Recognizing Indoor Scenes

Ariadna Quattoni
CSAIL, MIT
UC Berkeley EECS & ICSI
ariadna@csail.mit.edu

Antonio Torralba
CSAIL, MIT
32 Vassar St.,
Cambridge, MA 02139
torralba@csail.mit.edu

Abstract

Indoor scene recognition is a challenging open problem in high level vision. Most scene recognition models that work well for outdoor scenes perform poorly in the indoor domain. The main difficulty is that while some indoor scenes (e.g. corridors) can be well characterized by global spatial properties, others (e.g, bookstores) are better characterized by the objects they contain. More generally, to address the indoor scenes recognition problem we need a model that can exploit local and global discriminative information. In this paper we propose a prototype based model that can successfully combine both sources of information. To test our approach we created a dataset of 67 indoor scenes categories (the largest available) covering a wide range of domains. The results show that our approach can significantly outperform a state of the art classifier for the task.

1. Introduction

There are a number of approaches devoted to scene recognition that have been shown to be particulary successful in recognizing outdoor scenes. However, when these approaches are tested on indoor scene categories the results drop dramatically for most common indoor scenes. Fig. 1 shows results of a variety of state of the art scene recognition algorithms applied to a dataset of fifteen scene categories [10, 3, 8]. Common to all the approaches compared in this graph is their lower performance on indoor categories (RAW: 26.5%, Gist: 62.9%, Sift: 61.9%) in comparison with the performance achieved on the outdoor categories (RAW: 32.6%, Gist: 78.1%, Sift: 79.1%). [1]

[1]Note that the performances differ from the ones reported in [8]. The difference is that here we have cropped all the images to be square and with 256×256 pixels. The original dataset has images of different resolutions and aspect ratios that correlate with the categories providing non-visual discriminant cues.

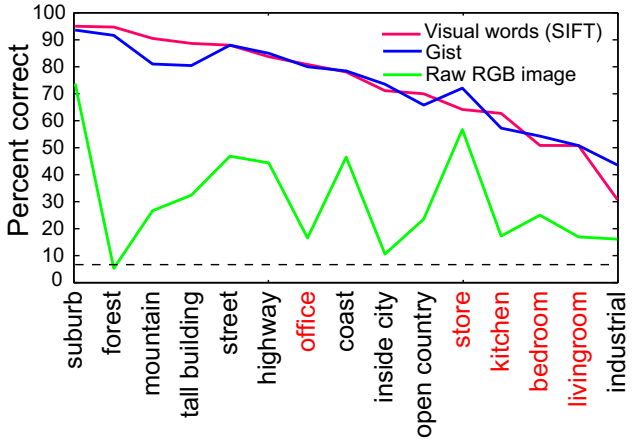

Figure 1. Comparison of Spatial Sift and Gist features for a scene recognition task. Both set of features have a strong correlation in the performance across the 15 scene categories. Average performance for the different features are: Gist: 73.0%, Pyramid matching: 73.4%, bag of words: 64.1%, and color pixels (SSD): 30.6%. In all cases we use an SVM.

There is some previous work devoted to the task of indoor scene recognition (e.g., [15, 16]), but to the best of our knowledge none of them have dealt with the general problem of recognizing a wide range of indoor scenes categories. We believe that there are two main reasons for the slow progress in this area. The first reason is the lack of a large testbed of indoor scenes in which to train and test different approaches. With this in mind we created a new dataset for indoor scene recognition consisting of 67 scenes (the largest available) covering a wide range of domains including: leisure, working place, home, stores and public spaces scene categories.

The second reason is that in order to improve indoor scene recognition performance we need to develop image representations specifically tailored for this task. The main difficulty is that while most outdoor scenes can be well characterized by global image properties this is not true of all

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

Figure 2. Summary of the 67 indoor scene categories used in our study. To facilitate seeing the variety of different scene categories considered here we have organized them into 5 big scene groups. The database contains 15620 images. All images have a minimum resolution of 200 pixels in the smallest axis.

indoor scenes. Some indoor scenes (e.g. corridors) can indeed be characterized by global spatial properties but others (e.g bookstores) are better characterized by the objects they contain. For most indoor scenes there is a wide range of both local and global discriminative information that needs to be leveraged to solve the recognition task.

In this paper we propose a scene recognition model specifically tailored to the task of indoor scene recognition. The main idea is to use image prototypes to define a mapping between images and scene labels that can capture the fact that images containing similar objects must have similar scene labels and that some objects are more important than others in defining a scene's identity.

Our work is related to work on learning distance functions [4, 7, 9] for visual recognition. Both methods learn to combine local or elementary distance functions. The are two main differences between their approach an ours. First, their method learns a weighted combination of elementary distance functions for each training sample by minimizing a ranking objective function. Differently, our method learns a weighted combination of elementary distance functions for a set of prototypes by directly minimizing a classification objective. Second, while they concentrated on object recognition and image retrieval our focus is indoor scene recognition.

This paper makes two contributions, first we provide a unique large and diverse database for indoor scene recognition. This database consists of 67 indoor categories covering a wide range of domains. Second, we introduce a model for indoor scene recognition that learns scene prototypes similar to start-constellation models and that can successfully combine local and global image information.

2. Indoor database

In this section we describe the dataset of indoor scene categories. Most current papers on scene recognition focus on a reduced set of indoor and outdoor categories. In contrast, our dataset contains a large number of indoor scene categories. The images in the dataset were collected from different sources: online image search tools (Google and Altavista), online photo sharing sites (Flickr) and the LabelMe dataset. Fig. 2 shows the 67 scene categories used in this study. The database contains 15620 images. All images have a minimum resolution of 200 pixels in the smallest axis.

This dataset poses a challenging classification problem. As an illustration of the in-class variability in the dataset, fig. 3 shows average images for some indoor classes. Note that these averages have very few distinctive attributes in comparison with average images for the fifteen scene categories dataset and Caltech 101 [11]. These averages suggest that indoor scene classification might be a hard task.

3. Scene prototypes and ROIs

We will start by describing our scene model and the set of features used in the rest of the paper to compute similarities between two scenes.

3.1. Prototypes and ROI

As discussed in the previous section, indoor scene categories exhibit large amounts of in-class appearance variability. Our goal will be to find a set of prototypes that best describes each class. This notion of scene prototypes has been used in previous works [12, 17].

In this paper, each scene prototype will be defined by a model similar to a constellation model. The main difference with an object model is that the root node is not allowed to move. The parts (regions of interest, ROI) are allowed to move on a small window and their displacements are independent of each other. Each prototype T_k (with $k = 1...p$) will be composed of m_k ROIs that we will denote by t_{kj}. Fig.4 shows an example of a prototype and a set of candidate ROIs.

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

Figure 3. Average images for a sample of the indoor scene categories. Most images within each category average to a uniform field due to the large variability within each scene category (this is in contrast with Caltech 101 or the 15 scene categories dataset [10, 3, 8]. The bottom 8 averages correspond to the few categories that have more regularity among examplars.

In order to define a set of candidate ROIs for a given prototype, we asked a human annotator to segment the objects contained in it. Annotators segmented prototype images for each scene category resulting in a total of 2000 manually segmented images. We used those segmentations to propose a set of candidate ROIs (we selected 10 for each prototype that occupy at least 1% of the image size).

We will also show results where instead of using human annotators to generate the candidate ROIs, we used a segmentation algorithm. In particular, we produce candidate ROIs from a segmentation obtained using graph-cuts [6].

3.2. Image descriptors

In order to describe the prototypes and the ROIs we will use two sets of features that represent the state of the art on the task of scene recognition.

We will have one descriptor that will represent the root node of the prototype image (T_k) globally. For this we will use the Gist descriptor using the code available online [10]. This results in a vector of 384 dimensions describing the entire image. Comparison between two Gist descriptors is computed using Euclidean distance.

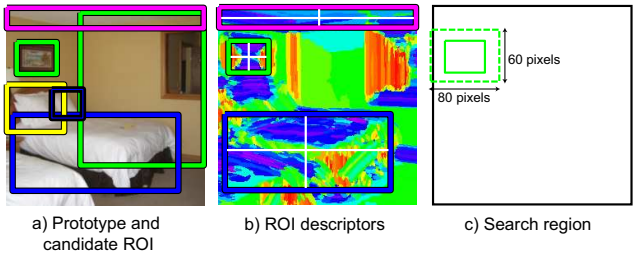

Figure 4. Example of a scene prototype. a) Scene prototype with candidate ROI. b) Illustration of the visual words and the regions used to compute histograms. c) Search window to detect the ROI in a new image.

