

# Indoor Scene Recognition from RGB-D Images by Learning Scene Bases

Shaohua Wan                      Changbo Hu                      J.K. Aggarwal  
 shaohuawan@utexas.edu      changbo.hu@gmail.com      aggarwaljk@mail.utexas.edu  
 Dept. of Electrical and Computer Engineering  
 The University of Texas at Austin  
 Austin, Texas 78712

**Abstract**—In this paper, we propose a RGB-D indoor scene recognition method that has mainly two advantages as compared to existing methods. First, by training object detectors using RGB-D images and recognizing their spatial interrelationships, we not only achieve better object localization accuracy than using RGB images alone, but also obtain details as to how the objects are related to each other in a spatial manner, thus resulting in a more effective high-level feature representation of the scene known as the Objects and Attributes (O&A) representation. Second, we learn class-specific sub-dictionaries that capture the high-order couplings between the objects and attributes. In particular, elastic net regularization and geometric similarity constraint is imposed to increase the discriminative power of the sub-dictionaries. The proposed method is evaluated on two RGB-D datasets, the NYUD dataset and the B3DO dataset. Experiments show that superior scene recognition rate can be obtained using our method.

## I. INTRODUCTION

Scene understanding is an active research topic in computer vision. In the past decade, scene understanding has mainly dealt with 2D RGB images. The Bag-of-Visual-Words (BoW) model, which extracts local features from interest points and aggregates the statistical properties of the appearance of the scene in a histogram, has achieved significant success ([1], [2], [3], [4]). Although promising results were shown from using these low-level representations, it is unclear why such a histogram representation should be optimal for the scene recognition problem. More recent research suggests that high-level representations, which explore the semantically meaningful components of a scene, such as objects and salient regions, are more effective for scene recognition ([5], [6], [7], [8]).

In this paper, we propose a new method for indoor scene recognition from RGB-D images. We use a high-level image representation, called *Objects and Attributes (O&A)*, that is built upon a set of object detectors and attribute classifiers<sup>1</sup>. Starting from publicly available RGB-D image sets annotated with object-level bounding boxes, we train *object* detectors using both the RGB and the depth information. RGB-D based object detection has been shown to produce superior localization results than that of using RGB image alone [10]. Then, to characterize the spatial layout of indoor scenes, we define a number of *attributes* to describe how the objects are

<sup>1</sup>The term *scene attributes* has been used with different meanings. In [9], scene attributes refer to scene properties that are related to materials, surface properties, lighting, etc. In this paper, scene attributes are used to describe the spatial relation between objects.

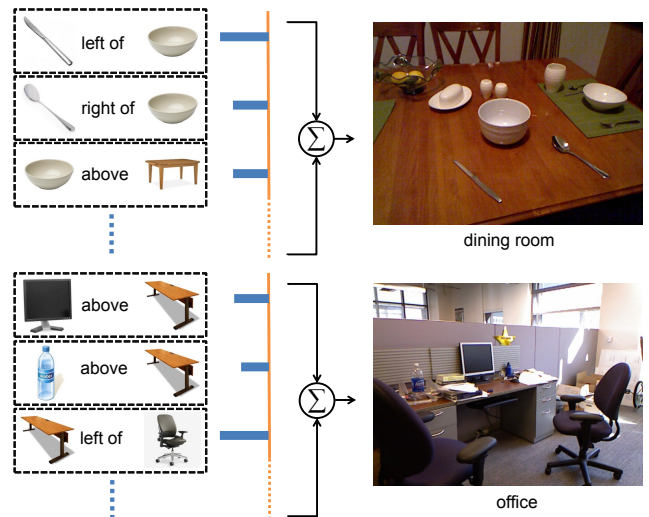


Fig. 1: We use Objects and Attributes (O&A) to represent a scene. Given a training set of images represented by the O&A, we then learn a set of class-specific sub-dictionaries, each of which encodes the coupling between highly correlated objects and attributes of each scene class. For a new scene image, its O&A representation can thus be reconstructed as a sparse weighted sum of the scene bases in these sub-dictionaries. In this figure, the left part indicates the detected objects and attributes. The height of the blue bars indicate the importance of the coupling of the corresponding object pair and attribute in the scene basis.

related to each other in a spatial manner. The objects and attributes complement each other in describing indoor scenes. A possible description of the office scene in Figure 1 using the Objects and Attributes representation is the following: *The monitor is above the desk. The water bottle is above the desk. The desk is to the left of the chair and close to it.*

Equipped with the Objects and Attributes representation of a scene, the task of scene recognition is solved by learning a dictionary of scene bases. Instead of learning one over-complete dictionary for all classes, we learn class-specific sub-dictionaries to increase the discrimination. In contrast to existing class-specific dictionary learning methods that are based on  $l_1$ -norm sparsity constraint, we impose the elastic net

regularizer to ensure that feature vectors are well reconstructed by scene bases from the same class. Moreover, the geometric similarity between the features are incorporated during the process of dictionary learning, so that features with high similarity will tend to have similar coefficients.

Figure 1 illustrates our approach. We capture the coupling between the objects and attributes by learning a set of class-specific sub-dictionaries. Given a new indoor scene image represented as a vector of detected objects and attributes, we reconstruct it as a sparse weighted sum of the scene bases of these sub-dictionaries. Since we use the learned scene bases to reconstruct the scene, our method is expected to correct false detections of objects and attributes. For example, in the learned scene bases, the "beside" attribute is more likely to be the relationship between a "table" and a "chair", than between a "computer" and a "chair".

The rest of this paper is organized as follows. Related work are described in Sec.2. The Objects and Attributes based representation of indoor scenes and the class-specific dictionary learning are elaborated in Sec.3 and Sec.4 respectively. Experimental results are given in Sec.5.

## II. RELATED WORK

Numerous efforts have been devoted to the area of scene recognition. The *Bag-of-visual-words* model (BoW) [1], [3], [4], which represents an image as an orderless collection of local features (e.g. SIFT [11]), has demonstrated impressive scene recognition performance because they can be reliably detected and matched across objects or scenes under different viewpoints or lighting conditions. [12], [13], [14] take into account the collocation patterns of visual words and learn "visual phrases" for high-level scene/object recognition. [15] incorporates the spatial information of local features into the BoW representation of images for better scene recognition performance. Topic models, which take a three-layer hierarchical Bayesian view of the generation of image, recognize scene classes by incorporating a semantic layer of latent topics in between the scene classes and visual words [16], [17], [18], [19].

In contrast to the above-mentioned methods which recognize scene classes using low-level features, Li et al. [5] propose a model called the Object Bank (OB), which represents an image as the response map of a large number of pre-trained generic object detectors. Zheng et al. [7] propose an object part model, named Hybrid Parts, for scene recognition based on the detection of fine-grained local object parts. Pandey et al. [6] optimally select the Region of Interest (ROI) to be those containing salient objects, and classify a novel scene based on the fact that similar scenes contain similar ROIs.

Motivated by the recent work of modeling the spatial relations between objects using language constructs such as "prepositions" for the task of object annotation [20], we propose an object-level image representation for the task of scene recognition which models the relations between objects via a set of attributes (e.g. above, below). Furthermore, we learn a set of class-specific sub-dictionaries and use these sub-dictionaries to robustly reconstruct each scene. Both the elastic net regularization and geometric similarity constraint

is incorporated to improve the discriminative power of sub-dictionaries. Extensive experiments on two existing RGB-D datasets demonstrate the effectiveness of our method on indoor scene recognition tasks.

## III. OBJECTS AND ATTRIBUTES

Each image in the training set is represented by the objects and attributes detected in the image. We now describe how to train classifiers to automatically recognize these objects and attributes.

### A. Object Detection

To detect objects within RGB-D images, we follow the implementation of [21]. In particular, we use a concatenation of the HOG features extracted from both the RGB and the depth channel as input to the SVM classifier.

The linear SVM classifier scans across the RGB-D image at all positions and scales. In total, we use a pyramid of 20 image scales to capture the visual and depth features ranging from coarse to fine. The score function of the linear SVM classifier for object category  $o \in \{1, \dots, O\}$  can be written as

$$f_o(r, p) = \mathbf{w}_o^T g(r, p) + b_o \quad (1)$$

where  $r$  and  $p$  denotes the scale and position of the image respectively,  $g$  denotes the features extracted from  $(r, p)$ , and  $\mathbf{w}_o$  and  $b_o$  is the weight vector and bias respectively. The score map  $s_o(p)$  is the maximum linear SVM classifier score at all scales, i.e.  $s_o(p) = \max_r f_o(r, p)$ .

We fit a sigmoid function to  $s_o(p)$  to compute a probabilistic map  $h_o(p)$  which can be seen as the detector confidence of an instance of object class  $o$  at position  $p$

$$h_o(p) = \frac{1}{1 + \exp\{a \cdot s_o(p) + b\}} \quad (2)$$

where  $a$  and  $b$  are constants that are learned from the training data using maximum likelihood estimation. Detections in the probabilistic map that are less than a specific threshold are discarded. We further remove overlapping detections via non-maximum suppression. The remaining detections are accepted as valid detections. See Figure 2 for example object detection results in different indoor scenes.

### B. Attribute Classification

Inspired by theory of spatial relation comprehension developed by Logan et al. [22], we use ten predefined attributes to describe the spatial relations between a pair of objects in a scene: *in front of*, *behind*, *right of*, *left of*, *above*, *below*, *contain*, *contained by*, *close to*, *far from*. Note that these attributes can be easily expanded to account for more complex spatial relations such as *surround*, *hang from*, *protrude from*, etc.

Following [23], we develop ten spatial templates to map the geometric information estimated from the RGB-D data to a confidence score indicating the likelihood of a pair of objects having a particular attribute. Specifically, the ten spatial templates are applied sequentially to the pair of objects and the

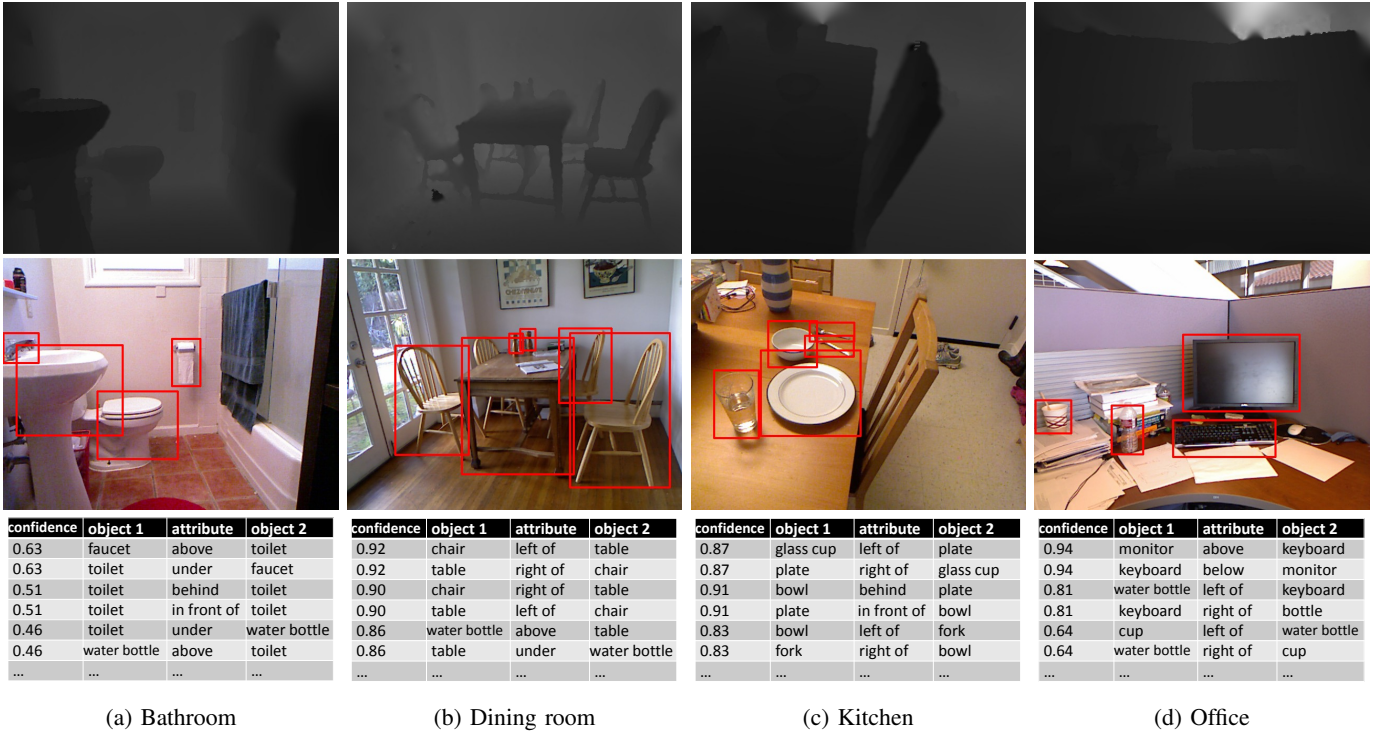


Fig. 2: The objects and attributes detected in different scenes. The first row shows the depth images of each scene, and the red bounding boxes in the RGB images in the second row shows the detected objects. The third row shows the O&A representation of each scene output by our object and attribute classifier. The object and attribute classifiers are not perfect, as demonstrated by some false detections.

attribute with the highest confidence score is kept. A spatial template is defined as a fuzzy membership function that returns a value indicating the applicability of an attribute for a pair of objects, behaving rather like a receptive field.

### C. Objects and Attributes Representation

With the detected objects and attributes, each image is represented by a feature vector  $\mathbf{y} \in \mathcal{R}^M$ , where  $M$  is the number of all possible object and attribute interactions, and each dimension is the probability of a pair of objects having a particular attribute, which is the product of the detector confidence of the two objects and their attribute. The third row of Figure 2 shows example O&A representations for indoor scenes from the B3DO dataset [24].

## IV. LEARNING DISCRIMINATIVE SCENE BASES

### A. Background of Dictionary Learning

Given a training set  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N] \in \mathcal{R}^{M \times N}$ , dictionary learning [25] aims to learn a dictionary of bases that best reconstruct the training examples:

$$\min_{\mathbf{D}, \mathbf{X}} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{x}_i\|_1 \quad (3)$$

where  $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_K] \in \mathcal{R}^{M \times K}$  is the dictionary with  $K$  bases,  $\mathbf{x}_i \in \mathcal{R}^K$  are the reconstruction coefficients for  $\mathbf{y}_i$ . Different from the K-means clustering that assigns each

training example to the nearest cluster center, Eq. 3 learns an overcomplete dictionary  $\mathbf{D}$  and represents each training example as a sparse linear combination of the bases in the dictionary.

To learn a dictionary that is well-suited for supervised classification tasks, class-specific dictionary learning methods have been proposed that learn a sub-dictionary for each class [26], which is formulated as

$$\min_{\mathbf{D}, \mathbf{X}} \sum_{c=1}^C \|\mathbf{Y}_c - \mathbf{D}_c \mathbf{X}_c\|_2^2 + \lambda \|\mathbf{X}_c\|_1 \quad (4)$$

where  $\mathbf{Y}_c = [\mathbf{y}_1^c, \dots, \mathbf{y}_{N_c}^c]$ ,  $\mathbf{X}_c = [\mathbf{x}_1^c, \dots, \mathbf{x}_{N_c}^c]$ , and  $\mathbf{D}_c = [\mathbf{d}_1^c, \dots, \mathbf{d}_{K_c}^c]$  are the training set, reconstruction coefficients, and sub-dictionary for class  $c$ , respectively.

As pointed out in [27], the sub-dictionaries learned above typically share common (correlated) bases, thus  $D$  may not be sufficiently discriminative for classification tasks and the sparse representation will be sensitive to the variations in features. Even though an incoherence promoting term  $\sum_{i \neq j} \|\mathbf{D}_i^T \mathbf{D}_j\|_2^2$  can be included in the dictionary learning objective function, correlated bases still exist [26].

### B. Our method

Our method learns high-order couplings between the objects and attributes of a scene in the form of a set of class-specific sub-dictionaries under the elastic net regularization

and geometric similarity constraint, which is formulated as

$$\min_{\mathbf{D}, \mathbf{X}} \sum_{c=1}^C \left[ \|\mathbf{Y}_c - \mathbf{D}_{\in c} \mathbf{X}_c\|_2^2 + \|\mathbf{D}_{\notin c} \mathbf{X}_c\|_2^2 + \lambda_1 \|\mathbf{X}_c\|_1 + \lambda_2 \|\mathbf{X}_c\|_2^2 \right] + \lambda_3 \sum_{l=1}^L \sum_{i=1}^N \|\alpha_l - \mathbf{x}_i\|_2^2 w_{li} \quad (5a)$$

$$s.t. \quad \|\mathbf{d}_k\|_2^2 = 1, \forall k = 1, \dots, K \quad (5b)$$

where  $\mathbf{D}_{\in c} = [\mathbf{0}, \dots, \mathbf{D}_c, \dots, \mathbf{0}]$  and  $\mathbf{D}_{\notin c} = \mathbf{D} - \mathbf{D}_{\in c}$ ,  $\|\mathbf{Y}_c - \mathbf{D}_{\in c} \mathbf{X}_c\|_2^2$  minimizes the reconstruction residual of the training examples of class  $c$  using the  $c^{th}$  sub-dictionary,  $\|\mathbf{D}_{\notin c} \mathbf{X}_c\|_2^2$  penalizes the reconstruction of the training examples using sub-dictionaries from different classes,  $\lambda_1 \|\mathbf{X}_c\|_1 + \lambda_2 \|\mathbf{X}_c\|_2^2$  is the elastic net regularizer, and  $\lambda_3 \sum_{l=1}^L \sum_{i=1}^N \|\alpha_l - \mathbf{x}_i\|_2^2 w_{li}$  is the geometric similarity constraint.

The elastic net regularizer is a weighted sum of the  $l_1$ -norm and the square of the  $l_2$ -norm of the reconstruction coefficients. Compared to a pure  $l_1$ -norm regularizer, the elastic net regularizer allows the selection of groups of correlated features even if the group is not known in advance. Besides enforcing grouped selection, elastic net regularizer is also crucial to the stability of the sparse reconstruction coefficients with respect to the input examples [28].

In the geometric similarity constraint,  $\alpha_l$  are the reconstruction coefficients for the "template"  $\mathbf{t}_l$ . Here, the templates are the cluster centers obtained by running K-means clustering on the training examples of each class.  $L$  is the total number of templates from all classes. In particular, the coefficients  $\alpha_l$  belonging to class  $c$  can be calculated as follows

$$\beta_l = \min_{\beta_l} \|\mathbf{t}_l - \mathbf{D}_c \beta_l\|_2^2 + \lambda_1 \|\beta_l\|_1 + \lambda_2 \|\beta_l\|_2^2 \quad (6a)$$

$$\alpha_l = \left[ \underbrace{\mathbf{0}}_{\mathbf{K}_1}, \dots, \underbrace{\mathbf{0}}_{\mathbf{K}_{c-1}}, \beta_l, \underbrace{\mathbf{0}}_{\mathbf{K}_{c+1}}, \dots, \underbrace{\mathbf{0}}_{\mathbf{K}_C} \right] \quad (6b)$$

In  $\alpha_l$ , only the coefficients corresponding to the bases from the same class  $c$  are non-zero. The weight  $w_{li}$  is defined to be the Gaussian similarity between the template  $\mathbf{t}_l$  and the training example  $\mathbf{y}_i$

$$w_{li} = \exp(-\|\mathbf{t}_l - \mathbf{y}_i\|_2^2 / 2\sigma^2) \quad (7)$$

The geometric similarity constraint defined in this way encourages two similar input examples to also have similar reconstruction coefficients.

We use a similar iterative algorithm in [29] to solve for the optimal dictionary  $\mathbf{D}$  and the reconstruction coefficients  $\mathbf{X}$  in Eq. 5, which iteratively optimizes over  $\mathbf{D}$  or  $\mathbf{X}$  while fixing the other.

### C. Classification

After learning the discriminative dictionary  $\mathbf{D}$ , the reconstruction coefficients for an input example  $\mathbf{y}$  can be calculated by solving the following optimization problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2 + \lambda_3 \sum_{l=1}^L \|\alpha_l - \mathbf{x}\|_2^2 w_l \quad (8)$$

scene category	number of examples	scene category	number of examples
bathroom	121	kitchen	225
bedroom	383	living room	221
bookstore	36	office	128
classroom	49	playroom	31
dining room	117	study	25
furniture store	27		

TABLE I: The statistics of the scene categories in NYUD dataset after validation.

scene category	number of examples	scene category	number of examples
bath room	43	kitchen	28
bedroom	15	living room	63
dining room	24	office	130

TABLE II: The statistics of the scene categories in B3DO dataset after validation.

After obtaining the reconstruction coefficients for a novel image, a linear SVM classifier is used to obtain the scene category.

## V. EXPERIMENTS

### A. Experiment settings

The proposed method is evaluated on two indoor scene RGB-D datasets, the NYUD dataset [30] and the B3DO dataset [24].

- NYUD is a dataset of 1449 RGB-D images, covering 27 diverse indoor scenes taken from 3 cities. A dense per-pixel object annotation is available, which features 35,064 distinct objects, spanning 894 different classes. In our experiment, we discard any scene category that has less than 20 images, and merge "home office" into "office" as we find that these two scenes are very much the same. This gives us a dataset the statistics of which are summarized in Table I.
- B3DO is a dataset of 849 RGB-D images taken in domestic and office settings. Object-level annotation is provided in the form of bounding boxes on the RGB image. Since no scene class labels are available, we manually classify the RGB-D images into one of six scene classes, including bathroom, bedroom, kitchen, dining room, living room, office. During this process, we discard those images that do not belong to any of the six scene classes, resulting in a scene dataset summarized in Table II.

For both datasets, we split the images in each scene class by a factor of 0.5, and use the first half for training and the rest for testing. All experiments are repeated ten times with random split of each scene class. The final performance metric is reported as the mean of the results from the individual runs.

In our implementation, the parameters are set as follows: (i) the  $l_1$ -norm weight  $\lambda_1 = 0.34$ ; (ii) the  $l_2$ -norm weight  $\lambda_2 =$

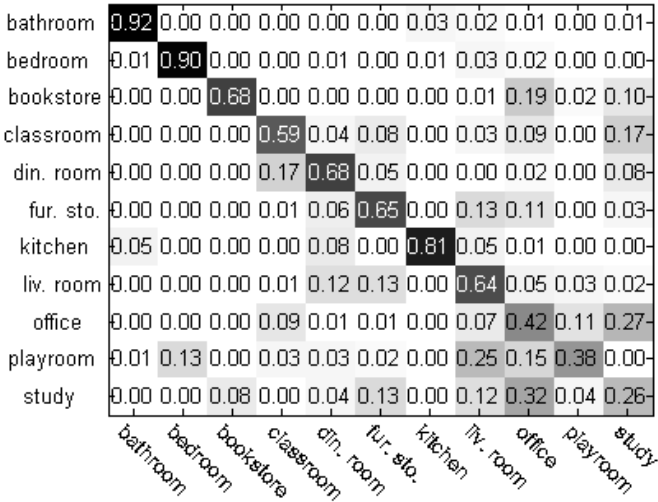


Fig. 3: The confusion matrix of the O&A-DL on the NYUD dataset.

0.17; (iii) the geometric similarity weight  $\lambda_3 = 1.03$ ; (iv) the similarity kernel width  $\sigma = 1.17$ . (Preliminary experimental results show that satisfactory performance is achieved when  $\lambda_1 \in [0.3 \sim 0.4]$ ,  $\lambda_2 \in [0.1 \sim 0.2]$ ,  $\lambda_3 \in [1.0 \sim 1.1]$ , and  $\sigma \in [1.0 \sim 2.0]$ .) The number of bases for each sub-dictionary is set to 15. For the geometric similarity constraint, the number of templates for each class is set to be 1/10 of the total number of training examples in that class. Compared to the total number of training examples, the number of templates is relatively small.

### B. Comparison with the State-of-the-art Methods

In this experiment, we investigate the effectiveness of the proposed method (denoted as O&A-DL) for indoor scene recognition from RGB-D images. Specifically, we compare O&A-DL to the Object Bank Model (OBM) [5] and the Deformable Part based Model (DPM) [6]. In addition, we also include the recognition performance from directly using the O&A representation without using their reconstruction coefficients (denoted as O&A).

As noted in previous work [5], scene recognition based on high-level representations generally outperforms low-level representation based scene recognition ([1], [2], [4], [15]). Therefore, we do not include comparison results with low-level scene recognition methods.

For all four methods, we train one-vs-all linear SVM classifiers to recognize the scene category, i.e. a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

Figure 3 and Figure 4 show the confusion matrix obtained using the O&A-DL on the NYUD dataset and the B3DO dataset, respectively. Figure 5 and Figure 6 show the per-class recognition accuracy of the four compared methods on the NYUD dataset and the B3DO dataset, respectively.

As can be seen, O&A-DL is significantly better than OBM and DPM. This is due to our explicitly modeling spatial layout relations between individual objects using the predefined

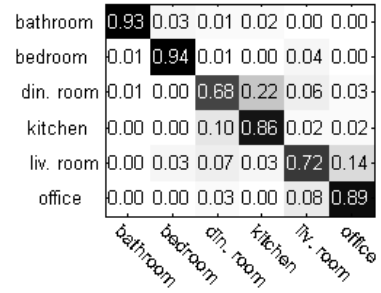


Fig. 4: The confusion matrix of the O&A-DL on the B3DO dataset.

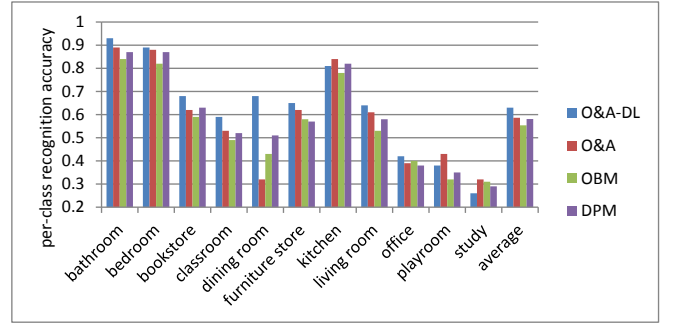


Fig. 5: Comparison of the O&A-DL to the state-of-the-art methods on the NYUD dataset.

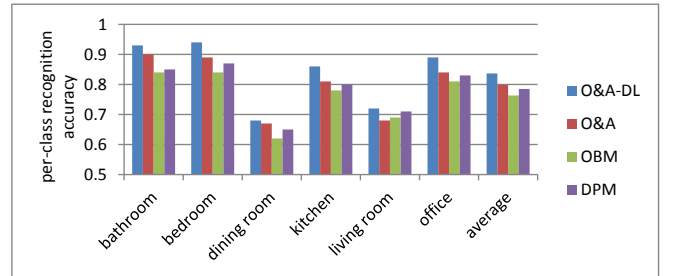


Fig. 6: Comparison of the O&A-DL to the state-of-the-art methods on the B3DO dataset.

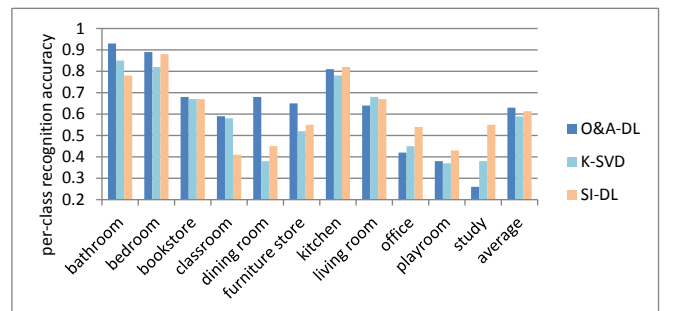


Fig. 7: Comparison of the O&A-DL to other dictionary learning based recognition methods on the NYUD dataset.

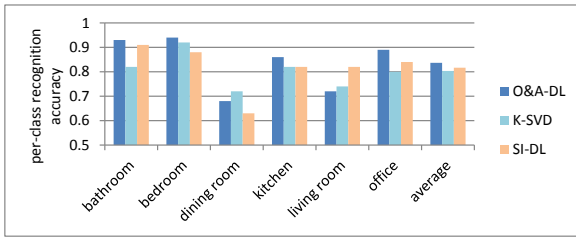


Fig. 8: Comparison of the O&A-DL to other dictionary learning based recognition methods on the B3DO dataset.

attributes. It is also clear from Figure 5 and Figure 6 that O&A-DL consistently outperforms O&A, which is simply a vector of the confidence score of object detectors and attribute classifiers. This demonstrates that scene recognition benefits from learning discriminative scene bases and classifying novel scene images using the sparse reconstruction coefficients.

### C. Comparison with Other Dictionary Learning Methods

In this section, we evaluate the O&A-DL against other dictionary learning based scene recognition methods, including K-SVD [31] and dictionary learning with structured incoherence (SI-DL) [27]. For all these three methods, the linear SVM classifier with the sparse reconstruction coefficients as input is used for classification.

Figure 7 and Figure 8 show the per-class average recognition accuracy on the NYUD and B3DO dataset, respectively. As can be seen, O&A-DL outperforms the other two methods on these two datasets. Also, the class-specific dictionary learning method, SI-DL, performs better than K-SVD, which demonstrates the discriminative power of class-specific dictionaries.

## VI. CONCLUSION

This paper presents a new method for indoor scene recognition from RGB-D images. In order to better represent a scene, we train a set of object detectors and attribute classifiers from RGB-D images, which not only describes the constituent objects of a scene but also captures their spatial interrelationships. Moreover, to discover the high-order couplings between the objects and attributes, we develop a class-specific sub-dictionary learning method under the elastic net regularization and geometric similarity constraint. Experimental results demonstrate that the proposed scene recognition method achieves better accuracy than the state of the art. We note that the application of the proposed dictionary learning method is not limited to scene recognition. Our future work would consist of further evaluation of the proposed method on recognition problems such as face recognition and activity recognition.

## REFERENCES

[1] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *ICCV*, 2005.  
 [2] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, 2001.

[3] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *ICCV*, 2003.  
 [4] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan, "Categorizing nine visual classes using local appearance descriptors," in *In ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.  
 [5] Y. L. Li-Jia Li, Hao Su and L. Fei-Fei, "Objects as attributes for scene classification," in *ECCV*, 2010.  
 [6] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.  
 [7] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *ECCV*, 2012.  
 [8] S. Wan and J. K. Aggarwal, "Scene recognition by jointly modeling latent topics," in *WACV*, March 24-26, 2014.  
 [9] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012.  
 [10] E. S. Ye, "Object detection in rgb-d indoor scenes," Master's thesis, EECS Department, University of California, Berkeley, Jan 2013.  
 [11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91-110, 2004.  
 [12] Y. Huang, K. Huang, C. Wang, and T. Tan, "Exploring relations of visual codes for image classification," in *CVPR*, 2011.  
 [13] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *CVPR*, 2007.  
 [14] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Visual synset: Towards a higher-level visual representation," in *CVPR*, 2008.  
 [15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing scene categories," in *CVPR*, 2006.  
 [16] I. González-Díaz, D. García-García, and F. Díaz-de María, "A spatially aware generative model for image classification, topic discovery and segmentation," in *ICIP 09*, pp. 781-784.  
 [17] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *CVPR '12*.  
 [18] Z. Niu, G. Hua, and Q. Tian, "Spatial-disclda for visual recognition," in *CVPR 11*.  
 [19] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *CVPR '09*.  
 [20] A. Gupta and L. Davis, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers," in *ECCV*, 2008.  
 [21] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *ICRA*, 2011.  
 [22] G. Logan and D. Sadler, "A computational analysis of the apprehension of spatial relations," in *In Language and space language speech and communication*. MIT Press, 2013.  
 [23] S. Ouellet and J. Davies, "Using relations to describe three-dimensional scenes: A model of spatial relation apprehension and interference," in *International Conference on Cognitive Modeling*, 2013.  
 [24] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3-d object dataset: Putting the kinect to work," in *ICCV Workshops*, 2011.  
 [25] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *NIPS*, 2008.  
 [26] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *ECCV*, 2012.  
 [27] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *CVPR*, 2010.  
 [28] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *J. Complex.*, 2009.  
 [29] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *ICCV*, 2013.  
 [30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *ECCV*, 2012, pp. 746-760.  
 [31] M. Aharon, M. Elad, and A. Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, 2006.