t^m(n,k)\}, \qquad (14)$$

where $n = 1, \ldots, T$ and $k = 1, \ldots, K^2$. Then, instead of using the individual score $\alpha_t^\star(n,k)$, we use the fused score $\alpha_t^{com}(n,k)$ and $\mathbf{CV}^\star(n,k)$ in Eq. (4) to compute the spatial feature $\mathbf{L}_t^\star(n)$ for different streams.

Next, we use the obtained $\mathbf{L}_t^\star(n)$ to compute its frame-score in the temporal attention, $\beta_t^\star(n)$ in Eq. (5)-(6). Subsequently, we fuse the frame-scores of temporal attention to highlight the important temporal frames in both appearance and motion streams. This can be achieved by taking the maximum of $\beta_t^a(n)$ (appearance) and $\beta_t^m(n)$ (motion),

$$\beta_t^{com}(n) = \max\{\beta_t^a(n), \beta_t^m(n)\}. \qquad (15)$$

Instead of using the individual score $\beta_t^\star(n)$, we use the fused score $\beta_t^{com}(n)$ and $\mathbf{L}_t^\star(n)$ in Eq. (7) to obtain $\Phi_t^\star$.

Finally, we apply $\Phi_t^\star$ to LSTM in Eq. (8)-(13), and obtain the current hidden states in both appearance and motion streams, i.e., $\mathbf{h}_t^a$ and $\mathbf{h}_t^m$. To train our RSTAN in an end-to-end fashion, we integrate both $\mathbf{h}_t^a$ and $\mathbf{h}_t^m$ into softmax,

$$\hat{\mathbf{y}}_t = softmax(\mathbf{W}_a\mathbf{h}_t^a + \mathbf{W}_m\mathbf{h}_t^m + \mathbf{b}_{am}), \qquad (16)$$

where $\hat{\mathbf{y}}_t$ is the prediction vector, $\{\mathbf{W}_a, \mathbf{W}_m, \mathbf{b}_{am}\}$ are parameters. Consequently, we use cross-entropy with weight decay as the main loss of training RSTAN,

$$\mathcal{L}_{main} = -\Sigma_{t=1}^{T}\Sigma_{c=1}^{C}(\mathbf{y}_{t,c}\log\hat{\mathbf{y}}_{t,c}) + \lambda_\Theta \parallel \Theta \parallel_2, \qquad (17)$$

where C is the number of action classes, T is the number of total time steps, Θ represents all the model parameters, λ_Θ is the coefficient for weight decay, and $\mathbf{y}_t$ denotes the ground-truth (one-hot label vector).

With our attention-driven fusion strategy, appearance and motion streams can be integrated into a unified framework, where LSTMs with their S-T attention modules can be jointly trained in an end-to-end fashion. In this case, *one stream can leverage the complementary action characteristics in the other stream to improve its action modeling capacity*.

D. Actor-Attention (A-A) Regularization

Next, we develop an actor-attention (A-A) regularization for the proposed network. This is mainly motivated by the fact that actions often occur in the regions around actors. As shown in Fig. 5, the regions indicated by the silhouette of actors are clearly important to recognize different actions.

TABLE I

SPATIAL-TEMPORAL ATTENTION. BASELINE: STANDARD LSTM WITHOUT ATTENTION. *S-T*: OUR RSTAN WITH SPATIAL-TEMPORAL ATTENTION. FOR (APPEARANCE+MOTION), THE RESULTS OF BASELINE NET AND OUR *S-T* NET ARE OBTAINED BY FUSING THE PREDICTION SCORES OF APPEARANCE AND MOTION STREAMS

Appearance	UCF101	HMDB51	JHMDB
Baseline Net	79.8	52.9	45.5
Our *S-T* Net	**80.2**	**53.4**	**47.0**
Motion	UCF101	HMDB51	JHMDB
Baseline Net	85.3	61.0	58.6
Our *S-T* Net	**86.9**	**63.1**	**62.0**
(Appearance+Motion)	UCF101	HMDB51	JHMDB
Baseline Net	89.5	65.7	61.2
Our *S-T* Net	**90.4**	**67.7**	**65.7**

TABLE II

IMPORTANT MODEL PROPERTIES EVALUATION. *S-T*: SPATIAL-TEMPORAL ATTENTION. *A-M*: ATTENTION-DRIVEN APPEARANCE-MOTION FUSION. $A\text{-}M_{(MEAN)}$ IS *A-M* FUSION WITH THE MEAN OPERATION. $A\text{-}M_{(MAX)}$ IS *A-M* FUSION WITH THE MAX OPERATION WHICH WE INTRODUCE FOR OUR RSTAN IN EQ. (14) AND (15). *A-A*: ACTOR-ATTENTION REGULARIZATION. $A\text{-}A_{(CE)}$ IS *A-A* REGULARIZATION WITH CROSS-ENTROPY LOSS. $A\text{-}A_{(L2)}$ IS *A-A* REGULARIZATION WITH L2 LOSS WHICH WE INTRODUCE FOR OUR RSTAN IN EQ. (18). ADDITIONALLY, THE PREDICTION OF *S-T* NET IS OBTAINED BY SCORE FUSION OF INDEPENDENT APPEARANCE AND MOTION STREAMS, AS THERE IS NO ATTENTION-DRIVEN *A-M* FUSION IN THIS *S-T* NET. FINALLY, WE USE JHMDB TO EVALUATION OUR RSTAN WITH A-A REGULARIZATION, DUE TO THE FACT THAT HUMAN SILHOUETTES ARE ANNOTATED IN JHMDB

Our RSTAN (Appearance+Motion)	UCF101	HMDB51	JHMDB
S-T	90.4	67.7	65.7
$S\text{-}T + A\text{-}M_{(MEAN)}$	**92.2**	68.9	67.2
$S\text{-}T + A\text{-}M_{(MAX)}$	**92.2**	**70.9**	67.4
$S\text{-}T + A\text{-}M_{(MAX)} + A\text{-}A_{(CE)}$	n/a	n/a	70.2
$S\text{-}T + A\text{-}M_{(MAX)} + A\text{-}A_{(L2)}$	n/a	n/a	**73.5**

