SUN database: Large-scale scene recognition from abbey to zoo

SUN Database: Large-scale Scene Recognition from Abbey to Zoo

Jianxiong Xiao James Hays[†] Krista A. Ehinger Aude Oliva Antonio Torralba
jxiao@csail.mit.edu hays@cs.brown.edu kehinger@mit.edu oliva@mit.edu torralba@csail.mit.edu
Massachusetts Institute of Technology [†]Brown University

Abstract

Scene categorization is a fundamental problem in computer vision. However, scene understanding research has been constrained by the limited scope of currently-used databases which do not capture the full variety of scene categories. Whereas standard databases for object categorization contain hundreds of different classes of objects, the largest available dataset of scene categories contains only 15 classes. In this paper we propose the extensive Scene UNderstanding (SUN) database that contains 899 categories and 130,519 images. We use 397 well-sampled categories to evaluate numerous state-of-the-art algorithms for scene recognition and establish new bounds of performance. We measure human scene classification performance on the SUN database and compare this with computational methods. Additionally, we study a finer-grained scene representation to detect scenes embedded inside of larger scenes.

1. Introduction

Whereas the fields of computer vision and cognitive science have developed several databases to organize knowledge about object categories [10, 28], a comprehensive database of real world scenes does not currently exist (the largest available dataset of scene categories contains only 15 classes). By "scene" we mean a place in which a human can act within, or a place to which a human being could navigate. How many kinds of scenes are there? How can the knowledge about environmental scenes be organized? How do the current state-of-art scene models perform on more realistic and ill-controlled environments, and how does this compare to human performance?

To date, computational work on scene and place recognition has classified natural images within a limited number of semantic categories, representing typical indoor and outdoor settings [17, 7, 23, 32, 21, 3, 29]. However, any restricted set of categories fails to capture the richness and diversity of environments that make up our daily experience. Like objects, scenes are associated with specific functions and behaviors, such as eating in a restaurant, drinking in a pub, reading in a library, and sleeping in a bedroom. Scenes, and their associated functions, are closely related to the visual features that structure the space. The function of environments can be defined by their shape and size (a narrow corridor is for walking, an expansive arena is for public events), by their constituent materials (snow, grass, water, wood), or by embedded objects (table and chairs, displays of jewelry, laboratory equipment).

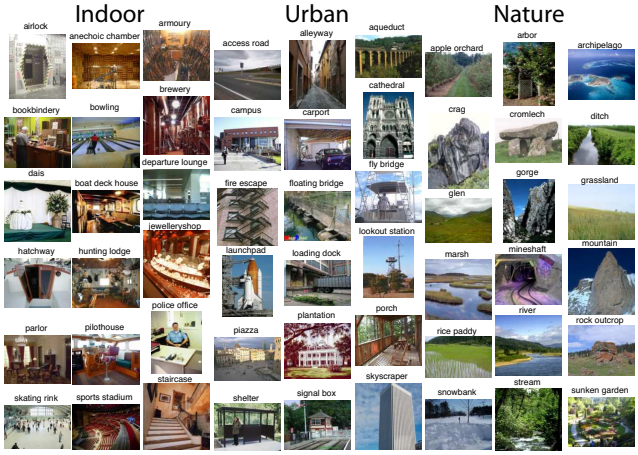


Figure 1. Examples of scene categories in our dataset.

The spatial layout and structure of a place often constrain human activities: for instance, a long open corridor affords walking or running, a classroom affords seating. Previous attempts to characterize environmental scenes have capitalized on uncovering a manageable set of dimensions, features or objects that are correlated with the semantic of purpose of a space [21, 17, 7, 32, 12].

This paper has the following four objectives. First, we seek to quasi-exhaustively determine the number of different scene categories with different functionalities. Rather than collect all scenes that humans experience – many of which are accidental views such as the corner of an office or edge of a door – we identify all the scenes and places that are important enough to have unique identities in discourse, and build the most complete dataset of scene image cate-

gories to date. Second, we measure how accurately humans can classify scenes into hundreds of categories. Third, we evaluate the scene classification performance of state of the art algorithms and establish new bounds for performance on the SUN database and the 15 scene database using a kernel combination of many features. Finally, we study the possibility of detecting scenes embedded inside larger scenes.

