3D Indoor Object Recognition by Holistic Scene Understanding

Kaicheng Wang
Stanford University
kwang2@stanford.edu

Shutong Zhang
Stanford University
zhangst@stanford.edu

Yixin Wang
Stanford University
wyixin@stanford.edu

Abstract

In this project, we jointly solve the problem of indoor object recognition and scene classification. The joint problem is modeled by a Conditional Random Field (CRF). Specifically, we implement unary potential on scene appearance, unary potential on object appearance, and binary potential on geometry between objects. The two appearance potentials are obtained by training two neural networks respectively and extracting features from them. Binary potential on object-object geometry relationship is inferred by projecting one object onto the other one's surface and perceiving the overlap. Other potentials include object cuboid geometry, binary potential on scene-object concurrence, and binary potential on object-object concurrence.

Experiments on the challenging NYU-RGBD v2 dataset show that the approach of jointly solving object detection and scene classification problems by integrating several feature representations achieves better performance than vanilla classification. Meanwhile, improving the potentials on scene appearance and object-object geometry relationship achieve fairly satisfactory improvements compared with the performance in [12]. We achieve object recognition precision of 0.3859 and scene classification accuracy of 0.6208 among 21 classes of objects and 13 classes of scenes. Both improve obviously comparing to 0.3829 and 0.6086 in baseline [12]. This result is also comparable to the state-of-art method implemented on the identical dataset [18].

1. Introduction

One core vision problem in indoor robotics is 3D recognition. Accurate object detection as well as classification is essential for robot to navigate and interact with the indoor environment. Towards solving this problem, we start our project by implementing the object detection system in [12].

Classification of indoor objects is not an easy task in that indoor objects usually have very similar local features. Most classes have cube-like shape and are very similar in surface textures. Classification of indoor scenes also suffer similar problems. However, these two problems can be solved jointly, in hope that the interaction of these two models could solve some ambiguities in both object and scene classification problems. The idea is to utilize object-object and object-scene relationships. For example, the probability of a lamp placed next to a bed should be much higher than the probability of a microwave oven placed next to a bed (object-object relationship). Also, the probability of a bed appearing in a bedroom scene should be much higher than it appearing in a living-room scene (object-scene relationship). Thus, when the classifier makes mistakes on the class of one object, these relationship constraints might save the day by slightly modifying the distribution of scores.

Meanwhile, depth information is helpful especially in indoor settings because it helps us capture the geometric features of an object and the scene. In the holistic model, it also helps capture the geometric relationship in object-object or object-scene. Indoor objects are not placed randomly in space. Beds are usually on the floor and against the wall. Utensils appear more often in the same image with refrigerator than television. Considering the physical and statistical interactions between the objects and the environment, we should be able to label objects in agreement with the scene.

We develop our model on the labeled portion of NYU-RGBD v2 dataset[13], which consists of RGB-D images with dense labels. The dataset is split into 795 training images and 654 dev/test images. The model detects 3D cuboid candidates at first, and then labels them via joint reasoning with Conditional Random Field (CRF).

2. Related Work

Indoor 3D object recognition from multi-view RGB-D images has achieved quite a success in the past few years [10][8][4]. However, scanning objects generally takes longer time needed for real-time applications [6], or the performance just deteriorates as the camera moves faster [11]. There are also circumstances where we might not have access to multiple cameras observing the same object. Therefore, recognition from single-view is still worth exploration. Context information is usually essential to boost recognition

performance for indoor scenarios [12][18].

The pipeline is usually to propose object candidates and classify scene at first. Then we decide the object categories accordingly. A most recent work [7] skips the proposal step by predefined 3D scene template, but the scenes are restricted to merely 4 categories. We review three major parts of this problem as follows.

2.1. Previous Work

1. **3D object proposal:**

   We want to generate object candidates as cubes for recognition. In this way, it is easier to force the surfaces parallel to walls and ground. One typical approach is to generate bottom-up region candidates and sorts them by "objectness" scores [2]. Segmentation is made by computing similarities between adjacent pixels [1]. In particular, depth information is also considered [14][16][15].

   Another option is to utilize sliding window algorithm. We can thus observe the whole object [17] instead of only part of it due to bottom-up nature. The state-of-the-art result is achieved by 3D ConvNets, which replaces hand-crafted features.

2. **Scene understanding:**

   Human beings can recognize different kinds of scenes even by glancing blurred images. Although essentially we do not need much information to understand the scene category of an image, it is quite helpful for object recognition since scene-object are often related statistically. The published state-of-the-art scene understanding was implemented by neural network trained by *SUN* database. This was first proposed in [21] and improved in [20]. After that, a 60 times larger database came along, namely *Places* [23]. To the best of our knowledge, the latest work (also trained on *Places*), which was released just 3 weeks ago (with paper coming soon), is even more powerful [22].

3. **Contextual object classification:**

   Contextual models mainly focused on segmentation with RGB-D datasets. [15] reasons about spatial transitions between superpixels based on RGB-D information. Our baseline model [12] reasons at higher levels, i.e. object-object and object-scene, by combining various potentials in CRF. [18] combines object features and scene understanding by neural network, yet their test result on NYU-RGBD v2 dataset is comparable to ours.

2.2. Our Contributions

Just to clarify, we were not able to improve the object detection performance from [12]. Instead, we get better performance by modifying the CRF model. In particular, we implement transferable training based on the pretrained network for better scene understanding [22]. We also try to extract segmentation features from fine-tuned CNN [9]. Geometric context is further reasoned by considering more complex spatial relationship between object pairs.

3. Methods

3.1. Problem Description

This problem takes an input of an RGB-D image in the NYU-RGBD v2 dataset. The goal is to perform 3D object recognition and indoor scene classification. These two tasks are tackled jointly in our holistic model, that is, some constraints are placed among scenes and these objects when reasoning about object label, which helps boost the performance of scene classification and object recognition.

The problem can be divided into two parts. The first part mainly generates 3D region candidates in a bottom-up manner, and then fits cuboids (analogies to bounding boxes in 2D) around the candidate regions. The second part is the incorporation of several features, either extracted from the image/semantics/cuboid geometry or from the output score of other classifiers, e.g. SVM. We then define a Conditional Random Field (CRF) and use these features as potentials of the CRF model. This has made a possible holistic understanding of the scene, which will likely boost recognition performance than state-of-art detection methods including part-based models[3] and exerting elaborate constraints that are obtained via depth information[19].

We use the same framework (cuboid fitting - feature extraction - CRF) as [12]. However, the feature extraction part is different where we focus on the power of deep structures and more geometric information. We explore more features as well as features from more powerful classifiers, in hope of further boosting performance.

3.2. 3D Detection

3.2.1 Generating 3D region candidates bottom-up

The model [2] we used for generating candidates uses parametric min-cut to generate a wide variety of foreground candidates from equally spaced seeds. The objective is to minimize the energy E over pixel labels $\{x_1, x_2, ..., x_N\}$:

$$E^{\lambda} = \sum_{u \in V} C_{\lambda}(x_u) + \sum_{(u,v) \in \varepsilon} V_{uv}(x_u, x_v)$$

with $\lambda \in R$, V the set of all pixels, ε the edges between neighboring pixels, and C_{λ} the unary potentials.

3.2.2 Fitting cuboids

After generating region candidates, we then generate cuboids from the candidate results [2]. Specifically, top k

candidate regions ranked by the objectness [2] scores are selected after performing non-maxima suppression. Then each candidate cuboid is fitted by a 3D cube around candidate region.

3.3. Holistic scene model using CRF

After obtaining cuboids of interest, we assign class labels to the detected cuboids, which would been our main focus for this project. Many feature-based approaches exist that extract features of each cuboid and use them as input to some classifiers to predict class labels. However, such approaches suffers occlusion and viewpoint change. More importantly, they discard contextual information which might be helpful to recognition. The CRF model, on the other hand, learns scene configuration in a holistic manner. Apart from appearance (which is obtained via image features), it models geometry, object relations and spatial configurations, in hope of achieving better recognition accuracy by incorporation these additional information.

3.3.1 Model

Conditional random fields are depicted by potentials. The probability of an assignment $\mathbf{x}$

$$p(\mathbf{x}) = \frac{1}{Z}\Pi_C \psi_C(x_C),$$

where x_C is a clique in the graph, and Z is the partition function for normalization.

The potentials ψ_C is depicted by energy $E(\cdot)$ of an assignment

$$\Psi_C(x_C) = \exp(-E(x_C)).$$

In a given image, denote the present objects as $y_i \in 0, 1, ..., C$ and scene variable (e.g. kitchen, bathroom) as $s \in 1, 2, ..., S$. There are $C + 1$ classes of objects, where class 0 is "other", i.e. false positives. There are S classes of scenes. In our scenario, the probability of an assignment of y_i, s is

$$p(y_i, s) = \frac{1}{Z}\exp\Big(w_s\psi_s(s) + \sum_t w_t \sum_i \psi_t(y_i) + \sum w_m \sum_i \phi_m(s, y_i) + \sum w_p \sum_{i,i'} \phi_p(y_i, y_{i'})\Big).$$

There are four kinds of potentials. $\psi_s(s)$ is unary potential of scene s. $\{\psi_t(y_i), t = 1, 2, 3...\}$ is a set of unary potentials for object y_i. $\{\phi_m(s, y_i), m = 1, 2, 3...\}$ is a set of binary potentials capturing the contextual relationship between scene s and object y_i. $\{\phi_p(y_i, y_{i'}, p = 1, 2, 3...\}$ is a set of binary potentials capturing the relationship between objects $y_i, y_{i'}$.

