Coherent Occlusion Reasoning for Instance Recognition

Edward Hsiao and Martial Hebert

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

***Abstract*—Occlusions are common in real world scenes and are a major obstacle to robust object detection. In this paper, we present a method to coherently reason about occlusions on many types of detectors. Previous approaches primarily enforced local coherency or learned the occlusion structure from data. However, local coherency ignores the occlusion structure in real world scenes and learning from data requires tediously labeling many examples of occlusions for every view of every object. Other approaches require binary classifications of matching scores. We address these limitations by formulating occlusion reasoning as an efficient search over occluding blocks which best explain a probabilistic matching pattern. Our method demonstrates significant improvement in estimating the mask of the occluding region and improves object instance detection on a challenging dataset of objects under severe occlusions.**

I. Introduction

Recognizing object instances from a single image is essential for many applications in visual search, robotics and augmented reality. The task is to detect an object under arbitrary viewpoint in a cluttered scene, where often only a single image per view is provided for training [1]–[3]. In real world scenes, occlusions are common and pose a major obstacle for robust object detection. While feature-based approaches [2], [3] can be used to recognize texture-rich objects under occlusions, these methods fail when presented with objects exhibiting large uniform regions [1]. Current approaches for recognizing these objects rely on matching their shape [1], [4], [5]. However, many object shapes are very simple and are easily confused with background clutter. In the presence of occlusions, the ambiguity increases and the performance of many detection algorithms decrease rapidly. In this work, we exploit the fact that occlusions in the real world are not random [6], [7] to reject false positives. The main contributions of this paper are threefold: 1) formulating occlusion reasoning as efficient search, 2) providing a coherent method for probabilistic reasoning on multiple cues, and 3) scoring the matching pattern of an object detector. Our approach provides a more accurate estimate of the occlusion mask (Figure 1) and improves object detection performance on a very challenging dataset.

In the past, significant research has been dedicated to occlusion reasoning for object detection. Occlusions are typically classified as regions that are inconsistent with object statistics [8], [9]. To handle classification noise, local coherency is enforced with methods such as Markov Random Field [10] and mean-shift [9]. In general, these approaches reason either on only the boundary [6] or on only coarse region grid cells [9], [11]. However, grid cells on the object boundary are easily corrupted by background clutter, and using the boundary alone ignores the fact that occlusions are by solid objects.

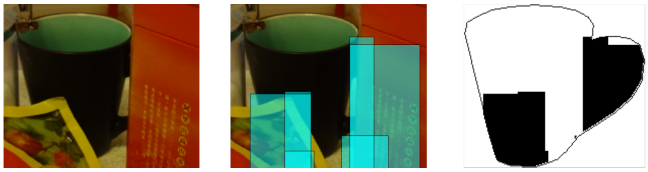

Fig. 1. Example occlusion predictions. Given a hypothesis detection from a cup detector, the proposed Occlusion Estimation by Subwindow Search (OESS) method predicts the occlusion mask and determines how likely it belongs to a true detection. From left to right we show (1) detection, (2) predicted occluder boxes and (3) predicted occlusion mask.

In addition, modeling any inconsistent region as an occlusion ignores the structure of occlusions for many objects [6], [7]. Some methods attempt to learn the structure from labeled data. Gao et al. [11] use structured-SVMs, while Kwak et al. [12] learn patch likelihoods. Given enough labeled examples of occlusions, these methods can obtain an accurate model. However, obtaining many labeled examples for each view of every object is very tedious. Recently, Hsiao and Hebert [6] proposed an occlusion model for object detection under arbitrary viewpoint. Their method, however, requires the object detector to return a binary decision on whether a boundary point is matched or not matched.

In this paper, we formulate occlusion reasoning as an efficient search for occluding blocks which best explain the matching pattern returned by an object detector. We show that our method can probabilistically reason about multiple types of object cues together. Our results show improvement in occlusion prediction and object instance detection over state-of-the-art methods on the challenging CMU Kitchen Occlusion Dataset [6].

II. Occlusion Model

Let the 2D view of an object be represented by k markers[1] $\mathcal{Z} = \{z_1, ..., z_k\}$. Each marker z_i captures the local information centered around coordinate (x_i, y_i) on the object. These markers can capture any type of information from local shape to texture and color. A marker, for example, can be the center of a HOG cell [13], a SIFT keypoint [14], a LINE2D edge

[1]We introduce the term *marker* to deliberately avoid using the term *feature* which has been significantly overloaded in the literature.

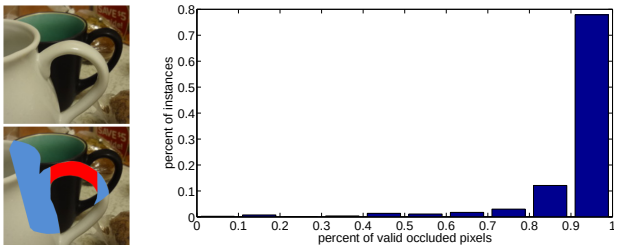

Fig. 2. Validity of occlusion model. (left) We show in blue, the occluded pixels which satisfy our approximation, and in red, those that do not. (right) For each object instance in the CMU_KO8 dataset [6], we evaluate the percentage of occluded pixels which satisfy our approximation that they can be explained by an occlusion box which touches the object base. For 80% of the images, over 90% of the occluded pixels can be explained with our model.

point [1], or a Hough voting patch [15]. By using this object representation, our occlusion model can augment any object detector which returns the probability, p_i, that each marker, z_i, matches every image location. Here, we follow previous research and assume that the matching probability, p_i, is a good indicator of how likely a marker is visible [8], [9]. In addition, let each marker, z_i, have a weight, w_i, which indicates its importance and influence. In the following, we refer to the set of tuples $\mathcal{M} = \{(z_i, p_i, w_i) | 1 \leq i \leq k\}$ as the *matching pattern*. The hypothesis is that if $\mathcal{M}$ can be explained well by a set of valid occluders, it is more likely to be a true detection than those matching patterns that cannot.

In the real world, objects which occlude each other are usually on the same support surface. Thus, the base of an occluder in an image is usually below the base of the object. To capture this insight, we approximate occluding objects as bounding boxes [6], [13] and consider valid occluders as those boxes which touch the base of the object. Boxes with a base lower than this do not need to be considered as we only care about matching probabilities on the object. This approximation is consistent with the occlusion types observed by Dollar et al. in [7]. Figure 2 further validates the approximation on the recently proposed CMU Kitchen Occlusion Dataset [6] which contains images of objects in natural scenes with groundtruth occlusion labels. An occluded pixel is consistent with the approximation if there are no un-occluded object pixels below it. For 80% of the images, over 90% of the occluded pixels are consistent.

Given this model of occluders, the goal is to search for the best set of occluder boxes b^* (oboxes) to explain $\mathcal{M}$. Each obox is parameterized by its top, left and right coordinates (t, l, r) with the bottom fixed to the base of the object.

We define the value of each marker z_i to be:

$$v_i = w_i \cdot (2p_i - 1). \tag{1}$$

For uniform marker weights (i.e., $w_i = 1$), definitely visible markers have a value of $v_i = 1$ and definitely occluded markers have a value of $v_i = -1$. When a marker z_i falls inside any obox, b, its value is negated, essentially rewarding markers more likely to be occluded and penalizing markers more likely to be visible. An object occlusion is thus represented by a set of oboxes, b. We define the occlusion quality function $q : \mathcal{B} \rightarrow \mathbb{R}$ to be:

$$q(\mathrm{b}) = \sum_{z_i \notin \mathrm{b}_\cup} v_i - \sum_{z_j \in \mathrm{b}_\cup} v_j, \tag{2}$$

where $\mathcal{B}$ is the set of all possible oboxes, $\mathrm{b} \subset \mathcal{B}$, and $\mathrm{b}_\cup$ is the union of all oboxes $b \in \mathrm{b}$. The first term considers all markers outside the union of the obox set, giving positive score to visible markers and penalizing occluded markers. The second term considers all markers inside the union of the obox set, giving positive score to occluded markers that are explained and penalizing visible markers. The best obox set is given by:

$$\mathrm{b}^* = \underset{\mathrm{b} \subset \mathcal{B}}{\operatorname{argmax}}\, q(\mathrm{b}). \tag{3}$$

