Occlusion Reasoning for Object Detection under Arbitrary Viewpoint

Edward Hsiao and Martial Hebert
Robotics Institute, Carnegie Mellon University, USA
{ehsiao,hebert}@cs.cmu.edu

Abstract

We present a unified occlusion model for object instance detection under arbitrary viewpoint. Whereas previous approaches primarily modeled local coherency of occlusions or attempted to learn the structure of occlusions from data, we propose to explicitly model occlusions by reasoning about 3D interactions of objects. Our approach accurately represents occlusions under arbitrary viewpoint without requiring additional training data, which can often be difficult to obtain. We validate our model by extending the state-of-the-art LINE2D method for object instance detection and demonstrate significant improvement in recognizing texture-less objects under severe occlusions.

1. Introduction

Occlusions are common in real world scenes and are a major obstacle to robust object detection. While texture-rich objects can be detected under severe occlusions with distinctive local features, such as SIFT [13], many man-made objects have large uniform regions. These texture-less objects are characterized by their contour structure, which are often ambiguous even without occlusions. Instance detection of texture-less objects compounds this ambiguity by requiring recognition under arbitrary viewpoint with severe occlusions as shown in Figure 1. While much research has addressed each component separately (texture-less objects [21], arbitrary viewpoint [7], occlusions [4]), addressing them together is extremely challenging. The main contributions of this paper are (1) a concise model of occlusions under arbitrary viewpoint without requiring additional training data and (2) a method to capture global visibility relationships without combinatorial explosion.

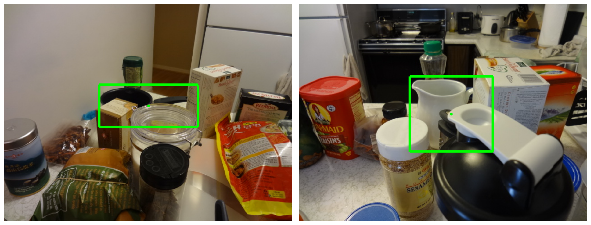


Figure 1: Example detections of (*left*) saucepan and (*right*) pitcher under severe occlusions.

In the past, occlusion reasoning for object detection has been extensively studied [6, 15, 17]. One common approach is to model occlusions as regions that are inconsistent with object statistics. Girshick *et al.* [5] use an occluder part in their grammar model when all parts cannot be placed. Wang *et al.* [22] use the scores of individual HOG filter cells, while Meger *et al.* [14] use depth inconsistency from 3D sensor data to classify occlusions. Local coherency of occlusions are often enforced with a Markov Random Field [3] to reduce noise in these classifications.

While assuming any inconsistent region is an occlusion is valid if occlusions happen uniformly over an object, it ignores the fact there is structure to occlusions for many objects. For example, in real world environments, objects are usually occluded by other objects resting on the same surface. Thus it is often more likely for the bottom of an object to be occluded than the top of an object [2].

Recently, researchers have attempted to learn the structure of occlusions from data [4, 10]. With enough data, these methods can learn an accurate model of occlusions. However, obtaining a broad sampling of occluder objects is usually difficult, resulting in biases to the occlusions of a particular dataset. This becomes more problematic when considering object detection under arbitrary view [7, 18, 20]. Learning approaches need to learn a new model for each view of an object. This is intractable, especially when recent studies [7] have claimed that approximately 2000 views are needed to sample the view space of an object. A key contribution of our approach is to represent occlusions under arbitrary viewpoint without requiring additional training data of occlusions. We demonstrate that our approach accurately models occlusions, and that learning occlusions from data does not give better performance.

Researchers have shown in the past that incorporating 3D geometric understanding of scenes [1, 9] improves the performance of object detection systems. Following these approaches, we propose to reason about occlusions by explicitly modeling 3D interactions of objects. For a given environment, we compute physical statistics of objects in

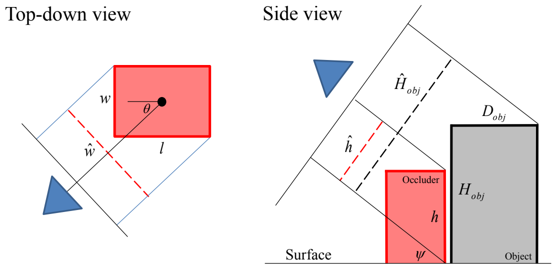


Figure 2: Occlusion model. (*left*) Projected width of occluder, $\hat{w}$, for a rotation of θ. (*right*) Projected height of occluder, $\hat{h}$, and projected height of object, $\hat{H}_{obj}$, for an elevation angle of ψ. An occluder needs a projected height of $\hat{h} \geq \hat{H}_{obj}$ to fully occlude the object.

the scene and represent an occluder as a probabilistic distribution of 3D blocks. The physical statistics need only be computed once for a particular environment and can be used to represent occlusions for many objects in the scene. By reasoning about occlusions in 3D, we effectively provide a unified occlusion model for different viewpoints of an object as well as different objects in the scene.

We incorporate occlusion reasoning with object detection by: (1) a bottom-up stage which hypothesizes the likelihood of occluded regions from the image data, followed by (2) a top-down stage which uses prior knowledge represented by the occlusion model to score the plausibility of the occluded regions. We combine the output of the two stages into a single measure to score a candidate detection.

