A Category-Level 3-D Object Dataset: Putting the Kinect to Work

Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, Trevor Darrell
UC Berkeley and Max-Plank-Institute for Informatics
{allie, sergeyk, jiayq, barron, saenko, trevor}@eecs.berkeley.edu, mfritz@mpi-inf.mpg.de

Abstract

Recent proliferation of a cheap but quality depth sensor, the Microsoft Kinect, has brought the need for a challenging category-level 3D object detection dataset to the fore. We review current 3D datasets and find them lacking in variation of scenes, categories, instances, and viewpoints. Here we present our dataset of color and depth image pairs, gathered in real domestic and office environments. It currently includes over 50 classes, with more images added continuously by a crowd-sourced collection effort. We establish baseline performance in a PASCAL VOC-style detection task, and suggest two ways that inferred world size of the object may be used to improve detection. The dataset and annotations can be downloaded at http://www.kinectdata.com.

Figure 1. Two scenes typical of our dataset.

1. Introduction

Recently, there has been a resurgence of interest in available 3-D sensing techniques due to advances in active depth sensing, including techniques based on LIDAR, time-of-flight (Canesta), and projected texture stereo (PR2). The Primesense sensor used on the Microsoft Kinect gaming interface offers a particularly attractive set of capabilities, and is quite likely the most common depth sensor available worldwide due to its rapid market acceptance (8 million Kinects were sold in just the first two months).

While there is a large literature on instance recognition using 3-D scans in the computer vision and robotics literatures, there are surprisingly few existing datasets for category-level 3-D recognition, or for recognition in cluttered indoor scenes, despite the obvious importance of this application to both communities. As reviewed below, published 3-D datasets have been limited to instance tasks, or to a very small numbers of categories. We have collected and describe here the initial bulk of the Berkeley 3-D Object dataset (B3DO), an ongoing collection effort using the Kinect sensor in domestic environments. The dataset already has an order of magnitude more variation than previously published datasets. The latest version of the dataset is available at http://www.kinectdata.com

As with existing 2-D challenge datasets, our dataset has considerable variation in pose and object size. An important observation our dataset enables is that the actual world size distribution of objects has less variance than the image-projected, apparent size distribution. We report the statistics of these and other quantities for categories in our dataset.

A key question is what value does depth data offer for category level recognition? It is conventional wisdom that ideal 3-D observations provide strong shape cues for recognition, but in practice even the cleanest 3-D scans may reveal less about an object than available 2-D intensity data. Numerous schemes for defining 3-D features analogous to popular 2-D features for category-level recognition have been proposed and can perform in uncluttered domains. We evaluate the application of HOG descriptors on 3D data and evaluate the benefit of such a scheme on our dataset. We also use our observation about world size distribution to place a size prior on detections, and find that it improves detections as evaluated by average precision, and provides a potential benefit for detection efficiency.

2. Related Work

There have been numerous previous efforts in collecting datasets with aligned 2D and 3D observations for object recognition and localization. We review the most pertinent ones, and briefly highlight how our dataset is different. We also give a brief overview of previous work targeting the integration of the 2D appearance and depth modalities.

2.1. 3D Datasets for Detection

We present an overview of previously published datasets that combine 2D and 3D observation and contrast our dataset from those previous efforts:

RGBD-dataset of [20]: This dataset from Intel Research and UW features 300 objects in 51 categories. The category count refers to nodes in a hierarchy, with, for example, *coffee mug* having *mug* as parent. Each category is represented by 4-6 instances, which are densely photographed on a turntable. For object detection, only 8 short video clips are available, which lend themselves to evaluation of just 4 categories (bowl, cap, coffee mug, and soda can) and 20 instances. There does not appear to be significant viewpoint variation in the detection test set.

UBC Visual Robot Survey [3, 18]: This dataset from UBC provides training data for 4 categories (mug, bottle, bowl, and shoe) and 30 cluttered scenes for testing. Each scene is photographed in a controlled setting from multiple viewpoints.

3D table top object dataset [23]: This dataset from University of Michigan covers 3 categories (mouse, mug, stapler) and provides 200 test images with cluttered backgrounds. There is no significant viewpoint variation in the test set.

Solutions in Perception Challenge [2]: This dataset from Willow Garage forms the challenge which took place in conjunction with International Conference on Robotics and Automation 2011, and is instance-only. It consists of 35 distinct objects such as branded boxes and household cleaner bottles that are presented in isolation for training and in 27 scenes for test.

Other datasets: Beyond these, other datasets have been made available which do include simultaneous capture of image and depth but serve more specialized purposes like autonomous driving [1], pedestrian detection [9] and driver assistance [24]. Their specialized nature means that they cannot be leveraged for the multi-object category localization task that is our goal.

In contrast to all of these datasets, our dataset contains both a large number of categories and many different instances per category, is photographed "in the wild" instead of in a controlled turntable setting, and has significant variation in lighting and viewpoint throughout the set. For an illustration, consider Figure 4, which presents a sample of examples of the "chair" category in our dataset. These qualities make our dataset more representative of the kind of data that can actually be seen in people's homes; data that a domestic service robot would be required to deal with and do online training on.

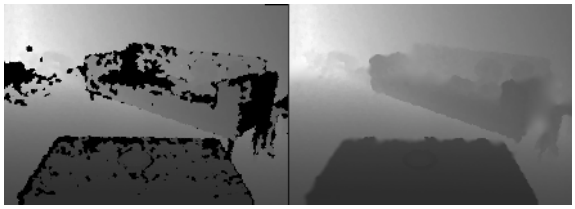

Figure 2. Illustration of our depth smoothing method.

2.2. 3D and 2D/3D Recognition

A comprehensive review of all 3-D features proposed for recognition is beyond the scope of our work. Briefly, notable prominent techniques include spin images [19], 3-D shape context [14], and the recent VFH model [22]–but this list is not exhaustive.

A number of 2D/3D hybrid approaches have been recently proposed, and our dataset should be a relevant testbed for these methods. A multi-modal object detector in which 2D and 3D are traded off in a logistic classifier is proposed by [15]. Their method leverages additional handcrafted feature derived from the 3D observation such as "height above ground" and "surface normal", which provide contextual information. [23] shows how to benefit from 3D training data in a voting based method. Fritz et al. [13] extends branch & bound efficient detection to 3D and adds size and support surface constraints derived from the 3D observation.

Most prominently, a set of methods have been proposed for fusing 2D and 3D information for the task of pedestrian detection. The popular HOG detector [8] to disparity-based features is extended by [17]. A late integration approach is proposed by [21] for combining detectors on the appearance as well as depth image for pedestrian detection. Instead of directly learning on the depth map, [24] uses a depth statistic that learns to enforce height constraints of pedestrians. Finally, [9] explores pedestrian detection by using stereo and temporal information in a hough voting framework also using scene constraints.

3. Our Dataset

We have compiled a large-scale dataset of images taken in domestic and office settings with the commonly available Kinect sensor. The sensor provides a color and depth image pair, and is processed by us for alignment and inpainting. The data was collected by many members of our research community, as well as Amazon Mechanical Turk (AMT) workers, enabling us to have impressive variety in scene and

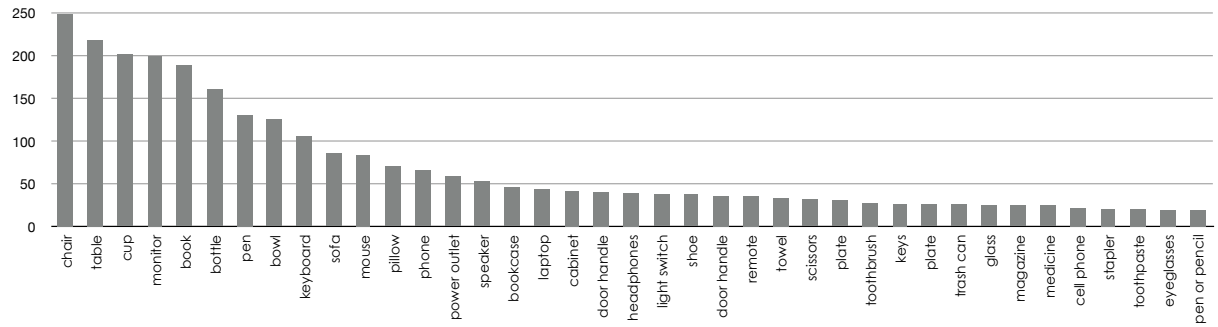

Figure 3. Object frequency for 39 classes with 20 or more examples.

object appearance. As such, the dataset is intended for evaluating approaches to category-level object recognition and localization.

The size of the dataset is not fixed and will continue growing with crowd-sourced submissions. The first release of the dataset contains 849 images taken in 75 different scenes. Over 50 different object classes are represented in the crowd-sourced labels. The annotation is done by Amazon Mechanical Turk workers in the form of bounding boxes on the color image, which are automatically transferred to the depth image.

3.1. Data Collection

We use crowd sourcing on AMT in order to label the data we collect and collect data in addition to the data collected in-house. AMT is a well-known service for "Human Intelligence Tasks" (HIT), which are typically small tasks that are too difficult for current machine intelligence. Our labeling HIT gives workers a list of eight objects to draw bounding boxes around in a color image. Each image is labeled by five workers for each set of labels in order to provide sufficient evidence to determine the validity of a bounding box. A proposed annotation or bounding box is only deemed valid if at least one similarly overlapping bounding box is drawn by another worker. The criteria for similarity of bounding boxes is based on the PASCAL VOC [10] overlap criterion (described in more detail in section 4.1), with the acceptance threshold set to 0.3. If only two bounding boxes are found to be similar, the larger one is chosen. If more than two are deemed similar, we keep the bounding box with the most overlap with the others, and discard the rest.

After the initial intensive in-house data collection, our dataset is now in a sustained effort to collect Kinect data from AMT workers as well as the local community. This AMT collection task is obviously more difficult than simple labeling, since workers must own a Kinect and be able and willing to set their Kinect up properly. Despite this, we were able to collect quality images from AMT workers. With the growing popularity of the Kinect, the pool of potential data contributors keeps getting bigger.

3.2. The Kinect Sensor

The Microsoft Xbox Kinect sensor consists of a horizontal bar with cameras, a structured light projector, an accelerometer and an array of microphones mounted on a motorized pivoting foot. Since its release in November 2010, much open source software has been released allowing the use of the Kinect as a depth sensor [7]. Across the horizontal bar are three sensors: two infrared laser depth sensors with a depth range of approximately 0.6 to 6 meters, and one RGB camera (640 x 480 pixels) [4]. Depth reconstruction uses proprietary technology from Primesense, consisting of continuous infrared structured light projection onto the scene.

The Kinect color and IR cameras are a few centimeters apart horizontally, and have different intrinsic and extrinsic camera parameters, necessitating their calibration for proper registration of the depth and color images. We found that the calibration parameters differ significantly from unit to unit, which poses a problem to totally indiscriminate data collection. Fortunately, the calibration procedure is made easy and automatic due to efforts of the open source community [7, 6].

3.3. Smoothing Depth Images

The structured-light method we use for recovering ground-truth depth-maps necessarily creates areas of the image that lack an estimate of depth. In particular, glass surfaces and infrared-absorbing surfaces can be missing in depth data. Tasks such as getting the average depth of a bounding box, or applying a global descriptor to a part of the depth image therefore benefit from some method for "inpainting" this missing data.

Our view is that proper inpainting of the depth image requires some assumption of the behavior of natural

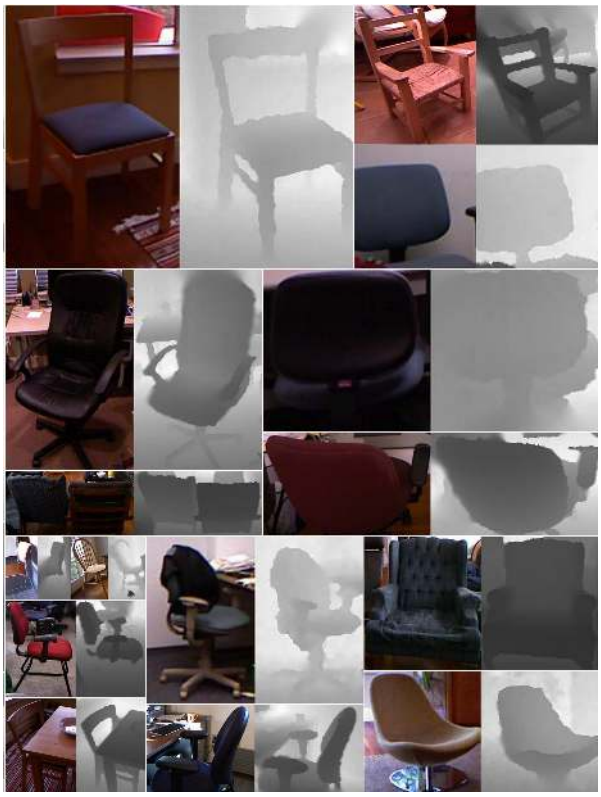

Figure 4. Instances of the "chair" class in our dataset, demonstrating the diversity of object types, viewpoint, and illumination.

shapes. We assume that objects have second order smoothness (that curvature is minimized)—a classic prior on natural shapes [16, 26]. In short, our algorithm minimizes $\|h * Z\|_F^2 + \|h^{\mathrm{T}} * Z\|_F^2$ with the constraints $Z_{x,y} = \hat{Z}_{x,y}$ for all $(x, y) \in \hat{Z}$, where $h = [-1, +2, -1]$, is an oriented 1D discrete Laplacian filter, $*$ is a convolution operation, and $\|\cdot\|_F^2$ is the squared Frobenius norm. The solution to this optimization problem is a depth-map Z in which all observed pixels in $\hat{Z}$ are preserved, and all missing pixels have been filled in with values that minimize curvature in a least-squares sense.

Figure 2 illustrates this algorithm operating on a typical input image with missing depth in our dataset to produce the smoothed output.

3.4. Data Statistics

The dataset described here, which will be able to be used for benchmarking for recognition tasks, is the first release of the B3DO dataset. As our collection efforts are ongoing, subsequent releases of data will include even more variation and larger quantities of data. The distribution of objects in household and office scenes as represented in our dataset is shown in Figure 3. The typical long tail of unconstrained datasets is present, and suggests directions for targeted data collection. At this time, there are 12 classes with more than 70 examples, 27 classes with more than 30 examples, and over 39 classes with 20 or more examples.

Unlike other 3D datasets for object recognition, our dataset features large variability in the appearance of object class instances. This can be seen in Figure 4, presenting random examples of the chair class in our dataset; the variation in viewpoint, distance to object, frequent presence of partial occlusion, and diversity of appearance in this sample poses a challenging detection problem.

The apparent size of the objects in the image, as measured by the bounding box containing them, can vary significantly across the dataset. Our claim is that the real-world size of the objects in the same class varies far less, as can be seen in Figure 5. As proxy for the real-world object size, we use the product of the diagonal of the bounding box l and the distance to the object from the camera D, which is roughly proportional to the world object size by similar triangles (of course, viewpoint variation slightly scatters this distribution–but less so than for the bounding box size). We find that mean smoothed depth is roughly equivalent to the median depth of the depth image ignoring missing data, and so use this to measure distance. The Gaussian was found to be a close fit to these size distributions, allowing us to estimate size likelihood of a bounding box as $\mathcal{N}(x|\mu, \sigma))$, where μ and σ are learned on the training data. This will be used in section 4.3.

4. Detection Baselines

The cluttered scenes of our dataset provide for a challenging object detection task, where the task is to localize all objects of interest in an image. We constrain the task to finding eight different object classes: chairs, monitors, cups, bottles, bowls, keyboards, computer mice, and phones. These object classes were among the most well-represented in our dataset.[1]

4.1. Sliding window detector

Our baseline system is based on a standard detection approach of sliding window classifiers operating on a gradient representation of the image [8, 12, 25]. Such detectors are currently the state of the art on cluttered scene datasets of varied viewpoints and instance types, such as the PASCAL-VOC challenge [10]. The detector considers windows of a fixed aspect ratio across locations and scales of an image pyramid and evaluates them with a score function, outputting detections that score above some threshold.

Specifically, we follow the implementation of the Deformable Part Model detector [12], which uses the LatentSVM formulation $f_\beta(x) = \max_z \beta \cdot \Phi(x, z)$ for scoring candidate windows, where β is a vector of model pa-

[1] We chose not to include a couple of other well-represented classes into this test set because of extreme variation in interpretation of instances of object by the annotators, such as the classes of "table" and "book."

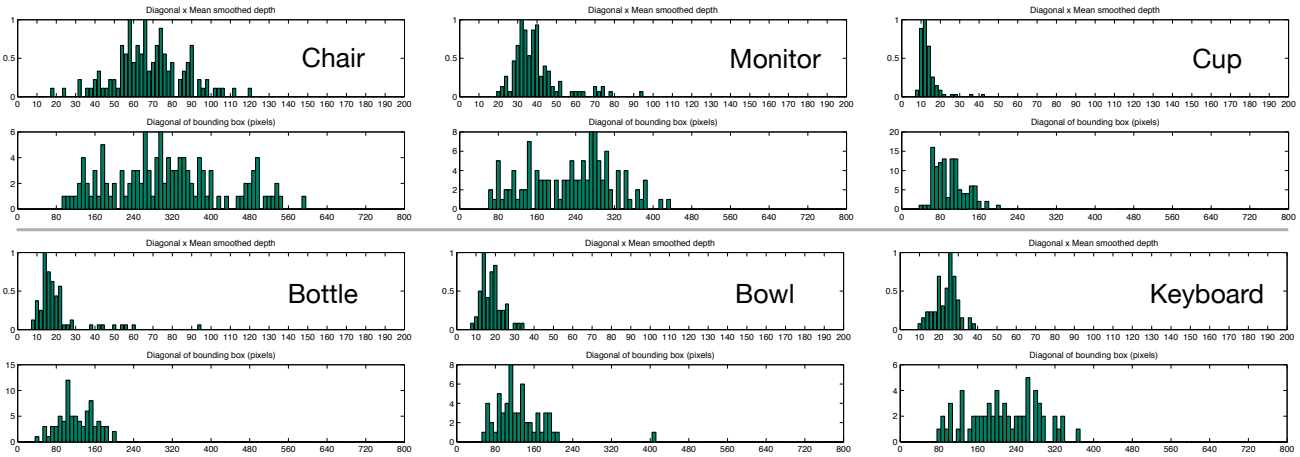

Figure 5. Statistics of object size. For each object class, the top histogram is inferred world object size, obtained as the product of the bounding box diagonal and the average depth of points in the bounding box. The bottom histogram is the distribution of just the diagonal of the bounding box size.

rameters and z are latent values (allowing for part deformations). Optimizing the LatentSVM objective function is a semi-convex problem, and so the detector can be trained even though the latent information is absent for negative examples.

Since finding good negative examples to train on is of paramount importance in a large dataset, the system performs rounds of data mining for small samples of hard negatives, providing a provably exact solution to training on the entire dataset.

To featurize the image, we use a histogram of oriented gradients (HOG) with both contrast-sensitive and contrast-insensitive orientation bins, four different normalization factors, and 8-pixel wide cells. The descriptor is analytically projected to just 31 dimensions, motivated by the analysis in [12].

We explore two feature channels for the detector. One consists of featurizing the color image, as is standard. For the other, we apply HOG to the depth image (Depth HOG), where the intensity value of a pixel corresponds to the depth to that point in space, measured in meters. This application of a gradient feature to depth images has little theoretical justification, since first-order statistics do not matter as much for depth data (this is why we use second-order smoothing in section 3.3). Yet this is an expected first baseline that also forms the detection approach on some other 3D object detection tasks, such as in [20].

Detections are further pruned by non-maximum suppression, which greedily takes the highest-scoring bounding boxes and rejects boxes that sufficiently overlap with an already selected detection. This procedure results in a reduction of detections on the order of ten, and is important for our evaluation metric, which penalizes repeat detections.

4.2. Evaluation

Evaluation of detection is done in the widely adopted style of the PASCAL detection challenge, where a detection is considered correct if $\frac{area(B \cap G)}{area(B \cup G)} > 0.5$ where B is the bounding box of the detection and G is the ground truth bounding box of the same class. Only one detection can be considered correct for a given ground truth box, with the rest considered false positives. Detection performance is represented by precision-recall (PR) curves, and summarized by the area under the curve–the average precision (AP). We evaluate on six different splits of the dataset, averaging the AP numbers across splits.

Our goal is category, not instance-level recognition. As such, it is important to keep instances of a category confined to either training or test set. This makes the recognition task much harder than if we were allowed to train on the same instances of a category as exist in the test set (but not necessarily same the views of them). We enforce this constraint by ensuring that images from the same scene or room are never in both sets. This is a harder constraint than needed, and is not necessarily perfect (for example many different offices might contain the same model laptop), but as there is no realistic way to provide per-instance labeling of a large, crowd-sourced dataset of cluttered scenes, we settle for this method, and keep the problem open for further research.

Figure 6 shows the detector performance on 8 different classes. We note that Depth HOG is never better than HOG on the 2D image. We attribute this to the inappropriateness of a gradient feature on depth data, as mentioned earlier, and to the fact that due to the limitations of the infrared structured light depth reconstruction, some objects tend to be missing depth data.

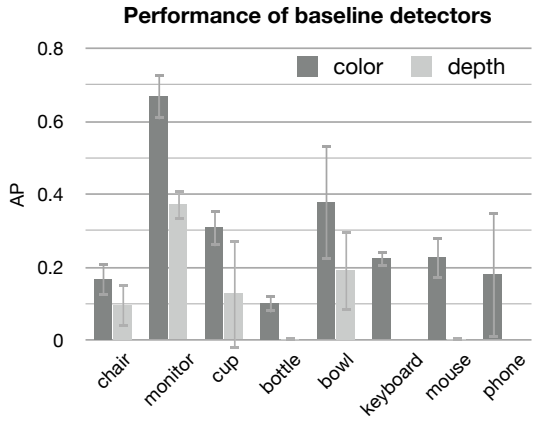

Figure 6. Performance of the baseline detector on our dataset, as measured by the average precision. Depth HOG fails completely on some categories, for reasons explained in the text.

4.3. Pruning and rescoring by size

In section 3.4, we made the observation that true object size, even as approximated by the product of object projection in the image and median depth of its bounding box, varies less than bounding box size. We therefore investigate two ways of using approximated object size as an additional source of discriminative signal to the detector.

Our first way of using size information consists of pruning candidate detections that are sufficiently unlikely given the size distribution of that object class. The object size distribution is modeled with a Gaussian, which we found is a close fit to the underlying distribution; the Gaussian parameters are estimated on the training data only. We prune boxes that are more than $\sigma = 3$ standard deviations away from the mean of the distribution.

Figure 7 shows that the pruning results provide a boost in detection performance, while rejecting from 12% to 68% of the suggested detection boxes (on average across the classes, 32% of candidate detections are rejected). This observation can be leveraged as part of an "objectness" filter or as a thresholding step in a cascaded implementation of this detector for detection speed gain [5, 11]. The classes chair and mouse are the two classes most helped by size pruning, while monitors and bottle are the least helped.

Using bounding box size of the detection (as measured by its diagonal) instead of inferred world size results in no improvement to AP performance on average. Two classes that are most hurt are bowl and plate; two that are least hurt by the bounding box size pruning are bottle and mouse.

The second way we use size information consists of learning a rescoring function for detections, given their SVM score and size likelihood. We learn a simple combination of the two values:

$$s(x) = \exp(\alpha \log(w(x)) + (1-\alpha)\log(\mathcal{N}(x|\mu,\sigma))) \quad (1)$$

where $w(x) = 1/(1+\exp(-2f_\beta(x)))$ is the normalized SVM score, $\mathcal{N}(x|\mu,\sigma))$ is the likelihood of the inferred world size of the detection under the size distribution of the object class, and α is a parameter learned on the training set. This corresponds to unnormalized Naive Bayes combination of the SVM model likelihood and object size likelihood. Since what matters for the precision-recall evaluation is the ordering of confidences and whether they are normalized is irrelevant, we are able to evaluate $s(x)$.

As Figure 7 demonstrates, the rescoring method works better than pruning. This method is able to boost recall as well as precision by assigning a higher score to likely detections in addition to lowering the score (which is, in effect, pruning) of unlikely detections.

5. Discussion

We presented a novel paradigm for crowd-sourced data collection that leverages the success of the Kinect depth sensor. Its popularity has been encouraging, and we think it is time to "put the Kinect to work" for computer vision. The main contribution of this paper is a novel category-level object dataset, which presents a challenging task and is far beyond existing 3-D datasets in terms of the number of object categories, the number of examples per category, and intra-category variation. Importantly, the dataset poses the problem of object detection "in the wild", in real rooms in people's homes and offices, and therefore has many practical applications.

References

[1] Ford campus vision and lidar dataset. http://robots.engin.umich.edu/Downloads. 2

[2] Solution in perception challenge. http://opencv.willowgarage.com/wiki/SolutionsInPerceptionChallenge. 2

[3] UBC Robot Vision Survey. http://www.cs.ubc.ca/labs/lci/vrs/index.html. 2

[4] Introducing Kinect for Xbox 360. http://www.xbox.com/en-US/Kinect/, 2011. 3

[5] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 6

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 3

[7] N. Burrus. Kinect RGB Demo v0.4.0. http://nicolas.burrus.name/index.php/Research/KinectRgbDemoV4?from=Research.KinectRgbDemoV2, Feb. 2011. 3

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005. 2, 4

[9] A. Ess, K. Schindler, B. Leibe, and L. V. Gool. Object detection and tracking for autonomous navigation in dynamic environments. *International Journal on Robotics Research*, 2010. 2

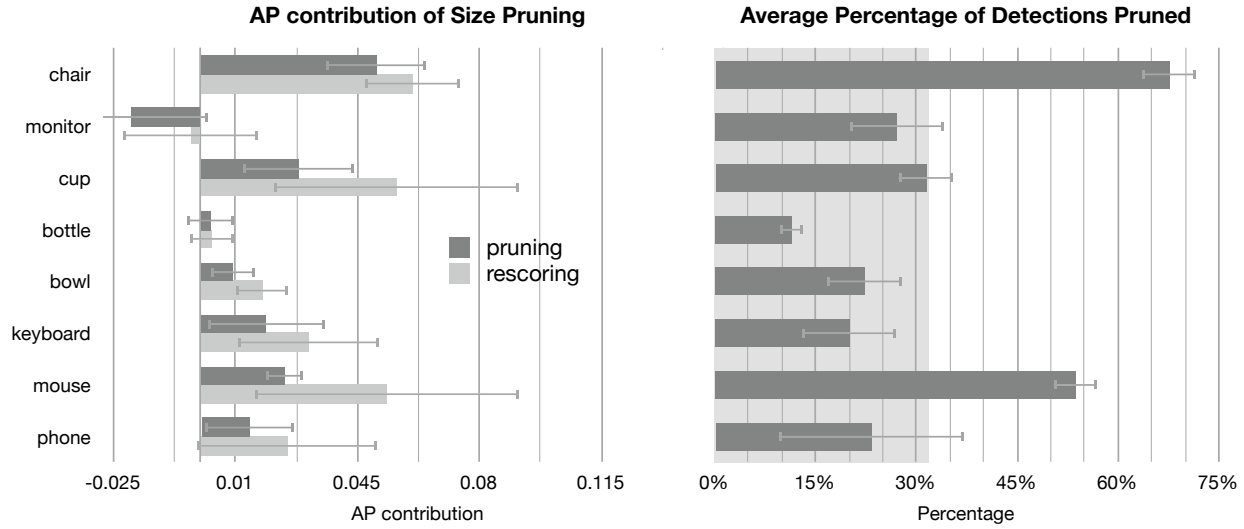

Figure 7. Left: Effect on the performance of our detector shown by the two uses of object size we consider. Right: Average percentage of past-threshold detections pruned by considering the size of the object. The light gray rectangle reaching to 32% is the average across classes. In both cases, error bars show standard deviation across six different splits of the data.

[10] 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. 3, 4

[11] P. F. Felzenszwalb, R. B. Girschick, and D. McAllester. Cascade object detection with deformable part models. *CVPR*, Mar 2010. 6

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, Jul 2009. 4, 5

[13] M. Fritz, K. Saenko, and T. Darrell. Size matters: Metric visual search constraints from monocular metadata. In *Advances in Neural Information Processing Systems 23*, 2010. 2

[14] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *ECCV04*, pages Vol III: 224–237, 2004. 2

[15] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. 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. 2

[16] W. Grimson. *From images to surfaces: A computational study of the human early visual system*. MIT Press, 1981. 4

[17] H. Hattori, A. Seki, M. Nishiyama, and T. Watanabe. Stereo-based pedestrian detection using multiple patterns. In *Proceedings of British Machine Vision Conference*, 2009. 2

[18] S. Helmer, D. Meger, M. Muja, J. J. Little, and D. G. Lowe. Multiple viewpoint recognition and localization. In *Proceedings of Asian Conferene on Computer Vision*, 2010. 2

[19] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5):433 –449, May 1999. 2

[20] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. *ICRA*, Feb 2011. 2, 5

[21] M. Rohrbach, M. Enzweiler, and D. M. Gavrila. High-level fusion of depth and intensity for pedestrian classification. In *Annual Symposium of German Association for Pattern Recognition (DAGM)*, 2009. 2

[22] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 10/2010 2010. 2

[23] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *ECCV*, 2010. 2

[24] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. In *ECCV*, 2010. 2

[25] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. *ICCV*, Jul 2009. 4

[26] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *PAMI*, 2009. 4