Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes

Stefan Hinterstoisser[1], Vincent Lepetit[3], Slobodan Ilic[1], Stefan Holzer[1], Gary Bradski[2], Kurt Konolige[2], and Nassir Navab[1]

[1] CAMP, Technische Universität München (TUM), Germany
[2] Industrial Perception, Palo Alto, CA, USA
[3] CV-Lab, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract. We propose a framework for automatic modeling, detection, and tracking of 3D objects with a Kinect. The detection part is mainly based on the recent template-based LINEMOD approach [1] for object detection. We show how to build the templates automatically from 3D models, and how to estimate the 6 degrees-of-freedom pose accurately and in real-time. The pose estimation and the color information allow us to check the detection hypotheses and improves the correct detection rate by 13% with respect to the original LINEMOD. These many improvements make our framework suitable for object manipulation in Robotics applications. Moreover we propose a new dataset made of 15 registered, 1100+ frame video sequences of 15 various objects for the evaluation of future competing methods.

1 Introduction

Many current vision applications, such as pedestrian tracking, dense SLAM [2], or object detection [1], can be made more robust through the addition of depth information. In this work, we focus on object detection for Robotics and Machine Vision, where it is important to efficiently and robustly detect objects and estimate their 3D poses, for manipulation or inspection tasks. Our approach is based on LINEMOD [1], an efficient method that exploits both depth and color images to capture the appearance and 3D shape of the object in a set of templates covering different views of an object. Because the viewpoint of each template is known, it provides a coarse estimate of the pose of the object when it is detected.

However, the initial version of LINEMOD [1] has some disadvantages. First, templates are learned online, which is difficult to control and results in spotty coverage of viewpoints. Second, the pose output by LINEMOD is only approximately correct, since a template covers a range of views around its viewpoint. And finally, the performance of LINEMOD, while extremely good, still suffers from the presence of false positives.

In this paper, we show how to overcome these disadvantages, and create a system based on LINEMOD for the automatic modeling, detection, and tracking

K.M. Lee et al. (Eds.): ACCV 2012, Part I, LNCS 7724, pp. 548–562, 2013.

Fig. 1. 15 different texture-less 3D objects are simultaneously detected with our approach under different poses on heavy cluttered background with partial occlusion. Each detected object is augmented with its 3D model. We also show the corresponding coordinate systems.

of 3D objects with RGBD sensors. Our main insight is that a 3D model of the object can be exploited to remedy these deficiencies. Note that accurate 3D models can now be created very quickly [2–5], and requiring a 3D model beforehand is not a disadvantage anymore. For industrial applications, a detailed 3D model often exists before the real object is even created.

Given a 3D model of an object, we show how to generate templates that cover a full view hemisphere by regularly sampling viewpoints of the 3D model. We also show how the 3D model can be used to obtain a fine estimate of the object pose, starting from the one provided by the templates. Together with a simple test based on color, this allows us to remove false positives, by checking if the object under the recovered pose aligns well with the depth map. Moreover, we show how to define the templates only with the most useful appearance and depth information, which allows us to speed up the template detection stage. The end result is a system that significantly improves the original LINEMOD implementation in performance, while providing accurate pose for applications.

In short, we propose a framework that is easy to deploy, reliable, and fast enough to run in real-time. We also provide a dataset made of 15 registered, 1100+ frame video sequences of 15 various objects for the evaluation of future competing methods. In the remainder of this paper we first discuss related work, briefly describe the approach of LINEMOD, introduce our method, represent our dataset and present an exhaustive evaluation.

2 Related Work

3D object detection and localization is a difficult but important problem with a long research history. Methods have been developed for detection in photometric images and range images, and more recently, in registered color/depth images. We discuss these below.

Camera Images. We can divide image-based object detection into two broad categories: learning-based and template approaches. Learning-based systems