Geometric Features Informed Multi-person Human-object Interaction Recognition in Videos Supplementary Material

Tanqiu Qiao[1], Qianhui Men[2], Frederick W. B. Li[1], Yoshiki Kubotani[3], Shigeo Morishima[3], and Hubert P. H. Shum[1†]

[1] Durham University, United Kingdom
{tanqiu.qiao, frederick.li, hubert.shum}@durham.ac.uk
[2] University of Oxford, United Kingdom qianhui.men@eng.ox.ac.uk
[3] Waseda Research Institute for Science and Engineering, Japan
yoshikikubotani@akane.waseda.jp, shigeo@waseda.jp

1 Cross-validation Results

Table 1, Table 2 and Table 3 present joined segmentation and label recognition task results for each subject cross-validation group on CAD120, MPHOI-72 and Bimanual Actions Datasets, respectively. We compare 2G-GCN with ASSIGN to show our improvement for each subject.

Table 1. Joined segmentation and label recognition task results for each subject cross-validation group on CAD120 dataset.

Model	Sub-activity				Object Affordance			
	Subject1	Subject3	Subject4	Subject5	Subject1	Subject3	Subject4	Subject5
ASSIGN	85.2	90.2	88.3	88.2	90.8	93.7	91.4	**92.0**
2G-GCN	**88.1**	**92.1**	**89.5**	**88.4**	**91.0**	**95.0**	**92.7**	90.8

Table 2. Joined segmentation and label recognition task results for each subject cross-validation group on our proposed MPHOI-72 dataset.

Model	Sub-activity; F_1@10			Sub-activity; F_1@25		
	Subject14	Subject25	Subject45	Subject14	Subject25	Subject45
ASSIGN	48.8	52.5	76.0	33.7	45.6	73.6
2G-GCN	**64.9**	**58.0**	**82.8**	**52.3**	**54.6**	**75.3**

† Corresponding author

Table 3. Joined segmentation and label recognition task results for each subject cross-validation group on the Bimanual Actions dataset.

Model	Sub-activity; F_1@10					
	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6
ASSIGN	**82.5**	84.2	80.7	84.3	85.2	87.1
2G-GCN	81.6	**85.5**	**83.7**	**85.3**	**85.3**	**88.8**

2 Ablation Study on MPHOI-72

Table 4 shows the ablation study result on MPHOI-72, where rows (1) - (4) represent the model drops human skeleton features, object features, embedding function and similarity matrix in the geometric-level graph, respectively; rows (5) - (7) represent the model disables the attention connection between the pair of human-human, human-object and object-object in the fusion-level graph, respectively; row (8) represents the model has an extra attention connection between human and geometry features in the fusion-level graph, while (9) 2G-GCN does not.

Table 4. Ablation study on MPHOI-72. GG and FG denote the geometric-level graph and the fusion-level graph, respectively.

Model	Sub-activity	
	F_1@10	F_1@25
(1) GG (w/o skeletons) & FG	66.8	60.2
(2) GG (w/o objects) & FG	66.7	59.8
(3) GG (w/o embedding) & FG	62.2	56.5
(4) GG (w/o similarity) & FG	66.1	58.9
(5) GG & FG (w/o human-human)	67.2	59.6
(6) GG & FG (w/o human-object)	58.6	51.7
(7) GG & FG (w/o object-object)	65.7	60.2
(8) GG & FG (w human-geometry)	65.6	60.7
(9) 2G-GCN	**68.6**	**60.8**

3 Visualisations of Confusion Matrix

Fig. 1 is the visualisation of confusion matrices of our 2G-GCN evaluated on the MPHOI-72 and Bimanual Actions datasets in this section. The diagonal elements denote the probability of the number of sub-activities whose recognition labels are equal to the ground-truth, while the off-diagonal elements are those sub-activities that are misidentified. The higher the diagonal value of the confusion matrix, the better, representing numerous correct recognitions.

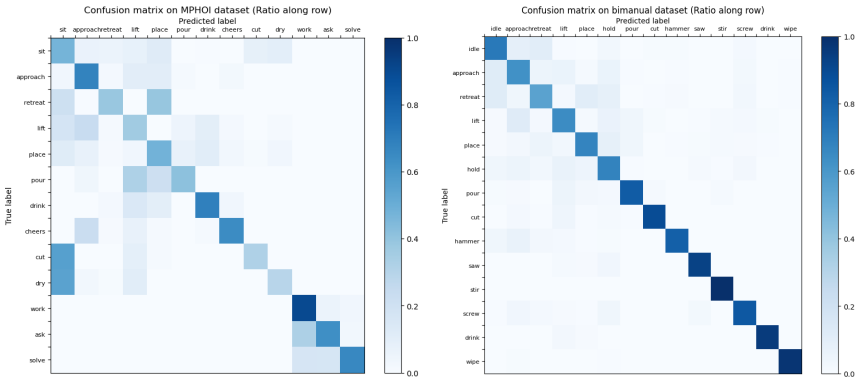

Fig. 1. The confusion matrix of 2G-GCN evaluated on the MPHOI-72 and Bimanual Actions dataset by class support size.