Object Detection by 3D Aspectlets and Occlusion Reasoning

Yu Xiang
University of Michigan
yuxiang@umich.edu

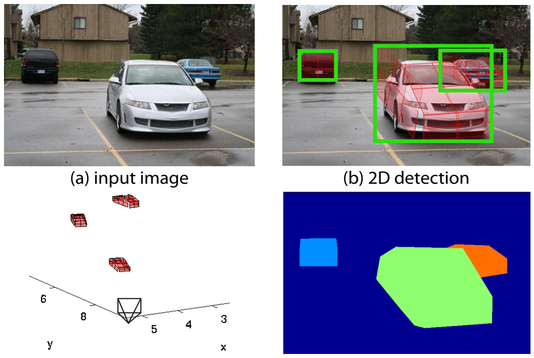
Silvio Savarese
Stanford University
ssilvio@stanford.edu

Abstract

We propose a novel framework for detecting multiple objects from a single image and reasoning about occlusions between objects. We address this problem from a 3D perspective in order to handle various occlusion patterns which can take place between objects. We introduce the concept of "3D aspectlets" based on a piecewise planar object representation. A 3D aspectlet represents a portion of the object which provides evidence for partial observation of the object. A new probabilistic model (which we called spatial layout model) is proposed to combine the bottom-up evidence from 3D aspectlets and the top-down occlusion reasoning to help object detection. Experiments are conducted on two new challenging datasets with various degrees of occlusions to demonstrate that, by contextualizing objects in their 3D geometric configuration with respect to the observer, our method is able to obtain competitive detection results even in the presence of severe occlusions. Moreover, we demonstrate the ability of the model to estimate the locations of objects in 3D and predict the occlusion order between objects in images.

1. Introduction

The traditional object detection methods (e.g., [22], [5] and [8]) detect each object in an input image independently without considering the environment of the object. However, objects are not isolated in the real world. The contextual information around the objects plays an important role in object recognition [17]. Recently, different types of contextual information have been utilized to help object detection, such as 3D scene geometry [12] and 2D object co-occurrence [6]. Despite these efforts, the contextual cues that arise by considering object occlusions have not been fully explored yet. When objects occlude each other or are truncated by other scene elements, only limited portions of the objects are visible and some of the cues which we typically use to recognize the objects may not be available (e.g., the wheels of the blue car in Fig. 1(a)). In these cases, detecting each object independently is likely to fail (the detection score of the blue car in Fig. 1(a) would be low).



(a) input image (b) 2D detection (c) 3D spatial layout (d) 2D object mask

Figure 1. Illustration of our spatial layout model. Given an input image (a), our model detects the objects in the image (b), estimates their 3D spatial layout (c), and predicts the 2D object mask (d) which shows the occlusion order between objects.

Detecting objects under occlusions is challenging due to various occlusion patterns in the image that can take place between objects. These occlusion patterns depend on the relative locations of objects in 3D with respect to the camera and also the shape and pose of the objects. Without considering these factors, methods which reason about occlusions based on 2D image features only, such as [23] and [9], are fragile to the uncertainty of the image evidence. In this paper, we handle occlusions in object detection from a 3D perspective. We design a novel framework that, from just one single image (Fig. 1(a)), is capable to jointly detect objects (Fig. 1(b)), determine their 3D spatial layout (Fig. 1(c)) and interpret which object occludes which (Fig. 1(d)). We call this model the Spatial Layout Model (SLM). First, inspired by the aspect part representation in [27], we propose a new 3D object representation using piecewise planar parts. These parts are fine-grained and suitable for occlusion reasoning in the sense that they can be approximated as either visible or non-visible. Second, inspired by the poselet framework for human detection [3], we group the planar parts in 3D to represent portions of the object. We call each group a "3D aspectlet", which is generated automatically. 3D aspectlets are able to provide more robust evidence of partial observations as opposed to the planar parts themselves. Finally, we generate hypotheses of the locations and poses of objects and camera in 3D (Fig. 1(c)), and then verify these hypotheses by combining prior knowledge and evidence from 3D aspectlets. This is achieved by a Markov Chain Monte Carlo (MCMC) sampling strategy, where different kinds of moves are designed to explore the hypothesis space efficiently. In this process, 3D aspectlets are weighted according to the occlusion patterns induced by the 3D hypotheses (Fig. 1(d)). Consequently, we combine the bottom-up evidence from 3D aspectlets and the top-down occlusion reasoning to help object detection. Experiments are conducted on two new challenging datasets, i.e., an outdoor-scene dataset with cars and an indoor-scene dataset with furniture, where multiple objects are observed under various degrees of occlusions. We demonstrate that our method is able to obtain competitive detection results even in the presence of severe occlusions. Besides, our method has the ability to estimate the spatial layouts of objects in 3D and predict the occlusion order between objects in images.

