SciForum
MOL2NET

Trajectory-pooled Spatial-temporal Structure of Deep Convolutional Neural Networks for Video Event Recognition

Yonggang Li[1,2], Xiaoyi Wan[1], Zhaohui Wang[1], Shengrong Gong[5,1,*] , Chunping Liu[1,3,4,*]

1.School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006
2. College of mathematics physics and information engineering, Jiaxing University, Jiaxing, Zhejiang, 314001
3. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin,130012
4. Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, Jiangsu,210046
5. School of Computer Science and Engineering, Changshu Institute of Science and Technology, Changshu, Jiangsu,215500

* Corresponding author email: shrgong@suda.edu.cn, cpliu@suda.edu.cn

Abstract:

Video event recognition according to content feature faces great challenges due to complex scenes and blurred actions for surveillance videos. To alleviate these challenges, we propose a spatial-temporal structure of deep Convolutional Neural Networks for video event recognition. By taking advantage of spatial-temporal information, we fine-tune a two-stream Network, then fuse spatial and temporal feature at a convolution layer using a conv fusion method to enforce the consistence of spatial-temporal structure. Based on the two-stream Network and spatial-temporal layer, we obtain a triple-channel structure. We pool the trajectory to the fused convolution layer, as the spatial-temporal channel. At the same time, trajectory-pooling is conducted on one spatial convolution layer and one temporal convolution layer, to form another two channels: spatial channel and temporal channel. To combine the merits of deep feature and hand-crafted feature, we implement trajectory-constrained pooling to HOG and HOF features. Trajectory-pooled HOG and HOF features are concatenated to spatial channel and temporal channel respectively. A fusion method on triple-channel is designed to obtain the final recognition result. The experiments on two surveillance video datasets including VIRAT 1.0 and VIRAT 2.0, which involves a suit of challenging events, such as person loading an object to a vehicle, person opening a vehicle trunk, manifest that the proposed method can achieve superior performance compared with other methods on these event benchmarks.

Our contribution including:

1. We utilize two-stream Network to extract spatial feature and temporal feature, and fuse spatial and temporal feature at a convolution layer by using a conv fusion method, which can enforce the consistence of spatial-temporal structure.

2. To combine the merits of deep feature and hand-crafted feature, we implement trajectory-constrained pooling to HOG and HOF features, which can more accurately represent local feature of the happening actions.

3. We design a trajectory-pooled triple-channel structure. Triple-stream structure can model the spatial-temporal information better.

4. We conduct an extensive set of experiments, which demonstrates that our method can obtain excellent performance.

[1] Wang X, Ji Q. Video event recognition with deep hierarchical context model[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4418-4427.
[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.
[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[5] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
[6] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4305-4314.
[7] Wu Z, Jiang Y G, Wang X, et al. Fusing Multi-Stream Deep Networks for Video Classification[J]. arXiv preprint arXiv:1509.06086, 2015.
[8] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems. 2014: 568-576.
[9] Wang H, Schmid C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 3551-3558.
[10] Wang L, Qiao Y, Tang X. MoFAP: A multi-level representation for action recognition[J]. International Journal of Computer Vision, 2015: 1-18.
[11] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv preprint arXiv:1507.02159, 2015.
[12] Wang L, Wang Z, Du W, et al. Object-scene convolutional neural networks for event recognition in images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 30-35.
[13] Sánchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice[J]. International journal of computer vision, 2013, 105(3): 222-245.
[14] Xu Z, Yang Y, Hauptmann A G. A discriminative CNN video representation for event detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1798-1807.
[15] Wu Z, Wang X, Jiang Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 461-470.
[16] Feichtenhofer C, Pinz A, Zisserman A. Convolutional Two-Stream Network Fusion for Video Action Recognition[J]. arXiv preprint arXiv:1604.06573, 2016.
[17] Chang X, Yu Y L, Yang Y, et al. They Are Not Equally Reliable: Semantic Event Search using Differentiated Concept Classifiers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
[18] Cheng Y, Fan Q, Pankanti S, et al. Temporal sequence modeling for video event detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2227-2234.

[19] Lai K T, Felix X Y, Chen M S, et al. Video event detection by inferring temporal instance labels[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2251-2258.
[20] Li W, Yu Q, Divakaran A, et al. Dynamic pooling for complex event recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 2728-2735.
[21] Feichtenhofer C, Pinz A, Wildes R. Dynamic Scene Recognition with Complementary Spatiotemporal Features[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, Dec. 2016, vol. 38:2389-2401.
[22] Zhang X, Zou J, He K, et al. Accelerating very deep convolutional networks for classification and detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, Oct. 2016, vol. 38:1943-1955.

Acknowledgements
This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61272258, 61170124, 6130129, 61272005), Provincial Natural Science Foundation of Jiangsu (Grant No. BK20151254, BK20151260), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Collaborative Innovation Center of Novel Software Technology and Industrialization.