Attention-based Temporal Weighted Convolutional Neural Network for Action Recognition

Jinliang Zang[1], Le Wang[1], Ziyi Liu[1], Qilin Zhang[2], Zhenxing Niu[3], Gang Hua[4], and Nanning Zheng[1]

[1]Xi'an Jiaotong University, Xi'an, Shannxi 710049, P.R.China
[2]HERE Technologies, Chicago, IL 60606, USA
[3]Alibaba Group, Hangzhou, Zhejiang 311121, P.R.China
[4]Microsoft Research, Redmond, WA 98052, USA

Abstract. Research in human action recognition has accelerated significantly since the introduction of powerful machine learning tools such as Convolutional Neural Networks (CNNs). However, effective and efficient methods for incorporation of temporal information into CNNs are still being actively explored in the recent literature. Motivated by the popular recurrent attention models in the research area of natural language processing, we propose the Attention-based Temporal Weighted CNN (ATW), which embeds a visual attention model into a temporal weighted multi-stream CNN. This attention model is simply implemented as temporal weighting yet it effectively boosts the recognition performance of video representations. Besides, each stream in the proposed ATW framework is capable of end-to-end training, with both network parameters and temporal weights optimized by stochastic gradient descent (SGD) with backpropagation. Our experiments show that the proposed attention mechanism contributes substantially to the performance gains with the more discriminative snippets by focusing on more relevant video segments.

Key words: Action recognition, Attention model, Convolutional neural netwoks, Video-level prediction, Temporal weighting

1 Introduction

Action recognition and activity understanding in videos are imperative elements of computer vision research. Over the last few years, deep learning techniques dramatically revolutionized research areas such as image classification, object segmentation [7–9] and object detection [1–6]. Likewise, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been popular in the action recognition task [6, 10–17]. However, various network architectures have been proposed with different strategies on the incorporation of video temporal information. However, despite all these variations, their performance improvements over the finetuned image classification network are still relatively small.

Unlike image classification, the most distinctive property of video data is the variable-length. While Images can be readily resized to the same spatial resolution, it is difficult to subsample videos temporally. Therefore, it is difficult for the early 3D ConvNet [1] to achieve action recognition performance on par with the sophisticated hand-crafted iDT [18] representations.

In addition, some of the legacy action recognition datasets (*e.g.*, KTH [19]) only contain repetitive and transient actions, which are rarely seen in everyday life and therefore have limited practical applications. With more realistic actions included (with complex actions, background clutter and long temporal duration), the more recent action recognition dataset, *e.g.*, YouTube's sports, daily lives videos (UCF-101 [20]) and isolated activities in movies (HMDB-51 [21]), offer much more realistic challenges to evaluate modern action recognition algorithms. Therefore, all experimental results in this paper are based on the UCF-101 and HMDB-51 datasets.

Previous multi-stream architecture, such as the two-stream CNN [10], suffers from a common drawback, their spatial CNN stream is solely based on a single image randomly selected from the entire video. For complicated activities and relatively long action videos (such as the ones in the UCF-101 and HMDB-51 datasets), viewpoint variations and background clutter could significantly complicate the representation of the video from a single randomly sampled video frame. A recent remedy was proposed in the Temporal Segment Network (TSN) [12] with a fusion step which incorporates multiple snippets[1].

Inspired by the success of the attention model widely used in natural language processing [22] and image caption generation [23, 24], the Attention-based Temporal Weighted CNN (ATW) is proposed in this paper, to further boost the performance of action recognition by the introduction of a benign competition mechanism between video snippets. The attention mechanism is implemented via temporal weighting: instead of processing all sampled frames equally, the temporal weighting mechanism automatically focuses more heavily on the semantically critical segments, which could lead to reduced noise. In addition, unlike prior P-CNN [15] which requires additional manual labeling of human pose, a soft attention model is incorporated into the proposed ATW, where such additional labeling is eliminated. Each stream of the proposed ATW CNN can be readily trained end-to-end with stochastic gradient descent (SGD) with backpropagation using only existing dataset labels.

The major contributions of this paper can be summarized as follows. (1) An effective long-range attention mechanism simply implemented by temporal weighting; (2) each stream of the proposed ATW network can be optimized end-to-end, without requiring additional labeling; (3) state-of-the-art recognition performance is achieved on two public datasets.

