Large-scale Continuous Gesture Recognition Using Convolutional Neutral Networks

August 13, 2016

1 Team details

- Team name: AMRL
- Team leader name: Pichao Wang
- Team leader address, phone number and email: 3/68, Robsons Road, Wollongong, Australia, (+61)405278871, pw212@uowmail.edu.au
- Rest of the team members: Wanqing Li, Song Liu, Yuyao Zhang, Zhimin Gao and Philip Ogunbona
- Affiliation: University of Wollongong

2 Contribution details

- Title of the contribution:Large-scale Continuous Gesture Recognition Using Convolutional Neutral Networks
- Final score: 0.2655
- General method description:This paper addresses the problem of continuous gesture recognition with convolutional neutral networks (ConvNets) using depth maps sequences. Unlike the common isolated recognition scenario, the gesture boundaries are here unknown, and one has to solve two problems: segmentation and recognition. For segmentation, we first obtained the begin and end frames of each gesture based on quantity of movement (QOM) and then proposed one compact representations for depth sequences, called Improved Depth Motion Map (IDMM), which converts each depth sequence into one image, to recognize the gestures using ConvNets. This method enables the use of existing ConvNets models directly on video data with fine-tuning, without introducing much parameters to be learned.
- References:

References

[1] F. Jiang, S. Zhang, S. Wu, Y. Gao, and D. Zhao, “Multi-layered gesture recognition with kinect,” *Journal of Machine Learning Research*, vol. 16, no. 2, pp. 227–254, 2015.

[2] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 9–14.

[3] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[4] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” in *Proc. ACM international conference on Multimedia (ACM MM)*, 2012, pp. 1057–1060.

[5] O. Oreifej and Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 716–723.

[6] X. Yang and Y. Tian, “Super normal vector for activity recognition using depth sequences,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 804–811.

[7] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, “Mining mid-level features for action recognition based on effective skeleton representation,” in *Proc. International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2014, pp. 1–8.

[8] C. Lu, J. Jia, and C.-K. Tang, “Range-sample depth feature for action recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 772–779.

[9] R. Vemulapalli and R. Chellappa, “Rolling rotations for recognizing human actions from 3d skeletal data,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1–9.

[10] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. O. Ogunbona, “Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring,” in *Proc. ACM international conference on Multimedia (ACM MM)*, 2015, pp. 1119–1122.

[11] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, “Action recognition from depth maps using deep convolutional neural networks,” *Human-Machine Systems, IEEE Transactions on*, vol. 46, no. 4, pp. 498–509, 2016.

[12] P. Wang, Z. Li, Y. Hou, and W. Li, “Action recognition based on joint trajectory maps using convolutional neural networks,” in *Proc. ACM international conference on Multimedia (ACM MM)*, 2016, pp. 1–5.

[13] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[14] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4041–4049.

[15] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *The 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[16] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from rgb-d data using bag of features," *Journal of Machine Learning Research*, vol. 14, pp. 2549–2582, 2013.

[18] J. Wan, V. Athitsos, P. Jangyodsuk, H. J. Escalante, Q. Ruan, and I. Guyon, "Csmmi: Class-specific maximization of mutual information for action and gesture recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3152–3165, July 2014.

[19] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan, "Principal motion components for one-shot gesture recognition," *Pattern Analysis and Applications*, pp. 1–16, 2015.

[20] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.

[21] Y. M. Lui, "Human gesture recognition on product manifolds," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3297–3321, Nov 2012.

[22] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from rgbd images," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 7–12.

[23] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702.

[24] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[25] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[27] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[28] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] B. R. Abidi, Y. Zheng, A. V. Gribok, and M. A. Abidi, "Improving weapon detection in single energy X-ray images through pseudocoloring," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 36, no. 6, pp. 784–796, 2006.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding." in *Proc. ACM international conference on Multimedia (ACM MM)*, 2014, pp. 675–678.

[32] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1626–1639, Aug 2016.

- Representative image / diagram of the method:

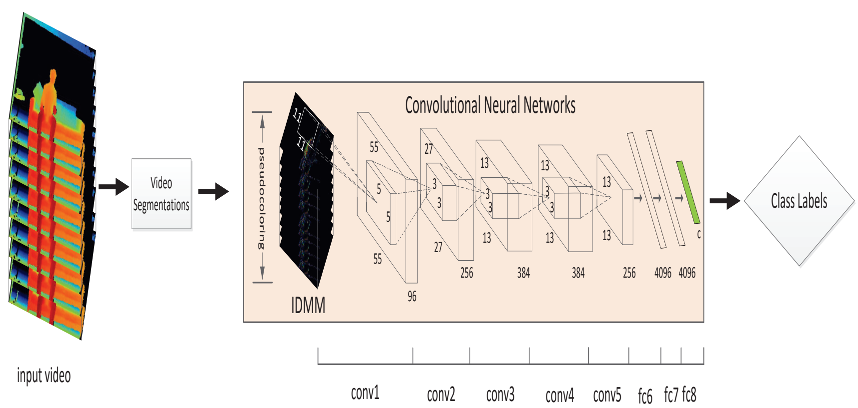

Figure 1: The framework for proposed method.

- Describe data preprocessing techniques applied (if any): None

3 Visual Analysis

3.1 Gesture Recognition (or/and Spotting) Stage

3.1.1 Features / Data representation

Describe features used or data representation model FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any): Deep learned features using ConvNets.

3.1.2 Dimensionality reduction

Dimensionality reduction technique applied FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any):None

3.1.3 Compositional model

Compositional model used, i.e. pictorial structure FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any):None

3.1.4 Learning strategy

Learning strategy applied FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any): Using ConvNets to learn.

3.1.5 Other techniques

Other technique/strategy used not included in previous items FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any): We used Improve Depth Motion Maps (IDMM) for the input of ConvNets to learn the features.

3.1.6 Method complexity

Method complexity FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE: Real-time

3.2 Data Fusion Strategies

List data fusion strategies (how different feature descriptions are combined) for learning the model / network: Single frame, early, slow, late. (if any): None

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any): We use pre-trained models on ILSVRC-2012 for Alexnet.
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any): We only used depth data.

- Qualitative advantages of the proposed solution: simple yet effective
- Results of the comparison to other approaches (if any):

Table 1: Comparative accuracy of proposed method and baseline methods on the ChaLearn LAP ConGD Dataset.

Method	Set	Mean Jaccard Index $\overline{J_S}$
MFSK	Validation	0.0918
MFSK+DeepID	Validation	0.0902
Proposed Method	Validation	**0.2403**
MFSK	Testing	0.1464
MFSK+DeepID	Testing	0.1435
Proposed Method	Testing	**0.2655**

- Novelty degree of the solution and if is has been previously published: incremental

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements): Matlab + Caffe + Python. GPU memory required no less than 3 GB.
- Human effort required for implementation, training and validation? Easy.
- Training/testing expended time? Less than one hour.
- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?: Very good and hope to see more contestants.