2. A Large Database for Scene Recognition

In order to get a quasi-exhaustive list of environmental categories, we selected from the 70,000 terms of WordNet [8] available on the tiny Images dataset [28] all the terms that described scenes, places, and environments (any concrete noun which could reasonably complete the phrase I am in a *place*, or Let's go to the *place*). Most of the terms referred to basic and entry level places [30, 24, 25, 14] with different semantic descriptions. We did not include specific place names (like Grand Canyon or New York) or terms which did not seem to evoke a specific visual identity (territory, workplace, outdoors). Non-navigable scenes (such as a desktop) were not included, nor were vehicles (except for views of the inside of vehicles) or scenes with mature content. We included specific types of buildings (skyscraper, house, hangar), because, although these can be seen as objects, they are known to activate scene-processing-related areas in the human brain. [5]. We also maintained a high tolerance for vocabulary terms that may convey significance to experts in particular domains (e.g. a baseball field contains specialized subregions such the pitcher's mound, dugout, and bullpen; a wooded area could be identified as a pine forest, rainforest, orchard, or arboretum, depending upon its layout and the particular types of plants it contains). To the WordNet collection we added a few categories that seemed like plausible scenes but were missing from WordNet, such as jewelry store and mission[1].

This gave about 2500 initial terms of space and scene, and after bundling together synonyms (provided by WordNet, and separating scenes with different visual identities such as indoor and outdoor views of churches), the final dataset reaches 899 categories and 130,519 images. We refer to this dataset as "SUN" (Scene UNderstanding) database[2].

[1] An alternate strategy to obtain a list of relevant scene categories is to record the visual experience of an observer and to count the number of different scene categories viewed. This procedure is unlikely to produce a complete list of all scene categories, as many scene categories are only viewed in rare occasions (e.g., a cloister, a corn field, etc.). However, we use this to validate the completeness of the list provided by WordNet. A set of 7 participants where asked to write down every half an hour the name of the scene category in which they were. They reported scenes for a total period of 284 hours. In that period, they reported 52 different scene categories, all of them within the set of scenes covered by our dataset.

[2] All the images and scene definitions are available at http://groups.csail.mit.edu/vision/SUN/.

Figure 2. SUN categories with the highest human recognition rate.

For each scene category, images were retrieved using WordNet terminology from various search engines on the web [28]. Only color images of 200×200 pixels or larger were kept. Each image was examined to confirm whether or not it fit a detailed, verbal definition for its category. For similar scene categories (e.g. "abbey", "church", and "cathedral") explicit rules were formed to avoid overlapping definitions. Degenerate or unusual images (black and white, distorted colors, very blurry or noisy, incorrectly rotated, aerial views, noticeable borders) were removed. All duplicate images, within and between categories, were removed.

For many of the 899 SUN categories the Internet search returns relatively few unique photographs[3]. For all experiments in this paper we use *only* the 397 categories for which there are at least 100 unique photographs.

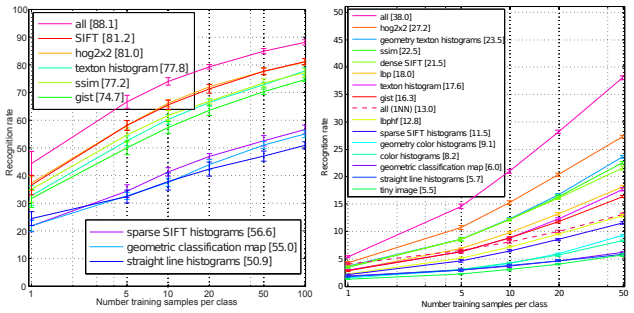
3. Human Scene Classification

In the previous section we provide a rough estimate on the number of common scene types that exist in the visual world and built the extensive SUN database to cover as many of those scenes as possible. In this section, we measure human scene classification performance on the SUN database. We have two goals: 1) to show that our database is constructed consistently and with minimal overlap between categories 2) to give an intuition about the difficulty of 397-way scene classification and to provide a point of comparison for computational experiments (Section 4.2).

