Mobile 3D Object Detection in Clutter

David Meger and James J. Little

Abstract— **This paper presents a method for multi-view 3D robotic object recognition targeted for cluttered indoor scenes. We explicitly model occlusions that cause failures in visual detectors by learning a generative appearance-occlusion model from a training set containing annotated 3D objects, images and point clouds. A Bayesian 3D object likelihood incorporates visual information from many views as well as geometric priors for object size and position. An iterative, sampling-based inference technique determines object locations based on the model. We also contribute a novel robot-collected data set with images and point clouds from multiple views of 60 scenes, with over 600 manually annotated 3D objects accounting for over ten thousand bounding boxes. This data has been released to the community. Our results show that our system is able to robustly recognize objects in realistic scenes, significantly improving recognition performance in clutter.**

I. INTRODUCTION

This paper considers the problem of locating objects specified by category in 3D, based on images and point cloud data collected from several viewpoints by a mobile platform. We focus on cluttered indoor scenes, where objects are often not completely visible from a single viewpoint. Here, occlusion is a primary mode of failure for state-of-the-art visual recognizers and geometry-based shape models alike. We propose a method to integrate information from many views and to reason explicitly about occlusion in each view. Our system can reliably locate objects in 3D even when they are occluded, as demonstrated by the sample result in figure 1.

Our approach augments a state-of-the-art visual category recognizer by modeling its performance as a function of visibility. We learn an occlusion model from a novel dataset of registered images and point clouds where the 3D location and category label of objects have been manually annotated. We are motivated by the observation that even the best current object detectors often produce false negatives for objects that are only partially visible. This is a primary failure mode in cluttered areas typical for real homes. By combining visual object models, images, point clouds, and annotated 3D object regions, our system learns a generative model for the expected drop in detector score under occlusion, which allows optimal use of the weaker appearance information that is present. The final component of our method is an inference procedure that uses sampling and iterative refinement to locate the 3D objects that best explain the sensory evidence – effectively lifting 2D detections to 3D object regions. Note that we localize objects based on semantic categories where the test instances have not been given to the system during training rather than recognizing specific object instances seen both in training and test.

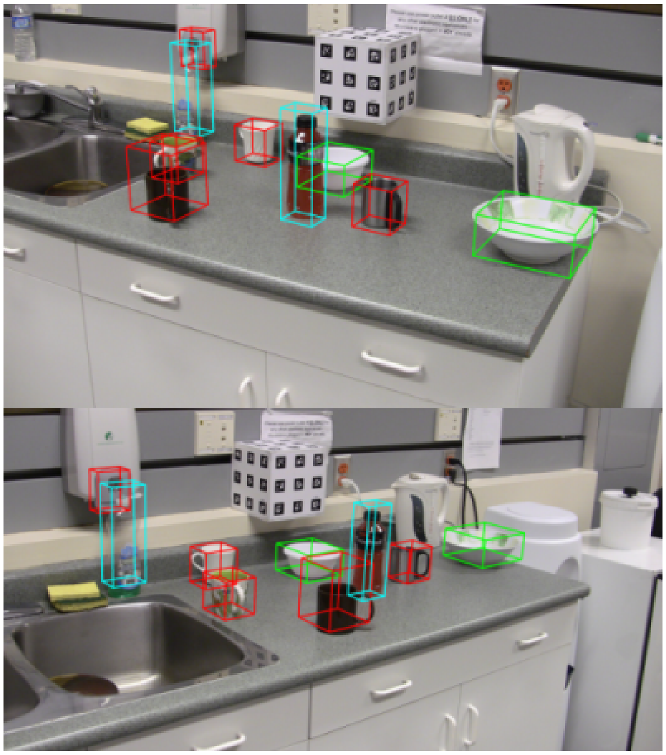


Fig. 1. An example result from our 3D object detection system evaluated on 10 views of a cluttered kitchen. This result is thresholded at one false positive per image and shows mugs in red, bowls in green and bottles in blue. Best viewed in colour.

Few previous robotic recognition methods have been made repeatable for other researchers since they are tied to their target platform. To improve this situation, we have assembled a new robot-collected database of images and scans from a tilting laser range finder and made our data publicly available. The UBC Visual Robot Survey (VRS[1]) includes 60 environments measured from numerous viewpoints. Data from each viewpoint is registered geometrically by use of a visual fiducial system. This allows us to focus only on locating objects in 3D. We assume visual SLAM systems will soon be robust and off-the-shelf so that this is a reasonable reduction in complexity. Instances of object categories have been manually annotated in both 2D and 3D for purposes of training and evaluation.

All authors are with the Department of Computer Science, University of British Columbia. Contact {dpmeger,little}@cs.ubc.ca

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

II. Related Work

Our system must learn the visual appearance of object categories from training data. We have utilized the state-of-the-art visual recognizer named *Deformable Parts Model* (DPM) [1] for this task. While our method is compatible with any recognizer that produces scored bounding boxes from test images, DPM is open-source and has placed first place in several recent Pascal Visual Object Categories challenges, so it provides a strong baseline for comparison and improvement.

Several previous approaches have combined information from many images in order to locate objects. This is a common approach for tracking moving objects where examples include [2], [3], [4], [5] among many others. Our method shares several characteristics with the work of Wojek *et al.* [6] as they also lift 2D detections of pedestrians to 3D in order to perform scene level inference. However, our approach handles more varied environment geometry, and we target indoor robotics where sensed point cloud information is also available.

Other authors have also considered fusing information between depth sensors and images, for example: [7], [8], [9]. However, we are not aware of other work that has directly reasoned about occlusion and fused this with image appearance reasoning.

Integrating imagery from many viewpoints of the same scene has also been studied. Many papers in the Active Vision literature involve similar high-level formulations to our own. Laporte *et al.* [10] is one particularly similar example where a Bayesian formulation is used. Even more similar are approaches targeted to recognizing objects in indoor scenes, such as [11] and our previous work [12], [13]. We have previously proposed the use of a generative model to combine detections, but this is our first attempt to include occlusion reasoning as a variable within the model, which increases the performance in high clutter.

Occlusion has been studied in the context of other Computer Vision problems, such as determining optical flow [14], segmentation [15], and recognition from single images [16]. Our method has been inspired by these approaches, but we do not believe any has constructed a generative appearance-occlusion model, as we require for 3D object inference.

Several robotics competitions have recently been proposed to address the difficulty in evaluating robotic recognition approaches. The Semantic Robot Vision Challenge [17] proposed recognizing object categories in simple environments based on training data from the world wide web. The scenes contained in our dataset are much more realistic and cluttered than those that were used for the contest, and our data is publicly available. More recently the Solutions in Perception Challenge [18] was held with the goal of evaluating performance on recognizing specific object instances using both images and depth data from the Microsoft Kinect sensor. Specific instance recognition is typically easier than categorization, which we study, since there is no need to deal with intra-category appearance variability.

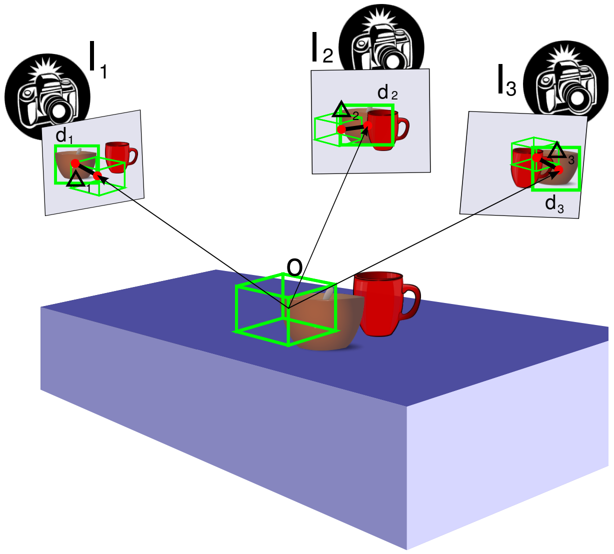


Fig. 2. An overall view of our model. Each inferred 3D object is projected into many views. The proposed region projects into all 3 views and associates to image-space detections with alignment errors Δ_i. The likelihood of this hypothesis is expressed based on the 3D geometry, the score of matching detections modulated by occlusion and the error in projection.

III. 3D Object Model

A. Overview

At a high level, our approach attempts to infer which objects exist in an environment based on sensed 3D information, (i.e. point clouds from a tilting laser range-finder or depth camera), and visual imagery, collected from many viewpoints. We attempt to locate objects as 3D volumes, and do not study pose estimation in this work. To achieve a manageable scope, we assume that a SLAM component external to our system estimates the robot's pose and that camera and robot calibration are available, which allows all measurements to be related in a common coordinate frame.

Our system hypothesizes many possible objects (both category labels and locations). As shown in figure 2, each potential object explains the visual appearance in all views through projection into image space and data association to a candidate detection bounding box returned by a recognizer. Note that when an object is occluded from some viewpoints, the data association for those views may fail, or the recognizer may report a low probability of the object existing in that view. If left un-modeled, this results in poor system performance in clutter, so we have explicitly addressed this problem by adding an occlusion term to our appearance model. This term and a 3D geometry likelihood for the hypothesized position are both derived from sensed 3D geometry.

B. Object Likelihood

We infer the set of objects present in an environment based on images, I_t, and point clouds, P_t, collected by a robot over time, from many views of a static scene. We denote all inputs for one time-step as $z_t = \{I_t, P_t\}$. We use superscript notation to represent the set of all data up to a given time: $Z^t = z_1, z_2..., z_t$.

Each static 3D object, $o = \{pos, sc, c\}$, is described by a category label c and a 3D location determined by a position, pos, and scale, sc. Our model includes a set of learned parameters θ to be described shortly. We will use θ as a generic term to describe all system parameters, and use subscripts when we refer to specific parameters. For example, θ_{geom} describes the parameters used to match an object to sensed 3D geometry. Therefore, we model $p(o|Z, \theta)$. In order to exploit problem structure and learned generative components, we apply Bayes rule and assume the information from each view is independent:

$$\begin{aligned} p(o|Z^t, \theta) \quad &\alpha \quad p(o, \theta)p(Z^t|o, \theta) &(1)\\ &= \underbrace{p(o, \theta)}_{object\ prior} \prod_t \underbrace{p(z_t|o, \theta)}_{data\ likelihood} &(2)\end{aligned}$$

Our prior for objects, $p(o, \theta_{geom})$, is a size distribution of instances from each object category. We modeled our objects with a normal distribution for both the height and radius, as they are all nearly cylindrical in nature. Our model is not fixed to the assumption of cylindrical shape; it has merely been effective for the categories we study. In principle, any other 3D shape prior emitting a size likelihood can also be used and more informative shape models such as [19] would further improve our method.

We decompose the data likelihood term, $p(z_t|o, \theta)$ into two generative sub-models that describe how well an object explains the image appearance I_t and the point cloud geometry, P_t. In order to model the occlusion in an image based on the sensed 3D geometry, we factor the probability such that the image appearance is conditioned on the 3D information:

$$\begin{aligned} p(z_t|o, \theta) &= p(P_t, I_t|o, \theta) &(3)\\ &= \underbrace{p(P_t|o, \theta)}_{geometry}\underbrace{p(I_t|P_t, o, \theta)}_{appearance} &(4)\end{aligned}$$

C. Geometric Reasoning

We use a *depth map* representation of the sensed geometry (i.e. the projection of the 3D point cloud onto a 2D image plane from the perspective of each viewpoint) to reason about two spatial properties for every object volume: its per-view occlusion and whether it overlaps occupied space. We project the hypothesized 3D objects into the same frame as the depth map using registration information, and evaluate the range agreements. There are 3 possible outcomes for each pixel within the image region covered by the hypothesized object:

1) The sensed depth is closer than the object, indicating the proposed location is **occluded**.
2) The depths match, so the proposed location is in the image **foreground**.
3) The sensed depth is farther than the object, indicating the proposed location is **unoccupied**.

Several examples of the pixels marked **occluded** by this process are shown in figure 3. In order to form the likelihood models described above, two ratios summarize each depth pixel labeling: $free(o, P_t)$, or free space within the object volume, is the ratio of the number of unoccupied pixels to the number of pixels; and $occ(o, P_t)$, or occlusion of the volume, is computed as the ratio of the number of occluded pixels to the total number of pixels.

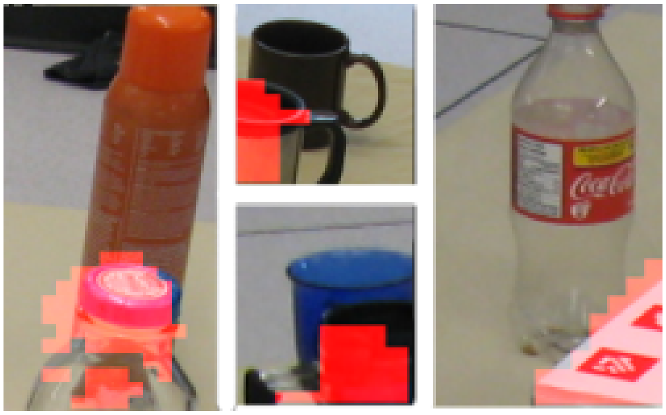


Fig. 3. Occlusion masks generated by our spatial reasoning procedure are shown in red. In all of the displayed examples, the target object is the one seen in the background and the occlusion mask is generated by our system to reason about where the target is not visible.

The geometry term from equation (4), $p(P_t|o, \theta_{geom})$ is derived based on $free(o, P_t)$. We model the sensor error as a zero-mean normal distribution with variance parameter θ_{geom}. If a large fraction of the volume described by o is free space, it is unlikely that an object exists in the proposed location. This reasoning allows our system to discard many false object hypotheses suggested by visual recognition that are not consistent with sensed 3D geometry.

In the next section, we continue to describe the visual detection and occlusion model, which will leverage the occlusion ratio computed here.

D. Appearance and Occlusion Likelihood

We model visual appearance using a set of bounding boxes detected by a visual object recognizer. We denote the set of detections from the image at time t as $D_t = D(I_t)$ and we write each detection as $d_{tj} = \{score, x, y, sc\} \in D_t$, for detector confidence, $score$, image location, (x, y), and scale, sc. 3D object hypotheses are projected into each image and a data association function, $a(d, o)$, produces one matching detection for each object in each image: $d_{to} = \{d_{tj}|a(o, d_{tj})\}$. The appearance term $p(d_{tj}|o, P_t, \theta)$ models both the image space agreement and the detector's prediction score. The intuition is that correctly inferred objects will project near to areas given high scores by the image-based recognizer, so we penalize error in re-projection agreement and also low detection scores for each view.

$$\begin{aligned} p(I_t|P_t, o, \theta) &= p(d_{to}|o, P_t, \theta) &(5)\\ &= \underbrace{p(\Delta(d_{to}, o)|\theta_\Delta)}_{location}\underbrace{p(d_{to}|occ(o, P_t), \theta_{detocc})}_{detector\ score}\end{aligned}$$

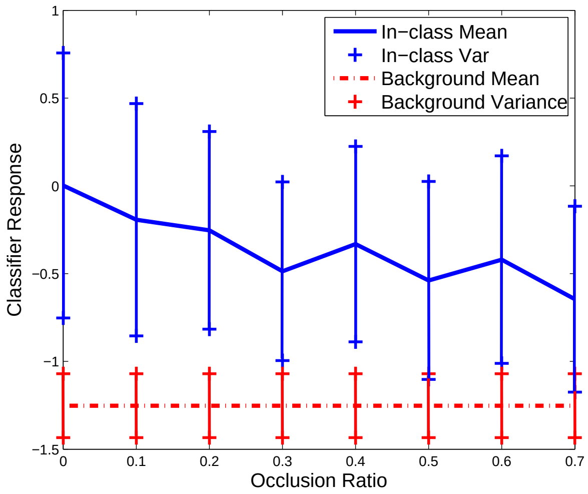


Fig. 4. The generative detection model obtained for the DPM [1] mug detector. The in-class mean shows the average response of the classifier on objects at various occlusion levels. The background distribution is plotted in red for comparison. This reflects the distribution of all false-positive detector responses (i.e. detections reported when the category is not present, or is fully occluded.)

```
**Input**: $Z^t$,$\theta$
**Output**: $o_i \in O$: inferred set of objects
$O_{final} = \emptyset$;
foreach $cat \in Categories$ do
    $O_{cat}$ = Bootstrapped set of candidate object regions;
    while assignments not converged do
        foreach $o_i \in O$, order by descending likelihood
        do
            foreach $z_t \in Z^t$ do
                $d_{it}$ = $assoc(o_i, D(I_t))$;
            end
            Refine($o_i$);
            if changed($o_i$) then
                Re-compute $occ(o_i, P_t)$ for all views;
                Re-compute $p(o_i|Z^t, \theta)$ by equation (1);
            end
        end
    end
    Append $O_{cat}$ to $O_{final}$;
end
$O_{final}$ = NMS($O_{final}$);
```

Algorithm 1: The 3D Object Inference Procedure

The function $\Delta(d, o)$ measures the error in image location and scale between the projection of the 3D object volume and the detected bounding box. This alignment error is displayed in figure 2 and should be near zero for correct 3D regions. Image location and scale disagreement are both penalized. For fair comparison, we normalize location by the scale of the detection bounding box. A zero-mean Gaussian, with variances for each of x, y, and $scale$[2], models the alignment error.

The second component of the visual appearance model is derived from the detector's confidence score. This is our system's primary way to reason about the correct category label for a region. The object detector will typically produce high scores for un-occluded object instances. The score is often reduced in the presence of occlusion. Recall that $occ(o, P)$, the occlusion statistic, was described previously in section III-C, and is visualized in figure 3. The intuition is that every occluded pixel (marked red in the figure) hides object evidence and accounts for a lower detector score. The parameter θ_{detocc} models this relation, and we will now describe our method for learning this from the annotations, images and point clouds in the UBC VRS dataset.

IV. MODEL LEARNING

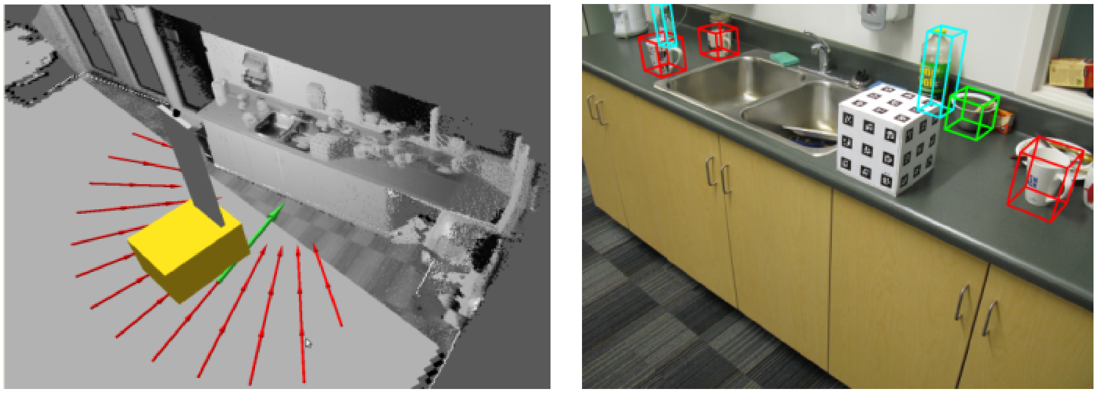
This section describes a method to learn the parameters for the appearance-occlusion distribution: $p(d_{tj}|occ(o, P_t), \theta_{detocc})$. The learning procedure requires a visual object detector to be evaluated on a set of data where objects are annotated in 3D and where we have access to point cloud representations of the environment. We obtain both of these from the validation set of the UBC VRS database. Then, our learning procedure associates image space detections on the validation set to 3D annotated objects by assigning the closest object if there is significant overlap. If no overlapping object exists, a false detection is indicated. The occlusion mask for each 3D object is automatically computed, using the approach we described above. This leads to a large set of $(score, occ)$ pairs, and also many false positive detection responses.

We define a model relating the terms and regress the parameter θ_{detocc} to fit the model to the collected data. Our model is a unique $1D$ Gaussian for each value of the occlusion ratio. For robustness in learning the parameters, we bin training data with bin width 0.1 and vote into neighboring bins with weight inverse to the distance from the bin centre. To obtain final estimates over the entire parameter space, we linearly interpolate parameters between bins. This allows computation of the likelihood for any detector response over the full space of occlusions. However, we have found this process to become unreliable when less than 0.3 of the object is visible, since our visual detector tends not to produce any bounding boxes. We therefore truncate the model and label all objects with $occ(o, P) > 0.7$ as fully occluded (i.e. we expect the detector to respond as if no object was present). Figure 4 displays the model we obtain by applying this process for the *mug* category.

V. 3D OBJECT INFERENCE

The model and learned parameters described previously are used to infer objects as a robot explores an environment. An exact inference technique would involve evaluating all object regions in 6 continuous dimensions (3D location and scale). One could also consider all possible near-intersections of 2D detections, but number of near-intersections is the

[2] The optimal values projection agreement are detector dependent. For the DPM detector, we have used variances 0.25 for position and 0.5 for scale.



(a) (b)

Sub-set	Scenes	Views	Instances	Boxes	Mugs	Bottles	Bowls
Training	19	453	184	4026	56/1405	42/1119	17/378
Validation	11	295	116	2701	33/711	33/839	17/346
Test	30	334	303	3466	85/935	57/589	64/681
UBC VRS	60	1082	603	10193	174/3051	132/2547	98/1405

(c)

Fig. 5. The UBC VRS Dataset. (a) A sample point cloud, and poses from the survey path followed by the robot. (b) A sample image with 3D wireframes projected to display user-annotated ground truth volumes. (c) Summary statistics of the annotations available for the UBC VRS database. The final 3 columns represent the (unique instances / number of bounding boxes) that are present for the specified category.

same as the cardinality of the power set of detections in the worst case (with cardinality exponential in the number of detections). We propose an approximation with complexity linear in the number of collected views that still allows robust inference in clutter.

We start by proposing a coarsely sampled set of object volumes. This is a 3D analog to the *sliding window* approach commonly used in visual detection (an apt name for our 3D sampler is *sliding volume*). These volumes will initially tend to have large alignment errors with detections, because a coarse spatial sampling is unlikely to include exactly the correct region. Therefore, we iteratively refine the proposed regions by alternating two steps: (1) performing data association that connects the closest detections to the projected volume in each image; and (2) geometric refinement based on gradient ascent, that maximizes the appearance likelihood term. The outside-loop iteration is needed because as 3D object volumes are moved to new positions, the greedily chosen data associations may change, calling for another step of refinement. We have found that this process stabilizes very rapidly. Four iterations was the maximum needed across the evaluation performed in the next section. A crucial step has been running the gradient ascent process to convergence for each association, rather than stopping after a few steps. The optimal object position given an association quickly leads to better associations and avoids oscillation between two associations that suggest nearby locations. Algorithm 1 describes the procedure. We continue by describing our experimental validation.

VI. EXPERIMENTAL EVALUATION

In this section, we will study the performance of our system for detecting mugs, bottles and bowls in 3D, on the test set of the UBC VRS database, which contains 30 realistic cluttered indoor scenes. Like the majority of modern object recognizers, the outputs of our system are scored with probabilistic confidence estimates. In practice, a physical system would likely choose a single confidence threshold and act only on the objects detected over this threshold. Here, we evaluate across a range of thresholds, using standard detector evaluation practices. Qualitative results are shown for the threshold equal to one false positive per image and recall versus false-positive-per-image curves are used for quantitative results. The remainder of this section describes our experimental evaluation. Next, we will describe the UBC VRS data and annotations that have been used for comparison.

A. UBC VRS Description and Testing Protocol

The UBC VRS database was collected by directing a mobile platform, previously described in [20], to follow a *survey trajectory* for a number of realistic scenes. For each scene, the robot moves in a circular arc, collecting data from vantage points evenly spaced throughout the traversable area. Figure 5(a) shows the waypoints that make up the *survey trajectory*, for one kitchen environment contained in the dataset.

A fiducial-based registration system [21] has been used to obtain accurate position information that mimics the outputs of a SLAM algorithm. A human annotator has labeled each image with bounding boxes containing three categories (mugs, bowls and bottles) along with many other objects that are not studied here. Each object's 3D location is also annotated as a bounding volume. This is done via triangulation of numerous user-clicked points using the geometric

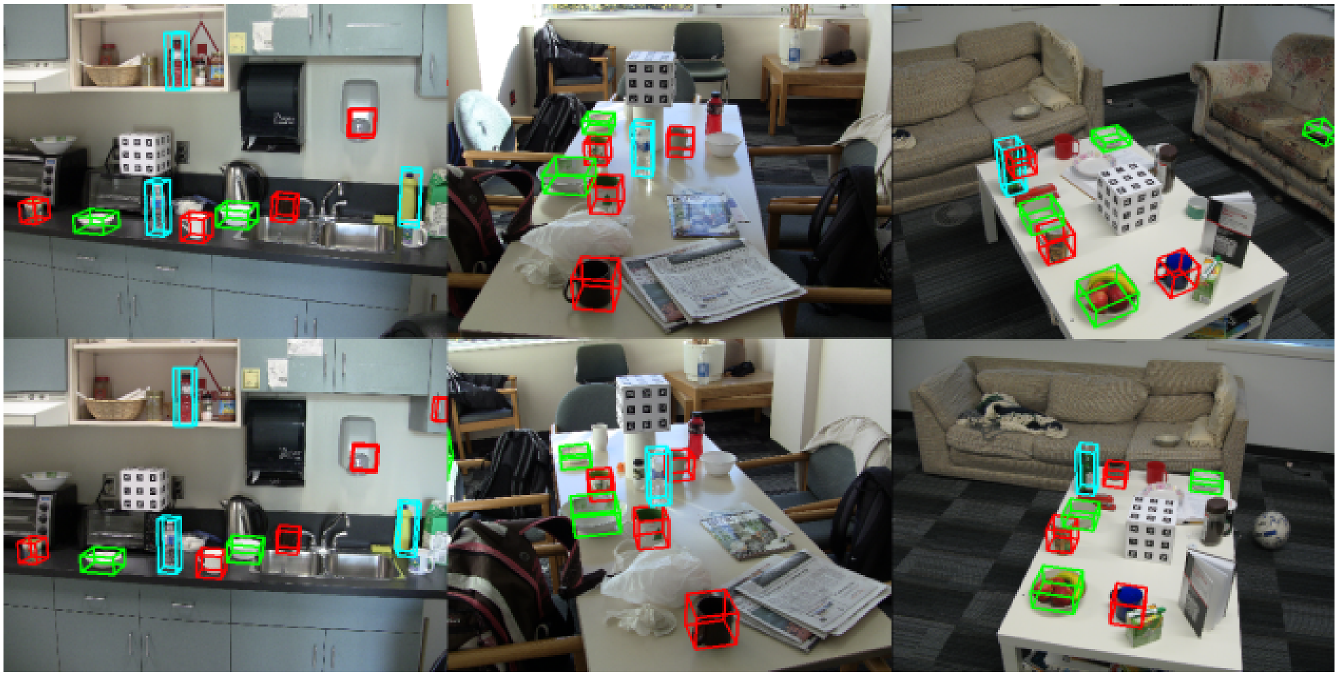


Fig. 6. 3D objects inferred by our system on a number of the UBC VRS scenes after 10 views. Displayed objects score over a threshold equivalent to one false positive per image. Red volumes indicate mug hypotheses, green are bowls and blue are bottles.

registration. Figure 5(b) shows example human-annotated 3D volumes.

All further results in this section involve the task of inferring the set of objects present in the test portion of the data set. The table in figure 5(c) summarizes the number of annotations that form this set. Our testing protocol provides data from a series of consecutive viewpoints along the *survey path* for each scene to the inference system. This is exactly the information that a robot could collect after moving in a short path while exploring an environment.

The ground truth annotations are used to evaluate performance using a methodology similar to that used in multi-camera object tracking. Specifically, we use the criteria outlined in the recent Performance Evaluation of Tracking and Surveillance (PETS) workshop [22]. There, tracking systems can access $N - 1$ views to produce results and the N^{th} view is designated for evaluation. Estimates are projected into the evaluation view and compared to ground truth annotations.

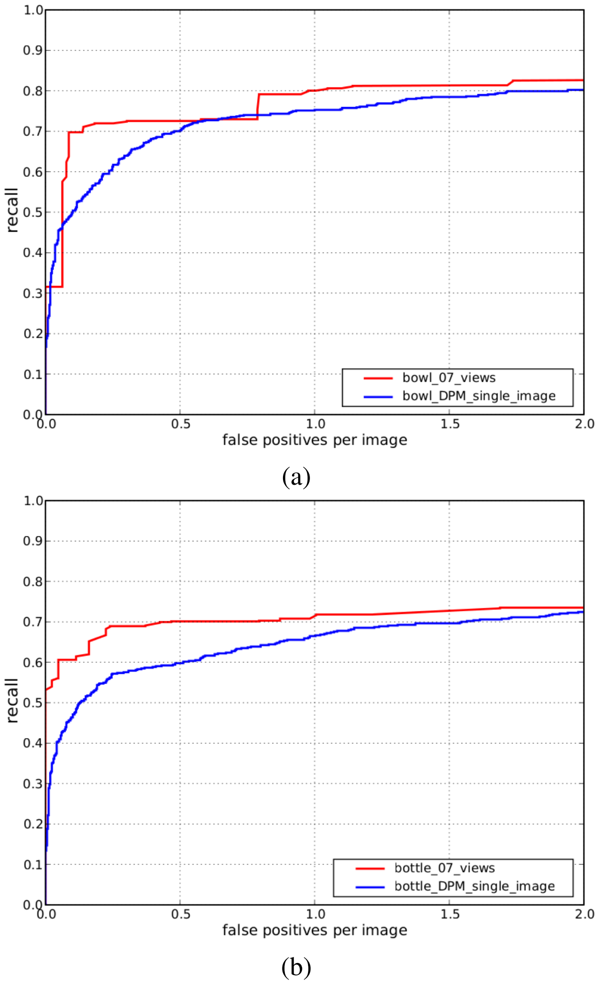
B. Sample Results

Figure 6 shows a thresholded set of object detections after ten views have been considered. Our system is often able to find correct 3D positions and achieves partial separation between true objects and false positives. That is, many objects can correctly be returned before the first false positive. However, mistakes are present. For example, the first column of the figure shows a soap dispenser labeled as "mug". This instance is nearly unavoidable for our system because the visual recognizer returns strong mug responses in every view for this object, the 3D region has nearly the same size as a typical mug, and the image rays intersect on physical structure. At the displayed thresholding, several false negatives are also visible: (first column) a large bowl on the left and a mug on the right of each image; two bottles, a mug and a bowl in the middle column, in the background where the small image size makes the object recognizer less confident in the image evidence; and two mugs and a bowl in the rightmost column, which are recognized, only not with a confidence below the display threshold. We also note that the 3D position of the suggested regions is typically accurate enough to visually represent the object well. However, in this paper we have not considered grasping or other precise robotic tasks. In that domain, a further refinement step is likely necessary.

C. Quantitative Evaluation

This section reports a quantitative summary of our method evaluated on all 30 test scenes in the UBC VRS dataset using the testing protocol described above. The "recall vs false-positive-per-image (fppi)" plot is a standard technique in Computer Vision. Recall refers to the ratio of the number of annotated objects correctly detected to the total number of objects annotated (1.0 is perfect performance). Fppi is a measure of precision. Each data point reflects the system's output thresholded at one specific confidence rate and the threshold is varied to produce the curve.

We begin by comparing the performance of our system to the state-of-the-art visual category recognizer by Felzenszwalb *et al.* [1] in figure 7. For both bowls and bottles our method achieves higher recall for nearly every fppi value. In some cases, our system recognizes five to ten percent more of the instances present with the same miss rate as the image space detector. However, we acknowledge that this



(a)

(b)

Fig. 7. Performance of the (a) bowl and (b) bottle detector on the UBC VRS test set compared with a state-of-the-art Deformable Parts Model (DPM) [1]. Note, that this detector has been scanned on the evaluation view directly while our 3D results are projected after the robot has seen 7 of the other views. Best viewed in colour.

comparison is problematic for several reasons. Our system has access to imagery from many views as well as point cloud data, while the visual DPM detector sees only the evaluation image. Also, DPM reports its results directly as bounding boxes on the evaluation image, while our technique reports 3D volumes that must be projected, so incorrect 3D location of objects can account for misses of our system. We have provided this analysis simply to note that our system does generally return more correct detection results than DPM.

We next analyze the performance of our method as an increasing number of views is made available to the inference procedure. Data from a larger number of views allows the visual appearance of 3D objects to be verified against a larger number of image patches. It also leads to a larger maximum baseline between views, since the path between first and last view is longer. Larger baseline constrains the object's position more tightly and leads to less error in projection of 3D volumes. Figure 8 displays our system's performance after path lengths, $p \in \{3, 7, 10, 12\}$. Performance generally increases with a larger number of views, but there are local deviations. We have investigated these and note that more available views leads to a larger search space and more accidental intersections between detections.

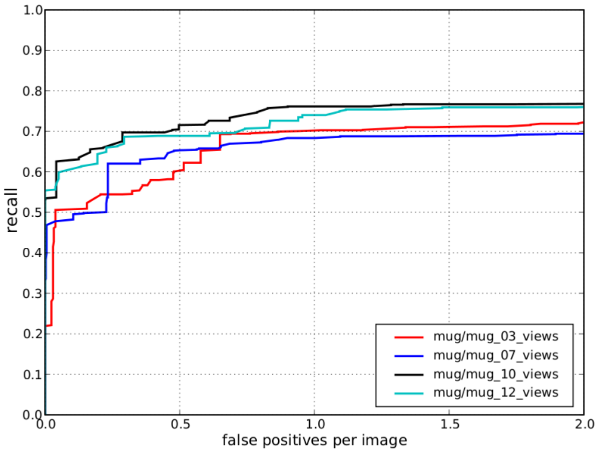


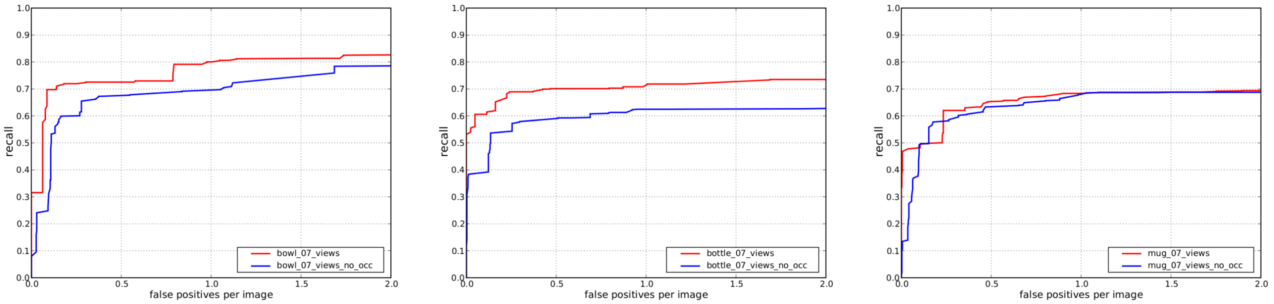
Fig. 8. Performance of our mug detector on the UBC VRS test set as the number of input views available is varied. Best viewed in colour.

The final quantitative evaluation involves disabling a primary system component. We remove the effect of point cloud data on the object likelihood, making for an entirely image-driven approach. This is a simple change to our system involving two small modifications: the geometry likelihood is set to 1.0 for all objects, giving no information; and the appearance-occlusion model is always evaluated as if the object is fully un-occluded. Figure 9 demonstrates that substantial performance is lost for bowls and bottles, while the performance of the mug detector is only affected slightly. The variation across categories appears to come from the fact that the visual detector has learned a slightly more reliable template for mugs than for the other two classes.

We have identified a number of true positive objects whose score is lowered when the occlusion variable is not present. This appears on the curves in figure 9 as the difference in starting value of recall where the curves intersect the Y-axis. It can be explained by the fact that the projection of the 3D objects into occluded views are now penalized the same as in un-occluded views. When detections are missed, due to occlusion, the object's score is penalized.

VII. CONCLUSION

We have presented a method that locates objects accurately in cluttered indoor scenes. 3D annotations and point cloud data are used during a learning phase to characterize the effect of occlusion on the output of a visual recognizer. The resulting model is integrated over many viewpoints to produce accurate final estimates of the objects in the scene, in 3D, even in the face of substantial clutter and frequent object occlusion. Our results show improved performance over one of the best current visual detectors, and that 3D detections are made accurately even when only a small number of views are available.



(a) (b) (c)

Fig. 9. Performance of our system on the UBC VRS test set with and without the occlusion terms of the likelihood model for (a) bowls, (b) bottles and (c) mugs. Best viewed in colour.

In future work, we plan to address the computational challenges inherent in extending our model to operate in an incremental fashion. Also, we currently summarize occlusion masks as a single ratio. This ignores the fact that some parts of an object's appearance are more discriminative than others, and we believe that continuing to exploit part-level information along with 3D geometric constraints will allow much enhanced understanding of object locations, even when they are only partially visible, and will allow our system to make a strong contribution to the field of indoor perception.

References

[1] 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*, vol. 32, 2010.

[2] B. L. Andreas Ess, Konrad Schindler and L. V. Gool, "Object detection and tracking for autonomous navigation in dynamic environments," *The International Journal of Robotics Research*, vol. 29, pp. 1707–1725, 2010.

[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[4] D. Schulz, W. Burgard, D. Fox, , and A. Cremers, "People tracking with a mobile robot using sample-based joint probabilistic data association filters," *International Journal of Robotics Research (IJRR)*, vol. 22, pp. 99 – 116, 2003.

[5] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2010.

[6] C. Wojek, S. Roth, K. Schindler, and B. Schiele, "Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes," in *In proceedings of European Conference on Computer Vision (ECCV)*, 2010. [Online]. Available: http://www.d2.mpi-inf.mpg.de/monocular-3d-scene

[7] S. Helmer and D. Lowe, "Object recognition using stereo," in *In proceedings of the International Conference on Robotics and Automation (ICRA)*, 2010.

[8] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller, "Integrating visual and range data for robotic object detection," in *In proceedings of the ECCV workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, 2008.

[9] M. Fritz, K. Saenko, and T. Darrell, "Size matters: Metric visual search constraints from monocular metadata," in *In proceedings of Neural Information Processing Systems (NIPS)*, 2010.

[10] C. Laporte and T. Arbel, "Efficient discriminant viewpoint selection for active bayesian recognition," *International Journal of Computer Vision*, vol. 68, no. 3, pp. 267–287, 2006.

[11] B. Rasolzadeh, M. Bjorkman, K. Huebner, and D. Kragic, "An active vision system for detecting, fixating and manipulating objects in the real world," *The International Journal of Robotics Research*, vol. 29, pp. 133 – 154, 2010.

[12] S. Helmer, D. Meger, M. Muja, J. J. Little, and D. G. Lowe., "Multiple viewpoint recognition and localization," in *Proceedings of the Asian Computer Vision Conference*, 2010.

[13] D. Meger, A. Gupta, and J. J. Little, "Viewpoint detection models for sequential embodied object category recognition," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2010.

[14] A. D. Jepson, D. J. Fleet, and M. J. Black, "A layered motion representation with occlusion and compact spatial support," in *In proceedings of the European Conference on Computer Vision (ECCV)*, 2002, pp. 692 – 706.

[15] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes, "Layered object detection for multi-class segmentation," in *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

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

[17] Website: http://www.semantic-robot-vision-challenge.org/.

[18] Website: http://opencv.willowgarage.com/wiki/SolutionsIn Perception-Challenge.

[19] K. Lai and D. Fox, "Object detection in 3d point clouds using web data and domain adaptation," *International Journal of Robotics Research, Special Issue from RSS 2009*, 2010.

[20] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, D. G. Lowe, and B. Dow, "Curious George: An attentive semantic robot," *Robotics and Autonomous Systems Journal Special Issue on From Sensors to Human Spatial Concepts*, vol. 56(6), pp. 503–511, 2008.

[21] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *CVPR'05*, vol. 1, 2005, pp. 590 – 596.

[22] Website: http://pets2010.net/.