SciForum
MOL2NET

Fusing Augmented Spatio-temporal Features for Action Recognition

Rui Ge[1], Xiaoyi Wan[1], Yi Ji[1], Chunping Liu[1,2,3], Shengrong Gong[4,1],

1. School of computer science and technolgoy, Soochow University, Suzhou 215000
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministryof Education, Jilin University, Changchun 130012*
3. Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210046
4. School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500

* Corresponding author email: Shengrong Gong shrgong@suda.edu.cn; Chunping Liu, email: cpliu@suda.eud.cn

Abstract:

Visual features are vitally important for action recognition in videos. However, traditional features fail to effectively recognize actions for two reasons: on one hand, spatial features are not powerful enough to capture appearance information of complex video actions; on the other hand, important temporal details are always ignored when pooling and encoding. In this paper, we present a new architecture that fuses multiple augmented spatio-temporal features. In order to strengthen spatial features, we conduct crop and horizontal flip on original frame images. Then we feed these processed images into deep Two-Stream network to produce robust spatial representations. To get powerful temporal features, we employ fourier temporal pyramid (FTP) to capture three different levels of video context, including short-term level, medium-range level, and global-range level. At last, we fuse these augmented spatio-temporal features using canonical correlation analysis (CCA) method, which is capable to capture the correlation between these features. Experimental results on UCF101 dataset show that our method can achieve excellent performance for action recognition.

Conclusions

In this paper, we propose to fuse multiple augmented spatio-temporal features for better action recognition. The enhanced spatial features are extracted by feeding multiple crop images into VGG networks. Then through a three-level FTP, the features are capable to capture different level temporal context information about action. Finally, the method is capable to improve the performance effectively by CCA fusion. Our experimental results show that the model can achieve comparable accuracy to the state of the art methods.

References

[1] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]. Computer Vision and Pattern Recognition, 2011: 3169 - 3176.

[2] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[C]. Advances in Neural Information Processing Systems, 2014: 568 - 576.

[3] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91 - 110.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005: 886 - 893.

[5] Laptev I, Marszalek M, Schmid C. Learning realistic human actions from movies[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1 - 8.

[6] Laptev, I. On space-time interest points[J]. International Journal of Computer Vision, 2005, 64(2-3): 107 - 123.

[7] Wang H, Schmid C. Action recognition with improved trajectories[C]. Proceedings of the IEEE International Conference on Computer Vision, 2013: 3551 - 3558.

[8] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. European Conference on Computer Vision, 2014: 818 - 833.

[9] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[10] Wang J, Liu Z, Wu Y. Learning actionlet ensemble for 3D human action recognition[M]. Human Action Recognition with Depth Cameras. Springer International Publishing, 2014: 11 - 40.

[11] Cai Z, Wang L, Peng X, Qiao Y. Multi-view super vector for action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 596 - 603.

[12] Karpathy A, Toderici G, Shetty S. Large-scale video classification with convolutional neural networks. Computer Vision and Pattern Recognition, 2014: 1725 - 1732.

[13] 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.

[14] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv preprint arXiv:1507.02159, 2015.

[15] Wang L M, Qiao Y, Tang X. Motionlets: Mid-level 3d parts for human motion recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2674 - 2681.

[16] Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2625 - 2634.

[17] Ma S, Sigal L, Sclaroff S. Learning activity progression in lstms for activity detection and early detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1942 - 1950.

[18] Li Y, Li W, Mahadevan V, et al. VLAD3: Encoding Dynamics of Deep Features for Action Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[19] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of the 22nd ACM international conference on Multimedia, 2014: 675 - 678.

[20] Sun Q S, Zeng S G, Liu Y, et al. A new method of feature fusion and its application in image recognition[J]. Pattern Recognition, 2005, 38(12): 2437 - 2448.

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.