The focus of this paper is to demonstrate that a relatively simple model of 3D interaction of objects can be used to represent occlusions effectively for instance detection of texture-less objects under arbitrary view. Recently, there has been significant progress in simple and efficient template matching techniques [7, 8] for instance detection. These approaches work extremely well when objects are largely visible, but degrade rapidly when faced with strong occlusions in heavy background clutter. We evaluate our approach by extending the state-of-the-art LINE2D [7] system, and demonstrate significant improvement in detection performance on a challenging occlusion dataset.

2. Occlusion Model

Occlusions in real world scenes are often caused by a solid object resting on the same surface as the object of interest. In our model, we approximate occluding objects by their 3D bounding box and demonstrate how to compute occlusion statistics of an object under different camera viewpoints, c, defined by an elevation angle ψ and azimuth θ.

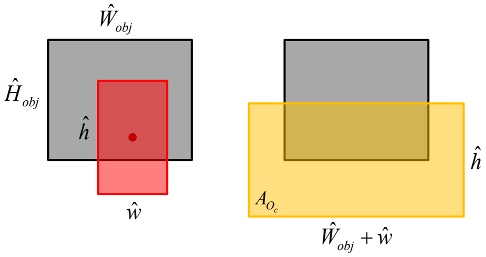


Figure 3: Computation of A_{O_c}. (*left*) We consider the center positions of a block (red) which occlude the object. The base of the block is always below the object, since we assume they are on the same surface. (*right*) The set of positions is defined by the yellow rectangle which has area A_{O_c}.

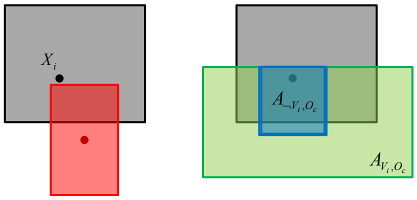


Figure 4: Computation of A_{V_i,O_c}. (*left*) We consider the center positions of a block (red) which occlude the object while keeping X_i visible. (*right*) The set of positions is defined by the green region which has area A_{V_i,O_c}.

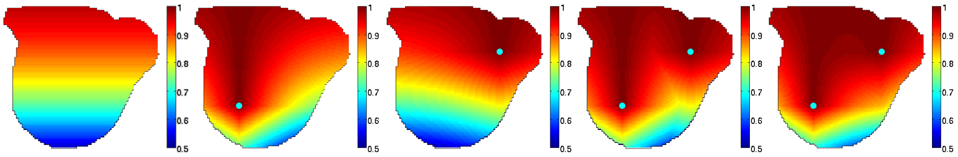
Let $\mathcal{X} = \{X_1, ..., X_N\}$ be a set of N points on the object with their visibility states represented by a set of binary variables $\mathcal{V} = \{V_1, ..., V_N\}$ such that if $V_i = 1$, then X_i is visible. For occlusions O_c under a particular camera viewpoint c, we want to compute occlusion statistics for each point in $\mathcal{X}$. Unlike other occlusion models which only compute an occlusion prior $P(V_i|O_c)$, we propose to also model the global relationship between visibility states, $P(V_i|\mathcal{V}_{\text{-}i}, O_c)$ where $\mathcal{V}_{\text{-}i} = \mathcal{V} \backslash V_i$. Through our derivation, we observe that $P(V_i|O_c)$ captures the classic intuition that the bottom of the object is more likely to be occluded than the top. More interesting is $P(V_i|\mathcal{V}_{\text{-}i}, O_c)$ which captures the structural layout of an occlusion. The computation of these two occlusion properties both reduce to integral geometry [16], and for explanation, we illustrate the derivation of the occlusion prior $P(V_i|O_c)$.

We make a couple of approximations to tractably derive the occlusion statistics. Specifically, since objects which occlude each other are usually physically close together, we approximate the objects to be on the same support surface and we approximate the perspective effects over the range of object occlusions to be negligible.

2.1. Representation under different viewpoints

The likelihood that a point on an object is occluded depends on the angle the object is being viewed from. Most methods that learn the structure of occlusions from data [4] require a separate occlusion model for each view of every object. These methods do not scale well when considering detection of many objects under arbitrary view.

In the following, we propose a unified representation of occlusions under arbitrary viewpoint of an object. Our



(a) (b) (c) (d) (e)

Figure 5: Example of (*a*) occlusion prior $P(V_i|O_c)$, (*b,c*) conditional likelihood $P(V_i|V_j, O_c)$ and $P(V_i|V_k, O_c)$ given two separate points X_j and X_k individually, (*d*) approximate conditional likelihood $P(V_i|V_j, V_k, O_c)$ from Equation 12, and (*e*) explicit conditional likelihood $P(V_i|V_j, V_k, O_c)$ from Equation 10.

method requires only the statistics of object dimensions, which is obtained once for a given environment and can be shared across many objects for that environment.

The representation we propose is illustrated in Figure 2. For a specific viewpoint, we represent the portion of a block that can occlude the object as a bounding box with dimensions corresponding to the projected height $\hat{h}$ and the projected width $\hat{w}$ of the block. The projected height and width are the observed height and width of a block to the viewer.

The object of interest, on the other hand, is represented by its silhouette in the image. Initially, we derive our model using the bounding box of the silhouette with dimensions $\hat{H}_{obj}$ and $\hat{W}_{obj}$, and then relax our model to use the actual silhouette (Section 2.4).

