Explicit Occlusion Reasoning for 3D Object Detection

David Meger[1]
dpmeger@cs.ubc.ca
Christian Wojek[2]
cwojek@mpi-inf.mpg.de
Bernt Schiele[2]
schiele@mpi-inf.mpg.de
James J. Little[1]
little@cs.ubc.ca

[1] Laboratory for Computational Intelligence
University of British Columbia
Vancouver, Canada

[2] Computer Vision and Multimodal Computing
Max-Planck Institut fur Informatik
Saarbruecken, Germany

Abstract

This paper presents a technique to locate objects in 3D that adapts visual appearance models using explicit visibility analysis. We formulate a Bayesian model for 3D object likelihood based on visual appearance, 3D geometry such as that available from RGB-depth sensors, and structure-from-motion. Learned visual appearance templates for portions of an object allow for strong discrimination even under occlusion. We describe an efficient inference procedure based on data-driven sampling with geometric refinement. Our 3D object detection technique is demonstrated on the publicly available robot-collected UBC Visual Robot Survey dataset, as well as with data from the Microsoft Kinect. Results show that our method improves robustness to occlusion when compared to a state-of-the-art visual category detector.

1 Introduction

Object recognition is a key competency for intelligent systems, as it links sensors to semantic concepts. However, typical operating environments, such as kitchens in homes, are often cluttered to the point where even humans struggle to find what they are looking for (e.g. the lost keys scenario). This paper describes a system for locating objects in cluttered indoor environments using the sensor sequence available from a moving intelligent system. To overcome the occlusion problem, we use sensed 3D geometry to compute the expected visibility of objects and compare learned partial-object appearance templates against the visible regions. Structure-from-motion (SfM) is used to register collected views, which allows probabilistic data fusion, and increases robustness. As demonstrated by the sample result in figure 1, the method can be reliably locate objects 3D, even in significant clutter and occlusion.

Consider the problem of recognizing an object that is partially occluded in an image. The visible portions are likely to match learned appearance models for the object, but hidden portions will not. This is a primary cause of poor recognition performance for modern

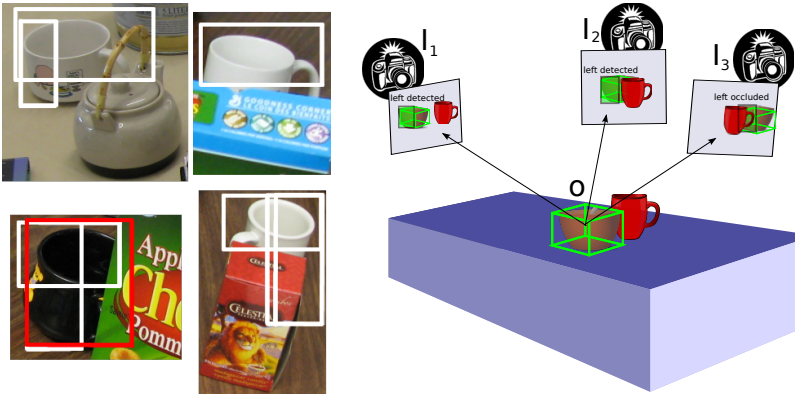
Figure 1: Sample results on two real home scenes (left) and a synthetic lab scene (right). 3D wireframes indicate mugs (red) and bowls (green). Thresholded at 90% precision. All figures are best viewed in colour.

approaches. The (hypothetical) ideal system would consider *only* the visible object information, correctly ignoring all occluded regions. In purely 2D recognition, this requires inferring the occlusion present, which is a significant challenge since the number of possible occlusion masks is, in principle, exponential. We simplify the problem, considering only a small subset of the most likely occlusions (top, bottom, left, and right halves) and noting that some mis-match is tolerable. We train partial-object detectors tailored exactly to each of these few cases. In addition, we reason about objects in 3D and incorporate sensed geometry, as from an RGB-depth camera, along with visual imagery. This allows explicit occlusion masks to be constructed for each object hypothesis. The masks specify how much to trust each partial template, based on their overlap with visible object regions. This comes close to the intuition - only the visible evidence contributes to our object reasoning.

We will continue by discussing related work. Section 2 presents a probabilistic model to explain 3D object presence. Section 3 describes an efficient inference procedure to optimize this model. Section 4 describes our datasets and experimental procedures and results are discussed in section 5.

