Incorporating Scene Context and Object Layout into Appearance Modeling

Hamid Izadinia*, Fereshteh Sadeghi*, Ali Farhadi
University of Washington
{izadinia,fsadeghi,ali}@cs.washington.edu

Abstract

A scene category imposes tight distributions over the kind of objects that might appear in the scene, the appearance of those objects and their layout. In this paper, we propose a method to learn scene structures that can encode three main interlacing components of a scene: the scene category, the context-specific appearance of objects, and their layout. Our experimental evaluations show that our learned scene structures outperform state-of-the-art method of Deformable Part Models in detecting objects in a scene. Our scene structure provides a level of scene understanding that is amenable to deep visual inferences. The scene structures can also generate features that can later be used for scene categorization. Using these features, we also show promising results on scene categorization.

1. Introduction

What is behind the black box in Figure 1? What are the cues that enables human vision to make an intelligent guess about the object behind the box? How can human vision go beyond category prediction and reason about details of pose, style, and material? Such predictions require complex reasoning about several interlacing components that define a scene. The fact that this picture shows a dinning room scene suggests the existence of dinning chairs, dinning table, walls, and windows. By considering the layout of the room and relative locations of the table, walls, windows and other chairs we expect to see a chair behind the black box. But do we expect to see an office chair? How about a rocking chair? By knowing the layout and the scene category we can also make strong predictions about fine-grained categories in the scene. Is the chair behind the box facing the camera? Or we expect to see a lateral view of the chair? The layout of the scene, along with the appearance of other parts in the scene suggests that the chair should be at the $3/4$ view, facing away from the camera and to the table.

We argue that the *scene* type imposes a tight distribution over the categories of *objects* in the scene, their *layout*, and also their context-specific *appearance*. For example, in an office scene we expect to see a desk, an office chair, and a monitor. The desk should probably be located against the wall, the monitor should be on the desk, and the chair should face the desk. The appearance of the chair is affected by its pose (side views of chairs look different from their front views), type (office chairs typically have one central leg whereas dinning room chairs have four wooden legs), and possible predictable occlusions that one should expect in an office (in an office, chairs are occluded by desks or other office chairs in predictable patterns).

Figure 1. What is behind the black box? Human observer can make predictions about the category of the object behind the box, its orientation, pose, material and style. By joint reasoning over scene categories, objects, their type, layout and context-specific appearances our method can make correct predictions about what is hidden behind the black box. Four image patches at the bottom are the ones selected by our method based on how well they can fill in the black box.

In this paper, we propose joint learning of scene categories, the context-specific layout of the prominent objects in the scene, and their context-specific appearance. To this

*The authors contributed equally to this work.

end, we need to know the underlying structure of object layouts. This structure can be discovered by exploiting spatially consistent relations among objects with similar appearance.

We cast the problem of joint learning of the scene category, the object layout and their appearance models as a structure learning problem where both the topology of the structure and its parameters are to be learned. The structure corresponds to the spatial relationships between objects; For example, monitors tend to appear on the desks. The layout corresponds to the locations and scales of objects in a scene. Learning the structure of the layout requires optimizing challenging objective functions that aim at incorporating the layout topology, the layout parameters, and the appearance of objects all together. In this paper we propose approximate solutions to this challenging problem.

Our experimental evaluations show that joint learning of scene categories, object categories, their context-specific layouts and appearance models improves not only object detection but also scene classification. Our method outperforms state-of-the-art Deformable Part Models (DPM [3]) as well as context based object detection [1] and provides a level of scene understanding that allows deeper inferences about scenes such as those necessary to do Black Box Test (BBT). Also, our method outperforms a strong baseline that observes the content of the image behind the box in BBT.

2. Related Work

Our work is related to the efforts done in scene recognition. Space does not allow a comprehensive review of the literature. Here, we briefly mention few related work. Scene recognition has long been regarded as a global classification task and several methods have utilized holistic features for scene categorization [21, 7, 12, 5, 23]. However, as shown in [14], holistic image features fail to provide detailed scene information that is amenable for discrimination between large number of scenes with various configuration of similar objects. Large scale indoor scene classification is an example of such a task. [14] proposed training a classifier using global features combined with local features captured from manually segmented salient regions of the scenes. Following this approach, [13, 17, 15] explore the problem of scene recognition through automatically discovering discriminative scene parts using various techniques. Pandy and Lazebnik [13], utilized Deformable Part Models (DPM) object detector [3] to automatically find the salient scene regions. Sadeghi and Tappen [15] represent a scene via discriminatively discovered scene parts called Latent Pyramidal Regions (LPR). [17] and [2] learn scene parts in a jointly unsupervised/weakly-supervised manner. All these recent approaches [13, 15, 17] seek the use of *parts* as an intermediate representation of scenes but do not provide any semantics for the discovered parts. [19] presents an exemplar-based approach to image parsing. In an earlier effort, Xiao et. al. proposed the extensive Scene UNderstanding (SUN) dataset with 899 scene categories [22]. In [22] a subset of well-sampled categories (397 categories) of SUN are used to evaluate the state-of-the-art holistic features in scene recognition. To the best of our knowledge, non of the recent scene classification algorithms has considered SUN dataset for evaluation. This is mainly due to the large number of categories and images in this challenging dataset.

On the other side of the spectrum, an object-centric approach represent an image as a pool of pre-trained object detectors (called Object Bank [8]). Scene recognition is performed by learning a scene category classifier using the object score-map as new features [8]. The main limitation of this approach is that the object detectors fail to provide accurate object localization because object models are learned independent from each other and the scene information. In addition, several scene components (e.g. sky, grass, wall, road) can hardly be modeled with object templates while they can be efficiently recognized if the context model is taken into account. In our method we show that learning the scene structure and the layout of prominent scene objects (semantic scene parts) can boost the performance of both scene recognition and object localization.

For improving object detection, [1] proposes an extension of non-maximum suppression that uses contextual information in the form of a single cooccurrences relation tree. [1] takes independently trained detector responses as input and smartly prunes out contextually irrelevant ones which results in improveing precision but not recall. However, we have scene specific trees that takes into account the scene specific appearances of objects (chairs in offices look different from chairs in family rooms) and their spatial/contextual relationships in a single framework. Also, we jointly train our detectors and contextual model and improve both object detection and scene recognition.

3. Our Approach

Learning the underlying structure of the layout entails reasoning about the appearance of the scene, the context-specific appearance and layout of prominent objects in the scene, and the topology of objects layouts. We cast this joint learning problem as learning the underlying layout structure and the structure parameters.

To setup the notations, assume that $\mathcal{X} = \{x_1, ..., x_n\}$ is the set of training images that belong to one of the scene categories $\mathcal{C} = \{c_1, ..., c_m\}$. Each image x_i might depict a set of objects $\mathcal{O}_i = \{o_i^1, ..., o_i^{q_i}\}$ where q_i corresponds to the number of objects in the i^{th} image. Each training image has a ground truth layout $\mathcal{H}_i$ that indicates the location of bounding boxes of each object $\mathcal{H}_i = \{h_{i,1}, h_{i,2}, ..., h_{i,q_i}\}$, $h_{i,1}$ corresponds to the location and scale of the first object in the i^{th} image. Each scene category c imposes a latent structure (topology) $\mathcal{G}_c$ over the layout. A scene structure $\mathcal{S}$ for the scene c correspond to the layout of objects, their relative locations, and their appearance models $\mathcal{S}^c = \{\mathcal{W}_c^a, \mathcal{W}_c^d, \mathcal{G}_c\}$ where $\mathcal{W}_c^a = \{W_{c,1}^a, W_{c,2}^a, ..., W_{c,p_c}^a\}$ is the

set of p_c appearance models for the objects in the c^{th} scene, and $\mathcal{W}_c^d$ corresponds to the set of weight vectors that encode relative locations of objects $\mathcal{W}_c^d = \{W_{c,j,k}^d | j,k = 1 : p_c\}$. The function $\mathcal{D}$ measures how well a scene structure $\mathcal{S}^c$ is aligned with an observation x_i:

$$\mathcal{D}(x_i, \mathcal{H}_i, \mathcal{W}_c^a, \mathcal{W}_c^d, \mathcal{G}_c) = \sum_{j=1}^{p_c}(W_{c,j}^a \phi(x_i, \mathcal{H}_i, \mathcal{G}_c) + \sum_{k=1}^{p_c} W_{c,j,k}^d \psi(\mathcal{H}_{i,j}, \mathcal{H}_{i,k}, \mathcal{G}_c)) \quad (1)$$

