Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Attention with Structure Regularization for Action Recognition

Yuhui Quan[a,c], Yixin Chen[a], Ruotao Xu[a,**], Hui Ji[b]

[a] *School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China*
[b] *Department of Mathematics, National University of Singapore, Singapore 119076, Singapore*
[c] *Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, China.*

ABSTRACT

Recognizing human action in video is an important task with a wide range of applications. Recently, motivated by the findings in human visual perception, there have been numerous attempts on introducing attention mechanisms to action recognition systems. However, it is empirically observed that an implementation of attention mechanism using attention mask of free form often generates ineffective distracted attention regions caused by overfitting, which limits the benefit of attention mechanisms for action recognition. By exploiting block-structured sparsity prior on attention regions, this paper proposed an $\ell_{2,1}$-norm group sparsity regularization for learning structured attention masks. Built upon such a regularized attention module, an attention-based recurrent network is developed for action recognition. The experimental results on two benchmark datasets showed that, the proposed method can noticeably improve the accuracy of attention masks, which results in performance gain in action recognition.

1. Introduction

Human action recognition is a challenging yet important task which has been receiving increasing attention in recent years. Recognizing human actions is about identifying human activities (*e.g.* running, walking, or dancing) in video sequences or images. It is an essential tool that enables effective analysis on human behaviors as well as efficient interactions between humans and vision systems. Thus, human action recognition can see its usage in a wide range of applications, including surveillance, video retrieval, human activity prediction, content-based summarization, electronic entertainment, automated cinematography, and many others. See *e.g.* (Moeslund et al., 2006; Poppe, 2010) for more discussions. Meanwhile, human action recognition is also a very challenging task due to significant variations in human actions, in terms of personal styles, human appearance, camera viewpoints, varying background and other environmental changes.

In the past decades, there has been an enduring effort on the development of effective action recognition systems, and they are quite successful under well-controlled environments (Poppe, 2010). However, for action recognition in a single video sequence taken under unconstrained scenarios, it remains a challenging problem with limited success. Over the past years, many manually-crafted features have been proposed for action recognition to exploit various types of cues. To name a few, human poses (Lv and Nevatia, 2007; Thurau and Hlavác, 2008; Raptis and Sigal, 2013), skeletons from depth cameras (Wang et al., 2012; Du et al., 2015), local space-time patterns (Niebles et al., 2008; Yeffet and Wolf, 2009; Scovanner et al., 2007), trajectories of interest points (Wang et al., 2011; Wang and Schmid, 2013), motion patterns from optical flow (Fathi and Mori, 2008; Laptev et al., 2008; Carreira and Zisserman, 2017), and additional features from external non-visual cues such as video attributes (Yao et al., 2011) and movie scripts (Laptev et al., 2008). To further improve the performance in complex scenarios, these features are often combined together to achieve a better representation; see *e.g.* (Wang and Schmid, 2013; Bilen et al., 2016; Wang et al., 2016; Feichtenhofer et al., 2016a; Carreira and Zisserman, 2017).

More recently, with great advance in deep learning, there has been rapid progress on applying deep neural network to solve

[**] Corresponding author
e-mail: csyhquan@scut.edu.cn (Yuhui Quan), yx.chen.cs@foxmail.com (Yixin Chen), xu.ruotao@mail.scut.edu.cn (Ruotao Xu), matjh@nus.edu.sg (Hui Ji)

the problem of action recognition. Most existing studies adopt two types of neural networks architectures. One is Convolutional Neural Network (CNN) (Ji et al., 2013; Simonyan and Zisserman, 2014; Feichtenhofer et al., 2016b), and the other is Recurrent Neural Network (RNN) (Baccouche et al., 2011; Wang et al., 2012; Du et al., 2015; Zhu et al., 2016). By replacing manually-crafted features using adaptive features learned from data, these neural-network-based approaches for action recognition showed impressive improvement over traditional approaches. Regarding action recognition in videos, the RNN with Long-Short Term Memory (LSTM) cells (Yeung et al., 2015; Du et al., 2015; Zhu et al., 2016) is particularly appealing, as it allows the NN to exploit one crucial property of human actions, *i.e.* long-term dependency of temporal patterns of actions (Duchenne et al., 2009; Hoai and Zisserman, 2014; Fernando et al., 2016). Nevertheless, it is empirically observed in many studies (Feichtenhofer et al., 2016a; Carreira and Zisserman, 2017) that despite its theoretical advantage on capturing the long-term temporal dependencies of human action, LSTM network does not lead to noticeable performance gain over other neural network architectures,

One solution proposed for exploiting the potential of LSTM networks in action recognition is the introduction of attention mechanism; see *e.g.* (Sharma et al., 2015; Yang et al., 2018; Yan et al., 2017; Girdhar and Ramanan, 2017; Wang et al., 2017; Du et al., 2018). Such a solution is motivated by the findings that attention plays a pervasive role in human visual perception in cluttered scenes; see *e.g.* (Itti and Koch, 2001; Sun, 2008), which enables human to focus more on the objects of interest and omit irrelevant background. Such a mechanism certainly makes the task of object recognition in cluttered scenes much easier.

The possible performance gain of a built-in attention module to a LSTM network in video action recognition largely depends on how accurately the introduced attention module can identify the regions of interest relevant to the subject and target of an action. Or equivalently, it depends on whether the attention module in a trained neural network can be generalized well. Current approaches of introducing the attention module to the LSTM network for action recognition are done by adding a soft attention mask to each temporally-recurrent layer in the network, and the masks are completely determined by the training data without any constraint; see *e.g.* (Sharma et al., 2015). Unfortunately, for action recognition in a single video, significant variations of regions of interest, as well as limited amount of available training data, make it challenging to train an attention module with good generalization. It is empirically observed that the attention module trained using a reasonable amount of data often generate diffusing attention masks on test data, as shown in Fig. 4.

This paper aims at developing a LSTM network with a new attention module to address the weakness of existing approaches listed above. The idea is to introduce a spatially structure prior to the attention module with the motivation from human visual perception. In human visual perception, the recognition of a human action can be done by only focusing on certain key parts of human body or of the target of the action. For instance, we can easily identify the action of walking by looking at the feet. See Fig. 1 for some illustrations. In other words, the regions of attention that facilitate action recognition can be assumed to be those connected regions.

More specifically, we propose an approach for regularizing the attention mechanism in a LSTM network by a block-wise sparsity prior on attention masks, *i.e.*, only few blocks of an attention mask are activated in the attention module of network. The introduction of regularized attention mechanism allows the LSTM network to be trained with better generalization, as shown in the experiments. It is empirically observed that with the proposed structure regularization, the generation of attention masks is more accurate on the locations of action instances as it is less sensitive to cluttered backgrounds and reduce the possibility of gating the salient patterns associated with certain actions in long video sequences. It is noted that in addition to the LSTM network, the proposed mechanism of structured attention can also be directly plugged into other architectures of RNNs for further performance improvement on action recognition.

The rest of this paper is organized as follows. Section 2 gives a brief review on existing approaches to action recognition. The LSTM network with a structure-regularized attention mechanism for action recognition is presented in Section 3. Section 4 is devoted to experimental evaluation of the proposed method. Section 5 concludes the paper.

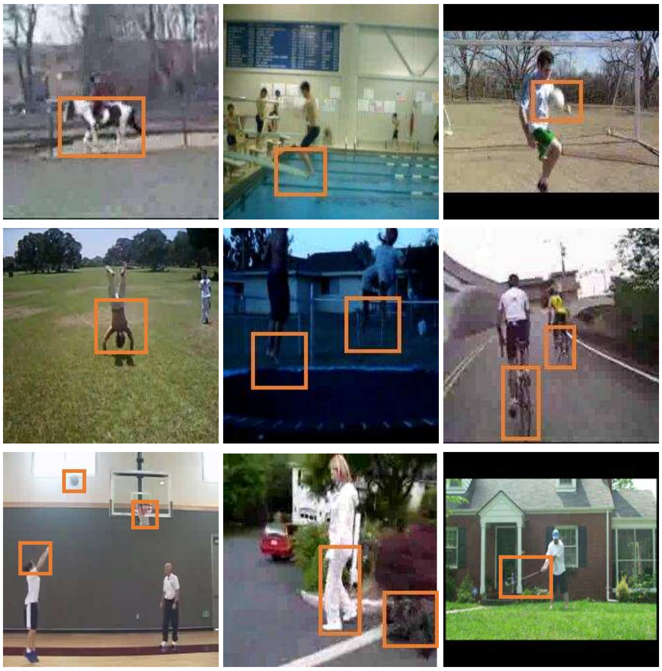


Fig. 1. Illustration of the block-structured sparsity prior on the attention regions in videos. It can be seen that the recognition of a human action can be done by only focusing on certain key parts of human body or of the target of the action.