III. Combining with Object Detection

Given the matching pattern $\mathcal{M}$, our method, termed Occlusion Estimation by Subwindow Search (OESS), returns the best obox set, b^*, and its occlusion quality $q(\mathrm{b}^*)$. While this occlusion quality can be used by itself as the score of the detection, it does not account for how well the object is matched. A detection with all the markers being occluded (i.e., $p_i = 0$) would receive a very high occlusion quality score since it can be fully explained with an obox that covers the whole object. However, no object markers are matched. To incorporate the occlusion quality q with the raw score s returned by the object detector, we learn a linear weighting using an Exemplar SVM (ESVM) [16] between, s, q and their product sq:

$$score = \alpha_1 s + \alpha_2 q + \alpha_3 sq. \tag{4}$$

The single positive example (s^+, q^+) is the ideal detection where s^+ is the score of the detector on the training image, and $q^+ = \sum w_i$ is the maximum occlusion quality when all points are visible. The goal of the ESVM is to determine the weighting which best separates the false positives from this point. An ideal detection would thus have both a high matching score as well as a high occlusion quality. The parameters for training the ESVM are the same as [16] and we use three iterations of hard negative mining. Since the ESVM output is not calibrated between different detectors, it is difficult to choose the best scoring template for object recognition. We calibrate the scores using the Extreme Value Theory [17], as it does not require positive examples, which are often difficult to obtain. Negative data are easily obtained by sampling background images.

A. Probability Calibration

Most object detectors do not return calibrated marker scores that can be interpreted as matching probabilities (e.g., the decomposed cell-wise score [9] of HOG and the point-wise similarity metrics of LINE2D [1] and rLINE2D [6]). To calibrate the raw marker scores, we again use the Extreme Value Theory [17] because obtaining many positive examples of occlusions is tedious. Each marker is calibrated independently using its raw matching scores on randomly sampled detections in background clutter as negative data.

B. Weighting

Different marker types capture different spatial extent of information around their positions. Boundary cues, such as

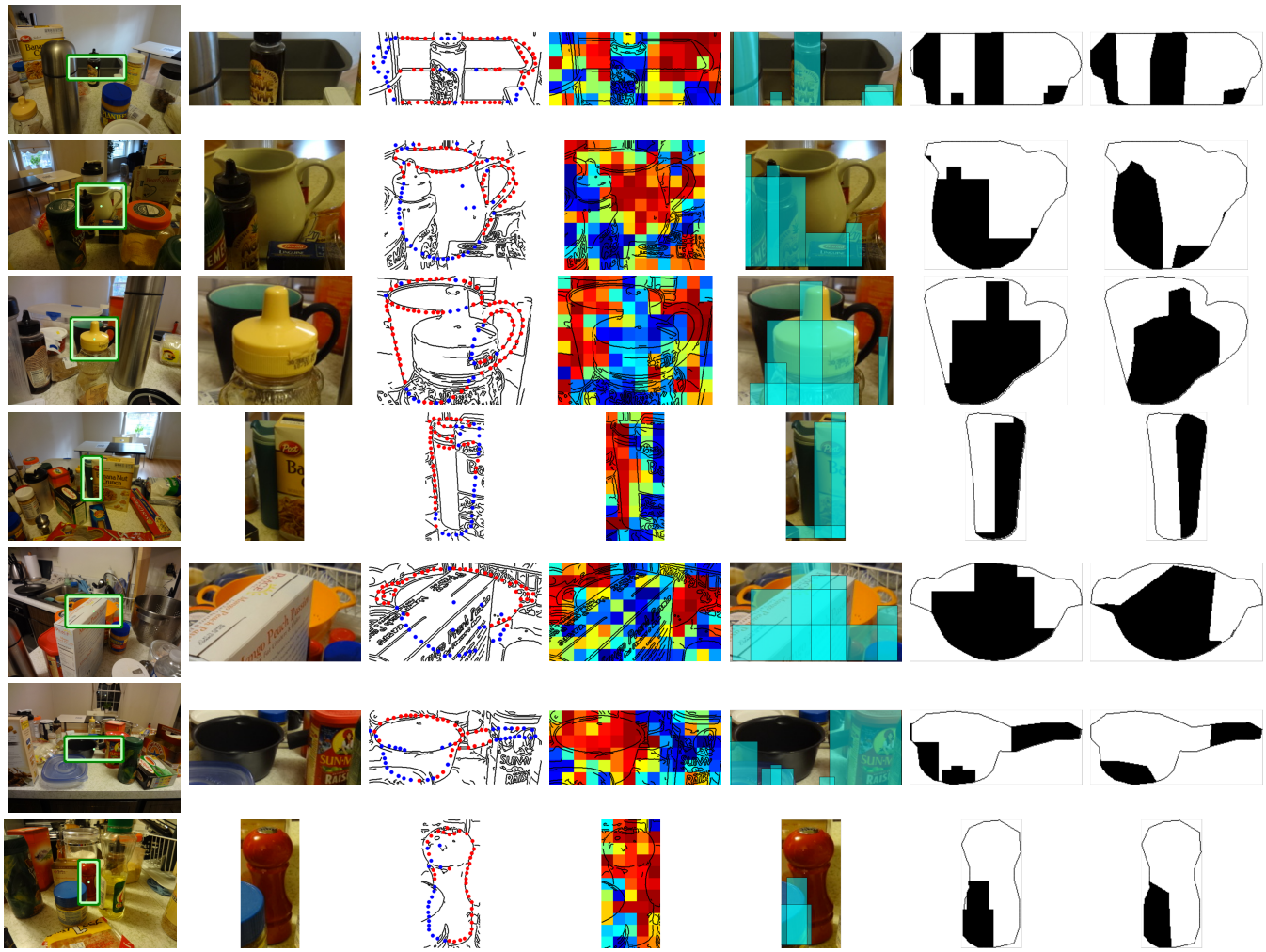

Fig. 3. Example detections and occlusion reasoning. From left to right, we show (1) the original image with bounding box of detection, (2) the zoomed in view of the detection, (3) boundary matches, (4) activation scores of region cells using texture and color, (5) hypothesized oboxes, (6) predicted occlusion mask, and (7) groundtruth occlusion mask. For columns 3 and 4, the hotter the color, the better the match. To be consistent, red points indicate matched boundary points and blue points indicate points that are not matched.

LINE2D, use sampled edge points which only consider information very locally. Grid-based approaches, such as HOG, cover a much larger area for each grid cell. Intuitively, we want to give more weight to points that have a larger region of influence. We weight each marker by the area in pixels of the region it represents. For grid based methods, this is the area of the cell. For point-based methods, this is the area of the sampling circle.

IV. Evaluation

We evaluate our occlusion model's performance in object instance detection by conducting two sets of experiments. The first evaluates the algorithm's accuracy in predicting occlusions, and the second evaluates the algorithm's ability to detect objects. We systematically analyze the combination of boundary (B), texture (T), and color (C) cues and their effect on occlusion reasoning. We set $\gamma = 16$ for all of the experiments.

A. Dataset

Most object instance detection datasets contain objects on monotone backgrounds [18] or have very little occlusion [19]. We evaluate our algorithm on the recent CMU Kitchen Occlusion Dataset [6] because it provides occlusion labels for images of objects in more realistic scenes with both severe clutter and occlusions. The dataset contains 1600 images of 8 common household objects, with the single view portion containing 800 images and the multiple view portion containing 800 images of objects under 25 viewpoints. There is roughly equal amounts of partial occlusions (1-35%) and heavy occlusions (35-80%) in the dataset, making it particularly challenging.

B. Templates

Template-based approaches have been used extensively in the literature [20]–[22] for object detection. The recent development of efficient template matching techniques [1], [23], [24] has made it a viable method for object detection under arbitrary viewpoint. In the following, we describe the templates we use to represent the boundary, texture and color

of an object. These templates are chosen as examples to illustrate our approach, and many other methods and types of cues can be used similarly.

a) Boundary (B): The object boundary is the most distinguishing characteristic of texture-less objects. Shape matching approaches range from using sparse edge points [1], [25] and curves [4], [5] to edge histograms [26], [27]. For efficient matching under many object viewpoints, we choose to use rLINE2D [6]. This method is based on LINE2D [1] which represents an object by a set of sparse edge points, each with a quantized orientation. While the LINE2D method uses the cosine of the orientation difference as the similarity measure, Hsiao and Hebert [6] observed that it works well primarily when objects are largely visible.

b) Texture (T): The lack of texture is often viewed as uninformative and ignored for recognition. However, many objects have simple shape which easily align to background clutter as shown by the false positives in Figure 4. These false positives can be filtered using the appearance of the object interior. We represent the texture using the popular HOG [13] template. For our experiments, we fix the size of the HOG cells to be the standard 8×8 for all the objects and learn the weights using ESVM [16].

c) Color (C): Color can be a very informative cue for object detection. We use a simple color representation to illustrate our approach. Using the same grid structure as the texture template, we compute the average color in L*a*b space for each cell. We use the squared difference between each extracted color of a cell in the image and its corresponding model cell as the feature. We again learn the weighting of the features using ESVM.

