Building a Taxonomy of Attributes for Fine-Grained Scene Understanding

Anonymous FGVC submission

Paper ID 21

Abstract

This paper presents the first effort to discover and exploit a diverse taxonomy of scene attributes. Starting with the fine-grained SUN database, we perform crowd-sourced human studies to find over 100 attributes that discriminate between scene categories. We construct an attribute-labeled dataset on top of the SUN database [7]. This "SUN Attribute database" spans more than 700 categories and 14,000 images and has potential for use in high-level scene understanding, attribute-based hierarchy construction, and fine-grained scene recognition.

1. Introduction

High-level scene understanding is a fundamental challenge in computer vision. Traditionally, computer vision algorithms have explained visual phenomena (objects, faces, actions, scenes, etc.) by giving each instance a categorical label. For scenes, this model has two significant problems: the space of scenes cannot be described by a well-defined taxonomy of non-overlapping categories, and simple category recognition does not provide any deep understanding or information about interesting inter-category and intra-category variations.

In the past two years there has been significant interest in *attribute-based* representations of visual phenomena [3, 1].In the domain of scenes, an attribute-based algorithm might describe an image with 'tiled floor', 'crowded', 'shopping', and 'shiny' in contrast to a categorical label such as 'store'. Attributes could be considered as an alternative to categorical descriptions of scenes, or they could be used to reinforce fine-grained classification techniques.

Scenes are difficult to model because instances in the same category have an incredible variety of layout, illumination, contents, occurrence, etc. Unlike with objects, people, or faces it is difficult to identify discriminative attributes, and it is more difficult to reliably isolate the same attributes in many instances of a scene. For example, eyes are a salient feature of a face, but what are the salient features of a mall? Can those mall features be identified for all malls?

It is also true that many scenes don't have a clear membership in any category, and many scenes seem to qualify for membership in several categories simultaneously. Ideally the boundary between attribute states is clearer. Even if a given scene does have a few ambiguous attributes, those that are not will still facilitate scene understanding. For this reason, one might expect attribute-based representations to fail more gracefully than strict categorical taxonomies.

2. Building a Taxonomy of Scene Attributes from Human Descriptions

The results of [5, 4] indicate that global scene attributes as well as local attributes are probably necessary for creating a discriminative set of scene attributes. For this initial endeavor into identifying scene attributes we limit ourselves to *global, binary* attributes. Still, the space of such attributes is effectively infinite. The vast majority of attributes (e.g., "Was these photo taken on a Tuesday", "Does this scene contain air?") are neither interesting nor discriminative among scene types. To determine relevant scene attributes, we conducted experiments with human users of Amazon's Mechanical Turk (AMT) service.

We will discover attributes by having humans describe and compare scenes. To ensure a maximally diverse set of probe scenes, we use the most prototypical image of each scene category in the SUN database as found by Ehinger et al. [2]. These 707 prototype images were the basis for our human experiments. In our first experiments we asked participants to list attributes for various individual prototypical scenes. From the thousands of responses, we were able to determine the most common categories of attributes. Below is a list of the attribute categories we identified in this experiment, along with a brief description of each.

- **Materials**: the material components, surface properties, or lighting found in a scene.
- **Functions or affordances**: activities that typically occur in a scene or that a scene may make possible, e.g. playing baseball in on a baseball field or thinking in a library.
- **Spatial envelope attributes**: these address global characteristics of a scene, for example the symmetry of a scene or a scene's degree of enclosure.
- **Objects**: the items commonly found in a particular scene.

Within these broad categories we want to focus on *discriminative* attributes - those that differentiate scene categories. Inspired by the "splitting task" of [5], we show participants two sets of scenes and ask them to list attributes that are present in one set but not the other. The images that make up these sets are prototypes from distinct, random categories. In the simplest case, with only one scene in each set, we found that participants would focus on trivial, happenstance objects or attributes. Such attributes would not be broadly useful for describing other scenes. At the other extreme, with many category prototypes in each set, it is rare that any attribute would be shared by one set and absent from the other. We found that having two scenes in each set produced a diverse but broadly applicable set of scene attributes. Figure 1 shows an example interface.

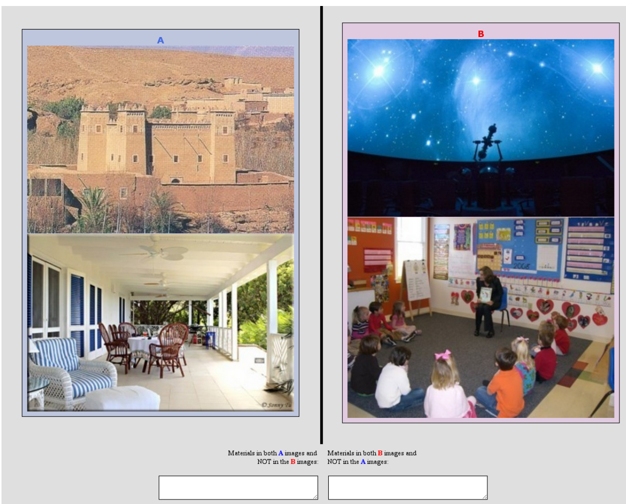


Figure 1. Mechanical Turk Human Intelligence Task - workers are asked to compare the images on the left to those on the right. Workers must attribute tags for left or right images into the text boxes at the bottom of the page.

The attribute gathering task was repeated more than 6000 times. From the thousands of raw discriminative attributes reported by participants, we collapse nearly synonymous attributes (e.g. dirt and soil) and then create our final taxonomy from the most frequently reported attributes. Some common emotional attributes (e.g. happy) were not used in order to focus our initial experiments on attributes that have a strong visual presence in scenes. The final list of attributes can be seen on the supplemental poster.

2.1. Labeling the Dataset

Now that we have a taxonomy of attributes we wish to create a large database of attribute-labeled scenes. In order to study the interplay of attribute and category-based representations, we build the "SUN Attribute database" on top of the fine-grained SUN categorical database. Building an attribute dataset on top of an existing fine-grained image dataset was successfully demonstrated by Russakovsky and Fei-Fei in [6] for the object domain.

We use Mechanical Turk to annotate 20 images from 717 scene categories. Participants are shown 20 scenes and asked to mark all the scenes that contained a specified attribute. The images are randomized to encourage the participants to examine each scene individually. Figure 2 shows an example interface.

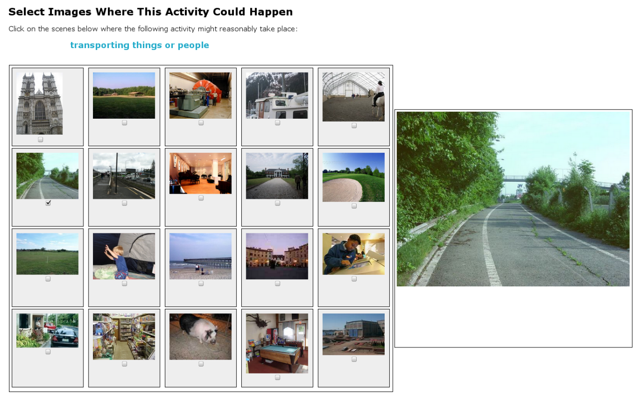


Figure 2. Attribute Labeling Interface for MTurk - workers are instructed to click on any of the 20 thumbnail-sized images that contain the given attribute (displayed in blue at the top of the page). Workers are able to mouse over a thumbnail and see the full-sized image in the review window on the right.

3. Future Work

The human experiments described in this paper are the first forays into a deep and interesting new domain. It remains to be seen how well attributes can be recognized and how useful such attributes will be for fine-grained categorization. One unexplored question is whether a principled hierarchy of the scene categories could be constructed by clustering based on attributes. Would the resulting categories resemble the lexicographical taxonomy used in the SUN database? It would also be interesting to see if attribute-based representations of scenes help explain human behaviors in studies of scene perception.

References

[1] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. *Computer Vision–ECCV 2010*, pages 663–676, 2010. 1

[2] K. Ehinger, A. Torralba, and A. Oliva. A taxonomy of visual scenes: Typicality ratings and hierarchical classification. *Journal of Vision*, 10(7):1237, 2010. 1