2. Related Work

Action recognition has been extensively studied in past. As this paper is about action recognition in a single RGB video sequence captured by conventional camera, our discussion focuses on the methods that work in the same setting. Other methods that work for videos from multiple cameras or depth cameras are not covered here, owing to space limitation.

Most early methods follow the same framework as that of traditional image classification, and generalize the local image features (*e.g.*bag of visual words) in spatial domain to spatial-temporal domain. For example, histograms of oriented space-time gradients (Scovanner et al., 2007; Klaser et al., 2008), 3D Harris (Laptev et al., 2008) and local trinary patterns (Yeffet and Wolf, 2009). Recently, by tracing each feature point in a short time interval and using the trajectories of feature points as local descriptors, the trajectory-based methods (Wang et al., 2011; Wang and Schmid, 2013) showed their performance gain over those traditional features. All the aforementioned methods share the same weakness, *i.e.* spatiotemporal configuration of feature points is not well exploited. To address this weakness, many approaches have been proposed to use generative models to characterize human actions, *e.g.* constellation models (Fanti et al., 2005; Niebles et al., 2008) and hidden Markov models (Wang and Mori, 2011; Tang et al., 2012). More recently, generative models with hierarchical structures are more preferred (*e.g.* (Wang and Mori, 2011; Lan et al., 2015)), due to its capability in encoding human actions with multi-level granularity.

Motivated by great advance on deep learning and its success in many vision tasks, several neural network-based approaches have been proposed for action recognition in recent years. In (Baccouche et al., 2011; Ji et al., 2013), a video clip is treated as a 3D volume, and then the CNN designed for 2D image classification is modified from processing 2D arrays to handling 3D volumes done by introducing 3D convolutions in spatiotemporal domain. Such a 3D-CNN architecture does not scale well with large-scale data. To speed up training process, a multi-resolution CNN architecture is introduced in (Karpathy et al., 2014). The performance of these generic spatiotemporal CNNs on action recognition is not satisfactory, even in comparison to those recent methods that are built upon handcrafted features (*e.g.* (Wang and Schmid, 2013)). Indeed, their performance is close to those CNNs defined in the spatial domain (Feichtenhofer et al., 2016b). This indicates that temporal cues are not well exploited in those approaches.

To exploit temporal cues of actions in videos, a two-stream architecture is proposed in (Simonyan and Zisserman, 2014), which decouples the spatiotemporal CNN into a spatial stream on each single frame and a temporal stream on optical flow across frames, and then fuses the classification score in the final layer. Such a decoupling scheme allows independently learning temporal structures from optical flow. In (Feichtenhofer et al., 2016a), the residual connections are injected into the two-stream CNN for further improvement. Instead of factorizing the network, the C3D method (Tran et al., 2015) constructs a very deep CNN by factorizing 3D convolution into a 2D spatial convolution and a 1D temporal convolution, which yields much better results than that from (Baccouche et al., 2011; Ji et al., 2013; Karpathy et al., 2014). In (Wang et al., 2016), the Siamese network architecture with convolution layers is used to model the relationship between actions and effects. While they do well on discovering local patterns within a small time window, such structured spatiotemporal CNNs ignore the long-term motion patterns which may be crucial for identifying certain types of actions, *e.g.*the action with periodic motion. The rank pooling (Bilen et al., 2016; Fernando et al., 2015, 2017) developed for spatiotemporal CNNs addressed this problem by firstly learning a ranker to rank video frames in feature space and secondly using the learned parameters of the ranker as the representation of video.

Another approach of exploiting long-term motion patterns in video is to adopt RNN with the LSTM units (Hochreiter and Schmidhuber, 1997). The resulting LSTM network is capable of discovering long-range temporal patterns by storing, modifying and accessing internal states of network layers. In the early work (Baccouche et al., 2010), the LSTM network with BoVW features as input is used to analyze the actions in soccer video. In (Zhu et al., 2016; Du et al., 2015), a deep LSTM network is applied to skeleton-based action recognition with good performance, but it is not applicable to generic RGB video. For action recognition in generic RGB video, a LSTM network is used in (Baccouche et al., 2011), together with a sequence of spatiotemporal CNN descriptors extracted from each frame. In (Yue-Hei Ng et al., 2015), a LSTM network is used, together with the aforementioned two-stream CNN (Simonyan and Zisserman, 2014). In parallel, an end-to-end network is proposed in (Donahue et al., 2015) which concatenates a deep LSTM network to a CNN on raw data, which showed impressive results without using any auxiliary input, *e.g.*optical flow used in (Yue-Hei Ng et al., 2015). The conventional LSTM networks cannot guarantee memorizing discriminative motion patterns. Thus, a differential gating scheme is proposed for the LSTM neural network in (Veeriah et al., 2015), which emphasizes the changes of information gain caused by the salient motions in successive video frames. The LSTM network can also be used in an unsupervised setting. See (Srivastava et al., 2015) for an auto-encoder inspired LSTM architecture for unsupervised learning of temporal features.

The performance of the LSTM networks on action recognition can be further improved by introducing attention mechanism (*e.g.* (Sharma et al., 2015; Yang et al., 2018; Yan et al., 2017; Girdhar and Ramanan, 2017; Wang et al., 2017; Du et al., 2018)), which is also the focus of this paper. The very related works are (Yeung et al., 2015; Sharma et al., 2015). In (Yeung et al., 2015), a dense action labeling is done using a temporal accumulated attention model on the input-output context. In (Sharma et al., 2015), an attention mechanism is introduced into the LSTM network, implemented by the soft masks that weight spatial locations using a softmax layer. These methods do not consider spatially structure prior of attention masks, which indeed is very important for neural network trained for action recognition to have good generalization.

3. Main body

In this paper, we propose a neural network with a structure-regularized attention mechanism for action recognition in a single video. The proposed neural network is composed of three

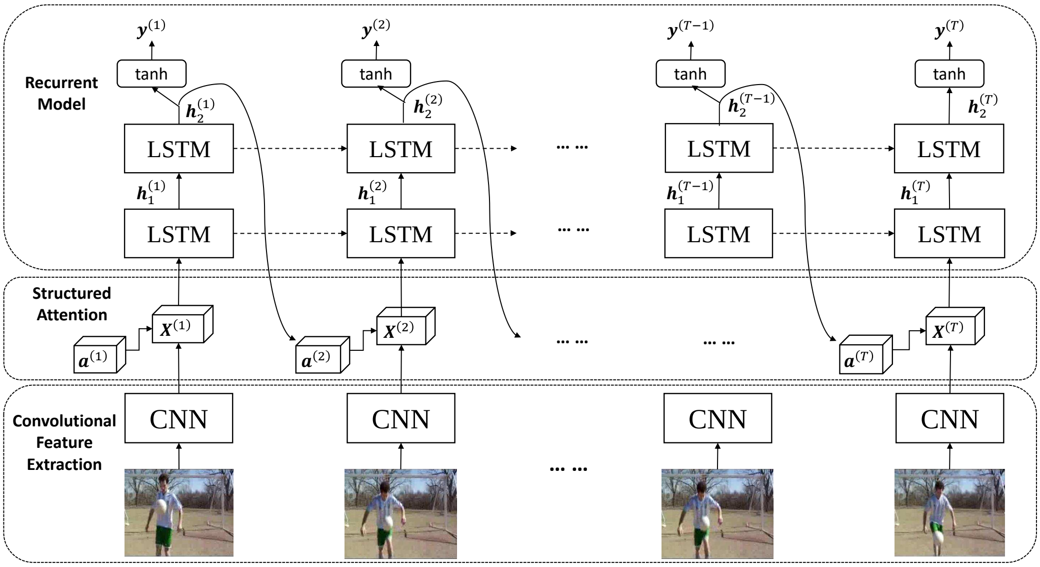


Fig. 2. Framework of the proposed model. The CNN module (bottom) is used as a feature extractor to yield the feature $X^{(t)}$ of each video frame $V^{(t)}$ for $t = 1, \cdots, T$. Thereafter, a module of structured attention (middle) attenuates the features of each frame by element-wise multiplication with the soft attention mask $a^{(t)}$. The RNN module (upper) takes the feature $X^{(t)}$ for all t as input and predicts the label of each video frame by analyzing sequential (temporal) patterns of the features.

modules: CNN, RNN, and attention. The CNN module is used as a feature extractor for each video frame. The RNN module takes the CNN features as input and extracts sequential (temporal) information from the features. The regularized attention mechanism is introduced such that the attention of the network concentrates on most relevant pivotal regions. See Fig. 2 for the diagram of the proposed method.

