Deep Recognition on First-person Videos

Yu-Cheng Lin and Wei-Chih Chen

As the first-person cameras like GoPro become popular for some sports hobbyists and geeks, there would be more first-person videos in the near future. However, the research of video action recognition on first-person videos is still in the beginning. Given that there are so many videos on the Internet, which are mostly third-person point-of-view, we are going to leverage the abundance of these third-person videos to train and recognize first-person videos.

Convolutional neural networks (CNNs) have got great improvement on visual recognition since AlexNet [1]. In addition to applying CNNs for image recognition, detection, segmentation and retrieval, some papers use CNNs to combine image information across videos [2][3].

On the other hand, there are still few papers talking about the relationships and recognition about the first-person and the third-person videos. The only paper that is kind of similar to our work is focusing on the recognition of each video for its action class and view point. We would like to use video data that is time-synchronized and leverage this relationship between the first-person and the third-person videos t0 train a model that can recognize first-person videos much better.

The video data we would use is provided by the Indiana university. This dataset comprises seven action classes. Each class has several videos from the first-person point-of-view and the third-person point-of-view from two different angles. These videos are all time-synchronized.

We would start from the popular deep CNN model, Caffe's Imagenet pre-train model [5], which adopts the same architecture as AlexNet. Using this model as the baseline, we would like to try the siamese network [6] and temporal pooling to improve the performance. In addition, we would use t-SNE library [7] to visualize the feature representation for further discovering the relationships between the different view videos.

The rough job partition is as below:
architecture/performance result discussion: both
architecture/source coding and simulation: Yu-Cheng
tools for feature fetching/result collection and analysis: Wei-Chih

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In NIPS.
[2] Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. arXiv preprint arXiv:1503.08909.
[3] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).
[4] Narayan, S., Kankanhalli, M. S., & Ramakrishnan, K. R. (2014, June). Action and Interaction Recognition in First-person videos. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on (pp. 526-532). IEEE.
[5] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia (pp. 675-678). ACM.
[6] Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In Computer vision and pattern recognition, 2006 IEEE computer society conference on (Vol. 2, pp. 1735-1742). IEEE.
[7] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.