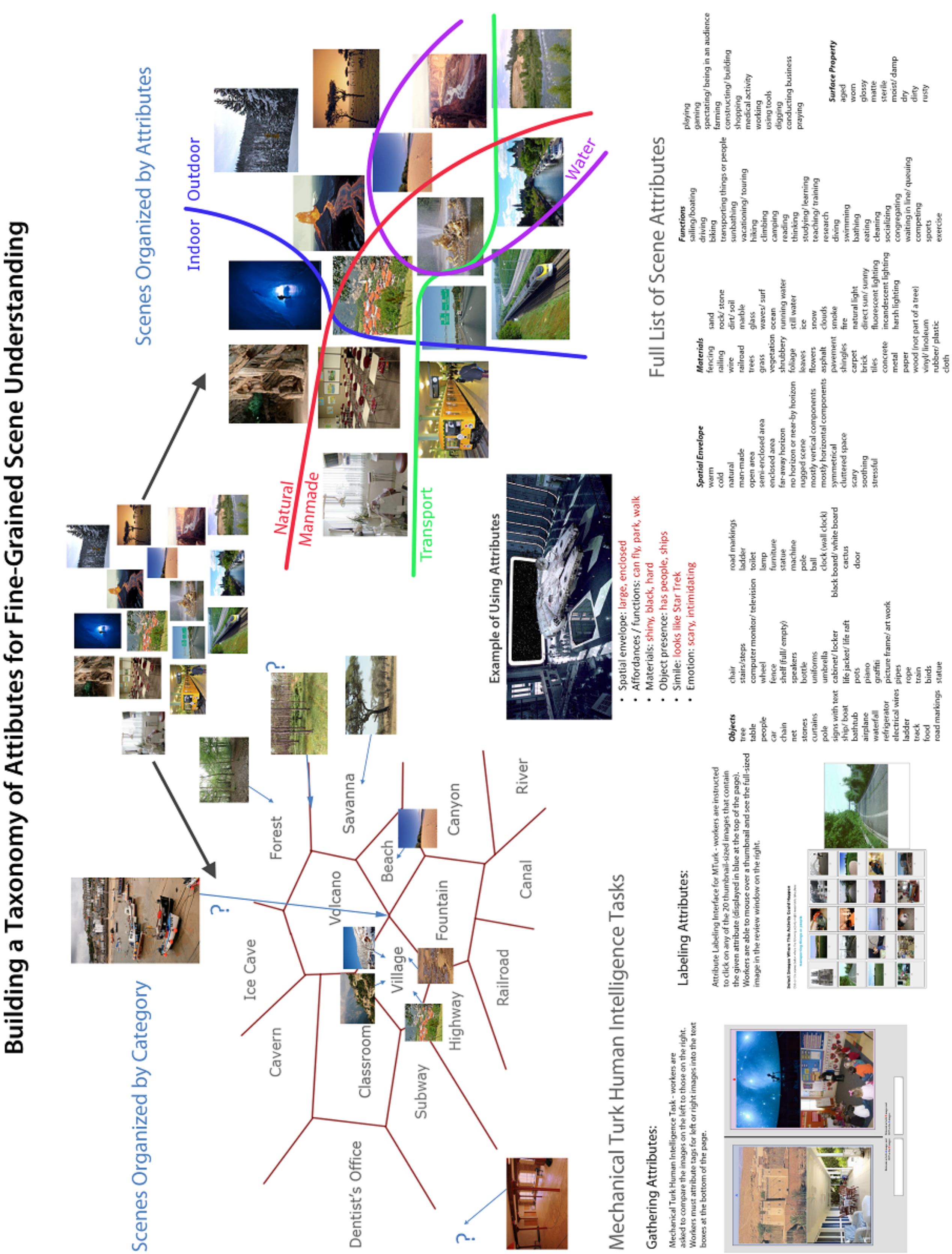
[3] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE, 2010. 1

[4] M. Greene and A. Oliva. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive psychology*, 58(2):137–176, 2009. 1

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 1

[6] O. Russakovsky and L. Fei-Fei. Attribute learning in largescale datasets. In *ECCV 2010 Workshop on Parts and Attributes*. Citeseer, 2010. 2

[7] J. Xiao, J. Hays, K. 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. 1


Building a Taxonomy of Attibutes for Fine-Grained Scene Understanding
Scenes Organized by Category
Forest
Savanna
Canyon
River
Beach
Volcano
Ice Cave
Fountain
Canal
Village
Railroad
Highway
Cavern
Classroom
Subway
Dentist's Office
Scenes Organized by Attributes
Indoor
Outdoor
Natural
Manmade
Transport
Water
Example of Using Attributes
Spatial envelope: large, enclosed
Affordances / functions: can fly, park, walk
Materials: shiny, black, hard
Object presence: has people, ships
Simile: looks like Star Trek
Emotion: scary, intimidating
Full List of Scene Attributes
Mechanical Turk Human Intelligence Tasks
Gathering Attributes:
Labeling Attributes: