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.

1 INTRODUCTION

Future mobile robotics rely heavily on robust sensing schemes in order to bring the success of industrial robotic applications in controlled environments to the unstructured and everyday changing scenario in our homes. 3D Perception has been one of the key technologies to provide a rich capture of indoor scenes that facilitates data driven segmentation, grasp planning, and much more. While the steadily improving sensing technology has provided us with more accurate and reliable data, we haven't — and probably will not — see a single sensor performing well across every conceivable condition. This calls for robust integration of multiple sensing schemes to complement for each others' short comings.

With the introduction of the Microsoft Kinect sensor, a highly performant yet low cost 3D sensor was made available that rivals much more costly solutions available to robotics. And with one million units sold in the first week and ten million units to date, it is probably

BMVC 2011 http://dx.doi.org/10.5244/C.25.116

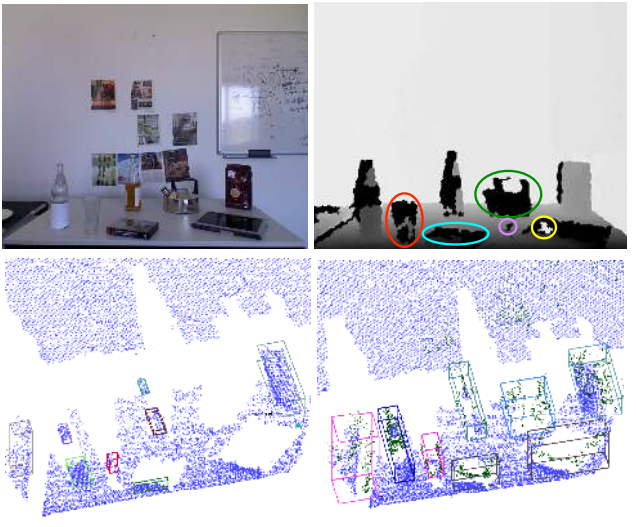


Figure 1: (top) Failure cases of Kinect 3D sensor: (red) transparency, (green) specularity, (cyan) dark objects under flat viewing angle, (yellow) reflections of the projected dot patterns, (violet) interference of dot patterns by reflections. (bottom) We evaluate on a object segmentation task, left: result on Kinect point cloud only, right: strongly improved result on our fused estimate proposed in this paper. (blue points: kinect; green points: cross-modal stereo)

one of the most distributed as well. Despite being originally designed as a gaming interface, soon after its release there was strong interest from hobbyists over enthusiasts to robotics researchers trying to stretch the envelope of possible application scenarios beyond its original intended use case.

In this paper we are particularly interested in the robotics scenario and resulting shortcomings when using the Kinect in those settings. In home environments object properties differ beyond the well behaved properties of cloth and skin where the Kinect depth estimation performs admirably well. We realize that in particular specular, transparent and reflective objects cause serious problems — and not rarely lead to a complete failure (See Fig. 1). Yet objects like glasses, bottles, tea kettles, pans are objects of daily living and therefore they are in the core set of objects home robotics and assisted living want to address [2].

Our main contribution is a more reliable depth estimation scheme using an off-the-shelf Kinect sensor without modifying hardware nor requiring any additional sensors. We complement the built-in active depth sensing scheme with a passive stereo path which we establish between the RGB and IR sensor. We investigate several fusion schemes to provide more reliable sensor data, improving drastically in particular on transparent and specular objects. In order to approximate IR and RGB image information and preserve similarities between modalities, we learn a channel combination. Besides qualitative improvements we provide empirical evidence for strong improvement on a data-driven object detection task in a table top scenario. Code and the database will be made available at the time of publication to further stimulate research that pushes the envelope on this interesting sensor on scenarios relevant to real-world applications.