[1] Snippets are multi-modal data randomly sampled from non-overlapping video segments, see Fig. 1. Typically a video is divided into 1 to 8 segments. Segments are typically much longer than "clips" used by 3D CNN literature, *e.g.*, the 16-frame clip in C3D [14].

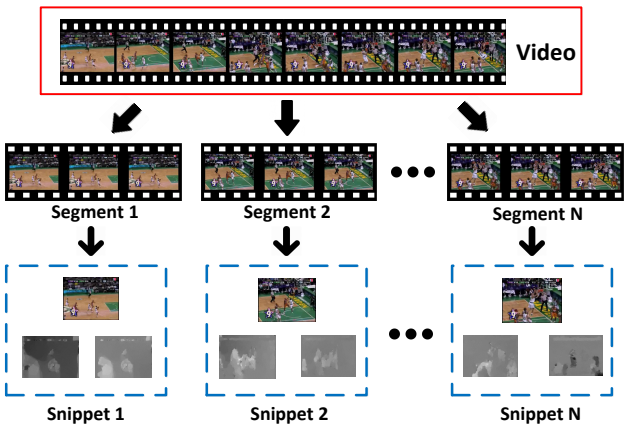

Fig. 1. Snippet generation with a fixed target number (N) of chunks. A video is evenly portioned into N non-overlapping segments. Each segment contains approximately the same number of video frames. As shown above, 2 additional modalities derived from RGB video frames are also included, *i.e.*, optical flows and warped optical flows. RGB, optical flow and warped optical flow images sampled from the same segment are grouped in a snippet.

2 Related Works

Human action recognition has been studied for decades, which were traditionally based on hand-crafted features, such as dense trajectories [18, 25] and sparse space-time interest points [26]. In the past few years, CNN based techniques have revolutionized the image/video understanding [1, 2, 4, 6, 10–13, 16, 27]. Per the data types used for action recognition, deep neural networks based methods can be categorized into two groups: (1) RGBD camera based action recognition, usually with skeleton data and depth/3D point clouds information [15, 28, 29]; (2) conventional video camera based action recognition.

RGBD camera based action recognition offers 3D information, which is a valuable addition to the conventional RGB channels. Such datasets are usually captured by the Microsoft Xbox One Kinect Cameras, such as The Kinetics dataset [15]. Despite its obvious advantage, there are some limiting factors which restrict such model from wide applications. RGBD video datasets are relatively new and labelled ones are not always readily available. A huge backlog of videos captured by conventional RGB camcorders cannot be parsed by such methods due to modality mismatch [30]. In addition, pure pose/skeleton based pipelines rarely achieve recognition accuracy on par with RGB video frame based pipelines [31, 32], making them more suitable for an auxiliary system to existing ones.

Inspired by the success of computer vision with still RGB images, many researchers have proposed numerous methods for the conventional RGB video camera based action recognition. Ji *et al.* [1] extend regular 2D CNN to 3D, with

promising performances achieved on small video datasets. Simonyan *et al.* [10] propose the two-stream CNN, with each steam being a regular 2D CNN. The innovation is primarily in the second CNN steam, which parses a stack of optical flow images that contain temporal information. Since then, optical flow is routinely used as the secondary modality in action recognition. Meanwhile, 3D CNN has evolved, too. Tran *et al.* [14] modified traditional 2D convolution kernels and proposed the C3D network for spatiotemporal feature learning. Feichtenhofer *et al.* [16] discovered one of the limiting factors in the two-stream CNN architecture, only a single video frame is randomly selected from a video as the input of the RGB image stream. They proposed five variants of fusing spatial CNN stream and two variants for the temporal steam. Additionally, Donahue *et al.* [13] developed a recurrent architecture (LRCN) to boost the temporal discretion. Consecutive video frames are loaded with redundant information and noises, therefore, they argue that temporal discretion via LRCN is critical to action recognition. Some recent literature also proposed new architectures with special considerations for temporal discretion [12, 17, 33, 34].

3 Formulation

Firstly, the temporally structured video representation is introduced, followed by the temporal attention model and the proposed ATW framework.

3.1 Temporally Structured Representation of Action

How do various CNN based architectures incorporate the capacity to extract semantic information in the time domain? According to the previous two-stream CNN [10] literature, there are generally 3 sampling strategies: (1) dense sampling in time domain, the network inputs are consecutive video frames covering the entire video; (2) spare sampling one frame out of τ ($\tau \geq 2$) frames, i.e., frames at time instants $0, t, t+\tau, t+2\tau, \cdots, t+N\tau$ are sampled; (3) with a target number of N segments[2], non-overlapping segments are obtained by evenly partition the video into N such chunks, as illustrated in Fig. 1.

As noted by [12, 13, 16], the dense temporal sampling scheme is suboptimal, with consecutive video frames containing redundant and maybe irrelevant information, recognition performance is likely to be compromised. For the sparse sampling strategy with τ intervals, the choice of τ is a non-trivial problem. With τ too small, it degrades to the dense sampling; with τ too large, some critical discriminative information might get lost. Therefore, the third sampling scheme with fixed target segments is arguably the advisable choice, given the segment number N is reasonably chosen.

Suppose a video V is equally partitioned into N segments, *i.e.*, $V = \{S_k\}_{k=1}^{N}$, where S_k is the k-th segment. Inspired by [10, 12, 35], multi-modality processing

[2] Typical N values are from 1 to 8.

is beneficial. Therefore, three modalities (RGB video frame, optical flow image and warped optical flow image[3]) are included in our proposed ATW network.

One RGB video frame, five optical flow image and five warped optical flow images are randomly sampled from each segment S_k, as illustrated in Fig. 1, and respectively used as the inputs to the spatial RGB ResNet stream, temporal flow ResNet stream, and temporal warped flow ResNet stream, as shown in Fig. 2. RGB, optical flow and warped optical flow images sampled from the same video segment are grouped in a snippet. Each snippet is processed by the proposed 3-stream ATW network and a per-snippet action probability is obtained. After processing all snippets, a series of temporal weights are learned via the attention model, which are used to fuse per-snippet probabilities into video-level predictions.

3.2 Temporal Attention Model

The proposed ATW network architecture is presented in Fig. 2. Our base CNN is the ResNet [36] or BN-Inception [37], which are both pretrained on the ImageNet dataset [38]. During the training phase, every labeled input video V is uniformly partitioned into N segments, *i.e.*, $V = \{M_i^{RGB}, M_i^F, M_i^{WF}, y\}_{i=1}^N$, where $M_i^{RGB}, M_i^F, M_i^{WF}$ represent the RGB, optical flow and warped optical flow images from the ith snippet, with y being the corresponding action label. The 3 CNN stream ($\mathcal{C}_{RGB}$, $\mathcal{C}_F$ and $\mathcal{C}_{WF}$) map each input to corresponding feature vector as

$$\begin{aligned} &\mathcal{C}_{RGB}(M_i^{RGB}) = \mathbf{a}_i^{RGB}, \\ &\mathcal{C}_F(M_i^F) = \mathbf{a}_i^F, \\ &\mathcal{C}_{WF}(M_i^{WF}) = \mathbf{a}_i^{WF}, \\ &i = 1, \cdots, N, \end{aligned} \tag{1}$$

where we call these $\mathbf{a}_{att}^{RGB}$, $\mathbf{a}_{att}^F$, $\mathbf{a}_{att}^{WF}$ action feature vectors, and use $\mathbf{a}_i$ to represent any given one from the 3 modalities. Note that w_i is the expected importance value of the ith snippet relative to the entire video. Evidently, if $w_i \equiv \frac{1}{N}$, the attention model degrades to naive averaging. The weight w_i is computed by the attention model f_{att} by a multi-layer perceptron conditioned on the previous fully-connected hidden state (*i.e.*, $\mathbf{w}_{att}$). The value of weight w_i decides which part of the segments should to pay attention to. Formally, the attention model f_{att} is defined as

$$e_i = f_{att}(\mathbf{w}_{att}, \mathbf{a}_i) = \mathbf{w}_{att}^{\mathbf{T}} \mathbf{a}_i. \tag{2}$$

The weight w_i of each action vector is computed by

$$w_i = \frac{\exp e_i}{\sum_j \exp e_j}, \tag{3}$$

[3] As in [18], warped optical flow is obtained by compensating camera motion by an estimated homography matrix.

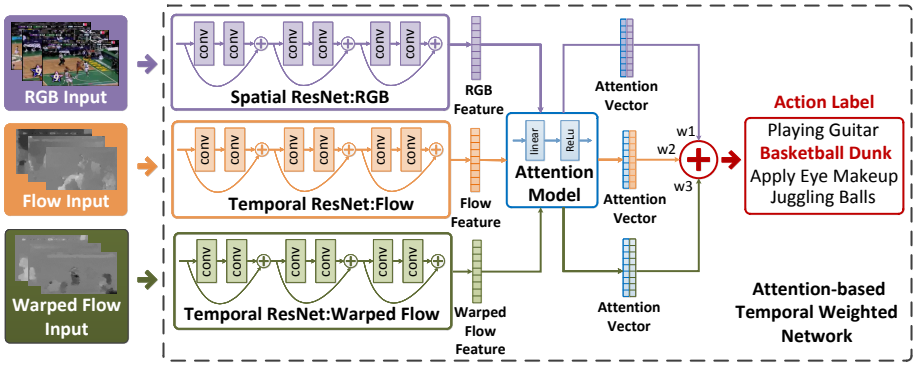

Fig. 2. Proposed ATW network architecture. Three CNN streams are used to process spatial RGB images, temporal optical flow images, and temporal warped optical flow images, respectively. An attention model is employed to assign temporal weights between snippets for each stream/modality. Weighted sum is used to fuse predictions from the three streams/modalities.

where each w_i are normalized by passing through a softmax function, which guarantees they are positive with $\sum_i w_i = 1$. Finally, the attention mechanism φ is implemented with a linear layer followed by a rectifier (ReLu), which serve as a temporal weighting function that aggregates all the per-snippet prediction probabilities into a per-video prediction. After training, the attention model obtains a set of non-negative weights $\{w_i\}_{i=1}^{N}$, so the weighted attention feature is obtained by

$$\begin{aligned}
\mathbf{A}_{att}^{RGB} &= \varphi(\mathbf{a}_1^{RGB}, \cdots, \mathbf{a}_N^{RGB}) = \sum_i w_i \mathbf{a}_i^{RGB}, \\
\mathbf{A}_{att}^{F} &= \varphi(\mathbf{a}_1^{F}, \cdots, \mathbf{a}_N^{F}) = \sum_i w_i \mathbf{a}_i^{F}, \\
\mathbf{A}_{att}^{WF} &= \varphi(\mathbf{a}_1^{WF}, \cdots, \mathbf{a}_N^{WF}) = \sum_i w_i \mathbf{a}_i^{WF}.
\end{aligned} \tag{4}$$

For better readability, we give this new action feature vector $\mathbf{A}_{att}$ a name as attention vector. To emphasize, the attention model directly computes a soft alignment, so that the gradient of the loss function is trained by backpropagation.

3.3 Implementation Details

During the training phase, images from all three modalities (RGB, optical flow and warped optical flow) are cropped to 224×224. We employ cross modality pre-training [12]. Firstly, the spatial stream (ResNet or BN-Inception) is pre-trained on the ImageNet image classification dataset. Subsequently, these pre-trained weights are used to initialize all 3 streams in the ATW. Each stream of the proposed ATW is trained independently. We use a single frame (1) and a stack

of (5) consecutive (warped) optical flow frame as inputs. Based on the standard cross-entropy loss function, the SGD algorithm is used with a mini-batch size of 128 videos. We use an initial learning rate of 0.001 for the spatial stream and 0.005 for both temporal streams. For spatial stream, the learning rate is multiplied by a factor of 0.1 every 2000 iterations. For both temporal streams, the learning rate decay is divided into stages. Learning rates are multiplied by 0.1 at iterations 12000 and 18000. All momentums are fixed at 0.9.

During the testing phase, with each testing video, a fixed number of snippets (80 in our experiments) are uniformly sampled. We use weighted average fusion $(1, 1, 0.5$ for the spatial stream, optical flow stream, and warped optical flow stream, respectively) to generate a per-video prediction.

Pytorch [39] is used in our experiments with optical flow and warped optical flow extracted via OpenCV with CUDA 8.0. To speed up training, 2 NVIDIA Titan Xp GPUs are used.