Measuring human classification accuracy with nearly 400 categories is difficult. We do not want to penalize humans for being unfamiliar with our specific scene taxonomy, nor do we want to require extensive training concerning the definitions and boundaries between scenes (however, such training was given to those who built the database). Ideally our scene categories would be clear and distinct enough such that an untrained participant can unambiguously assign a categorical label to each scene in our database given a list of possible scene types. To help par-

[3] Examples of under-sampled categories include "airlock", "editing room", "grotto", "launchpad", "lava flow", "naval base", "oasis", "ossuary", "salt plain", "signal box", "sinkhole", "sunken garden", "winners circle"

Figure 3. Top row: SUN categories with the lowest human recognition rate. Below each of these categories, in the remaining three rows, are the most confusing classes for that category.



(a) 15 scene dataset (b) SUN database

Figure 4. Recognition performance on the 15 scene dataset[21, 17, 7], and our SUN database. For the 15 scene dataset, the combination of all features (88.1%) outperforms the current state of the art (81.4%) [17].

ticipants know which labels are available, we group the 397 scene categories in a 3-level tree, and the participants navigate through an overcomplete three-level hierarchy to arrive at a specific scene type (e.g. "bedroom") by making relatively easy choices (e.g. "indoor" versus "outdoor natural" versus "outdoor man-made" at the first level). Many categories such as "hayfield" are duplicated in the hierarchy because there might be confusion over whether such a category belongs in the natural or man-made sub-hierarchies. This hierarchy is used strictly as a human organizational tool, and plays no roll in our experimental evaluations. For each leaf-level SUN category the interface shows a prototypical photograph of that category.

We measure human scene classification accuracy using Amazon's Mechanical Turk (AMT). For each SUN category we measure human accuracy on 20 distinct test scenes, for a total of $397 \times 20 = 7940$ experiments or HITs (Human Intelligence Tasks in AMT parlance). We restricted these HITs to participants in the U.S. to help avoid vocabulary confusion.

On average, workers took 61 seconds per HIT and achieved 58.6% accuracy at the leaf level. This is quite high considering that chance is 0.25% and numerous categories are closely related (e.g. "dining room", "dining car", "home dinette", and "vehicle dinette"). However, a significant number of workers have 0% accuracy – they do not appear to have performed the experiment rigorously. If we instead focus on the "good workers" who performed at least 100 HITs and have accuracy greater than 95% on the relatively easy first level of the hierarchy the leaf-level accuracy rises to 68.5%. These 13 "good workers" accounted for just over 50% of all HITs. For reference, an author involved in the construction of the database achieved 97.5% first-level accuracy and 70.6% leaf-level accuracy. Therefore, these 13 good workers are quite trustable. In the remainder of the paper, all evaluations and comparisons of human performance utilize only the data from the good AMT workers.

Figures 2 and 3 show the SUN categories for which the good workers were most and least accurate, respectively. For the least accurate categories, Figure 3 also shows the most frequently confused categories. The confused scenes are semantically similar – e.g. abbey to church, bayou to river, and sandbar to beach. Within the hierarchy, indoor sports and leisure scenes are the most accurately classified (78.8%) while outdoor cultural and historical scenes were least accurately classified (49.6%). Even though humans perform poorly on some categories, the confusions are typically restricted to just a few classes.