TABLE III

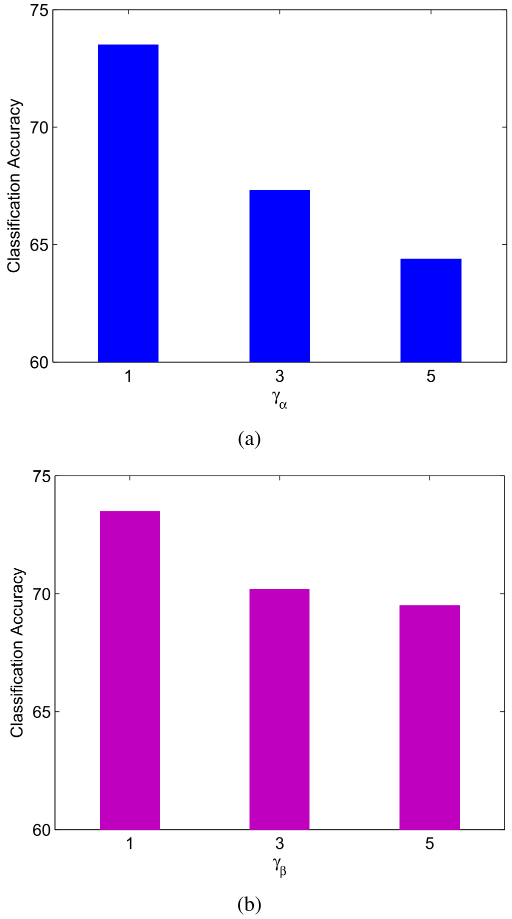
CURRENT-FRAME ATTENTION VS. SPATIAL-TEMPORAL ATTENTION. THE CURRENT-FRAME ATTENTION NET IS OUR RSTAN ONLY WITH ATTENTION TO THE CURRENT FRAME AT EACH TIME STEP

Our RSTAN	UCF101	HMDB51	JHMDB
Current-Frame Attention	91.5	68.4	70.9
Spatial-Temporal Attention	**92.2**	**70.9**	**73.5**

Hence, we propose an A-A regularization by penalizing the t-th spatial attention map with the human silhouette at the t-th step. Inspired by human pose estimation in images and videos [57], [58], we use L2 loss (between the spatial attention map and the ground-truth human silhouette) as A-A regularization to capture spatial multi-modality of human actions,

$$\mathcal{L}_{AA} = \Sigma_{t=1}^{T} \Sigma_{k=1}^{K^2} (\alpha_t^{com}(t,k) - \mathbf{M}(t,k))^2, \tag{18}$$

where $\alpha_t^{com}(t,k)$ is the appearance-motion-fused score at the k-th location of the t-th spatial attention map for the t-th step. $\mathbf{M}(t,\cdot)$ is the ground-truth with regard to the human silhouette of actor at the t-th step. To obtain $\mathbf{M}(t,\cdot)$, we resize the video frame into $K \times K$ and find the region in $\mathbf{M}(t,\cdot)$ which corresponds to the ground-truth silhouette of the original frame. Suppose that there are N_{box} pixels within the silhouette region of $\mathbf{M}(t,\cdot)$, we use the mean $1/N_{box}$ as the value for each of these pixels. The values for other regions in $\mathbf{M}(t,\cdot)$ are zero.



(a)

(b)

Fig. 6. Evaluation on sharpness parameters in our spatial-temporal attention. Note that, when we change γ_α in the spatial-attention (or γ_β in the temporal-attention), we fix $\gamma_\beta = 1$ in the temporal-attention (or $\gamma_\alpha = 1$ in the spatial-attention). One can see that, when γ_α (or γ_β) is increasing, the spatial (or temporal) attention is often concentrated on few important regions (or frames). As a result, the action prediction may be fragmented to decrease accuracy. (a) Accuracy when changing γ_α in the spatial-attention. (b) Accuracy when changing γ_β in the temporal-attention.

With our A-A regularization, the total loss is written as,

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda_{AA}\mathcal{L}_{AA}, \tag{19}$$

where λ_{AA} is the coefficient. Based on this total loss, *our attention mechanism does not only exploit the complementary spatial-temporal cues from the past and future frames to assist the current prediction, but it also leverages A-A regularization as additional supervision which guides our attention learning to focus on the important action-relevant regions.*

IV. EXPERIMENTS

In this section, we evaluate our recurrent spatial-temporal attention network (RSTAN) on three benchmark data sets, to show its effectiveness on action recognition in videos. First, we introduce the evaluation data sets and the implementation details. Then, we comprehensively investigate a number of model properties in our RSTAN. Next, we compare our RSTAN with other state-of-the-art approaches. Finally, we

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART ON UCF101 AND HMDB51 (AVERAGE OVER THREE SPLITS). THE RESULT OF SOFT ATTENTION MODEL [54] IS FROM THE RE-IMPLEMENTATION IN [11], AS [54] DID NOT REPORT THE RESULT ON UCF101. OUR RSTAN$_{GPCNN}$ DENOTES THAT CNN FEATURES ARE EXTRACTED FROM GPCNN [29]. OUR RSTAN$_{TSN}$ DENOTES THAT CNN FEATURES ARE EXTRACTED FROM TSN [42]. NOTE THAT, THE ORIGINAL GPCNN DOES NOT PERFORM ON HMDB51. HENCE, WE USE TSN AS DEEP FEATURE EXTRACTOR FOR HMDB51. ONE CAN SEE THAT, OUR RSTAN OUTPERFORMS THE RECENT RNN-BASED APPROACHES AND ACHIEVES THE COMPARABLE RESULTS TO THE STATE-OF-THE-ART

