Master thesis proposal

Texture-less Object Detection in RGB-D Images

Tomás Golvano García

Supervisor: Prof. Jiří Matas[1]
Advisor: Prof. Julián Fiérrez[2]

Prague, March 2016

[1] Center for Machine Perception, Czech Technical University
[2] Escuela Politécnica superior, Universidad Autónoma de Madrid

Assignment

1. Review the state-of-the-art on object detection.
2. Consider their application to texture-less object detection.
3. Implement the selected method.
4. Perform an experimental evaluation of the method on selected datasets, and compare the obtained results with other methods.
5. Propose possible improvements.

Contents

1 Introduction

Even though they are omnipresent in many environments, recognition and localization of texture-less objects is challenging in several respects. Appearance of a texture-less object is dominated by its shape, its material properties and by the configuration of light sources. Such objects present significant challenges to contemporary visual object detection methods. This is mainly because common local appearance descriptors are not discriminative enough to provide reliable correspondences.

In 2D images, the most sensible solution to describe the texture-less objects is to use a representation amenable to the object edges, i.e. the points where image characteristics change sharply. These edges correspond mainly to objects' outline in the case of texture-less objects.

Depth images used as additional input can simplify the detection task. This kind of images, known as RGB-D images, have aligned color and depth, simultaneously describing the appearance and geometry of the scene. The RGB-D images can be obtained using Kinect-like sensors. Extra information obtained with the 3D shape allows a more detailed description, expecting to decrease the number of hallucinations.

The rest of this document is organized as follows. Section 2 reviews the existing methods suitable for detection of texture-less objects, either in RGB or RGB-D images. Section 3 presents a work plan and the goals of the thesis.

2 Related Work

We review four categories of methods for detection of texture-less objects: template matching methods, shape matching methods, methods based on dense features, and deep learning methods.

One of the earliest techniques applied to object detection in images was template matching. In this methods every object is represented by a set of templates whom capture possible global object ocurrences thoroughly. Every template (RGB or RGB-D) captures the 3D pose of the object. Is usual to have also templates that feature occlusion, background cluttering and conditions under different light sources. Correlation coefficients express the similarity between a window and a template, if the correlation is high enough its score means a likely match. Correlation is suitable for edges, image gradiens, color, depth and also 3D shapes. One example of this methods can be found in Hinterstoisser et al. [9] and a sub-linear complexity method achieved in Cai et al [5].

Shape matching methods aspire to represent the object shape by relative relationships between 2D or 3D shape features, either within local neighborhoods or globally over the whole image. Therefore, a set of feature

descriptors use to represent each object. When there are correspondences between training and test descriptors, the detection hypotheses are generated. This correspondences are voted in a similar way as Hough detector does. This methods operate bottom-up and they use traditional detection methods, using local appearance features [16] being used for 3D shape-based method in Drost et al. [10].

Another category of bottom-up methods is based on dense features, where every pixel is involved in prediction about the detection output. The pixel is described either by a descriptor of local patch surrounding the pixel or simple measurements in the local pixel's neighborhood. Since local 2D features are not descriptive enough in the case of texture-less objects, these methods can be propitiously used only if depth information is available, which allows for a richer description of the local neighborhood. This approaches showed to be suitable for texture-less objects [12].

Convolutional neural networks (CNN) yield remarkable results in many computer vision fields. Up to now, in the field of texture-less object detection, Wohlhart et al. [17] used CNN to obtain descriptors of object views that efficiently capture both the object identity and 3D pose. The CNN was trained by enforcing simple similarity and dissimilarity constraints between the descriptors, untangling the images from different objects and different views into clusters that are not only well-separated but also structured as the corresponding sets of poses. The method can work with either RGB or RGB-D images and outperforms the state-of-the-art methods on the dataset of Hinterstoisser et al. [8]. Another method using CNN is the one presented by David Held et al. [13], where they also outperformed the state-of-the-art using a CNN, they introduced a new approach for recognition with limited training data, in which they used multi-view dataset to train their network to be robust to viewpoint changes, being able to improve the recognition of objects with a single image training for each object.

3 Work Plan and Goals

The first two tasks from the assignment have been completed – a brief review of the state-of-the-art methods together with discussion about their application to texture-less objects can be found in Section 2.

