Supplementary material for: Deep Dual Relation Modeling for Egocentric Interaction Recognition

Haoxin Li[1,3,4], Yijun Cai[1,4], Wei-Shi Zheng[2,3,4,*]
[1]School of Electronics and Information Technology, Sun Yat-sen University, China
[2]School of Data and Computer Science, Sun Yat-sen University, China
[3]Peng Cheng Laboratory, Shenzhen 518005, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
lihaoxin05@gmail.com, caiyj6@mail2.sysu.edu.cn, wszheng@ieee.org

1. FLOPs

To evaluate the complexity of our model, we calculate the inference FLOPs (floating point operations) per video of variations of our model, the results are shown in Table 1. It is observed that our relation modeling components would not increase FLOPs too much but still boost performance.

Table 1. Recognition accuracy (%) and FLOPs of our model.

Variations of model	PEV	inference FLOPs
Concat(no relation)	60.8	8.4971×10^{10}
Interaction with sym. blocks	62.7	8.4974×10^{10}
Interaction with rel. branch	63.0	8.4982×10^{10}
Interaction with both	64.2	8.4984×10^{10}

2. Visualization results

To visualize the effect of our attention module and motion module, we illustrate some learned human masks and motions here.

Figure 1 to Figure 8 illustrate the learned human masks and motions of some samples. In each figure, the upper left and upper right are frame $\boldsymbol{I}_{n-1}$ and frame $\boldsymbol{I}_n$, respectively. The middle left is the learned human mask; the middle right is the reconstructed frame $\hat{\boldsymbol{I}}_{n-1}$ from frame $\boldsymbol{I}_n$ in which the global motion is reflected. The bottom left and the bottom right are the vertical and horizontal local motion field, respectively, where the amplitudes to the bottom and the right are proportional to the brightness of the motion fields. In addition, those motion vectors outside the learned human masks are discarded. For example, in Figure 1, the reconstructed frame shows a slight head motion to the right, the vertical and horizontal local motion field together show the interactor's hands moving towards the upper right.

It is observed that our attention module could learn to localize the interactor, and the motion module could capture the global and local motions in most cases, except those with violent shakes as shown in Figure 8.

Figure 1.

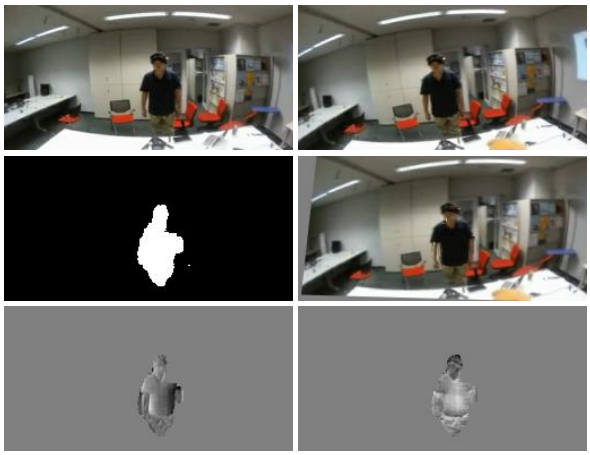

Figure 2.

*Corresponding author

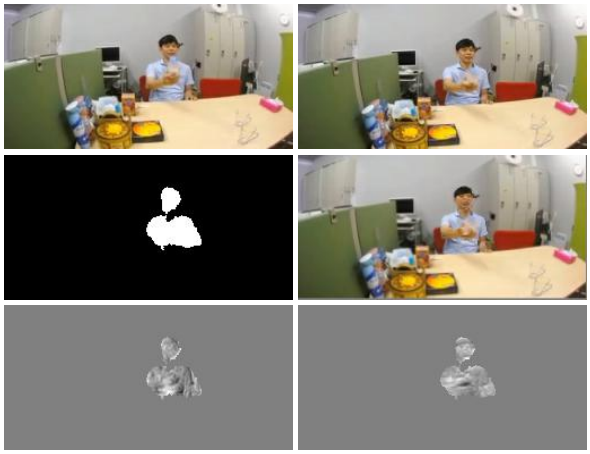

Figure 3.

Figure 4.

Figure 5.

Figure 6.

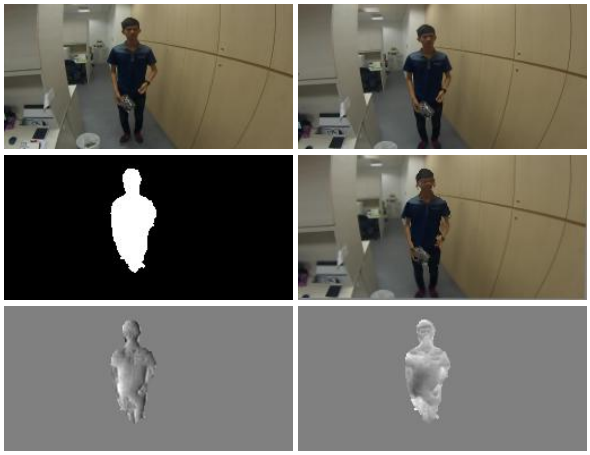

Figure 7.

Figure 8.