Improving the Kinect by Cross-Modal Stereo

Wei-Chen Chiu
walon@mpi-inf.mpg.de
Ulf Blanke
blanke@mpi-inf.mpg.de
Mario Fritz
mfritz@mpi-inf.mpg.de

Max-Planck-Institute for Informatics
Saarbrücken, Germany

Abstract

The introduction of the Microsoft Kinect Sensors has stirred significant interest in the robotics community. While originally developed as a gaming interface, a high quality depth sensor and affordable price have made it a popular choice for robotic perception.

Its active sensing strategy is very well suited to produce robust and high-frame rate depth maps for human pose estimation. But the shift to the robotics domain surfaced applications under a wider set of operation condition it wasn't originally designed for. We see the sensor fail completely on transparent and specular surfaces which are very common to every day household objects. As these items are of great interest in home robotics and assistive technologies, we have investigated methods to reduce and sometimes even eliminate these effects without any modification of the hardware.

In particular, we complement the depth estimate within the Kinect by a cross-modal stereo path that we obtain from disparity matching between the included IR and RGB sensor of the Kinect. We investigate how the RGB channels can be combined optimally in order to mimic the image response of the IR sensor by an early fusion scheme of weighted channels as well as a late fusion scheme that computes stereo matches between the different channels independently. We show a strong improvement in the reliability of the depth estimate as well as improved performance on a object segmentation task in a table top scenario.

Method

Transparent and specular phenomena have been proven notoriously hard to capture [1], in particular in unconstrained scenarios where prior information about lighting and geometry of the scene can rarely be assumed.

The main focus in this paper is to explore cross-modal stereo so that we can have an active and passive depth sensing path, similar to [2, 3], but in a single sensor unit and that is less sensitive to transparent and specular phenomena.

Early integration We look at several baselines using naive combinations of color channels and the IR channel. Furthermore, we develop a method for estimating optimal weighted combination of the RGB channels to create an IR-like image for improved stereo matching:

Given a IR image I^{ir} and a RGB image $(I_r^{rgb}, I_g^{rgb}, x_b^{rgb})$, we would like to obtain a weighting $w = (w_r, w_g, w_b)$ of the channels such that the converted RGB image is more similar to IR image and has more corresponding points during the stereo matching. We evaluate the performance of stereo matching by calculating the number of found correspondences: $num\_of\_stereo\_match(I^{rgb}, I^{ir})$. The optimization reads:

$$\begin{aligned} \max_{w_r,w_g,w_b} \quad & num\_of\_stereo\_match(w_r * I_r^{rgb} + w_g * I_g^{rgb} + w_b * I_b^{rgb}, I^{ir}) \\ \text{subject to} \quad & w_r + w_g + w_b = 1. \end{aligned} \tag{1}$$

Figure 2 shows the disparity map from converted RGB image with learnt weights comparing to the result from original RGB image in gray level. More details are preserved after applying the learnt weighting.

Late integration We also investigate a late integration scheme where we delay the combination of the different color channels and compute stereo correspondences w.r.t. the IR image independently. We then fuse the resulting depth estimate with depth sensed by the Kinect by union of the point clouds.

Experiments

In order to evaluate our approach, we utilize an object segmentation task operating on the estimated point clouds. A new dataset has been collected in order to test the Kinect sensor on more challenging scenarios. The best result of the proposed fusion schemes achieves an average precision of 76.6%. Comparing to the average precision of 48.8% of the built-in Kinect depth estimate, we achieve a significant improvement of nearly 30%. See Fig. 1(e) and 1(f) for illustrations.

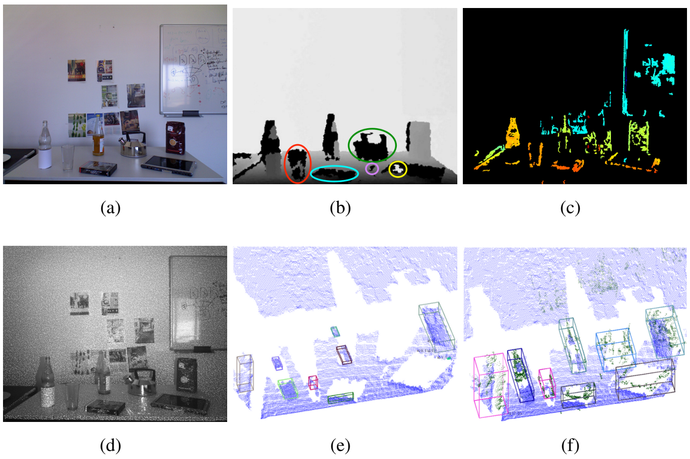

Figure 1: (a)RGB image. (b)Depth map from Kinect with some failure cases highlighted. (c)Disparity map computed from the stereo pair of (a)RGB and (d)IR image. (e)Detection result on Kinect point cloud only. (f)Detection result on proposed fused estimate.

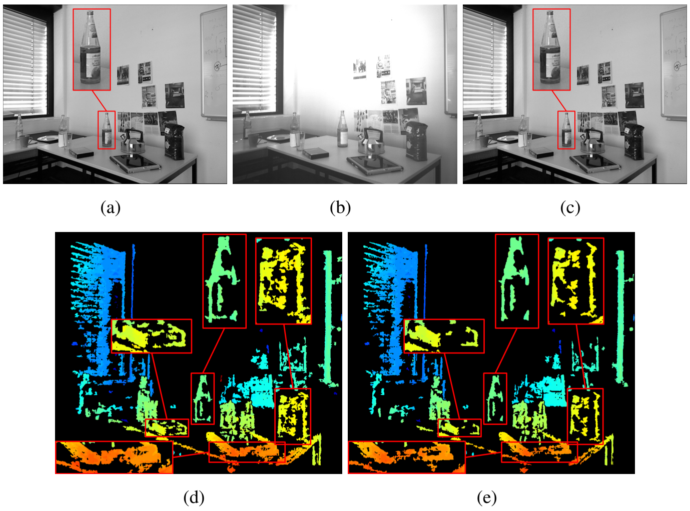

Figure 2: (a)RGB image converted according to the optimized color channel weights. (b)IR image with covered. (c)RGB image in gray scale (averaged channels with equal weight).(d)Disparity map from (a) and (b). (e)Disparity map from (c) and (b).

References

[1] Ivo Ihrke, Kiriakos N. Kutulakos, Hendrik P. A. Lensch, Marcus Magnor, and Wolfgang Heidrich. State of the art in transparent and specular object reconstruction. In *STAR Proceedings of Eurographics*, 2008.

[2] R.G. Yang J.J. Zhu, L. Wang and J.E. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *CVPR*, 2008.

[3] 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 *3DIM09*, 2009.