The CRF structure is illustrated in Figure 1.
For CRF learning, we will be using the primal-dual approach as described in [12].

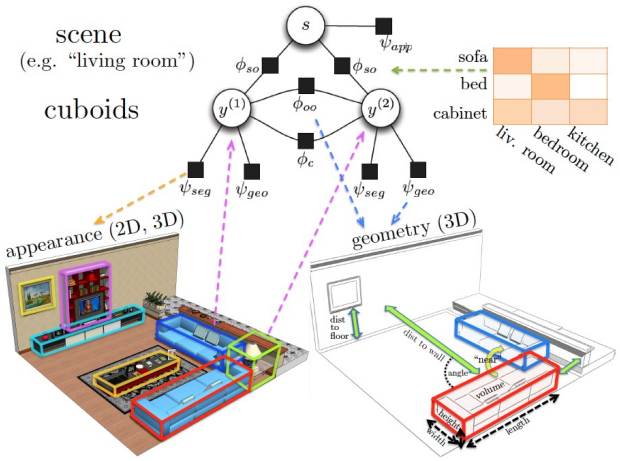


Figure 1. Illustration of conditional random field model on indoor scene

3.3.2 Potentials

As the CRF model shown in Figure 1, we have defined six different kinds of potentials to capture the relationships between scenes, objects, objects and objects, as well as objects and scenes. We would introduce the details of each potential in this section.

1. **Scene appearance:**

   This is a unary potential defined over a scene, which models the likelihood of a scene s of being class u by defining

   $$\psi_s(s = u) = \sigma(t_u),$$

   where t_u is the classifier score for the scene of being class u and σ is the sigmoid function to normalize the scores to probability space. While [12] takes the output score from an SVM using SIFT/GIST features as input, we believe a more powerful description might be obtained using fine-tuned CNN or stacked fully-connected layers.

2. **Scene context:**

   This is a binary potential defined over a scene and an object modeling the likelihood of a particular object appearing in a particular scene. The intuition is that an oven is more likely to appear in a kitchen than in a bedroom. In [12], it takes in the predicted label of the scene and an object and outputs the potential, and finally takes a weighted summation over all objects. However, we believe that because we are not fully confident with the object label, taking the most probable two or three labels might be more sensible than merely taking the most-likely predicted label.

3. **Object context:**

This is a binary potential defined over two objects, modeling the likelihood of them appearing together in a scene. The intuition is similar to the previous potential. For every pair of predicted labels of objects, it outputs the potential and finally takes a weighted summation over all pairs.

4. **Segmentation potential:**

   Segmentation potentials are used as unary potentials for cuboid hypotheses. The approach trains a classifier on the kernel descriptors aggregated over superpixels [14] and obtain a classification score for each superpixel. The cuboids are then projected to a image plane, and using a convex hull we can derive the potential for each cuboid.

5. **Object geometry:**

   Geometric properties of a $3D$ cuboid are represented using a vector with ten dimensions, which describe not only the inner characteristic of objects, but also their relation between the object and the scene. We train a SVM classifier with RBF kernel with these features, and the geometry unary potential is defined as the score of each class.

6. **Geometric context:**

   Geometric context models the geometric relationship between objects. [12] models two kinds of relationships "close to" and "on top of". We believe increasing the types of relationships might help capture more information.

3.4. Improved Potentials

We focus on three potentials, scene appearance, segmentation potentials and geometric context, where we consider deep structures as classifiers and take specific geometric information into account to boost the performance of the overall scene model.

The improvement can be summarized in Figure 2.

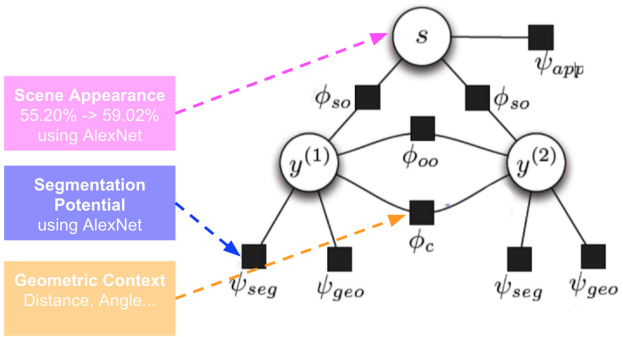


Figure 2. Illustration of Improved CRF

1. **Scene Appearance:**

   In [12], unary scene appearance potentials are defined using normalized classification scores where we obtained by modified SVM classifier. Based on the strength of the internal data patterns, neural network have been widely explored in many scenarios, including scene understanding.