The next step will be to implement and analyze a baseline method for texture-less object detection in 2D images. As the baseline method, we selected a sliding window approach using histogram of oriented gradients (HOG) to describe the window content and support vector machines (SVM) to classify the HOG descriptors. This approach was successfully applied to pedestrian detection by Dalal and Triggs [7].

HOG is an image descriptor presented by Dalal and Triggs [7], which counts occurrences of gradient orientation in localized portions of an image.

The HOG descriptor effectively encodes the object's shape. The shape is one of the dominating properties of texture-less objects and thus the HOG descriptor seems to be a suitable choice to describe this type of objects.

Additionally, we would like to extend the HOG descriptor to depth images and compare its discriminative power to its original version. Beside the HOG + SVM approach, the plan is to experiment also with the convolutional neural networks (CNN). Given its uprising in many fields of computer vision, in which it achieves the state-of-the-art results, it is nearly a must to try them also for the texture-less object detection.

4 Tools

4.1 Software

- MATLAB
- VLFeat: Cross-platform open source collection of vision algorithms with a special focus on visual features and clustering. It bundles a MATLAB toolbox, a clean and portable C library and a number of command line utilities.

4.2 Hardware

For the experimental evaluation, datasets captured with the following sensors will be used:

- Primesense Carmine 1.09 (Short Range)
- Microsoft Kinect v2
- Canon Digital IXUS 950 IS

References

[1] T. Hodaň, J. Matas, *Texture-less object detection– PhD thesis proposal.* Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic. September 2015.

[2] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, J. Matas, *Detection and Fine 3D Pose Estimation of Texture-less Objects in RGB-D Images* International Conference on Intelligent Robots and Systems (IROS) 2015, Hamburg, Germany

[3] A. Collet, M. Martinez, S. Srinivasa, *The MOPED framework: object recognition and pose estimation for manipulation.* I. J. Robotic Res., vol. 30, no. 10, pp. 1284–1306, 2011

[4] T. Tuytelaars, K. Mikolajczyk, *Local invariant feature detectors: A survey.* Found. Trends. Comput. Graph. Vis., vol. 3, no. 3, pp. 177–280, July 2008.

[5] H. Cai, T. Werner, and J. Matas. *Fast detection of multiple textureless 3-D objects.* ICVS, volume 7963 of LNCS, pages 103–112. 2013.

[6] F. Tombari, A. Franchi, L. Di, *BOLD features to detect textureless objects.* ICCV, 2013, pp. 1265–1272.

[7] N. Dalal, B. Triggs. *Histograms of oriented gradients for human detection.* International Conference on Computer Vision & Pattern Recognition (CVPR '05), Jun 2005, San Diego, United States. IEEE Computer Society, 1, pp.886–893, 2005, . ¡10.1109/CVPR.2005.177¿.

[8] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. *Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes.* ACCV, 2012.

[9] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. *Gradient response maps for real-time detection of textureless objects.* IEEE PAMI, 2012.

[10] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. *Model globally, match locally: Efficient and robust 3D object recognition.* CVPR, pages 998–1005, 2010.

[11] E. Brachmann, A. Krull, F. Michel, S. Gumhold, and J. Shotton. *Learning 6D object pose estimation using 3D object coordinates.* ECCV, 2014.

[12] E. Brachmann, A. Krull, F. Michel, S. Gumhold, and J. Shotton. *Learning 6D object pose estimation using 3d object coordinates.* ECCV, 2014.

[13] D. Held, S. Thrun, S. Savarese. *Deep learning for single-view instance recognition.* Stanford University arXiv:1507.08286v1 [cs.CV] 29 Jul 2015

[14] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. *Object detection with discriminatively trained part based models.*PAMI, 2009.

[15] Jan Fischer , R. Bormann1 , G. Arbeiter and A. Ver. *A feature descriptor for texture-less object representation using 2D and 3D cues from RGB-D data.* 2013 IEEE International Conference on Robotics and Automation (ICRA) Karlsruhe, Germany, May 6-10, 2013

[16] D.G. Lowe. *Object recognition from local scale-invariant features.* In ICCV, volume 2, pages 1150–1157, 1999.

[17] P. Wohlhart, V. Lepetit. textitLearning descriptors for object recognition and 3D pose estimation. arXiv preprint arXiv:1502.05908, 2015.