supplementary Materials: Human Action Recognition Using A Multi-Modal Hybrid Deep Learning

BMVC 2017 Submission # 384

1 Experimental Results and Discussions

We performed two types of experiments as in [6]: cross-view (CV) and cross-subject (CS). Table 1 summarizes the obtained results for our experiment compared to the state of the art.

	Method	CS	CV
1	HOG^2	32.24%	22.27%
2	Super Normal Vector	31.82%	13.61%
3	HON4D	30.56%	7.26%
4	Lie Group	50.08%	52.76%
5	Skeletal Quads	38.62%	41.36%
6	FTP Dynamic Skeletons	60.23%	65.22%
7	HBRNN-L	59.07%	63.97%
8	1 Layer RNN	56.02%	60.24%
9	2 Layer RNN	56.29%	64.09%
10	1 Layer LSTM	59.14%	66.81%
11	2 Layer LSTM	60.69%	67.29%
12	1 Layer P-LSTM	62.05%	69.40%
13	2 Layer P-LSTM	62.93%	70.27%
14	ST-LSTM (Joint Chain)	61.7 %	75.5%
15	ST-LSTM (Tree Traversal)	65.2%	76.1%
16	ST-LSTM (TT +TG)	69.2%	77.7%
17	**CL1D**	**62.99%**	**70.11%**
18	**MCL (upper body)**	**70.03%**	**78.01%**
19	**MCL (body parts)**	**73.76%**	**78.4%**

Table 1: Cross subjects and Cross views accuracies in NTU RGB+D dataset

Table 1 shows the results of nineteen methods/models that were tested in both the CS and CV scenarios. Some of them used hand-crafted features, and others used deep learning methods to automatically extract features. The results of the first 13 methods were obtained from [6]. The methods from 1 to 6 in the table used hand-crafted features based on the depth and/or 3D skeleton data. HOG^2 [4], Super Normal Vector [8], and HON4D [5] achieved their highest score (32.24%, 32.82%, and 30.56 % respectively) in the CS scenario because these representations are not view-point invariant. On the other hand, Lie Groups [7], Skeleton Quads [1], and FTP Dynamic Skeletons [2] achieved better scores (52.76 %, 41.56%, and 65.22 %) in the CV scenario because these representations are view-point invariant and hence perform better in the CV scenario because in this scenario the same subject may appear in training and testing, which make the problem easier. Deep learning techniques are used from method 7 to the end of the table. The best scores that were achieved in both CV and CS scenarios, which are 62.99 % and 70.27 %, respectively, are obtained by using 2-Layer P-LSTM [6]. Methods 14, 15 and 16 are the work of [3], which is considered the current state of the art with scores outperforming prior methods.

The last three rows in the table contain our results. When CL1D is used to classify actions based on the body motion alone, the recognition rate for both CS and CV experiments are (62.99% and 70.11%), respectively. However, when MCL is used to capture both the body motion and part shape, the results go up to **73.76 %** and **78.4%**, for the CS and CV scenarios, respectively, which are superior to the current state of the art.

The table contains an extra row for the results of MCL trained on the upper body part as whole after being cropped and reduced to the size of 128×128. This model was trained for the same number of epochs as the MCL (body parts) model. However, 20 frames are sampled per sequence in this model because training with less samples caused under-fitting. Despite the extra information provided to this model, it took much longer training time to exceed the state of the art results, and yet fell clearly behind the MCL model with body parts, especially in the CS scenario. This verifies our initial hypothesis that combining and integrating modalities, as well as leveraging the powers of CNN and LSTM are effective mechanisms in action recognition.

References

[1] Georgios Evangelidis, Gurkirt Singh, and Radu Horaud. Skeletal quads: Human action recognition using joint quadruples. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4513–4518. IEEE, 2014.

[2] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[3] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[4] Eshed Ohn-Bar and Mohan Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 465–470, 2013.

[5] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2013.

[6] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *arXiv preprint arXiv:1604.02808*, 2016.

[7] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[8] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811, 2014.