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

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

Introduction

- Recently, there has been a resurgence
  techniques due to advances in active
- including techniques based on LIDAR, t
  projected texture stereo (PR2). The Prim
  Microsoft Kinect(8 million Kinects were s
- While there is a large literature on instar
  the computer vision and robotics literat
  existing datasets for category-level 3-D
- Numerous schemes for defining 3-D fea
  features for category-level recognition
  perform in uncluttered domains

Figure 1. Two scenes typical of our dataset.

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

Related Work

- **RGBD-dataset of [21]:** This dataset from Intel Research and UW features 300 objects in 51 categories 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, 19]:** 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.

Related Work

- **3D table top object dataset [24]:** 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.

Related Work

- **Other datasets :** they cannot be leveraged for the mult localization task that is our goal.
- **Our dataset :** our dataset contains both a large numbe many different instances per category, is photographe instead of in a controlled turntable setting . presents a examples of the "chair" category in our dataset. These dataset more representative of the kind of data that c in people's homes

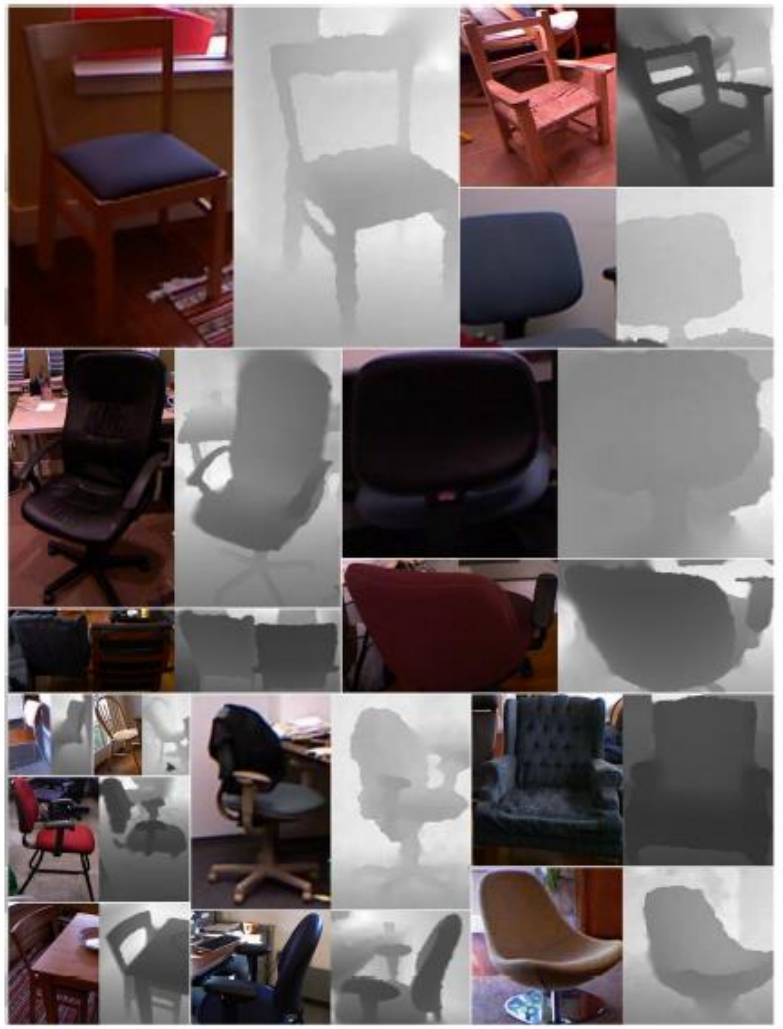

Figure 4. Instances of the "chair" class in our dataset, demonstrating the diversity of object types, viewpoint, and illumination.

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

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
- 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

Data Collection

- 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
- 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

The Kinect Sensor

- Since its release in November 2010, much open source software has been released allowing the use of the Kinect as a depth sensor
- two infrared laser depth sensors with a depth range of approximately 0.6 to 6 meters
- one RGB camera (640 x 480 pixels)
- Depth reconstruction uses proprietary technology from Primesense, consisting of continuous infrared structured light projection onto the scene

Smoothing Depth Images

- In particular, glass surfaces and infrared-absorbing surfaces can be missing in depth data

$$\|h * Z\|_F^2 + \|h^{\mathrm{T}} * Z\|_F^2 \text{ with the constraints } Z_{x,y} = \hat{Z}_{x,y}$$

$$\text{for all } (x, y) \in \hat{Z}$$

- where h = [-1, +2,-1], is an oriented1D discrete Laplacian filter
- * is a convolution operation
- $\|\cdot\|_F^2$ is the squared Frobenius norm $\|A\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2$

