Explicit Occlusion Reasoning for 3D Object Detection

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

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

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

Consider the problem of recognizing an object that is partially occluded in an image. The visible portions are likely to match learned appearance models for the object, but hidden portions will not. The (hypothetical) ideal system would consider *only* the visible object information, correctly ignoring all occluded regions. In purely 2D recognition, this requires inferring the occlusion present, which is a significant challenge since the number of possible occlusion masks is, in principle, exponential. We simplify the problem, considering only a small subset of the most likely occlusions (top, bottom, left, and right halves) and noting that some mismatch is tolerable. We train partial-object detectors tailored exactly to each of these few cases. In addition, we reason about objects in 3D and incorporate sensed geometry, as from an RGB-depth camera, along with visual imagery. This allows explicit occlusion masks to be constructed for each object hypothesis. The masks specify how much to trust each partial template, based on their overlap with visible object regions. Only the visible evidence contributes to our object reasoning.

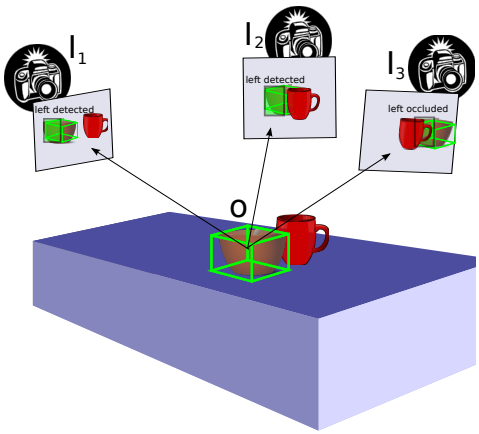

Figure 1: A 3D object is projected into many views and associated with partial detections, as indicated by visibility reasoning.

The implementation of our model, shown in figure 1 is as a Bayesian likelihood that explains the potential existence of an object o at every 3D region by the observed data and priors. Briefly, the input data from each view is $Z_t = \{I_t, C_t, P_t\}$, where I_t is the image, C_t is a point cloud and P_t is the structure-from-motion information. We factor the likelihood across views and data-types as:

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

$$= \underbrace{p(o)}_{object\ prior} \prod_t \underbrace{p(I_t|o,C_t,P_t)}_{appearance}\underbrace{p(C_t|o,P_t)}_{geometry}\underbrace{p(P_t|o)}_{registration} \quad (2)$$

Each of the components of our model is described in detail in the paper. Most notably, for the appearance model, we adapt the approach previously described in Wojek *et al.* [3] which weighted the contribution of each partial object detector by the expected visibility of the object in each region. Previous work required all occluding objects to be detected as well (i.e. the case of a pedestrian occluding a pedestrian) in order to compute occlusions, but we do this based on depth data from the sensor. The visibility-weighted mixture-of-experts formulation is:

$$p(I_t|o,C_t,P_t) = \frac{\sum_i v_{it}\delta(v_{it}>\theta)\Psi_s(d_{it}(o))\Psi_g(P_t\cdot o, d_{it}(o))}{\sum_i v_{it}\delta(v_{it}>\theta)} \quad (3)$$

which includes v, the expected visibility, Ψ_s, a potential function over the detector's score, and Ψ_g, a potential function for geometric agreement of the projected object with the image evidence. Our paper describes a method to sample likely 3D regions in a data-driven fashion that avoids linear search over the six dimensional output space.

Figure 2: Partial-object detectors respond on visible portions even when the object is occluded.

Our approach relies on visual object detectors trained both for entire objects, as well as for the portions of objects that are likely to be visible upon partial occlusion. We have trained detectors using the Deformable Parts Model (DPM) [1] where the training data given was automatically restricted to only a portion of the object. This is notably different than training a single full-object template and later restricting it during test time. The difference comes during the training procedure, where the learned partial template is likely to produce a different set of hard negatives. These important support vectors can be added to the training process and this leads to accurate learned models, even for small portions of the object. Figure 2 demonstrates the output of these learned models on several example images.

We have evaluated our method on the *UBC Visual Robot Survey*[1] dataset, which is a large, publicly available robot-vision resource containing images and point clouds from many views of realistic indoor scenes. This data was previously described in [2]. Our results demonstrate that the 3D object localization method is robust to clutter and occlusion. We outperform the image-based DPM [1] on the task of recognizing mugs and bowls in this dataset. There is also a noticeable improvement in results gained by using the partial-object templates.

[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*, 32, 2010.

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

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

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