Object Reconstruction and Recognition leveraging an RGB-D camera

Nicolas Burrus Mohamed Abderrahim Jorge Garcia Luis Moreno
Department of Systems Engineering and Automation, Carlos III University
Av. Universidad 30, 28911 Leganés (Madrid), Spain
E-mail: firstname.lastname@uc3m.es

Abstract

Recently, sensing devices capable of delivering real-time color and depth information have become available. We show how they can benefit to 3D object model acquisition, detection and pose estimation in the context of robotic manipulation. On the modeling side, we propose a volume carving algorithm capable of reconstructing rough 3D shape with a low processing cost. On the detection side, we find that little robustness can be directly added to classical feature-based techniques, but we propose an interesting combination with traditionally less robust techniques such as histogram comparison. We finally observe that 3D pose estimates can also be greatly improved using the depth measurements.

1 Introduction

Recently, so-called RGB-D cameras have become available, capable of delivering synchronized color (RGB) and depth (D) information in real-time. The depth information is dense, and comes at negligible additional processing cost for the host CPU. They avoid the complexity of robust disparity map computation of stereo systems, and are much faster than laser scanning techniques. Thus, these sensors are very attractive for the computer vision community and their benefits to classical applications are worth investigating.

In this paper, we study how they can benefit to some of the computer vision tasks involved in robotic object manipulation. More specifically, we focus on how the depth information can simplify the acquisition of new 3D object models, improve object recognition robustness, and make the estimation of the 3D pose of detected objects more accurate.

Our particular hardware setup and limitations are introduced in Section 2. In Section 3, we briefly review the state-of-the-art of RGB-D based model acquisition and propose a simple but efficient algorithm based on volume carving. We then study in Section 4 how the provided depth information can benefit to feature-based object recognition in cluttered scenes. We show that their ability to roughly segment object surfaces can make the combination with classically less robust techniques such as color histograms more attractive. We finally observe in Section 5 that 3D pose estimation is also greatly improved when few feature matches are available.

2 Hardware setup and calibration

Several RGB-D cameras are currently available. Most of them are based on a Time-of-Flight (ToF) principle, and measure the time taken by an emitted infrared (IR) signal to reach back an array of IR sensors. Kolb [5] gives an overview of the existing cameras and technical details.

Figure 1. Left: PMD Camcube 2.0 coupled with a color camera. Right: Turn-table setup using a PA10 robot.

The ToF camera used for all our experiments is a PMD Camcube 2.0. It is a ToF camera with a spatial resolution of 204x204 pixels, and a depth range of 30cm-7m. Depth precision is ∼2 cm and repeatability is ∼3 mm. It provides synchronized depth and grayscale images. To capture detailed textures of the scene, we coupled the camera with a classical color webcam, as shown on Figure 1. The resolution of the grayscale output of the PMD camera is high enough to enable classical stereo calibration, such as Bouguet's method. It is thus possible to get color information for each pixel of the depth image by back-projecting the pixel to 3D using the estimated depth, and then projecting it onto the color image.

For model acquisition, the setup includes a turntable driven by an robotic arm. This enables repeatable and calibrated points of view.

3 Object Model Acquisition

3.1 Previous work

Only a few methods have been proposed for object model acquisition using RGB-D cameras. In [4], classical multi-view stereo is combined with a ToF camera to reconstruct poorly textured regions. Using only a single ToF camera, Cui [2] focuses on providing high quality models with costly super-resolution techniques. Using a different kind of RGB-D camera, Krainin [6] proposes a surface-based technique relying on "Surface Elements". This technique however requires a high depth precision that is not currently provided by existing ToF cameras.

3.2 Reconstruction based on volume carving

Observing that the ToF camera is however very good at delivering silhouettes of the objects in a scene, we propose to rely on a silhouette-based technique. Silhouettes have been used extensively in the literature [7, 10], but their computation is still problematic using classical cameras. Uniform or easily discriminated

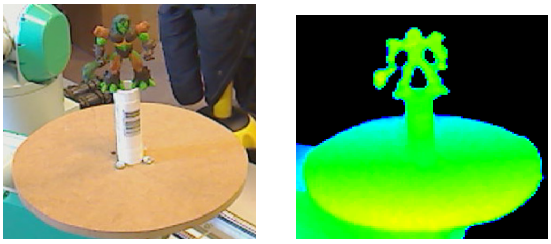

Figure 2. Silhouette extraction using a ToF camera and simple depth thresholding. The depth image is color-encoded.

backgrounds are usually required, resulting into a less flexible system.