The notations used in the paper are as follows. Bold upper letters are used for matrices, bold lower letters for column vectors, light upper letters for constants, light lower letters for scalars, and both hollow letters and calligraphy letters for sets.

3.1. Feature extraction by CNN

In our approach, the first step is mapping video frames to some feature space for better representation. In this step, we focus on the spatial features of videos, and the mapping is done by extracting image features on each video frame. Motivated by classification performance of GoogLeNet (Szegedy et al., 2016) model on ImageNet dataset (Deng et al., 2009), we adopt GoogleNet for the task of feature extraction. Consider a video $\mathcal{V} = \{\boldsymbol{V}^{(t)}\}_{t=1}^{T}$ with T successive frames. The feature extraction process is done by feeding each $\boldsymbol{V}^{(t)}$ to the GoogLeNet and using as the spatial features the feature maps output by the last convolutional layer of the GooLeNet.

Suppose there are D feature maps extracted from the last convolutional layer of GooLeNet, each of which is of size $K \times K$. Then, for the video frame $\boldsymbol{V}^{(t)}$, we have a feature cube of size $K \times K \times D$, which is expressed in the matrix form for convenience:

$$\boldsymbol{V}^{(t)} \longrightarrow \boldsymbol{X}^{(t)} = [\boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_{K^2}^{(t)}], \tag{1}$$

for $t = 1, \cdots, T$, where $\boldsymbol{x}_i^{(t)} \in \mathbb{R}^D$ denotes the coefficient vector formed by concatenating of the coefficients of each feature map on the i^{th} spatial location at the t^{th} time step. Thereafter, we use the attention mechanism presented in Section 3.3 to pool the feature cube $\boldsymbol{X}^{(t)} \in \mathbb{R}^{D\times K^2}$ into a frame feature $\boldsymbol{x}^{(t)} \in \mathbb{R}^D$. Then, $\boldsymbol{x}^{(t)}$ is fed to the RNN as the input in the next step.

3.2. Recurrent model

Taking the spatial feature of each video frame as input, we construct a recurrent network to capture the temporal dynamics among video frames. In order to extract the long-range temporal patterns of an action, a bi-level LSTM network (Zaremba et al., 2014) is adopted in our implementation. Let $\sigma(\cdot)$ denote the sigmoid function and $\odot$ denote the element-wise multiplication. Taking the features $\{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ extracted from the previous step as the input, the bi-level LSTM network is defined by the recurrent form as follows.

$$\begin{pmatrix} \boldsymbol{i}_1^{(t)} \\ \boldsymbol{f}_1^{(t)} \\ \boldsymbol{o}_1^{(t)} \\ \boldsymbol{g}_1^{(t)} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{P}_1 \begin{pmatrix} \boldsymbol{h}_1^{(t-1)} \\ \boldsymbol{x}^{(t)} \end{pmatrix}, \quad \begin{pmatrix} \boldsymbol{i}_2^{(t)} \\ \boldsymbol{f}_2^{(t)} \\ \boldsymbol{o}_2^{(t)} \\ \boldsymbol{g}_2^{(t)} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{P}_2 \begin{pmatrix} \boldsymbol{h}_2^{(t-1)} \\ \boldsymbol{h}_1^{(t)} \end{pmatrix}, \tag{2}$$

$$\boldsymbol{c}_1^{(t)} = \boldsymbol{f}_1^{(t)} \odot \boldsymbol{c}_1^{(t-1)} + \boldsymbol{i}_1^{(t)} \odot \boldsymbol{g}_1^{(t)}, \tag{3}$$

$$\boldsymbol{c}_2^{(t)} = \boldsymbol{f}_2^{(t)} \odot \boldsymbol{c}_2^{(t-1)} + \boldsymbol{i}_2^{(t)} \odot \boldsymbol{g}_2^{(t)}, \tag{4}$$

$$\boldsymbol{h}_1^{(t)} = \boldsymbol{o}_1^{(t)} \odot \tanh(\boldsymbol{c}_1^{(t)}), \tag{5}$$

$$\boldsymbol{h}_2^{(t)} = \boldsymbol{o}_2^{(t)} \odot \tanh(\boldsymbol{c}_2^{(t)}), \tag{6}$$

where $\boldsymbol{i}_k^{(t)}, \boldsymbol{f}_k^{(t)}, \boldsymbol{o}_k^{(t)}, \boldsymbol{g}_k^{(t)}, \boldsymbol{c}_k^{(t)}, \boldsymbol{h}_k^{(t)} \in \mathbb{R}^M$ denote the input gate, forget gate, output gate, intermediate result, cell state, and hidden state on the k^{th} ($k = 1, 2$) level respectively. The transform $\boldsymbol{P}_k \in \mathbb{R}^{4M\times(M+D)}$, $k = 1, 2$ denotes an affine transformation consisting of trainable parameters, and t denotes the time step.

At each time step, the LSTM network predicts the soft label $\boldsymbol{y}^{(t)} = [y_1^{(t)}, \cdots, y_C^{(t)}] \in \mathbb{R}^C$ regarding C action classes on the t^{th}

frame, by using the additional layers defined by

$$y_i^{(t)} = \frac{\exp(\tanh(\boldsymbol{b}_i^\top \boldsymbol{h}_2^{(t)}))}{\sum_{j=1}^{C} \exp(\tanh(\boldsymbol{b}_j^\top \boldsymbol{h}_2^{(t)}))}, \qquad i = 1, \cdots, C, \tag{7}$$

where $\boldsymbol{b}_i \in \mathbb{R}^M$ denotes the trainable parameters of the linear classifier on the i^{th} class.

3.3. Attention mechanism regularized by spatially structured-sparsity prior

Recall that in the stage of feature extraction described in Section 3.1, for each frame, the feature $\boldsymbol{x}^{(t)}$ is obtained from the feature cube $\boldsymbol{X}^{(t)}$ via the so-called attention mechanism:

$$\text{Attention}: \quad \boldsymbol{X}^{(t)} \in \mathbb{R}^{D \times K^2} \longrightarrow \boldsymbol{x}^{(t)} \in \mathbb{R}^D.$$

The introduction of such an attention mechanism is motivated from the findings in human visual perception, i.e., when recognizing the action of a specific object, human tends to focus his/her attention on that object while omitting the other less relevant parts, such as the background. In other words, for efficient action recognition, different regions of the image have different contributions to the features. Such a mechanism can be incorporated into the neural network by introducing the so-called *attention masks* defined on the original features derived from the image.

In the vector form, let $\boldsymbol{a}^{(t)} = [a_1^{(t)}, \cdots, a_{K^2}^{(t)}] \in \mathbb{R}^{K^2}$ denote the soft attention mask with respect to a $K \times K$ feature map in the spatial domain at the t^{th} time step. The soft mask can also be viewed as the relevance indicator with respect to the feature in spatial domain. Each attention degree $a_i^{(t)}$ in the attention mask is predicted according to $\boldsymbol{h}_2^{(t-1)}$, *i.e.* the hidden state at the last time step, through the softmax function:

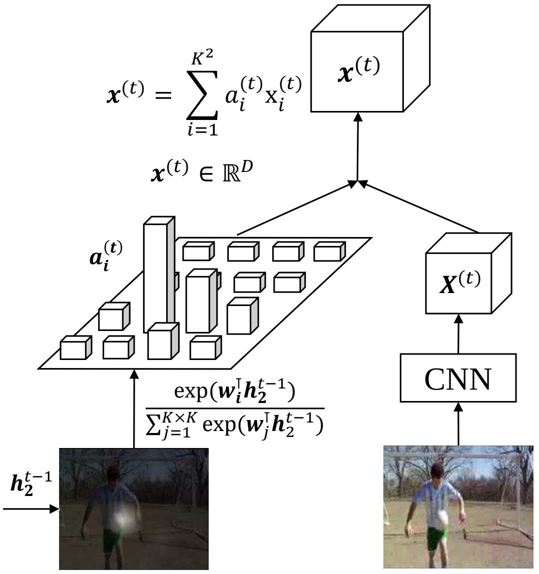
$$a_i^{(t)} = \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{h}_2^{(t-1)})}{\sum_{j=1}^{K \times K} \exp(\boldsymbol{w}_j^\top \boldsymbol{h}_2^{(t-1)})}, \qquad i = 1, \cdots, K^2, \tag{8}$$

where $\boldsymbol{w}_i \in \mathbb{R}^{K^2}$ denotes the transform with learnable weights that maps the last hidden state to a proper representation at the i^{th} spatial location. It can be seen that the entries of $\boldsymbol{a}_i^{(t)}$ are normalized by the softmax function and thus can be viewed as the relevance of the input frame to the recognition. After calculating the attention mask $\boldsymbol{a}^{(t)}$, the feature $\boldsymbol{x}^{(t)}$ of the t^{th} video frame is computed by the average of the feature vectors over $K \times K$ spatial locations weighted by the attention mask $\boldsymbol{a}^{(t)}$ as follows.

$$\boldsymbol{x}^{(t)} = \sum_{i=1}^{K^2} a_i^{(t)} \boldsymbol{x}_i^{(t)} = \boldsymbol{X}^{(t)} \boldsymbol{a}^{(t)}. \tag{9}$$

In our implementation, $\boldsymbol{a}^{(1)}$ is initialized as $\frac{1}{K^2}$, as each spatial location is equally relevant when no observation is available. See Figure 3 for the diagram of the attention mechanism.