Other State-of-the-art Approaches	Year	UCF101	HMDB51
iDT-FV [20]	2014	87.9	61.1
Large-scale ConvNet [4]	2014	65.4	-
Two-Stream ConvNet [9]	2014	88.0	59.4
C3D [8]	2015	85.2	-
(C3D+iDT)-FV [8]	2015	90.4	-
Factorized ConvNet [7]	2015	88.1	59.1
GPCNN [29]	2015	91.4	-
(TDD+iDT)-FV [22]	2015	91.5	65.9
Dynamic Image Network [65]	2016	89.1	65.2
TwoStream 3Dnet [66]	2016	90.2	-
AdaScan [36]	2016	89.4	54.9
AdaScan + iDT-FV [36]	2016	91.3	61.0
three-stream sDTD [30]	2016	92.2	65.2
LTC Network [67]	2016	91.7	64.8
LTC Network + iDT-FV [67]	2016	92.7	67.2
Key Volume ConvNet [33]	2016	93.1	63.3
Convolutional Fusion [34]	2016	92.5	65.4
Convolutional Fusion + iDT-FV [34]	2016	93.5	69.2
TSN (2 modalities) [42]	2016	94.0	68.5
ST-ResNet [44]	2016	93.4	66.4
ST-ResNet + iDT-FV [44]	2016	94.6	70.3
FV-VAE [68]	2016	94.2	-
Hidden Two-Stream [69]	2017	90.3	58.9
2ndOrder Pooling [41]	2017	89.4	67.8
Chained Multi-Stream [35]	2017	91.1	69.7
ActionVLAD [32]	2017	92.7	66.9
Temporal-Inception [45]	2017	93.9	67.5
(SVMP+NSVMP)+iDT-FV [38]	2017	94.6	70.6
RNN Based Approaches	Year	UCF101	HMDB51
LRCN [2]	2014	82.9	-
Composite LSTM Model [6]	2015	84.3	44.0
Beyond Short Snippets Model [5]	2015	88.6	-
Hybrid Network [51]	2015	91.3	-
Soft Attention Model [54]	2015	77.0	41.3
GRN-RCN [52]	2016	90.8	-
VideoLSTM [11]	2016	89.2	56.4
VideoLSTM [11]+ iDT-FV	2016	91.5	63.0
Hierarchical Attention [55]	2016	92.7	64.3
RMDN [56]	2017	82.8	-
TS-LSTM [45]	2017	94.1	69.0
Our Recurrent Net & Its Variants		UCF101	HMDB51
Our RSTAN$_{GPCNN}$		92.5	-
Our RSTAN$_{GPCNN}$ + iDT-FV		93.9	-
Our RSTAN$_{TSN}$		94.6	70.5
Our RSTAN$_{TSN}$ + iDT-FV		**95.1**	**79.9**

visualize our spatial-temporal attention mechanism to show the important contextual action regions captured by our RSTAN.

A. Data Sets

In our experiment, we choose three popular benchmarks for action recognition [9], [29], [42], [59], namely UCF101 [60], HMDB51 [61] and JHMDB [62]. UCF101 consists of 13,320 videos with 101 action classes, covering a broad range of activities such as sports and human-object interaction. HMDB51 consists of 6,766 videos with 51 action classes, collecting from movies and web sources (such as YouTube). JHMDB consists of 928 videos with 21 action classes. As human silhouettes are annotated for all video frames in JHMDB, it is a suitable choice to verify the effectiveness of our actor-attention (A-A) regularization. For all the data sets, we use the standard evaluation protocol and report accuracy over the predefined train/test splits [9].

TABLE V

COMPARISON WITH THE STATE-OF-THE-ART ON JHMDB (AVERAGE OVER THREE SPLITS)

State-of-the-art Approaches	Year	JHMDB
iDT-FV [17]	2013	65.9
iDT-(FV+Stacked FV) [59]	2014	69.0
Finding Action Tubes [70]	2015	62.5
Pose-based CNN (w/o test GT) [71]	2015	61.1
Multi-region Two-Stream R-CNN [72]	2016	71.1
Two-stream LSTM [53]	2017	69.0
GRP [40]	2017	70.6
2ndOrder Pooling [41]	2017	73.7
Chained Multi-Stream [35]	2017	76.1
Our Recurrent Net & Its Variants		JHMDB
Our RSTAN$_{GPCNN}$		72.0
Our RSTAN$_{TSN}$		**79.2**

B. Implementation Details

Unless stated otherwise, we perform our RSTAN with the following implementation details. Firstly, since our main novelty is the proposed recurrent spatial-temporal attention network instead of CNN feature design, we choose two widely-used two-stream CNN architectures as the standard feature extraction module, i.e., good-practice CNN (GPCNN) [29] and temporal segment net (TSN) [42]. In this case, we follow the same data argument strategies as used in these models, and feed RGB image and stacked optical flow respectively into appearance-stream and motion-stream CNNs for deep feature extraction. For each video frame, we extract the convolutional feature cube (GPCNN: last convolutional layer after pooling, $7 \times 7 \times 512$; TSN: inception-5a layer, $7 \times 7 \times 1024$) for our attention. In addition, we extract the feature vector (GPCNN: first fully-connected layer after ReLU, $4,096$ dimension; TSN: last pooling layer, 1024 dimension) as input to LSTM. The dimension of all hidden variables in LSTM of our RSTAN is 1024. Secondly, for UCF101/HMDB51, 16/32 videos are randomly chosen for each mini-batch, and 16/8 frames are randomly sampled from each video with equal interval. For JHMDB, we follow the multi-task training strategy in [9], due to the limited size of this data set. Specifically, we randomly select 8/8 videos from UCF101 / JHMDB in each mini-batch for multi-task learning. The training losses for UCF101 / JHMDB refer to Eq.(17) / Eq.(19), since human silhouettes annotated in JHMDB can be used for A-A regularization ($\lambda_{AA} = 0.5$ in Eq.(19)). The total loss of multi-task learning is the weighted sum of training losses, where the weight for UCF101 / JHMDB is 0.1/1. Thirdly, at each time step, our spatial-temporal attention is conducted on all sampled frames for UCF101 / HMDB51, while it is performed on the current and one-future-step frame for JHMDB. This is mainly because videos in JHMDB are truncated to a short duration. The sharpness parameters γ_α /