C. Occlusion Prediction

First, we evaluate the performance in predicting the occluded region. We convert the best obox set, b^*, of each detection window into a binary occlusion mask (Figure 3). The performance is evaluated using the standard intersection-over-union (IoU) metric between the predicted mask and the groundtruth mask from the dataset. We average the IoU for all the images of all the objects. We compare our method against thresholding the matching probabilities at 0.5, the mean-shift occlusion reasoning approach of [9] which enforces local coherency, and OCLP [6]. Since thresholding and mean-shift on B produce only point classifications, we dilate the classifications by the sampling radius of rLINE2D to produce an occlusion mask. This dilation captures the local region of influence of each point. In addition, OCLP is a scoring mechanism and does not predict an occlusion mask. We generate a mask by first thresholding the matching probabilities at 0.5 to get the matched points. Then, we evaluate the occlusion conditional likelihood [6] at all points on the object mask given these matched points and threshold it to predict the occlusion, as summarized in Figure 5.

From the figure, OESS significantly outperforms the other methods for all templates except for C. Contrary to what one would naturally believe, color is actually not a very good cue for whether a marker is occluded or not because many objects have very similar colors. When the color of the occluder is similar to the object, the matching probability will be incorrect

Fig. 4. Example false positives from using boundary alone. From left to right, we show (1) object, (2) false detection, (3) zoomed-in view and (4) boundary points that are matched in red and not matched in blue.

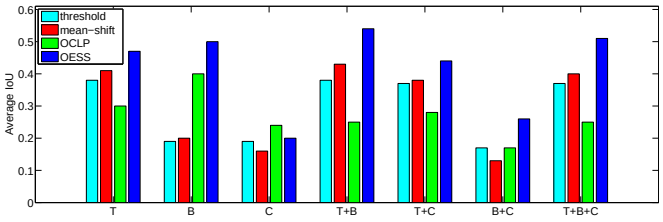

Fig. 5. Occlusion prediction performance. We compare OESS with thresholding the matching probabilities, using the mean-shift approach of [9] to enforce local occlusion coherency, and OCLP. We report the average intersection-over-union (IoU) between the predicted mask and the groundtruth. OESS significantly outperforms all the approaches.

and these poor confidences often hurt the occlusion reasoning performance. This suggests that most of the information is contained in T and B. In addition, thresholding and mean-shift perform significantly worse since they do not account for occlusion structure in the real world. OCLP captures some structure and works well for B, but is sensitive to binary misclassifications, especially when applied to approaches that use a dense grid. OESS operates directly on probabilities and captures higher level occlusion structure.

Figure 6 shows example failure cases of the full system (T+B+C with OESS). In the first case, the occluding object violates our bounding box assumption and we are unable to recover a good occlusion mask. In the second case, the matching probabilities are inaccurate. More robust templates which obtain more accurate matching probabilities will aid in improving the performance of occlusion reasoning and OESS.

D. Object Detection

We also need to verify that our method maintains or improves the detection performance while significantly improving the occlusion prediction. In the following experiments, an object is correctly detected if it satisfies the PASCAL overlap criterion [28] with the ground truth bounding box. We compare our occlusion reasoning approach with the OPP and OCLP approaches of Hsiao and Hebert [6]. We compare the occlusion reasoning on different combinations of B, T, and C cues. Table I and II summarize the precision-recall and the false-positive-per-image (FPPI) versus the detection rate (DR) curves respectively, using the average precision (AP) and DR at 1.0 FPPI. From the tables, OESS improves the performance over the baseline for all the cues. Importantly, it never performs worse, unlike OPP and OCLP. In addition, OESS outperforms OPP and OCLP in all cases for the typical recognition scenario

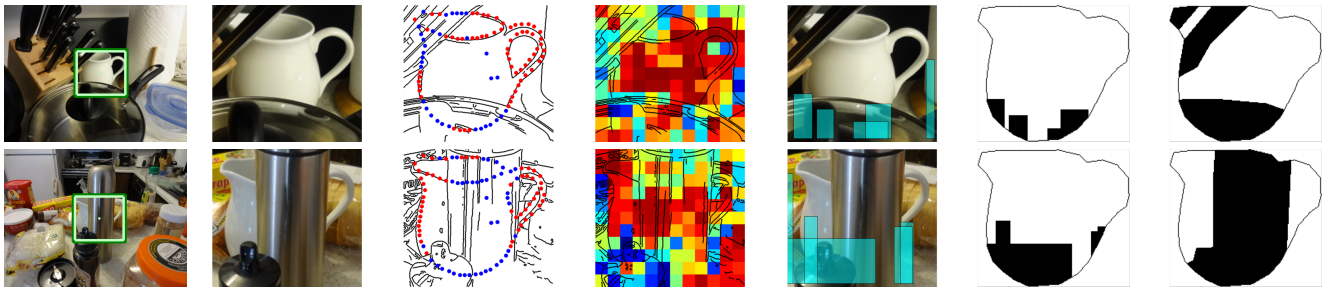

Fig. 6. Example failure cases for occlusion segmentation of a pitcher. In the first row, the occlusion does not satisfy the bounding box approximation of occluders. In the second row, the region template produces inaccurate activation scores.