The attention mechanism can be interpreted as a dynamic weighted average pooling, where the weights are predicted according to the hidden state of the LSTM network at the last time step. The attention mechanism can also be understood as a simulation of the same mechanism in human visual perception.



$\boldsymbol{x}^{(t)} = \sum_{i=1}^{K^2} a_i^{(t)} \mathrm{x}_i^{(t)}$
$\boldsymbol{x}^{(t)}$
$\boldsymbol{x}^{(t)} \in \mathbb{R}^D$
$\boldsymbol{a}_i^{(t)}$
$\boldsymbol{X}^{(t)}$
CNN
$\frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{h}_2^{t-1})}{\sum_{j=1}^{K \times K} \exp(\boldsymbol{w}_j^\top \boldsymbol{h}_2^{t-1})}$
$\boldsymbol{h}_2^{t-1}$

Fig. 3. Illustration of diagram of the proposed attention mechanism. The attention mask *a* is used as the weights for pooling the feature cube *X* to a feature vector *x*.

For action recognition, human attention usually concentrates on very related regions. For example, the attention should be focused on the football in juggling, and the bike should be paid more attention to in cycling.

However, overfitting is likely to happen when training a RNN with attention masks, especially compared to those without attention mechanisms. Intuitively, the overfitting of a network with attention may be reflected by one of the following phenomena of the attention masks: incorrect focus placed on the objects irrelevant to the action, or distracted attention with scattering non-zero entries of the mask. It is empirically observed that the second phenomenon is dominant. See Fig. 4 for an example. The overfitting of NN cancel off most benefits of the introduction of attention masks. This leads us to introduce a spatial prior to regularize the estimation of the attention masks in NN for alleviating the overfitting. The prior we proposed is based on the following observations. Firstly, the relevant regions in action recognition usually only take a small portion of the full image, *i.e.*, only a small percentage of the entries of the attention mask are significant. Secondly, the pixels of these regions are not randomly distributed in spatial domain. They are likely to be located in connected regions with similar blob-type shape. These two physical observations on effective attention regions motivated us to propose a block-wise sparsity prior for regularizing the attention masks in the LSTM network. Our implementation of the block sparsity regularization is done by minimizing the $\ell_{2,1}$-norm of attention mask that prompts its group sparsity.

Given an attention mask $\boldsymbol{a} = [a_1, \cdots, a_{K^2}]$, we first divide the $K \times K$ spatial locations into $N \times N$ non-overlapping regions, where the region length is $L = \lfloor K/N \rfloor$. The index set of each region is defined as follows.

$$\mathbb{S}_r = \left\{ p \;\middle|\; \begin{array}{l} ((r-1)\%N)L \le \ (p-1)\%K \ < ((r-1)\%N + 1)L \\ \lfloor (r-1)/N \rfloor L \ \ \le \lfloor (p-1)/K \rfloor < \lfloor (r-1)/N \ + 1 \rfloor L \end{array} \right\}$$

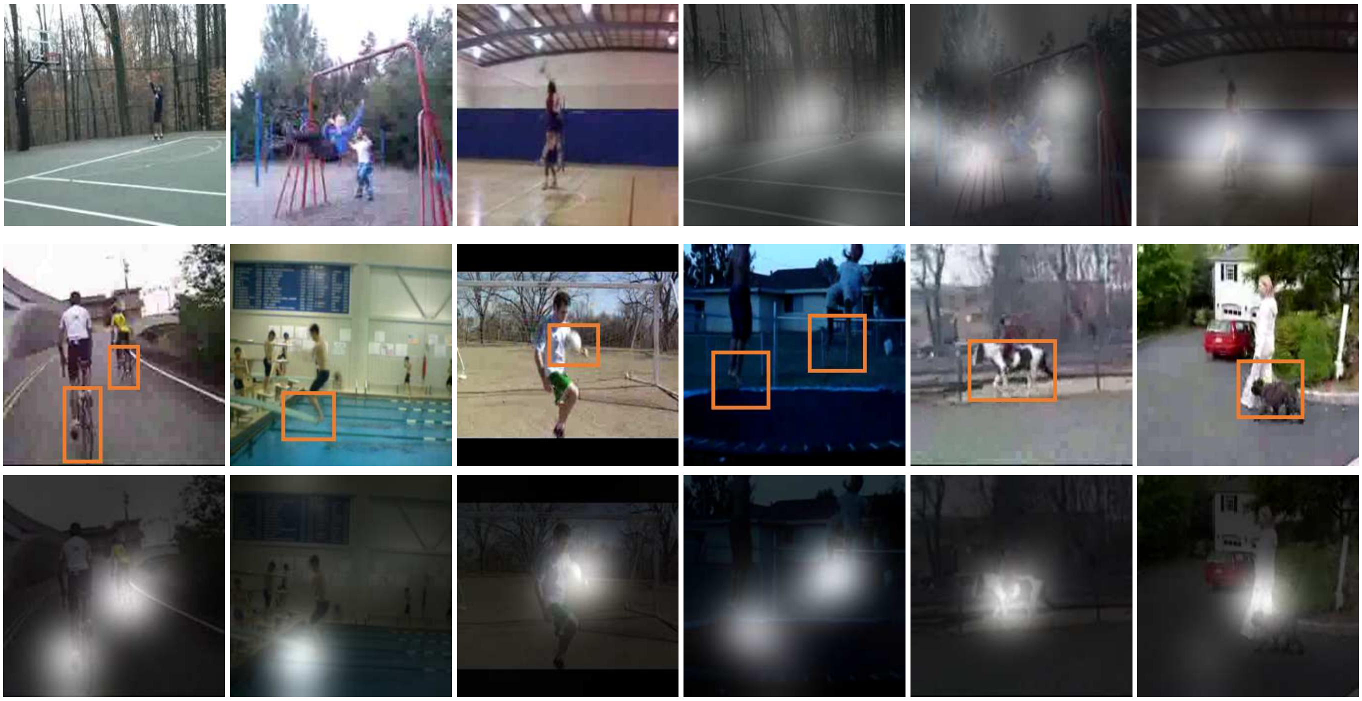


Fig. 4. Test video frames and their attention masks. First row: Phenomena of distracted attention when adopting an un-regularized attention mechanism. The left rows show one key frame of each test video frame and the right rows show the corresponding attention masks generated by the soft attention LSTM network (Sharma et al., 2015) with an un-regularized attention mechanism. It can be seen that the attention masks scatter over the background, while they are supposed to concentrate on the key parts of the objects related to the actions. The second and third rows: the key parts in the corresponding actions and the attention masks computed by the proposed attention mechanism which are consistent with human perception.

for $r = 1, \cdots, N^2$, where % denotes the modulo operation. Then we use the $\ell_{1,2}$ norm that prompting group sparsity to define the structure loss regarding an attention mask as follows.

$$\mathrm{r}(\boldsymbol{a}) = \sum_{r=1}^{N^2} \|\boldsymbol{a}_{\mathbb{S}_r}\|_2 = \sum_{r=1}^{N^2} \sqrt{\sum_{k\in\mathbb{S}_r} a_k{}^2}. \tag{10}$$

The structure loss $r(\cdot)$ applies the ℓ_2 norm within each block for smoothing the attention degrees in the corresponding region, while applying the ℓ_1 norm across the blocks to select important regions in the attention mask. In other words, the penalty by the structure loss in objective function can make the values of solution biased to a limited number of pivotal regions in which the values are well connected, and thus promoting the block sparsity we need. See Fig. 4 (b) for some results computed by the LSTM network with the proposed regularized attention mechanism.

3.4. Loss function and data preparation

In this section, we define the total loss function for training the network proposed in the previous sections. Recall that the proposed neural network accepts a video clip and predicts the label of each video frame in the clip. For convenience, let $g(\,\cdot\,, \boldsymbol{h}_2^{(t-1)}; \boldsymbol{\theta})$ denote the mapping that predicts the soft label $\boldsymbol{y}^{(t)}$ of the video frame $\boldsymbol{V}^{(t)}$ using the proposed model, *i.e.*

$$\boldsymbol{y}^{(t)} = g(\boldsymbol{V}^{(t)}, \boldsymbol{h}_2^{(t-1)}; \boldsymbol{\theta}),$$

where $\boldsymbol{h}_2^{(t-1)}$ is the hidden state generated at $(t-1)^{\text{th}}$ time step, and $\boldsymbol{\theta}$ contains all related model parameters. Let $\boldsymbol{a}^{(t)} = [\boldsymbol{a}_1^{(t)}, \cdots, \boldsymbol{a}_{K^2}^{(t)}]$ denote the attention mask , calculated by (8).at the t^{th} time step in the network. Given a video clip $\mathcal{V} = \{\boldsymbol{V}^{(t)}\}_{t=1}^{T}$ with T frames and its ground-truth label $\hat{\boldsymbol{y}}$, the loss function of the proposed model is defined as

$$L(\mathcal{V}, \hat{\boldsymbol{y}}, \boldsymbol{\theta}) = -\sum_{t=1}^{T}\sum_{i=1}^{C} \hat{y}_i \log y_i^{(t)} + \gamma \sum_i \theta_i^2 + \lambda \sum_{t=1}^{T}\sum_{r=1}^{N^2} \sqrt{\sum_{k\in\mathbb{S}_r} (a_k^{(t)})^2}, \tag{11}$$

where γ and λ are the weights that balance the contribution of each term. The first term is for measuring the discrimination error, which is defined by the cross-entropy between the predicted label and the ground-truth one. The second term is the weight decay regularization on the parameters of network for avoiding overfitting. The last term is the group sparsity regularization for imposing the block-wise sparsity prior on the attention masks.

Let $\{\mathcal{V}_p\}_{p=1}^{P}$ denote a set of P video clips for training the proposed model, and the corresponding ground-truth labels are denoted by $\{\hat{\boldsymbol{y}}_p\}_{p=1}^{P}$ where $\hat{\boldsymbol{y}}_p$ is an one-hot vector. The loss function on the training set $\{(\mathcal{V}_p, \hat{\boldsymbol{y}}_p)\}_{p=1}^{P}$ is defined as

$$\min_{\boldsymbol{\theta}} \sum_{p=1}^{P} L(\mathcal{V}_p, \hat{\boldsymbol{y}}_p, \boldsymbol{\theta}). \tag{12}$$

To enhance the stability of the proposed model, we use the following strategy of data augmentation to generate sufficient training data. Given a video of an action, we split the video into segments with a stride of d, each of which contains T consecutive frames. The video segments from all the videos are used as the video clips $\{\mathcal{V}_p\}_p$ for training.

3.5. Prediction

When a new video arrives for the prediction of action label, we first split the video into T-frame segments (clips),

$$\{\mathcal{V}_q : \mathcal{V}_q = \{\boldsymbol{V}_q^{(t)}\}_{t=1}^{T}\}_{q=1}^{Q},$$

using the same strategy as that in the data augmentation of training. Each segment is input to the proposed network, and the label of the segment is calculated as

$$\boldsymbol{y}_q = \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{V}_q^{(t)}, h_{t-1}; \boldsymbol{\theta}). \tag{13}$$

Then the predicted label of $\mathcal{V}_q$ is defined by

$$c_q = \max(\boldsymbol{y}_q(1), \boldsymbol{y}_q(2), \cdots, \boldsymbol{y}_q(C)). \tag{14}$$

Lastly, the label of the whole video is predicted by

$$c^* = \max_{c=1}^{C} \sum_{q=1}^{Q} \mathrm{I_c(c_q)}, \tag{15}$$

where $\mathrm{I_x}(\cdot)$ denotes the indicator function.

4. Experiments

4.1. Datasets

The proposed method for action recognition in a single video is evaluated on two public benchmark datasets: one is UCF-11 (Liu et al., 2009) and the other is HMDB-51 (Kuehne et al., 2011). See the following for a brief description of these two datasets.

- The **UCF-11** dataset is also called the "YouTube Action" dataset, which contains 1600 video clips from 11 action categories, including basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. Each category is divided into 25 groups, each group contains several video clips, and each video clip is of the frame rate 9.97fps that only associated with a single action. Since the data set does not provide a standard partition, we randomly extract 15 groups in each class as training data, and the remaining 10 groups are used for testing. Finally we used 948 videos for training and 652 videos for testing (*i.e.* 60% for training and 40% for testing). Due to the randomness of training/test split, the average of five runs are reported for the evaluation.

- The **HMDB-51** dataset is a large human motion dataset collected from movies, public databases, and web videos. It contains 6849 video clips with 51 action categories, such as clapping, drinking, hugging, jumping, somersaulting, throwing, and etc. Each category contains at least 101 video clips. The HMDB-51 dataset is more challenging than the UCF-11 dataset, as the scenes captured in it are more complex. In existing literature, there are three configurations to split the HMDB-51 dataset into the training set and test set. We followed the original setting (Kuehne et al., 2011), which used the first configuration and used 3570 videos (*i.e.* 70 videos per category) for training and used remaining 1530 videos (*i.e.* 30 videos per category) for testing. The training/test split is fixed in this setting for all the compared methods.

4.2. Implementation details

Recall that our method is composed of both CNN and RNN, and the training of the network with recurrent structure often does not converge when using the current training algorithm. Thus, we split the training procedure into two steps: first pre-training the CNN on images, and then learning the parameters of the RNN as well as the parameters of attention masks. The reason of pre-training CNN on images instead of videos is that we only consider spatial features in the CNN module of the proposed model. Indeed, pre-training CNN for feature extraction is a widely used practice in action recognition and video classification, *e.g.* (Sharma et al., 2015; Yang et al., 2018).

The training details are listed as follows. Each input video was split into segments with $T = 30$ and $d = 2$. The GoogLeNet model (Szegedy et al., 2016) trained on ImageNet (Deng et al., 2009) was adopted as the pre-trained CNN network. In the experiments, each video frame was resized to $224 \times 224 \times 3$, the number of output feature maps D was set to 2048, and the size of each feature map was set to 8×8. As a result, the size of $\boldsymbol{X}_t$ was 64×2048. Regarding the LSTM units, the dimensions of LSTM memory gates and hidden states were all set to 512. Regarding the module of structured attention, the spatial locations were divided into 4×4 non-overlapping regions with the length of 2, *i.e.* $N = 4$, $L = 2$. The penalty coefficients of the loss function in (11) were set as $\gamma = 10^{-5}$ and $\lambda = 10^{-4}$. The model was trained by Adam with the learning rate starting from 1×10^{-4}, and other parameters use default values. In the training, the batch size was set to 256 and the training process was stopped after 10 epochs. The dropout operation with ratio 50% was used at all non-recurrent connections.

The proposed neural network is implemented using TensorFlow with CUDA acceleration. The experiments were carried out on a workstation with a 3.5GHz Intel Core i7-5930K CPU, 64G RAM and an NVIDIA GeForce Titan-X GPU.

Table 1. Classification accuracy (%) by different methods.

Model	**UCF-11**	**HMDB-51**
(Gammulle et al., 2017)	89.20	-
(Meng et al., 2018)	89.70	-
(Li et al., 2018)	-	43.30
(Yan et al., 2017)	-	43.40
(Sharma et al., 2015)	84.86	41.31
(Yang et al., 2018)	89.70	**52.30**
Ours	**91.84**	48.81

4.3. Quantitative analysis

The performance of the proposed method is compared to several recent video-based action recognition methods, including (Gammulle et al., 2017; Meng et al., 2018; Li et al., 2018; Sharma et al., 2015; Yang et al., 2018). These methods were chosen for comparison since they are also built uopn LSTM networks or attention mechanisms. The results in terms of classification accuracy are shown in Table 1. It is clear that the proposed method performs the best among all the compared methods on the UCF-11 dataset. On the HMDB-51 dataset, our method exceeds all other methods except (Yang et al., 2018). It is noted that the method proposed in (Yang et al., 2018) uses optical flow as an additional cue for action recognition In contrast, our method does not use optical flow. By using less cues, our method still outperforms (Yang et al., 2018) on the UCF-11 dataset and is comparable on the HMDB-51 dataset. The soft-attention LSTM model (Sharma et al., 2015) is more related to the proposed method, which also employs a similar attention mechanism but without structure regularization. The effectiveness of the proposed regularized attention mechanism is justified by the improvement (*i.e.* 6.98% on UCF-11 and 7.5% on HMDB-51) of the proposed method over (Sharma et al., 2015).