2. Related Work

Recently, the use of context for object detection has received increasing attention. Desai et al. [6] formulate the multiple object detection as a structured labeling problem, where spatial interactions between objects in 2D are modeled. Hoiem et al. [12] introduce the idea of using 3D scene geometry to help 2D object detection, where objects are supposed to be on the ground plane with certain heights. The ground plane constraint is generalized to supporting planes of objects by Bao et al. [2]. Richer geometrical and physical constraints are also explored by different works. Hedau et al. [11] detect indoor-scene objects by considering the room layout. Choi et al. [4] propose 3D Geometric Phases to capture the semantic and geometric relationships between co-occurring objects in 3D. In this work, we demonstrate that by modeling the spatial context of objects in 3D, we can successfully enhance object detection and reason about occlusions between objects.

Previous works that reason about occlusions have mostly focused on image segmentation [24, 13], object tracking [25], single object instance recognition [16] and category-level object detection [26, 23, 9, 28, 18]. Methods for object detection have leveraged on 2D image features to predict whether an object is occluded or not, such as [23] and [9]. Very few works have addressed the problem from a 3D perspective. Two exceptions are [26] and [25], which reason about occlusions between humans by generating hypotheses of humans in 3D and verifying these hypotheses using part-based human detectors. Different from these, we do not model occlusions with a simplified 2.5D structure of depth layers, but rather a true 3D representation to predict occlusion patterns. Recently, [28] uses 2D masks to represent occlusion patterns, while [18] learns the occlusion patterns from training data. In both methods, the occlusion patterns are view-specific, and only limited number of occlusion patterns can be modeled. Our method infers the occlusion patterns from the 3D spatial layout of objects, which is general to handle various occlusion patterns.

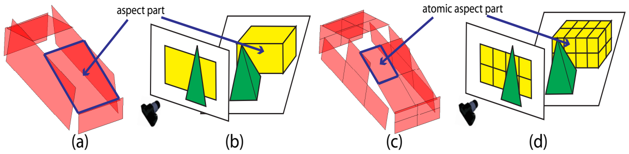


Figure 2. (a) Aspect part representation of car in [27] (b) A toy example shows that an AP is partially visible due to occlusion. (c) AAP representation of car in our model. (d) A toy example shows that an AAP can be approximated as either visible or non-visible.

3. Spatial Layout Model

We propose a novel Spatial Layout Model (SLM) which is able to model the interactions between objects, 3D scene and camera viewpoint, especially the occlusions between objects. Given an input image I, SLM predicts a set of objects $\mathbf{O} = \{O_1, \dots, O_M\}$ in the 3D world, their projections in the image plane $\mathbf{o} = \{o_1, \dots, o_M\}$ and the camera C, where M is the number of objects in the scene. SLM models the posterior probability distribution of 2D projections $\mathbf{o}$, 3D objects $\mathbf{O}$ and camera C as

$$P(\mathbf{o}, \mathbf{O}, C|I) = P(C)P(\mathbf{O})P(\mathbf{o}|\mathbf{O}, C, I) \qquad (1)$$

$$\propto P(C)P(\mathbf{O})\prod_{i=1}^{M} P(o_i|\mathbf{O}, C, I) \prod_{(i,j)} P(o_i, o_j|\mathbf{O}, C, I),$$

where $P(C)$ and $P(\mathbf{O})$ are the prior distributions over camera and 3D objects respectively, $P(o_i|\mathbf{O}, C, I)$ is the unary likelihood of 2D projection o_i given all the 3D objects, the camera and the image, and $P(o_i, o_j|\mathbf{O}, C, I)$ is the pairwise likelihood of a pair of 2D projections. Note that each 2D projection o_i depends on the configuration of all the 3D objects $\mathbf{O}$. This is because occlusions between objects in 3D affect the appearances of projections in 2D. SLM explicitly models the occlusions between objects.

3.1. 3D Object Representation

We represent the 3D objects inspired by the piecewise planar representation introduced in the Aspect Layout Model (ALM) [27]. In ALM, a 3D object consists of a set of Aspect Parts (APs). An aspect part is defined as "a portion of the object whose entire 3D surface is approximately either entirely visible from the observer or entirely non-visible" (Fig. 2(a)). While this definition is suitable for

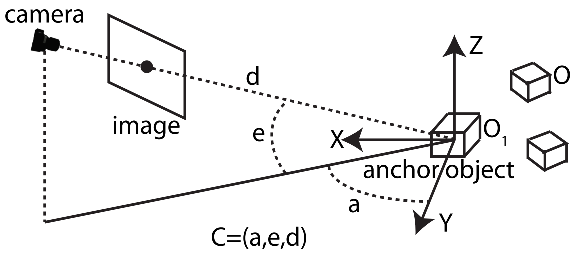


Figure 3. Camera and world coordinate system in our model.

modeling object self-occlusions (akin to those used in aspect graph representations), they are not flexible enough to handling occlusions caused by other objects in the scene (as we seek to do). For instance, it is very unlikely that an AP is entirely occluded by another object - most likely just a portion of it is occluded (Fig. 2(b)). So we propose to represent a 3D object as a collection of Atomic Aspect Parts (AAPs) which are obtained by decomposing the original APs into smaller planar parts (Fig. 2(c)). Each AAP is approximated to be either visible or non-visible (Fig. 2(d)). This approximation is less coarse if AAPs are used as opposed to APs. As we can see, smaller AAPs are better for modeling occlusions. However, smaller AAPs are harder to detect due to the lack of visual features. So there is a trade-off between the ability of AAPs to model occlusions and the reliability to detect them in the image.

3.2. Camera Prior

In SLM, 3D objects are rendered using the same internal virtual camera calibration matrix. As a result, the unknown camera parameters are the external camera matrix with respect to the world coordinate system. To define the world coordinate system, we choose one 3D object in the scene as the “anchor object”, and define the world coordinate origin as the center of the anchor object. The axes of the world coordinate system are aligned with the dominating directions of the anchor object. Then the camera location in the world coordinate system can be specified by its azimuth a, elevation e and distance d. By assuming the camera is always looking at the world coordinate origin, the unknown camera parameters to be estimated are the azimuth, elevation and distance of the camera pose, i.e., $C = (a, e, d)$. A 3D object O_i can be represented by its coordinates in the world coordinate system (X_i, Y_i, Z_i) and its relative orientation in the X-Y plane with respect to the anchor object Θ_i, i.e., $O_i = (X_i, Y_i, Z_i, \Theta_i)$. Fig. 3 illustrates the camera representation and the world coordinate system in our model. Note that different anchor objects result in different coordinates of the camera and the 3D objects. The locations of the 2D projections in the image, however, are not affected. So we can choose an arbitrary 3D object as the anchor object. We define the camera prior as

$$P(C) = P(a)P(e)P(d), \tag{2}$$

where $P(a)$, $P(e)$ and $P(d)$ are the prior distributions for the azimuth, elevation and distance respectively. We assume uniform priors for the three variables:

$$a \sim \mathcal{U}(0, 2\pi), e \sim \mathcal{U}(0, \pi/2), d \sim \mathcal{U}(d_{\text{min}}, d_{\text{max}}), \tag{3}$$

where d_{min} and d_{max} are the minimum and maximum distances of the camera we considered in the model.

3.3. 3D Objects Prior

We design the following prior to impose two constraints to a set of M objects in 3D: i) all the objects lie on the “ground plane”; ii) two objects can not occupy the same space in 3D. We model the prior distribution of 3D objects using a Markov Random Field (MRF):

$$P(\mathbf{O}) \propto \exp\Big(\sum_{i=1}^{M} V_1(O_i) + \sum_{(i,j)} V_2(O_i, O_j)\Big), \tag{4}$$

where V_1 and V_2 are the unary potential and pairwise potential respectively. Recall that the world coordinate system is defined on one of the 3D objects. If all the 3D objects lie on the “ground plane”, their Z-coordinates should be close to zero (Fig. 3). By assuming a Gaussian distribution for the objects’ Z-coordinates, we design the unary potential as

$$V_1(O_i) = -\frac{Z_i^2}{2\sigma^2}, \tag{5}$$

where σ is the standard deviation of the Gaussian distribution. Note that we do not estimate the real ground plane of the scene. The unary potential constrains that the 3D objects are all at similar heights. The pairwise potential penalizes overlapping between two 3D objects, which is defined as

$$V_2(O_i, O_j) = -\rho \frac{O_i \bigcap O_j}{O_i \bigcup O_j}, \tag{6}$$

where ρ is the parameter controlling the strength of the penalty, $\bigcap$ and $\bigcup$ denote the intersection and union between the volumes of two 3D objects. We represent the 3D objects using voxels, based on which we compute the intersection and union of two volumes (refer to [1] for details).

3.4. 3D Aspectlets

In order to obtain evidence of partial observations of objects, we introduce the concept of “3D aspectlet” inspired by the poselet framework for human detection [3]. A 3D aspectlet is defined as a portion of the 3D object, which consists of a set of the AAPs in our case. Not all the combinations of AAPs can form 3D aspectlets. We require the

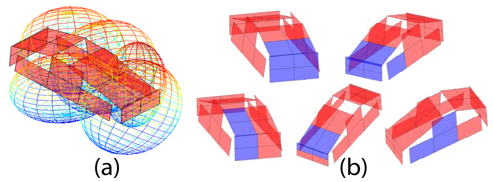


Figure 4. (a) Generating 3D aspectlet candidates by sampling ellipsoids in the space of the 3D object. (b) Examples of 3D aspectlets generated, where blue AAPs belong to the 3D aspectlets.

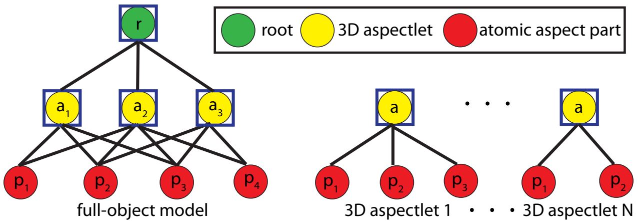


Figure 5. The graph structures of the full-object model and the 3D aspectlets, where the blue squares indicate that the bounded nodes can contain more than one template.