   Among many different deep structures, the performance of AlexNet [9] have outperformed many state-of-the-art methods on image classification. The structure of AlexNet[9] is shown in Figure 3. The net contains eight layers with weights, among them the first five are convolutional layers and the remaining three are fullyconnected layers. Each convolutional layer consists of the combination of convolutional, relu and max-pooling layer. The output of the last fully-connected layer can be considered as the probability of each class. AlexNet maximizes the multinomial logistic regression objective, which is equivalent to maximizing the average across training cases of the log-probability of the correct label under the prediction distribution, and by minimizing the loss we can train all the parameters using back propagation.

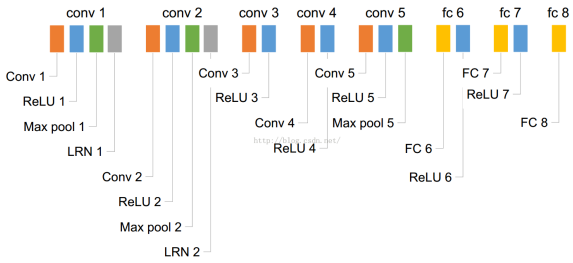


Figure 3. The structure of AlexNet

Thus we perform fine-tune on AlexNet [9] using the pre-trained caffe model[5] on a largescale database called *Places*. [22] has constructed Places that contains more than 10 million images comprising 400+ unique scene categories, and 5000 to 30,000 training images per class.

Using the pre-trained weights of places database, we fix the weights from conv1 layer to fc6 layer, and fine-tune fc7 layer and output layer on our training set. We have re-sized the images in our dataset from $480 * 640$ to $227 * 227$ to fit into AlexNet. The dimension of output layer has reduced from 365 classed to 13 classes as well.

2. **Segmentation Potentials:**

   [12] computes segmentation potentials using six types of RGB-D kernel descriptors: gradient, color, local binary pattern, depth gradient, spin/surface normal, and

KPCA/self-similarity. [12] thinks that training a good classifier is not likely because the number of training samples is limited. However, we think this problem could be tackled via transfer learning, that is, fine-tuning a trained model of neural network will require much fewer training samples than other methods to achieve competitive results. We performed fine-tune on AlexNet [9], using its parameters and training the final fully-connected layer from scratch. The training/test loss and test accuracy over iterations is shown in Figure 4 and 5. The model seems to begin to overfit from iteration 6000, at which point we cut off the training process.

The training samples are obtained by cutting out the objects with its detection bounding box, adding a padding of 25 pixels, and resizing the cropped image to 227x227. In this way we obtain our own training and test set for deep network. The deep network is trained and tested using Caffe framework[5].

The improved segmentation potential is defined as follows:

$$\psi_{seg}(o = v) = \Pr{}_{\text{CNN}}(o = v),$$

where $\psi_{seg}(o = v)$ denotes the segmentation potential of object o being of class label v. $\Pr{}_{\text{CNN}}(o = v)$ denotes the probability of object o being of class label v given its appearance feature (i.e. RGB image), which is exactly the output of the softmax layer of CNN because softmax layer maps output scores to probability space.

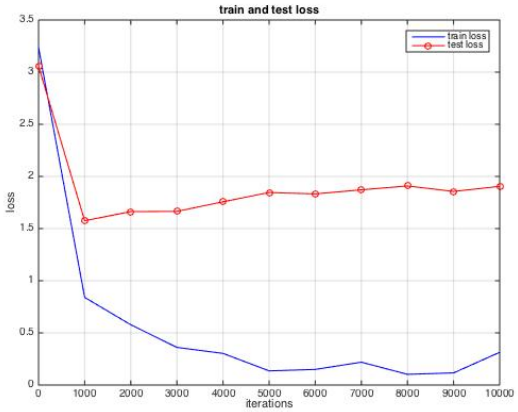


Figure 4. Training and test loss versus iterations

3. **Geometric context:**

Instead of "close to" relationship reasoned in [12], we believe that objects are more relative if their surfaces are parallel. For example, sofas usually face to tea table no matter which side of wall they are facing to. Moreover, that "close to" relationship might overlap with the object-object potential. Therefore, we only consider objects that are neither too close nor too far. In other words, we are implementing a stricter standard in terms of related object pairs.

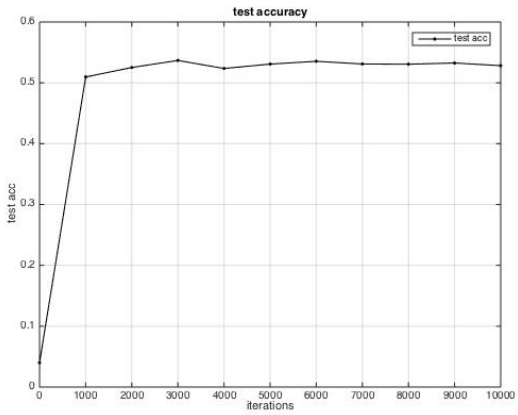


Figure 5. Test accuracy versus iterations

We need to tune hyperparameters for distance range and angle range. In practice, we identify surfaces to "parallel" when their angle of intersection is at most 15 degrees. Objects that are close to 0.2 or farther than 2 are ruled out from this relationship, because we believe that they should be in appropriate distance if they are related. As a matter of fact, this pair of parameters works best on validation set.

4. Experiments

4.1. Dataset

We divide our dataset [13] into 381 training samples, 414 validation samples and 654 test samples. We also preprocess them into 13 different scene categories and 21 different object categories.

4.2. Dataset Visualization

To better understand the NYU-RGBD v2 dataset [13], we have visualized depth map for different scenes. Figure 6 shows the visualization for the raw data of the bathroom scene.

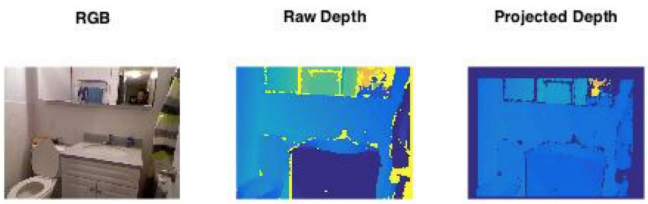


Figure 6. Visualization of the RGB-D image in NYU-RGBD v2 dataset[13]

We also visualize the generated object bounding boxes obtained by models proposed by [2] to better understand the object detection results. Comparing to groundtruth bounding boxes, we can clearly see some false positives in the generated ones in Figure 7.

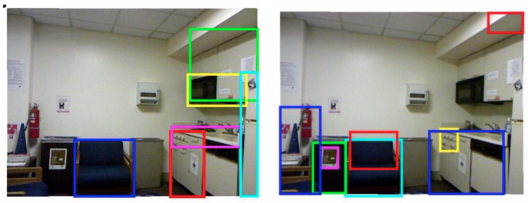
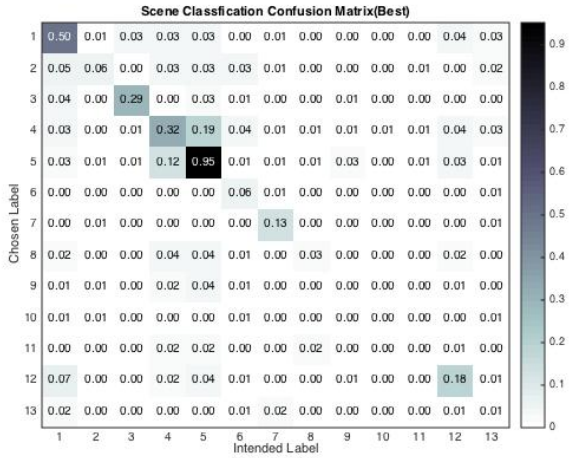


Figure 7. Visualization of the generated object bounding box (**left**: groundtruth, **right**: k=8 candidates per image)



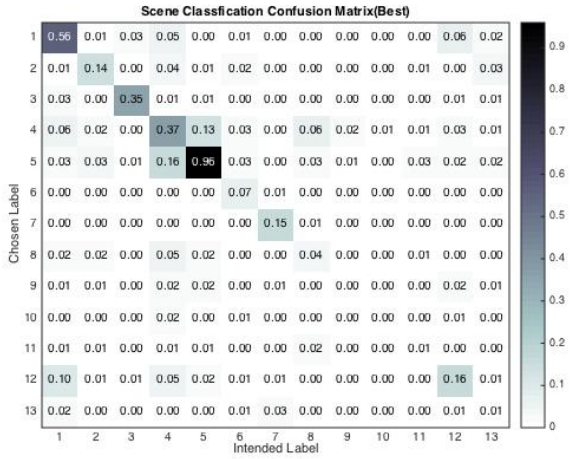


Figure 8. Class-specific analysis of each scene category (**upper**: [12] using sce. + seg. + geo. + cpmc + sce.-obj. + obj.-next. + obj.-top, **lower**: our best CRF model using **new-sce.** + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top)

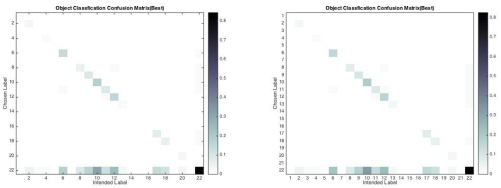


Figure 9. Class-specific analysis of each object category (**left**: [12] using sce. + seg. + geo. + cpmc + sce.-obj. + obj.-next. + obj.-top, **right**: our best CRF model using **new-sce.** + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top)

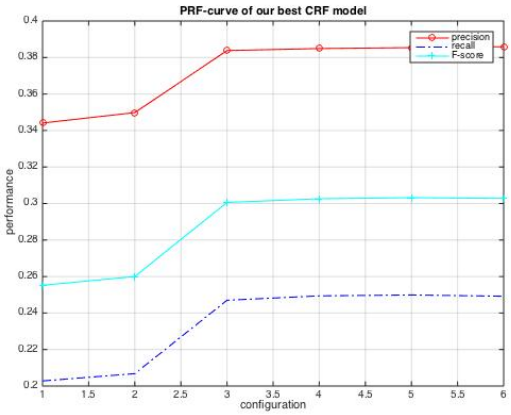


Figure 10. Precision, Recall and F-score with our best CRF model under six different configurations same as the configuration in Table 1

4.3. Detection Recall and Classification Accuracy

We first set up and run the CRF model with potentials proposed by [12]. Then using their model as baseline, we incorporate more powerful potentials. The results and analysis are provided in this section. For both baseline model and our improved model, we report classification performance on both ground truth detection and $k = 8$ best detected candidate cuboids per image, using detection framework described in Section 3.2. Testing on ground truth detection tells us the capability of potentials and the classifier model, while testing on cuboid candidates is closer to real-life scenarios.

4.3.1 Notation

We denote experiments using eight best cuboid candidates per image as $k = 8$ *situation*. We denote experiments using ground truth detection as *ground-truth situation*. A *recall* is defined as more than half of the detected object overlaps the ground truth bounding box in the same label, which means both detection and classification have to be good enough.

configuration	object	scene
scene appearance + segmentation	0.5446	0.5520
sce. + seg. + geometry	0.5921	0.5520
sce. + seg. + geo. + sce.-object	0.6039	0.5841
sce. + seg. + geo. + sce.-obj. + obj.-next	0.6085	0.5856
sce. + seg. + geo. + sce.-obj. + obj.-next. + obj.-obj.	0.6089	0.5872
sce. + seg. + geo. + sce.-obj. + obj.-next. + obj.-obj. + obj.-top	**0.6134**	**0.5917**

Table 1. Classification performance on ground truth detection, testing potentials proposed by [12]

configuration	object	scene	recall
scene appearance + segmentation	0.3440	0.5520	0.2028
sce. + seg. + geometry	0.3498	0.5520	0.2068
sce. + seg. + geo. + cpmc	0.3842	0.5520	0.2470
sce. + seg. + geo. + cpmc + sce.-object	0.3830	0.6040	0.2494
sce. + seg. + geo. + cpmc + sce.-obj. + obj.-next	0.3828	0.6101	0.2396
sce. + seg. + geo. + cpmc + sce.-obj. + obj.-next. + obj.-top	**0.3829**	**0.6086**	**0.2401**

Table 2. Classification performance taking eight cuboid candidates ($k = 8$) from detection results, testing potentials proposed by [12]

4.3.2 Baseline and analysis

First, we run the CRF model on ground truth cuboids to evaluate the joint classification performance only. Then we add the 3D detection part, and all results are shown in Table 1, 2. Notice that the number of candidates per image, i.e. k, is a parameter of the model. For now, we are demonstrating only one circumstances where $k = 8$.

For ground-truth situation, the classification results are shown in Table 1. Accuracy becomes higher as we take into account more kinds of potentials. With all potentials, the best classification accuracy is reached, i.e. 0.6134 for object classification and 0.5917 for scene classification.

For $k = 8$ situation (see 4.3.1 for meaning of notation $k = 8$), recall is also shown in Table 2 (see 4.3.1 for definition of recall). The existence of many false positive cuboids during detection brings down the recognition accuracies a lot. But still, we test our CRF in the same way, where potentials are added one by one to model. Since we have more cuboids, the runtime is significantly longer than that of ground truth. The choice of k is still under exploration. More candidates means higher noise. Since our detection is sometimes noisy, larger k does not guarantee better results. Also, since we have even less data due to wrong detection, the accuracies do not essentially get higher when the model contains more potential. We assume that it is because of over-fitting. Considering object precision, scene accuracy and detection recall all together, we think that the last row in Table 2 represents the best performance for baseline model.

4.3.3 Improved object-object geometry relationship

Implementation details for improved binary potential of geometry context is elaborated in 3.4. The performance with this improved potential is shown in Table 4 and 5, denoted as **new-obj.-next**.

