Content-Aware Attention Network for Action Recognition

Ziyi Liu, Le Wang(✉), and Nanning Zheng

Institute of Artificial Intelligence and Robotics,
National Engineering Laboratory for Visual Information Processing and Applications,
Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China
lewang@xjtu.edu.cn

Abstract. This paper presents a content-aware attention network (CatNet) for action recognition task, which can leverage attention mechanism to aggregate frame-level features into a compact video-level representation. Unlike most previous methods that consider every video frame equally, our CatNet contains an attention module which can adaptively emphasize the representative frames, and thus can benefit the action recognition task. Moreover, the CatNet can take an action video with arbitrary length yet produce a compact video representation with fixed length. The attention module consists of two cascaded blocks, an adaptive attention weighting block and a content-aware weighting block. The experiments are carried on two challenging video action datasets, *i.e.*, the UCF-101 dataset and HMDB-51 dataset. Our method achieves significantly improvements on both datasets compared with existing methods. The results show that our proposed CatNet is able to focus on the representative frames corresponding to a specific action category, and meanwhile significantly improve the recognition performance.

Keywords: Action recognition · Attention mechanism
Content-aware

1 Introduction

Recognizing human actions in various videos is a challenging task, and has received significant attention in the computer vision community [1–12]. From hand-crafted features based methods [4,5], to deep learning based methods [6–12], impressive progresses have been achieved in recent years. Similar with other computer vision tasks, the performance of action recognition has been significantly improved due to the emerging deep learning, especially the convolutional neural networks (CNN), based methods.

However, compared with the successes achieved by CNN in still image classification field [13–15,17], the action recognition task has not been fully explored yet. There still remains many challenges need to be further addressed. In order to

L. Iliadis et al. (Eds.): AIAI 2018, IFIP AICT 519, pp. 109–120, 2018.
https://doi.org/10.1007/978-3-319-92007-8_10

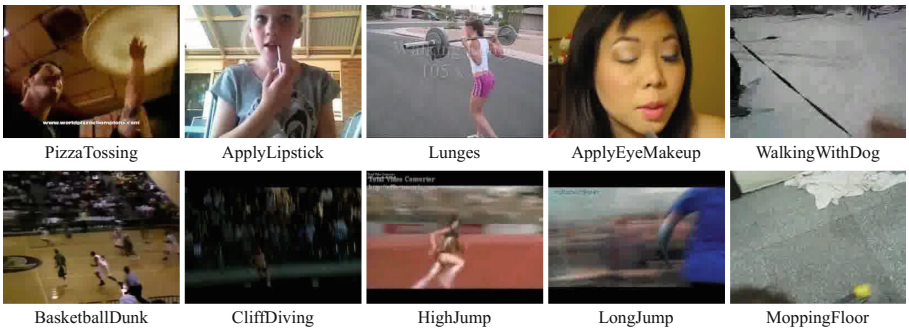

Fig. 1. Some example frames alone with their action category labels that are not suited for the action recognition task. In these cases, it is very difficult to predict the video category by a single frame. The first row consists of frames with semantic ambiguity, which can be easily mistaken for PlayingDaf, BrushingTeeth, CleanAndJerk, ApplyLipstick and Skiing from left to right. The second row consists of frames in poor condition, such as motion blur and poor illumination.

better recognize the action categories inside a variety of videos, the action/video representation should be discriminative and, more importantly, compact. Thus, one of the key issues is *how to construct a discriminative and compact video-level representation*. Hand-crafted feature based methods usually employ encoding methods, such as Fisher Vectors [16], to aggregate local hand-crafted descriptors into a global video representation [4]. With the benefit from the deep CNN, early CNN based methods propose to use a single frame to represent the whole video [6], or feed multiple frames into CNN and aggregate them with average or max pooling strategy [9,18]. In addition, [19] proposed to employ a long short-term memory (LSTM) network upon the CNN, which can model the temporal correlations among frames into a fixed-length representation. To further explore both the spatial and temporal correlations among video frames simultaneously, [20] first proposed the 3D convolution pipeline to handle related tasks in videos.