2 RELATED WORK

Transparent and specular phenomena have been proven notoriously hard to capture [9], in particular in unconstrained scenarios where prior information about lighting and geometry of the scene can rarely be assumed. Only very recently some initial success towards practical systems for detecting transparent objects has been reported on visual object detection tasks [6] and multi-view lidar based object detection [12]. [6] learns object models for glasses and [12] improves transparent object detection by integrating two sensors of the same type. In contrast, our work uses a single off-the-shelf unit combining an active and a passive approach and we show improved results on wide range of effects like transparency, specularity and highly absorbent surfaces as they occur on many household objects such as tea kettles, mirrors, displays, bottles.

But also stereo algorithms are effected by more complex surface properties. E.g. [15] provides an analysis of such effects and presents a sophisticated model for recovering multi-layered scene structure. In practice, we see stereo correspondences still being preserved at least on borders of objects with complex surface properties. Therefore we use a simple, computationally efficient block matching algorithm as implemented in OpenCV [1].

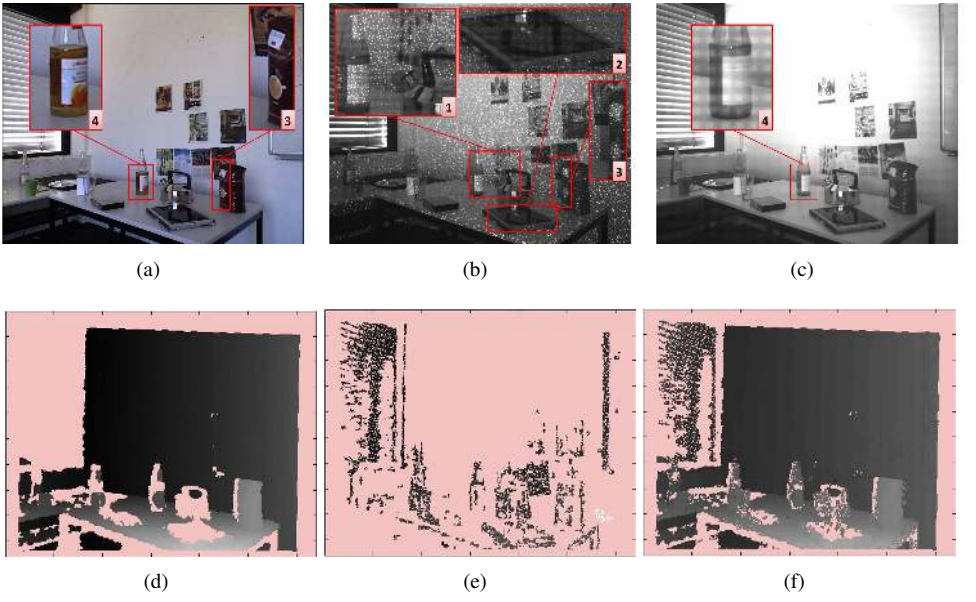
As we seek to find correspondences between a RGB and IR sensor, we face a problem of different data domains. Most recently, related problems have been successfully addressed in a metric learning formulation for visual category recognition from different data sources like images from the web, DSLRs and webcams [14]. While this approach is based on Information Theoretic Metric Learning (ITML) [4] much simpler formulations based on large-margin classifiers [3] have been proposed from which we drew some inspiration. However, the latter approach is only applicable for classification while we learn a transformation that is directly applicable to the image without any change to the stereo algorithm.

Previous work investigated fusion techniques for depth measurements originating from time of flight cameras and stereo cameras [10, 11]. In contrast, our main focus in this paper is to explore cross-modal stereo so that we can have an active and passive depth sensing path in a single sensor unit. Therefore the previous investigations are orthogonal to ours. Furthermore, our focus is on a object segmentation task with an emphasis on problematic cases containing specular and reflective surfaces.

In our evaluation we use an object segmentation scheme based on a support plane assumption which is common to many recent systems using 3D information, e.g. [7, 8, 13].

3 METHODS

As mentioned previously the Kinect depth estimate fails on specular, transparent or reflective surfaces. Depth is calculated from an IR-pattern that is projected from by the Kinect sensor unit (Fig. 2(b)). On reflective objects however the pattern is not visible or being reflected, causing holes in the depthmaps or potential interferences (Fig. 2(d)). In Contrast, stereo vision enables to detect disparities at edges of transparent or reflective objects, but has difficulties finding correspondences on textureless areas, such as the wall or the desk (Fig. 2(e)). Since the Kinect features two cameras (IR and RGB) we propose a cross-modal stereo approach that we combine with the built-in 3D estimate of the Kinect in order to compensate for the problems of the individual sensors. In the following we describe our method that can be run on an off-the-shelf Kinect sensor without any hardware modifications.



(a) (b) (c)

(d) (e) (f)

Figure 2: (a) *RGB image*. (b) *IR image with Kinect's IR-projector pattern*. 1. pattern not visible on reflective surfaces, 2. reflected pattern on display, 3. Shadow from projector spot differs to RGB lighting condition (c) *IR image with covered IR-projector.* Red texture invisible to IR-sensor (d) *Rectified depth map from Kinect*. (e) *Disparity map from stereo* and (f) *Fused depth maps*.

3.1 Stereo and alignment to Kinect depth

We first briefly describe our stereo calibration algorithm for the IR and RGB camera and the fusion step with Kinect depth information. Then, we describe the alignment procedure that enables the combination of stereo and Kinect depth information by simple union operator on the two point clouds. Since we have to deal with two different data domains, we introduce an optimization step in order to obtain improved stereo correspondences.

Stereo calibration Given the images from IR and RGB sensors, the extrinsic and intrinsic parameters can be computed with standard stereo calibration technique. Once the calibration parameters of IR and RGB cameras are obtained, we apply Bougeuet's algorithm to do the rectification toward the stereo image pairs and make them row-aligned. Then we utilize the SAD (Sum of Absolute Difference) block matching to find corresponding pixels between IR and RGB images. The offset between such pixels marks the disparity in image coordinates. As a result, we obtain the disparity maps for every pair of IR and RGB images, for which depth can be calculated as shown in Fig. 2(e).

Fusing Stereo depth with Kinect depth In order to combine the disparity map from our stereo setting and the depth map from the Kinect, we need to align the image planes. Since the disparity map from the stereo is rectified, we apply the same rectification to the depth map from the Kinect as shown in Figure 2(d).

After converting the disparity from stereo into depth measurements, we can directly compare depth values obtained from stereo and the Kinect. From a set of calibration scenes we obtain scaling and offset parameters that align the depth values.We use least squares to estimate these parameters. Figure 2(f) shows an example of the aligned depth measurements.

In order to evaluate our depth estimate on an object detection task, we generate a point cloud by means of the reprojection matrix obtained from stereo calibration. The fusion of stereo and Kinect depth is carried out in 3D by simply taking the union of the point clouds as displayed in Figure 1.

3.2 Cross-Modal Adaptation for IR-RGB-Stereo

Early integration As described in the previous section we search for stereo correspondences in the ir-rgb-image pairs. The channels of those images correspond to different sensor characteristics that are only receptive over a range of wavelengths. Due to correlations in the sensor and material characteristics running stereo across the modalities is expected to produce at least some correspondences. Yet, it seems more appropriate to find a better combination of the channels that would make the two signals more similar. We do so by employing a global optimization approach.

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 IR image and has more corresponding points during the stereo matching. We evaluate the performance of stereo matching by simply calculating the number of corresponding points we can find, which we denote by: $num\_of\_stereo\_match(I^{rgb}, I^{ir})$. The resulting optimization problem reads:

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

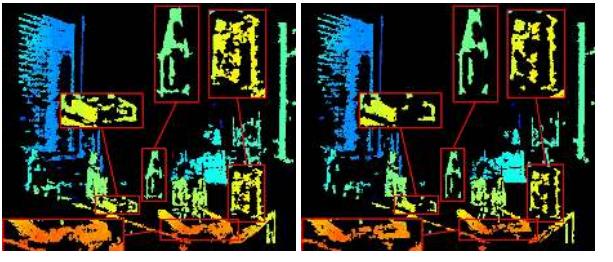
We use the IR image with covered IR-projector to avoid the effect from the projected pattern. This optimization problem is solved by grid search with uniformly sampling of the plane $w_r + w_g + w_b = 1$. Figure 3 shows the disparity map from the RGB image converted with learnt channel weights and compared to the disparity from original RGB image in gray level. We observe that 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 proceed as with the 3D data from the Kinect and fuse the results in 3D space by forming the union over the point clouds.

3.3 Point cloud based object segmentation

In order to quantify the improved depth information for potential detection tasks, we implemented a simple object detection system based on point clouds from Sec. 3.1. Since objects of interest are often located on tables or fixed at walls, we first segment space into a support surface and a background surface. We then cluster the residual point cloud, which is neither part of the table's nor the wall's surface, into potential objects.

(a) (b) (c)



(d) (e)

Figure 3: (a) *RGB image converted according to the optimized color channel weights,* $(w_r, w_g, w_b) = (0.368421, 0.473684, 0.157895)$. (b) *IR image with covered IR-projector.* (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).*

Support surface extraction We apply an iterative RANSAC-algorithm to extract the support surface and the background wall from the point cloud. We assume that the scene contains two surfaces: the background wall and the table, where objects of interest can be spread out. The residual points, which are neither inliers of the background wall and nor the table surface, are then clustered into potential objects.

Point cloud clustering We first partition the residual pointcloud from the step above using kmeans-clustering. We set the number of centers to $K = 850$. This effectively reduces the complexity for further calculation. The kmeans-centers are then further grouped by agglomerative clustering. Based on grouped kmeans-centers we calculate a 3D-bounding box as in Fig. 1. For the evaluation in the following section, we backproject 3D-coordinates of each group member into the image coordinates using the transformation from Sec. 3.1. Upon image pixels of each group a rectangular bounding box is fitted and compare to the ground truth annotations of objects. In order to score each bounding box for precision-recall analysis we use the number of points associated with each box, respectively group, as a score.

4 EXPERIMENTS

We evaluate the success of our approach on a new database that we have collected in order to test the kinect sensor on more challenging scenarios. The dataset consists of 106 objects in 19 images. All objects are annotated with 2D bounding boxes. We follow the evaluation criterion of the Pascal challenge [5] and compute precision-recall curves based on the overlap

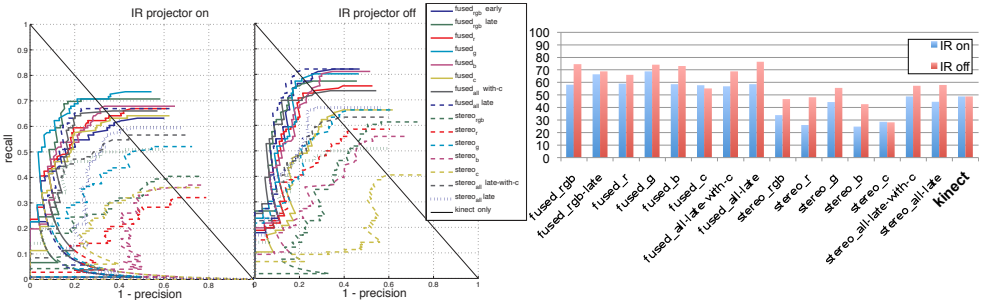


Figure 4: Precision Recall and average precision for the table-dataset

criterion $a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$, where B_p is the predicted bounding box and B_{gt} the ground truth bounding box and where a_0 must exceed 50%.

We evaluate several fusion schemes under different conditions, and settings for the cross-modal stereo matching. Following standard stereo matching, we average RGB into a grayscale image (stereo only: $stereo_{rgb}$ and fused with kinect depth: $fused_{rgb}\ early$). We also combine each channel individually (stereo only: $stereo_{\{r,g,b\}}$ and fused with kinect depth: $fused_{\{r,g,b\}}$). Given a disparity map for each channel invididually, we also combine these into a late fusion scheme with the kinect depthmap ($fused_{rgb}\ late$). For a weighted combination of the R, G and B-channel as presented in Sec 3.2 we denote the index c.

4.1 Results

Table 4 shows average precision for different fusion schemes. We expect a strong influence of the IR projector on the correspondence matching of the stereo vision. Therefore, we captured stereo pairs under two conditions, first by covering the emitting IR projector and second under normal operating condition with the IR projector switched on. Note, that since the Kinect depth estimate does not operate without the IR projector, we captured the images consecutively in the first setting.

The built-in Kinect depth estimate achieves 48.8% of average precision. Combining stereo with the Kinect depth results in significant improvement of nearly 30%. Overall the maximum average precision of 76.6% is achieved by fusing all channel-specific depth maps (*fusion_rgb_late*). When turning the Kinect into normal operation mode that is with emitting IR projector, overall performance of different combination schemes drecreases, but still improves the Kinect depth about 10-20%. Interestingly, the best result (68.8%) is achieved by fusing Kinect depth and stereo depth from the green channel, which is closely followed by fusion with channel-specific depth estimates (*fusion_rgb_late:* 66.5%). As expected in this scenario, stereo only with projected IR pattern, performs worse in all different combination schemes than Kinect only.

Fig. 5 shows example depth maps and detections using the pointcloud segmentation from Sec. 3.3. Fig. 5 (a) shows a comparison between the Kinect-only depthmaps and *fusion_all_late* scheme. Based on kinect depth estimation, we observe that nearly all transparent or reflective objects are either missed or over-segmented by their opaque parts (e.g. the bottle labels). Also interferences occur, e.g., the reflected IR-pattern from the wall results in false depth on tablet's display.

Fig. 5 (b) shows a comparison of fused depth maps with operating IR projector (middle)

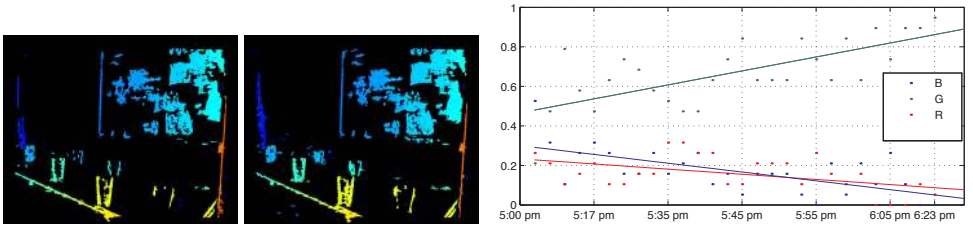
Figure 5: Example images from dataset. (a) Top: RGB image, Middle: Kinect only, Bottom *fusion_all_late*, (b) Top: RGB images, Middle: *fusion_all_late* with covered IR projector, Bottom: *fusion_all_late*

and with covered IR projector (bottom). It can be seen that the object segmentation merges individual objects into one object only with operating IR projector. Besides the interfering IR pattern on the object's edges, lighting conditions differ. While the IR projector behaves similar to a spotlight and causes hard shadows around objects in a scene, the RGB images are effected by environmental illumination only. As a results shadows differ between IR and RGB images (see Fig. 2(a) and 2(b), box 3). This leads to increased smearing effects and which leads to point cloud connections between nearby objects. Here a more sophisticated segmentation approach or statistical outlier removal techniques can remedy this effect. The left-most depth maps show a rather pathologic case, where the Kinect is directed to a mirror. The emitted IR pattern is reflected back to the camera causing a glare. Only when switching off the IR-pattern depth is revealed by stereo vision.

4.2 Discussion

We can see that stereo vision across modalities is feasible and improves object detection based on the Kinect's depth estimation up to 30% without any modification of the hardware. Overall, stereo matching between the ir channel and each color channel individually combined with 3D from the Kinect performs best in all considered settings. Using depth maps based on the green channel performs surprisingly well. Since the red and infrared are close in wavelength, the infrared camera has a similar sensitivity. We intuitively expected a good correspondence matching for this channel. In fact however, red textures on white background do not contrast, and red becomes invisible. This effect can be seen on the bottle label in Fig. 2(a) and 2(c). Green texture however is preserved and represented by low intensities in the IR channel. Then, stronger gradients facilitate the correspondence matching.

Learning a weighted RGB-channel-combination scheme to obtain an "IR-like" image, turns out to be highly sensitive to environmental change. Colors of objects or varying environmental IR exposure influences the choice of weights significantly. Fig. 6(c) shows a series of captures during varying daylight conditions and corresponding optimized weights. Although we could find more correspondences using a weighted scheme Fig. 6(b) compared



(a) (b) (c)

Figure 6: (a) Disparity from grayscale RGB and IR, (b) Disparity from weighted RGB and IR, and (c) Learned weights for obtaining IR like image from RGB under different lighting conditions (Date of capture 31. March, sunset 7:56pm).

to grayscale RGB only Fig. 6(a), overall segmentation results did not reflect the improvement. Since, we estimated the parameters on a training set, they did not generalize well to our detection dataset. Dynamic weight adaption based on image statistics, such as illumination and white balance, might lead to improvement, which is subject to further investigation.

Practical issues Kinect does not allow simultaneous grabbing of the RGB and IR stream. The OpenNI framework, as well as libfreenect offer functionality of switching the streams programmatically (and asynchronously). The speed depends on the buffer writing speed. When switching too fast the buffer is not entirely written. We did initial stress tests to tune the framerate. We observe that libfreenect shows faster performance and yields about 1.5-2fps for taking an IR and RGB pair, which leaves space for estimating stereo disparity maps and provides reasonable update rates for many robotics applications. The OpenNI framework achieves far less than 1fps.

5 CONCLUSIONS

We presented a simple and effective cross-modal stereo vision approach for combination with Kinect depth estimates, which can be applied without any further hardware requirements. We provide empirical evidence for drastic improvement to the Kinect 3D sensing capabilities.

We presented a cross-modal adaptation scheme that allows for improved correspondence matching between RGB and IR cameras and show general feasibility of their combination. The value of our improved 3D sensing scheme is validated by a generic, data-driven object detection task. Our combination method produces depthmaps that include sufficient evidence for reflective and transparent objects, and preserves at the same time textureless objects, such as tables or a walls.

We expect this work to have a high impact in the robotics community due to the wide spread use of Kinect sensors and the ubiquitous problem of capturing transparent objects for detection, recognition and manipulation. In the future we plan to investigate further cross-modal adaption schemes and also to include more sophisticated fusion schemes.

References

[1] Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[2] Y.S. Choi, T. Deyle, T. Chen, J.D. Glass, and C.C. Kemp. A list of household objects for robotic retrieval prioritized by people with als. In *ICORR*. IEEE, 2009.

[3] Hal Daumé, III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, 2010.

[4] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[5] M. Everingham and J. Winn. The PASCAL visual object classes challenge development kit (VOC 2007). Technical report, University of Leeds, 2007.

[6] Mario Fritz, Michael Black, Gary Bradski, Sergey Karayev, and Trevor Darrell. An additive latent feature model for transparent object recognition. In *NIPS*, 2009.

[7] Mario Fritz, Kate Saenko, and Trevor Darrell. Size matters: Metric visual search constraints from monocular metadata. In *NIPS*, 2010.

[8] Stephen Gould, Paul Baumstarck, Morgan Quigley, Andrew Y. Ng, and Daphne Koller. Integrating visual and range data for robotic object detection. In *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, 2008.

[9] 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.

[10] 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.

[11] 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.

[12] Ulrich Klank, Daniel Carton, and Michael Beetz. Transparent object detection and reconstruction on a mobile platform. In *IEEE International Conference on Robotics and Automation (ICRA).*, 2011.

[13] Zoltan-Csaba Marton, Radu Bogdan Rusu, Dominik Jain, Ulrich Klank, and Michael Beetz. Probabilistic categorization of kitchen objects in table settings with a composite sensor. In *Proceedings of the IEEE/RSJ international conference on Intelligent robots and systems*, 2009.

[14] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[15] Yanghai Tsin, Sing Bing Kang, and Richard Szeliski. Stereo matching with linear superposition of layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:290–301, 2006.