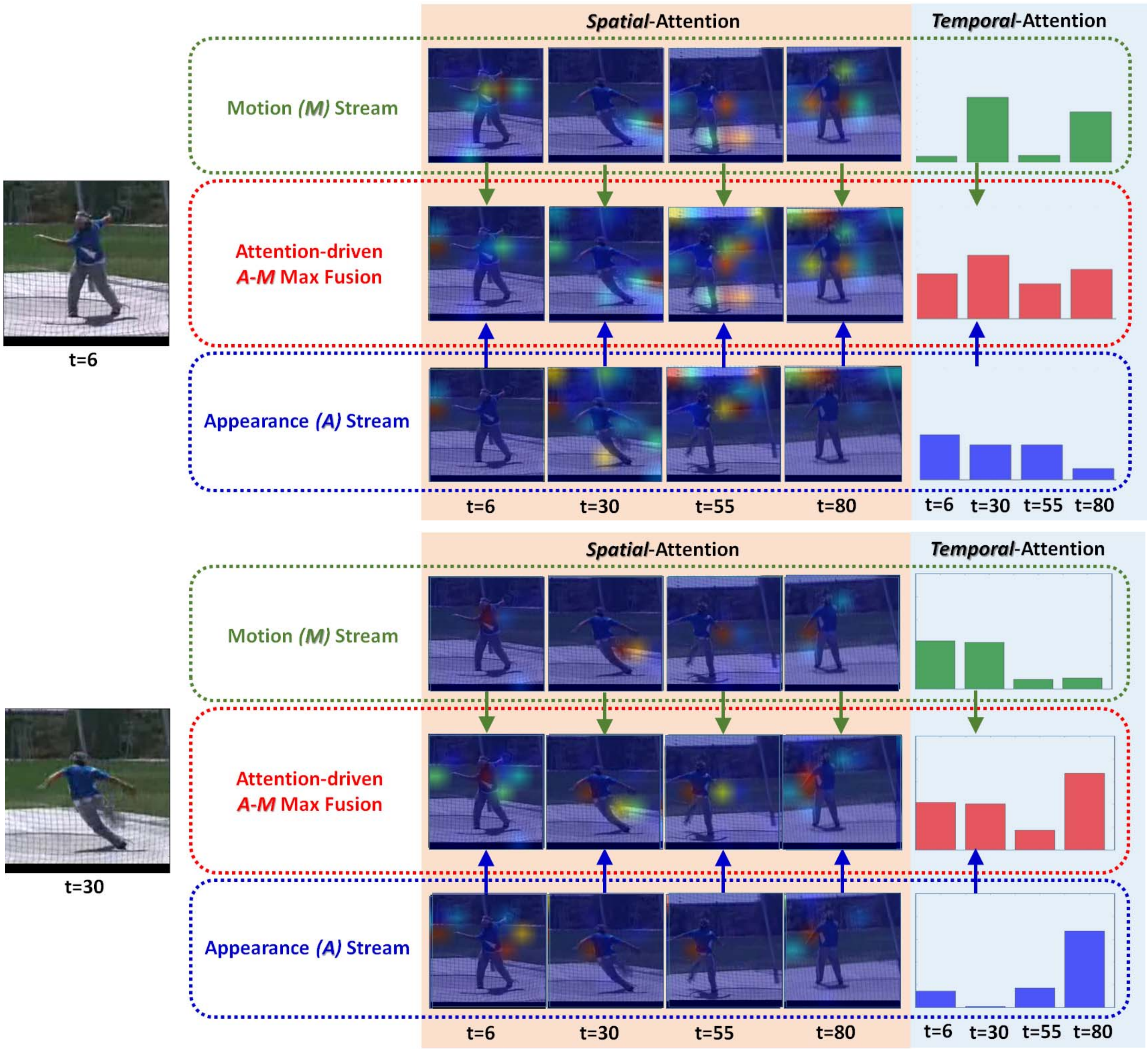


Fig. 7. Illustration of our attention-driven appearance-motion (A-M) fusion for *ThrowDiscus* in UCF101. For the 6th frame, spatial attention maps at the 6th, 30th, 55th, 80th frames of appearance stream are highly-complementary to the ones at the corresponding frames of motion stream. Similarly, temporal attentions in two streams also exhibit the complementary properties. As a result, our attention-driven fusion strategy (Section III-C) can integrate two streams into a unified framework, in order to enhance action representation for the 6th frame. Fusion for the 30th frame is the similar case.

γ_β in our attention are set to five/ten for UCF101 and HMDB51, one/one for JHMDB. Finally, we train our RSTAN with the mini-batch stochastic gradient descent for all the data sets, where the momentum is 0.9, the coefficient for weight decay λ_Θ is 5×10^{-4}, the learning rates for both streams are set to 10^{-2} initially, reduced to 10^{-3} after 12K iterations. The training procedure stops at 20K iterations. We implement our network using theano [63], with multi-GPU parallel implementation of the BPTT algorithm [64].

C. Properties of Our Network

We evaluate a number of important properties in our RSTAN to show its effectiveness. To be fair, when we explore different aspects of one property in our RSTAN, all other properties are with the basic settings as follows. Deep features are extracted from GPCNN [29] (for UCF101 and JHMDB) and TSN [42] (for HMDB51), since the model of GPCNN is not published for HMDB51. Furthermore, we select 32 frames from each testing video with equal sampling interval, where each frame is center-cropped with size of 224×224. We use the last-frame prediction of our recurrent network to report test accuracy (split one for all these data sets).

1) Spatial-Temporal (S-T) Attention: We examine the proposed spatial-temporal (S-T) attention module (Section III-B) on the appearance, motion, and two-stream (appearance+motion) streams respectively. As shown in Table I, our S-T net consistently outperforms the baseline net (LSTMs without attention) on all the data sets. It indicates