where ϕ encodes unary appearance features (in our case vectorized HOG features) and ψ corresponds to binary deformation features (in our case quadratic distance transform function [3]). Discovering scene structures for a scene category c can be formulated as a structure discovery problem:

$$\begin{aligned} &\min_{\mathcal{G}_c, \mathcal{W}_c^a, \mathcal{W}_c^d, \mathcal{H}, \xi} \quad \sum_{j=1}^{p_c} \|W_{c,j}^a\|_2^2 + \\ &\sum_{j,k=1}^{p_c} \|W_{c,j,k}^d\|_2^2 + \lambda_1 \sum_{i=1}^{n} \xi_i + \lambda_2 \|\mathcal{G}_c\|_\bullet \\ &\mathcal{D}(x_i, \mathcal{H}_i, \mathcal{W}_c^a, \mathcal{W}_c^d, \mathcal{G}_c) \geq \\ &\max_{\mathcal{H}^*} \mathcal{D}(x_i, \mathcal{H}_i^*, \mathcal{W}_c^a, \mathcal{W}_c^d, \mathcal{G}_c) + \Delta(\mathcal{H}_i, \mathcal{H}_i^*) - \xi_i \quad \forall i \\ &\xi_i \geq 0 \quad \forall i \end{aligned} \quad (2)$$

where $\|\mathcal{G}\|_\bullet$ is a form of complexity regularizer over the topology of the structure, and Δ is a form of structured loss.

Joint optimization over the topology of the structure and the parameter of the structure, the appearance and deformation models of objects, is extremely challenging. This is an optimization over highly interleaving parameters $\mathcal{G}$ and $\mathcal{H}$. In fact, this is an NP-Hard problem. To approximate this hard optimization we decouple the optimization over $\mathcal{G}$ from the rest of the parameters. If $\mathcal{G}$ is known, we can rewrite the optimization 2 as a form of margin based structure learning problem. Fixing $\mathcal{G}$ and putting $\mathcal{W}^a$ and $\mathcal{W}^d$ together into $\mathcal{W}$ results in:

$$\begin{aligned} &\min_{\mathcal{W}, \mathcal{H}, \xi} \quad \|\mathcal{W}\|_2^2 + \lambda_1 \sum_{i=1}^{n} \xi_i \\ &\mathcal{D}(x_i, \mathcal{H}_i, \mathcal{W}) \geq \max_{\mathcal{H}^*} \mathcal{D}(x_i, \mathcal{H}_i^*, \mathcal{W}) + \Delta(\mathcal{H}_i, \mathcal{H}_i^*) - \xi_i \\ &\xi_i \geq 0 \quad \forall i. \end{aligned} \quad (3)$$

We use hamming loss for the structured loss Δ and solve this optimization problem by cutting plane method [18, 24].

Optimizing for $\mathcal{G}$ is a challenging problem and requires approximation. Learning for $\mathcal{G}$ entails reasoning about which objects will make it to the final layout and what is the topology of the graph connecting these prominent objects. To approximate this, we make use of domain knowledge. Distribution of objects in scenes is known to follow Zipf's law [22]. Meaning that there are large number of objects which occur very rarely in each scene and have small correlation with other objects in the scene. At the same time, a limited subset of objects exposes strong correlation across instances of a scene category. This suggests pruning scene-specific objects that are rare and relations (edges) that are spatially inconsistent. To this end, we form a Scene-Object graph $SOG_c = (V_c, E_c)$ whose nodes correspond to objects that appeared in scene c. The edges E_c correspond to spatial consistency of two objects with respect to each other across samples of scene c. Starting from a full graph SOG_c, discovering $\mathcal{G}$ can be formulated as selecting a set of nodes and edges that maximizes the prominence of objects and spatial consistency of their relations. Since the discovered structure will be used for further inferences, a crucial constraint would be to avoid loops in the resultant structure. More formally, the discovered $\mathcal{G}^*$ can be represented as the optimized set of nodes V^* and edges E^* such that:

$$\begin{aligned} &\max_{\sigma_e, \sigma_v} \quad \sum_{v \in V} \sigma_v \Omega(v) + \sum_{e \in E} \sigma_e \Gamma(e) \\ &\text{Subject to} \\ &\sum_{v \in V} \sigma_v \leq p_c \\ &\sum_{e \in E(\mathcal{N})} \sigma_e \leq |\mathcal{N}| - 1, \quad \forall \mathcal{N} \subset V, \mathcal{N} \neq \emptyset \\ &\sigma_e \in \{0,1\}, \sigma_v \in \{0,1\} \end{aligned} \quad (4)$$

where σ_e and σ_v are binary indicator variables that indicate which nodes and edges will make it to the final scene structure, $\Omega(v)$ is proportional to the prominence of each object in a scene, and $\Gamma(e)$ is proportional to the spatial consistency between two objects in a scene and p_c is the total number of objects in scene c. The second constraint avoids loop in the final structure. This optimization can be reduced to a form of weighted maximum spanning tree problem. We initialize the tree with the vertex with maximum $\Omega(v)$ and grow the tree by adding one edge at a time which brings the maximum gain to the equation 4. We stop this process until there is no edge that increases the total gain more than a certain threshold. In our experiments the prominence function $\Omega(v)$ corresponds to the frequency of object v in the instances of scene c. The spatial consistency $\Gamma(e)$ is set to $\frac{1}{\sigma}$ where σ is the variance of the Gaussian distribution for the relative spatial location of each pair of objects.

Our learned scene structures include a set of context-specific appearance filters for prominent objects in a scene category and their corresponding deformation models. To be more expressive, we use mixture models for the appearance and deformations for prominent objects. This allows our model to capture context-specific variations within a scene category. Our model encodes the relationships between the layout of the objects and their appearance. For example, the pose of a chair affects the appearance model of the nearby desk.

Inference involves computing equation 1 for all possible scene structures and picking the highest scoring one for each image. More specifically, we first find the part convo-

lution scores for all mixtures and take the maximum among mixtures for computing part score in every location of the image. Then distance transform is used to efficiently pass messages from child nodes to their parents. The maximum collected score in the root node is considered as the best scene structure. In the message passing, we save the location of maximum score for each node and use them as the detection of all scene parts for the best inferred structure.

This provides not only the scene category labels but also the layout of prominent objects and their spatial relationships. The scene structures provide a level of understanding for scenes that is much richer than just scene category labels. Our experimental evaluations show that our model can localize objects in a scene significantly more accurate than state-of-the-art object detectors that trained independently.

4. Experiments

We train our scene structures using the SUN dataset that includes 397 scene categories. We use 390 categories which contained enough annotated images. We experimentally evaluate the benefits of using the discovered scene structures in object detection, the Black Box Test and also scene recognition.

4.1. Object Detection

At inference, the maximum scoring scene structure contains information about objects, their appearance and also layout. Layout contains bounding box information about prominent objects in a scene.

To better encode the appearance and deformations of stuff we consider a three part model for each stuff. This allows our model to detect stuff more reliably by leveraging the structured boundaries of stuff with other prominent objects in the scene. For example, for the abbey scene, our model can reliably localize sky be leveraging the fact that the boundaries of sky with the abbey structure produces a very context-specific pattern (Fig. 2). Fig. 3 also includes interesting detection results for stuff and also objects.

We compare the performance of object detection using our scene structure versus that of Deformable part models trained in Object Bank (OB) [8]. We use 7249 annotated images in SUN database from all 390 scene categories. For evaluation, in each image we only consider the labeled objects for which there is a node in our method and also an object detector trained in OB. The precision of object localization is measured by computing the intersection over union of object localization mask B_l and ground truth polygon P_{gt} (i.e. $\frac{B_l \cap P_{gt}}{B_l \cup P_{gt}}$).

For DPM models in OB, we apply models on each image and compute the score map in different levels of feature pyramid. Then the score of each level is propagated in a window with the same size as that of HOG filter as object mask. The score of all levels is pooled over each pixel using max pooling. Similar to standard object detection criteria we threshold the object score map with different thresholds and compute the object localization precision for each threshold. The threshold with the best precision is used for comparison with our method in Fig. 4.

