Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks

Yanghao Li[1], Cuiling Lan[2⋆], Junliang Xing[3], Wenjun Zeng[2],
Chunfeng Yuan[3], and Jiaying Liu[1⋆]

[1] Institute of Computer Science and Technology, Peking University
[2] Microsoft Research Asia [3] Institute of Automation, Chinese Academy of Sciences
{lyttonhao, liujiaying}@pku.edu.cn,
{culan, wezeng}@microsoft.com, {jlxing, cfyuan}@nlpr.ia.ac.cn

1 Experimental Results on the OAD Dataset

1.1 Evaluation of the Soft Selector Module

In this section, we evaluate the influence of the Soft Selector Module, which described in Section 4.3. We implement a variant version of our method by removing the Soft Selector module and directly linking the FC2 and FC3 layers. Table 1 and Table 2 compare the $F1-$Score, $SL-$ and $EL-$Scores of our method with and without soft selctor on the OAD Dataset. Note that we use the same parameters for the two settings. We can see that the incorporation of the Soft Selector can bring significant improvement.

Table 1. $F1-$Score on OAD dataset.

Actions	w/o Soft Selector	with Soft Selector
drinking	0.445	**0.574**
eating	**0.542**	0.523
writing	0.714	**0.822**
opening cupboard	**0.526**	0.495
washing hands	0.688	**0.718**
opening microwave	0.697	**0.703**
sweeping	0.547	**0.643**
gargling	0.478	**0.623**
throwing trash	**0.608**	0.459
wiping	0.691	**0.780**
overall	0.616	**0.653**

⋆ Corresponding author. This work was done at Microsoft Research Asia.

Table 2. $SL-$ and $EL-$Scores on the OAD dataset.

Scores	w/o Soft Selector	with Soft Selector
$SL-$	0.389	**0.418**
$EL-$	0.413	**0.443**

1.2 Sliding Window Size of the Baselines

Since the baseline methods SVM-SW and RNN-SW [1] are based on the sliding window schemes, we report the $F1-$Score of these methods using different window size (number of frames) in Table 3 and 4. The correponding $SL-$ and $EL-$ scores are shown in Table 5 and 6. In the experiments, the stride of the window is set to be half of the window size.

Table 3. $F1-$Score of using different window sizes (ws) for SVM-SW method on OAD dataset.

Actions	SVM-SW				JCR-RNN
	ws=5	ws=10	ws=20	ws=40	
drinking	0.291	0.146	0.000	0.000	**0.574**
eating	0.507	0.465	**0.548**	0.050	0.523
writing	0.671	0.645	0.792	0.542	**0.822**
opening cupboard	0.284	0.308	0.352	0.033	**0.495**
washing hands	0.501	0.562	0.650	0.308	**0.718**
opening microwave	0.521	0.607	0.590	0.492	**0.703**
sweeping	0.434	0.461	0.515	0.498	**0.643**
gargling	0.466	0.437	0.433	0.083	**0.623**
throwing trash	0.475	0.554	0.315	0.000	**0.459**
wiping	**0.867**	0.857	0.748	0.433	0.780
overall	0.525	0.540	0.565	0.334	**0.653**

2 Experimental Results on the G3D Dataset

In this section, we show experimental results on all the seven categories of videos in the G3D dataset [2]. Note that we have only shown the results on the first two categories of videos in the paper due to the space limitation. We evaluate the performance in terms of the $F1-$Score and $SL-$ and $EL-$Score as described in this paper, and summarize them in Table 7 and Table 8. Table 9 shows the comparsion of these methods using the evaluation metric of action-based $F1$ as defined in [2].

Table 4. $F1-$Score of using different window sizes (ws) for RNN-SW [1] method on OAD dataset.