First, we compute the projected width $\hat{w}$ of an occluder with width w and length l as shown by the top-down view in Figure 2. In our convention, $\hat{w} = w$ for an azimuth of $\theta = 0$. Using simple geometry, the projected width is,

$$\hat{w}(\theta) = w \cdot |\cos\theta| + l \cdot |\sin\theta|. \tag{1}$$

Since θ is unknown for an occluding object, we obtain a distribution of $\hat{w}$ assuming all rotations about the vertical axis are equally likely. The distribution of $\hat{w}$ over $\theta \in [0, 2\pi]$ is equivalent to the distribution over any $\frac{\pi}{2}$ interval. Thus, the distribution of $\hat{w}$ is computed by transforming a uniformly distributed random variable on $[0, \frac{\pi}{2}]$ by Equation 1. The resulting probability density of $\hat{w}$ is given by,

$$p_{\hat{w}}(\hat{w}) = \begin{cases} \frac{2}{\pi}\left(1 - \frac{\hat{w}^2}{w^2+l^2}\right)^{-\frac{1}{2}} & w \le \hat{w} < l \\ \frac{4}{\pi}\left(1 - \frac{\hat{w}^2}{w^2+l^2}\right)^{-\frac{1}{2}} & l \le \hat{w} < \sqrt{w^2+l^2}. \end{cases} \tag{2}$$

Next, we compute the projected height $\hat{h}$ of an occluder as illustrated by the side view of Figure 2. For an elevation angle ψ and occluding block with height h, the projected height $\hat{h}$ is

$$\hat{h}(\psi) = h \cdot \cos\psi. \tag{3}$$

This corresponds to the maximum height that can occlude the object given our assumptions.

The projected height of the object, $\hat{H}_{obj}$, is slightly different in that it accounts for the apparent height of the object silhouette. An object is fully occluded vertically only if $\hat{h} \ge \hat{H}_{obj}$. To compute $\hat{H}_{obj}$, we need the distance, D_{obj}, from the closest edge to the farthest edge of the object. Following the computation of the projected width $\hat{w}$, we have $D_{obj}(\theta) = W_{obj} \cdot |\sin\theta| + L_{obj} \cdot |\cos\theta|$. The projected height of the object at an elevation angle ψ is then given by,

$$\hat{H}_{obj}(\theta, \psi) = H_{obj} \cdot |\cos\psi| + D_{obj}(\theta) \cdot |\sin\psi|. \tag{4}$$

Finally, the projected width of the object $\hat{W}_{obj}$ is computed using the aspect ratio of the silhouette bounding box.

2.2. Occlusion Prior

Given the representation derived in Section 2.1, we want to compute a probability for a point on the object being occluded. Many systems which attempt to address occlusions assume that they occur randomly and uniformly across the object. However, recent studies [2] have shown that there is structure to occlusions for many objects.

We begin by deriving the occlusion prior using an occluding block with projected dimensions $(\hat{w}, \hat{h})$ and then extend the formulation to use a probabilistic distribution of occluding blocks. The occlusion prior specifies the probability $P(V_i|O_c)$ that a point on the object $X_i = (x_i, y_i)$ is visible given an occlusion of the object. This involves estimating the area, A_{O_c}, covering the set of block positions that occlude the object (shown by the yellow region in Figure 3), and estimating the area, A_{V_i,O_c}, covering the set of block positions that occlude the object while keeping X_i visible (shown by the green region in Figure 4). The occlusion prior is then just a ratio of these two areas,

$$P(V_i|O_c) = \frac{A_{V_i,O_c}}{A_{O_c}}. \tag{5}$$

From Figure 3, a block (red) will occlude the object if its center is inside the yellow region. The area of this region, A_{O_c}, is

$$A_{O_c} = (\hat{W}_{obj} + \hat{w}) \cdot \hat{h}. \tag{6}$$

Next, from Figure 4, this region can be partitioned into a region where the occluding block occludes X_i (blue) and a region which does not (green). A_{V_i,O_c} corresponds to the area of the green region and can be computed as

$$A_{V_i,O_c} = \hat{W}_{obj} \cdot \hat{h} + \hat{w} \cdot \min(\hat{h}, y_i). \tag{7}$$

Now that we have derived the occlusion prior using a particular occluding block, we extend the formulation to a distribution of blocks. Let $p_{\hat{w}}(\hat{w})$ and $p_{\hat{h}}(\hat{h})$ be distributions of $\hat{w}$ and $\hat{h}$ respectively. To simplify notation, we define $\mu_{\hat{w}} = \mathbb{E}_{p_{\hat{w}}(\hat{w})}[\hat{w}]$ and $\mu_{\hat{h}} = \mathbb{E}_{p_{\hat{h}}(\hat{h})}[\hat{h}]$ to be the expected width and height of the occluders under these distributions, and define $\beta_y(y_i) = \int \min(\hat{h}, y_i) \cdot p_{\hat{h}}(\hat{h})\, d\hat{h}$. The average areas, A_{O_c} and A_{V_i,O_c}, are then given by

$$A_{O_c} = (\hat{W}_{obj} + \mu_{\hat{w}}) \cdot \mu_{\hat{h}}, \tag{8}$$

$$A_{V_i,O_c} = \hat{W}_{obj} \cdot \mu_{\hat{h}} + \mu_{\hat{w}} \cdot \beta_y(y_i). \tag{9}$$

This derivation assumes that the distribution $p_{\hat{w},\hat{h}}(\hat{w}, \hat{h})$ can be separated into $p_{\hat{w}}(\hat{w})$ and $p_{\hat{h}}(\hat{h})$. For household objects, we empirically verified that this approximation holds. In practice, the areas are computed by discretizing the distributions and Figure 5(a) shows an example occlusion prior.