As shown in Fig. 4 the scene layout information encoded in our scene structures help object detection. In fact, our method outperforms DPM by large margins. Our mean Average Precision is 24.10 compared to 19.13 of DPM. Examples of detected objects along with their best scoring scene structures that produces those detections are shown in Fig. 3. It is interesting to see that some of the occluded objects have been correctly detected using our method. Scenes in our experiments vary from indoor to outdoor, from scenes with more dominant layout such as "Street" to scenes with very complex layout like "Dorm room". Our method can also localize stuff such as sky, grass, road, etc. See Fig. 3 for examples. We have also compared our method with the context modeling approach of [1] using their publicly available context tree model for object detection on SUN dataset. The average precision of [1] is 22.22% compared to 24.10% of our method.

4.2. The Black Box Test

Our scene structures provide a level of scene understanding that enables deep inferences such as the one used in Black Box Test. This test is inspired by Antonio Torralba's Context challenge [20] and has also been studied in [9]. To be successful in BBT one needs to understand the scene category, and the context-specific layout and appearance models. For example, in Fig. 1, our model understands that because of the location of the wall and the table the expected chair should be in $3/4$ view. To test the performance of our model in BBT, we randomly select 120 images and black out examples of stoves, chairs, windows, beds, sofas, and cars in these images. We then use these blacked out images to find the highest scoring scene structures. To avoid artifacts due to the black box, we cancel out the effects (both appearance and deformation) of nodes that have any overlap with the black box. The best scoring scene structure contains information about the layout of the scene. This structure also encodes information about the missing part. For example, if there is a wall in a specific location and there is desk at the wall, there should be a chair with a specific pose at the desk. We use the part filters in the best scoring structure that has enough overlap with the black box to retrieve image patches. Fig. 1 and 6 show samples of image patches suggested by our method. Note that, our method not only finds the object of the right category but also its right pose.

To evaluate the quality of suggested patches by our method, we compare it against a baseline that can see what is behind the box. We also use the information about what objects we expect to see. This optimistic baseline uses the HOG2x2 descriptors of the whole image (including the object behind the box) to retrieve images of similar appearance. We then use the corresponding DPM model to obtain the best scoring patches among the retrieved images with similar appearance.