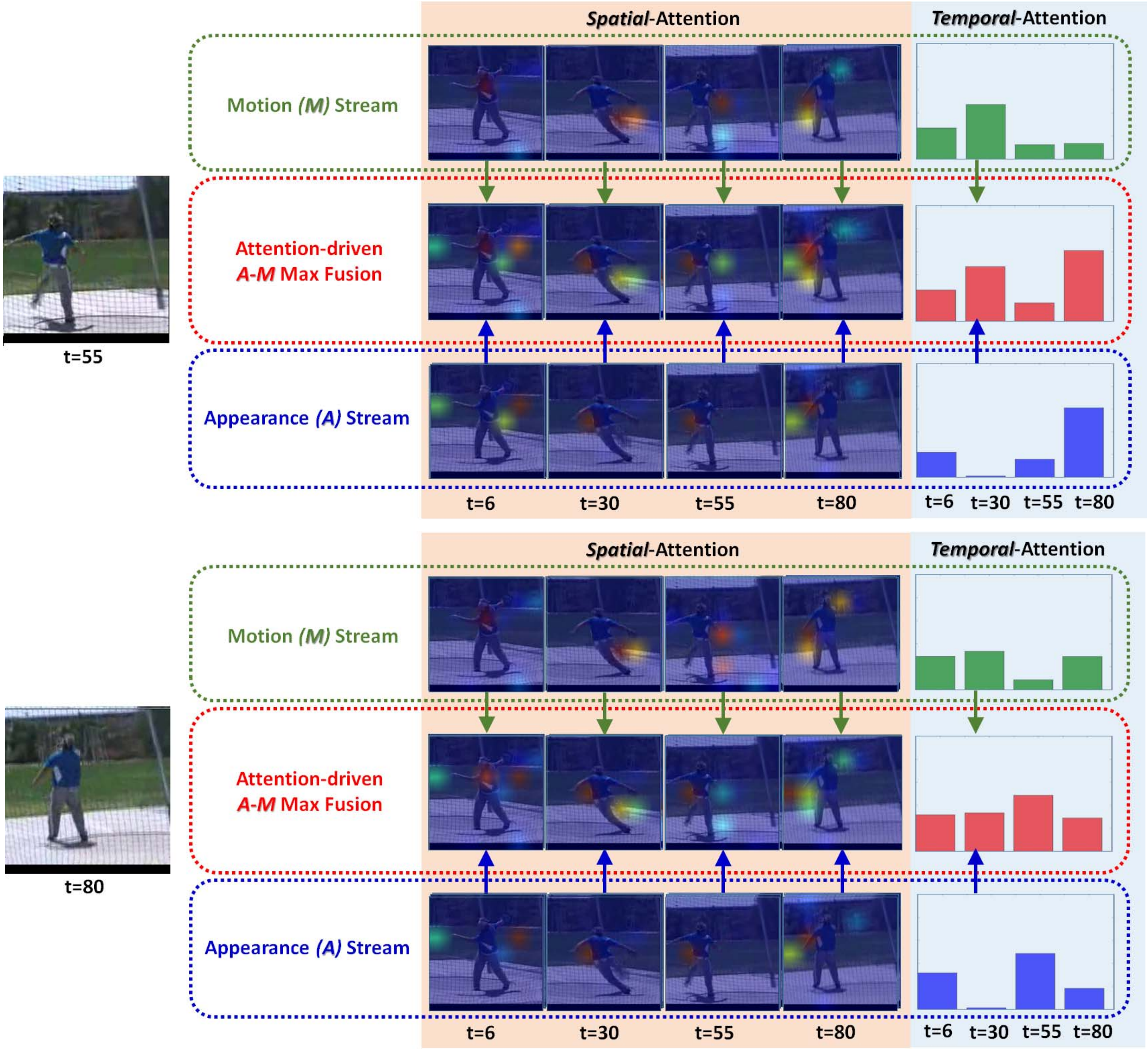


Fig. 8. Illustration of our attention-driven appearance-motion (A-M) fusion for *ThrowDiscus* in UCF101. Similar to Fig. 7, the proposed attention-driven fusion strategy (Section III-C) can integrate two streams into a unified framework, in order to enhance action representation for the 55th and 80th frame.

that *our S-T attention can effectively exploit the detailed spatial-temporal feature from the global context, which is complementary to the high-level input feature vector of LSTM for action representation enhancement*.

2) Attention-Driven Appearance-Motion (A-M) Fusion: We verify our attention-driven appearance-motion (A-M) fusion strategy (Section III-C). In Table II, our *S-T* + *A-M* net outperforms the *S-T* net (with prediction score fusion) on all the data sets. It shows that *our attention-driven A-M fusion can effectively integrate both streams into a unified framework, where one stream can take advantage of the complementary action characteristics from the other stream to improve its discriminative power*. Additionally, we perform A-M fusion with the mean operation for comparison. We find that A-M fusion is robust to different operations (i.e., mean or max). For consistency, we use the max operation in the following experiments.

3) Actor-Attention (A-A) Regularization: We evaluate our actor-attention (A-A) regularization (Section III-D). Note that, we perform A-A regularization for JHMDB, since human silhouettes are annotated in this data set. As a result, we use human silhouettes in the training set of JHMDB to penalize our attention in the model training procedure. In Table II, our *S-T* + *A-M* + *A-A* net outperforms *S-T* + *A-M* net on JHMDB, illustrating that *A-A regularization can guide our attention mechanism to focus on the important action regions around actors*. Furthermore, we perform A-A regularization with cross-entropy loss for comparison. We find that A-A regularization with our L2 loss in Eq. (18) is more effective for classification accuracy improvement.