2.3. Occlusion Conditional Likelihood

Most occlusion models only account for local coherency and the prior probability that a point on the object is occluded. Ideally, we want to compute a global relationship between all visibility states $\mathcal{V}$ on the object. While this is usually infeasible combinatorially, we show how a tractable approximation can be derived in the following section.

Let $\mathcal{X}_{\mathcal{V}_{-i}}$ be the visible subset of $\mathcal{X}$ according to $\mathcal{V}_{-i}$. We want to compute the probability $P(V_i|\mathcal{V}_{-i}, O_c)$ that a point X_i is visible given the visibility of $\mathcal{X}_{\mathcal{V}_{-i}}$. Following Section
2.2, the conditional likelihood is given by

$$P(V_i|\mathcal{V}_{-i}, O_c) = \frac{A_{V_i,\mathcal{V}_{-i},O_c}}{A_{\mathcal{V}_{-i},O_c}}. \tag{10}$$

We first consider the case where we condition on one visible point, X_j (i.e., $\mathcal{X}_{\mathcal{V}_{-i}} = \{X_j\}$). To compute $P(V_i|V_j, O_c)$, we already have A_{V_j,O_c} from Equation 9, so we just need A_{V_i,V_j,O_c}. The computation follows from Section 2.2, so we omit the details and just provide the results below. If we let $\beta_x(x_i, x_j) = \int \min(\hat{w}, |x_i - x_j|) \cdot p_{\hat{w}}(\hat{w})\, d\hat{w}$, then

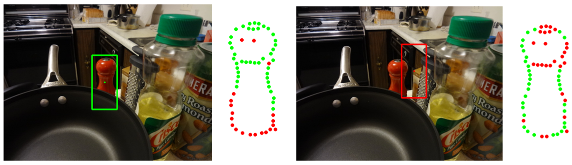
$$\begin{aligned} A_{V_i,V_j,O_c} &= (\hat{W}_{obj} - |x_i - x_j|) \cdot \mu_{\hat{h}} + \\ &\left(\int_0^{|x_i - x_j|} (|x_i - x_j| - \hat{w}) \cdot p_{\hat{w}}(\hat{w})\, d\hat{w} \right) \cdot \mu_{\hat{h}} + \\ &\beta_x(x_i, x_j) \cdot \beta_y(y_i) + \mu_{\hat{w}} \cdot \beta_y(y_j). \end{aligned} \tag{11}$$

We can generalize this to k visible points (i.e., $|\mathcal{X}_{\mathcal{V}_{-i}}| = k$) by counting as above, however, the number of cases increases combinatorially. We make the approximation that the point $X_j \in \mathcal{X}_{\mathcal{V}_{-i}}$ with the highest conditional likelihood $P(V_i|V_j, O_c)$ provides all the information about the visibility of X_i. This observation assumes that $V_i \perp \{\mathcal{V}_{-i} \backslash V_j\} | V_j$ and allows us to compute the global visibility relationship $P(V_i|\mathcal{V}_{-i}, O_c)$ without combinatorial explosion. The approximation of $P(V_i|\mathcal{V}_{-i}, O_{-i})$ is then

$$P(V_i|\mathcal{V}_{-i}, O_c) \approx P(V_i|V_j^*, O_c), \tag{12}$$

$$V_j^* = \operatorname*{argmax}_{V_j \in \mathcal{V}_{-i}} P(V_i|V_j, O_c). \tag{13}$$

For example, Figure 5(d,e) shows the approximate conditional likelihood and the exact one for $|\mathcal{X}_{\mathcal{V}_{-i}}| = 2$.



(a) True positive

(b) False positive

Figure 6: Examples of occlusion hypotheses. (*a*) For a true detection, the occluded points (red) are consistent with our model. (*b*) For a false positive, the top of the object is hypothesized to be occluded while the bottom is visible, which is highly unlikely according to our model.

2.4. Arbitrary object silhouette

The above derivation can easily be relaxed to use the actual object silhouette. The idea is to subtract the area, A_s, covering the set of block positions that occlude the object bounding box but not the silhouette from the areas described in Sections 2.2 and 2.3. The occlusion prior and conditional likelihood are then given by,

$$P(V_i|O_c) = \frac{A_{V_i,O_c} - A_s}{A_{O_c} - A_s}, \tag{14}$$

$$P(V_i|\mathcal{V}_{-i}, O_c) = \frac{A_{V_i,\mathcal{V}_{-i},O_c} - A_s}{A_{\mathcal{V}_{-i},O_c} - A_s}. \tag{15}$$

3. Object Detection

Given our occlusion model from Section 2, we augment an object detection system by (1) a bottom-up stage which hypothesizes occluded regions using the object detector, followed by (2) a top-down stage which measures the consistency of the hypothesized occlusion with our model. We explore using the occlusion prior and occlusion conditional likelihood for scoring and show in our evaluation that both are informative for object detection.

3.1. Occlusion Hypothesis

We follow previous methods [3, 5, 22] and consider regions that do not match well with the object statistics to be occluded. Hypothesizing these regions depends on the individual object detector. For HOG, Wang *et al.* [22] use the score of individual filter cells to classify occlusions. In the following, we propose a similar approach for LINE2D.

The LINE2D method represents an object by a template of sampled edge points, each with a quantized orientation. For every scanning window location, a similarity score is computed between the gradient of each model point and the image. We consider a point to be occluded if the image gradient and the model gradient have different quantized orientations. Figure 6 shows example occlusion hypotheses.