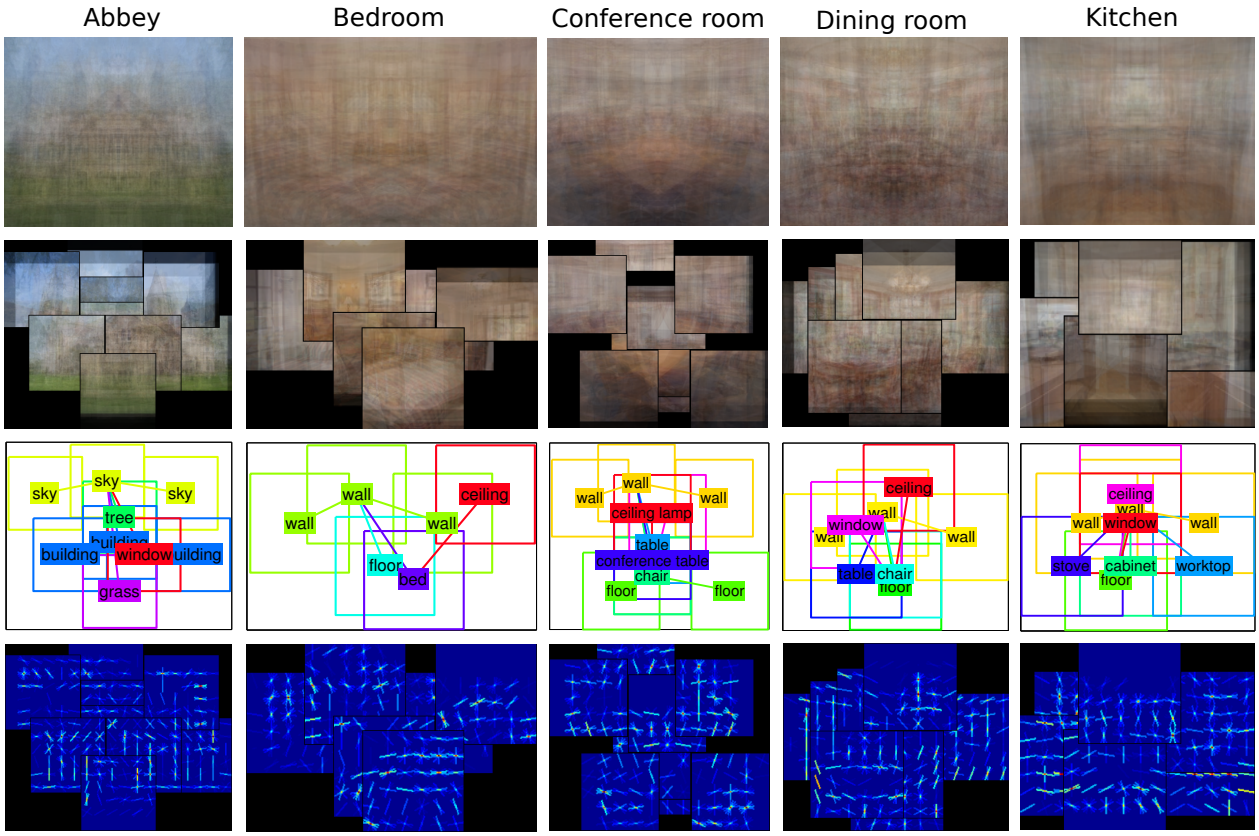

Figure 2. The visualization of the model for five scenes. Each scene is shown in different column. Row 1: the average of training images which shows if we use one filter for the entire image the model would be vague. Row 2: the average of image patches after finding best scene layout in each training image. The constellation of scene parts shows the discriminative shapes in each scene such as bed in bedroom and table in Dining room. Row 3: the semantic object label for each scene part and their learned layout as a tree. Row 4: visualization of the appearance models learned for each scene. Note that both the appearance models and their learned locations are context-specific.

To compare the quality of the patches retrieved by our method to that of the baseline we perform a human subject forced choice task where subjects were asked to choose between the patches produced by our method and those of the baseline. For each image, we gather between 3 to 4 annotations. On average, the annotators preferred our patches to those of the baseline on 74.74% of cases. Fig. 5 shows the results of the human subject test on BBT for different categories of objects. For objects such as Stove and window that appear in more structured scenes like kitchen and rooms our method shows larger gain compared to objects like cars in less structured scenes such as streets. Fig. 6 shows qualitative results of the BBT.

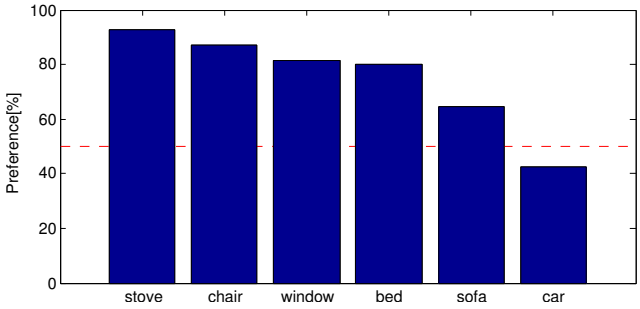

Figure 5. Human subject experiments for the Black Box Test: In our forced choice human subject task, annotators prefered our method on 75% of the cases compared to a baseline that can actually see what is behind the box. For objects like Stove and Bed which appear in more structured scenes our method produces better results compared to objects like cars that tend to appear in less structured scenes.

4.3. Scene Category Recognition

We also exploit the scene structures to generate features that can later be used for scene recognition. To this end, we run all of our scene structure models on all test images and pick the k best scoring structures per scene category. We then record the structure scores of the k best structures as features and append the convolution scores of the objects in the best structures, and their normalized locations. We also include relative locations of objects using the parents in the best scoring structures. Then, for each scene category, we train an SVM with HI kernel using this feature vector.

To evaluate the performance of the structures in scene categorization, we use SUN as well as the MIT indoor-67 dataset as test beds. In both these datasets we report the average per class accuracy obtained by our method and compare it with other state-of-the-art methods. We use half

Figure 3. Sample of best scoring scene structures that lead to accurate object detection in several challenging cases. In this image we show four different categories: Street, Office, Kitchen and Dorm room. Each row shows samples of one scene category. Discovered scene structures are superimposed into each image and objects are color-coded according to the legend on the right most column. Our model can detect the objects accurately even in very different layouts of a scene. Stuffs such as buildings and walls are also detected precisely using our flexible mixture of parts. The edges between objects correspond to the layout discovered by our method. For example, our method discovered that chairs are typically at desks and desks are typically located by the wall or windows in office scenes.

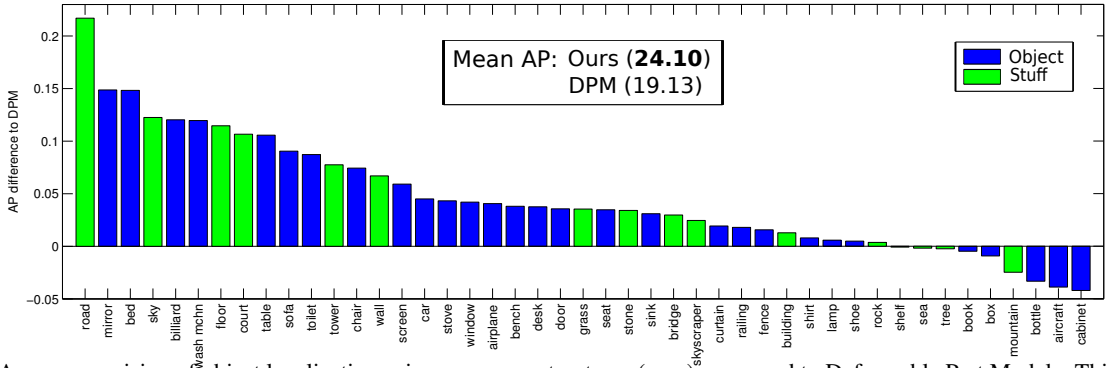

Figure 4. Average precision of object localization using our scene structures (ours) compared to Deformable Part Models. This plot shows the gain over DPM. Positive values corresponds to the case where our method outperforms DPM and negative values correspond to cases where DPM works better than our method. Our method outperforms DPM by significant margin. The biggest gain corresponds to objects or stuffs that are hard to detect but can be detected with the help of contextual information encoded in our scene structure. The green bars correspond to stuffs and blue bars to objects. For very small objects like bottle or objects that typically appear in less structured scenes like airplanes DPM performs better. For most of the stuff our method outperforms DPM.

of the annotated images in each SUN category for learning scene structures. For categorization task on SUN, we randomly choose 100 (50 test, 50 train) images from the unannotated portion of each category which is not used in the training set of our structure learning. For the MIT indoor-67 we use the standard train/test split which is available in [14]. In scene recognition task on MIT indoor-67 dataset we use the scene structures trained on SUN. It is not possible to train our scene structures for MIT indoor 67 because object-level annotations are not available.

Tables 1 and 2 compare our results with the state-of-the-art methods in scene recognition. Following [22, 13, 17] we also combine our result with other state-of-the-art holistic features to encode global scene appearance information. We

Figure 6. Black Box Test: What is behind the black box? Our scene structures provide a level of scene understanding that allows deep inferences such as the one necessary to complete the back box test. Our method not only predicts what is behind the box but it also provides interesting detailed information about object poses.

use Locally-constrained Linear Coding(LLC) [21], Self-Similarity (SSIM) [16], Local Binary Patterns (LBP) [11] and Texton [10] features. We observe that combining our feature vectors with other global features boosts our recognition performance in both MIT indoor-67 and SUN database. Our scene structures provide information orthogonal to the state-of-the-art holistic features.

According to the results of Table 1 scene structures produce promising results in scene categorization (accuracy of 45.91%). When combined with other state-of-the-art holistic features, the accuracy boosts to 52.41% .Note that our scene structures are trained on SUN dataset and tested on MIT indoor-67 dataset whereas other methods have trained their models on the MIT indoor-67 dataset.

Table 2 compares our scene recognition results with state-of-the-art models on SUN database. We have provided the recognition accuracy of the state-of-the-art holistic features using our test and train splits. The best single-feature scene classification accuracy is 27.2% which is obtained by HOG2x2 [22]. As reported in [22], the recognition accuracy can be boosted up to 38% by combining 15 different feature types. According to Table 2, our method obtain an accuracy of 28.45% in SUN. After combining our method with four holistic features (LLC, SSIM, Texton and LBP) we have reached the accuracy of 35.95% which is on par with the combination of 15 features. Note that [6, 2] are using very high dimensional features to produce state-of-the-art results.

5. Discussion

We introduced a model that can link up the scene categories, the context-specific layout and appearance models for prominent objects and stuff in a scene. Our experimental evaluations show that the learned scene structures are capable of localizing objects significantly better that state

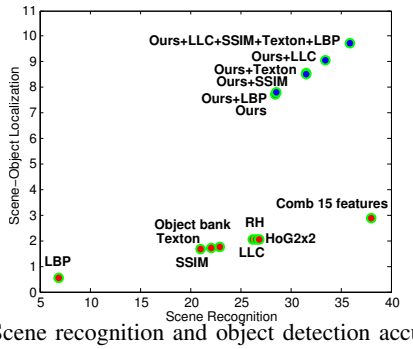

Figure 7. Scene recognition and object detection accuracy spectrum. Our method can recognize the scene label and detect objects simultaneously. For other methods since they do not have object detection step, we use object detectors of Object bank for their object detection step.

of the art detectors. We also show that our scene structure enables the level of scene understanding that is amenable to deep inferences such as those required in BBT. We compared our method with a baseline that sees what is behind the black box and show significant improvement in a forced choice human subject task. We also show promising results in scene recognition.

Our scene structures enables both scene categorization and object localization. To better understand the space of scene understanding methods in terms of both object localization and scene categorization we propose to consider a joint evaluation. Fig. 7 compares the performance of our method in the scene recognition and scene-object localization tasks with other state-of-the-art scene recognition methods. In this graph, the x-axis shows the scene recognition accuracy whereas the y-axis represents the scene-object localization precision. This can be measured as multiplication of scene recognition performance and scene-object localization. Most conventional scene recognition methods focus on the categorization and do not provide object-level

Method	Accuracy	Method	Accuracy
LBP	18.12	GIST-color+SP+DPM [13]	43.1
HOG [13]	22.8	LPR [15]	44.84
ROI-GIST [14]	26.5	Ours	45.91
GIST-color [13]	29.7	Ours+LBP	47.64
DPM [13]	30.4	Ours+Texton	49.36
SSIM	33.45	Ours+LLC	49.38
Spatial Pyramid (SP)[7]	34.4	MLD Patches+GIST+SP+DPM [17]	49.4
Texton	35.98	Ours+SSIM	49.62
Object bank [8]	37.6	Ours+LLC+SSIM+Texton+LBP	52.41
LLC	37.53	BoP+IFV [6]	63.10
MLD Patches [17]	38.1	Midlevel elements+IFV [2]	66.87

Table 1. Scene categorization results on MIT indoor: The average per-class accuracy results on MIT indoor-67 dataset.

Method	Accuracy	Method	Accuracy
LBP	6.84	Ours	28.45
SSIM	21.06	Comb 15 features [22]	38
Texton	22.04	Ours+LBP	28.59
Object bank [8]	22.93	Ours+Texton	31.57
LLC	26.23	Ours+SSIM	31.58
RH [4]	26.9	Ours+LLC	33.45
HoG2x2 [22]	27.2	Ours+LLC+SSIM+Texton+LBP	35.95

Table 2. Scene categorization results on SUN: The average per-class accuracy results on SUN database.

information. To make a fair comparison, we use the available pre-trained detectors of OB [8] as the object detection step for state-of-the-art methods. We assume that for other methods we first run the scene recognition for each image and then use scene label for applying the corresponding object detectors on that image. As shown in Fig. 7, our method outperforms other state-of-the-art methods in terms of both scene recognition accuracy and object detection precision. In this plot an ideal method would be located on the top right corner. Our method takes advantage of the scene layout to improve the object localization and uses the object localization to improve scene recognition.

Acknowledgement

The authors would like to thank Larry Zitnick for helpful discussions and insightful remarks. This research was in part supported by NSF IIS-1218683, and ONR N00014-13-1-0720.

References

[1] M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *IEEE Trans. PAMI*, 2012.

[2] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*. 2013.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 2010.

[4] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[5] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[6] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[8] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[9] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS*, 2009.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[11] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 2002.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[13] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[14] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[15] F. Sadeghi and M. F. Tappen. Latent pyramidal regions for recognizing scenes. In *ECCV*, 2012.

[16] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[17] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. 2012.

[18] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. *NIPS*, 2004.

[19] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[20] A. Torralba. The context challenge `http://web.mit.edu/torralba/www/carsAndFacesInContext.html`.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[22] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[23] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[24] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. PAMI*, 2013.