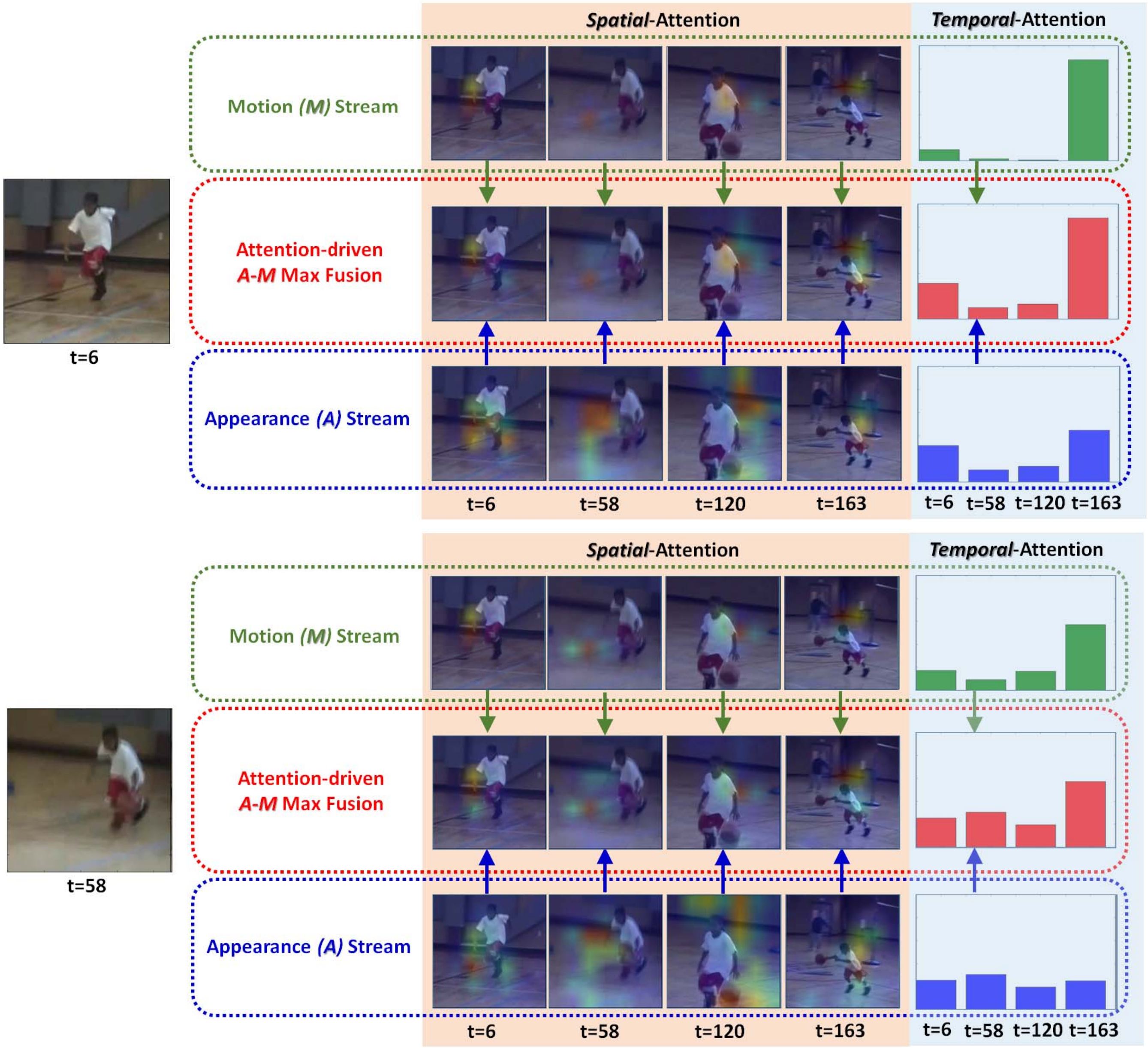


Fig. 9. Illustration of our attention-driven appearance-motion (A-M) fusion for *Dribble* in HMDB51. For the 6th frame, spatial attention maps at the 6th, 58th, 120th, 163th frames of appearance stream are highly-complementary to the ones at the corresponding frames of motion stream. Similarly, temporal attentions in two streams also exhibit the complementary properties. As a result, our fusion strategy (Section III-C) can integrate two streams into a unified framework, in order to enhance action representation for the 6th frame. Fusion for the 58th frame is the similar case.

4) Exploratory Experiments for Attention: We further investigate our spatial-temporal attention from the following perspectives. (1) **Current-Frame-Attention vs. Spatial-Temporal Attention**. we compare our RSTAN with the current-frame-attention net, where the current-frame attention net is our RSTAN but only with current-frame-attention at each time step. In Table III, our RSTAN achieves a better accuracy than the current-frame-attention net, indicating that *current-frame-attention at each step is not sufficient to capture the detailed contextual information of complex actions. On the contrary, our spatial-temporal attention contains important global context to improve the prediction at the current step*. (2) **Sharpness Parameters in Spatial-Temporal Attention**. To evaluate the influence of sharpness parameters in our spatial-temporal attention, we take JHMDB (split one) as example and change γ_α in Eq. (3) and γ_β in Eq. (6) to report classification accuracy. One can see in Fig. 6 that, when γ_α (or γ_β) is increasing, the spatial (or temporal) attention is often concentrated on few important regions (or frames). As a result, the action prediction may be fragmented to decrease accuracy.

D. Comparison With State-of-the-Art

We now evaluate our RSTAN, compared to the state-of-the-art. Since videos in UCF101 and HMDB51 are relatively long, we propose a spatial-temporal multi-scale testing strategy for

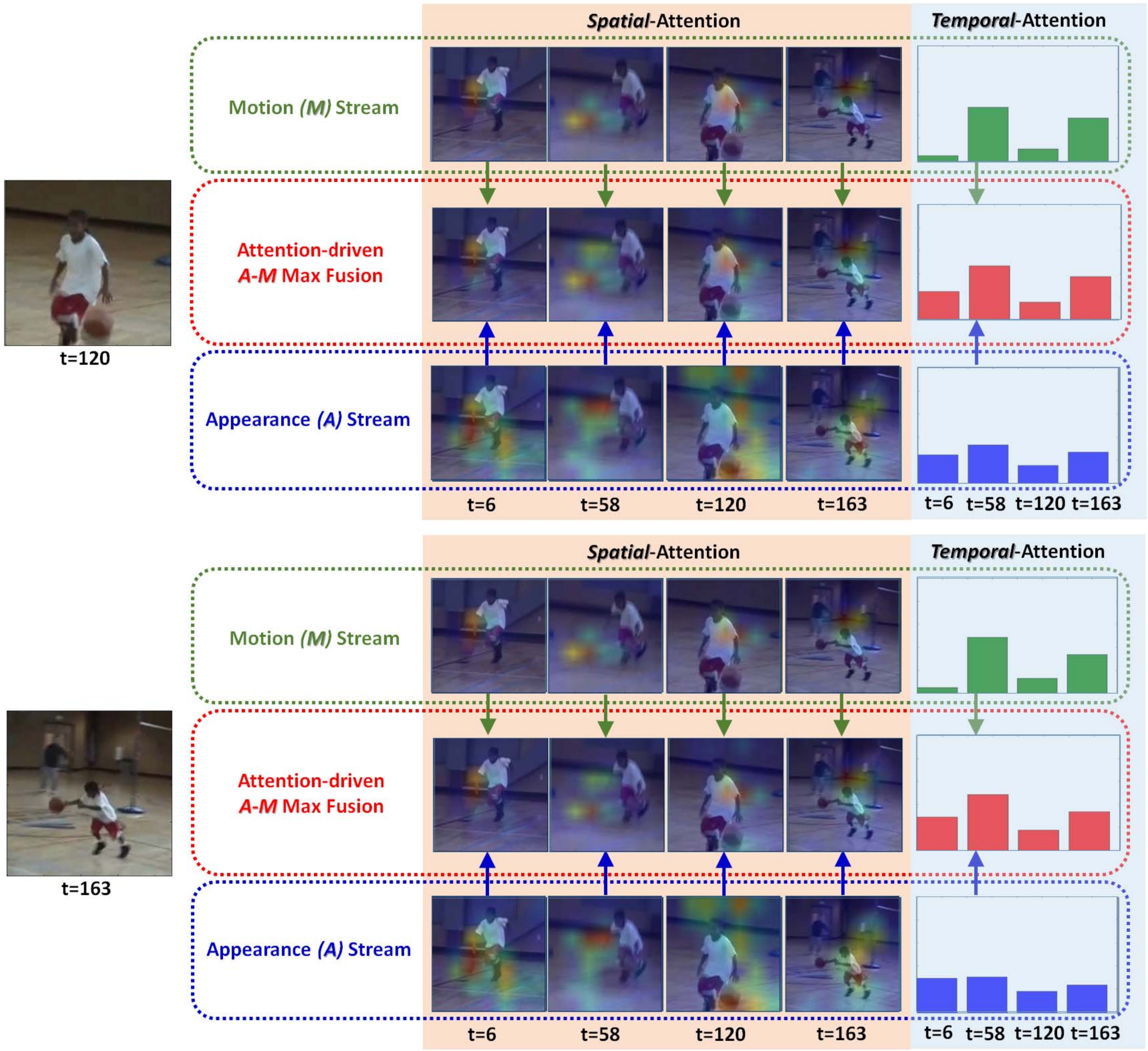


Fig. 10. Illustration of our attention-driven appearance-motion (A-M) fusion for *Dribble* in HMDB51. Similar to Fig. 9, the proposed attention-driven fusion strategy (Section III-C) can integrate two streams into a unified framework, in order to enhance action representation for the 120th and 163th frame.