However, using depth information, it becomes easy to discriminate an object of interest from the background, and silhouettes can be obtained with a simple depth threshold, as illustrated on Figure 2. To demonstrate this, we developed a simple space carving technique based on the depth measurements. It is similar in spirit to [10] but here no photo-consistency step is applied. This prevents the reconstruction of small concave details but the reconstruction is faster and the obtained models are accurate enough for manipulation tasks.

The proposed algorithm iteratively carves a 3D discrete volume. Each voxel is represented as a cube whose size is used-defined in function of the level of details required. For each view taken by the ToF camera, voxels are eliminated according to their depth compatibility with the new view. This is done by projecting each voxel onto the depth image, and comparing its depth d_{voxel} with the measured depth d_{view}. If $d_{view} > d_{voxel} + \delta$, it means that the voxel is not consistent in the scene, and it is discarded. δ is the tolerated margin and depends on the sensor precision. In all our experiments it is set to 3 cm to be conservative and avoid removing voxels with the Camcube camera. This value is not very sensitive since most of the carving will come for the edges. A small value however enables the reconstruction of concave structures whose depth is greater than δ.

Depending on the voxel size, the projection of one voxel can overlap several pixels on the depth image. The actual overlap is approximated by computing the projected width and height of the voxel and comparing d_{voxel} with the depth measurements in the corresponding neighborhood in the depth image. If at least one d_{view} is compatible with d_{voxel}, the voxel is kept.

3.3 Post-processing and results

The output of the algorithm is a rather dense set of cube-shaped voxels. For manipulation tasks, it is more useful to get a surface representation of the object. This can be achieved by first removing all the inside voxels with a simple neighborhood test, and then run a surface reconstruction algorithm such as Poisson [3]. Some results are given in Figure 3 and 4. These models were acquired using 35 views captured by rotating the turntable by 10° steps. The camera was at a distance of about 50 cm from the object and the chosen voxel size is 1 mm. Processing time is currently less than 10 seconds on a 2Ghz computer for 36 views, and real-

Figure 3. Example of acquired model: cup. Top left: object to scan, top right: carved volume, bottom left: Poisson reconstruction, bottom right: reprojection on a color image with known pose.

time performance should be reachable using a careful implementation.

4 Object Recognition

In this section we investigate how the ToF camera could also benefit to the recognition and 3D localization of the object models in new scenes. Many techniques have been proposed for recognition, but since the influential work of [8], methods based on local feature matching and clustering have become very popular since they give good results for textured-enough objects. We thus choose a method derived from [8] as our baseline algorithm, and in this paper we focus on how a ToF camera can contribute to improve detection rates.

4.1 Baseline detection algorithm

We recall here the main principles of the object detection method. First, SIFT points are extracted on each color view of the acquired models and stored in a database. Then, SIFT points are extracted on the image to analyze, and are matched to the closest ones in the database. Each single feature point association results in a candidate pose for the corresponding model. Compatible pose candidates are then clustered, and a statistical criterion is used to make reliable detections.

4.2 First attempts

Some applications of an RGB-D camera to feature matching seem straightforward. However, we observed that due to the lack of precision of the measurements, some direct improvements turned out to bring only a marginal benefit. We find it interesting to comment some of them:

1. Use depth to accelerate SIFT point matching and to discard points whose scale/depth ratio is not coherent. Only about 10% of wrong SIFT matches

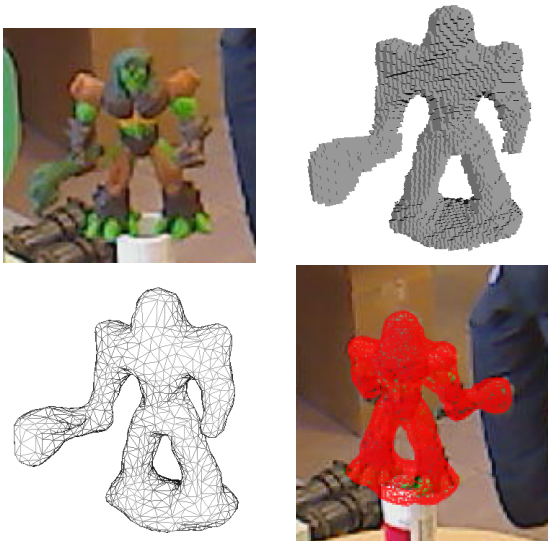

Figure 4. Example of acquired model: toy. Top left: object to scan, top right: carved volume, bottom left: Poisson reconstruction, bottom right: reprojection on a color image with known pose.

could be actually removed using this optimization. Indeed, combining the imprecision of SIFT scale estimation and depth measurements, and considering that most points appear at the same scale, very few points can actually get safely discarded.

2. Use depth to discard outliers during the feature clustering phase. The benefit of such a filtering was also very limited, for two reasons. First, geometrical filters described in [8] are already quite effective to remove outliers. Second, depth measurements are often strongly biased when there is a specularity in the object. In theses areas, valid features may be wrongly discarded, somehow compensating the potential benefits of the filter.

3. Take into account depth information to reduce background influence on SIFT descriptors. This is known to be a serious problem for small object recognition, as pointed out by [9]. Instead of using 2D gaussian to compute the neighborhood influence, we experimented a 3D gaussian including depth differences. However, the low spatial resolution of the depth maps results into fuzzy borders in the color image and the influence of the background remains significant. Errors due to specularities here also compensate the small benefit and the resulting gain is marginal.

4.3 Combination with depth-segmented histogram comparison

These unsuccessful attempts to directly improve the core of feature based methods led us to investigate more regional attributes. We base our study on [1], where SIFT feature matching is combined with regional histogram comparison using only a color image. A major weakness of regional features is however their sensitivity to the surrounding background and to occluding objects.

Figure 5. Effect of depth filtering to avoid the influence of background and occluding pixels in histogram computation. Left, in red: full candidate region. Right: filtered region using depth dissimilarity.

Figure 6. One of the 192 annotated test images used for performance evaluation. Bounding boxes are shown around detected objects, along with local features matches against the database.

In [1], histogram in the analyzed image are computed in the bounding box of a candidate area determined by the feature clustering step. To make histogram computation more robust, we propose to include only pixels whose depth is similar enough to the median depth of the feature points. This discards pixels belonging to further background objects but also belonging to occluding objects if their depth is different enough. The depth similarity threshold has been empirically set to 5cm in all experiments. Figure 5 shows how this filtering method can significantly reduce the amount of clutter in the histogram computation.

To quantify the improvement brought by the depth filter, we built a test dataset. First, 17 models were acquired using the method of Section 3. Then test images were recorded by locating the objects into various configurations, including multiple-objects scenes, heavy background clutter, occlusions and light variations, resulting into 192 images. These images were finally annotated with object bounding boxes to establish a ground truth and evaluate quantitatively the performance of the recognition algorithm. Figure 6 gives an example of test image and Figure 7 presents the detection improvements with depth filtering on this data set. Detection rate is significantly improved, especially for objects with poor texture and salient colors.

5 Improving Pose Estimation

Once an object has been detected, its precise 3D pose has to be estimated. The most common approaches

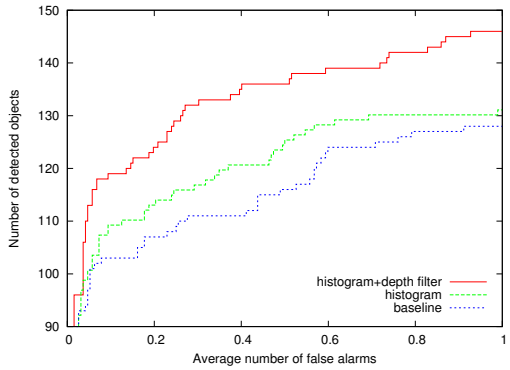

Figure 7. Detection rate of baseline method, baseline combined with color histogram and baseline combined with depth filtered color histogram.

are RANSAC and least square minimization of feature reprojection error. RANSAC performs very well at discarding outliers but requires a high number of feature matches. Since the recognition algorithm can detect objects with as few as 2 or 3 matches, we rely on iterative least square only to minimize the reprojection error when the number of matches is too small. Using only 2D feature coordinates, pose estimation with few points is often under-constrained or unreliable. We observed that introducing the depth information to compute a 3D reprojection error can significantly improve the results. This can be done by defining the following error function:

$$err = \sum_{i=0}^{n} \Delta(H(F_i), C_i)^2$$

with $H(F_i)$ the 3D coordinates of the model feature i projected on the current image according to the pose estimate H, and C_i the (x, y, d) coordinates of the corresponding detected feature point in the analyzed image. When depth is available, the metric distance function Δ is computed as:

$$\Delta(p_1, p_2) = \left[\frac{(x_1 - x_2)}{f_x}\right]^2 + \left[\frac{(y_1 - y_2)}{f_y}\right]^2 + (d_1 - d_2)^2$$

with p_1 and p_2 two 3D points with respective coordinates (x_1,y_1,d_1) and (x_2,y_2,d_2), and f_x, f_y the estimated horizontal and vertical focal lengths of the camera assuming a standard pin-hole model.

The obtained improvements on a scene with occlusion is illustrated on Figure 8.

6 Conclusion

This paper presents some preliminary results regarding the benefits of using a ToF camera for 3D object model acquisition and recognition, in the context of robotic manipulation. On the modeling side, we proposed a volume carving method to rapidly acquire new models. The acquisition setup is quite flexible and does not require a special background. It compares favorably to previously proposed surface-based techniques when the camera depth precision is low.

On the recognition side, we observed that little improvements can be directly added to the core of

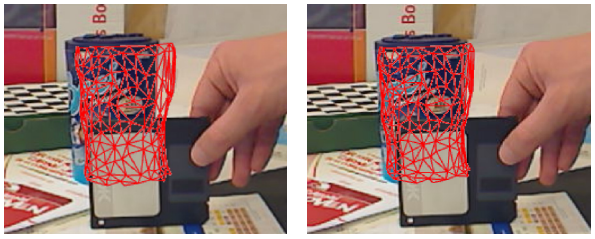

Figure 8. Pose estimation (projected mesh in red) can be improved using ToF measurements (right), especially when few feature points are detected, e.g when objects are occluded.

methods based on local feature clustering. However, depth information can make regional similarity measures more robust and we showed significant improvements on top of state-of-the-art algorithms on a real dataset. Pose estimation also appears to be greatly improved using a 3D reprojection error when only few feature matches are available, but quantitative evaluation is still ongoing work.

Acknowledgements

The research leading to these results has been partially funded by the HANDLE European project (FP7/2007-2013) under grant agreement ICT 231640.

References

[1] N. Burrus, T.M. Bernard, and J.-M. Jolion. Bottom-up and top-down object matching using asynchronous agents and *a contrario* principles. In *Proc. of ICVS*, pages 343–352, 2008.

[2] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3D shape scanning with a time-of-flight camera. In *Proc. of CVPR*, pages 1173–1180, 2010.

[3] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proc. of Eurographics Symposium on Geometry processing*, pages 61–70, 2006.

[4] Y.M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view Image and ToF Sensor Fusion for Dense 3D Reconstruction. In *Proc. of 3DIM*, 2009.

[5] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-Flight Cameras in Computer Graphics. *Computer Graphics Forum*, 29(1):141–159, 2010.

[6] M. Krainin, P. Henry, X. Ren, and D. Fox. Manipulator and Object Tracking for In Hand Model Acquisition. In *ICRA Mobile Manipulation Workshop*, 2010.

[7] K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000.

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] A. Stein and M. Hebert. Incorporating background invariance into feature-based object recognition. In *Proc. of Applications of Computer Vision*, 2005.

[10] G. Walck and M. Drouin. Progressive 3D reconstruction of unknown objects using one eye-in-hand camera. In *Proc. of Robotics and Biomimetics*, pages 971–976, 2010.