SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes

Genevieve Patterson and James Hays
Brown University

Abstract

In this paper we present the first large-scale scene attribute database. First, we perform crowd-sourced human studies to find a taxonomy of 102 discriminative attributes. Next, we build the "SUN attribute database" on top of the diverse SUN categorical database. Our attribute database spans more than 700 categories and 14,000 images and has potential for use in high-level scene understanding and fine-grained scene recognition. We use our dataset to train attribute classifiers and evaluate how well these relatively simple classifiers can recognize a variety of attributes related to materials, surface properties, lighting, functions and affordances, and spatial envelope properties.

1. Introduction

Scene representations are vital to enabling many data-driven graphics and vision applications. There is important research on *low-level* representations of scenes (i.e. visual features) such as the gist descriptor [14] or spatial pyramid [12], but there has been little investigation into *high-level* representations of scenes (e.g. attributes or categories). The standard category-based recognition paradigm has gone largely unchallenged. In this paper, we explore a new, attribute-based representation of scenes.

Traditionally, computer vision algorithms describe visual phenomena (e.g. objects, faces, actions, scenes, etc.) by giving each instance a categorical label (e.g. cat, Halle Berry, drinking, downtown street, etc.). For scenes, this model has several significant issues, visualized in Figure 1: (1) The extent of scene understanding achievable is quite shallow – there is no way to express interesting *intra*-category variations. (2) The space of scenes is continuous, so hard partitioning creates numerous ambiguous boundary cases. (3) Images often simultaneously exhibit characteristics of multiple distinct scene categories. (4) A categorical representation can not generalize to types of scenes which were not seen during training.

In the past several years there has been interest in *attribute-based* representations of objects [7, 10, 6, 5, 1, 18, 21], faces [9], and actions [24, 13] as an alternative or complement to category-based representations. However, there has been only limited exploration of attribute-based representations for scenes, *even though scenes are uniquely poorly served by categorical representations*. For example, an object usually has unambiguous membership in one category. One rarely observes *issue 2* (e.g. this object is on the boundary between sheep and horse) or *issue 3* (e.g. this object is both a potted plant and a television).

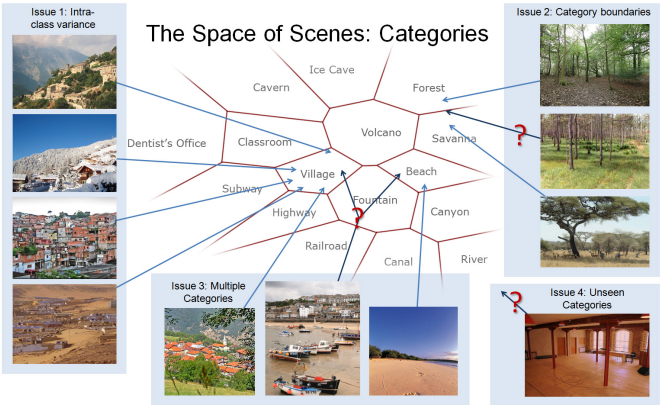


Figure 1: Visualization of a hypothetical space of scenes embedded in 2D and partitioned by categories. Categorical scene representations have several potential shortcomings: (1) Important intra-class variations such as the dramatic differences between four "village" scenes can not be captured, (2) hard partitions break up the continuous transitions between many scene types such as "forest" and "savanna", (3) an image can depict multiple, independent categories such as "beach" and "village", and (4) it is difficult to reason about unseen categories, whereas attribute-based representations lend themselves towards zero-shot learning [17].

In the domain of scenes, an attribute-based representation might describe an image with "concrete", "shopping", "natural lighting", "glossy", and "stressful" in contrast to a categorical label such as "store". Figure 2 visualizes the space of scenes partitioned by attributes rather than categories. Note, the attributes do not follow category boundaries. Indeed, that is one of the appeals of attributes – they can describe intra-class variation (e.g. a canyon might have

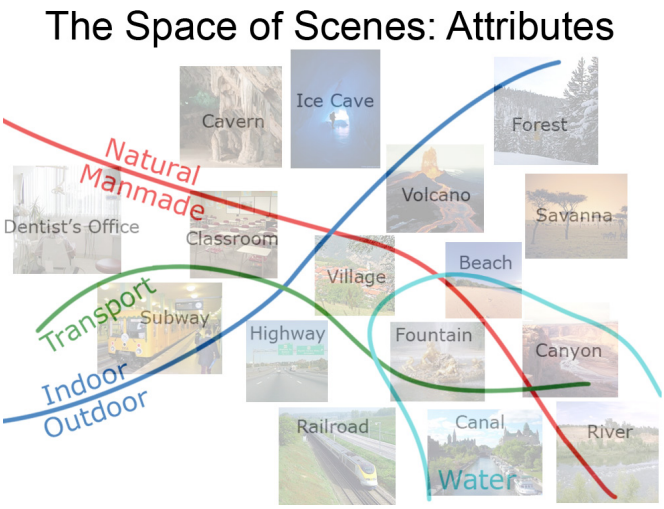


Figure 2: Hypothetical space of scenes partitioned by attributes rather than categories. In reality, this space is much higher dimensional and there are not clean boundaries between attribute presence and absence.

water or it might not) and inter-class relationships (e.g. both a canyon and a beach could have water).

A small set of scene attributes were explored in Oliva and Torralba's seminal "gist" paper [14] and follow-up work [15]. Eight "spatial envelope" attributes were found by having participants manually partition a database of eight scene categories. These attributes such as openness, perspective, and depth were predicted based on the scene gist representation. In [8], it was shown that these global scene attributes are predictive of human performance on a rapid basic-level scene categorization task.

More recently, Parikh and Grauman [16] argue for "relative" rather than binary attributes. They demonstrate results on the eight category outdoor scene database, but their training data is limited – they do not have per-scene attribute labels and instead provide attribute labels at the category level (e.g. all highway scenes should be more "natural" than all street scenes). This undermines one of the potential advantages of attribute-based representations – the ability to describe intra-class variation. In this paper we discover, annotate, and recognize 15 times as many attributes using a database spanning 90 times as many categories where *every scene* has independent attribute labels.

Paper Outline. This paper primarily focuses on the creation and verification of our SUN attribute database in the spirit of analogous database creation efforts such as ImageNet [3], LabelMe [19], and Tiny Images [22]. First, we derive a taxonomy of more than 100 scene attributes from crowd-sourced experiments (Section 2). Next, we use crowd-sourcing to construct our attribute-labeled dataset on top of a significant subset of the SUN database [23] spanning more than 700 categories and 14,000 images (Section 3). We visualize the distribution of scenes in attribute space (Section 4) and measure how well our attributes predict scene category (Section 5). Finally we measure how accurately we can predict attributes using existing image representations (Section 6).

2. Building a Taxonomy of Scene Attributes from Human Descriptions

Our first task is to establish a taxonomy of scene attributes for further study. The space of attributes is effectively infinite but the majority of possible attributes (e.g., "Was this photo taken on a Tuesday", "Does this scene contain air?") are not interesting. We are interesting in finding discriminative attributes which are likely to distinguish scenes from each other (not necessarily along categorical boundaries). We limit ourselves to *global, binary* attributes. This limitation is primarily economic – we collect millions of labels and annotating binary attributes is more efficient than annotating real-valued or relative attributes. Nonetheless, by averaging the binary labels from multiple annotators we produce a real-valued confidence for each attribute.

To determine which attributes are most relevant for describing scenes we perform open-ended image description tasks on Amazon Mechanical Turk (AMT). First we establish a set of "probe" images for which we will collect descriptions. We want a set of images which is maximally diverse and representative of the space of scenes. For this reason we use the images which human participants found to be most typical of 707 SUN database categories [4]. We first ask AMT workers to provide text descriptions of individual scenes. From thousands of such tasks (hereafter HITs, for human intelligence tasks) it emerges that people tend to describe scenes with five types of attributes: (1) Materials (e.g. cement, vegetation), (2) surface properties (e.g. rusty) (3) functions or affordances (e.g. playing, cooking), (4) spatial envelope attributes (e.g. enclosed, symmetric), and (5) object presence (e.g. cars, chairs).

Within these broad categories we focus on *discriminative* attributes. To find such attributes we develop a simplified, crowd-sourced version of the "splitting task" used by [14]. We show AMT workers two groups of scenes and ask them to list attributes of each type (material, surface property, affordance, spatial envelope, and object) that are present in one group but not the other. The images that make up these groups are typical scenes 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 (e.g. "treadmill" or "yellow shirt"). 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 random scene prototypes in each set elicited a diverse, broadly applicable set of attributes. Figure 3 shows an example interface.

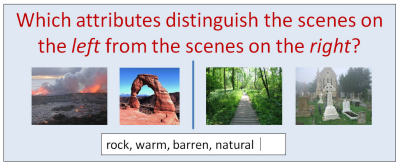


Figure 3: Mechanical Turk interface for discovering discriminative attributes.

The attribute gathering task was repeated over 6000 times. From the thousands of raw discriminative attributes reported by participants we collapse nearly synonymous responses (e.g. dirt and soil) into single attributes. We omit attributes related to aesthetics rather than scene content. For this study we also omit the object presence attributes from further discussion because prediction of object presence, i.e. object classification, has been thoroughly investigated (Additionally, the SUN database already has dense object labels for most scenes). Our participants did not report all of the spatial envelope attributes found by [14], so we manually add binary versions of those attributes so that our taxonomy is a superset of prior work. In total, we find 38 material, 11 surface property, 36 function, and 17 spatial envelope attributes.

3. Building the SUN Attribute Database

With our taxonomy of attributes finalized we create the first large-scale database of attribute-labeled scenes. We build the SUN attribute database on top of the existing SUN categorical database [23] for two reasons: (1) to study of the interplay between attribute-based and category-based representations and (2) to ensure a diversity of scenes. We annotate 20 scenes from each of the 717 SUN categories that contain at least 20 instances. Each scene has 102 attributes and each attribute will have multiple annotations. In total we gather more than four million labels. This necessitates a crowdsourced annotation strategy and we once again utilize AMT.

3.1. The Attribute Annotation Task

The primary difficulty of using a large, non-expert workforce is ensuring that the collected labels are accurate while keeping the annotation process fast and economical [20]. From an economic perspective, we want to have as many images labeled as possible for the lowest price. From a quality perspective, we want workers to easily and accurately label images. We find that particular UI design decisions and worker instructions significantly impacted throughput and quality of results. After several iterations, we choose a design where workers are presented with a grid of 4 dozen images and are asked to consider only a single attribute at a time. Workers click on images which exhibit

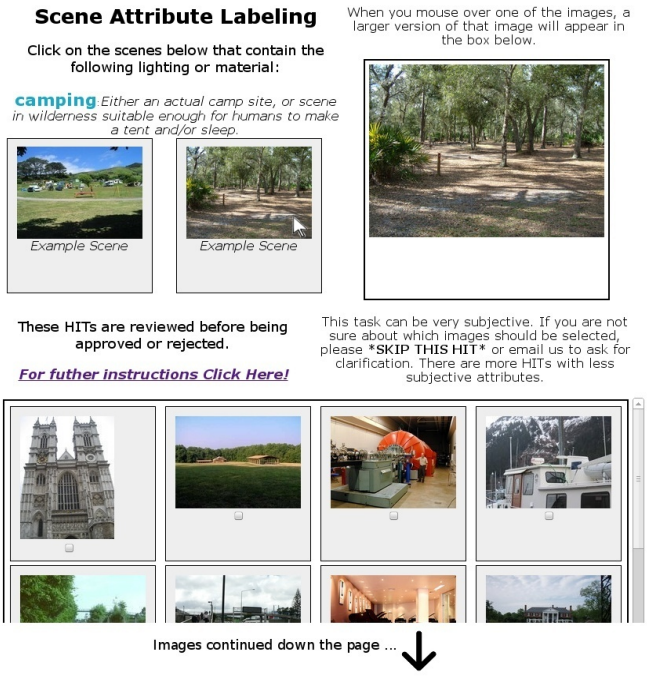
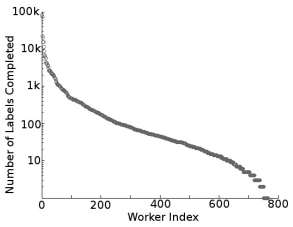


Figure 4: *Annotation interface for AMT workers*. The particular attribute being labeled is prominently shown and defined. Example scenes which contain the attribute are shown. The worker can not scroll these definitions or instructions off of their screen. When workers mouse over a thumbnail a large version appears in the preview window in the top right corner.

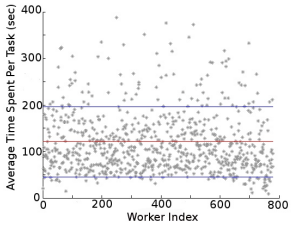
the attribute in question. Before working on our HITs, potential annotators are required to pass a quiz covering the fundamentals of attribute identification and image labeling. An example of our HIT user interface is shown in Figure 4.

Even after the careful construction of the annotation interface and initial worker screening, many workers' annotations are unreasonable. We use several techniques to filter out bad workers and then cultivate a pool of *trusted* workers:

Filtering bad workers. Deciding whether or not an attribute is present in a scene image is sometimes an ambiguous task. This ambiguity combined with the financial incentive to work quickly leads to sloppy annotation from some workers. In order to filter out those workers who performed poorly, we flag HITs which are outliers with respect to annotation time or labeling frequency. Some attributes, such as "ice" or "fire", rarely appear and are visually obvious and thus those HITs can be completed quickly. Other attributes, such as "man-made" or "natural light", occur in more than half of all scenes thus the expected completion time per HIT is higher. Any worker whose average number of labels or work time for a given attribute is greater than one standard deviation away from the average for all workers is added to a list of workers to manually review. We review by hand a



(a) The total number of labels completed by each worker.



(b) The average time (sec) each worker spent on a HIT of image labels.

Figure 5: These plots visualize our criteria for identifying suspicious workers to grade. Figure 5a shows the heavy-tailed distribution of worker contributions to the database. The top workers spent hundreds of hours on our HITs. The red line in plot 5b demarcates the average work time across all workers, and the blue lines mark the positive and negative standard deviation from the mean. Work time statistics are particularly useful from identifying scam workers as they typically rush to finish HITs.

fraction of the HITs for each suspicious worker as well as a random sampling of non-suspicious workers. Any worker whose annotations are clearly wrong is added to a blacklist. They are paid for their time, but none of their labels become part of the final dataset.

Cultivating good workers. The pay per HIT is initially \$0.03 but increases to \$0.05 plus 10% bonus after workers have a proven track record of accuracy. The net result of our filtering and bonus scheme is that we cultivate a pool of trained, efficient, and accurate annotators as emphasized by [2]. In general, worker accuracy rose over time and we omit over one million early annotations from the final database.

After labeling the entire dataset once with the general AMT population, we identify a smaller group of 38 trusted workers out of the ∼800 who participated. We repeat the labeling process two more times using only these trusted workers. The idea of finding and heavily utilizing *good* workers is in contrast to the "wisdom of the crowds" crowdsourcing strategy where consensus outweighs expertise, but is consistent with recent research such as [11] where good workers were shown to be faster *and* more accurate than the average of many workers. Figure 5 shows the contributions of all workers to our database.

Figure 6 qualitatively shows the result of our annotation process. To quantitatively assess accuracy we manually grade ∼600 random positive and ∼600 random negative AMT annotations in the database. For both types of annotation, we find ∼93% of labels to be reasonable. Negative labels are more common, making up 92% of the annotations. This does not seem to be an artifact of our interface (which defaults to negative), but rather it seems that scene attributes follow a heavy-tailed distribution with a few being very common (e.g. "natural") and most being rare (e.g. "wire").

In the following sections, our experiments rely on the consensus of multiple annotators rather than individual annotations. This increases the accuracy of our labels. We manually grade 5 scenes for each of our 102 attributes where the consensus was positive (2 or 3 votes) and likewise for negative (0 votes). In these 1020 scenes we find that the consensus annotation is reasonable 95% of the time for both positive and negative labels.

4. Exploring Scenes in Attribute Space

Now that we have a database of attribute-labeled scenes we can attempt to visualize that space of attributes. In Figure 7 we show all 14,340 of our scenes projected onto the two highest variance PCA bases. We sample several points in this space to show the types of scenes present as well as the nearest neighbors to those scenes in attribute space. For this analysis the distance between scenes is simply the Euclidean distance between their real-valued, 102-dimensional attribute vectors.

5. Predictive Power of Attributes

In this section we measure how well we can predict scene category from *ground truth* scene attributes. While the goal of this paper and our database is not necessarily to improve the task of scene categorization, this analysis does give some insight into the interplay between scene categories and scene attributes.

One hundred binary attributes could potentially predict membership in 700 hundred categories if the attributes were (1) independent and (2) consistent within each category, but neither of these are true. Many of the attributes are correlated (e.g. "farming" and "open area") and there is significant attribute variation within categories. Furthermore, many groups of SUN database scenes would require very specific attributes to distinguish them (e.g. "forest_needleleaf" and "forest_broadleaf"), so it would likely take several hundred attributes to very accurately predict scene categories.

Figure 8 shows how well we can predict the category of a scene with *known* attributes as we increase the number of training examples per category. Each image is represented by the ground truth average attribute labels with no visual features. We compare this to the classification accuracy of visual features [23] on the same data set. With 1 training example per category, attributes are roughly twice as accurate as visual features. Performance equalizes as the number of training examples approaches 20 per category. The performance of our attribute-based classifiers hints at the viability of zero-shot learning techniques which have access to attribute distributions for categories but no visual examples. The fact that category prediction accuracy increases signif-

Figure 6: The images in the table above are grouped by the number of positive labels (votes) they received from AMT workers. From left to right the visual presence of each attribute increases. Note that for functional / affordance attributes, AMT workers are instructed to positively label an image if the attribute is *likely to occur* in that image, not just if it is actually occurring.

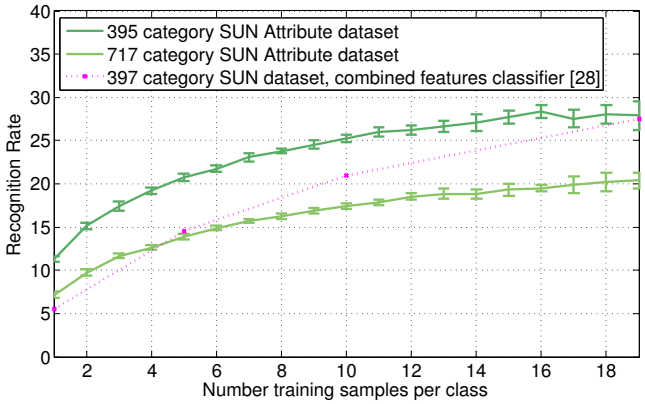


Figure 8: *Category recognition from ground truth attributes using an SVM*. We plot accuracy for the 717 category SUN Attribute dataset and for a subset of 395 categories which roughly match the evaluation of the SUN 397 dataset [23] (two categories present in [23] are not part of the SUN Attribute dataset). We compare attribute-based recognition to visual recognition by plotting the highest accuracy from [23] (pink dotted line).

icantly with more training examples may be a reflection of intra-class attribute variations.

6. Recognizing Scene Attributes

A motivation for creating the SUN Attribute dataset is to enable deeper understanding of scenes. For scene attributes to be useful they need to be machine recognizable. To assess the difficulty of scene attribute recognition we perform experiments using the features and kernels which achieve state of the art category recognition on the 15 scene database and SUN database. Xiao et al. in [23] show that a combination of several scene descriptors results in a significantly more powerful classifier than any individual feature. Accordingly, our SVM classifiers use a combination of kernels generated from gist, HOG 2x2, self-similarity, and geometric context color histogram features (See [23] for feature and kernel details). These four features were chosen because they are each individually powerful and because they can describe distinct visual phenomena.

To recognize attributes in images, we create an individual classifier for each attribute using random splits of the SUN Attribute dataset for training and testing data. Note that our training and test splits are category agnostic – for the purpose of this section we simply have a pool of 14,340 images with varying attributes. We treat an attribute as present if it receives at least two votes and absent if it receives zero votes. As shown in Figure 6, images with a

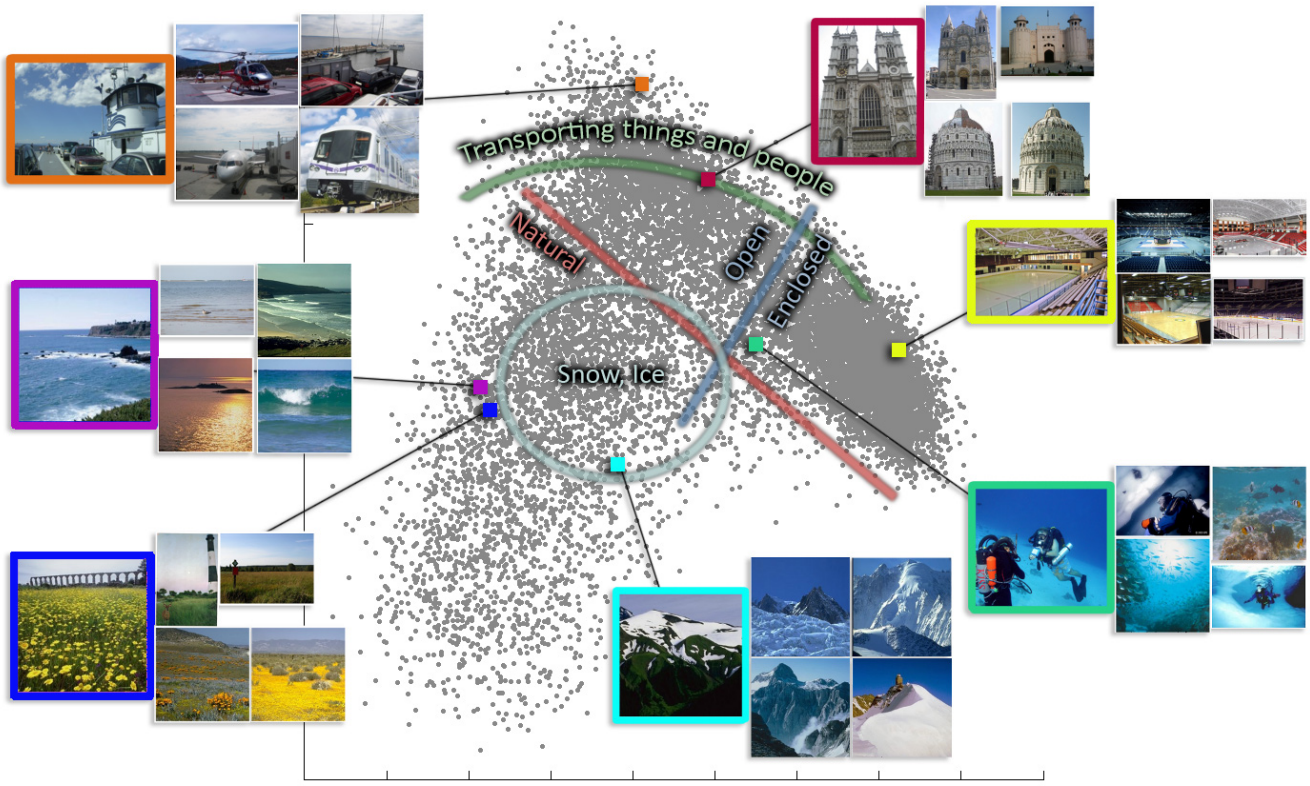


Figure 7: *2D visualization of the SUN Attribute dataset*. Each image in the dataset is represented by the projection of its 102-dimensional attribute feature vector onto two PCA dimensions. There are 7 groups of nearest neighbors, each designated by a color. The centers of the nearest neighbor groups are marked with a square. Interestingly, while the nearest-neighbor scenes in attribute space are semantically very similar, for most of these examples (underwater_ocean, abbey, coast, ice skating rink, field_wild) *none* of the nearest neighbors actually fall in the same SUN database category. The colored border lines deliniate the approximate separation of images with and without the attribute associated with the border. Figure best viewed in color.

single vote tend to be in a transition state between the attribute being present or absent so they are excluded from these experiments. We train and evaluate independent classifiers for each attribute even though correlation between attributes could make "multi-label" classification methods advantageous. For each attribute we wish to recognize we train an SVM with each of our four features. We calculate the average precision (AP) of each classifier and construct a combined kernel from a combination of individual feature kernels. Each kernel is normalized and then weighted in proportion to that feature's AP for a given attribute. Each classifier is trained on 300 images and tested on 50 images and AP is computed over five random splits. Each classifier's train and test sets are half positive and half negative even though most attributes are sparse (i.e. usually absent). We fix the positive to negative ratio so that we can compare the intrinsic difficulty of recognizing each attribute without being influenced by attribute popularity.

Figure 9 shows that the average performance of our classifiers is fairly good (AP .88) and the combined classifier

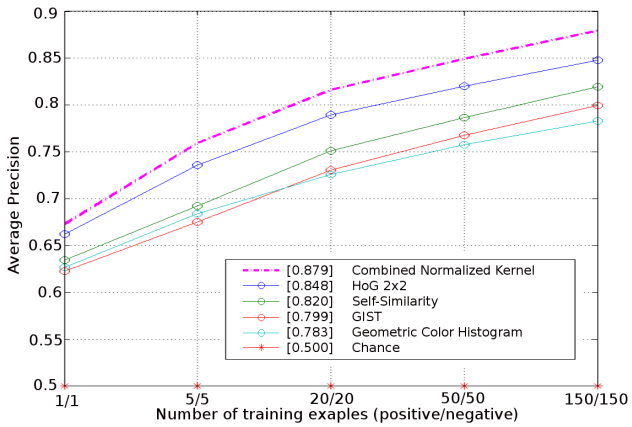


Figure 9: *Average Precision values averaged for all attributes*. The combined feature classifier is more accurate than any individual feature classifier. Average Precision steadily increases with more training data.

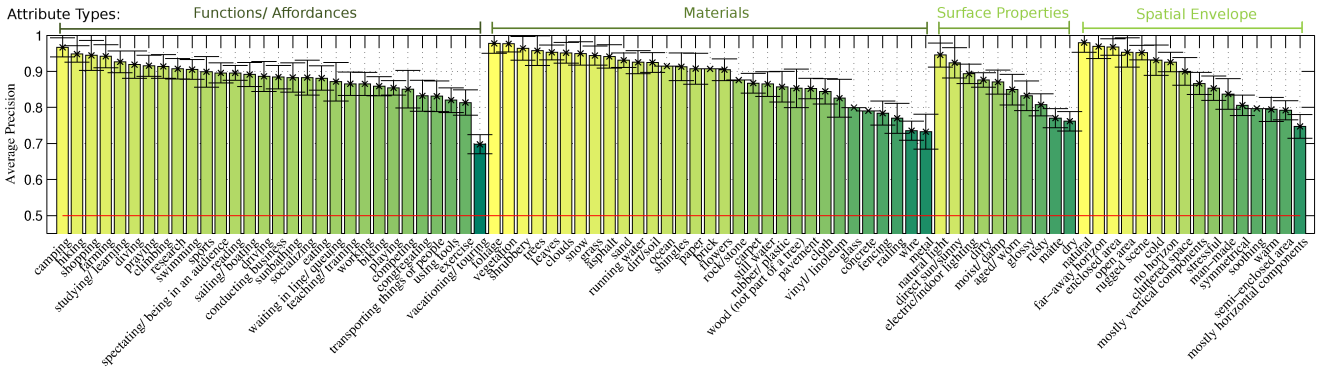


Figure 10: *Average Precision for all of the attributes in our dataset*. The AP of chance selection is marked by the red line. All attributes can be recognized significantly better than chance, even when the visual manifestation of such attributes tends to be quite subtle.

outperforms any individual feature. Not all attributes are equally easy to recognize – Figure 10 plots the average precision for each attribute's combined feature SVM. It is clear from Figure 10 that certain attributes, especially some surface properties and spatial envelope attributes, are particularly difficult to recognize with our global image features.

We show qualitative results of our attribute classifiers in Figure 11. Our attribute classifiers perform well at recognizing attributes in a variety of contexts. Most of the attributes with strong confidence are indeed present in the images. Likewise, the lowest confidence attributes are clearly not present. It is particularly interesting that function/affordance attributes and surface property attributes are often recognized with the stronger confidence than other types of attributes even though functions and surface properties are complex concepts that may not be easy to define visually. For example the golf course test image in Figure 11 shows that our classifiers can successfully identify such abstract concepts as "sports" and "competing" for a golf course, which is visually quite similar to places where no sports would occur. Abstract concepts such as "praying" and "aged/worn" are also recognized correctly in both the abbey and mosque scenes in Figure 11. Figure 12 shows three failure cases.