We argue that to effectively aggregate frame-level descriptors and construct a compact video-level representation, an adaptive content-aware aggregation method is vital. The motivation behind this idea is quite intuitive, *i.e.*, there should exist a subset of the video frames, which are more tightly related to the action category. Thus, we should emphasize these representative frames when aggregating the frame-level descriptors, in order to make the aggregated video-level representation discriminative. Figure 1 gives some example frames alone with their action category labels that are not suitable for the recognition task, due to semantic ambiguity or motion blur. These "bad" samples will introduce noisy information into the inference stage of the neural network [21]. These ambiguities should be suppressed during the inference process. Therefore, it is a natural sense to utilize the attention mechanism for action recognition, which can enable the whole network to focus on the representative frames and suppress the noises.

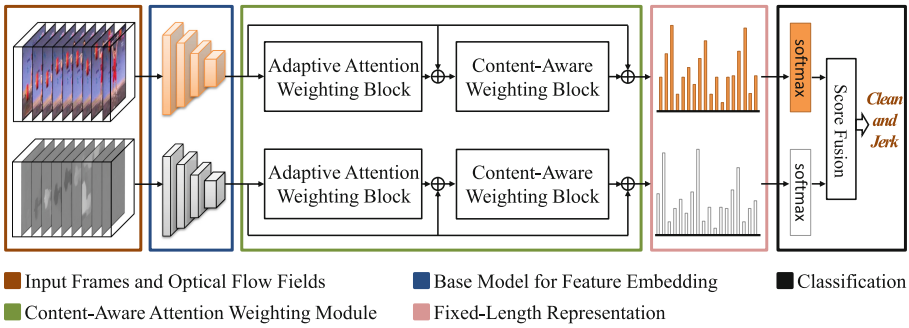

Fig. 2. Content-aware attention network. The inputs are video frames and optical flows. First, they are embedded by a CNN (base model). Then, the extracted features are aggregated by the proposed content-aware attention weighting module, and a fixed-length representation can be obtained. Finally, we use this representation to classify the input video, and adopt score fusion to obtain the prediction of the video.

We propose a content-aware attention network (CatNet) embedding an effective attention module to identify the representative frames from the noisy ones. The framework of the CatNet is illustrated in Fig. 2. The attention module consists of two cascaded blocks, *i.e.*, an adaptive attention weighting block which can adaptively weight all frames fed into the CatNet based on their extracted features, and a content-aware weighting block which is added to constrain the aggregation weights to be more consistent with the video content. To capture both the appearance and motion information, a standard two-stream structure is adopted, where each stream can be trained in an end-to-end manner.

The main contributions of this paper are:

- We propose a novel structure, namely CatNet, for action recognition, which achieves state-of-the-art performance.
- We introduce the attention mechanism to action recognition and validate it is beneficial for the action recognition task.

The rest of the paper is organized as follows. Section 2 reviews previous work related to action recognition. Section 3 presents the content-aware attention network for action recognition. Section 4 introduces the experiments and discussions. Section 5 presents the conclusion and future work.

2 Related Work

In this section, we review the related work in action recognition, mainly including hand-crafted feature based methods, and deep learning based methods.

2.1 Hand-Crafted Feature Based Methods

Before the prevalent of deep learning, hand-crafted feature based methods have dominated the action recognition filed [4,22–24]. Plenty of image local descriptors have been extended to the video domain, such as 3D-sift [25], HOG3D [26], and motion boundary histograms (MBH) [27]. The improved Dense Trajectory (iDT) [4] achieves state-of-the-art performance among these hand-crafted feature based methods. The iDT consists of multiple local descriptors extracted along with the dense trajectories where camera motion is compensated. To perform action recognition, the descriptors are then aggregated into a video-level representation by using encoding methods, such as Fisher vector [16].

2.2 Deep Learning Based Methods

Aggregating Frame-Level Features. Recently, the CNN based image descriptors have emerged as the state-of-the-art generic descriptors for visual recognition. To obtain a discriminative frame-level representation, the recent work on action recognition almost always extract the CNN feature as the frame-level descriptor, and then aggregate them into a global video-level representation. The aggregation methods can be roughly divided into two categories. The first one is to employ recurrent neural network (RNN) upon a frame-level feature extractor, such as LSTM [19]. By plugging RNN on top of CNN, temporal correlations within the action video can be easily captured, and a compact fixed-length video representation is then obtained. The other one aims to aggregate the frame-level features via different pooling methods, such as the average or max pooling over time [9], the vector of locally aggregated descriptors (VLAD) [10,28], and the temporal pyramid pooling (TPP) [29]. All these methods treat the frame-level features equally during aggregation, which may inevitably over-weight the noisy frames. To eliminate the negative influence from noisy frames, we propose to embed an attention module into action recognition, which can adaptively emphasize the representative frames while suppressing the influence from noisy frames.

Spatio-temporal CNN. The spatio-temporal CNN was first proposed in [20] and named 3D CNN, and later a number of its variants were proposed [8,11, 30,31]. The 3D CNN based action recognition methods take the video clip as input, and aim to model both the spatial and temporal correlations among the video content. Considering the 3D convolution brings extra kernel parameters, to fully train 3D CNN model usually requires massive video data and is time consuming, and thus 3D CNN is unsuitable to handle small dataset [11,18].

3 Content-Aware Attention Network

This section details the proposed content-aware attention network (CatNet). As shown in Fig. 2, our proposed CatNet takes a video with arbitrary length

as input, and outputs a fixed-length video representation for subsequent action recognition task. The frame-level feature embedding is based on the CNN model, which is followed by an adaptive attention weighting block and a content-aware weighting block. These two blocks enable the CatNet to adaptively emphasize the representative frames and meanwhile suppress the noisy ones.

3.1 Frame-Level Feature Embedding

The frame-level features are extracted using the deep CNN, which embeds each frame of an action video to a fixed-length vector. Here, we adopt the Inception with Batch Normalization [32] as the feature extractor (base model). Note that our proposed attention module is not limited to a specific CNN model, other CNN models can also be used. The extracted d-dimension CNN features are first normalized by $L2$ norm and then fed into the attention module. Formally, given a video $\mathbf{V} = \{f_t\}_{t=1}^{T}$ with T frames, where f_t denotes the tth frame, the feature embedding can be formulated as

$$\mathbf{x}_t = \mathcal{F}(f_t; \mathbf{W}), \tag{1}$$

$$\bar{\mathbf{x}}_t = \frac{\mathbf{x}_t}{||\mathbf{x}_t||}, \tag{2}$$

where $\mathcal{F}$ denotes the base model and $\mathbf{W}$ denotes the parameters of $\mathcal{F}$. $\mathbf{x}_t \in R^d$ represents the extracted d-dimensional feature of f_t. $\bar{\mathbf{x}}_t$ is a normalized feature.

3.2 Adaptive Attention Weighting Block

As obtained the feature $\mathbf{x}_t$ of each frame f_t by feature embedding, our goal is to obtain a fixed-length video representation for video $\mathbf{V}$ by aggregating its frame-level descriptors $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$. Our adaptive attention weighting block first computes a corresponding weight w_t for each frame f_t, and then aggregates the frame-level descriptors into a fix-length video-level representation by

$$\mathbf{v} = \sum_{t=1}^{T} w_t \bar{\mathbf{x}}_t. \tag{3}$$

Here, the key is how to compute an appropriate weight w_t for $\mathbf{x}_t$ according to its importance. If $w_t = \frac{1}{T}$, then this block will degrade to average pooling.

There are two issues need to be considered when building the adaptive attention weighting block. First, the block should be able to handle videos with arbitrary lengths. Second, the block should be differentiable, *i.e.*, can be easily plugged into the current networks to perform an end-to-end training. Our proposed solution is to introduce a learnable kernel $\mathbf{k}$ with the same dimension as $\mathbf{x}$. Then, w_t can be calculated by

$$w_t = \mathbf{k}^T \bar{\mathbf{x}}_t. \tag{4}$$

Here, $\mathbf{k}$ actually serves as a scoring function. It is expected that if $\mathbf{x}_t$ is discriminative, w_t will be larger, and vice versa. In this way, the video representation $\mathbf{v}$ calculated by Eq. (3) will adaptively emphasize the representative frames and suppress the noisy frames during aggregation.