Human and computer performance are compared extensively in Section 4.2. It is important to keep in mind that the human and computer tasks could not be completely equivalent. The "training data" for AMT workers was a text label, a single prototypical photo, and their lifetime visual experience. For some categories, a lifetime of visual experience is quite large (e.g "bedroom") while for others it is quite small (e.g. "medina"). On the other hand, the computational methods had (up to) 50 training examples. It is also likely the case that human and computer failures are qualitatively different – human misclassifications are between semantically similar categories (e.g. "food court" to "fast-food restaurant"), while computational confusions are more likely to include semantically unrelated scenes due to spurious visual matches (e.g. "skatepark" to "van interior"). In Figure 8 we analyze the degree to which human and computational confusions are similar. The implication is that the human confusions are the most reasonable possible confusions, having the shortest possible semantic distance. But human performance isn't necessarily an upper bound – in fact, for many categories the humans are less accurate than the best computational methods (Figure 6).

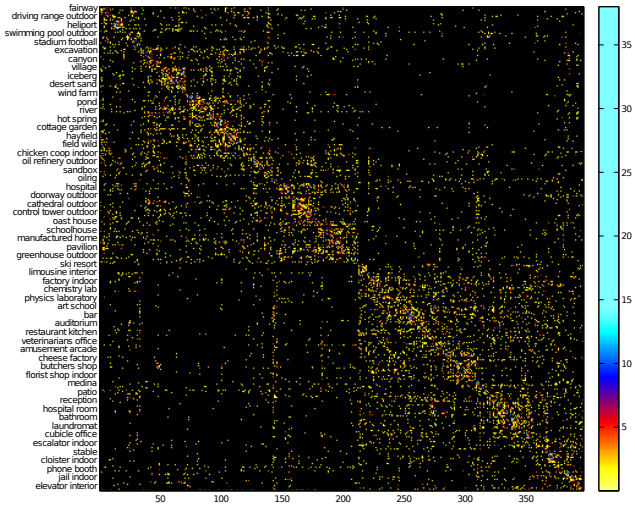


Figure 5. This figure shows the pattern of confusion across categories. The classes have been re-arranged to reveal the blocky structure. For clarity, the elements in the diagonal have been set to zero in order to increase the contrast of the off-diagonal elements. On the Y axis we show a sample of the scene categories. Confusions seem to be coherent with semantic similarities across classes. The scenes seem to be organized as indoor (top), urban (center) and nature (bottom).

Figure 6. Categories with similar and disparate performance in human and "all features" SVM scene classification. Human accuracy is the left percentage and computer performance is the right percentage. From top to bottom, the rows are 1) categories for which both humans and computational methods perform well, 2) categories for which both perform poorly, 3) categories for which humans perform better, and 4) categories for which computational methods perform better. The "all features" SVM tended to outperform humans on categories for which there are semantically similar yet visually distinct confusing categories. E.g. sandbar and beach, baseball stadium and baseball field, landfill and garbage dump.

4. Computational Scene Classification

In this section we explore how discriminable the SUN categories are with a variety of image features and kernels paired with 1 vs. all support vector machines.

4.1. Image Features and Kernels

We selected or designed several state-of-art features that are potentially useful for scene classification. GIST features [21] are proposed specifically for scene recognition tasks. Dense SIFT features are also found to perform very well at the 15-category dataset [17]. We also evaluate sparse SIFTs as used in "Video Google" [27]. HOG features provide excellent performance for object and human recognition tasks [4, 9], so it is interesting to examine their utility for scene recognition. While SIFT is known to be very good at finding repeated image content, the self-similarity descriptor (SSIM) [26] relates images using their internal layout of local self-similarities. Unlike GIST, SIFT, and HOG, which are all local gradient-based approaches, SSIM may provide a distinct, complementary measure of scene layout that is somewhat appearance invariant. As a baseline, we also include Tiny Images [28], color histograms and straight line histograms. To make our color and texton histograms more invariant to scene layout, we also build histograms for specific geometric classes as determined by [13]. The geometric classification of a scene is then itself used as a feature, hopefully being invariant to appearance but responsive to layout.

GIST: The GIST descriptor [21] computes the output energy of a bank of 24 filters. The filters are Gabor-like filters tuned to 8 orientations at 4 different scales. The square output of each filter is then averaged on a 4×4 grid.