The results listed above demonstrated performance gain of our approach to attention mechanism in neural network. We also compared our method with some state-of-the-art methods whose architectures and mechanisms are different from ours and the above compared methods. Table 2 shows the results of the comparison. The performance of our method is close to I3D (Carreira and Zisserman, 2017) but worse than MiCT-Net (Zhou et al., 2018) and SMNet (Feichtenhofer et al., 2017). However, we note that some of these methods use additional sources such as optical flow as input, and our attention mechanism can also be incorporated into these methods for further improvement. While not performing as good as those recent state-of-the-art ones, our approach still has its value for attention mechanism, one important technique in action recognition.

Table 2. Classification accuracy (%) by ours and some state-of-the-art methods on HMDB-51.

Model	**HMDB-51**
I3D (Carreira and Zisserman, 2017)	49.80
MiCT-Net (Zhou et al., 2018)	63.80
SMNet (Feichtenhofer et al., 2017)	**68.90**
The proposed method	48.81

4.4. Parameter influence analysis and more studies

The parameter λ in the loss function of (11) is one important parameter in the proposed method, as it controls the contribution of the block-structured sparsity regularization on the attention masks. To test the impact of the different settings of λ to the performance, we conducted the experiments using $\lambda = 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}$ respectively. The results are shown in Table 3. It can be seen from the table that within a reasonable range, the performance of the proposed model arises as λ is increased. Such results also demonstrate the effectiveness of introducing the structured attention mechanism. When λ exceeds the range, the performance of our model decreases. It is reasonable because very large value of λ imposes too much regularization on attention masks, and the resulting model could be less relevant to the data.

Table 3. Classification accuracies (%) of proposed method with different values of λ.

Model	**UCF-11**	**HMDB-51**
Proposed model ($\lambda = 10^{-3}$)	89.54	47.62
Proposed model ($\lambda = 10^{-4}$)	**91.84**	**48.81**
Proposed model ($\lambda = 10^{-5}$)	90.46	47.56
Proposed model ($\lambda = 10^{-6}$)	89.23	47.29

The block size N in constructing the group sparsity regularization in our method is also an important factor, as it may influence the way to divide image into regions and determine the impact of the block sparsity constraint. The block size is related to the expected spatial region over which an action is to occur, and the blocks with proper size can well fit the spatial regions related to the action. The ablation study on the value of N is done by conducting the experiments using another two values of N. Moreover, we also tested the performance of using overlapping blocks instead of the non-overlapping ones in our method, since intuitively using overlapping blocks may have better expressive power to fit more complex shape of attention masks. In Table 4, we list the results of using different values of N and different strides of overlapping block sampling. From the results we can see that using block size $N = 4$ yields better results than using bigger/smaller values. In practice, setting N too large may make a block to cover many pixels that are not related to the action, which significantly decreases the performance. Meanwhile, too small N may weaken the regularization of structured sparsity, which leads to noticeable impact on performance. Also, it can be seen from Table 4 that using overlapping regions is worse than using non-overlapping regions. The reason is probably that using overlapping regions has less regularization effects on the network as it has more freedoms to fit the attention well.

Table 4. Classification accuracy (%) by our method using different block sizes and sampling strides on UCF-11.

	Non-overlapping			**Overlapping**		
Stride	2	4	6	1	2	2
Blk. Size	2	4	6	2	4	6
Accuracy	89.17	**91.84**	85.22	86.71	88.46	82.62

To verify the effectiveness of our attention mechanism, we compare our method to the version without attention, as well as the version with classic un-regularized attention Sharma et al. (2015). See Table 5 for the comparison It can be seen that our attention mechanism outperformed the other two compared ones demonstrates the benefits of the attention mechanism regularized by group-sparsity prior in action recognition.

The attention map we generated only takes previous state of LSTM as input. One possible improvement is taking account

Table 5. Classification accuracy (%) using different attention mechanisms.

Method	UCF-11	HMDB-51
Without attention	82.37	38.46
Classic attention	84.98	41.42
Our attention	**91.84**	**48.81**

of the feature of current frame when computing the attention map. With such a purpose, we modified our network as follows: (i) applying a 1×1×2048 convolution to $\boldsymbol{X}^{(t)}$ to form a 2D map; (ii) applying an MLP (multi-layer perceptron) to the 2D map to form a 512-dimensional feature $\boldsymbol{w}^{(t)}$; (iii) replacing $\boldsymbol{h}^{(t-1)}$ with the concatenation of $\boldsymbol{h}^{(t-1)}$ and $\boldsymbol{w}^{(t)}$ in the attention module. However, we found that the performance has no further improvement but even gets worse with more than 0.5 accuracy decrease. The reason is probably that additional operations for utilizing current frame makes the network much more complex.

4.5. *Visual illustration*

This section is for visual illustration of some practical examples using the regularized attention mechanism proposed in this paper. See Fig. 5 for the demonstration of sample attention masks obtained in our method, which includes several video frames and the corresponding attention masks produced by the proposed method. In Fig. 5 (a), the attention mask concentrates on the regions of the soccer ball in the video frames. Note that the soccer ball is the key object in the action of soccer juggling. In Fig. 5 (b), the attention concentrates on the pedal as well as the feet of the biker in the video frames, which are the key parts of the subject and object in the action of biking. Fig. 5 (c) shows another example of biking. In this example, in addition to major attention on the bike, there is also partial attention locating at the road. This result is reasonable, as biking and road are highly correlated to each other, considering the fact that biking often occurs on road. In Fig. 5 (d), it can be seen that the proposed attention mechanism can capture the key parts of golfing, *i.e.*swinging hand.

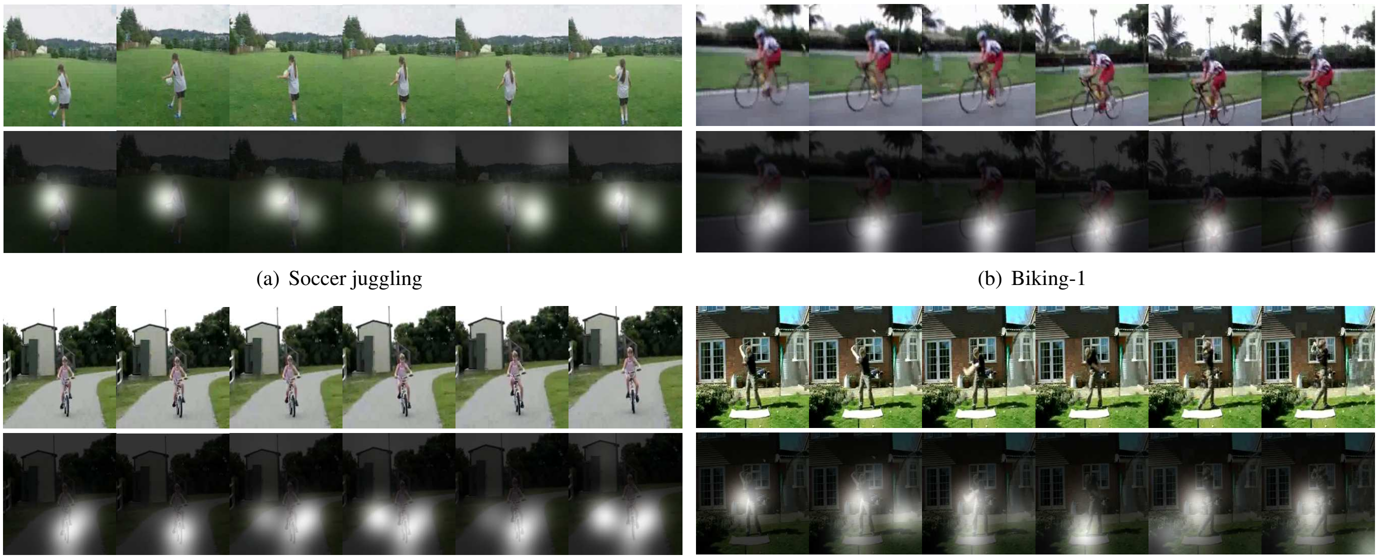
In the next, we present some examples to illustrate how the proposed block-structured sparsity regularization impacts the performance of the attention mechanism. In this demonstration, the attention masks are generated by the attention mechanism with and without regularization, which can be done by setting the parameter $\lambda = 0$) of the proposed method. In Fig. 6 (a), the un-regularized version incorrectly classifies the soccer juggling video as golf swing, while in Fig. 6 (b) the proposed regularized version yields a correct prediction. It can be seen that the pixels of the regions produced by the un-regularized version are scattering over the frames in the video. In contrast, the attention regions predicted by the proposed regularized version mainly concentrate on the objects and subjects of action, which certainly provide more relevant information for the recognition of action. In addition, it can be seen in Fig. 6 (a) and Fig. 6 (b) that the attention to the right person gradually becomes weaker while the attention to the left person becomes more prominent over the time. This shows that over time, the proposed regularized attention mechanism can effectively focus on the object related to the action instead of simply detecting human body. See Fig. 6 (c) and Fig. 6 (d) for the illustration of another example, in which the un-regularized attention mechanism generates the attention masks that spread out such that most are wrongly labeled. It is noted that the attention is effective to some degree at the beginning, but it gradually diffuses over time, which makes the LSTM network gate the discriminative information of the action. In contrast, the proposed regularized one is capable of classifying the input video correctly using the attention masks with higher accuracy.