3.2. Occlusion Scoring

Given the hypothesized visibility labeling $\mathcal{V}$ for a detection window Z from Section 3.1, we want a metric of how

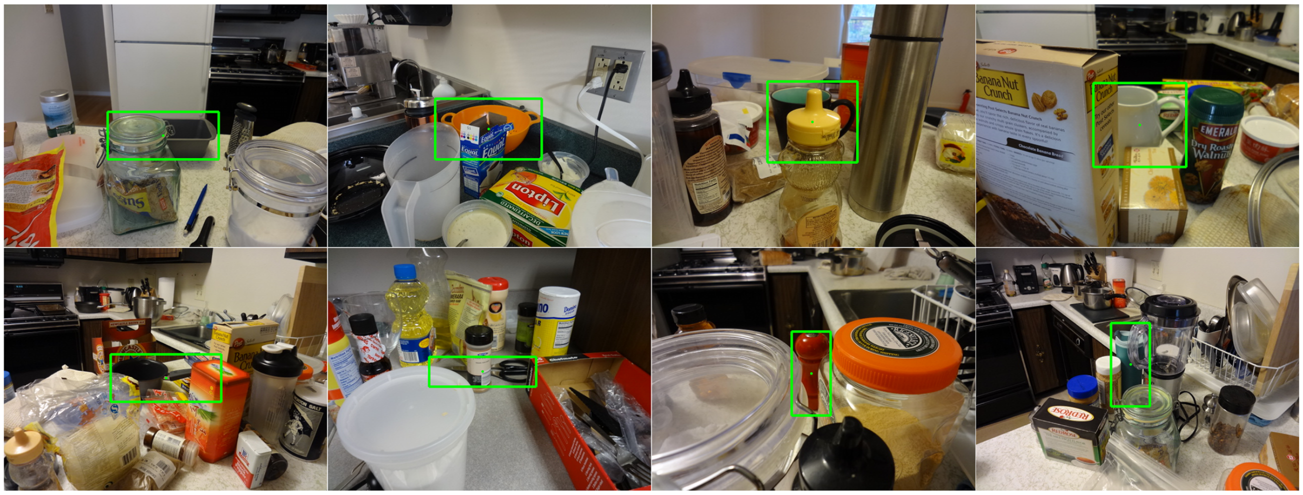


Figure 7: Example detection results under severe occlusions in cluttered household environments.

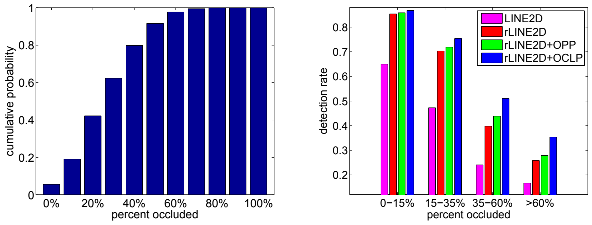


Figure 8: Dataset occlusion and performance. (*left*) Our dataset contains roughly equal amount of partial occlusions (1-35%) and heavy occlusions (35-80%). (*right*) While our methods improve performance under all levels of occlusions, we see larger gains under heavy occlusions.

well the occluded regions agree with our model. Intuitively, we should penalize points that are hypothesized to be occluded by the object detector (Section 3.1) but are highly likely to be visible according to our occlusion model. From this intuition, we propose the following detection score,

$$\text{score}_f(Z, \mathcal{V}) = \frac{1}{N} \sum_{i=1}^{N} V_i - f(Z, \mathcal{V}), \tag{16}$$

where $f(Z, \mathcal{V})$ is a penalty function for occlusions. A higher score indicates a more confident detection, and for detections with no occlusion, the score is 1. For detections with occlusion, the penalty $f(Z, \mathcal{V})$ is higher the more occluded points which are inconsistent with the model. In the following, we propose two penalty functions, $f_{\text{OPP}}(Z, \mathcal{V})$ and $f_{\text{OCLP}}(Z, \mathcal{V})$, based on the occlusion prior and occlusion conditional likelihood of Section 2.

3.2.1 Occlusion Prior Penalty

The occlusion prior penalty (OPP) gives high penalty to locations that are hypothesized to be occluded but have a high prior probability $P(V_i|O_c)$ of being visible. Intuitively, once the prior probability drops below some level λ, the point should be considered part of a valid occlusion and should not be penalized. This corresponds to a hinge loss function $\Gamma(P, \lambda) = \max\left(\frac{P-\lambda}{1-\lambda}, 0\right)$. The linear penalty we use is then,

$$f_{\text{OPP}}(Z, \mathcal{V}) = \frac{1}{N} \sum_{i=1}^{N} \left[(1 - V_i) \cdot \Gamma\left(P(V_i|O_c), \lambda_p\right) \right]. \tag{17}$$

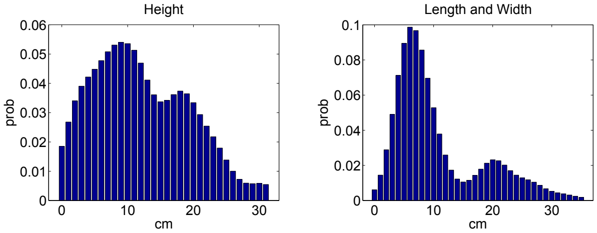


Figure 9: Distribution of (*left*) heights and (*right*) length and width of occluders in household environments.

3.2.2 Occlusion Conditional Likelihood Penalty

The occlusion conditional likelihood penalty (OCLP), on the other hand, gives high penalty to locations that are hypothesized to be occluded but have a high probability $P(V_i|\mathcal{V}_{\text{-}i}, O_c)$ of being visible given the visibility labeling of all other points $\mathcal{V}_{\text{-}i}$. Using the same penalty function formulation as the occlusion prior penalty, we have that,

$$f_{\text{OCLP}}(Z, \mathcal{V}) = \frac{1}{N} \sum_{i=1}^{N} \left[(1 - V_i) \cdot \Gamma\left(P(V_i|\mathcal{V}_{\text{-}i}, O_c), \lambda_c\right) \right]. \tag{18}$$

4. Evaluation

In order to evaluate our occlusion model's performance for object instance detection, two sets of experiments were conducted; the first for a single view of an object and the second for multiple views of an object. While in practice, one would only detect objects under multiple views, it is important to tease apart the effect of occlusion from the effect of viewpoint. We validate our approach by extending the LINE2D [7] method, a current state-of-the-art system for instance detection under arbitrary viewpoint.

	LINE2D	rLINE2D	rLINE2D with OPP	rLINE2D with OCLP
baking pan	0.44	0.51	0.51	**0.55**
colander	0.43	0.65	0.68	**0.74**
cup	0.40	0.60	0.63	**0.69**
pitcher	0.21	0.62	0.64	**0.69**
saucepan	0.48	0.67	0.65	**0.67**
scissors	0.32	0.46	0.46	**0.51**
shaker	0.18	0.35	0.40	**0.48**
thermos	0.43	0.73	0.80	**0.80**
Average	0.36	0.57	0.60	**0.64**
Gain	0.00	0.21	0.24	**0.28**

Table 1: Single view. Detection rate at 1.0 FPPI.

	LINE2D	rLINE2D	rLINE2D with OPP	rLINE2D with OCLP
baking pan	0.26	0.36	0.36	**0.41**
colander	0.37	0.61	0.60	**0.64**
cup	0.29	0.52	0.55	**0.58**
pitcher	0.28	0.46	0.51	**0.54**
saucepan	0.42	0.69	0.68	**0.69**
scissors	0.31	0.39	0.38	**0.39**
shaker	0.19	0.21	**0.30**	0.29
thermos	0.30	0.59	0.66	**0.69**
Average	0.30	0.48	0.51	**0.53**
Gain	0.00	0.18	0.21	**0.23**

Table 2: Multiple views. Detection rate at 1.0 FPPI.

In each set of experiments, we explore the benefits of (1) using only the bottom-up stage and (2) incorporating prior knowledge of occlusions with the top-down stage. When evaluating the bottom-up stage, we hypothesize the occluded region using the method of Section 3.1 and consider the score of only the visible portions of the detection. This score is equivalent to the first term of Equation 16. We will refer to this system as robust LINE2D (rLINE2D).

The parameters for LINE2D were the same as [7], and in our implementation, we use random edge points for the object template. We tested our implementation on a subset of the dataset provided by the authors of [7] and observed negligible difference in performance. The parameters of our occlusion model were calibrated on images not in the dataset and were kept the same for all objects and all experiments. The occlusion parameters were set to $\lambda_p = 0.5$ and $\lambda_c = 0.05$. We ran each experiment 10 times using different random edge points, and report the average results.

4.1. Dataset

Many object recognition algorithms work well in controlled scenes, but fail when faced with real-world conditions exhibiting strong viewpoint and illumination changes, occlusions and clutter. Current datasets for object detection under multiple viewpoints either contain objects on simple backgrounds [19] or have minimal occlusions [7, 11]. For evaluation under a more natural setting, the dataset we collected consists of common household objects in real, cluttered environments under various levels of occlusion. Our dataset contains 1600 images of 8 objects and is split evenly into two parts; 800 for a single view of an object and 800 for multiple views of an object. The single-view part contains ground truth labels of the occlusions and Figure 8 shows that our dataset contains roughly equal amounts of partial occlusion (1-35%) and heavy occlusions (35-80%) as defined by [2], making this dataset very challenging.

For multiple-view evaluation, we focus our viewpoint variation to primarily the elevation angle as relative performance under different azimuth angles is similar. We use 25 model images for each object which is the same sampling density as [7]. Each model image was collected with a calibration pattern to ground truth the camera viewpoint (ψ, θ) and to rectify the object silhouette to be upright. The test data was collected by changing the camera viewpoint and the scene around a stationary object. A calibration pattern was used to ground truth the position of the object.

4.2. Distribution of Occluder Sizes

The distribution of object sizes varies in different environments. For a particular scenario, it is natural to only consider objects as occluders if they appear in that environment. The statistics of objects can be obtained from the Internet [12] or, in the household scenario, simply from 100 common household items. Figure 9 shows the distributions for household objects.

From real world dimensions, we can compute the projected width and height distributions, $p_{\hat{w}}(\hat{w})$ and $p_{\hat{h}}(\hat{h})$, for a given camera viewpoint. The projected width distribution is the same for all viewpoints and is obtained by computing the probability density from Equation 2 for each pair of width and length measurement. These densities are discretized and averaged to give the final distribution of $\hat{w}$.

The projected height distribution, on the other hand, depends on the elevation angle ψ. From Equation 3, $\hat{h}$ is a factor $\cos\psi$ of h. Thus, the projected height distribution, $p_{\hat{h}}(\hat{h})$, is computed by subsampling $p_h(h)$ by $\cos\psi$.

