Large-scale Isolated Gesture Recognition Using Convolutional Neutral Networks

August 16, 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, Zhimin Gao, Chang Tang and Philip Ogunbona
- Affiliation: University of Wollongong

2 Contribution details

- Title of the contribution:Large-scale Isolated Gesture Recognition Using Convolutional Neutral Networks
- Final score : 55.57%
- General method description: In this paper we proposed three simple, compact yet effective representations of depth sequences for gesture recognition in the context of convolutional neutral networks (ConvNets). The three representations are called Dynamic Depth Image (DDI), Dynamic Depth Normal Image (DDNI) and Dynamic Depth Motion Normal Image (DDMNI). They are all based on bidirectional rank pooling method converting the depth sequences into images. Such representations enables the use of existing ConvNets models directly on video data with fine-tuning without introducing large parameters to learn. The three representations represent the posture and motion in different levels and they are complementary to each other and improve the recognition accuracy largely.
- References:

References

[1] S. Escalera, V. Athitsos, and I. Guyon, "Challenges in multimodal gesture recognition," *Journal of Machine Learning Research*, vol. 17, no. 72, pp. 1–54, 2016.

[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] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 1351–1357.

[7] 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.

[8] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 742–757.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1–9.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[27] 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.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] 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.

[30] 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:

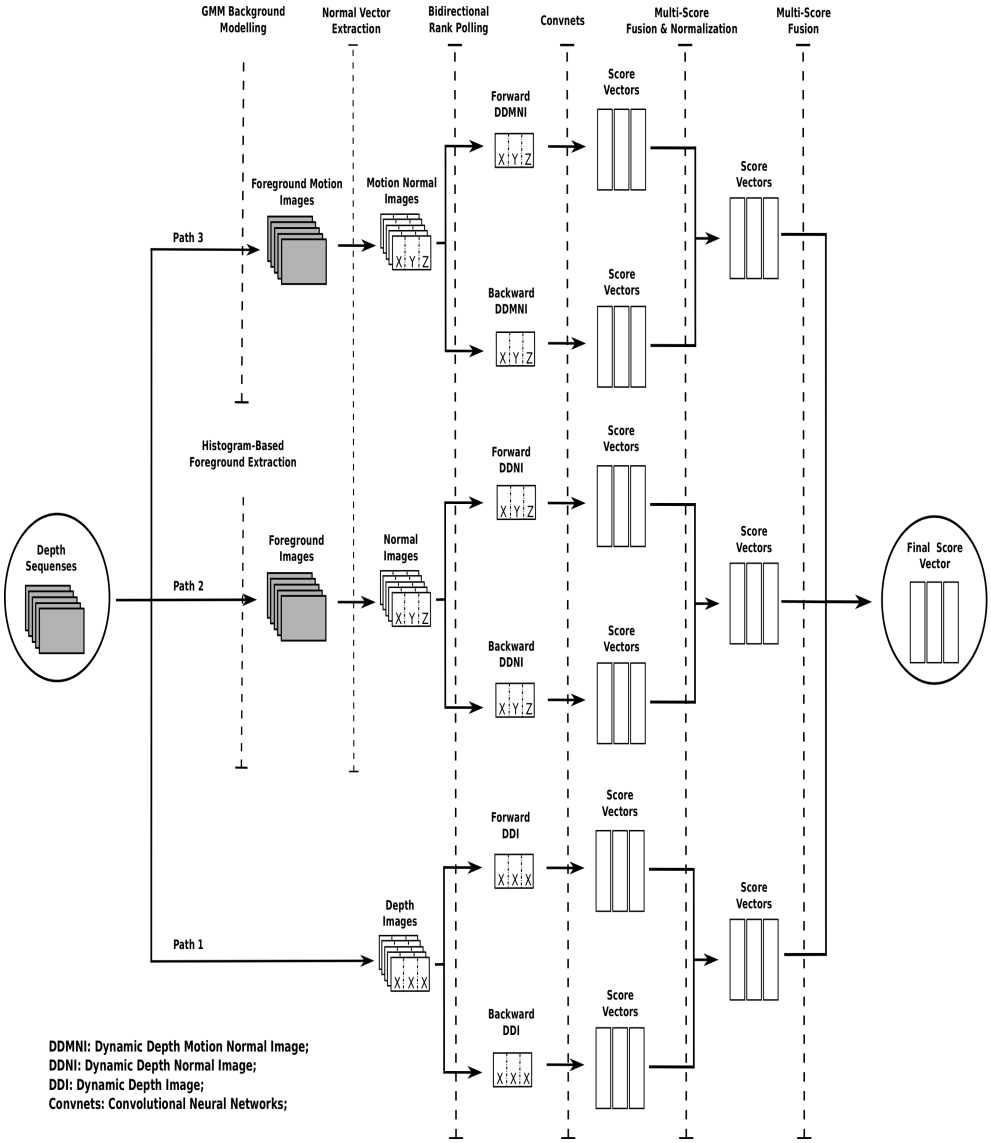

Figure 1: The framework for proposed method.

- 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): ConvNets learned features

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): ConvNets

3.1.5 Other techniques

Other technique/strategy used not included in previous items FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any):None

3.1.6 Method complexity

Method complexity FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE: Somewhat complicated

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): Score fusion

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any): VGG-16 Models
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any): We only use depth data.
- Qualitative advantages of the proposed solution: Good results

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

Method	Set	Recognition rate r
MFSK	Validation	18.65%
MFSK+DeepID	Validation	18.23%
Proposed Method	Validation	**39.23%**(AlexNet)
MFSK	Testing	24.19%
MFSK+DeepID	Testing	23.67%
Proposed Method	Testing	**55.57%** (VGG-16)

- Results of the comparison to other approaches (if any):
- Novelty degree of the solution and if is has been previously published: Novel

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements): Matlab + Caffe + Python. No less than 8G GPU memory under Ubuntu14.04.
- Human effort required for implementation, training and validation?: Easy
- Training/testing expended time?: Less than 10 hours.
- 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 challenging.