3.3 Content-Aware Weighting Block

To leverage context information, which is popular in image segmentation field [33,34], a content-aware weighting block is introduced to let the attention module select the discriminative features by considering the content throughout the video instead of single frames. Inspired by [35] for language modeling, we borrow the ideas from [35,36] and adjust it to the action recognition task. The content-aware weighting block can be formulated as

$$\mathbf{k}_c = \mathcal{C}(\mathbf{v}; (\mathbf{W}_c, \mathbf{b})) = \sigma(\mathbf{W}_c \mathbf{v} + \mathbf{b}), \tag{5}$$

$$w_t^c = \mathbf{k}_c^T \mathbf{x}_t, \tag{6}$$

$$\mathbf{v}_c = \sum_{t=1}^{T} w_t^c \mathbf{x}_t, \tag{7}$$

where σ denotes the sigmoid function and $\mathbf{k}_c$ serves as a new weighting kernel which is content-aware. $\mathbf{W}_c \in R^{d \times d}$ and $\mathbf{b} \in R^d$ are trainable parameters of this block. $\mathbf{v}_c$ is the final fixed-length representation of the input video.

3.4 Two-Stream Structure

It is critical to capture temporal information for action recognition. A common way to capture temporal information is adopting the two-stream structure [6,37]. We also employ this validated effective structure to combine the spatial and temporal features. Each of the two streams is constructed as described above, and can be trained in an end-to-end way. To fuse the scores from these two streams, the simplest weighted average fusion strategy is utilized.

4 Experiments and Discussions

We evaluate the performance of the proposed CatNet on two challenging action datasets, including UCF-101 dataset [38] and HMDB-51 dataset [39]. The UCF-101 dataset contains 13320 action videos in 101 action categories. Our evaluation on UCF-101 dataset follows the scheme of the THUMOS-13 challenge [40]. We use all the three training/testing splits, and report the average accuracy on them. The HMDB-51 dataset contains 6766 videos in 51 action categories. We follow the standard evaluation scheme with three training/testing splits, and report the average accuracy on them. We proceed to introduce the implementation details of our method, and then explore the efficacy of our attention module and compare with baseline method. Finally, our CatNet is compared with the state-of-the-art methods.

4.1 Implementation Details

We use the stochastic gradient descent optimizer to train the network, by setting momentum to be 0.9 and batch size to be 32. We implement a two-stream CNN framework with the spatial stream for RGB image inputs and temporal stream for optical flow inputs, as multi-modality inputs offer more information [9,41, 43,45]. We choose the Inception with Batch Normalization (BN-Inception) [32] as the building block for both spatial and temporal stream, because of its good balance between accuracy and efficiency. We adopt the partial BN with extra dropout layer proposed in [9] to avoid over-fitting. For the spatial stream, the weights are initialized by pre-trained models from ImageNet [42]. While for the temporal stream, we use the cross modality pre-training proposed in [9]. As to data augmentation, we employ the scale jittering technique [44] and random horizontal flipping. For the computation of optical flow, we use the TVL1 optical flow algorithm [46], which is implemented in OpenCV with CUDA. When fusing the scores from the two streams, we average the final prediction score of each stream with a weight 1 for spatial stream and a weight 1.5 for temporal stream.

4.2 Evaluation of the Proposed Attention Module

To validate the efficacy of the proposed attention module, we compared it with a baseline aggregation strategy, *i.e.* average pooling. The average accuracies for action recognition of them are summarized in Table 1. They showed that our aggregation strategy outperforms the average pooling on both of the UCF-101 dataset [38] and HMDB-51 dataset [39]. This clearly manifests that the proposed attention module can improve the action recognition performance.

It can be also seen that the performance improvement on HMDB-51 is larger than that on UCF-101. This is mainly because the videos in HMDB-51 contain more noisy frames that should be suppressed. Note that the improvement on temporal stream is less than that on spacial stream. This is because the optical flow fields contain less noise and are more discriminative than RGB images, especially for the action recognition task.

Table 1. The average accuracies by using average pooling and our aggregation strategy on the UCF-101 dataset [38] and HMDB-51 dataset [39].