The visual inspection on Fig. 4 and Fig. 5 might lead to the concern that the attention masks may be too sharply focused on sub-parts of the actor that do not necessarily define the action in and of themselves. For example, in the case of bicycle riding the cycles are much more highlighted than the feet of the rider, and in the case of horseback riding, the horse is much more focused than than the feet of the rider. In other words, it might be confused with stationary bicycles that have no no riders. Indeed, such a concern is not a severe issue. Firstly, human visual perception allows one to recognize a human action by mainly focusing on certain key parts of the human body or of the target of the action. Secondly, in the cases of bicycle riding and horse riding, the attention mask still has small focusing part on the feet of rider. Lastly, when the bicycle and horse are stationary, the attention mask may be not focused on them, as the attention is based on temporal cues exploited by the LSTM. Thus, our method will not take stationary bicycles into account when recognizing an action. See Fig. 7 for a demonstration.

Before ending the section, we also show some cases that the proposed method fail. See Fig. 8 for the illustration. In Fig. 8 (a), the video of playing basketball is mis-recognized as the action of swinging, and the attention is misplaced on the basketball court. One possible cause of such failure might be that the basketball player is not conspicuous enough and too small to satisfy the block size when calling block-sparsity regularization. Setting a proper block size in the group sparsity regularization can help alleviating such error. Another example is shown in Fig. 8 (b), where the video of football juggling is mis-recognized as golf swing. In this example, the attention is placed on the player at the beginning, but later the attention leaves the human body due to a large movement occurring in a short time interval, which eventually leads to incorrect classification. This issue can not be easily fixed in the current implementation of the proposed mechanism, and it will be one problem we focus on in future work.

5. Conclusion

Attention mechanism is an important technique in neural network based recognition systems by telling the system where to focus. However, existing implementation of attention mechanism in action recognition often suffers from over-fitting such that its benefits are mostly canceled out. Motivated by the observation that the parts of attention usually only concentrate on a few connected regions relevant to the subject and object of an action, this paper proposed an LSTM network for action recognition with attention mechanism regularized by a block-



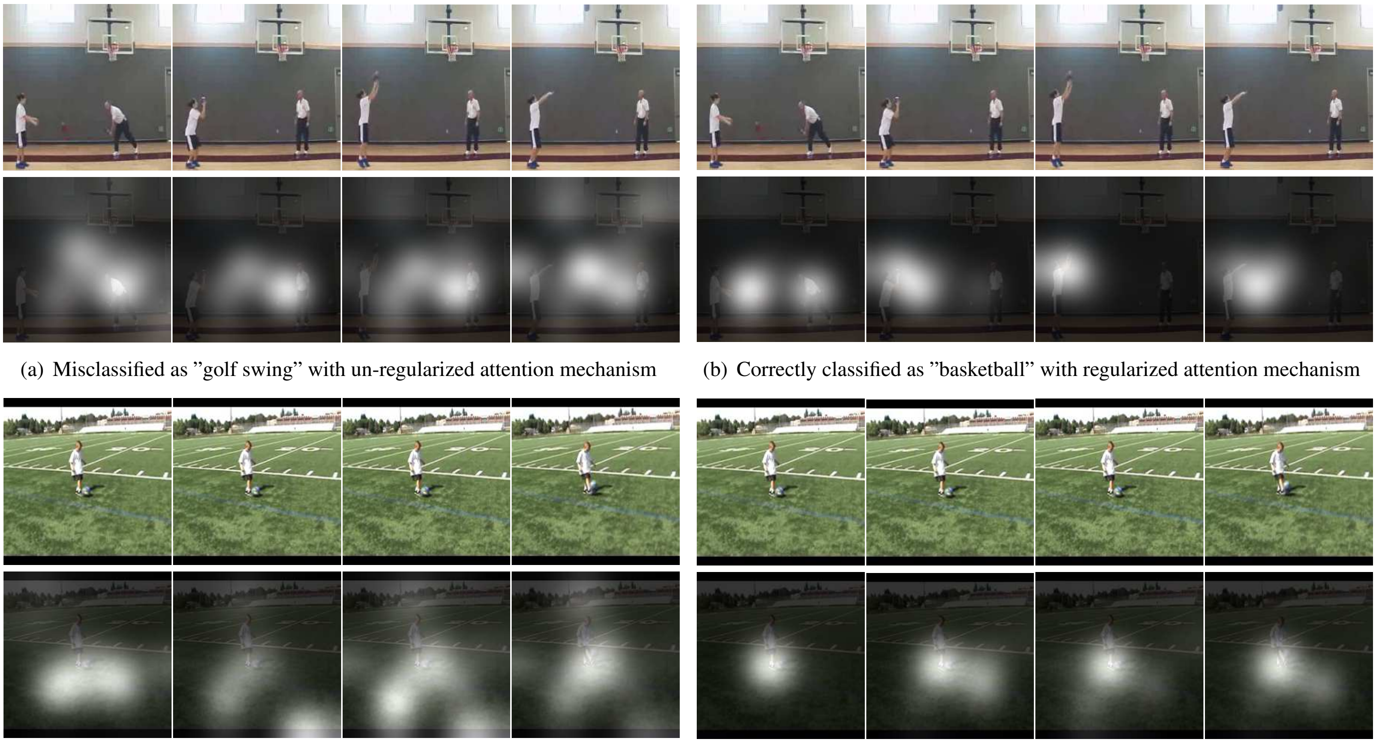
(a) Soccer juggling

(b) Biking-1

(c) Biking-2

(d) Golfing

Fig. 5. Video frames (upper rows) and the corresponding attention masks (bottom rows) produced by the proposed method.



(a) Misclassified as ”golf swing” with un-regularized attention mechanism

(b) Correctly classified as ”basketball” with regularized attention mechanism

(c) Misclassified as ”volleyball spiking” with un-regularized attention mechanism

(d) Correctly classified as ”soccer juggling” with regularized attention mechanism

Fig. 6. Video frames (upper rows) and the corresponding attention masks (bottom rows) produced by the baseline method and the proposed method. (a) and (c) correspond to the results by the baseline method that uses the un-regularized attention mechanism, where the videos are mis-classified. (b) and (d) correspond to the results by the proposed method which uses the regularized attention mechanism, where the videos are correctly classified.

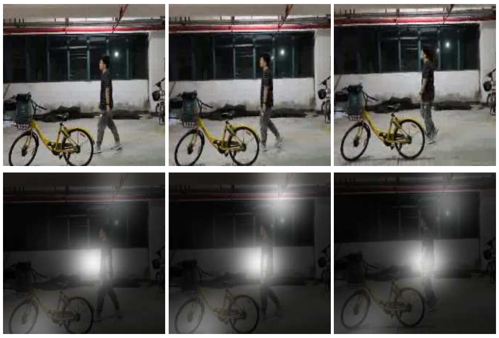


Fig. 7. Key frames of a video (upper row) and its attention mask (bottom row) generated by our method. The video contains a walking man and a stationary bike. It can be seen that the attention mask generated by our mechanism mainly focuses on the walking man instead of the stationary bike.

wise sparsity prior, In the proposed method, the prior is implemented by imposing a $\ell_{2,1}$-norm on attention mask that prompts its block-wise sparsity. With such a built-in regularized attention mechanism that measures the relevance of image pixels to action recognition, the proposed neural network showed good performance in the experiments on the UCF-11 and HMDB-51 datasets. In future, we would like to investigate how to improve the robustness of the attention mechanism in the case where large motions of objects/subjects of actions occur, and how to introduce the idea of regularized attention mechanism to other recognition tasks for better performance.

Acknowledgments

Yuhui Quan would like to acknowledge the support by National Natural Science Foundation of China (61602184, 61872151, 61672241), Natural Science Foundation of Guangdong Province (2017A030313376), Cultivation Project of Major Basic Research of NSF-Guangdong Province (2016A030308013), and Fundamental Research Funds for Central Universities (x2js-D2181690). Hui Ji would like to acknowledge the support from Singapore MOE AcRF (R146000229114, MOE2017-T2-2-156).

References

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A., 2010. Action classification in soccer videos with long short-term memory recurrent neural networks. Proceedings of International Conference on Artificial Neural Networks , 154–159.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A., 2011. Sequential deep learning for human action recognition, in: International Workshop on Human Behavior Understanding, Springer. pp. 29–39.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S., 2016. Dynamic image networks for action recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 3034–3042.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. Proceedings of Conference on Computer Vision and Pattern Recognition .

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248–255.

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

Du, W., Wang, Y., Qiao, Y., 2018. Recurrent spatial-temporal attention network for action recognition in videos. IEEE Transactions on Image Processing 27, 1347–1360.

Du, Y., Wang, W., Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 1110–1118.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J., 2009. Automatic annotation of human actions in video, in: Proceedings of International Conference on Computer VisionCV, IEEE. pp. 1491–1498.

Fanti, C., Zelnik-Manor, L., Perona, P., 2005. Hybrid models for human motion recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1166–1173.

Fathi, A., Mori, G., 2008. Action recognition by learning mid-level motion features, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–8.

Feichtenhofer, C., Pinz, A., Wildes, R., 2016a. Spatiotemporal residual networks for video action recognition, in: Proceedings of Conferences on Advances in Neural Information Processing Systems, pp. 3468–3476.

Feichtenhofer, C., Pinz, A., Wildes, R.P., 2017. Spatiotemporal multiplier networks for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4768–4777.

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

Fernando, B., Anderson, P., Hutter, M., Gould, S., 2016. Discriminative hierarchical rank pooling for activity recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 1924–1932.

Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T., 2017. Rank pooling for action recognition. IEEE Transactions on Pattern Analysis and Machine Inteligence 39, 773–787.

Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T., 2015. Modeling video evolution for action recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 5378–5387.

Gammulle, H., Denman, S., Sridharan, S., Fookes, C., 2017. Two stream lstm: A deep fusion framework for human action recognition, in: Proceedings of Winter Conference on Applications of Computer Vision, IEEE. pp. 177–186.

Girdhar, R., Ramanan, D., 2017. Attentional pooling for action recognition, in: Proceedings of Conferences on Advances in Neural Information Processing Systems, pp. 34–45.

Hoai, M., Zisserman, A., 2014. Improving human action recognition using score distribution and ranking, in: Proceedings of Asian Conference on Computer Vision, Springer. pp. 3–20.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780.

Itti, L., Koch, C., 2001. Computational modelling of visual attention. Nature Reviews Neuroscience 2, 194–203.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Inteligence 35, 221–231.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

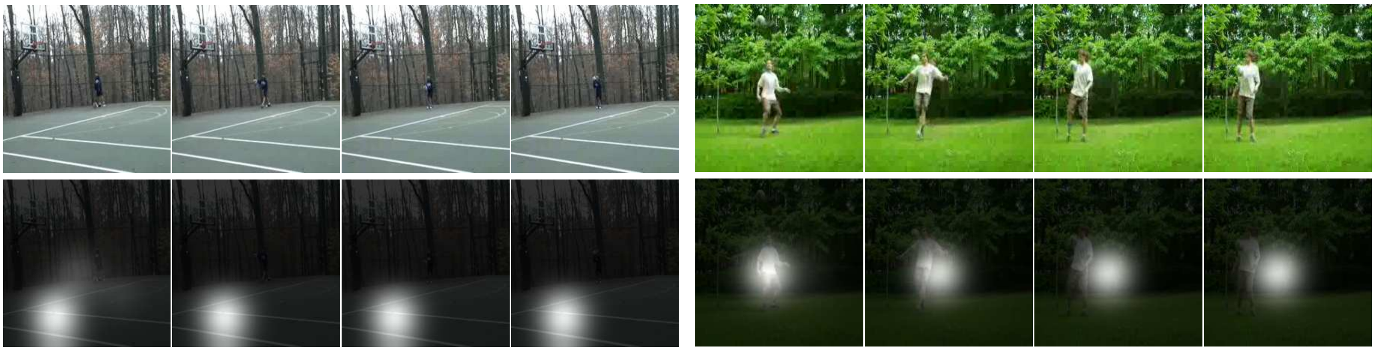
Klaser, A., Marszałek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of British Machine Vision Conference, British Machine Vision Association. pp. 275–1.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. Hmdb: a large video database for human motion recognition, in: Proceedings of International Conference on Computer Vision, IEEE. pp. 2556–2563.

Lan, T., Zhu, Y., Roshan Zamir, A., Savarese, S., 2015. Action recognition by hierarchical mid-level action elements, in: Proceedings of International Conference on Computer Vision, pp. 4552–4560.

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies, in: Proceedings of Conference on Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–8.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G., 2018. Videolstm convolves, attends and flows for action recognition. Computer Vision and



(a) Incorrectly classified as ”swing”

(b) Incorrectly classified as ”golf swing”

Fig. 8. Incorrectly-classified examples by the proposed method. The upper rows show the original video frames. The bottom rows show the attention masks produced by the proposed method. (a) The model only focuses on the corner of the basketball court; (b) The model completely loses its focus on the motion.

Image Understanding 166, 41–50.

Liu, J., Luo, J., Shah, M., 2009. Recognizing realistic actions from videos “in the wild”, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1996–2003.

Lv, F., Nevatia, R., 2007. Single view human action recognition using key pose matching and viterbi path searching, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–8.

Meng, B., Liu, X., Wang, X., 2018. Human action recognition based on quaternion spatial-temporal convolutional neural network and lstm in rgb videos. Multimedia Tools and Applications , 1–18.

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126.

Niebles, J.C., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79, 299–318.

Poppe, R., 2010. A survey on vision-based human action recognition. Image and Vision Computing 28, 976–990.

Raptis, M., Sigal, L., 2013. Poselet key-framing: A model for human activity recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 2650–2657.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of ACM International Conference on Multimedia, ACM. pp. 357–360.

Sharma, S., Kiros, R., Salakhutdinov, R., 2015. Action recognition using visual attention. arXiv preprint arXiv:1511.04119 .

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos, in: Proceedings of Conferences on Advances in Neural Information Processing Systems, pp. 568–576.

Srivastava, N., Mansimov, E., Salakhudinov, R., 2015. Unsupervised learning of video representations using lstms, in: International Conference on Machine Learning, pp. 843–852.

Sun, R., 2008. The Cambridge handbook of computational psychology. Cambridge University Press.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tang, K., Fei-Fei, L., Koller, D., 2012. Learning latent temporal structure for complex event detection, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1250–1257.

Thurau, C., Hlavác, V., 2008. Pose primitive based human action recognition in videos or still images, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–8.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of International Conference on Computer Vision, pp. 4489–4497.

Veeriah, V., Zhuang, N., Qi, G.J., 2015. Differential recurrent neural networks for action recognition, in: Proceedings of International Conference on Computer Vision, pp. 4041–4049.

Wang, C.Y., Chiang, C.C., Ding, J.J., Wang, J.C., 2017. Dynamic tracking attention model for action recognition, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, IEEE. pp. 1617–1621.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2011. Action recognition by dense trajectories, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3169–3176.

Wang, H., Schmid, C., 2013. Action recognition with improved trajectories, in: Proceedings of International Conference on Computer Vision, pp. 3551–3558.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1290–1297.

Wang, X., Farhadi, A., Gupta, A., 2016. Actions˜ transformations, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 2658–2667.

Wang, Y., Mori, G., 2011. Hidden part models for human action recognition: Probabilistic versus max margin. IEEE Transactions on Pattern Analysis and Machine Inteligence 33, 1310–1323.

Yan, S., Smith, J.S., Lu, W., Zhang, B., 2017. Cham: action recognition using convolutional hierarchical attention model. arXiv preprint arXiv:1705.03146 .

Yang, H., Zhang, J., Li, S., Lei, J., Chen, S., 2018. Attend it again: Recurrent attention convolutional neural network for action recognition. Applied Sciences 8, 383.

Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L., 2011. Human action recognition by learning bases of action attributes and parts, in: Proceedings of International Conference on Computer Vision, IEEE. pp. 1331–1338.

Yeffet, L., Wolf, L., 2009. Local trinary patterns for human action recognition, in: Proceedings of International Conference on Computer Vision, IEEE. pp. 492–497.

Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L., 2015. Every moment counts: Dense detailed labeling of actions in complex videos. International Journal of Computer Vision , 1–15.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G., 2015. Beyond short snippets: Deep networks for video classification, in: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 4694–4702.

Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 .

Zhou, Y., Sun, X., Zha, Z.J., Zeng, W., 2018. Mict: Mixed 3d/2d convolutional tube for human action recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X., et al., 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks., in: Proceedings of Coference on Association for the Advancement of Artificial Intelligence, p. 8.