Supplementary Material for "Early Action Prediction by Soft Regression"

Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang

Abstract—In this supplementary document, we provide more experimental investigations and analysis on the soft regression-based early action prediction system developed in our main submission, which is excluded from the main submission due to space limitation.

In our main submission, we developed our early action prediction system based on the proposed soft regression framework. In the soft regression framework, we learn both the soft labels and predictor jointly from linear to deep models (SLR, SRNN and M-SRNN). We have demonstrated the effectiveness of our approach on three RGB-D benchmark datasets and a unconstrained RGB action set and show that the proposed soft regression-based early action prediction model outperforms existing models significantly. In the following, we provide more experimental investigations and analysis.

1 More Discussion

Complement to the HOF features. In the main manuscript, our LAFF feature is mainly built on the HOG descriptors extracted from both RGB and depth frames. Here, we would like to illustrate that the LAFF can also be constructed based on the HOF descriptors, which can capture more motion cues. The detailed evaluation results are presented in Table R.1. As shown, combining the HOG and HOF features (denoted by ALL) can obtain an AUC of 75.1%, which is comparable with the performance of using two-stream CNN features (75.4%) and outperforms the HOG-only features (denoted by HOG) by 3.5%. However, it is worth noting that extracting HOF features from both RGB and depth videos is quite time consuming, and the overall prediction speed is about 2 fps, which is much slower than HOG-only systems (34 fps). We also observe that only using the HOF features extracted from RGB (depth) videos can achieve a performance of 54.8% (54.2%), which is much lower than that HOG features (61.3% and 60.8%). This demonstrates that the flow information is less discriminative than appearance in early action prediction, although combining them together outperforms each individually.

TABLE R.1
More evaluations on the LAFF features. (HOG: HOG only features; HOF_RGB (DEP): flow features from RGB (depth) videos; ALL: HOG+ HOF_RGB+ HOF_DEP).

Observation ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	AUC
HOG	47.5	56.7	66.7	75.4	78.3	80.4	81.7	82.5	81.7	79.6	71.6
HOF_RGB	28.8	36.7	52.1	53.8	59.6	62.1	65	65.4	64.2	60.8	54.8
HOF_DEP	25.8	37.9	50.8	56.3	57.9	60.4	62.5	64.2	63.8	62.1	54.2
HOG+HOF_RGB	**50.8**	59.6	67.9	75.4	77.1	79.6	80.8	83.8	82.1	79.6	73.7
HOG+HOF_DEP	45	57.5	66.3	75.4	77.9	**83.3**	84.6	**85**	**85**	**82.1**	74.2
ALL	50	**60.8**	**68.3**	**77.5**	**79.2**	82.1	**85**	**85**	82.9	80.4	**75.1**

What actions can/cannot be early predicted in the three datasets, and why? For instance, we examine the detailed prediction results (top 10 and bottom 10) in the NTU set in Figure R.1. We can find that actions with interactions, such as human-object interactions and human-human interactions, can be better predicted at early stages, especially when the manipulated object is salient in the observed sequences. This is as expected, as the manipulated object often appears when the action is in progress and it plays an important role for the definition of human actions. Actions interacting with similar objects (for examples, wearing on grasses and taking off grasses) or with identical gesture/pose (e.g., eating meal, sneezing/coughing, reading) are a bit more challenging for early prediction, as their discrimination requires fine-grained motion information, which often becomes apparent at later stages.