AAPs of a 3D aspectlet to have the two properties: i) they are geometrically close to each other in 3D; ii) there exists at least one viewpoint from which all the AAPs are visible, i.e., not self-occluded. If property ii) is not satisfied, we can represent the set of AAPs by smaller 3D aspectlets. To generate a 3D aspectlet with the two properties, we first randomly sample an ellipsoid in the 3D space of the 3D object (Fig. 4(a)), and select the AAPs inside the ellipsoid to form the 3D aspectlet. Then we check whether property ii) is satisfied. If not, we keep sampling ellipsoids until it is satisfied. Fig. 4(b) shows some 3D aspectlets of car generated in this way, where the blue AAPs belong to the 3D aspectlets.

To obtain evidence of objects from the image, we propose to represent the whole object and the 3D aspectlets an ensemble of tree models $\{\mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_N\}$. Fig. 5 illustrates the graph structures of the trees. One of the tree models $\mathcal{T}_0$ represents the whole object, which is called the full-object model. The other N tree models $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ correspond to N 3D aspectlets, which represent portions of the object. The full-object model has a three-level tree structure which consists of the root level, the 3D aspectlet level and the AAP level. The root connects to all the 3D aspectlets in the mid-level, while a 3D aspectlet connects to all the AAPs it contains. By introducing 3D aspectlets as the mid-level, the full-object model is more robust to noises in the image. In theory, all the 3D aspectlets can be placed in the mid-level level. However, this would produce a complicated tree structure which makes the training and inference infeasible. Instead, 3D aspectlets which are not in the full-object model are represented by independent two-level tree structures. In our experiments, the 3D aspectlets in the full-object model correspond to the original APs in [27].

In the tree models, the AAPs are view-invariant, which means we only need to train one part template for each AAP regardless of the number of viewpoints. This is achieved by using rectified HOG features as in [27]. But the root and the 3D aspectlets are viewpoint dependent. We train multiple templates for them, where each template captures the visual appearance of the object from a specific view section. For example, we train eight templates for the root with each template covering $45°$ azimuth. The number of templates for a 3D aspectlet depends on the range of its visible view section (i.e., not self-occluded). The blue squares in Fig. 5 indicate that there are multiple templates in these nodes. During inference, given a specific viewpoint hypothesis, only one template for each node is activated according to whether the given viewpoint hypothesis is inside its view section or not.

3.5. 2D Projection Likelihood

The 2D projection likelihood measures the compatibility between the hypothesis of the locations and poses of objects and camera in 3D and the image evidence. Let the 2D projection o_i denote the 2D location of the ith object in the image plane, i.e., $o_i = (x_i, y_i)$. We model the unary 2D projection likelihood as

$$P(o_i|\mathbf{O}, C, I) \propto P_0(o_i|O_i, C, I) + \sum_{k=1}^{N} w_k(\mathbf{O}, C) P_k(o_i|O_i, C, I), \text{ s.t. } \sum_{k=1}^{N} w_k(\mathbf{O}, C) = 1, \quad (7)$$

where $P_0(o_i|O_i, C, I)$ is the likelihood of object O_i's 2D location from the full-object model, $P_k(o_i|O_i, C, I)$ is the likelihood of object O_i's 2D location from the kth 3D aspectlet, and $w_k(\mathbf{O}, C)$ is the weight of the kth 3D aspectlet. The weights measure the reliability of the 3D aspectlets, which relates to the visibility of the 3D aspectlets. Based on the observation that 3D aspectlets with more visible AAPs are more reliable, we set the weight of a 3D aspectlet proportional to the number of visible AAPs in it and constrain that all the weights sum to one. To test the visibility of AAPs, we project the 3D objects $\mathbf{O}$ to the image plane in the order of increasing distances of the objects from the camera. During the projection, the visibility test can be performed by checking whether the 2D regions of the AAPs are occupied by some frontal objects or not (refer to the 2D object mask in Fig. 1). Consequently, different occlusion patterns between objects result in different likelihoods. Note that in the unary 2D projection likelihood, the full-object model contributes equally with all the 3D aspectlets.

To define the likelihood of a 3D aspectlet for the object's 2D location, we perform a Hough transform from the 3D aspectlet's 2D location to the object's 2D location. Let $o_{ik} = (x_{ik}, y_{ik})$ be the 2D location of the kth 3D aspectlet.

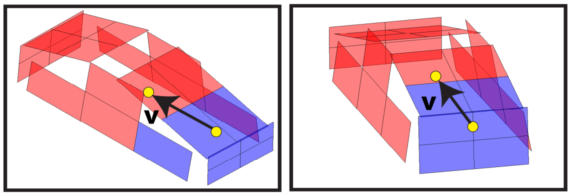


Figure 6. Illustration of the transform from the 3D aspectlet's 2D location to the object's 2D location, where the 2D projections of the two 3D aspectlets are shown in blue, and the yellow dots denote the 2D locations of the 3D aspectlets/objects in the projection.

