Occlusion-Aware Object Localization, Segmentation and Pose Estimation

Samarth Brahmbhatt
http://www.cc.gatech.edu/~sbrahmbh
Henri Ben Amor
http://henibenamor.weebly.com
Henrik Christensen
http://www.cc.gatech.edu/~hic

Figure 1: Object localization and segmentation example. Left: Image, Right: Refined mask from SD-HOP

We present a learning approach for localization and segmentation of objects in an image in a manner that is robust to partial occlusion. Our algorithm, Segmentation and Detection using Higher-Order Potentials (SD-HOP) produces a bounding box around the full extent of the object and labels pixels in its interior that belong to the object. This is different from semantic segmentation, which does not provide information about the spatial position of labelled pixels inside the object.

A common theme in the literature is to model occlusion geometrically or appearance-wise, thereby allowing it to contribute to the detection process. The former often make simplifying assumptions about occluder and scene geometry. Our appearance-based approach avoids these assumptions and performs better than existing appearance-based approaches due to the use of higher-order potentials for modelling neighbour influence and a loss function that targets both localization and segmentation.

SD-HOP discriminatively learns HOG templates for objects and occlusion. Whereas the object templates model the objects of interest, the occlusion templates provide discriminative support and do not model a specific occluder. Segmentation is done by considering the response of patches to these templates, and influence of neighbouring patches through a CRF with higher-order connections. The training phase requires a set of patches to these templates, and influence of neighbouring patches through the spatial position of labelled pixels inside the object.

Inference is performed by finding the labelling that minimizes the dot-product energy: $\mathbf{y} = \arg\min_{\mathbf{y}} \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y})$. Due to the linear parametrization of energy and decomposability of the loss function over the unary terms, inference is efficient. At every bounding box location in a pyramid, it is performed by a single $s^2$ mincut on a graph constructed as described in [1] and [2].

We implemented SD-HOP in Matlab, with MCV search and inference implemented in CUDA since they are massively parallel problems. Inference on a 640x480 image with 11 scales takes 3s for a single object with a single viewpoint on our 3.4 GHz CPU and NVIDIA GT730 GPU. SD-HOP achieves 13.52% segmentation error and 0.81 area under the false-positive per image vs. recall curve on average over the challenging CMU Kitchen Occlusion Dataset. This is a 42.44% decrease in segmentation error and a 16.13% increase in localization performance compared to the state-of-the-art. Figure [3] shows a sample output on this dataset.

We demonstrate that the segmentation output of SD-HOP can be used to ignore edges produced by occlusion, thereby making model-based 3D pose estimation robust to partial occlusion as shown in Figure [2].

Learning involves determining linear weights $\mathbf{w}$ such that the score $\mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y})$ of any ground truth labelled image $\mathbf{x}$ must be smaller than the score $\mathbf{w}^T \Psi(\mathbf{x}, \hat{\mathbf{y}})$ of any other labelling $\hat{\mathbf{y}}$, by the distance between the two labellings $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ minus the slack variable $\xi_i$, where $|\mathbf{w}|_2$ and $\xi_i$ are minimized. Hence we learn $\mathbf{w}$ by solving the following constrained Quadratic Program:

$$\min_{\mathbf{w}} \frac{1}{2} |\mathbf{w}|_2^2 + C \sum_{i=1}^{N} \xi_i$$

s.t. $\mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y}) + \xi_i \geq \Delta(\mathbf{y}, \hat{\mathbf{y}}) \forall i, \hat{\mathbf{y}} \in Y_i$

$\xi_i \geq 0 \forall i$

$D^T \mathbf{w} \geq 0$

$D^T$ is a second order curvature constraint on the K+1 weights for the higher-order potentials, which forces them to make a concave lower envelope. Training is performed by using the cutting plane training algorithm of [3], with adaptation for training higher-order potentials as described in [2]. The loss function between two labels $\mathbf{y}$ and $\hat{\mathbf{y}}$ depends on the amount of overlap between the two bounding boxes and the Hamming distance between the visibility labellings:

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \left( \frac{1 - \text{area}(\mathbf{p} \cap \hat{\mathbf{p}})}{\text{area}(\mathbf{p})} + \frac{\text{area}(\mathbf{p} \cap \hat{\mathbf{p}})}{\text{area}(\mathbf{p})} \right) \cdot H(\mathbf{v}, \hat{\mathbf{v}})$$

We demonstrate that the segmentation output of SD-HOP can be used to ignore edges produced by occlusion, thereby making model-based 3D pose estimation robust to partial occlusion as shown in Figure [2].

