Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd

Andreas Doumanoglou[1,2], Rigas Kouskouridas[1], Sotiris Malassiotis[2], Tae-Kyun Kim[1]

[1]Imperial College London
[2]Center for Research & Technology Hellas (CERTH)

Abstract

Object detection and 6D pose estimation in the crowd (scenes with multiple object instances, severe foreground occlusions and background distractors), has become an important problem in many rapidly evolving technological areas such as robotics and augmented reality. Single shot-based 6D pose estimators with manually designed features are still unable to tackle the above challenges, motivating the research towards unsupervised feature learning and next-best-view estimation. In this work, we present a complete framework for both single shot-based 6D object pose estimation and next-best-view prediction based on Hough Forests, the state of the art object pose estimator that performs classification and regression jointly. Rather than using manually designed features we a) propose an unsupervised feature learnt from depth-invariant patches using a Sparse Autoencoder and b) offer an extensive evaluation of various state of the art features. Furthermore, taking advantage of the clustering performed in the leaf nodes of Hough Forests, we learn to estimate the reduction of uncertainty in other views, formulating the problem of selecting the next-best-view. To further improve 6D object pose estimation, we propose an improved joint registration and hypotheses verification module as a final refinement step to reject false detections. We provide two additional challenging datasets inspired from realistic scenarios to extensively evaluate the state of the art and our framework. One is related to domestic environments and the other depicts a bin-picking scenario mostly found in industrial settings. Our framework significantly outperforms state of the art both on public and on our datasets.

1. Introduction

Detection and pose estimation of everyday objects is a challenging problem arising in many practical applications, like robotic manipulation, tracking and augmented reality. Low-cost availability of depth data facilitates pose estimation significantly, but still one has to cope with many challenges such as viewpoint variability, clutter and occlusions. When objects have sufficient texture, techniques based on key-point matching [5, 6] demonstrate good results, yet when there is a lot of clutter in the scene they depict many false positive matches which degrades their performance. Also, holistic template-based techniques provide superior performance when dealing with texture-less objects [4], but suffer in cases of occlusions and changes in lighting conditions, while the performance also degrades when objects have not significant geometric detail. In order to cope with the above issues, a few approaches use patches [7] or simpler pixel based features [2] along with a Random Forest classifier. Although promising, these techniques rely on manually designed features which are difficult to make discriminative for the large range of everyday objects.

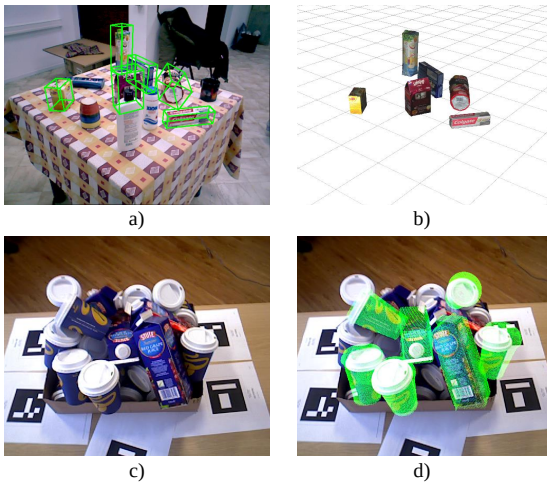

Figure 1: Sample photos from our dataset. a) Scene containing objects from a supermarket, b) our system's evaluation on a), c) Bin-picking scenario with multiple objects stacked on a bin, d) our system's evaluation on c).

Last, even when the above difficulties are partly solved, multiple objects present in the scene, occlusions and distructors can make the detection very challenging from a

