MixedFusion: 6D Object Pose Estimation from Decoupled RGB-Depth Features

Hangtao Feng, Lu Zhang, Xu Yang, Zhiyong Liu
fenghangtao2018, zhanglu2016, xu.yang, zhiyong.liu}@ia.ac.cn

Problem
Estimating the 6D pose of objects is an important process for intelligent systems to achieve interaction with the real-world. As the RGB-D sensors become more accessible, the fusion-based methods have prevailed, since the point clouds provide complementary geometric information with RGB values. However, due to the difference in feature space between color image and depth image, the network structures that directly perform point-to-point matching fusion do not effectively fuse the features of the two. We argue that the spatial correspondence of color and point clouds could be decoupled and reconnected, thus enabling a more flexible fusion scheme.

Contributions

Contribution 1
An decouple-module is added to expand the point cloud receptive field, break the point-to-point correspondence, and realize the decoupling of point cloud features and RGB features (MAX represents max-pooling layer and FC represents fully connected layers).

Contribution 2
We design a mixed fusion module, and it concatenates the point embeddings and the color embeddings. We also concatenate the mixed fusion feature embeddings and the global feature embeddings as the input feature map of the posenet [2] to fully mine the performance of the network.

Experiments
Visualized results on the YCB-Video Dataset. All the methods are tested with the same segmentation masks, and we use twice optimization iterations for DenseFusion and ours. We get the poses of the point cloud based on the outputs by these networks, project them to an RGB image and visualize the differences have been highlighted with circles. From this figure, we can see that in most cases, the visualization of our model is better than DenseFusion and PoseCNN+ICP.

References