Actions	RNN-SW [1]				JCR-RNN
	ws=5	ws=10	ws=20	ws=40	
drinking	0.495	0.441	0.142	0.033	**0.574**
eating	**0.525**	0.550	0.532	0.183	0.523
writing	0.665	**0.859**	0.837	0.725	0.822
opening cupboard	0.331	0.321	0.413	0.217	**0.495**
washing hands	0.588	0.668	**0.867**	0.517	0.718
opening microwave	0.628	0.665	0.680	0.647	**0.703**
sweeping	0.475	0.590	**0.811**	0.783	0.643
gargling	0.426	0.550	0.534	0.108	**0.623**
throwing trash	0.434	**0.674**	0.325	0.050	0.459
wiping	0.734	0.747	**0.797**	0.550	0.780
overall	0.512	0.600	0.627	0.476	**0.653**

Table 5. $SL-$ and $EL-$Scores of using different window sizes (ws) for SVM-SW method on OAD dataset.

Scores	SVM-SW				JCR-RNN
	ws=5	ws=10	ws=20	ws=40	
$SL-$	0.288	0.316	0.339	0.182	**0.418**
$EL-$	0.300	0.325	0.343	0.184	**0.443**

Table 6. $SL-$ and $EL-$Scores of using different window sizes (ws) for RNN-SW [1] method on OAD dataset.

Scores	RNN-SW [1]				JCR-RNN
	ws=5	ws=10	ws=20	ws=40	
$SL-$	0.287	0.366	0.393	0.276	**0.418**
$EL-$	0.291	0.376	0.401	0.274	**0.443**

Table 7. $F1-$Score on the G3D Dataset.

Action Category	SVM-SW	RNN-SW [1]	CA-RNN	JCR-RNN
Fighting	0.486	0.613	0.700	**0.735**
Golf	0.680	0.745	0.900	**0.967**
Tennis	0.598	0.480	0.774	**0.788**
Bowling	0.667	0.889	1.000	**1.000**
FPS	0.571	**0.581**	0.378	0.523
Driving	1.000	1.000	1.000	**1.000**
Misc	0.712	0.742	0.813	**0.862**

Table 8. $SL-$ and $EL-$Score on the G3D Dataset.

Action Category	Scores	SVM-SW	RNN-SW [1]	CA-RNN	JCR-RNN
Fighting	$SL-$	0.318	0.412	0.512	**0.528**
	$EL-$	0.328	0.419	0.525	**0.557**
Golf	$SL-$	0.553	0.635	0.789	**0.793**
	$EL-$	0.524	0.656	0.791	**0.836**
Tennis	$SL-$	0.444	0.338	0.605	**0.665**
	$EL-$	0.460	0.333	0.617	**0.667**
Bowling	$SL-$	0.612	0.777	0.933	**0.959**
	$EL-$	0.550	0.713	0.816	**0.861**
FPS	$SL-$	0.351	**0.388**	0.183	0.311
	$EL-$	0.353	**0.393**	0.199	0.327
Driving	$SL-$	**0.991**	0.983	0.957	0.955
	$EL-$	0.975	0.975	0.964	**0.975**
Misc	$SL-$	0.487	0.593	0.609	**0.614**
	$EL-$	0.515	0.612	0.690	**0.766**

Table 9. Action-based $F1$ [2] on the G3D Dataset.

Action Category	G3D [2]	SVM-SW	RNN-SW [1]	CA-RNN	JCR-RNN
Fighting	58.54	76.72	83.28	94.00	**96.18**
Golf	11.88	45.00	55.00	50.00	**70.00**
Tennis	14.85	37.57	36.68	59.04	**62.38**
Bowling	31.58	22.22	44.44	44.44	**66.67**
FPS	13.65	35.35	**39.89**	23.69	33.85
Driving	2.5	39.99	50.00	19.99	**50.00**
Misc	18.13	53.32	65.24	73.81	**86.19**

3 Demo

The attached video shows some results of the proposed method for online action detection on our Online Action Dataset (OAD).

References

1. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: AAAI Conference on Artificial Intelligence. (2016)
2. Bloom, V., Makris, D., Argyriou, V.: G3d: A gaming action dataset and real time action recognition evaluation framework. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops. (2012) 7–12