these two data sets. In the spatial multi-scale strategy, we crop each testing frame from center and four corners with size of 224×224 and the mirror of these crops. and thus there are 10 crops for each testing frame. In the temporal multi-scale strategy, we sample 32 frames with equal interval respectively from the first 2/3 part of video, the last 2/3 part and the whole video of size 256×256, 192×192 with center crops and their mirrors. Hence, there are $10+3\times2\times2 = 22$ sequences for each testing video, and each sequence consists of 32 frames. We compute the score of last-frame prediction for each sequence, and average the scores over all sequences for each testing video in UCF101 and HMDB51. Since videos in JHMDB are relatively short, the test for JHMDB uses the same strategy as the one in Section IV-C.

For UCF101 and HMDB51 in Table IV, our RSTAN outperforms other recent RNN-based approaches, and achieves the comparable results to the state-of-the-art by prediction score fusion with iDT-FV. It indicates that *our RSTAN can identify spatial-temporal contextual information in a recurrent manner, and action representation of our RSTAN is complementary to the hand-crafted spatial-temporal features*. For JHMDB in Table V, our RSTAN achieves the state-of-the-art, compared with recent published works. It illustrates that *our RSTAN can use A-A regularization as guidance to exploit key action regions around actors. Hence, it outperforms other pose-estimation-based or action-detection-based approaches in Table V*.

E. Spatial-Temporal Attention Visualization

We first visualize our attention-driven appearance-motion (A-M) fusion strategy (Section III-C) in Fig. 7-10. One can see that, for a given frame of *ThrowDiscus* (or *Dribble*), spatial attention maps at all sampled frames of appearance stream are highly-complementary to the ones at the corresponding frames of motion stream. Similarly, temporal attentions in two streams also exhibit the complementary properties. Consequently, our fusion strategy can effectively take advantage of important appearance and motion characteristics in a unified framework, to enhance action representation for this given frame. The spatial-temporal attention (after A-M-fusion) for *ThrowDiscus* is also shown in Fig. 1.

We next visualize our actor-attention (A-A) regularization (Section III-D) in Fig. 5. One can see that, attention without A-A regularization may either overestimate the action-irrelevant regions (for the *Pick* action) or underestimate the action-relevant regions (for the *SwingBaseball* action), which could lead to the wrong predictions. On the contrary, attention with A-A regularization can reasonably highlight the important regions around actors for action recognition. Furthermore, it can enhance the relevant contextual information in these regions (such as baseballs in *Swing Baseball*) as well as reduce the influence of irrelevant regions (such as trees and grasses in *Pick*), in order to assist our network to make the correct predictions.

V. Conclusion

In this paper, we designed a recurrent spatial-temporal attention network (RSTAN) for action recognition in videos. First, our spatial-temporal (S-T) attention can leverage the global context for identifying important S-T cues which are strongly-relevant to the current frame. Since our contextual feature is complementary to the FC feature of CNNs, their cooperation can enhance action representation at each time step of LSTM. Second, we developed an attention-driven appearance-motion (A-M) fusion strategy to integrate the two streams into a unified framework, where LSTMs and their S-T attention modules can be jointly trained in an end-to-end fashion. Consequently, one stream can leverage the complementary properties in the other stream to improve its discriminative power. Finally, we proposed an actor-attention (A-A) regularization to guide our attention to focus on important action regions around actors. We evaluated our network on three benchmark data sets for action recognition. The results demonstrated that our RSTAN can outperform other recent RNN-based approaches on UCF101 and HMDB51, and achieve the state-of-the-art on JHMDB.

References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[2] J. Donahue *et al.* (Nov. 2014). "Long-term recurrent convolutional networks for visual recognition and description." [Online]. Available: https://arxiv.org/abs/1411.4389

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. CVPR*, 2014, pp. 1725–1732.

[5] J. Y.-M. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. (Apr. 2015). "Beyond short snippets: Deep networks for video classification." [Online]. Available: https://arxiv.org/abs/1503.08909

[6] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. ICML*, 2015, pp. 843–852.

[7] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. ICCV*, 2015, pp. 4597–4605.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, 2015, pp. 4489–4497.

[9] K. Simonyan and A. Zisserman. (Nov. 2014). "Two-stream convolutional networks for action recognition in videos." [Online]. Available: https://arxiv.org/abs/1406.2199

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Z. Li, E. Gavves, M. Jain, and C. G. M. Snoek. (Jul. 2016). "VideoLSTM convolves, attends and flows for action recognition." [Online]. Available: https://arxiv.org/abs/1607.01794

[12] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. (Jul. 2015). "Every moment counts: Dense detailed labeling of actions in complex videos." [Online]. Available: https://arxiv.org/abs/1507.05738

[13] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. BMVC*, 2008, pp. 275-1–275-10.

[14] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[15] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. ACMMM*, 2007, pp. 357–360.

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, 2011, pp. 3169–3176.

[17] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. ICCV*, 2013, pp. 3551–3558.

[18] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. ECCV*, 2008, pp. 650–663.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[20] X. Peng, L. Wang, X. Wang, and Y. Qiao. (May 2014). "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice." [Online]. Available: https://arxiv.org/abs/1405.4506

[21] Y. Yang, R. Liu, C. Deng, and X. Gao, "Multi-task human action recognition via exploring super-category," *Signal Process.*, vol. 124, pp. 36–44, Jul. 2016.

[22] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. CVPR*, 2015, pp. 4305–4314.

