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

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva and Antonio Torralba

Source Code and Database Available http://groups.csail.mit.edu/vision/SUN/

Motivation

Given a picture, the task for scene categorization is to identify the place that it depicts. Algorithms have previously been evaluated on the incrementally built 15 scene database. But our visual world is far more diverse than this.

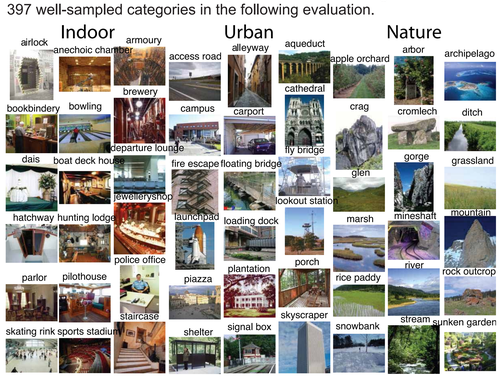
Oliva and Torralba, 2001

Fei Fei and Perona, 2005 | Lazebnik et al., 2006

Database Construction

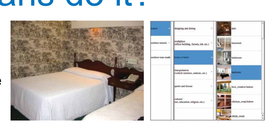
All Scene Terms | Search Images | Expert Screening

To get an exhaustive list of environmental categories, we selected from the 70,000 terms of all the terms of WordNet that described scenes, places, and environments. For each scene category, images were retrieved from various search engines. Each image was manually examined to confirm whether or not it fit a detailed, verbal definition for its category. After clean up, the dataset reaches 899 categories and 130,519 image. And we use 397 well-sampled categories in the following evaluation.

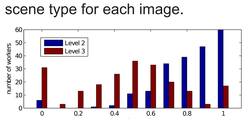


Can humans do it?

Experiment on Mechanical Turk. To show that our database is constructed consistently and with minimal overlap between categories, we hire human workers to navigate through an over-complete three-level hierarchy to arrive at a specific scene type for each image.



Turk GUI for 3-level Selection

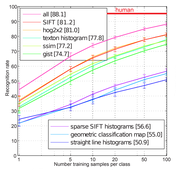


Worker Performance Distribution

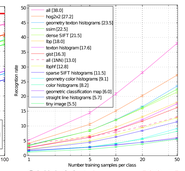
Most Accurate Categories

Computational Scene Classification

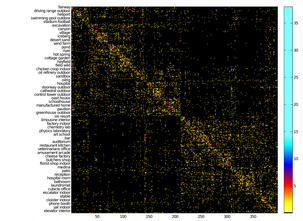
We selected or designed several state-of-art features that are potentially useful for scene classification, and several different kind of kernels including Histogram Intersection, Chi-square, RBF, L1, etc. We train classifiers with one-vs-all Support Vector Machines. The "all features" classifier is 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.



15-scene database



SUN 397 human [68.5]



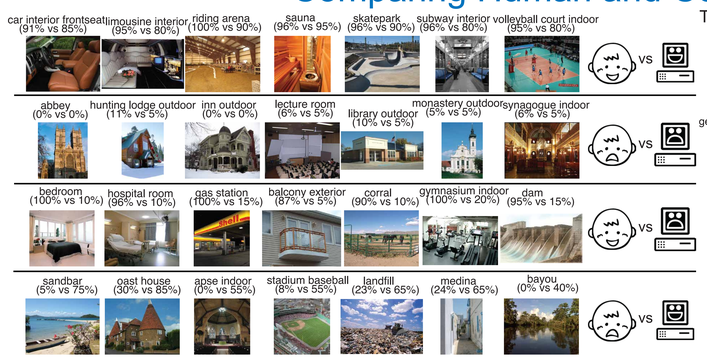
Confusion Matrix for Combined "all" Kernels

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.

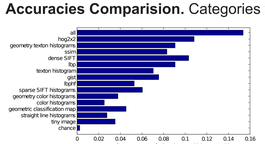
Computational performance is best for outdoor natural scenes (43.2%), and then indoor scenes (37.5%), and worst in outdoor man-made scenes (35.8%). Within the hierarchy, indoor transportation (vehicle interiors, stations, etc.) scenes are the most accurately classified (51.9%) while indoor shopping and dining scenes were least accurately classified (29.0%).

Example Classification Result

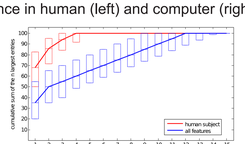
Comparing Human and Computer Errors



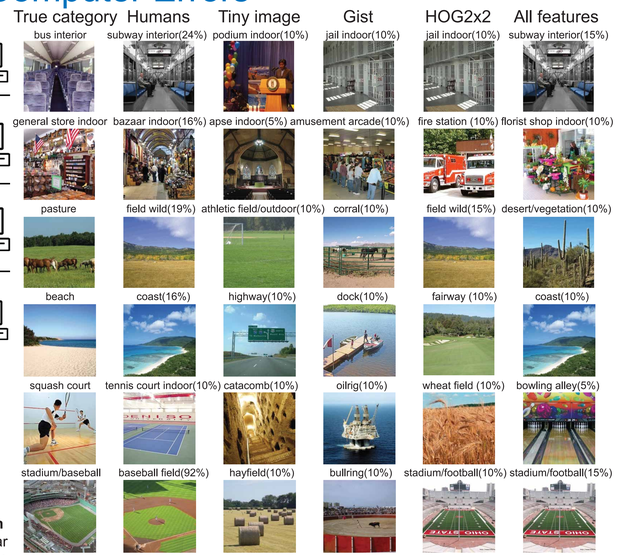
Accuracies Comparision. Categories with performance in human (left) and computer (right).



Similarity of Human & Computer. For each feature, the proportion of categories for which the largest off-diagonal confusion is the same category as the largest human confusion.



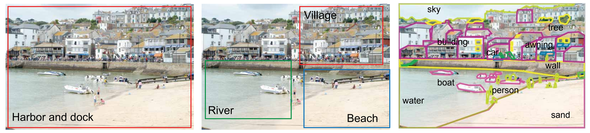
The cumulative sum of the n largest entries in each row of confusion matrix. Humans have far fewer confusions than the best performing computational approach.



Most Confused Catogories

Seeing Scenes in Scenes

There are many complementary levels of image understanding. One can understand images at the global scene level (left) or the local object level (right). Here, we introduce the concept of detection of local subscenes (middle).



Scene Classification | **Scene Detection** | **Object Segmentation**

Scene Detection: Given an image, the task of scene detection is to determine whether or not the specified kind of scene is present, and, if present, determine the locations and sizes of that scene.

We densely generate crops and label them in Turk to obtain masks for ground truth.

Example Training Data | **Example Testing Data** | **Average Precision-Recall**

Scene detection result examples. This figure shows the two most confident detections for several images. The detections with a red bounding box are incorrect detections and the green bounding boxes denote correct detections.

Subscene detection results for certain classes with recall precision curves.