Smoothing Depth Images

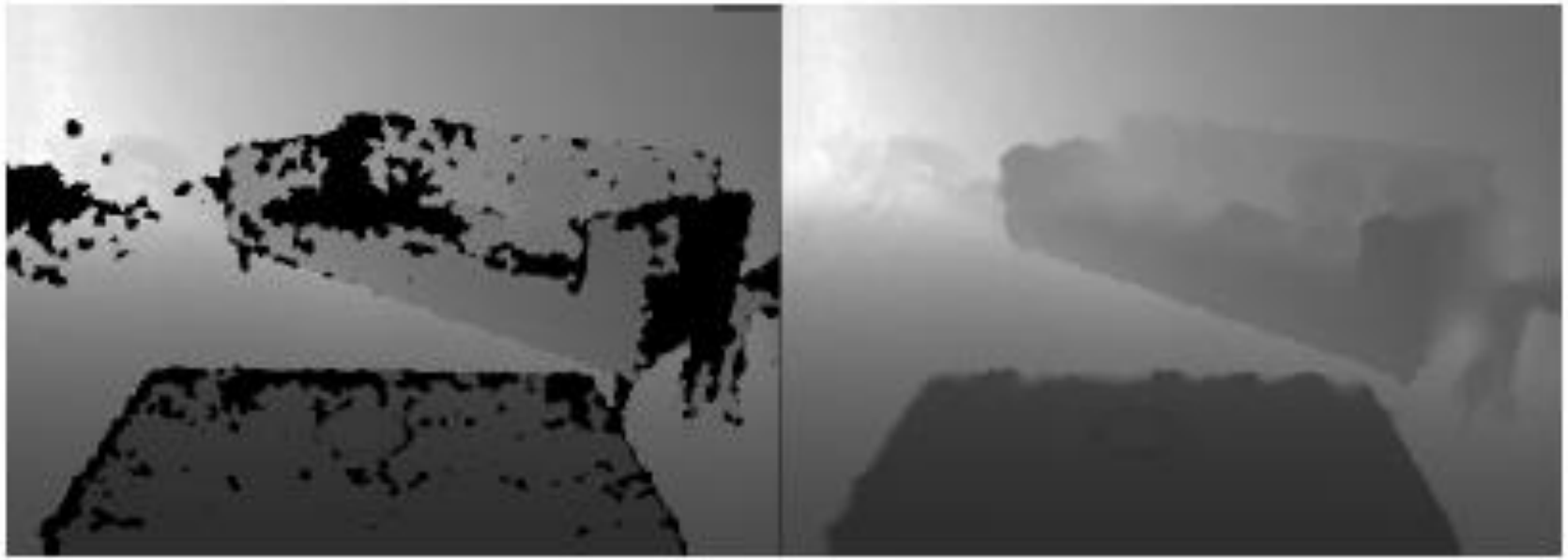

Figure 2. Illustration of our depth smoothing method.

Data Statistics

- 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

Data Statistics

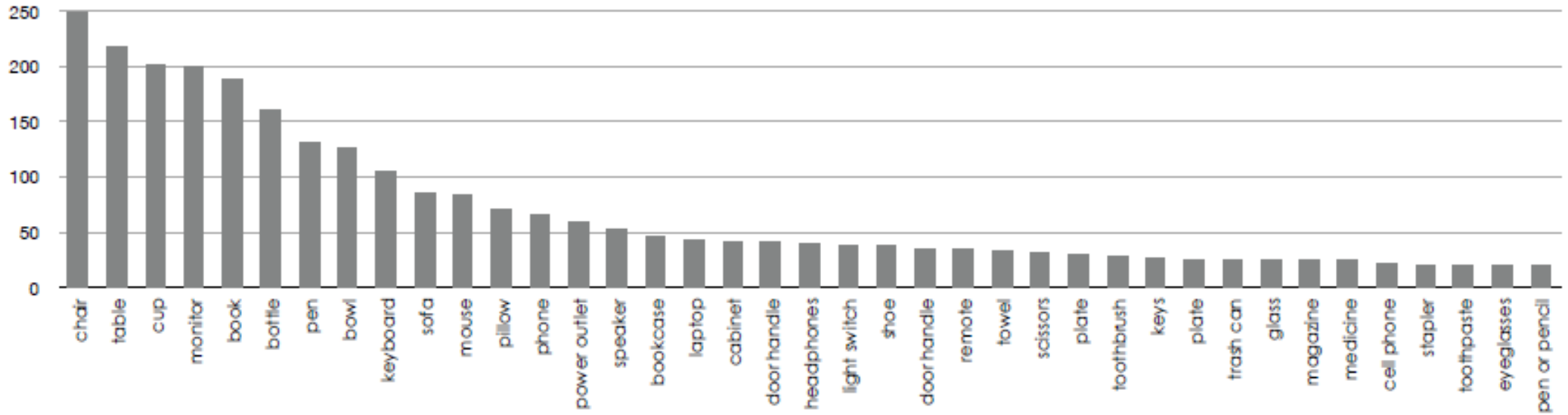

Figure 3. Object frequency for 39 classes with 20 or more examples.

Data Statistics

- 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
- 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
- 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

Data Statistics

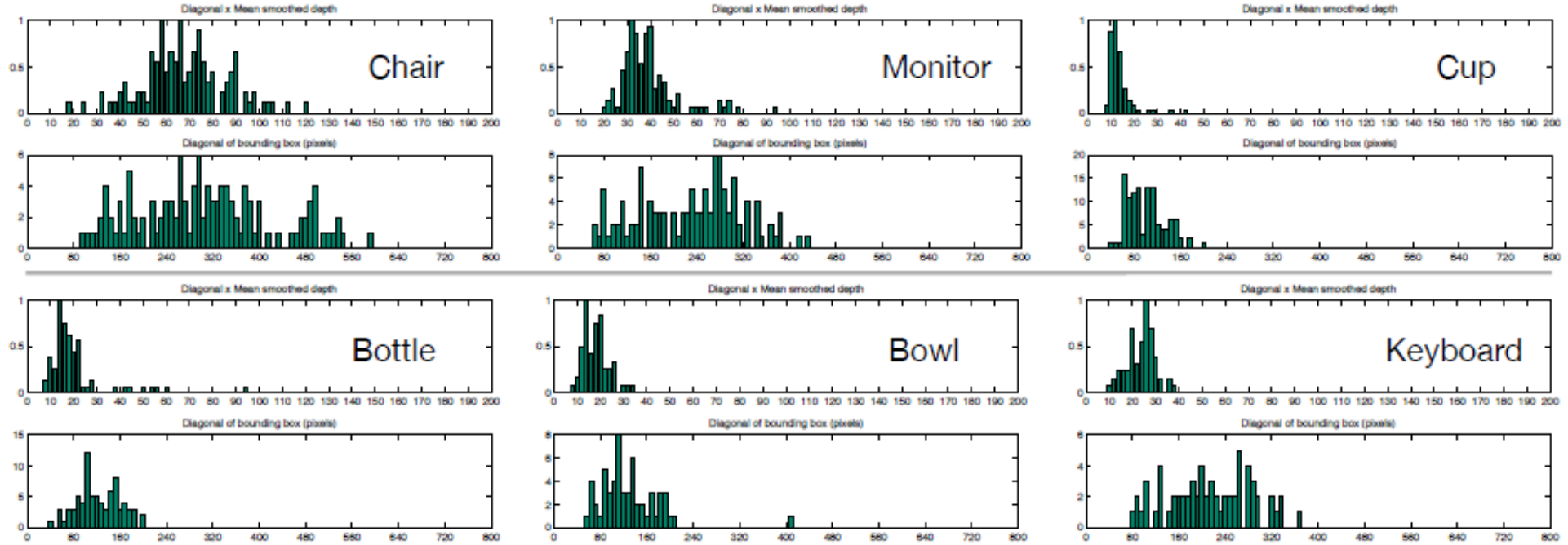

Figure 5. Statistics of object size. For each object class, the top histogram is inferred world object size, obtained as the product of the bounding box diagonal and the average depth of points in the bounding box. The bottom histogram is the distribution of just the diagonal of the bounding box size.

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

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
- we follow the implementation of the Deformable Part Model detector
- 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 parameter and z are latent values
- 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)

Evaluation

- 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
- 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

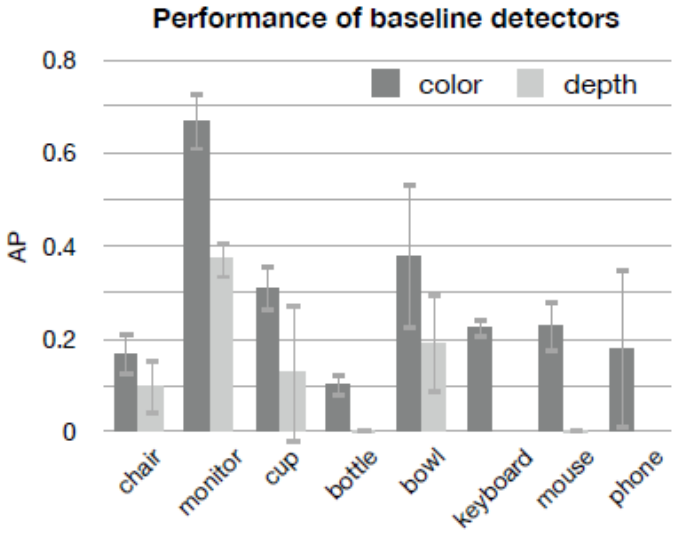

Figure 6. Performance of the baseline detector on our dataset, as measured by the average precision. Depth HOG fails completely on some categories, for reasons explained in the text.

Pruning and rescoring by size

- We therefore investigate two ways of using approximated object size as an additional source of discriminative signal to the detector
- Method 1 :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
- Method 2 :we use size information consists of learning a rescoring function for detections, given their SVM score and size likelihood

rescoring by size

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

$$w(x) = 1/(1 + \exp(-2 f_\beta(x)))$$ is 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

Pruning and rescoring by size

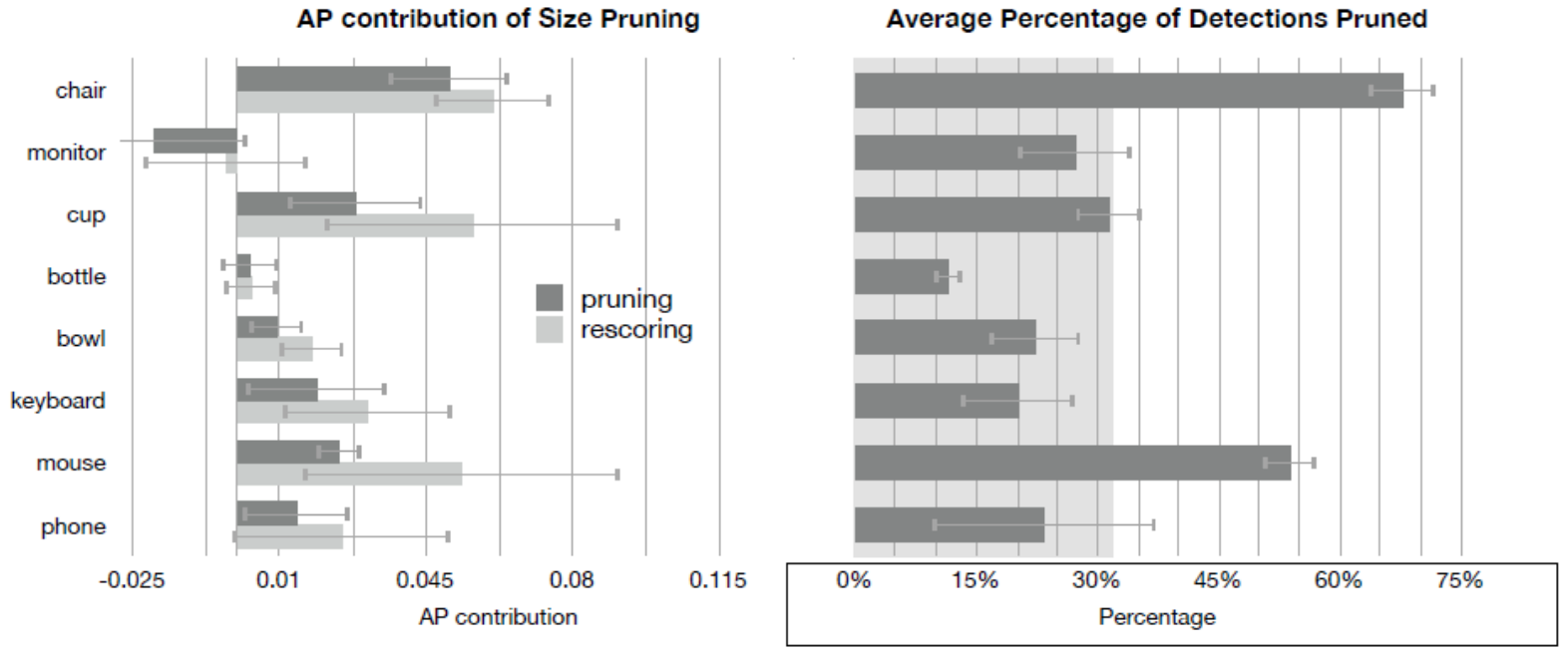

Figure 7. Left: Effect on the performance of our detector shown by the two uses of object size we consider. Right: Average percentage of past-threshold detections pruned by considering the size of the object. The light gray rectangle reaching to 32% is the average across classes. In both cases, error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

outline

- Introduction
- Related Work
- Our Dataset
- Detection Baselines
- Discussion

Discussion

- Its popularity has been encouraging, and we think it is time to "put the Kinect to work" for computer vision
- 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