[23] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao, "Latent max-margin multitask learning with skelets for 3-D action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 439–448, Feb. 2017.

[24] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3D action recognition," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 519–529, Mar. 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun. (Dec. 2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[27] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[28] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[29] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. (Jul. 2015). "Towards good practices for very deep two-stream convnets." [Online]. Available: https://arxiv.org/abs/1507.02159

[30] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.

[31] Z. Lan, Y. Zhu, and A. G. Hauptmann. (Jan. 2017). "Deep local video feature for action recognition." [Online]. Available: https://arxiv.org/abs/1701.07368

[32] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. C. Russell. (Apr. 2017). "ActionVLAD: Learning spatio-temporal aggregation for action classification." [Online]. Available: https://arxiv.org/abs/1704.02895

[33] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *Proc. CVPR*, 2016, pp. 1991–1999.

[34] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. CVPR*, 2016, pp. 1933–1941.

[35] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. (May 2017). "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection." [Online]. Available: https://arxiv.org/abs/1704.00616

[36] A. Kar, N. Rai, K. Sikka, and G. Sharma. (Nov. 2016). "AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos." [Online]. Available: https://arxiv.org/abs/1611.08240

[37] J. Wang, A. Cherian, and F. Porikli. (Apr. 2017). "Ordered pooling of optical flow sequences for action recognition." [Online]. Available: https://arxiv.org/abs/1701.03246

[38] J. Wang, A. Cherian, F. Porikli, and S. Gould. (Apr. 2017). "Action representation using classifier decision boundaries." [Online]. Available: https://arxiv.org/abs/1704.01716

[39] A. Cherian, P. Koniusz, and S. Gould. (Jan. 2017). "Higher-order pooling of CNN features via kernel linearization for action recognition." [Online]. Available: https://arxiv.org/abs/1701.05432

[40] A. Cherian, B. Fernando, M. Harandi, and S. Gould. (Jul. 2017). "Generalized rank pooling for activity recognition." [Online]. Available: https://arxiv.org/abs/1704.02112

[41] A. Cherian and S. Gould. (Apr. 2017). "Second-order temporal pooling for action recognition." [Online]. Available: https://arxiv.org/abs/1704.06925

[42] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*, 2016, pp. 20–36.

[43] A. Diba, V. Sharma, and L. Van Gool. (Nov. 2016). "Deep temporal linear encoding networks." [Online]. Available: https://arxiv.org/abs/1611.06678

[44] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. NIPS*, 2016, pp. 3468–3476.

[45] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. (Mar. 2017). "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition." [Online]. Available: https://arxiv.org/abs/1703.10667

[46] D. Bahdanau, K. Cho, and Y. Bengio. (Sep. 2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[47] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, 2014, pp. 1764–1772.

[48] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015, pp. 3156–3164.

[49] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[50] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. ICCV*, 2015, pp. 4507–4515.

[51] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. ACMM*, 2015, pp. 461–470.

[52] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. ICLR*, 2016, pp. 1–2.

[53] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. (Apr. 2017). "Two stream LSTM: A deep fusion framework for human action recognition." [Online]. Available: https://arxiv.org/abs/1704.01194

[54] S. Sharma, R. Kiros, and R. Salakhutdinov. (Nov. 2015). "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119

[55] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li. (Jul. 2016). "Hierarchical attention network for action recognition in videos." [Online]. Available: https://arxiv.org/abs/1607.06416

[56] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. ICLR*, 2017, pp. 1–17.

[57] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregle, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. NIPS*, 2014, pp. 1799–1807.

[58] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proc. ICCV*, 2015, pp. 1913–1921.

[59] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Proc. ECCV*, 2014, pp. 581–595.

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

[61] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. ICCV*, 2011, pp. 2556–2563.

[62] H. Jhuang, J. Gall, S. Zuffi, M. J. Black, and C. Schmid, "Towards understanding action recognition," in *Proc. ICCV*, 2013, pp. 3192–3199.

[63] T. D. Team. (May 2016). "Theano: A Python framework for fast computation of mathematical expressions." [Online]. Available: https://arxiv.org/abs/1605.02688

[64] W. Ding, R. Wang, F. Mao, and G. Taylor. (Dec. 2014). "Theano-based large-scale visual recognition with multiple GPUs." [Online]. Available: https://arxiv.org/abs/1412.2302

[65] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. CVPR*, 2016, pp. 3034–3042.

[66] A. Diba, A. M. Pazandeh, and L. Van Gool. (2016). "Efficient two-stream motion and appearance 3D CNNs for video classification." [Online]. Available: https://arxiv.org/abs/1608.08851

[67] G. Varol, I. Laptev, and C. Schmid. (Apr. 2016). "Long-term temporal convolutions for action recognition." [Online]. Available: https://arxiv.org/abs/1604.04494

[68] Z. Qiu, T. Yao, and T. Mei. (Nov. 2016). "Deep quantization: Encoding convolutional activations with deep generative model." [Online]. Available: https://arxiv.org/abs/1611.09502

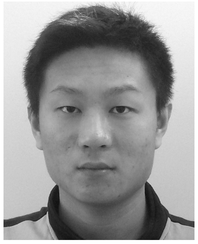
[69] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. (Oct. 2017). "Hidden two-stream convolutional networks for action recognition." [Online]. Available: https://arxiv.org/abs/1704.00389

[70] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. CVPR*, 2015, pp. 759–768.

[71] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. ICCV*, 2015, pp. 3218–3226.

[72] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. ECCV*, 2016, pp. 744–759.

Wenbin Du received the M.S. degree in software engineering from Donghua University, Shanghai, China, in 2014. He is currently pursuing the Ph.D. degree with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests are computer vision and deep learning.



Yali Wang received the Ph.D. degree in computer science from Laval University, Canada, in 2014. He is currently an Assistant Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests are deep learning and computer vision, machine learning, and statistics.

Yu Qiao (SM'13) received the Ph.D. degree from the University of Electro-Communications, Japan, in 2006. He was a JSPS Fellow and a Project Assistant Professor with the University of Tokyo, from 2007 to 2010. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He has authored over 140 papers in journals and conference including, PAMI, IJCV, TIP, ICCV, CVPR, ECCV, and AAAI. His research interests include computer vision, deep learning, and intelligent robots. He was a recipient of the Lu Jiaxi Young Researcher Award from the Chinese Academy of Sciences in 2012. He was the first Runner-Up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition and the recipient at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification.