6D Pose Estimation with Correlation Fusion

Yi Cheng, Hongyuan Zhu, Ying Sun, Cihan Acar, Wei Jing, Yan Wu, Liyuan Li, Cheston Tan, Joo-Hwee Lim
Institute for Infocomm Research, Agency for Science, Technology and Research

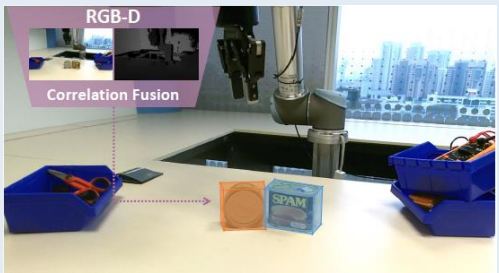
Introduction

6D pose estimation, which aims to predict the 3D rotation and translation from object space to camera space, is useful in 3D object detection and recognition, robot grasping and manipulation.

Limitations of existing methods:

- RGB-only methods ignores complementary information from depth modality, which are vulnerable to heavy occlusion and poor illumination.
- RGB-D based methods fail to adequately exploit the consistent and complementary information between the RGB and depth modalities.

Figure 1. Robot grasping system with our real-time 6D pose estimation algorithm.



Objective

To address the aforementioned limitations, we propose a novel Correlation Fusion (CF) framework which models the feature correlation within and between RGB and depth modalities to improve the final performance of 6D pose estimation.

Our contribution can be summarized as follows:

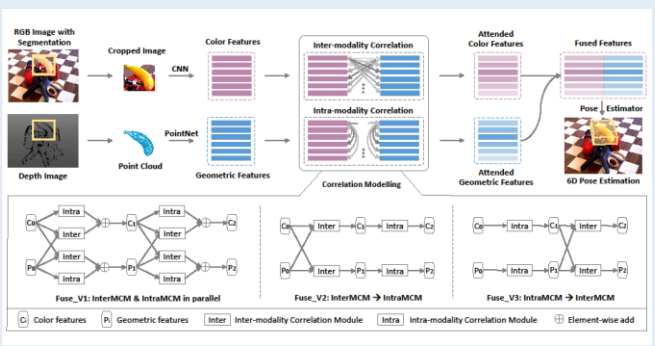
- We propose the intra- and inter-correlation modules to exploit the consistent and complementary information within and between RGB and depth modalities for 6D pose estimation.
- We explore multiple strategies for fusing the intra- and inter-modality information flow to learn discriminative multi-modal features.
- We demonstrate that the proposed method can achieve the state-of-the-art performance on widely-used benchmark datasets for 6D pose estimation, including LineMOD and YCB-Video datasets.
- We showcase that our method can benefit robot grasping tasks by providing an accurate estimation of object pose.

Methodology

The overall framework includes three stages:

1) Semantic Segmentation and Feature Extraction. We first segment the target objects in the image with an existing semantic segmentation architecture, and then generate the color and geometric features with the predicted segmentation maps.
2) Multi-modality Correlation Learning. It consists of Intra-modality Correlation Modelling (IntraMCM), Inter-Modality Correlation Modelling (InterMCM) and Multi-modality Fusion Strategies.
3) Iterative Pose Refinement. A refiner network is employed for iteratively refining the predicted object pose.

Figure 2. Overview of Correlation Fusion (CF) framework for 6D pose estimation..



Results

We compare our method with other state-of-the-art methods on YCB-Video and LineMOD dataset. The results on YCB-Video dataset can be found in the full paper.

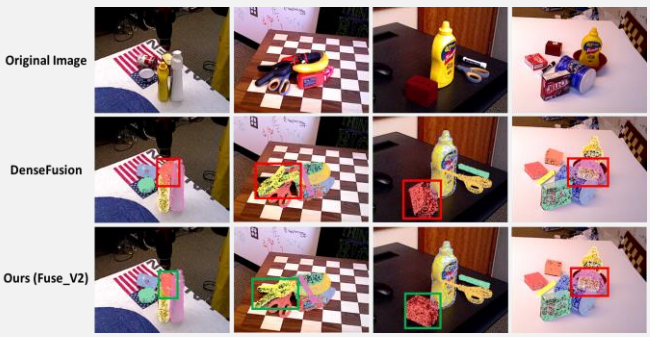
Table 1. The 6D pose estimation accuracy on the LINEMOD Dataset in terms of the ADD(-S) metric. The objects with bold name (glue and eggbox) are considered as symmetric. All the methods use RGB-D images as input.

	SSD6D [16]	BB8 [13]	DenseFusion [21]	OURS (IntraMCM)	OURS (InterMCM)	OURS (Fuse_V1)	OURS (Fuse_V2)	OURS (Fuse_V3)
ape	65	40.4	92.3	94.9	95.2	94.8	**95.6**	95.4
bench	80	91.8	93.2	93.7	94.0	96.1	**96.9**	96.1
camera	78	55.7	94.4	97.5	95.6	96.0	**97.9**	97.5
can	86	64.1	93.1	95.4	95.7	92.2	**96.0**	95.0
cat	70	62.6	96.5	98.4	98.8	**99.2**	97.8	99.1
driller	73	74.4	87.0	92.2	92.7	91.4	**95.6**	94.7
duck	66	44.3	92.3	**96.2**	95.1	95.7	95.7	95.8
eggbox	**100**	57.8	99.8	**100.0**	99.6	**100.0**	99.9	99.9
glue	**100**	41.2	**100.0**	99.8	99.8	99.8	99.7	99.8
hole	49	67.2	92.1	95.2	95.6	95.8	96.7	**97.1**
iron	78	84.7	97.0	95.8	96.2	97.4	97.8	**98.4**
lamp	73	76.5	95.3	95.4	96.3	96.5	**97.0**	96.8
phone	79	54.0	92.8	97.3	**97.5**	95.6	97.0	97.4
MEAN	77	62.7	94.3	96.3	96.3	96.2	**97.2**	97.1

Qualitative Results

We present some qualitative results on the examples from YCB-Video dataset, for both DenseFusion and our proposed method.

Figure 3. Visualizations of results on the YCB-Video Dataset. The first row is original RGB image, the second row is from DenseFusion, and third row is from our proposed method. The red boxes highlight the errors while the green boxes shows the superiority of our method.



Robotic Grasping Experiments

We also carry out robotic grasping experiments in both simulation and real world to demonstrate that our method is effective in assisting robots to correctly grasp objects by providing accurate object pose estimation.

Table 2. Success rate for the grasping experiments with robotic arm in simulation environment of Gazebo.

Success Attempts (%)	tomato_soup_can	mustard_bottle	banana	bleach_cleanser
DenseFusion [21]	80.0	70.0	55.0	65.0
Ours	90.0	85.0	75.0	80.0

Conclusion

In this paper, we have proposed a novel Correlation Fusion framework with intra- and inter-modality correlation learning for 6D object pose estimation. The IntraMCM module is designed to learn prominent modality-specific features and the InterMCM module is to capture complement modality features. Subsequently, multiple fusion schemes are explored to further improve the performance on 6D pose estimation. Extensive experiments conducted on YCB, LINEMOD dataset and a real-world robot grasping task demonstrate the superior performance of our method to several benchmarking methods.

Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A2b0046).