For ground-truth situation, the best model obtains accuracy of 0.6123 for object classification and 0.5971 for scene classification, with scene classification accuracy better than baseline model. For $k = 8$ situation, the best model obtains accuracy of 0.3850 for object classification, 0.6040 for scene classification, and object classification recall is 0.2492, with object classification accuracy and recall both higher than baseline model.

4.3.4 Improved scene classification potential

After fine-tuning on our training dataset with AlexNet structure and [22]'s pre-trained weights, we obtain a $1 * 13$ vector for each input sample, using normalized classification scores of 13 scene categories. Denoted as **new-sce.**, the performance of the new scene appearance potentials are shown in Table 4 and 5. For the ground truth bounding boxes, we obtain 0.6174 for object classification accuracy and 0.6055 for scene classification accuracy. For $k = 8$ situation, we obtain 0.3846 for object classification accuracy and 0.6208 scene classification accuracy, and object classification recall is 0.2414, all of which are higher than baseline model. In particular, scene classification accuracy is gaining a significant leap.

4.3.5 The two improvements above combined together

When combine improved object-object geometry binary potentials and improved scene classification potentials, we are able to obtain the best performance. In the rest of paper, we would denote this combination as **our best CRF model**. When using the configuration new-sce. + seg. + geo. + cpmc + sce.-obj. + new-obj.-next + obj.-top, we obtain the object precision to be 0.3859 and the recall to be 0.2492.

configuration	object	scene
scene appearance + **nn.obj.**	0.5387	0.5520
sce. + **nn.obj.** + geometry	0.5682	0.5520
sce. + **nn.obj.** + geo. + sce.-object	0.5708	0.5581
sce. + **nn.obj.** + geo. + sce.-obj. + obj.-next	0.5754	0.5581
sce. + **nn.obj.** + geo. + sce.-obj. + obj.-next. + obj.-obj.	0.5793	0.5749
sce. + **nn.obj.** + geo. + sce.-obj. + obj.-next. + obj.-obj. + obj.-top	0.5703	0.5800

Table 3. Classification performance on ground truth detection, using segmentation potentials obtained from fine-tuned Alexnet using pre-trained ImageNet weights (denoted as **nn.obj.**)

configuration	object	scene
sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next**	0.6072	0.5872
sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top	0.6123	0.5971
new-sce. + segmentation	0.5446	0.5902
new-sce. + seg. + geometry	0.5921	0.5902
new-sce. + seg. + geo. + sce.-object	0.6118	0.6024
new-sce. + seg. + geo. + sce.-obj. + obj.-next	0.6170	0.6009
new-sce. + seg. + geo. + sce.-obj. + obj.-next. + obj.-obj.	0.6164	0.6024
new-sce. + seg. + geo. + sce.-obj. + obj.-next. + obj.-obj. + obj.-top	**0.6174**	**0.6055**
new-sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next**	0.6134	0.5979
new-sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top	0.6144	0.5963

Table 4. Classification performance on ground truth detection, using improved object-object geometry relationship implemented by us (denoted as **new-obj.-next**) and improved scene appearance obtained from fine-tuned Alexnet using pre-trained model from [22] (denoted as **new-sce.**)

The scene classification has boosted from 0.5902 to 0.6208 after sampling the CRF model, which illustrates that object potentials and binary potentials could in turn improve scene recognition.

To better understand our classification results for object and scene to further improve our model, we visualize the confusion matrix in Figure 8 and 9. The confusion matrix is defined with two dimensions, actual label and predict label. Our best CRF model outperforms baselines for most of the classes. The ambiguity between several classes may lead to the low accuracy on specific class, where we can further explore how to distinguish these classes.

We illustrate the power of potentials using our best CRF model in Figure 10. As the number of potentials increases, we can see the precision increases as well, which indicates the potentials in the CRF structure can boost the final performance by considering the inner relationships among scene and objects within the scene. The more we capture the relationships, the better we would have for our scene model.

4.3.6 Improved segmentation potential

Improved segmentation potential is obtained by taking the softmax score from a Convolutional Neural Network initialized by AlexNet parameters [9] and fine-tuned on the images in our dataset cropped by bounding boxes. The performance with this modified potential is shown in Table 3 (denoted as **nn.obj.**). [12] states that because the training set is relatively small, training a 21-class classifier on objects might not be a good idea. We tackled the data limitation problem with the idea of transfer learning, i.e. fine-tuning instead of training from scratch. The learning rate of newly-trained final fully-connected layer is set as 10 times the learning rate of previous layers. Experiments on the ground-truth situation shows that the CNN model is able to achieve results comparible with results obtained by original segmentation potentials proposed by [12], but slightly lower. Therefore, we choose not to incorporate this modified potential in our final model.

5. Conclusions

Three major conclusions from this project are:

1. Jointly solving scene classification and object classification problem by embedding them in an inference model such as CRF can help boost performance. This is because this type of models capture a holistic understanding for the scene, making it possible to correct some mistakes by reasoning, which would not have been possible to correct in vanilla classification models.

2. More powerful potentials lead to better performance of the system. Improved representations of geometric relationship, as well as better potentials obtained via Convolutional Neural Networks, can both further

configuration	object	scene	recall
sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next**	0.3831	0.6040	0.2492
sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top	0.3850	0.6086	0.2497
new-sce. + segmentation	0.3441	0.5902	0.2028
new-sce. + seg. + geometry	0.3498	0.5902	0.2068
new-sce. + seg. + geo. + cpmc	0.3838	0.5902	0.2470
new-sce. + seg. + geo. + cpmc + sce.-object	0.3845	0.6177	0.2492
new-sce. + seg. + geo. + cpmc + sce.-obj. + obj.-next	0.3844	0.6269	0.2406
new-sce.. + seg. + geo. + cpmc + sce.-obj. + obj.-next. + obj.-top	0.3846	0.6208	0.2414
new-sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next**	0.3854	0.6147	0.2499
new-sce. + seg. + geo. + cpmc + sce.-obj. + **new-obj.-next** + obj.-top	**0.3859**	**0.6208**	**0.2492**

Table 5. Classification performance taking eight cuboid candidates ($k = 8$) from detection results, using improved object-object geometry relationship (denoted as **new-obj.-next**) and improved scene appearance potential obtained from fine-tuned CNN (denoted as **new-sce.**)

	mantel	counter	toilet	sink	bathtub	bed	headboar	table	shelf	cabinet	sofa	chair	chest	refriger	oven	microwav	blinds	curtain	board	monitor	printer
#samples	14	199	34	108	28	174	36	527	316	679	238	875	179	48	35	51	161	141	56	153	40
original	0	30.6	41.9	35.2	0	37.7	33.3	33.9	31.1	39.2	37.4	34.6	22.1	3.2	22.2	55.6	32.8	35.6	26.1	37.0	0
best	0	39.2	47.2	36.3	0	42.2	50.0	43.3	35.6	36.9	39.2	38.1	41.9	25.0	40.0	57.1	33.3	32.5	53.3	50.7	20.0

Table 6. Class-specific performances using potentials proposed by [12] ($k = 8$)

boost performance.

3. Transfer learning makes it possible to fit good models without massive dataset. The labelled NYU-RGBD v2 dataset we have for this project is relatively small, but the idea of transfer learning makes it possible to obtain a much better scene classifier than in [12] (59.02% vs. 55.20% accuracy), and an almost-as-good object classifier (53.9% vs. 54.4% accuracy, specifically, the object classifier on [12] is not trained on the NYU-RGBD v2 dataset, but on some larger datasets [14]).

5.1. Bottleneck and future work

The bottleneck for this framework appears at detection. As is shown in Figure 7, there are many spurious detections which correspond to none of our object categories. Improvements in detection is possible via deep networks such as [7]. However, [7] is based on templates of four indoor scene settings, which does not cover all 13 scene classes in our database. One future work would be training more scene templates, which requires considerable training time and thus we were not able to do it in this project. Nevertheless, approaching scene understanding directly via deep network would be an interesting topic.

5.2. Repository

Code for our project is available at `https://github.com/STZhang/cs231a_project`.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011.

[2] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1312–1328, 2012.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[4] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller. Integrating visual and range data for robotic object detection. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[6] Y. M. Kim, N. J. Mitra, D.-M. Yan, and L. Guibas. Acquiring 3d indoor environments with variability and repetition. *ACM Transactions on Graphics (TOG)*, 31(6):138, 2012.

[7] Y. Z. M. B. P. Kohli, S. Izadi, and J. Xiao. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. *arXiv preprint arXiv:1603.04922*, 2016.

[8] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in neural information processing systems*, pages 244–252, 2011.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[11] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3d scenes. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1330–1337. IEEE, 2012.

[12] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1417–1424, 2013.

[13] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[14] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766. IEEE, 2012.

[15] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 601–608. IEEE, 2011.

[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.

[17] S. Song and J. Xiao. Sliding shapes for 3d object detection in depth images. In *Computer Vision–ECCV 2014*, pages 634–651. Springer, 2014.

[18] S. Song and J. Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. 2016.

[19] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. In *Computer Vision–ECCV 2010*, pages 182–195. Springer, 2010.

[20] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, pages 1–20, 2014.

[21] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv*, 2016.

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.