Paper Review of Dense Object Nets
Object Descriptors By and For Robotic Manipulation

Alvin Sun

November 8, 2021

1 Paper Summary

This paper proposed a self-supervised method for learning dense visual descriptors that can be used for identifying good grasp locations. As the camera is mounted with the manipulator with known forward kinematics, a full 3D reconstruction can be constructed with multiple images taken from different views. The vertex correspondence across multiple views can then be labeled via re-projection without any supervision. The descriptor vector for each pixel is trained such that pixels that points to the same vertex projects to feature vectors that are very close in the embedding space. In addition to the proposed learning method, the authors also provided detailed training procedures, such as data augmentation and hard-negative scaling, to battle for higher data efficiency. Both the manipulation experiments and the visualization of the descriptors show that the learned visual representation can robustly identify reference grasp points with deformable objects in a wide range of configurations. It is also shown that the algorithms generalizes well to objects that are unseen during training.

2 What I Learned

1. Re-identification techniques can be used to train for dense pixel descriptors!

2. It seems like computational photographic methods should really take advantage of the known forward kinematics when used in manipulation settings. SLAM is in general quite hard, but with a robotic arm, it becomes much easier.

3 Opinions

3.1 Up Votes

1. I really like their data generation approach that the whole labeling process is self-supervised. This not only is faster, but also really systematically eliminates human which ensures the quality of the generated dataset. It also cleverly uses 3D reconstruction to solves the stereo correspondence problem that is otherwise quite hard in the usual settings.

2. I also strongly agree with their methods in tackling data efficiency, where they used many data augmentation techniques such as domain randomization and randomized lighting / view configurations. Combined with the cross object negative examples, these techniques really makes the training generalizable with reasonably small dataset size.

3.2 Down Votes

I don't quite agree with the author's claim on how strong this method is connected with manipulation tasks. Even though they solve this feature point query problem, a reference grasp point needs to be hand-specified nonetheless. This dense visual descriptor "might" be useful for manipulation tasks but I don't think the experiments carried in this paper show the connection. The problem of finding good grasp locations, which is one of the core problems of manipulation, remains unsolved.

4 Evaluations

This paper aims at obtaining dense visual representation for manipulatable objects that is robust against viewpoints and domains without any supervision. This is certainly a valid objective as most of the traditional feature descriptor algorithms rely heavily on hand-designed feature extractors that are sensitive to specific domains or viewing conditions. Moreover, the re-identification learning methods usually require hand labeling of matching pairs (of pixels or objects depending on the domain), while the proposed method in this paper is purely self-supervised, which makes the learning and inference fully automated.

The overall quality of this paper is very good. The experiments as well as the visualization show that the method is capable of generating dense descriptors that are robust at querying reference feature points across multiple viewing conditions and even backgrounds, while keeping a fully automated learning process that takes reasonably short amount of time. It is really impressive that they delivers this learning algorithms with minimal human intervention. However, one big implicit assumption they made is that a reference point is known or hand picked by a human. This really reduces the attraction of this methods even though it achieves very robust pixel-level querying performance. However, the major manipulation problem of finding good grasp locations remains unsolved.

5 Questions

1. It is quite well known that hard negative mining is really effective in re-identification tasks. Why do they opt for hard sample scaling instead?

2. How can this learned descriptor be used in a more automated way? (i.e. without needing a human to pick a reference grasp point)