Single	T [13]	B [6]	C	T+B	T+C	B+C	T+B+C
baseline	0.49	0.47	0.06	0.68	0.54	0.55	0.72
+OPP [6]	0.50	0.51	0.04	0.69	0.55	0.56	0.73
+OCLP [6]	0.51	0.59	0.04	**0.71**	0.55	0.55	0.74
+OESS	**0.53**	**0.59**	**0.07**	0.70	**0.58**	**0.56**	**0.74**
Multiple	T [13]	B [6]	C	T+B	T+C	B+C	T+B+C
baseline	0.47	0.30	0.08	0.57	0.51	0.38	0.59
+OPP [6]	0.48	0.34	0.06	0.57	0.52	0.39	0.60
+OCLP [6]	0.50	0.38	0.06	0.58	0.53	0.39	0.61
+OESS	**0.52**	**0.47**	**0.11**	**0.61**	**0.55**	**0.41**	**0.63**

TABLE I. OBJECT DETECTION : AVERAGE PRECISION.

Single	T [13]	B [6]	C	T+B	T+C	B+C	T+B+C
baseline	0.78	0.75	0.32	0.88	0.83	0.81	0.90
+OPP [6]	0.77	0.77	0.24	0.89	0.84	**0.82**	0.91
+OCLP [6]	0.79	0.84	0.21	**0.91**	0.85	0.81	0.92
+OESS	**0.80**	**0.84**	**0.33**	0.89	**0.86**	0.81	**0.92**
Multiple	T [13]	B [6]	C	T+B	T+C	B+C	T+B+C
baseline	0.74	0.66	0.28	0.84	0.77	0.71	0.86
+OPP [6]	0.75	0.68	0.22	0.84	0.78	0.72	0.86
+OCLP [6]	0.77	0.73	0.22	0.85	0.79	0.72	0.86
+OESS	**0.79**	**0.78**	**0.36**	**0.86**	**0.81**	**0.74**	**0.87**

TABLE II. OBJECT DETECTION : DR AT 1.0 FPPI.

with multiple object viewpoints.

V. CONCLUSION

The main contribution of this paper is to formulate occlusion reasoning as an efficient search over occluding blocks which best explain a probabilistic matching pattern. Our approach is able to coherently reason on matching patterns returned from multiple cues. Given a set of hypothesis object detections, we effectively score them based on how well the matching pattern can be explained by a set of valid occluding boxes. Our results on a challenging dataset of objects under severe occlusions and in heavy clutter demonstrate significant improvement over state-of-the-art methods.

REFERENCES

[1] 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.

[2] A. Collet, M. Martinez, and S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *IJRR*, 2011.

[3] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *IJCV*, 2006.

[4] P. Srinivasan, Q. Zhu, and J. Shi, "Many-to-one contour matching for describing and discriminating object shape," in *CVPR*, 2010.

[5] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *PAMI*, 2008.

[6] E. Hsiao and M. Hebert, "Occlusion reasoning for object detection under arbitrary viewpoint," in *CVPR*, 2012.

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

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

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

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

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

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

[13] N. Dalal, "Finding people in images and videos," Ph.D. dissertation, Institut National Polytechnique de Grenoble / INRIA Grenoble, 2006.

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

[15] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *CVPR*, 2009.

[16] T. Malisiewicz and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *ICCV*, 2011.

[17] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *CVPR*, 2012.

[18] S. Nene, S. Nayar, and H. Murase, "Columbia object image library," *TR CUCS-006-96, Columbia University*, 1996.

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

[20] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *CVPR*, 1997.

[21] D. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *ICCV*, 1999.

[22] P. Viola and M. Jones, "Robust real-time object detection," *IJCV*, 2001.

[23] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *CVPR*, 2008.

[24] 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.

[25] M. Leordeanu, M. Hebert, and R. Sukthankar, "Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features," in *CVPR*, 2007.

[26] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, 2002.

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

[28] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, 2010.