Then

$$P_k(o_i|O_i,C,I) = \sum_{o_{ik}} P(o_i|o_{ik},O_i,C)P_k(o_{ik}|O_i,C,I), \tag{8}$$

where $P(o_i|o_{ik},O_i,C)$ is the probability distribution of the object's 2D location conditioned on the 3D aspectlet's 2D location, the 3D geometry of the object and the camera viewpoint, and $P_k(o_{ik}|O_i,C,I)$ is the likelihood of the 3D aspectlet's 2D location. $P(o_i|o_{ik},O_i,C)$ is defined as a delta function induced from the 3D-2D projection:

$$P(o_i|o_{ik},O_i,C) = \begin{cases} 1, \text{ if } o_i = o_{ik} + \mathbf{v}_{ik}(O_i,C) \\ 0, \text{ otherwise,} \end{cases} \tag{9}$$

where $\mathbf{v}_{ik}(O_i,C)$ denotes the vector from the 3D aspectlet's 2D location to the object's 2D location in the projection of the 3D object O_i according to the camera C. Fig. 6 illustrates the transform. In practice, the equality test in Eq. (9) is performed by partitioning the image into grids and testing for inside the same grid.

The likelihood of the object's 2D location from the full-object model in Eq. (7) and the likelihoods of the 3D aspectlets' 2D locations in Eq. (8) are all modeled with the same type of Conditional Random Fields (CRFs) [15] on their own tree structures (Fig. 5):

$$P_k(o_{ik}|O_i,C,I) \propto \exp\Big(\sum_{p\in\mathcal{T}_k} V_1(o^p_{ik},O_i,C,I) + \sum_{(p,q)\in\mathcal{T}_k} V_2(o^p_{ik},o^q_{ik},O_i,C)\Big), k=0,1,\ldots,N, \tag{10}$$

where p and q index nodes in the kth tree, (p,q) indicates an edge in the kth tree. $P_0(o_i|O_i,C,I) = P_0(o_{i0}|O_i,C,I)$, since there is no transform needed for the full-object model. $o^p_{ik} = (x^p_{ik}, y^p_{ik})$ denotes the 2D location of the pth node, i.e., the 2D location of the root, the 3D aspectlet or the AAP depending on the type of the node. V_1 is the unary potential modeling 2D visual appearance and V_2 is the pairwise potential which constrains the 2D relative locations between two nodes. We utilize the unary and pairwise potentials used in [27]. The unary potential is defined as

$$V_1(o^p_{ik},O_i,C,I) = \begin{cases} \mathbf{w}^{pT}_k \phi(o^p_{ik},O_i,C,I), \text{ if node } p \text{ visible} \\ \alpha^p_k, \text{ if node } p \text{ self-occluded,} \end{cases} \tag{11}$$

where $\mathbf{w}^p_k$ is the template for node p, $\phi(o^p_{ik},O_i,C,I)$ is the rectified HOG features for the node extracted from the 2D image, and α^p_k is the weight for node p if it is self-occluded. The pairwise potential is defined as

$$\begin{aligned} & V_2(o^p_{ik},o^q_{ik},O_i,C) \\ & = -w_x\big(x^p_{ik} - x^q_{ik} + d^{pq}_{ik}(O_i,C)cos(\theta^{pq}_{ik}(O_i,C))\big)^2 \\ & - w_y\big(y^p_{ik} - y^q_{ik} + d^{pq}_{ik}(O_i,C)sin(\theta^{pq}_{ik}(O_i,C))\big)^2, \end{aligned} \tag{12}$$

where w_x and w_y are the parameters controlling the strength of the pairwise constraints, $d^{pq}_{ik}(O_i,C)$ is the computed distance between the two nodes after projecting the 3D object to the 2D image according to the camera, and $\theta^{pq}_{ik}(O_i,C)$ is the relative orientation between the two nodes computed from the 2D projection. Combining Eq. (7)-(12), we can obtain the form of the unary 2D projection likelihood.

For the pairwise 2D projection likelihood, it measures how likely the occlusion between a pair of objects induced from 3D is compatible with the 2D image evidence. We design the pairwise 2D projection likelihood to reflect the observation that the occluding object usually has higher unary 2D projection likelihood than the occluded object:

$$P(o_i,o_j|\mathbf{O},C,I) \propto \exp\Big(-\frac{P(o_j|\mathbf{O},C,I)}{P(o_i|\mathbf{O},C,I)}\Big) \tag{13}$$

if O_i occludes O_j and $P(o_i|\mathbf{O},C,I)$ is larger than some threshold to make sure O_i is a confident occluder. As a result, if the occluded object has higher unary 2D projection likelihood than the occluding object, the occlusion pattern is unlikely to be correct.