Stream	UCF-101	HMDB-51	Stream	UCF-101	HMDB-51
Spacial (Avg)	85.5	50.9	Spacial (Ours)	**86.1**	**51.6**
Temporal (Avg)	87.9	61.1	Temporal (Ours)	**88.0**	**61.4**
Fusion (Avg)	93.4	67.8	Fusion (Ours)	**93.8**	**68.6**

Besides, to further verify the efficacy of the proposed attention module, we visualize what has been learnt by the attention module. Figure 3 presents some example frames sorted by their attention weights. They showed that the proposed

Fig. 3. Some example frames sorted by their attention weights from high (left) to low (right) along with their action category labels. The frames of the left part are from the UCF-101 dataset, and the frame of the right part are from the HMDB-51 dataset. We can see that the representative frames are assigned with higher attention weights compared with frames in poor conditions, such as motion blur (*e.g.*, Biking and Knitting), shot changing (*e.g.*, Smile), irrelevant clips (*e.g.*, Shoot Gun and Run), partial observation (*e.g.*, Brushing Teeth and Haircut), and semantic ambiguity (*e.g.*, Tennis Swing and Hug).

attention module can automatically pay more attention to the representative frames while suppressing the frames in poor condition, without providing any extra supervision during training. For example, in video "Hug", it is difficult to judge it is either hugging or hand shaking by the frames with low attention. However, the frames with higher attention can clearly reveal the hugging action. Similar examples can be found in other videos shown in Fig. 3.

Moreover, it is interesting to observe that our attention module assigns high attention to frames containing action "Stand" in the video labelled as "Stand", while it assigns low attention to frames containing similar action in other videos (as shown in "Sit" and "Stand" videos in Fig. 3). This validates that our attention module is content-aware.

4.3 Evaluation of the Proposed CatNet for Action Recognition

After comparing with baseline methods to validate the efficacy of our proposed attention module, it is also very important to compare our CatNet with other

Table 2. The average accuracies of our CatNet and other state-of-the-art methods on the UCF-101 dataset [38] and HMDB-51 dataset [39].

Method	UCF101	HMDB51
iDT+FV [4]	85.9	57.2
Spatio-Temporal ConvNet [18]	65.4	-
LRCN [19]	82.9	-
C3D [8]	85.2	-
Factorized ConvNet [30]	88.1	59.1
Two-Stream ConvNet [6] (VGG-M)	88.0	59.4
Two-Stream + LSTM [7] (GoogLeNet)	88.6	-
Two-Stream Fusion [37] (VGG-16)	92.5	65.4
Transformations [47]	92.4	62.0
TPP [29]	89.1	63.1
ActionVLAD [10]	92.7	66.9
LTC [31]	92.7	67.2
CatNet Spacial-Stream	86.1	51.6
CatNet Temporal-Stream	88.0	61.4
CatNet Two-Stream	**93.8**	**68.6**

state-of-the-art methods for action recognition. We present the average accuracies of CatNet and a variety of recently proposed action recognition methods on both UCF-101 dataset [38] and HMDB-51 dataset [39] in Table 2.

The results showed that, the two-stream version of CatNet significantly outperforms other state-of-the-art methods, although we adopt the simplest weighted average fusion strategy. Moreover, both the single stream versions of our method can achieve competitive performances, when comparing with the hand-crafted based methods (*e.g.*, [4]), LSTM based methods (*e.g.*, [19]), and 3D CNN based methods (*e.g.*, [8]).

5 Conclusion and Future Work

This paper proposed a content-aware attention network for action recognition, which leverages an attention module to aggregate the frame-level features into a compact video-level representation. Experimental results on the UCF-101 dataset and HMDB-51 dataset validated the efficacy of the proposed attention module and also the whole action recognition method, and demonstrated that the attention module can lead the content-aware attention network to adaptively emphasize the representative frames while suppressing the noisy frames.

For future work, we aim to extend our content-aware attention network to handle untrimmed action videos, where we argue that the attention module can play a more significant role. We will conduct extensive experiments on

untrimmed videos to fully explore the efficacy of the attention module. Moreover, we only validate that the attention module can help improving the action recognition performance in this paper. In future work, we will extend the attention module to action localization or action segmentation task.

