MVA - Project Proposal for 6D Object Pose Estimation

Alexis GROSHENRY - alexis.groshenry@polytechnique.edu

Abstract

This report presents my project proposal for the MVA course Object Recognition and Computer Vision which will tackle the subject of 6D Object Pose Estimation. After motivationg the topic, I present the foreseen work plan and the considered approaches to extend the method to RGB-only images.

1. Motivation of the project

6D object pose estimation is an ancient and important topic at the core of many real-world applications which aims to produce a robust estimate of object position and orientation from input image. The DenseFusion method introduced in [4] has recently achieved state of the art performance by relying on deep-learning and on two new components : a pixel-wise approach combining pixel-level and image-level features to infer the pose of the object and an iterative refinement inspired from [2] to improve the performances of the approach.

This approach uses RGB-D images but the acquisition of the depth recquires dedicated devices whose measurements may eventually fail in some environment such as outdoor scenes. Besides, at each step of the iterative refinement phase, the target object needs to be 3D-rendered, which is computationally expensive and problematic to achieve real-time applications.

2. Plan of work

2.1. Reproducing results of Dense Fusion

I will start by reviewing the main article on DenseFusion and highlighting its main differences with the previous approaches and notably the DeepIM method [2]. I will use the official implementation [1] and pretrained models of the authors to reproduce the results presented in the article on a 5-objects subset from the Linemod dataset [1]. The choice of the susbset will be made to explore the performance of the approach as well as some failure cases.

[1] https://github.com/j96w/DenseFusion

2.2. Extension to RGB images

I will try to adapt the ideas from the Dense Fusion approach to attempt solving the problem of 6D Object Pose Estimation on RGB-only images. I will use the RGB images derived from the RGB-D images from the Linemod dataset [1]. This will allow the comparison of the performance of the new approach with those of the Dense Fusion method.

2.2.1 Depth estimation

A first idea is to try reconstructing the Depth dimension of the input image by using a deidcated depth-estimation network. The estimated RGB-D image can then be fed to the Dense Fusion model to produce the final 6D Object Pose Estimation.

This a very straightforward approach and easy to add to the previous framework. Besides, there are many pretrained models which achieve state of the art performance on the NYU Depth dataset [3] which contains indoor scenes similar to those from the Linemod dataset. For instance, a recent method based on transformers is introduced in [5] and its implementation is available online[2].

2.2.2 Pixel-wise 6D Pose Estimation

Another possible approach is to remove the iterative refiner and the point cloud embeddings and to leverage the pixel-wise component to produce pixel-wise predictions. An interesting aspect of this approach will be to evaluate the interest of a pixel-wise approach and to study the effect of different loss on the learning performance.

This second approach demands more work and adaptations of the initial framework but offers a larger variety of refinements and is a more innovative way of leveraging the existing method on RGB-D images to estimate 6D Object Pose Estimation for RGB images.

[2] https://github.com/ygjwd12345/TransDepth

References

[1] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 128:858–865, 2011. 1

[2] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128(3):657–678, Nov 2019. 1

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision, ECCV 2012 - 12th European Conference on Computer Vision, Proceedings*, number PART 5 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 746–760, 2012. 12th European Conference on Computer Vision, ECCV 2012 ; Conference date: 07-10-2012 Through 13-10-2012. 1

[4] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion, 2019. 1

[5] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction, 2021. 1