3.6. Training

Training aims at learning the CRFs of our 3D object detector, which is composed of two tasks: learning 3D aspectlets and estimating the model parameters. Since it is not feasible to use all the 3D aspectlets, we select the "good" 3D aspectlets automatically. We set up the following three criteria to measure the quality of a set of 3D aspectlets. i) *Discriminative power*: the selected 3D aspectlets are discriminatively powerful. To achieve this goal, we first sample a large number of 3D aspectlets according to the sampling process described in Sec. 3.4. Then we train and test the CRFs of the 3D aspectlets on the training dataset by cross-validation. The parameter estimation of the CRFs can be performed by the structural SVM optimization [21] in [27], while the inference is conducted by Belief Propagation on the tree structure of the 3D aspectlet. The discriminative power is measured by their detection performance, based on which we select the 3D aspectlets. ii) *Viewpoint coverage*: for a specific viewpoint, there are at least K 3D aspectlets visible. Because it would be difficult to detect an object under the viewpoint if too few 3D aspectlets are available. iii) *Atomic aspect part coverage*: an AAP is contained at least in one 3D aspectlet. Otherwise, the AAP is useless. According to the three criteria, we employ an greedy algorithm to select the 3D aspectlets. The algorithm starts with an empty set of 3D aspectlets. Then it keeps adding highly discriminative 3D aspectlets into the set until the viewpoint coverage and the atomic part coverage are satisfied.

3.7. Inference

The inference problem of our spatial layout model is to search for the most compatible configuration of 2D projections, 3D objects and camera given an input image:

$$(\mathbf{o}^*, \mathbf{O}^*, C^*) = \arg\max_{\mathbf{o},\mathbf{O},C} P(\mathbf{o}, \mathbf{O}, C|I), \qquad (14)$$

where $P(\mathbf{o}, \mathbf{O}, C|I)$ is the posterior distribution defined in Eq. (1). Due to the complexity of the posterior distribution, we resort to Markov Chain Monte Carlo (MCMC) simulation to solve the inference problem. MCMC generates samples from the posterior distribution using a Markov chain mechanism. Then, the mode of the distribution is approximated by the sample with the largest probability among all the generated samples. As in [6], we compute the log-odds ratio from the maximum a posteriori estimation as the 2D detection scores. Specifically, we exploit the reversible jump MCMC (RJMCMC) algorithm [10]. In RJMCMC, new samples are proposed by different moves from the proposal distributions. The proposed samples are either accepted or rejected according to the acceptance probabilities. The reversible moves enable the algorithm to explore spaces of different number of objects.

Initialization. We initialize the MCMC sampler with high confidence detections in the image, which are obtained by evaluating the unary 2D projection likelihood (Eq. (7)) without considering occlusions between objects. The 3D objects and the camera are initialized by back-projecting the 2D detections into 3D according to the internal virtual camera calibration matrix and the estimated viewpoints of the 2D detections. A candidate set of objects is also obtained by evaluating the unary 2D projection likelihood without considering occlusions, which is used in the add moves and delete moves described below.

Add moves. Add moves add a new object O_{M+1} to the scene, where M is the current number of objects. An object in the candidate set which has not been associated with any object in the scene is randomly chosen to be added. The proposal distribution is proportional to the unary 2D projection likelihood. Since the add moves change the dimension of the state variables, specific consideration needs to be taken when computing the acceptance ratio. We map the low dimensional distribution into high dimension by assuming a constant probability $P(O_{M+1})$ for the new object:

Table 1. Statistics of the objects in our new datasets.

Category	Car	Bed	Chair	Sofa	Table
# objects	659	202	235	273	222
# occluded	235	81	112	175	61
# truncated	135	86	41	99	80

$$\hat{P}(\mathbf{o}, \mathbf{O}, C|I) = P(\mathbf{o}, \mathbf{O}, C|I)P(O_{M+1}), \qquad (15)$$

where $\hat{P}$ denotes the expanded posterior distribution. In this way, distributions of different dimensions can be compared.

Delete moves. Delete moves are the reverse moves of add moves, which remove one object from the scene and return it back to the candidate set. We adopt a uniform proposal distribution for delete moves. Similar to add moves, we map the low dimension distribution into high dimension by using a constant probability for the deleted object.

Switch moves. The switch moves change the anchor object in the scene, which prevents the model from local maximums if the anchor object is badly chosen. For example, if an object which is at different height with the other objects is selected to be the anchor object, then the other objects are unlikely to be added to the scene. The proposal distribution for switch moves is a uniform distribution.

4. Experiments

4.1. Datasets and Evaluation Measures

As far as we know, there is no dataset designed to test the ability to reason about occlusions in object detection. So we collected a new outdoor-scene dataset with 200 images of cars and a new indoor-scene dataset with 300 images of furniture for experiments, where objects are observed under various degrees of occlusion. These images are collected from PASCAL VOC [7], LabelMe [19], ImageNet [14] and our own photos. Table 1 shows the statistics of the objects in the two datasets, from which we can see they include a large number of occluded and truncated objects. The new datasets are used for testing only. To learn the 3D aspectlets and train the CRFs, we utilize the 3DObject dataset in [20] for car and the ImageNet dataset in [27] for bed, chair, sofa and table. We use the detailed ground truth annotations in [27], where each object is annotated by discretized azimuth, elevation, distance, and AP locations. The ground truth locations of AAPs and 3D aspectlets can be computed accordingly. Negative samples are from PASCAL VOC [7]. The same training datasets are used for two baselines: Deformable Part Model (DPM) [8] and Aspect Layout Model (ALM) [27]. To measure the object detection performance, we use Average Precision (AP), where the standard 50% bounding box overlap criteria of PASCAL VOC [7] is used.