Acknowledgments. This work was supported partly by NSFC Grants 61629301, 61773312, 91748208, and 61503296, China Postdoctoral Science Foundation Grant 2017T100752, and key project of Shaanxi province S2018-YF-ZDLGY-0031.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
2. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P.: Advances in human action recognition: a survey. arXiv:1501.05964 Computer Vision and Pattern Recognition (2015)
3. Herath, S., Harandi, M.T., Porikli, F.: Going deeper into action recognition. Image Vis. Comput. **60**, 4–21 (2017)
4. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV, pp. 3551–3558 (2013)
5. Lan, Z., Lin, M., Li, X., Hauptmann, A.G., Raj, B.: Beyond Gaussian pyramid: multi-skip feature stacking for action recognition. In: CVPR, pp. 204–212 (2015)
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576 (2014)
7. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: CVPR, pp. 4694–4702 (2015)
8. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: CVPR, pp. 4489–4497 (2015)
9. 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: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
10. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: ActionVLAD: learning spatio-temporal aggregation for action classification. In: CVPR (2017)
11. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR (2017)
12. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. arXiv preprint arXiv:1801.10111 (2018)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
15. Ran, L., Zhang, Y., Wei, W., Zhang, Q.: A hyperspectral image classification framework with spatial pixel pair features. Sensors **17**(10) (2017)
16. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR, pp. 1–8. IEEE (2007)
17. Ran, L., Zhang, Y., Zhang, Q., Yang, T.: Convolutional neural network-based robot navigation using uncalibrated spherical images. Sensors **17**(6) (2017)

18. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732 (2014)
19. 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, pp. 2625–2634 (2015)
20. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
21. Lan, Z., Zhu, Y., Hauptmann, A.G., Newsam, S.: Deep local video feature for action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1219–1225. IEEE (2017)
22. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**(2–3), 107–123 (2005)
23. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72. IEEE (2005)
24. Sadanand, S., Corso, J.J.: Action bank: a high-level representation of activity in video. In: CVPR, pp. 1234–1241 (2012)
25. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, pp. 357–360. ACM (2007)
26. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC, pp. 275–1. British Machine Vision Association (2008)
27. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_33
28. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetvLAD: CNN architecture for weakly supervised place recognition. In: CVPR, pp. 5297–5307 (2016)
29. Wang, P., Cao, Y., Shen, C., Liu, L., Shen, H.T.: Temporal pyramid pooling-based convolutional neural network for action recognition. IEEE Trans. Circuits Syst. Video Technol. **27**(12), 2613–2622 (2017)
30. Sun, L., Jia, K., Yeung, D.Y., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: CVPR, pp. 4597–4605 (2015)
31. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
32. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456 (2015)
33. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue. In: ICCV, pp. 105–112 (2011)
34. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. IEEE Trans. Pattern Anal. Mach. Intell., 2074–2088 (2017)
35. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. arXiv preprint arXiv:1612.08083 (2016)
36. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. arXiv:1706.06905 (2017)
37. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR, pp. 1933–1941 (2016)

38. Soomro, K., Roshan Zamir, A., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. In: CRCV-TR-12-01 (2012)
39. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV, pp. 2556–2563. IEEE (2011)
40. Jiang, Y.G., Liu, J., Roshan Zamir, A., Laptev, I., Piccardi, M., Shah, M., Sukthankar, R.: THUMOS challenge: action recognition with a large number of classes (2013). http://crcv.ucf.edu/ICCV13-Action-Workshop/
41. Zhang, Q., Hua, G.: Multi-view visual recognition of imperfect testing data. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp. 561–570. ACM (2015)
42. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)
43. Zhang, Q., Hua, G., Liu, W., Liu, Z., Zhang, Z.: Auxiliary training information assisted visual recognition. IPSJ Trans. Comput. Vision Appl. **7**, 138–150 (2015)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
45. Zhang, Q., Hua, G., Liu, W., Liu, Z., Zhang, Z.: Can visual recognition benefit from auxiliary information in training? In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9003, pp. 65–80. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16865-4_5
46. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Pattern Recognition, pp. 214–223 (2007)
47. Wang, X., Farhadi, A., Gupta, A.: Actions transformations. In: CVPR (2016)