Previous Work

Our technique fuses information from numerous viewpoints and performs inference of 3D object locations. Many of the individual components have been previously studied. For example, for urban scenes (i.e. car-based cameras). Ess *et al.* [1] utilized stereo as well as SfM. Wojek *et al.* [22] perform 3D reasoning for objects tracked over several frames that have been registered using SfM. Helmer *et al.* [13] and Meger *et al.* [15] use registered visual information from many views to improve performance for indoor scenes, and [3] applies a similar approach for sports video. The most similar previous approaches are pedestrian detection methods, such as [5] and Wojek *et al.* [23], which inspire our mixture-of-experts formulation to combine partial object detectors. The primary difference is that we utilize sensed depth information, such as from an RGB-depth camera, while [5] uses motion discontinuities to pre-segment regions and [23] utilizes inter-object reasoning that requires all occluders to be detected with an appearance model.



(a) (b)

Figure 2: (a) Partial detectors (shown in white) often respond when full-object models (shown as red boxes) are missed due to occlusion. (b) An object is projected and associated with partial detections where available.

A number of authors combine imagery and 3D information from a single view. Sun *et al.* developed a depth-aware Hough Transform [18]. Both Lai *et al.* [14] and Quigly *et al.* [17] extract features from both visual and depth imagery for robotic recognition. While indoor objects are considered in these approaches, the authors do not focus on occlusion. Several authors have also considered using depth to reduce the set of scales to be searched at each image pixel [10] [12].

Occlusion reasoning has also been considered based on single images. Desai *et al.* [4] reasoned about the depth-ordering of a set of detections to achieve state-of-the-art performance on pixel-level segmentation. Vedaldi *et al.* [20] and Wang *et al.* [21] consider occlusion during object detection. Both of these methods have motivated us to study occlusion in a 3D context and their improved image-space performance is complementary to our approach.

2 Occlusion-aware 3D Object Model

In this section, we describe our model for the likelihood that an object exists at a given 3D location, given the available visual imagery, sensed depth information and SfM information between several viewpoints. Figure 2 demonstrates our system's view of a scene. Each candidate 3D object location projects into all views and is associated to image-space detections produced by visual category recognizers for the object's complete appearance as well as for a subset of the possible occlusions. Fully visible objects are likely to align well with strong detections in each image. However, occlusion can cause weak detection results, since there is little matching appearance evidence. If left un-modeled, this would lead to the system rejecting the 3D object candidate. Therefore, we use sensed depth information to estimate the occlusion of each part of the object in each view. The occlusion estimate is incorporated into the scene score, allowing our system to ignore meaningless appearance information from occluded regions and more faithfully representing the underlying geometry.

This section explains a model to compute the likelihood of any proposed 3D object, but does not consider how these objects should be proposed. That is left for following section, which outlines our sampling-based inference procedure.

2.1 Top-level Object Likelihood

We model the likelihood of an object o, which has 3D position and scale, given a stream of observed data gathered over time and space, as the system moves. We assume a synchronized image, I_t, and point cloud, C_t, are available. Also, we express registration information obtained from structure-from-motion as a projection matrix, P_t, from a common global coordinate frame into the camera's coordinate frame in view t. Hence, the observed data from each view is $Z_t = \{I_t, C_t, P_t\}$. We use super-script notation to denote the sequence of data from the beginning of time until the present: $Z^t = \{Z_1, Z_2, ...Z_t\}$. The likelihood of an object given the available data is expressed using Bayes rule and by making the Naive Bayes assumption to achieve independence between viewpoints:

$$p(o|Z^t) \propto p(o)p(Z^t|o) \approx p(o)\prod_t p(Z_t|o) \quad (1)$$

$$= p(o)\prod_t p(I_t, C_t, P_t|o) \quad (2)$$

$$= \underbrace{p(o)}_{object\ prior} \prod_t \underbrace{p(I_t|o, C_t, P_t)}_{appearance} \underbrace{p(C_t|o, P_t)}_{geometry} \underbrace{p(P_t|o)}_{registration} \quad (3)$$

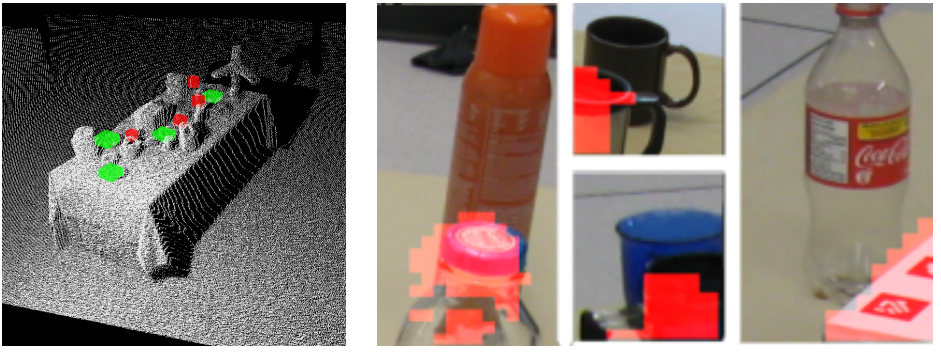
Note that the conditional independence structure between data types has been used to factor the distribution. We use a delta function centered at the SfM estimate as the registration model, as this was sufficiently accurate in our experiments. In future work, this could easily be replaced by a probabilistic model for registration error between views. $p(o)$ represents a size prior per object category. We model this as a normal distribution on both the height and radius of the object, which is appropriate given the cylindrical nature of the objects studied here. Other shape priors can easily be substituted. We will continue by describing the geometry and appearance likelihood terms for each viewpoint in detail.

2.2 Geometry Model

We model the likelihood of the observed depth data given an inferred object location. As shown in figure 3(a), inferred 3D object regions are placed in the same coordinate frame as measured 3D data. This allows us to reason about the agreement between the observed environment geometry, and our hypothesized object. For each pixel in the depth image within the projected object region, we note three discrete outcomes:

1. The measured depth is near to the inferred depth: they agree
2. The measured depth is greater than the inferred depth, which indicates the inferred object region is unoccupied: they conflict
3. The measured depth is less than the inferred depth: the object is occluded at this pixel

The geometry term in equation (3) is constructed from pixels that fall into the first and second outcome only, as occluded regions cannot tell us anything about the object's geometry. We employ a mixture of two Gaussians to model the expected error in the depth sensor and the rare occurrence of outliers far from the expected value. We compute the product of this model over all pixels expected to fall on the object. This model is common in geometric inference, and has been used previously in robotic mapping, for example in [19].



(a) (b)

Figure 3: Point clouds and inferred 3D objects, as shown in (a), allow computation of occlusion masks for each object in each image, as in (b). Occluded regions shown in red.

Figure 3(b) shows pixels marked with the third outcome above: occluded. The ratio of occluded pixels within the region considered by each partial-object appearance template forms a visibility score v used for the appearance model, as will be described in the next section.

2.3 Appearance Model

The likelihood of the image appearance given an object is evaluated using a learned model for the entire object, as well as for a number of sub-parts. We scan the image with each learned model to obtain a set of object detections: D_t. The 3D object, o, is projected into each image and assigned to nearby detections (for the whole, and the visible portions) greedily. Let $d_{it}(o) \in D_t$ represent the detection for object part i assigned to object o at time t. We use the mixture-of-experts model proposed in [23] to express the contribution of each partial detector (expert) weighted by the visibility in the corresponding region. We also enforce soft geometric consistency by penalizing error in alignment between the object and an associated detction. This is written symbolically:

$$p(I_t|o,C_t,P_t) = \frac{1}{\sum_i v_{it}\delta(v_{it}>\theta)} \sum_i v_{it}\delta(v_{it}>\theta)\Psi_s(d_{it}(o))\Psi_g(P_t\cdot o, d_{it}(o)) \quad (4)$$

Recall that visibility, v_{it}, is derived from the sensed depth within the region explained by the object (or object part). δ is an indicator function to completely discount contributions of parts that are more occluded than a hand-chosen threshold, θ. Ψ_s is a potential function related to the detector's score. We have implemented Ψ_s here with a linear mapping of detector score to the range $[-1,1]$ based on scores on a set of validation data. Platt Scaling, as in [16], is an alternative that provides a meaningful probabilistic interpretation, but gave no benefit in our experiments. Ψ_g measures the geometric agreement between a projected object, $P_t \cdot o$, and its associated detection. It is implemented as a three-dimensional Gaussian distribution on scale-normalized error in object center (both x and y position) as well as error in predicted image scale.

3 3D Object Inference

The 3D object likelihood model relates the presence of an object in a location to the observed data. However, for every test environment, we must infer the objects that maximize this likelihood. Linear search, such as the sliding-window method is often used for image-space object localization. This approach scales exponentially with dimensionality and we consider locating objects with 6 dimensions (3D position and 3D scale). So, we employ data-driven sampling of likely regions, followed by refinement of each sample and non-maxima suppression. This allows only the most promising regions to be considered and saves considerable computation. The remainder of this section describes our inference procedure in detail.

Data-Driven Sampling

While it is expensive to search for the global maximum that simultaneously explains all observed data, we can efficiently compute the local maxima relative to each view by considering the terms in the product of equation 3 one at a time. First, we draw a detection with probability proportional to the confidence score and constrain the 3D object center to align with the center of the detector's bounding box. This constrains the sampling-space to the ray in 3D over all (infinte) positive depth values. We must also choose a depth and a scale for each proposed 3D sample. Depths are drawn from the values in the depth map within the detector's bounding box. Scales are drawn from the prior on the object's size. This one sample is saved for further processing and the process begins again by selecting a new detection.

Position Refinement

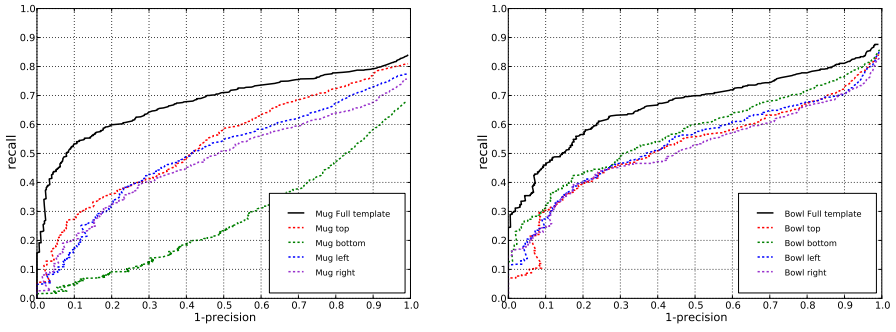
After the sampling stage, a set of 3D regions is available, and it is possible to score each region directly with equation 3. However, unless a large number of samples is used, the 3D localization accuracy is often poor. Therefore, we refine each samples's location and scale to locally maximize the likelihood of observed data in all views. This refinement involves coordinate descent, alternating between two steps. First, the 6 degree-of-freedom (3D position and scale) object pose is optimized given a fixed data association by descending the analytically computed gradient of image errors from projection. Second, the greedy data association is re-computed for the new object location. The projection and greedy data association portions of our object model are non-linear. Therefore, we can make no guarantee on the convergence of the optimization, but we have found the procedure works well in our empirical evaluation.

Non-Maxima Suppression

As with many detectors, our 3D object inference procedure tends to find many slightly shifted versions of each true object with high likelihood scores. We suppress detections which are not local maxima based on their overlap in 3D. Note that our approach can tolerate very cluttered scenes where two objects occupy nearly the same region in image-space. As long as these objects have different depths, we will be able to maintain both hypotheses (there is no overlap in 3D). This is in contrast to many detectors that apply image-space non-maxima suppression which performs poorly when two objects of the same category are nearby in the image.

4 Experimental Setup

We have implemented a complete 3D object detection system, instantiating each of the model components described above in a robust fashion in order to evaluate their performance in realistic, cluttered indoor scenes. This section will describe the practical details of the eval-



(a) (b)

Figure 4: Performance of full and partial detectors for (a) mugs and (b) bowls.

uation including the learned visual recognizers used as input, the physical scenarios used to gather test data, and the structure-from-motion algorithms employed.

4.1 Visual Detectors

We learn detectors for each of four half-sized templates: top, bottom, left, and right. Each partial-object detector is trained independently, as this allows the hard negatives for each template to be included, maximizing resulting detection performance. We employ the Deformable Parts Model of Felzenszwalb *et al.* [7] both for appearance learning and also for test-time detection in images. Both full object models and partial templates are learned from the same training data, which only requires modifying the positive annotations accordingly.

Figure 4 shows the performance of our full template and partial object recognizers over a set of validation images containing annotated examples of each category. The clear trend is that the complete template achieves the best performance overall, which is intuitive because it considers the largest image region and can therefore most strongly discriminate objects from background. We note that the performance of a partial detector does not change at all if an instance is occluded in regions ignored by the template (i.e. a left detector is unaffected by occlusion on the right), which can be leveraged during 3D inference.

4.2 Evaluation Scenarios

We evaluate our method on two scenarios involving indoor clutter and occlusion. First, we locate objects in a publicly available robot-vision dataset named *The UBC Visual Robot Survey*[1]. For each scene in this collection, depth information is available from the robot's tilting laser-rangefinder. The robot surveys the scenes by moving in a trajectory through the environment and collecting both images and point cloud data roughly every five degrees. Each object of interest is manually annotated as both a 2D bounding box and as a 3D location, although all evaluation was done in image-space for this paper using the Pascal [6] evaluation criteria. The UBC VRS dataset is described in more detail in [15].

Our second method for evaluation involves novel data collected for this paper with the Microsoft Kinect sensor in a number of real homes. Here, the sensor was hand-held and we performed marker-less position registration, as will be described below. We present

[1] http://www.cs.ubc.ca/labs/lci/vrs/index.html

Figure 5: Sample results of our 3D object detection method, thresholded at 90% precision. The leftmost image is from the Kinect sensor, the remainder are robot-collected.

our results from this portion of the data qualitatively, as insufficient annotations have been collected so-far to achieve meaningful quantitative comparison.

4.3 Structure from Motion

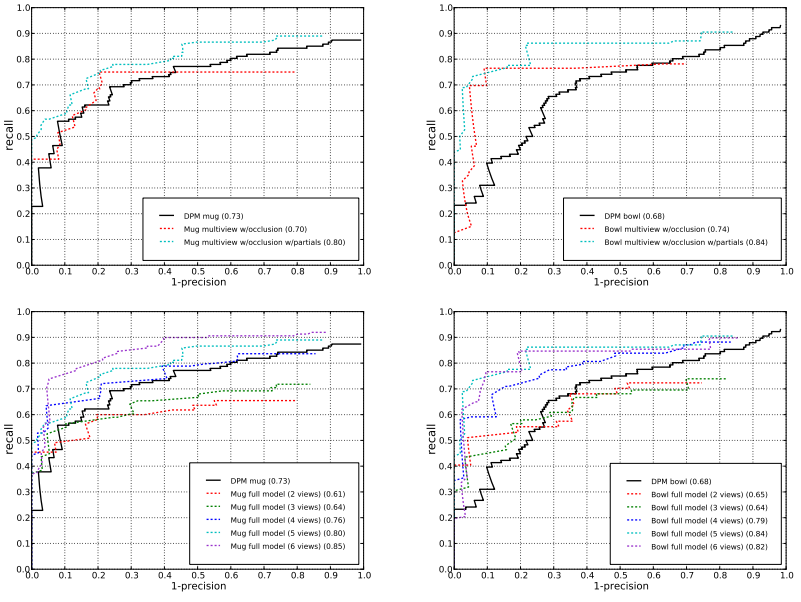
Our experiments include two separate structure-from-motion techniques. The robot-collected images in the UBC VRS are registered using a target made of Artag fiducial markers [8] to obtain exact point correspondences with a vanishingly small chance of false matches. In this case, a non-robust position estimator is initialized with the direct linear transform (DLT) and refined by iteratively minimizing re-projection error. Due to the nearly noiseless point correspondences, this simple method yields registration accurate to within one pixel. This enables careful study of detection performance without undue challenge to the system in recovering the camera's path.

Our Microsoft Kinect data contains unstructured, real home scenarios and no calibration target has been used. Here, we have employed an off-the-shelf technique named RGBD-SLAM - 6DOF SLAM for Kinect-style cameras[2]. Like many SfM solutions for hand-held cameras, Speeded-Up Robust Feature (SURF) [2] points are tracked between frames, and a set of geometrically consistent inliers is found with Randomized Sampling and Consensus (RANSAC) [4]. Long-range performance and loop-closure is achieved by refining poses globally using the technique described by [11]. Registration is not sub-pixel accurate in this scenario, which demonstrates robustness to errors in our geometry model.

5 Results

5.1 Qualitative Results

Figure 5 shows a number of example results from our method. In many scenes, our technique can leverage the reliable information available from visible parts of objects, and confidently locate their position in 3D, even in clutter. However, our system returns false positives on objects whose visual appearance and structure are similar to the searched category. The rightmost image in figure 5 shows a soap dispenser and the top of a bottle which are both labeled "mug". Additional geometric constraints may be able to filter these objects as being in unlikely positions (not resting on table).



(a) (b)

Figure 6: (Top) The performance of our method evaluated over 5 viewpoints vs. the state-of-the-art DPM model by [7]. (Bottom) Full model performance vs. number of viewpoints. Columns: (a) mugs, (b) bowls. The summary statistic is Average Precision.

5.2 Comparison to Visual Recognition

Figure 6(top) shows the results of our 3D object recognition approach, a variant without partial detectors, and the purely image-space DPM model [7]. In all cases, evaluation is on the test set of the UBC VRS dataset. For the 3D detection methods, object volumes are projected to form bounding boxes for scoring. In some cases, our complete model detects 40% more of the annotated objects, for the same miss rate, than the visual detector. The effect of partial detections is shown by the improvement of the complete model over the variant using only full appearance templates. Further inspection reveals that partial detections improve performance primarily on occluded objects.

5.3 Altering the Number of Viewpoints

Figure 6(bottom) shows the results of our method as the number of views considered is increased from two to six. The method must generally observe four or more views of the scene before it outperforms the image-space detector (although all multi-view models perform better at high precision). Visual inspection shows that the issue is poor 3D location of objects when only two or three views is available. We note that the baseline available to our system in this case is only five degrees, the spacing between consecutive frames in UBC VRS data. When 3D locations are poorly estimated, objects project to incorrect image-space locations, degrading performance. We have run a similar 2-view experiment allowing our method to observe non-consecutive views, in this case separated by ninety degrees. The performance

[2] http://openslam.org/rgbdslam.html

in this trial was 0.75 AP for mugs and 0.72 AP for bowls, which improves upon the DPM score, and rivals the four view approach run on consecutive frames.

6 Conclusion

We have developed a method to relate 3D objects to incoming image and geometry data from many views, including explicit occlusion reasoning and learned partial-object appearance models. This approach has the potential to perform robust detection in home scenarios, where intelligent systems will soon be deployed. Our approach is integrated with a structure-from-motion system, and in combination the techniques form a semantic mapping system suitable for object-centric tasks such as scene description to disabled users or mobile manipulation.

References

[1] Bastian Leibe Andreas Ess, Konrad Schindler and Luc Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *The International Journal of Robotics Research*, 29:1707–1725, December 2010.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006.

[3] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, and Esther Koller-Meier a nd Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009.

[5] Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Proceedings of Computer Vision and Pattern Recognition*, 2010.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.

[8] M. Fiala. Artag, a fiducial marker system using digital techniques. In *CVPR'05*, volume 1, pages 590 – 596, 2005.

[9] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381 – 395, 1981.

[10] Mario Fritz, Kate Saenko, and Trevor Darrell. Size matters: Metric visual search constraints from monocular metadata. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2010.

[11] Giorgio Grisetti, Rainer Kuemmerle, Cyrill Stachniss, Udo Frese, and Christoph Hertzberg. Hierarchical optimization on manifolds for online 2d and 3d mapping. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[12] Scott Helmer and David Lowe. Object recognition using stereo. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2010.

[13] Scott Helmer, David Meger, Marius Muja, James J. Little, and David G. Lowe. Multiple viewpoint recognition and localization. In *Proceedings of the Asian Computer Vision Conference*, Queenstown, New Zealand, November 2010.

[14] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Sparse distance learning for object recognition combining rgb and depth information. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[15] David Meger and James J. Little. Mobile 3d object detection in clutter. In *In proceedings of the IEEE/RSJ Conference on Robots and Intelligent Systems (IROS)*, 2011.

[16] A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *Proceedings of the Conference on Uncertainty and Artificial Intelligence*, 2005.

[17] Morgan Quigley, Siddharth Batra, Stephen Gould, Ellen Klingbeil, Quoc V. Le, Ashley Wellman, and Andrew Y. Ng. High-accuracy 3D sensing for mobile manipulation: Improving object detection and door opening. In *ICRA*, 2009.

[18] Min Sun, Bing-Xin Xu, Gary Bradski, and Silvio Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *ECCV*, Crete, Greece, 09/2010 2010.

[19] S. Thrun, C. Martin, Y. Liu, D. Hahnel, R. Emery-Montemerlo, D. Chakrabarti, and W. Burgard. A real-time expectation maximization algorithm for acquiring multi-planar maps of indoor environments with mobile robots. *IEEE Transactions on Robotics and Automation (TRO)*, 20:433 – 442, 2003.

[20] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009.

[21] X. Wang, T. Han, and S. Yan. A hog-lbp human detector with partial occlusion handling. In *Proceedings of ICCV*, 2009.

[22] Christian Wojek, Stefan Roth, Konrad Schindler, and Bernt Schiele. Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. In *Proceedings of ECCV*, 2010.

[23] Christian Wojek, Stefan Walk, Stefan Roth, and Bernt Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, 2011.