To represent each ROI we will use a spatial pyramid of visual words. The visual words are obtained as in [14]: we create vector quantized Sift descriptors by applying K-means to a random subset of images (following [8] we used 200 clusters, i.e. visual words). Fig. 4.b shows the visual words (the color of each pixel represents the visual word to which it was assigned). Each ROI is decomposed into a 2x2 grid and histograms of visual words are computed for each window [8, 1, 13]. Distances between two regions are computed using histogram intersection as in [8].

Histograms of visual words can be computed efficiently using integral images, this results in an algorithm whose computational cost is independent of window size. The detection a ROI on a new image is performed by searching around a small spatial window and also across a few scale changes (Fig. 4.c). We assume that if two images are similar their respective ROIs will be roughly aligned (i.e. in similar spatial locations). Therefore, we only need to perform the search around a small window relative to the original location. Fig. 5 shows three ROIs and its detections on new images. For each ROI, the figure shows best and worst matches in the dataset. The figure illustrates the variety of ROIs that we will consider: some correspond to well defined objects (e.g., bed, lamp), regions (e.g., floor, wall with paintings) or less distinctive local features (e.g., a column, a floor tile). The next section will describe the learning algorithm used to select the most informative prototypes and ROIs for each scene category.

4. Model

4.1. Model Formulation

In scene classification our goal is to learn a mapping from images x to scene labels y. For simplicity, in this section we assume a binary classification setting. That is, each $y_i \in \{1, -1\}$ is a binary label indicating whether an image belongs to a given scene category or not. To model the multiclass case we use the standard approach of training one-versus-all classifiers for each scene; at test, we predict the

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

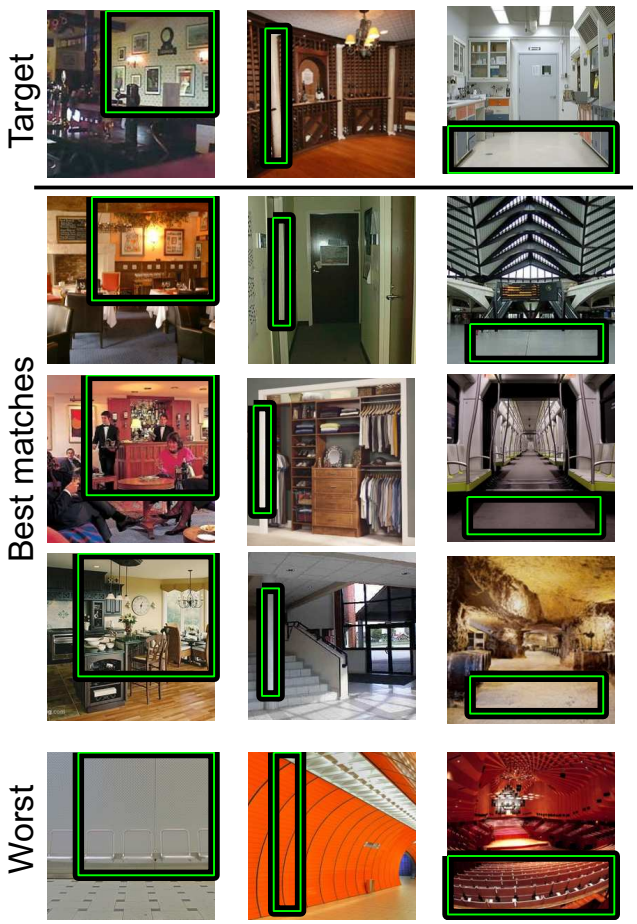

Figure 5. Example of detection of similar image patches. The top three images correspond to the query patterns. For each image, the algorithm tries to detect the selected region on the query image. The next three rows show the top three matches for each region. The last row shows the three worst matching regions.

scene label for which the corresponding classifier is most confident. However, we would like to note that our model can be easily adapted to an explicit multiclass training strategy.

As a form of supervision we are given a training set $D = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ of n pairs of labeled images and a set $S = \{T_1, T_2 \dots, T_p\}$ of p segmented images which we call prototypes. Each prototype $T_k = \{t_1, t_2, \dots, t_{m_k}\}$ has been segmented into m_k ROIs by a human annotator. Each ROI corresponds to some object in the scene, but we do not know their labels. Our goal is to use D and S to learn a mapping $h : X \rightarrow \mathbb{R}$. For binary classification, we would take the prediction of an image x to be $\text{sign}(h(x))$; in the multiclass setting, we will use directly $h(x)$ to compare it against other class predictions.

As in most supervised learning settings choosing an appropriate mapping $h : X \rightarrow \mathbb{R}$ becomes critical. In particular, for the scene classification problem we would like to learn a mapping that can capture the fact that images containing similar objects must have similar scene labels and that some objects are more important than others in defining a scene's identity. For example, we would like to learn that an image of a library must contain books and shelves but might or might not contain tables.

In order to define a useful mapping that can capture the essence of a scene we are going to use S. More specifically, for each prototype T_k we define a set of features functions:

$$f_{kj}(x) = \min_s d(t_{kj}, x_s) \tag{1}$$

Each of these features represents the distance between a prototype ROI t_{kj} and its most similar segment in x (see section 3 for more details of how these features are computed). For some scene categories global image information can be very important, for this reason we will also include a global feature $g_k(x)$ which is computed as the $L2$ norm between the Gist representation of image x and the Gist representation of prototype k. We can then combine all these feature functions to define a global mapping:

$$h(x) = \sum_{k=1}^{p} \beta_k \exp^{-\sum_{j=1}^{m_k} \lambda_{kj} f_{kj}(x) - \lambda_{kG} g_k(x)} \tag{2}$$

In the above formulation β and λ are the two parameter sets of our model. Intuitively, each β_k represents how relevant the similarity to a prototype k is for predicting the scene label. Similarly, each λ_{kj} captures the importance of a particular ROI inside a given prototype. We can now use the mapping h to define the standard regularized classification objective:

$$L(\beta, \lambda) = \sum_{i=1}^{n} l(h(x_i), y_i) + C_b||\beta||^2 + C_l||\lambda||^2 \tag{3}$$

The left term of equation 3 measures the error that the classifier incurs on training examples D in terms of a loss function l. In this paper we use the hinge loss, given by $l(h(x), y) = \max(0, 1 - yh(x))$ but other losses such as logistic loss could be used instead. The right hand terms of Equation 3 are regularization terms and the constants C_b and C_l dictate the amount of regularization in the model.

Finally, we introduce non-negativity constraints on the λ. Since each f_{kj} is a distance between image ROIs, these constraints ensure that their linear combination is also a global distance between a prototype and an image. This eases the interpretability of the results. Note that this global distance is used to induce a similarity measure in the classifier h.

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

4.2. Learning

In this section we describe how to estimate the model parameters $\{\beta^*, \lambda^*\} = \text{argmin}_{\beta,\lambda \geq 0} L(\beta, \lambda)$ from a training set D. The result of the learning stage will be the selection of the relevant prototypes for each class and the ROI that should be used for each prototype.

We use an alternating optimization strategy, which consists of a series of iterations that optimize one set of parameters given fixed values for the others. Initially the parameters are set to random values, and the process iterates between fixing β and minimizing L with respect to λ and fixing λ and minimizing L with respect to β.

We use a gradient-based method for each optimization step. Since our objective is non-differentiable because of the hinge loss, we use a sub-gradient of the objective, which we compute as follows:

Given parameter values, let Δ be the set of indices of examples in D that attain non-zero loss. Also, to simplify notation assume that parameter λ_{k0} and feature f_{k0} correspond to λ_{kG} and g_k respectively. The subgradient with respect to β is given by:

$$\frac{\partial L}{\partial \beta_k} = -\sum_{i \in \Delta} y_i \exp^{-\sum_{j=1}^{m_k} \lambda_{kj} f_{kj}(x_i)} + \frac{1}{2} C_b \beta_k$$

and the subgradient with respect to λ is given by:

$$\frac{\partial L}{\partial \lambda_{kj}} = \sum_{i \in \Delta} y_i \beta_k f_{kj}(x_i) \exp^{-\sum_{j=1}^{m_k} \lambda_{kj} f_{kj}(x_i)} + \frac{1}{2} C_l \lambda_{kj}$$

To enforce the non-negativity constraints on the λ we combine sub-gradient steps with projections to the positive octant. In practice we observed that this is a simple and efficient method to solve the constrained optimization step.