4 Experiments

Trimmed Action Datasets. We evaluate our approach on two popular action recognition benchmarks, namely UCF-101 [20] and HMDB-51 [21]. The UCF-101 dataset is one of the biggest action datasets containing 13320 videos clips distributed in 101 classes. HMDB-51 dataset is a very challenging dataset with 6766 videos (3570 training and 1530 testing videos) in 51 classes. Evaluation on these two trimmed datasets is performed using average accuracy over three training/testing splits.

Baselines. Throughout the following section, we compare our proposed ATW network with the standard base architecture, mostly two-stream with the single segment of a video ($N = 1$). For network architecture, we choose the traditional BN-Inception [37] for comparison in experiments.

Comparison with Different Consensus Functions. Firstly, we focus on comparing the attention model from two optional consensus functions: (1) max segmental consensus; (2) average segmental consensus. The max and weighted average consensus function is injected at last fully-connected layer, whereas, average consensus can be used after softmax layer. On the other hand, our attention model is set before softmax layer. The experimental performance is summarized in Table 1. We implement these four segmental consensuses with the BN-Inception ConvNet [37] on the first split of UCF-101. The number of segmentation N is set to 4. We use the weighted average fusion of three-stream outputs to generate the video-level prediction. Average segmental consensus performs slightly better than max function. The best result is obtained by the proposed attention model. Thus it can be seen the usage of attention model significantly improves temporal structure for action recognition.

Multi-Segment. Specially, in Table 3, we use RGB modality for training on multi-segment temporal structure with BN-Inception ConvNet [37]. Note that if $N < 3$, the model is oversimplified, and the performance on UCF-101 (split1) has seriously degraded. The attention model boosts the mAP with 85.80% on

Table 1. Exploration of different segmental consensus functions on the UCF-101 dataset (split1).

Consensus Function	Spatial ConvNets	Temporal ConvNets	Two-Stream
Max	85.0%	86.0%	91.6%
Average	85.0%	87.9%	93.4%
Attention Model	**86.7**%	**88.3**%	**94.6**%

Table 2. Experiments of different initialization strategies for initializing the attention layer's parameters and several traditional activation functions on the UCF-101 dataset (split1). Specifically, $weight = 1/N(N = 4)$ equivalent to average consensus.

Initialization	Spatial-Stream	Activation Function	Spatial-Stream
$weight = 1/N$	84.44%	tanh	84.91%
$weight = 1$	85.17%	sigmoid	85.29%
random gaussian	**85.80**%	relu	**85.80**%

the UCF-101 dataset and 53.88% on HMDB-51 dataset (N=4), resulting from the successfully reduced training error. This comparison verifies the effectiveness of the soft attention mechanism on long-range temporal structure.

Parameters Initialization and Activation Function. As we train the proposed ATW CNN, an appropriate initialization of the attention layer's parameters is crucial. We compare different initialization strategies: (1) the weight w_i is set to 1, bias is 0; (2) the weight w_i is set to $\frac{1}{N}$, bias is 0; (3) random Gaussian distribution initialization. In addition, on behalf of finding the most fitting activation functions, we tested several traditional activation functions in the attention layer. As shown in Table 2, on the UCF-101 dataset, 1 for weight and 0 for bias initialization achieves 85.80% on the top of the three.

Comparison with State-of-the-arts. We present a comparison of the performance of Attention-based Temporal Weighted CNN and previous state-of-the-art methods in Table 4, on UCF-101 and HMDB-51 datasets. For that we used a spatial ConvNet pre-trained on ImageNet, the temporal ConvNet was trained by cross-modality pretraining. We choose ResNet [36] for network architecture. As can be seen from Table 4, both our spatial and temporal nets alone outperform the hand-crafted architectures of [18, 40–42] by a large margin. The combination of attention improves the results and is comparable to the very recent state-of-the-art deep models [10, 12, 14, 43–48].

Visualization. To analyze the ability of the proposed attention model and to select key snippets from long-range temporal multi-segment, we visualize what the proposed model has learned on frame-level, which can help to understand the operation of attention interpreting the feature activity. We test our model on several videos to acquire the expected value of the action feature, which can map this attention back to the input temporal dimension. Fig. 3 presents what input

Table 3. Exploration of ATW CNN with more number of segments on the UCF-101 dataset and HMDB-51 dataset (split1).

Dataset	Spatial-Stream Accuracy							
	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8
UCF-101	83.33%	83.89%	84.80%	**85.80%**	85.29%	85.21%	85.04%	85.55%
HMDB-51	50.07%	53.33%	53.01%	**53.88%**	53.33%	55.36%	53.20%	53.14%

Table 4. Comparison of our method with other state-of-the-art methods on the UCF-101 dataset and HMDB-51 dataset.

HMDB-51		UCF-101	
Model	Accuracy	Model	Accuracy
DT [40]	55.9%	DT [40]	83.5%
iDT [18]	57.2%	iDT [18]	85.9%
BoVW [41]	61.1%	BoVW [41]	87.9%
MoFAP [42]	61.7%	MoFAP [42]	88.3%
Two Stream [10]	59.4%	Two Stream [10]	88.0%
VideoDarwin [47]	63.7%	C3D [14]	85.2%
MPR [48]	65.5%	Two stream +LSTM [11]	88.6%
$F_{ST}CN$ (SCI fusion) [43]	59.1%	$F_{ST}CN$ (SCI fusion) [43]	88.1%
TDD+FV [44]	63.2%	TDD+FV [44]	90.3%
LTC [45]	64.8%	LTC [45]	91.7%
KVMF [46]	63.3%	KVMF [46]	93.1%
TSN (3 modalities) [12]	69.4%	TSN (3 modalities) [12]	93.4%
Proposed ATW	**70.5%**	Proposed ATW	**94.6%**

images originally caused an attention value. The first row shows the top ranked four frames with their corresponding attention weights, and the second row shows the lowest ranked four frames. The attention model pays more attention to the relevant frames than irrelevant frames, and this means that the attention model always focuses on the foreground over time.

5 Conclusion

We presented the Attention-based Temporal Weighted Convolutional Neural Network (ATW), which is a deep multi-stream neural network that incorporates temporal attention model for action recognition. It fuses all inputs with a series of data-adaptive temporal weights, effectively reducing the side effect of redundant information/noises. Experimental results verified the advantage of the proposed method. Additionally, our ATW can be used for action classification from untrimmed videos, and we will test our proposed method on other action datasets in our future work.

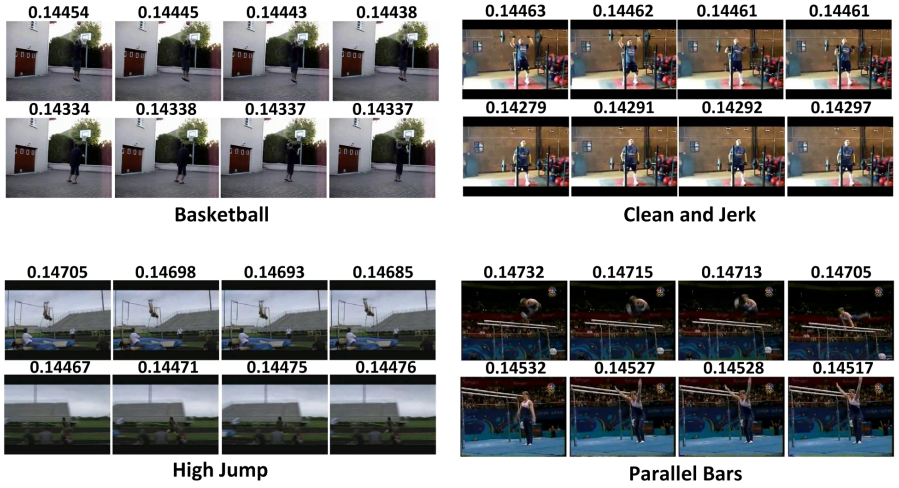

Fig. 3. Visualization of the focus of attention on four videos from UCF-101 dataset over temporal dimension. The model learns to focus on the relevant parts. The attention weight is given on top of each image. The higher the attention weight (w_i) of the frame, the more critical to classify the action.

Acknowledgment

This work was supported partly by NSFC Grants 61629301, 61773312, 91748208 and 61503296, China Postdoctoral Science Foundation Grant 2017T100752, and key project of Shaanxi provinceS2018-YF-ZDLGY-0031.