The performance difference for the early action prediction on the ORGBD and NTU sets. Compared with the ORGBD set, the NTU set is much more challenging for the early prediction of actions. The ORGBD set is a small set with only 224 samples from 7 action classes: (drinking, eating, using laptop, reading cell-phone, making phone call, reading book and using remote). Most of these actions can be identified from the different type of manipulated objects, which can often be observed at early stages, as shown in Figure R.2 (left). In contrast, the NTU set is more challenging. It contains a total of 56,880 action samples from 60 action categories, including body actions, gestures, human-object interactions, and even human-human interactions. Moreover, some actions in this set are easily confused with each other and contain

- *J.-F. Hu, W.-S. Zheng, and J. Lai are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. J.-F. Hu is also with the Guangdong Province Key Laboratory of Computational Science, Guangzhou, China. W.-S. Zheng is also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China, E-mail:hujianf5@mail.sysu.edu.cn, wszheng@ieee.org and stsljh@mail.sysu.edu.cn*
- *L.-Y. Ma is with Tencent AI Lab, Beijing, China. E-mail:leonlyma@tencent.com*
- *G. Wang is with Alibaba AI Labs. E-mail:gangwang6@gmail.com*
- *J. Zhang is with Computing, School of Science and Engineering, University of Dundee United Kingdom. E-mail:j.n.zhang@dundee.ac.uk*

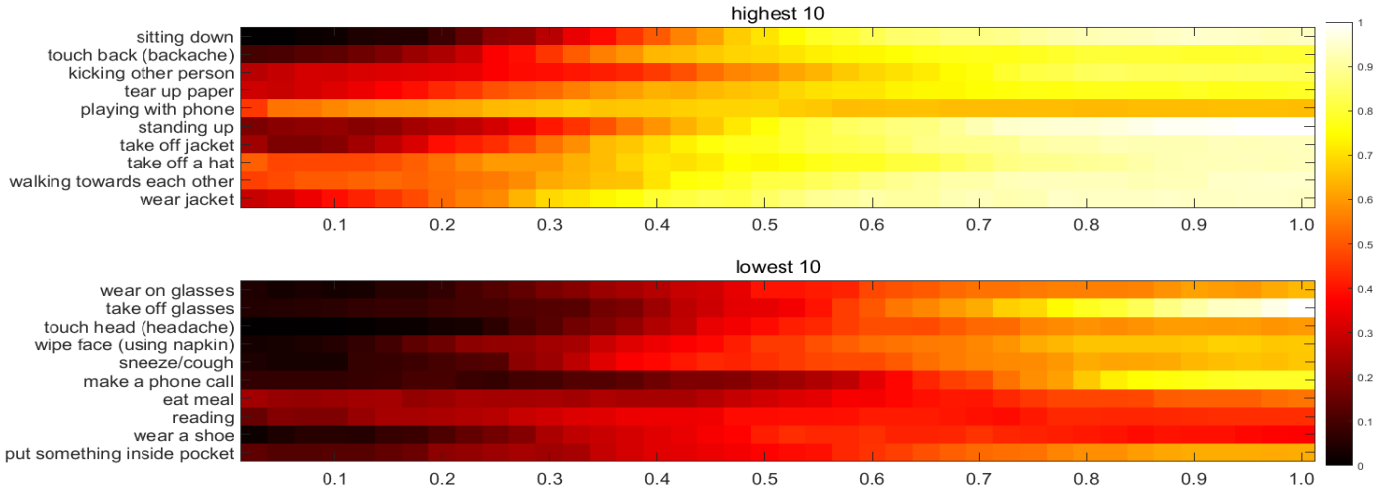


Fig. R.1. Detailed prediction results for the actions have the highest 10 (top) and the lowest 10 (bottom) AUCs. (best viewed in color)

TABLE R.2
Prediction results in a subset from NTU set.

Observation Ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	AUC
Accuracy (%)	43.3	51.3	62.5	74.8	83.0	86.8	89.0	90.1	91.3	91.1	74.4

similar appearance information (e.g., wearing on glasses vs. taking off glasses). From the examples in Figure R.2 (right), we can find that the object cues are not salient at early stages in some action samples. These elements make the appearance less useful for distinguishing different actions in the NTU set.