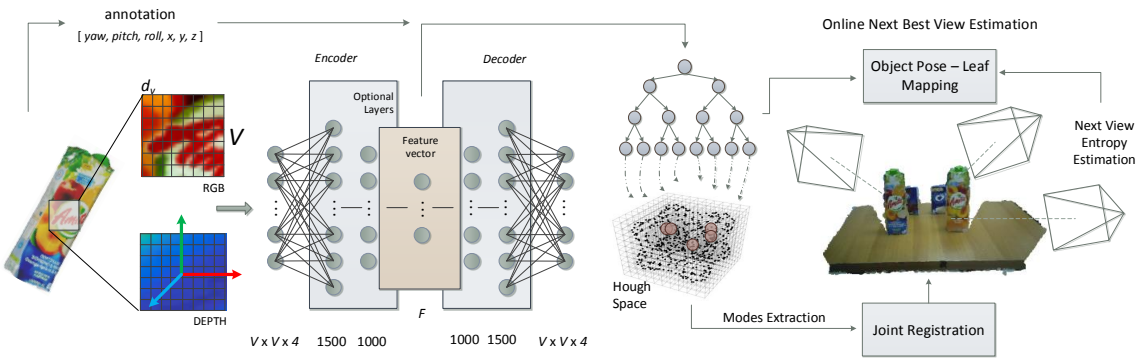

Figure 2: Framework Overview. After patch extraction, RGBD channels are given as input to the Sparse Autoencoder. The annotation along with the produced features of the middle layer are given to a Hough Forest, and the final hypotheses are generated as the modes of the Hough voting space. After refining the hypotheses using joint registration, we estimate the next-best-view using a pose-to-lead mapping learnt from the trained Hough Forest.

single viewpoint, resulting in many ambiguous hypotheses. When the setup permits, moving the camera to another viewpoint can be proved very beneficial for accuracy increase. However the problem is how to select the next best viewpoint, which is crucial for fast scene understanding.

2. 6 DoF Object Pose & Next-Best-View Estimation Framework

The above observations motivated us to introduce a complete framework for both single shot-based 6D object pose estimation and next-best-view prediction in a unified manner based on Hough Forests, a variant of Random Forest that performs classification and regression jointly [7]. We adopted a patch-based approach but contrary to [4, 7, 2] we learn features in an unsupervised way using deep Sparse Autoencoders. The learnt features are fed to a Hough Forest [3] to determine object classes and poses using 6D Hough voting. To estimate the next-best-view, we exploit the capability of Hough Forests to calculate the hypotheses entropy, i.e. uncertainty, at leaf nodes. Using this property we can predict the next-best-viewpoint based on current view hypotheses through an object-pose-to-leaf mapping. We are also taking into account the various occlusions that may appear from the other views during the next-best-view estimation. Last, for further false positives reduction, we introduce an improved joint optimization step inspired by [1]. To the best of our knowledge, there is no other framework jointly tackling feature learning, classification, regression and clustering (for next-best-view) in a patch-based inference strategy, and doing all by a deep network is not trivial.

Our object detection and pose estimation framework consists of two main parts: a) single shot-based 6D object detection and b) next-best-view estimation. In the first part, we render the training objects and extract depth-invariant RGB-D patches. The latter are given as input to a Sparse Autoencoder which learns a feature vector in an unsupervised manner. Using this feature representation, we train a Hough Forest to recognize object patches in terms of class and 6D pose (translation and rotation). Given a test image, patches from the scene pass through the Autoencoder followed by the Hough forest, where the leaf nodes cast a vote in a 6D Hough space indicating the existence of an object. The modes of this space represent our best object hypotheses. The second part, next-best-view estimation, is based on the previously trained forest. Using the training sample distribution in the leaf nodes, we are able to determine the uncertainty, i.e. the entropy, of our current hypotheses, and further estimate the reduction in entropy when moving the camera to another viewpoint using a pose-to-leaf mapping. Fig. 2 shows an overview of the framework.

Evaluation on single shot detection of various state of the art features and detection methods, shows that the proposed approach demonstrates a significant improvement on many challenging publicly available datasets. We also evaluate our next-best-view selection to various baselines and show its improved performance, especially in cases of occlusions. To demonstrate more explicitly the advantages of our framework, we provide an additional dataset consisting of two realistic, everyday scenarios shown in Fig. 1. Our dataset also reveals the weaknesses of the state of the art techniques to generalize to realistic scenes.

References

[1] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze. A global hypotheses verification method for 3d object recognition. In *ECCV*. 2012. 2

[2] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*. 2014. 1, 2

[3] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 2011. 2

[4] 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. In *ICCV*, 2011. 1, 2

[5] M. Martinez, A. Collet, and S. S. Srinivasa. Moped: A scalable and low latency object recognition and pose estimation system. In *ICRA*, 2010. 1

[6] J. Tang, S. Miller, A. Singh, and P. Abbeel. A textured object recognition pipeline for color and depth image data. In *ICRA*, 2012. 1

[7] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim. Latent-class hough forests for 3d object detection and pose estimation. In *ECCV*. 2014. 1, 2