5. Experiments

In this section we present experiments for indoor scene recognition performed on the dataset described in section 2. We show that the model and representation proposed in this paper give significant improvement over a state of the art model for this task. We also perform experiments using different versions of our model and compare manual segmentations to segmentations obtained by running a segmentation algorithm.

In all cases the performance metric is the standard average multiclass prediction accuracy. This is calculated as the mean over the diagonal values of the confusion matrix. An advantage of this metric with respect to plain multiclass accuracy is that it is less sensitive to unbalanced distributions of classes. For all experiments we trained a one versus all classifier for each of the 67 scenes and combined their scores into a single prediction by taking the scene label with maximum confidence score. Other approaches are possible for combining the predictions of the different classifiers.

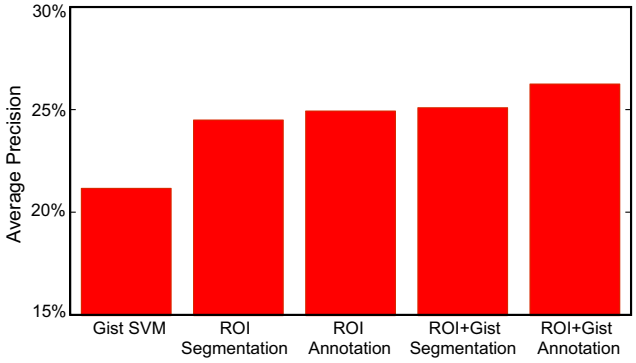

Figure 6. Multiclass average precision performance for the baseline and four different versions of our model.

We start by describing the four different variations of our model that were tested on these experiments. In a first setting we used the ROIs obtained from the manually annotated images and restricted the model to use local information only by removing the $g_k(x)$ features (ROI Annotation). In a a second setting we allowed the model to use both local and global features (ROI+Gist Annotation). In a third setting we utilized the ROIs obtained by running a segmentation algorithm and restricted the model to use local information only (ROI Segmentation). Finally, in the fourth setting we used the ROIs obtained from the automatic segmentation but allowed the model to exploit both local and global features (ROI+Gist Segmentation). All these models were trained with 331 prototypes.

We also compared our approach with a state of the art model for this task. For this we trained an SVM with a Gist representation and an RBF kernel (Gist SVM). In principle other features could have been used for this baseline but as it was shown in Fig. 1 Gist is one of the most competitive representations for this task.

To train all the models we used 80 images of each class for training and 20 images for testing. To train a one versus all classifier for category d we sample n positive examples and $3n$ negative examples.

6. Results

Figure 6 shows the average multiclass accuracy for the five models: Gist SVM, ROI Segmentation, ROI Annotation, ROI+Gist Segmentation and ROI+Gist Annotation.

As we can see from this figure combining local and global information leads to better performance. This suggests that both local and global information are useful for the indoor scene recognition task. Notice also that using automatic segmentations instead of manual segmentations

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

church inside 63.2%	studiomusic 36.8%	fastfood restaurant 23.5%	hairsalon 9.5%
elevator 61.9%	hospitalroom 35.0%	bar 22.2%	subway 9.5%
auditorium 55.6%	nursery 35.0%	clothingstore 22.2%	warehouse 9.5%
buffet 55.0%	trainstation 35.0%	casino 21.1%	meeting room 9.1%
classroom 50.0%	bathroom 33.3%	deli 21.1%	children room 5.6%
greenhouse 50.0%	laundromat 31.8%	bookstore 20.0%	shoeshop 5.3%
bowling 45.0%	stairscase 30.0%	waitingroom 19.0%	kindergarden 5.0%
cloister 45.0%	garage 27.8%	dining room 16.7%	restaurant 5.0%
concert hall 45.0%	gym 27.8%	bakery 15.8%	museum 4.3%
computerroom 44.4%	tv studio 27.8%	livingroom 15.0%	restaurant kitchen 4.3%
dentaloffice 42.9%	videostore 27.3%	movietheater 15.0%	jewelleryshop 0.0%
library 40.0%	gameroom 25.0%	bedroom 14.3%	laboratorywet 0.0%
inside bus 39.1%	pantry 25.0%	toystore 13.6%	mall 0.0%
closet 38.9%	poolinside 25.0%	operating room 10.5%	office 0.0%
corridor 38.1%	inside subway 23.8%	airport inside 10.0%	
grocerystore 38.1%	kitchen 23.8%	artstudio 10.0%	
locker room 38.1%	winecellar 23.8%	lobby 10.0%	
florist 36.8%		prison cell 10.0%	

Figure 7. The 67 indoor categories sorted by multiclass average precision (training with 80 images per class and test is done on 20 images per class).

causes only a small drop in performance.

Figure 7 shows the sorted accuracies for each class for the ROI+Gist-Segmentation model. Interestingly, five of the categories (greenhouse, computer-room, inside-bus, corridor and pool-inside) for which we observed some global regularity (see 3) are ranked among the top half best performing categories. But among this top half we also find four categories (buffet, bathroom, concert hall, kitchen) for which we observed no global regularity. Figure 8 shows ranked images for a random subset of scene categories for the ROI+Gist Segmentation model.

Figure 9 shows the top and bottom prototypes selected for a subset of the categories. We can see from these results that the model leverages both global and local information at different scales.

One question that we might ask is: how is the performance of the proposed model affected by the number of prototypes used? To answer this question we tested the performance of a version of our model that used global information only for different number of prototypes (1 to 200). We observed a logarithmic growth of the average precision as a function of the number of prototypes. This means that by allowing the model to exploit more prototypes we might be able to further improve the performance.

In summary, we have shown the importance of combining both local and global image information for indoor scene recognition. The model that we proposed leverages both and it can outperform a state of the art classifier for task. In addition, our results let us conclude that using automatic segmentations is similar to using manual segmentations and thus our model can be trained with a minimum amount of supervision.

7. Conclusion

We have shown that the algorithms that constitute the actual state of the art algorithms on the 15 scene categorization task [10, 3, 8] perform very poorly at the indoor recognition task. Indoor scene categorization represents a very challenging task due to the large variability across different exemplars within each class. This is not the case with many outdoor scene categories (e.g., beach, street, plaza, parking lot, field, etc.) which are easier to discriminate and several image descriptors have been shown to perform very well at that task. Outdoor scene recognition, despite being a challenging task has reached a degree of maturity that has allowed the emergence of several applications in computer vision (e.g. [16]) and computer graphics (e.g. [5]). However, most of those works have avoided dealing with indoor scenes as performances generally drop dramatically.

The goal of this paper is to attract attention to the computer vision community working on scene recognition to this important class of scenes for which current algorithms seem to perform poorly. In this paper we have proposed a representation able to outperform representations that are the current state of the art on scene categorization. However, the performances presented in this paper are close to the performance of the first attempts on Caltech 101 [2].

References

[1] A. Bosch, A. Zisserman, and X. Munoz.

[2] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[3] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *cvpr*, pages 524–531, 2005.

[4] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[5] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics*, 26, 2007.

[6] S. Jianbo and M. Jitendra. Normalized cuts and image segmentation. In *cvpr*, 1997.

[7] J. Krapac. Learning distance functions for automatic annotation of images. In *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, 2008.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *cvpr*, pages 2169–2178, 2006.

[9] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, June 2008.

[10] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal in Computer Vision*, 42:145–175, 2001.

[11] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *In Toward*

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

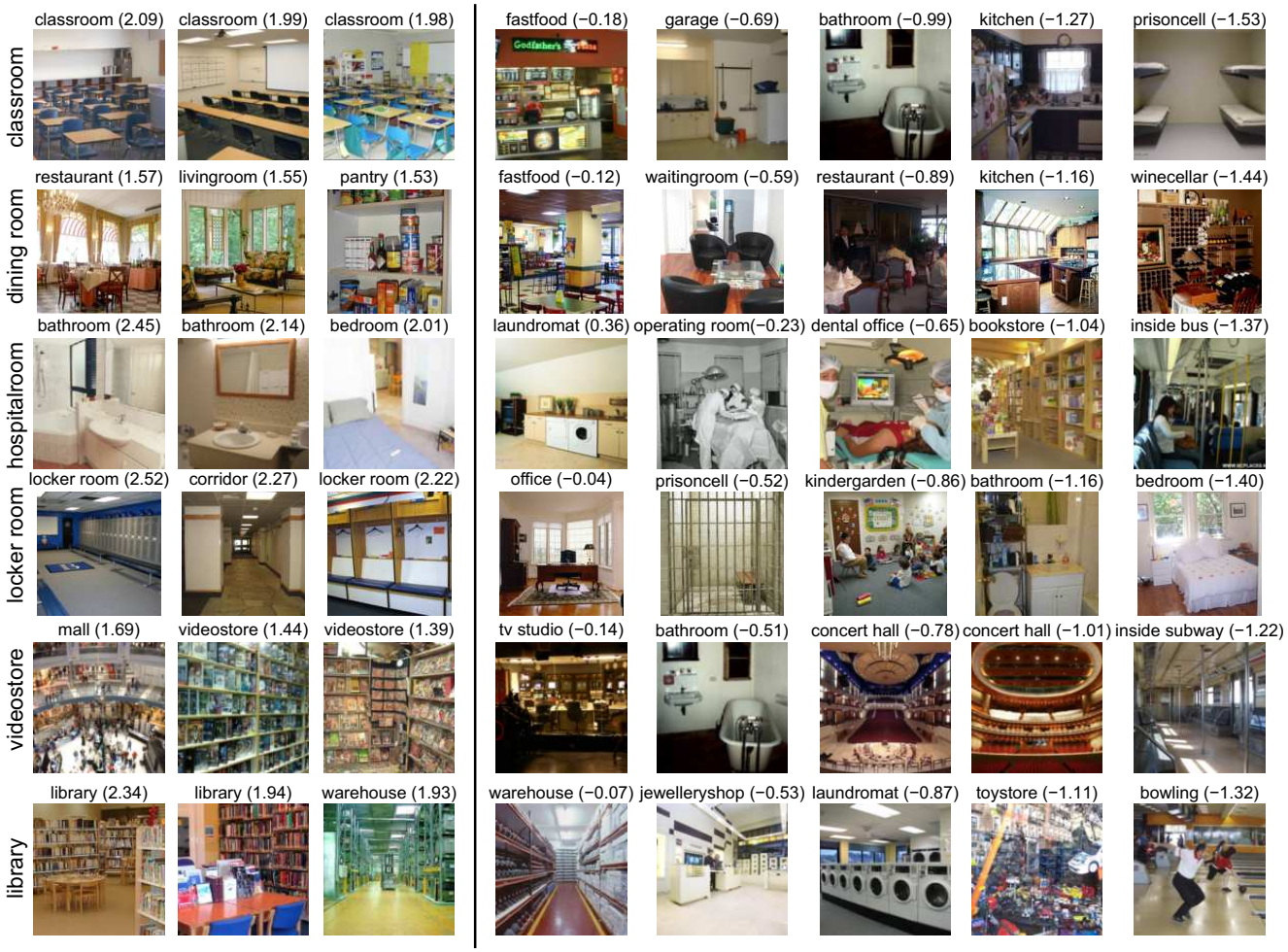

Figure 8. Classified images for a subset of scene categories for the ROI+Gist Segmentation model. Each row corresponds to a scene category. The name on top of each image denotes the ground truth category. The number in parenthesis is the classification confidence. The first three columns correspond to the highest confidence scores. The next five columns show 5 images from the test set sampled so that they are at equal distance from each other in the ranking provided by the classifier. The goal is to show which images/classes are near and far away from the decision boundary.

Category-Level Object Recognition, pages 29–48. Springer, 2006.

[12] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008.

[13] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[14] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006.

[15] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *CAIVD '98: Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, page 42, Washington, DC, USA, 1998. IEEE Computer Society.

[16] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *Intl. Conf. Computer Vision*, 2003.

[17] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766–786, October 2006.

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.

Figure 9. Prototypes for a subset of scene categories and sorted by their weight. The 7 first columns correspond to the highest rank prototypes and the last two columns show prototypes with the most negative weights. The thickness of each bounding box is proportional to the value of the weight for each ROI: λ_{kj}.

Authorized licensed use limited to: Korea University. Downloaded on June 20,2020 at 03:31:16 UTC from IEEE Xplore. Restrictions apply.