To experimentally verify that the actions in the NTU set are easily confused, we further cluster all the 60 action classes into a few groups (clusters) by affinity propagation clustering algorithm [1], where the similarity matrix is defined by the mean confusion matrices of predicting early actions of various progress levels. Therefore, each group contains the action classes that are easily confused. In our case, we have obtained 11 groups and some are presented in Figure R.3. As shown, the actions within each group are easily confused and they often have similar appearance context or motion cues. From the clustering results, it is reasonable to hypothesize that actions from different groups can be well separated. To verify this, we then formed an 11-class subset by selecting one action class from each group and conducted early action prediction on this new subset. We find that our model can obtain an AUC of 74.4% (see Table R.2 for details). Moreover, for the prediction of actions from the first 10% frames, our system can achieve an accuracy of 43.3%, which is quite consistent with the results obtained on the SYSU 3D HOI and ORGBD datasets.

Comparison of SLRs and SRNNs on the ORGBD and NTU Large Scale sets. Table R.3 and Table R.4 summarize the results on the datasets of NTU and ORGBD, including the detailed comparisons w.r.t the observation ratios for each dataset, respectively. Similar to that on SYSU 3D HOI set, it could be observed that when the observation ratio is smaller than 50%, MSRNN model, the further development of SRNN, consistently outperformed SLR on both the NTU and ORGBD datasets. The performances are improved by 1.5-2.8% in terms of AUC. The improvement of AUCs

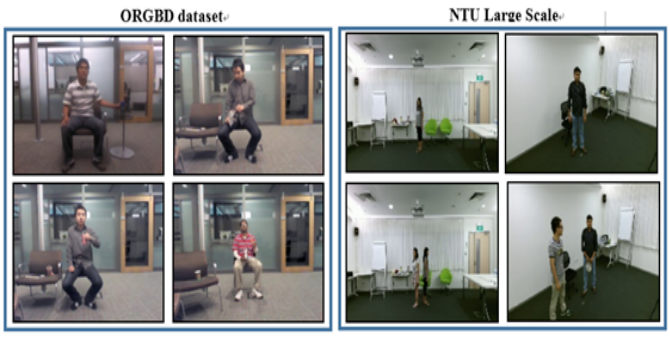


Fig. R.2. Snapshots from the first 10% frames of some actions in ORGBD (left) and NTU Large Scale action dataset (right). It could be seen that actions in ORGBD are often in the form of human-object interactions (where the manipulated objects are of different types for different actions), whilst NTU contains some actions performed without interacting with objects.

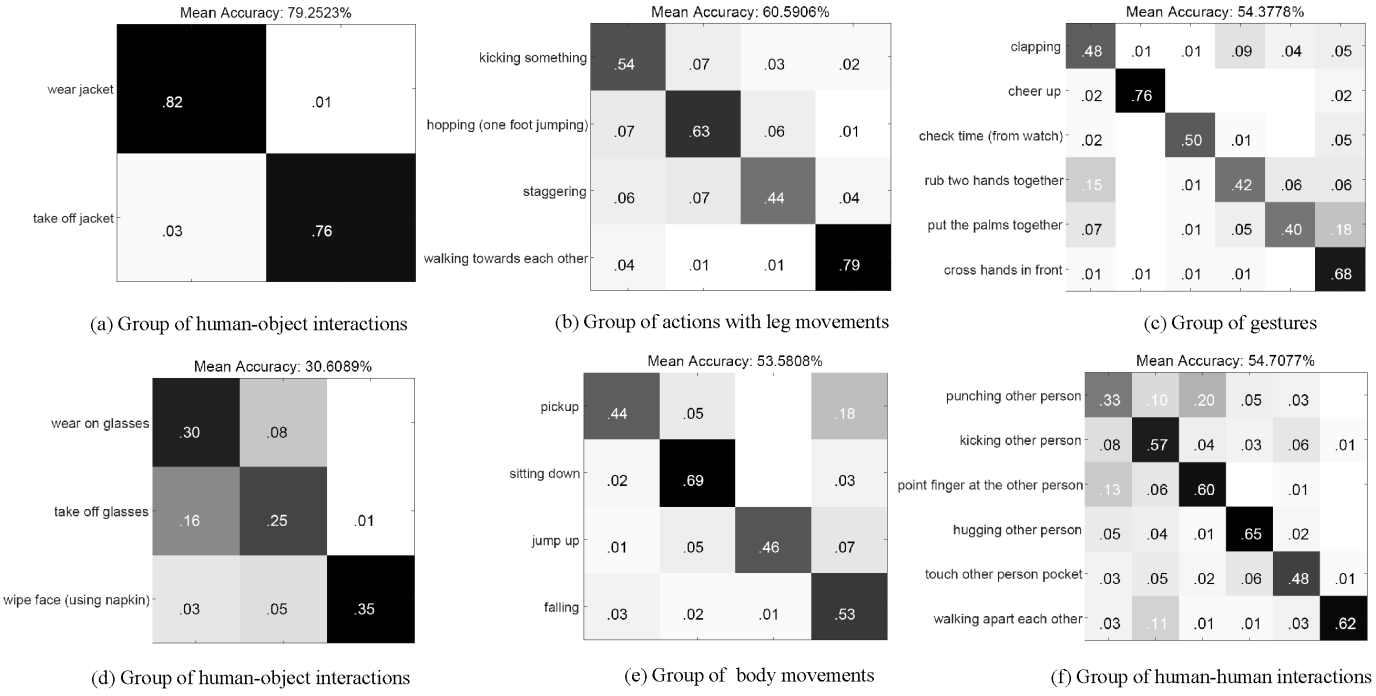
TABLE R.3
Evaluations of different soft regression models on NTU Large Scale Set (%).

Observation ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	AUC
SLR	14.4	18.5	26.5	37.6	48.2	56.1	61.7	64.6	65.9	64.8	43.8
MSLR	13.5	18.5	27.7	39.7	50	57.9	63	66.4	67.9	68.1	45.1
SRNN	13.4	18.2	26.7	38.7	49.4	57.6	63	66.2	67.4	67.6	44.7
MSRNN	**15.2**	**20.3**	**29.5**	**41.4**	**51.6**	**59.2**	**63.9**	**67.4**	**68.9**	**69.2**	**46.6**

on NTU is larger than that on ORGBD (2.8% vs. 1.5%). Compared to ORGBD, SYSU 3DHOI and especially NTU datasets are very challenging with larger scales and class diversity, and thus larger improvement on these two datasts indicates that MSRNN with multi-soft label learning, which means that learning a category-specific soft label vector for each action type, is beneficial for the early action prediction.

REFERENCES

[1] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.



(a) Group of human-object interactions (b) Group of actions with leg movements (c) Group of gestures

(d) Group of human-object interactions (e) Group of body movements (f) Group of human-human interactions

Fig. R.3. We selected six action groups/clusters from the 11 groups learned on NTU set by affinity propagation clustering algorithm. Each group is shown as a sub-block of the confusion matrix (i.e., similarity matrix) with the corresponding action classes. For each sub-block, mean accuracy is shown for that block indicating the difficulties of discriminating actions within that block. The left-hand side of confusion matrix shows the label of the actions belonging to the group. As seen, some actions are grouped based on interacting with similar object (e.g., wear on or take off glasses as shown in (a)); some are grouped based on hand gesture action (e.g., clapping, cheer-up, rub two hands together etc. as shown in (c)) and some involves body actions (e.g., jump up, sitting down etc. as shown in (e)). And different action groups could contain different number of actions.

TABLE R.4
Evaluations of different soft regression models on ORGBD Set (%).

Observation ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	AUC
SLR	57.1	63.4	64.3	66.5	70.1	70.1	70.5	72.8	**73.7**	**74.1**	67.2
MSLR	56.3	61.6	64.7	67.9	69.6	71.4	72.8	**73.2**	73.2	73.7	67.6
SRNN	57.9	62.5	**67**	68.8	71.9	**73.2**	**73.2**	72.3	73.2	72.8	68.4
MSRNN	**60.7**	**63.8**	**67**	**69.6**	**72.3**	71.4	71.9	72.8	73.2	73.2	**68.7**