4.3. Single view

We first evaluate the performance for single view object detection. An object is correctly detected if the intersection-over-union (IoU) of the predicted bounding box and the ground truth bounding box is greater than 0.5. Each object is evaluated on all 800 images in this part of the dataset and Figure 10 shows the false positive per image (FPPI) versus the detection rate (DR). To summarize the performance, we report the detection rate at 1.0 FPPI in Table 1. A few example detections are shown in Figure 7.

From the table, the rLINE2D method already significantly outperforms the baseline LINE2D method. One issue with the LINE2D gradient similarity metric (i.e., cosine of the orientation difference) is that it gives high score even to orientations that are very different, resulting in false posi-

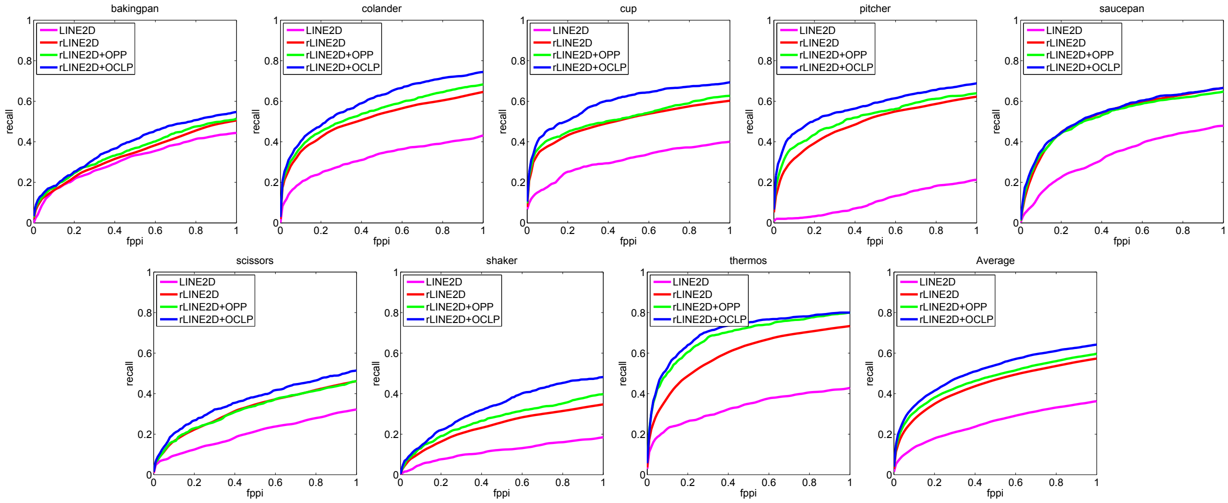


Figure 10: FPPI/DR results for single view. There is significant improvement in performance by using occlusion reasoning.

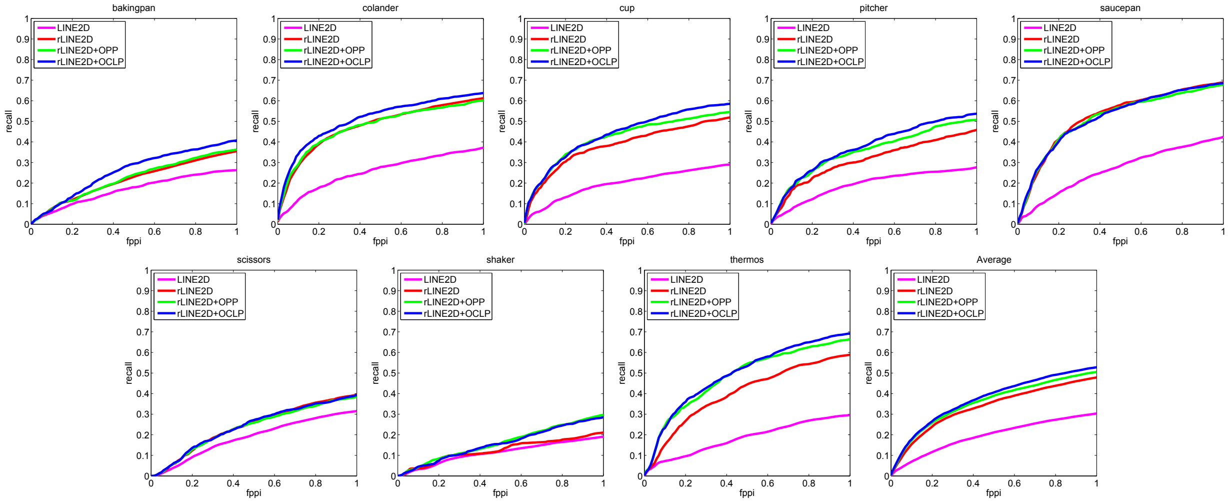


Figure 11: FPPI/DR results for multiple views. The overall performance is lower than the single view experiments due to more false positives, but importantly, we observe similar gains from using our occlusion reasoning

tives with high scores. The rLINE2D metric of considering only points with the same quantized orientation is more robust to background clutter in the presence of occlusions.

When rLINE2D is augmented with occlusion reasoning, there is an absolute improvement of 3% for OPP and 7% for OCLP. We performed a paired t-test and the results are significant at the $p = 10^{-11}$ level, indicating that both occlusion properties are informative for object detection. The disparity between the gains of OPP and OCLP suggests that accounting for global occlusion layout by OCLP is more informative than considering the *a priori* occlusion probability of each point individually by OPP. In particular, OCLP improves over OPP when one side of the object is completely occluded as shown in Figure 13. Although the top of the object is validly occluded, OPP assigns a high penalty.

Figure 8 shows the performance under different levels of occlusion. Here, the detection rate is the percentage of top detections which are correct. Our occlusion reasoning improves object detection under all levels of occlusion, but provides significantly larger gains for heavy occlusions.

To verify that our model accurately represents occlusions in real world scenes, we reran the above experiments with occlusion priors and conditional likelihoods learned from data. We use 5-fold cross-validation and Figure 12 shows the results from using different number of images for learning. From the plot, the performance of the learned occlusion properties, lOPP and lOCLP, both converge to their corresponding explicit counterparts, OPP and OCLP. This indicates that our model is able to represent occlusions accurately without requiring additional training data.

4.4. Multiple views

Next we evaluate the performance for object detection under multiple views. Figure 11 shows the FPPI/DR plots and Table 2 reports the detection rate at 1.0 FPPI. Again, we obtain significant improvement gains over the LINE2D system. Although the performance is lower overall due to more false positives from increasing the number of templates, the

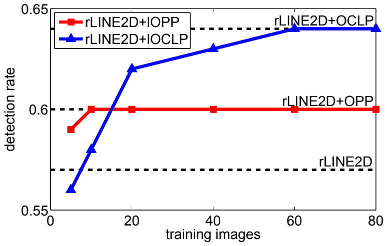


Figure 12: Learning the occlusion prior and conditional likelihood from data. Our model is able to explicitly represent occlusions accurately without additional training data.

relative gains at 3% for OPP and 5% for OCLP are similar to the single view case and are significant at the $p = 10^{-7}$ level. This demonstrates that our model is effective for representing occlusions under arbitrary view.

Figure 6 shows a typical false positive that can only be filtered by our occlusion reasoning. Although a majority of the points match well and the missing parts are largely coherent, the detection is not consistent with our occlusion model and is thus heavily penalized and filtered.

Figure 14 shows a couple of failure cases where our assumptions are violated. In the first image, the pot occluding the pitcher is not accurately modeled by its bounding box. In the second image, the occluding object rests on top of the scissor. Even though we do not handle these types of occlusions, our model represents the majority of occlusions and is thus able to increase the overall detection performance.

5. Conclusion

The main contribution of this paper is to demonstrate that a simple model of 3D interaction of objects can be used to represent occlusions effectively for object detection under arbitrary viewpoint without requiring additional training data. We propose a tractable method to capture global visibility relationships and show that it is more informative than the typical *a priori* probability of a point being occluded. Our results on a challenging dataset of texture-less objects under severe occlusions demonstrate that our approach can significantly improve object detection performance.

Acknowledgments

This work was supported in part by the National Science Foundation under ERC Grant No. EEEC-0540865.

References

[1] S. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *CVPR*, 2010.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2011.

[3] R. Fransens, C. Strecha, and L. Van Gool. A mean field em-algorithm for coherent occlusion handling in map-estimation prob. In *CVPR*, 2006.

[4] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 2011.

[5] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *NIPS*, 2011.

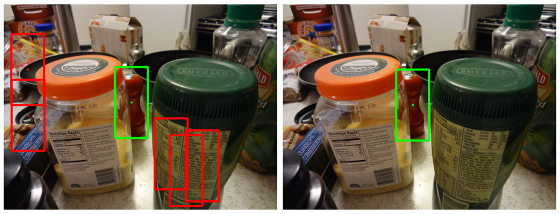


Figure 13: A typical case where OCLP performs better than OPP. (*left*) For OPP, the false positives in red have higher scores than the true detection in green. (*right*) For OCLP, the true detection is the top detection.

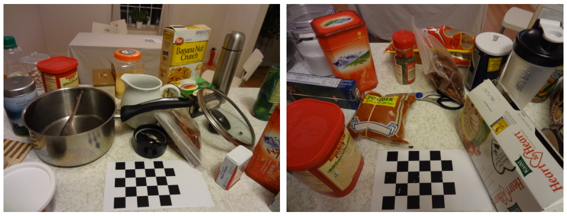


Figure 14: Typical failure cases of OCLP. (*left*) The pitcher is occluded by the handle of the pot which is not accurately modeled by a block. (*right*) The scissor is occluded by a plastic bag resting on top of it. In these cases, OCLP over penalizes the detections.

[6] W. Grimson, T. Lozano-Pérez, and D. Huttenlocher. *Object recognition by computer*. MIT Press, 1990.

[7] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of texture-less objects. *PAMI*, 2011.

[8] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab. Dominant orientation templates for real-time detection of texture-less objects. In *CVPR*, 2010.

[9] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008.

[10] S. Kwak, W. Nam, B. Han, and J. H. Han. Learn occlusion with likelihoods for visual tracking. In *ICCV*, 2011.

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

[12] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. In *SIGGRAPH*, 2007.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[14] D. Meger, C. Wojek, B. Schiele, and J. J. Little. Explicit occlusion reasoning for 3d object detection. In *BMVC*, 2011.

[15] H. Plantinga and C. Dyer. Visibility, occlusion, and the aspect graph. *IJCV*, 1990.

[16] L. Santalo. *Integral geometry and geometric probability*. Addison-Wesley Publishing Co., Reading, MA, 1976.

[17] M. Stevens and J. Beveridge. *Integrating Graphics and Vision for Object Recognition*. Kluwer Academic Publishers, 2000.

[18] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *ICCV*, 2009.

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

[20] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, 2006.

[21] A. Toshev, B. Taskar, and K. Daniilidis. Object detection via boundary structure segmentation. In *CVPR*, 2010.

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