References

1. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE T-PAMI **35**(1) (2013) 221–231
2. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. (2014) 1725–1732
3. Zhang, Q., Abeida, H., Xue, M., Rowe, W., Li, J.: Fast implementation of sparse iterative covariance-based estimation for source localization. The Journal of the Acoustical Society of America **131**(2) (2012) 1249–1259
4. Ran, L., Zhang, Y., Zhang, Q., Yang, T.: Convolutional neural network-based robot navigation using uncalibrated spherical images. Sensors **17**(6) (2017)
5. Abeida, H., Zhang, Q., Li, J., Merabtine, N.: Iterative sparse asymptotic minimum variance based approaches for array processing. Signal Processing, IEEE Transactions on **61**(4) (2013) 933–944
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 4724–4733
7. Le W., Jianru X., Nanning Z, Gang H.: Global contrast based salient region detection. In: ICCV. (2011) 105–112

8. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: T-PAMI. **39**(10) (2017) 2074-2088
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431-3440
10. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. (2014) 568–576
11. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR. (2015) 4694–4702
12. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. (2016) 20–36
13. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015) 2625–2634
14. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. (2015) 4489–4497
15. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: ICCV. (2015) 3218–3226
16. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR. (2016) 1933–1941
17. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. arXiv preprint arXiv:1801.10111 (2018)
18. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013) 3551–3558
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR. Volume 3. (2004) 32–36
20. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
21. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: Hmdb51: A large video database for human motion recognition. In: High Performance Computing in Science and Engineering. (2013) 571–582
22. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
23. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. (2015) 2048–2057
24. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS. (2014) 2204–2212
25. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011) 3169–3176
26. Laptev, I.: On space-time interest points. IJCV **64**(2-3) (2005) 107–123
27. Ran, L., Zhang, Y., Wei, W., Zhang, Q.: A hyperspectral image classification framework with spatial pixel pair features. Sensors **17**(10) (2017)
28. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. (2012) 1290–1297
29. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR. (2015) 1110–1118

30. Zhang, Q., Hua, G.: Multi-view visual recognition of imperfect testing data. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, ACM (2015) 561–570
31. Zhang, Q., Hua, G., Liu, W., Liu, Z., Zhang, Z.: Can visual recognition benefit from auxiliary information in training? In: Computer Vision – ACCV 2014. Volume 9003 of Lecture Notes in Computer Science., Springer International Publishing (2015) 65–80
32. Zhang, Q., Hua, G., Liu, W., Liu, Z., Zhang, Z.: Auxiliary training information assisted visual recognition. IPSJ Transactions on Computer Vision and Applications **7** (2015) 138–150
33. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: ICCV. (2015) 4507–4515
34. Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. IEEE T-PAMI **35**(11) (2013) 2782–2795
35. Zhang, Q., Abeida, H., Xue, M., Rowe, W., Li, J.: Fast implementation of sparse iterative covariance-based estimation for array processing. In: Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on, IEEE (2011) 2031–2035
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
37. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. (2015) 448–456
38. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009) 248–255
39. Paszke, A., Gross, S., Chintala, S., Chanan, G.: Pytorch (2017)
40. Cai, Z., Wang, L., Peng, X., Qiao, Y.: Multi-view super vector for action recognition. In: CVPR. (2014) 596–603
41. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CVIU **150** (2016) 109–125
42. Wang, L., Qiao, Y., Tang, X.: Mofap: A multi-level representation for action recognition. IJCV **119**(3) (2016) 254–271
43. Sun, L., Jia, K., Yeung, D.Y., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: ICCV. (2015) 4597–4605
44. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR. (2015) 4305–4314
45. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE T-PAMI (2017)
46. Zhu, W., Hu, J., Sun, G., Cao, X., Qiao, Y.: A key volume mining deep framework for action recognition. In: CVPR. (2016) 1991–1999
47. Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T.: Modeling video evolution for action recognition. In: CVPR. (2015) 5378–5387
48. Ni, B., Moulin, P., Yang, X., Yan, S.: Motion part regularization: Improving action recognition via trajectory selection. In: CVPR. (2015) 3698–3706