Masterthesis: Optimization of Augmented Reality Applications considering the Depth Information with Googles Project Tango

Technical Report

by **Steffen Tröster**
Cologne University of Applied Sciences

Prof. Dr. Hubert Randerath
Prof. Dr. Martin Eisemann

in cooperation with **inovex GmbH**

Project Tango is a new mobile platform by Google's Advanced Technology and Projects (ATAP) team which brings motion tracking, depth perception, and area learning to smartphones and tablets. This technology can be used to realize real world measurement applications, indoor navigation and virtual reality environments. With its motion tracking technology, Project Tango is also suitable for precise three dimensional augmented reality (AR) applications. The illusion in those applications can be realized by equating the extrinsic and intrinsic camera properties of the real and the virtual camera in a virtual scene. Motion tracking can then be used to update the virtual camera location and orientation continuously. But the illusion of the model projection in these AR applications is often disrupted when real objects in the scene are located in front of virtual projections, that are not getting occluded.

The presented thesis is focusing on this augmented reality problem and is comparing three occlusion mechanisms which can solve the virtual object occlusion with Project Tangos depth perception on mobile hardware and in real time. The idea of real world occlusion by the determined depth information was first introduced by Wloka and Anderson (1995). They used the z-buffer algorithm and a depth estimation with stereo cameras to prevent a rendering of virtual object parts which are occluded by real object depth information. The presented thesis is indicating three different approaches to fill the z-buffer with depth information captured by the infrared laser sensor, which is integrated in the Project Tango device. It is filled by the direct sensor data projection, by a TSDF based reconstruction called Chisel by Klingensmith et al. (2015) or by a self combined and implemented plane based reconstruction.

Project Tango is not producing a depth map which could be integrated into the z-buffer directly. It instead is giving a pointcloud with depth information of the current camera perspective. The first naive and already mentioned approach is the projection of the pointcloud to an image plane which then can be used as z-buffer. The depth sensor is limited to a range of $50cm$ to $4m$ and it also has issues capturing depth of complex or reflecting surfaces. Another limitation is the reception rate of only $5Hz$. Each of these issues is influencing the pointcloud projection due to noise and latency.

Breen et al. (1996) are mentioning that the z-buffer based occlusion can also be realized by rendering a primitive based reconstruction as a depth map. Therefore the second approach for solving the mentioned problems, is a plane based reconstruction which was developed during this thesis. It relies on the RANSAC plane estimation from pointclouds by Yang and Förstner (2010) and on the plane augmentation and plane range determination method which is used in a SLAM method by Trevor et al. (2012). In this approach all incoming points of the depth sensor get collected into an octree with a limited depth for spatial clustering. Then, the RANSAC algorithm is applied to all clusters with new points. It either augments existing planes in this cluster or creates new planes with a limit of three planes per cluster. The reach of each plane is calculated and triangulated by the convex hull.

The second reconstruction and third depth generating approach of this thesis is based on the real time reconstruction field of research. Unlike to offline reconstruction methods like the Poisson reconstruction by Kazhdan et al. (2006) or the approach of Hoppe et al. (1992), the challenge for real time reconstruction methods is the migration of continuous depth information from different perspectives into a single augmenting model. Klingensmith et al.

(2015) have presented a real time reconstruction method based on truncated signed distance functions (TSDF) focusing on the mobile application as CPU implementation called Chisel. In this TSDF approach the world is divided into voxel which contain the shortest distance to the next surface. Usually, this representation is rendered via raycasting on desktop environments like in KinectFusion by Newcombe et al. (2011). But Klingensmith et al. (2015) are using the marching cubes method to get a polygon based representation of the surface. They also integrate a spatial hash data structure presented by Nießner et al. (2013) to minimize the footprint of Chisel for mobile devices.

All theses depth map generating approaches producing errors because of noise and limitations in cluster or voxel sizes. Also the plane based reconstruction is producing gaps between planes, which lead to missing depth information in the depth map. Therefore, a guided filtering approach by He et al. (2010) is applied to the depth map. This filter can interpolate the depth according to the edges of the current color image frame captured by the Project Tango device.

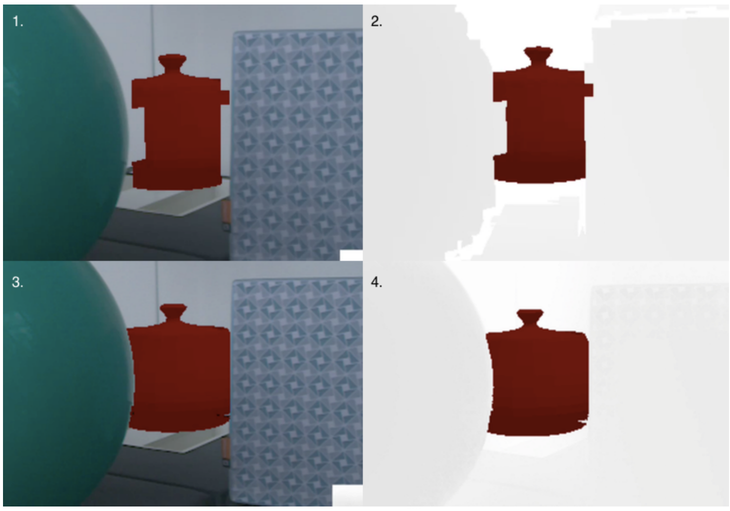

Figure 1: Protoype Screenshots - (1.) Occlusion by the Chisel Reconstruction (2.) Chisel Reconstruction Depth Map (3.) Guided Filter Result Occlusion (4.) Guided Filter Depth Map

During this thesis, all mentioned approaches where implemented or ported to the Project Tango development kit as a proof of concept. The final implemented prototype is containing all depth generating approaches and also the guided filter which can be combined and manipulated dynamically for an evaluation purpose. It is written in C++ and is using OpenGL and

OpenCV for the rendering and filtering. The final prototype is shown in figure 1 and is here rendering a scene which was used during the evaluation. Each combination was tested on this static and reproducible setup next to another more complex test setup to ensure the same input information for each tested approach. In addition, some not measurable dynamic tests have been performed to get an impression how the different approaches can be used in production.

All those mentioned six approaches (three depth map generating methods combined with the guided filter) can be used to achieve an augmented reality occlusion by real objects. The naive pointcloud projection has the already mentioned disadvantages of noise and the limited depth range because of sensor limitations. Noise can be successfully reduced by the guided filter which, however, is limited to $2 - 3Hz$ due to the OpenCV CPU image processing. Nevertheless, in combination the pointcloud projection could be used for more detailed but size constrained AR scenes.

The guided filter is also able to close the depth gaps of the developed plane based reconstruction. Although the plane based reconstruction was producing good results in the static tests, it is still rebuilding non-planar surfaces with rough plane approximations. This leads to bigger depth map errors in a more dynamic augmented reality scene where the camera position is not constrained. The cluster size also cannot be reduced, otherwise the RANSAC plane detection would produce statistically more false positives because of less measurement results inside each cluster.

Good results could be achieved by using the TSDF reconstruction Chisel as seen in figure 1. Although the voxel resolution was quite rough in this prototype, this reconstruction system could still be implemented on the GPU, whereby it would benefit on the parallel processing characteristics. By negotiating the reconstruction scale, voxel resolution and performance, the voxel resolution could be still reduced to the limits of the depth sensor.

The guided filter was always able to improve the quality of the real world occlusion in the static scenes. However, some artifacts where observed during the dynamic testing. When an edge in the color frame was just painted on a flat real world surface, which produced also a flat depth map, the filter was alternating the depth map with some artifacts caused by the underlying color structure. This should be investigated next to alternative depth map upsampling methods for mobile usage in future work, like the approach of Ferstl et al. (2013). Another future idea could be the integration of the guided filter into the OpenGL fragmentshader. This would make the OpenCV binding superfluous and could save the conversion time. The filter would also benefit from the parallel GPU computing characteristics implemented in OpenGL and would run much faster than in this prototype realization.

Since Lenovo is announcing a cooperation[1] with Google, Project Tango promises to get the standardization for motion tracking, depth perception and area learning on mobile devices. Lenovo will publish the first consumer hardware in summer 2016 with Google's Project Tango hard and software. Intel is cooperating[2] with Google as well and enables his competitive product RealSense$^{\text{TM}}$ to run Project Tango applications.

Bibliography

Breen, D. E., Whitaker, R. T., Rose, E., and Tuceryan, M. (1996). Interactive occlusion and automatic object placement for augmented reality. In *Computer Graphics Forum*, volume 15, pages 11–22. Wiley Online Library.

Ferstl, D., Reinbacher, C., Ranftl, R., Ruether, M., and Bischof, H. (2013). Image guided depth upsampling using anisotropic total generalized variation. In *The IEEE International Conference on Computer Vision (ICCV)*.

He, K., Sun, J., and Tang, X. (2010). Guided image filtering. In *Computer Vision–ECCV 2010*, pages 1–14. Springer.

Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., and Stuetzle, W. (1992). *Surface reconstruction from unorganized points*, volume 26. ACM.

Kazhdan, M., Bolitho, M., and Hoppe, H. (2006). Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7.

[1] Lenovo News - http://goo.gl/jFLNyn (21.03.16)

[2] Intel®RealSense$^{\text{TM}}$ Developer Kits - http://goo.gl/j4Y18A (21.03.16)

Klingensmith, M., Dryanovski, I., Srinivasa, S., and Xiao, J. (2015). Chisel: Real time large scale 3d reconstruction onboard a mobile device. In *Robotics Science and Systems 2015*.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE.

Nießner, M., Zollhöfer, M., Izadi, S., and Stamminger, M. (2013). Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 32(6):169.

Trevor, A. J., Rogers III, J. G., Christensen, H., et al. (2012). Planar surface slam with 3d and 2d sensors. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3041–3048. IEEE.

Wloka, M. M. and Anderson, B. G. (1995). Resolving occlusion in augmented reality. In *Proceedings of the 1995 symposium on Interactive 3D graphics*, pages 5–12. ACM.

Yang, M. Y. and Förstner, W. (2010). Plane detection in point cloud data. In *Proceedings of the 2nd int conf on machine control guidance, Bonn*, volume 1, pages 95–104.