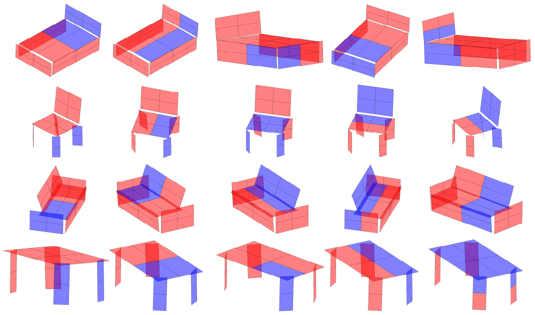


Figure 7. Sampled 3D aspectlets learnt in our experiments, where the blue AAPs belong to the 3D aspectlets.

Table 2. APs for the five categories in the two datasets.

Category	Car	Bed	Chair	Sofa	Table
ALM [27]	46.6	28.9	14.2	41.1	19.2
DPM [8]	57.0	34.8	14.4	38.3	15.1
SLM Aspectlets	59.2	35.8	15.9	45.5	24.3
SLM Full	**63.0**	**39.1**	**19.0**	**48.6**	**28.6**

Table 3. APs/mAPs on the two datasets with different test image sets according to the degrees of occlusions.

Dataset	Outdoor-scene			Indoor-scene		
% occlusion	$<.3$	.3-.6	$>.6$	$<.2$	.2-.4	$>.4$
# images	66	68	66	77	111	112
ALM [27]	72.3	42.9	35.5	38.5	25.0	20.2
DPM [8]	75.9	58.6	44.6	38.0	22.9	21.9
SLM Aspectlets	78.7	59.7	47.7	41.9	30.8	24.8
SLM Full	**80.2**	**63.3**	**52.9**	**45.9**	**34.5**	**28.0**

Table 4. 3D localization errors on the outdoor-scene dataset according to best recalls of ALM, DPM and SLM respectively.

Recall	54.8	64.6	76.8
ALM [27]	1.90	-	-
DPM [8]	2.07	2.39	-
SLM	**1.64**	**1.86**	2.33

4.2. Results

After the learning of 3D aspectlets, we obtain 50 3D aspectlets for car, and 32, 46, 24 and 25 3D aspectlets for bed, chair, sofa and table respectively. Fig. 4(b) and Fig. 7 show some learnt 3D aspectlets in our experiments, where the blue AAPs belong to the 3D aspectlets (refer to [1] for all the learnt 3D aspectlets). We compare the object detection performance of SLM with two baseline methods: the state-of-the-art object detector DPM [8] and the state-of-the-art object pose estimator ALM [27]. Table 2 shows the average precisions of SLM and the two baseline methods on the two datasets. "SLM Aspectlets" only uses our unary 2D projection likelihood for detection without considering the occlusions between objects. By using 3D aspectlets, we are able to achieve better performance than the two baseline methods, which we attribute to the ability of 3D aspectlets to detect occluded or truncated objects. However, 3D aspectlets also produce more false alarms compared with the full object model since less visual features are available. By reasoning about occlusions, our full model "SLM Full" is able to increase the detection scores of truly occluded objects and penalize false alarms which introduce wrong occlusion patterns. As a result, "SLM Full" consistently achieves the best performance on the five categories in the two datasets.

To clearly see the advantage of SLM in handling occlusions, we partition the test images in the two datasets into three sets according to the degrees of occlusions respectively, and evaluate the detection performance of SLM on each of the three sets. For an occluded object, we define its occlusion percentage as the percentage of area occluded by other objects. Then the degree of occlusion for one image can be measured by the maximum occlusion percentage of the objects in the image. Table 3 shows the number of images in each set and the APs/mAPs of the three methods on different test sets. In all the settings, SLM achieves the best performance. Besides, the improvement for the large occlusion sets is significant, which demonstrates the ability of SLM to detect occluded and truncated objects.

In order to evaluate the 3D localization accuracy, we back-project the ground truth annotations and the 2D detections into 3D respectively and obtain two spatial layouts. Since their coordinate systems can be different, we first compute the pairwise distances among objects in each layout, and then evaluate the absolute error between two corresponding pairwise distances across the two layouts. Finally, we use the mean error in the dataset as the measure for 3D localization. Since the 3D location of an object is evaluated only if it is correctly detected, we present the mean pairwise distance error according to different recalls. Table 4 shows the errors according to the best recalls of ALM, DPM and SLM on the outdoor-scene dataset, where unit one is the length of the 3D car model. SLM achieves better 3D localization at the highest recalls of both ALM and DPM.

Fig. 8 shows some anecdotal detection results from our method. The 2D detections are high confidence detections in the MAP estimations from the MCMC inference. The 3D plots show the 3D spatial layout of the objects and the camera. Based on the detected AAPs, we are able to generate the 2D mask of an object. Then according to the inferred occlusions between objects, we can refine the 2D mask to only contain the visible region of the object, from which it is possible to clearly see which object occludes which (refer to [1] for more results).

5. Conclusion

We have proposed a novel Spatial Layout Model (SLM) for multiple object detection and occlusion reasoning. SLM contextualizes objects in their 3D geometric configuration with respect to the observer to help object detection. By combining the bottom-up evidence from 3D aspectlets and the top-down occlusion reasoning, SLM is able to estimate the 3D spatial layout of objects and reason about occlu-

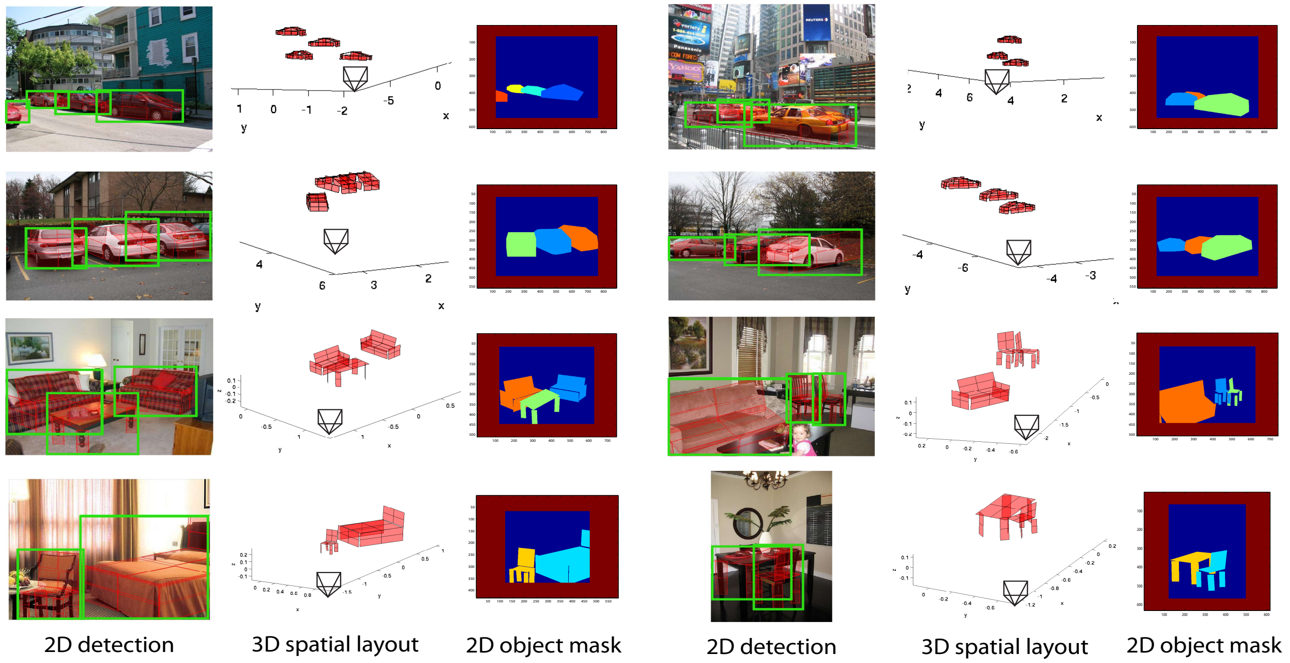


Figure 8. Anecdotal detection results on our datasets. The 2D detections show the detected objects in the images. The 3D plots show the spatial layout of objects and camera in 3D. The 2D object masks show the occlusion order in the images.

sions between objects. Experiments on two new challenging datasets with various degrees of occlusions demonstrate the ability of our model to detect objects under severe occlusions and predict the occlusion patterns in images.

Acknowledgments. We acknowledge the support of NSF CAREER grant #1054127, NSF CPS grant #0931474 and a HTC award.

References

[1] http://cvgl.stanford.edu/papers/xiang_3drr13_tr.pdf.

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

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[4] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *CVPR*, 2013.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

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

[10] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[11] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010.

[12] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[13] D. Hoiem, A. Stein, A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In *ICCV*, 2007.

[14] ImageNet. http://www.image-net.org.

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[16] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[17] A. Oliva, A. Torralba, et al. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

[18] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In *CVPR*, 2013.

[19] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008.

[20] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[21] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[22] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.

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

[24] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

[25] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR*, 2011.

[26] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005.

[27] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012.

[28] M. Z. Zia, M. Stark, and K. Schindler. Explicit occlusion modeling for 3d object class representations. In *CVPR*, 2013.