21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

RoμNect: Hand Mounted Depth Sensing Using a Commodity Gaming Sensor

Christian Reinbacher Matthias Rüther
Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology
<reinbacher,ruether,bischof>@icg.tugraz.at

Abstract

In this work, we investigate the applicability of the Kinect depth camera as a robot mounted measurement unit. In contrast to traditional head mounted robot sensors, Kinect is small, cheap and delivers robust depth measurements on a variety of scenes. In the course of applying it on a robot arm, we solve a number of problems: we reduce the sensor working distance to a few centimeters, replace the Laser projector unit by a focusable projector, and calibrate this sensor unit. We further exploit the motion capabilities of the robot arm to integrate multiple depth maps with 30 Hz *in a volumetric fusion approach. We show how this method considerably improves completeness of the scanned models, even under severe reflections and difficult surface properties. We employ our approach in a classical bin-picking setting, where the robot scans the object during its approaching motion, and picks it afterwards.*

1 Introduction

The problem of automated robotic picking from a random pile of objects is still an unsolved problem. Left aside the grasp point calculation and gripping, especially the problem of obtaining robust 3D information is crucial. In industrial settings, the sensing problem is traditionally solved by means of laser-based profile scanners, or time-multiplexed structured light systems. A few approaches rely on pure image information and formulate the problem as a recognition task, or a model fitting task. Recently, advances have been made in the discipline of area-based depth sensing. Time-of-Flight sensors generate dense depth information based on runtime measurement, and also projector-camera systems evolved to real-time capable sensing devices.

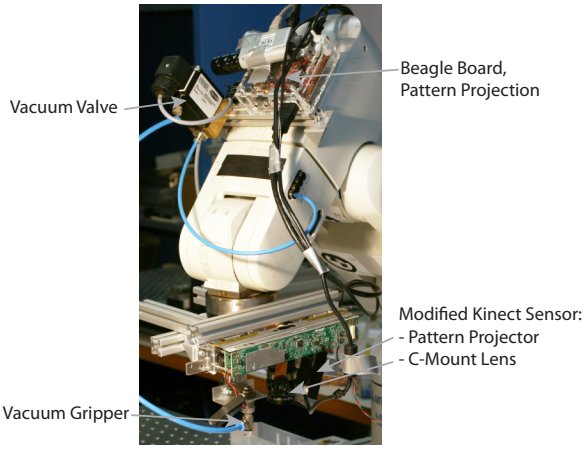


Figure 1. Hardware setup

Especially developments for the computer gaming market triggered new devices which are available at a much lower cost than traditional, industry grade units. Here, especially the Kinect System, manufactured by Primesense and distributed by Microsoft, showed to provide extremely robust depth measurements at a very modest price. While many research disciplines like augmented reality, mobile robotics, and close range photogrammetry, adopted the new device and showed impressive results, research was less active in industrial robotics. The main reason probably is the lack of flexibility in depth range and accuracy. While the Kinect is designed to provide robust depth measurements over a range from 0.5m to 8m, and an accuracy in the range of $10mm$, most industrial sensing problems require higher accuracy at a closer range.

In this paper we use a Kinect Depth sensor and modify it to work at a distance of 200mm, at an accuracy of 0.1mm for a single scan. The resulting system is cheap and lightweight, and delivers depth measurements at

978-4-9906441-1-6

30Hz. As a robot tool, it is applicable in diverse problem domains, e.g. grasping, bin picking or collision avoidance. We demonstrate the potential of our robot mounted RGBD camera by performing 3D reconstructions of bins containing random piles of objects. Due to the robustness of the obtained depth measurements, we are able to quickly perform volumetric fusion of depth measurements during the robot approaches the object to finally obtain a dense depth map.

2 Related Work

The problem of random bin picking is subject to extensive research. One of the very first approaches by Horn and Ikeuchi [2] dates back to 1983, where objects were identified using a Shape from Shading method. Later methods utilize depth cameras to obtain a surface reconstruction of the filled bin. Kristensen *et al*. [4] proposed a bin-picking application for robotic work cell using a solid state range camera mounted above the bin. Most recent approaches are based on dense depth measurements obtained from a head mounted depth sensor (*e.g*. [1]). 3D models of the goods to pick allow an alignment with the dense depth map. Utilizing the processing power of a modern GPU, Park *et al*. [5] show a real time recognition and pose estimation on range images. These approaches rely on an accurate and dense depth map obtained by expensive and highly specialized depth sensors.

In contrast to the existing methods, we apply a very lightweight, cheap and robust depth sensor and mount it directly to the robot hand. During the approach movement, the sensor is able to robustly scan a bin and provide dense depth information. The proposed depth acquisition method can be seen as a plug-in replacement for existing depth measurement systems.

3 Depth Map Creation

The typical working range of a robot mounted depth measurement device should be in the range of 10cm to 50cm to capture the majority of goods which are manipulated by a robot. This contradicts the usage of gaming depth cameras, because these are typically designed for detecting and tracking people in an indoor environment. We therefore modify the original depth sensor by applying a different lens to the camera, a micro-projector instead of the original laser, and modify the baseline to reduce the working distance and increase depth accuracy [6]. In the following section we discuss the hardware modifications made, in conjunction with a robot-based and fully automatic calibration method. We further exploit the extreme robustness of resulting depth measurements in a depth fusion framework, which allows us to scan larger parts of the bin during the robot approach and finally obtain a dense volumetric representation.

3.1 Sensor Hardware

The Kinect sensor relies on the projection of a static, pseudo-random dot grid pattern. A temperature-stabilized Laser diode generates a holographic projection through a diffractive optical element (DOE). This creates a depth of field which is far beyond what is possible with traditional optics, ranging from approximately 50cm in the short range to a maximal working distance of roughly 10m.

As a general disadvantage of this approach, the projected image cannot be refocused or zoomed, because this would destroy the pattern coherence. We therefore replace the Laser by a pico-projection unit. These units are cheap, small and lightweight. They rely on traditional optical projection, and can therefore be refocused and potentially also zoomed. It may be overly complex to employ a video beamer for the static projection of a dot pattern. In practice however, it is advantageous to be able to switch the pattern on and off, to project calibration patterns, and to project intensity-modulated dot grids [6].

Our hardware modifications are as follows: on the camera side, we added a standard C-mount lens. We removed the IR bandpass filter, mainly for practical reasons, but the system can also be realized in the near-infrared spectrum of light with a different projector. The baseline of the projector-camera system is approximately 5cm. With a working distance of 20cm, this results in a stereo angle of 15deg. The projector itself is a Texas Instruments DLP unit, which weighs 86g. An embedded PC (Beagleboard) performs the pattern projection. Camera, projector, Kinect processor and Beagle board are hand-mounted. A USB cable and an Ethernet cable connect the hand to a static PC, on which robot control and depth map fusion are performed. The robot mounted setup weighs 750g in total, where most of the weight stems from the high-quality camera lens, which is not absolutely necessary. The modified Kinect delivers depth maps at a rate of 30Hz, and observes a volume of $6 \times 4\text{cm}^2$, at a focused depth range of 2cm. Theoretically, a robot approaching the bin at 0.3m/s would receive a depth map every traveled centimeter, which is sufficient to perform depth map fusion and collision detection. Figure 1 shows the hardware setup used.

3.2 Calibration

Clearly, the hardware modification of the Kinect sensor destroys its geometric calibration. To obtain a valid

depth map, the following parameters need to be recovered: epipolar geometry between camera and projector, a pixel-wise metric scale, and a tool-hand calibration.

Our sensor comprises a special case of a projector-camera system, with slightly different calibration requirements. We start with a monocular calibration of the camera using the method of [9], to obtain lens distortion and poses relative to a fixed reference target. In a next step, we seek to determine the homography $\mathbf{H}_{\mathbf{P}}^{rec}$, which relates the default projector pattern to a rectified configuration with the camera view, such that the pattern is detectable and usable by the Kinect processor. We perform this calibration through a Hardware-in-the-loop optimization of the unknown homography $\mathbf{H}_{\mathbf{P}}^{rec}$:

$$\arg\min_{\mathbf{H}_{\mathbf{P}}^{rec}} \|\mathcal{F}_{img}(\mathbf{H}_{\mathbf{P}}^{rec} * \mathbf{x}_{\mathbf{P}}^{i}) - \mathbf{x}_{\mathbf{C}}^{i}\|, \tag{1}$$

where $\mathbf{x}_{\mathbf{P}}^{i}$ are points in in projector image space, and $\mathbf{x}_{\mathbf{C}}^{i}$ is the desired, rectangular point grid in camera image space. Function $\mathcal{F}_{img}(\mathbf{x})$ maps points from the projector to the camera image, and depends on scene structure. We realize the point mapping by physically projecting a grid of feature points onto a plane, acquiring it with the camera and determining feature locations in the image. The resulting reprojection cost (1) is minimized iteratively by using the Levenberg-Marquardt algorithm.

In the second step, we use the known camera poses from intrinsic calibration, and the corresponding robot hand poses, to perform tool-hand calibration according to [7].

At this point, as the Kinect processor still delivers a depth map scaled with its internal, hard coded geometric parameters, we identify a per-pixel linear scaling to generate a metric correct depth map. This step is based on a set of known camera poses above a reference plane π_{ref}, which are easily generated through the known tool-hand calibration. At each camera pose with projection matrix $\mathbf{P}$, a ground-truth depth map is generated by intersecting the camera projection rays with π_{ref}:

$$\hat{\mathbf{d}}_{\mathbf{i}} = \mathbf{P}_3 \begin{bmatrix} \mathbf{P} \\ \pi_{ref} \end{bmatrix}^{-1} \begin{bmatrix} x_i \\ y_i \\ 1 \\ 0 \end{bmatrix} \tag{2}$$

with $\mathbf{P}_3$ the third row of $\mathbf{P}$ and $||\mathbf{P}_{3,1\ldots3}|| = 1$. Consequently for each pixel location (x, y) a vector of true depths $\hat{\mathbf{D}} = (\hat{d}_0(x,y), ..., \hat{d}_n(x,y))$ and estimated depths $\mathbf{D} = (d_0(x,y), ..., d_n(x,y))$ is available. The linear relation between these is again established in a least squares sense by solving

$$\begin{bmatrix} \mathbf{D}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a}(x,y) \\ \mathbf{b}(x,y) \end{bmatrix} = \hat{\mathbf{D}}^T. \tag{3}$$

A correct depth map is finally retrieved by mapping an incoming depth value according to $\hat{d} = ad + b$.

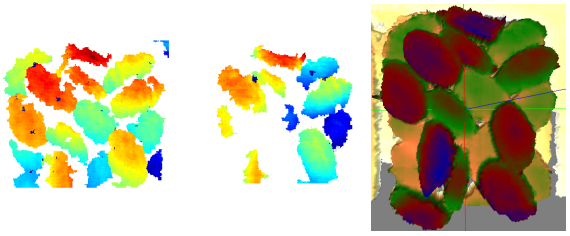
3.3 Depth Fusion

The modified Kinect depth measurement system delivers extremely robust results. It effectively eliminates the need to deal with gross outliers. On the other hand, depth maps may become sparse, if the object surfaces are extremely dark, or the geometric level of detail is below the projected pattern resolution. To produce dense 3D measurements even under adverse conditions, we employ a volumetric depth fusion method, which is related to the work of Izadi *et al*. [3]. We initialize a volumetric occupancy grid at a desired resolution, e.g. 0.25mm per voxel. For a given camera pose and depth measurement, we cast a ray back into the volume, and accumulate the number of hits or misses for each voxel the ray passes.

The final, fused depth map can be simply reconstructed by casting rays perpendicular to the z plane of the world coordinate system. For each ray, we keep the voxel center with the highest accumulated vote and encode its z coordinate in the depth map. In this way, an arbitrary number of depth maps can be fused. Please note that the depth maps could also be fused by a more sophisticated fusion technique (*e.g*. [8]). However, our experiments showed that the simple winner-takes-all strategy produces depth maps with sufficient density and accuracy. Figure 2 shows the comparison of a single depth map and a fused depth map from 30 depth images, taken while approaching the object. As can be clearly seen, the modified Kinect has a limited depth-of-field which is exceeded by the object to be measured. By combining depth maps taken from different ground levels, a dense depth map, suitable for further processing, can be created.

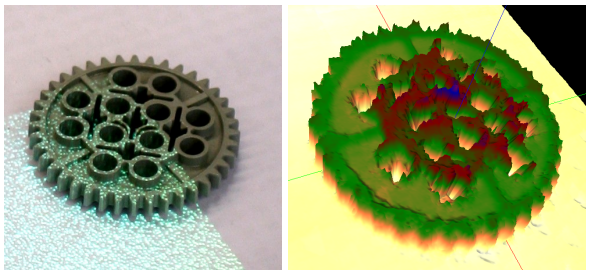
4 Experiments

In this Section we show how the incremental fusion of depth map increases the completeness of the resulting depth map. In the experiment we obtain a depth map of a region of $150 \times 80\text{mm}^2$. A bin ($55 \times 120 \times 35$mm) containing various objects, is placed in this region of interest. Depth measurements from the sensor are iteratively combined according to Section 3.3 with a resolution of 0.25mm. Table 4 shows the number of entries in the depth map for an increasing number of input depth maps. As can be seen, a single measurement results in depth maps with many empty entries. However, after



(a) (b) (c)

Figure 2. Example of depth map fusion. Two raw depth maps (a-b) and a fused result is shown. White in (a-b) depicts regions for which no depth measurement could be obtained.



(a) (b)

Figure 3. High quality depth map (b) of an untextured plastic part (a) obtained by fusing 50 single depth measurements.

Table 1. Completeness of the obtained depth maps.

Object	Completeness[%]		
	1 frame	20 frames	50 frames
keys	19.69	75.53	88.67
pillbox	14.97	56.06	86.31
puzzle	20.59	61.38	86.31
lego	15.19	74.89	89.61

approximately 50 fused frames the depth maps show a completeness of 85-90% independent of the object type.

On a PC with a 2.66 GHz Core i7 processor, the insertion of a single depth map takes **20** ms which allows to create the depth maps while approaching the object. The creation of the final depth map takes 6ms.

Figure 3(b) shows an example depth map obtained in the previous experiment. The real object with a part of the projected pattern can be seen in Fig. 3(a). Please note how the fine details of the gear could be recovered.

5 Conclusion

In this work we have shown how to modify and calibrate a consumer grade depth sensor to serve in industrial applications. By exploiting the robustness and price advantage, triggered by its original application in the gaming industry, the modified sensor delivers dense depth maps with 30 Hz with an accuracy which is adequate for most pick and place tasks for a fraction of the price of a conventional depth sensor. Through the fusion of multiple depth maps during robot motion, the effective measurement range and measurement completeness can be extended in real-time. The system will allow the cheap solution of pick-and-place tasks and pose estimation problems, and enable a broader application of vision systems in industrial settings.

Acknowledgements

We would like to thank Gerhard Reitmayr for providing an open-source implementation of KinectFusion. This work was supported by the Austrian Research Promotion Agency (FFG) under the project SILHOUETTE (825843).

References

[1] S. Fuchs, S. Haddadin, M. Keller, S. Parusel, A. Kolb, and M. Suppa. Cooperative bin-picking with time-of-flight camera and impedance controlled dlr lightweight robot iii. In *Intelligent Robots and Systems (IROS)*, 2010.

[2] K. Horn, Berthold K.P.; Ikeuchi. Picking parts out of a bin. Technical report, Massachussetts Institute of Technology, 1983.

[3] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, 2011.

[4] S. Kristensen, S. Estable, M. Kossow, and R. Brösel. Bin-picking with a solid state range camera. *Robotics and Autonomous Systems*, 35:143 – 151, 2001.

[5] I. Park, M. Germann, M. Breitenstein, and H. Pfister. Fast and automatic object pose estimation for range images on the gpu. *Machine Vision and Applications*, 21:749–766, 2010. 10.1007/s00138-009-0209-8.

[6] M. Ruther, M. Lenz, and H. Bischof. μnect: On using a gaming rgbd camera in micro-metrology applications. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011.

[7] R. Tsai and R. Lenz. Real time versatile robotics hand/eye calibration using 3d machine vision. In *ICRA*, 1988.

[8] C. Zach. Fast and high quality fusion of depth maps. In *3DPVT*, 2008.

[9] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *International Conference on Computer Vision*, volume 1, pages 666 –673 vol.1, 1999.