HOG2x2: First, histogram of oriented edges (HOG) descriptors [4] are densely extracted on a regular grid at steps of 8 pixels. HOG features are computed using the code available online provided by [9], which gives a 31-dimension descriptor for each node of the grid. Then, 2×2 neighboring HOG descriptors are stacked together to form a descriptor with 124 dimensions. The stacked descriptors spatially overlap. This 2×2 neighbor stacking is important because the higher feature dimensionality provides more descriptive power. The descriptors are quantized into 300 visual words by k-means. With this visual word representation, three-level spatial histograms are computed on grids of 1×1, 2×2 and 4×4. Histogram intersection[17] is used to define the similarity of two histograms at the same pyramid level for two images. The kernel matrices at the three levels are normalized by their respective means, and linearly combined together using equal weights.

Dense SIFT: As with HOG2x2, SIFT descriptors are densely extracted [17] using a flat rather than Gaussian window at two scales (4 and 8 pixel radii) on a regular grid at steps of 5 pixels. The three descriptors are stacked together for each HSV color channels, and quantized into 300 visual words by k-means, and spatial pyramid histograms are used as kernels[17].

Figure 7. Selected SUN scene classification results using all features.

LBP: Local Binary Patterns (LBP) [20] is a powerful texture feature based on occurrence histogram of local binary patterns. We can regard the scene recognition as a texture classification problem of 2D images, and therefore apply this model to our problem. We also try the rotation invariant extension version [2] of LBP to examine whether rotation invariance is suitable for scene recognition.

Sparse SIFT histograms: As in "Video Google" [27], we build SIFT features at Hessian-affine and MSER [19] interest points. We cluster each set of SIFTs, independently, into dictionaries of 1,000 visual words using k-means. An image is represented by two histograms counting the number of sparse SIFTs that fall into each bin. An image is represented by two 1,000 dimension histograms where each SIFT is soft-assigned, as in [22], to its nearest cluster centers. Kernels are computed with χ^2 distance.

SSIM: Self-similarity descriptors [26] are computed on a regular grid at steps of five pixels. Each descriptor is obtained by computing the correlation map of a patch of 5×5 in a window with radius equal to 40 pixels, then quantizing it in 3 radial bins and 10 angular bins, obtaining 30 dimensional descriptor vectors. The descriptors are then quantized into 300 visual words by k-means and we use χ^2 distance on spatial histograms for the kernels.

Tiny Images: The most trivial way to match scenes is to compare them directly in color image space. Reducing the image dimensions drastically makes this approach more computationally feasible and less sensitive to exact alignment. This method of image matching has been examined thoroughly by Torralba et al.[28] for the purpose of object recognition and scene classification.

Line Features: We detect straight lines from Canny edges using the method described in Video Compass [15]. For each image we build two histograms based on the statistics of detected lines– one with bins corresponding to line angles and one with bins corresponding to line lengths. We use an RBF kernel to compare these unnormalized histograms. This feature was used in [11].

Texton Histograms: We build a 512 entry universal texton dictionary [18] by clustering responses to a bank of filters with 8 orientations, 2 scales, and 2 elongations. For each image we then build a 512-dimensional histogram by assigning each pixel's set of filter responses to the nearest texton dictionary entry. We compute kernels from normalized χ^2 distances.

Color Histograms: We build joint histograms of color in CIE L*a*b* color space for each image. Our histograms have 4, 14, and 14 bins in L, a, and b respectively for a total of 784 dimensions. We compute distances between these histograms using χ^2 distance on the normalized histograms.

Geometric Probability Map: We compute the geometric class probabilities for image regions using the method of Hoiem et al. [13]. We use only the ground, vertical, porous, and sky classes because they are more reliably classified. We reduce the probability maps for each class to 8×8 and use an RBF kernel. This feature was used in [11].

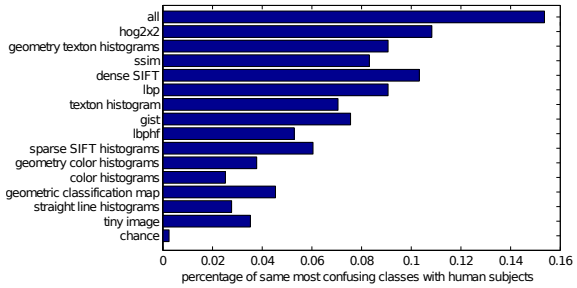


Figure 8. For each feature, the proportion of categories for which the largest *incorrect* (off-diagonal) confusion is the same category as the largest human confusion.

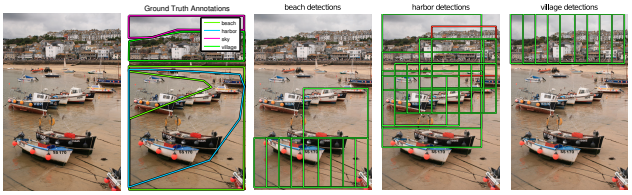
Geometry Specific Histograms: Inspired by "Illumination Context" [16], we build color and texton histograms for each geometric class (ground, vertical, porous, and sky). Specifically, for each color and texture sample, we weight its contribution to each histogram by the probability that it belongs to that geometric class. These eight histograms are compared with χ^2 distance after normalization.

4.2. Experiments and Analysis

With the features and kernels defined above, we train classifiers with one-vs-all Support Vector Machines. To better establish comparisons with other papers we start by providing results on the 15 scene categories dataset [21, 17, 7] (Figure 4(a)).

For the experiments with our SUN database, the performance of all features enumerated above is compared in Figure 4(b). For each feature, we use the same set of training and testing splits. For trials with fewer training examples, the testing sets are kept unchanged while the training sets are consistently decimated. The "all features" classifier is built from a weighted sum of the kernels of the individual features. The weight of each constituent kernel is proportional to the fourth power of its individual accuracy. As an additional baseline, we plot the performance of the "all features" kernel using one-nearest-neighbor classification. The 1-vs-all SVM has nearly three times higher accuracy. It is interesting to notice that with increasing amounts of training data, the performance increase is more pronounced with the SUN dataset than the 15 scene dataset. The confusion matrix of the "all features" combined classifier is shown in Figure 5.

The best scene classification performance with all features, 38%, is still well below the human performance of 68%. Computational performance is best for outdoor natural scenes (43.2%), and then indoor scenes (37.5%), and worst in outdoor man-made scenes (35.8%). Within the hierarchy, indoor transportation (vehicle interiors, stations, etc.) scenes are the most accurately classified (51.9%) while indoor shopping and dining scenes were least accurately classified (29.0%). In Figure 6, we examine the categories



(a) (b) (c) (d) (e)

Figure 9. (a) Examples of photographs that contain multiple scene types. (b) Ground truth annotations. (c)-(e): Detections of beach, harbor, and village scene categories in a single image. In all visualizations, correct detections are green and incorrect detections are red. Bounding box size is proportional to classifier confidence. For this and other visualizations, all detections above a constant confidence threshold are shown. In this result, one harbor detection is incorrect because it does not overlap with enough ground truth "harbor" annotation. "Beach" or "village" would have been acceptable.

for which human and machine accuracies are most similar and most dissimilar. In Figure 8 we examine the similarity in scene classification *confusions* between humans and machines. The better performing features not only tend to agree with humans on correct classifications, they also tend to make the same mistakes that humans make. However, humans have dramatically *fewer* confusions – for humans, on average, the three largest entries in each row of the confusion matrix sum to 0.95, while the "all features" SVM needs 11 entries to reach this mark.

5. Scene Detection

The dominant view in the scene recognition literature is that one image depicts one scene category. There are a few exceptions to this trend – In [31], each image is graded according to different scene categories. In [21], scenes are represented along continuous dimensions. However, current approaches assume that there is a unique scene label that can be assigned to an entire image, and this assumption is reasonable within a relatively narrow and disjoint selection of categories (e.g. the 15 in [17]), but the real world is not so neatly divided. Just as people can move continuously between scene categories (e.g. "office" into "corridor", "street" into "shopfront"), it is frequently the case that a single photograph depicts multiple scenes types at different scales and locations within the image (see Figure 9). By constructing a scene database with broad coverage we can explore the possibility of distinguishing all of the scenes types within single images.

We introduce the concept of *Scene Detection* – recognizing the scene type within image regions rather than entire images. Our terminology is consistent with the object detection literature[6] where object *classification* involves classifying entire images, while object *detection* requires

Scene Category	alley	balcony (exterior)	beach	boardwalk	bridge	building facade	cathedral (outdoor)	crosswalk	forest	gas station	harbor	hill	market (outdoor)	park	playground	plaza	restaurant patio	river	shopfront	sky	skyscraper	street	tower	village	Average
All Features	23.1	1.7	18.4	14.5	20.8	58.4	25.7	24.3	50.1	44.2	65.2	29.5	58.4	48.0	46.7	18.9	25.4	13.2	32.4	64.2	48.2	30.4	8.7	57.4	34.5
Tiny Images	5.4	1.2	5.0	5.4	7.7	38.7	9.1	5.3	34.0	4.7	23.9	13.8	18.0	27.4	12.4	12.7	7.7	6.0	9.8	59.8	12.1	18.2	8.1	10.6	14.9
Chance	3.1	1.0	4.2	2.2	6.6	30.8	5.1	3.0	17.9	3.5	12.4	4.8	4.8	17.1	4.8	9.2	4.0	5.2	6.6	35.9	5.7	13.4	2.3	6.7	8.8

Table 1. *Scene Detection Average Precision.* We compare the scene detection performance of our algorithm using all features and 200 training examples per class to baselines using only the "tiny images" feature and random guessing. "Sky", "Forest", and "Building Facade" make up a large portion of the test set and thus random guessing can achieve significant AP.

localizing and recognizing objects within an image.

As in object detection, we adopt a multiscale scanning-window approach to find sub-scenes. More specifically, we examine sub-images of a fixed aspect ratio at three different scales (1, 0.65. 0.42). We classify ~ 100 crops per image using the same features, training data, and classification pipeline used in Section 4. Finally we evaluate our detections against a novel, spatially annotated test set as described below.

5.1. Test Set and Evaluation Criteria

For these experiments we restrict ourselves to outdoor, urban environments to make the annotation process easier. We use 24 of the 398 well-sampled SUN categories. Our test set consists of 104 photographs containing an average of four scene categories each. In every photo we trace the ground truth spatial extent of each sub-scene and ensure that sub-scenes obey the same definitions used to construct the SUN database.

A ground truth annotation labeled "village" implies *if a sub-image bounding box (B_p) has at least $\mathcal{T}\%$ overlap with polygon (P_{gt}), it can be correctly classified as a "village"*. More precisely, a correct detection has $area(B_p \cap P_{gt})/area(B_p) \geq \mathcal{T}$. This notion of overlap is *not* symmetric as it is in object detection[6]. We do not care if the ground truth annotation is larger than the detection. A beach detection is correct even if the beach has much greater spatial extent than the detection. In this regard, scenes behave more like materials (or "stuff" [1]) in that they have unspecified spatial extent. The boundaries of a street scene or a forest are not as well defined as the boundaries of a chair.[4]

Annotations of differing categories can also spatially overlap – A "restaurant patio" can be wholly contained within a "plaza". Overlap can also occur where scene types transition or share dual meaning, e.g. the "beach" and "harbor" in Figure 9. In our experiments, we set the overlap threshold $\mathcal{T} = 15\%$. This is necessarily small because in some scene types, e.g. "tower" and "street", the defining element of the scene can occupy a relatively small percentage of pixels. A side effect of this low threshold, together with possibly overlapped annotations, is that for some sub-images there is more than one valid classification. At a given image location, the number of valid classifications can change with scale. In Figure 9, classifying the entire image "village" would be incorrect, but some sub-images are small enough such that the ground truth "village" annotation exceeds $\mathcal{T}\%$ overlap.

Unlike most object detection approaches, we have no "background" class. Under our overlap criteria, more than 90% of the sub-images in our test set have a valid scene classification. Those that do not are excluded from all evaluations. In order to achieve a perfect recall rate under our evaluation criteria an algorithm must densely classify nearly ~ 100 sub-images per test image (not just one correct detection per scene annotation). This is a difficult task because the detector must recognize sub-scenes that are transitioning from one category to another, even though such transition examples are infrequent in the SUN database because ambiguous scenes were filtered out. However, at lower recall rates our classifiers are quite accurate (on average, 80% accuracy at 20% recall).

5.2. Experiments and Analysis

To distinguish between the 24 scene detection classes we first train 1-vs-all SVMs using the same training data (50 per class), features, and distance metrics described in Section 4. For each class, we compute the average precision of the detector in a manner analogous to the PASCAL VOC evaluation, except for our differing overlap criteria. Average precision varies from 62% for "sky" to 2.8% for "balcony (exterior)" with an average of 30.1%. If we instead use more of the SUN database to train our detectors (200 exemplars per class), we get the performance shown in Table 1.

6. Conclusion

To advance the field of scene understanding we need datasets that encompass the richness and varieties of environmental scenes and knowledge about how scene categories are organized and distinguished from each other. In this work, we have proposed a quasi-exhaustive dataset of

[4] Along these same lines, non-maximum suppression of overlapped detections is not important in this domain. While it is reasonable to require that one chair should generate one detection, a forest can generate an arbitrary number of smaller forest scenes.

scene categories (899 environments). Using state-of-the-art algorithms for image classification, we have achieved new performance bounds for scene classification. We hope that the SUN database will help the community advance the state of scene understanding. Finally, we introduced a new task of scene detection within images.

Acknowledgments

This work is funded by NSF CAREER Awards 0546262 to A.O, 0747120 to A.T. and partly funded by BAE Systems under Subcontract No. 073692 (Prime Contract No. HR0011-08-C-0134 issued by DARPA), Foxconn and gifts from Google and Microsoft. K.A.E is funded by a NSF Graduate Research fellowship.

References

[1] E. H. Adelson. On seeing stuff: The perception of materials by humans. *Proceedings of the SPIE*, (4299), 2001.

[2] T. Ahonen, J. Matas, C. He, and M. Pietikäinen. Rotation invariant image description with local binary pattern histogram fourier features. In *SCIA*, 2009.

[3] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *J. of Machine Learning Research*, 3:1107–1135, Feb. 2003.

[4] N. Dalal and B. Triggs. Histogram of oriented gradient object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[5] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, 1998.

[6] M. Everingham, L. V. Gool, C. K. I. Williams, J. W. ands, and A. Zisserman. The pascal visual object classes (voc) challenge. *Intl. J. Computer Vision*, September 2009.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[8] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[11] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[12] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2005.

[13] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *Intl. J. Computer Vision*, 75(1), 2007.

[14] P. Jolicoeur, M. Gluck, and S. Kosslyn. Pictures and names: Making the connection. *Cognitive Psychology*, 16:243–275, 1984.

[15] J. Kosecka and W. Zhang. Video compass. In *Proc. European Conf. on Computer Vision*, pages 476–490, 2002.

[16] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2001.

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761 – 767, 2004.

[20] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Intl. J. Computer Vision*, 42:145–175, 2001.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[23] L. Renninger and J. Malik. When is scene recognition just texture recognition? *Vision Research*, 44:2301–2311, 2004.

[24] E. Rosch. Natural categories. *Cognitive Psychology*, 4:328–350, 1973.

[25] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[26] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[27] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2004.

[28] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large database for non-parametric object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, November 2008.

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

[30] B. Tversky and K. Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15:121–149, 1983.

[31] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *German Symposium on Pattern Recognition DAGM*, 2004.

[32] J. Vogel and B. Schiele. Semantic model of natural scenes for content-based image retrieval. *Intl. J. Computer Vision*, 72:133–157, 2007.