7. Discussion

Scene attributes are a fertile, unexplored recognition domain. Many attributes are visually quite subtle and nearly all scene descriptors in the literature were developed for the task of *scene categorization* and may not generalize well to attribute recognition. Even though all of our attribute labels are global, many attributes have clear spatial support (materials) while others may not (functions and affordances). Techniques from weakly supervised object recognition might have success at discovering the spatial support of our global attributes where applicable. Multi-label clas-

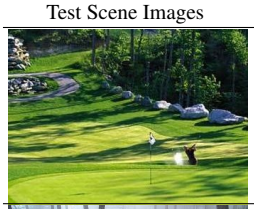
Test Scene Images	Detected Attributes
	*Most Confident Attributes*: vegetation, open area, sunny, sports, natural light, no horizon, foliage, competing, railing, natural *Least Confident Attributes*: studying, gaming, fire, carpet, tiles, smoke, medical, cleaning, sterile, marble
	*Most Confident Attributes*: shrubbery, flowers, camping, rugged scene, hiking, dirt/soil, leaves, natural light, vegetation, rock/stone *Least Confident Attributes*: shingles, ice, railroad, cleaning, marble, sterile, smoke, gaming, tiles, medical
	*Most Confident Attributes*: eating, socializing, waiting in line, cloth, shopping, reading, stressful, congregating, man-made, plastic *Least Confident Attributes*: gaming, running water, tiles, railroad, waves/ surf, building, fire, bathing, ice, smoke
	*Most Confident Attributes*: vertical components, vacationing, natural light, shingles, man-made, praying, symmetrical, semi-enclosed area, aged/ worn, brick *Least Confident Attributes*: railroad, ice, scary, medical, shopping, tiles, cleaning, sterile, digging, gaming
	*Most Confident Attributes*: vertical components, brick, natural light, praying, vacationing, man-made, pavement, sunny, open area, rusty *Least Confident Attributes*: ice, smoke, bathing, marble, vinyl, cleaning, fire, tires, gaming, sterile

Figure 11: *Attribute detection*. For each query, the most confidently recognized attributes (green) are indeed present in the test images, and the least confidently recognized attributes (red) are either the visual opposite of what is in the image or they are irrelevant to the image.

Test Images	Detected Attributes
	*Most Confident Attributes*: swimming, asphalt, open area, sports, sunbathing, natural light, diving, still water, exercise, soothing *Least Confident Attributes*: tiles, smoke, ice, sterile, praying, marble, railroad, cleaning, medical activity, gaming
	*Most Confident Attributes*: cold, concrete, snow, sand, stressful, aged/ worn, dry, climbing, rugged scene, rock/stone *Least Confident Attributes*: medical activity, spectating, marble, cleaning, waves/ surf, railroad, gaming, building, shopping, tiles
	*Most Confident Attributes*: carpet, enclosed area no horizon, electric/indoor lighting, concrete, glossy, cloth, working, dry, rubber/ plastic *Least Confident Attributes*: trees, ocean, digging, open area, scary, smoke, ice, railroad, constructing/ building, waves/ surf

Figure 12: *Failure cases*. In the top image, it seems the smooth, blue regions of the car appear to have created false positive detections of "swimming", "diving", and "still water". The bottom images, unlike all of our training data, is a close-up object view rather than a scene with spatial extent. The attribute classifiers seem to interpret the cat as a mountain landscape and the potato chips bag as several different materials - "carpet", "concrete", "glossy", and "cloth".

sification methods, which exploit the correlation between attributes, might also improve accuracy when recognizing attributes simultaneously. We hope that the scale and variety of our dataset will enable many future explorations in the exciting space of visual attributes.

Acknowledgements. We thank Vazheh Moussavi (Brown Univ.) for his insights and contributions in the data annotation process. Genevieve Patterson is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This work is also funded by NSF CAREER Award 1149853 to James Hays.

References

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

[2] D. Chen and W. Dolan. Building a persistent workforce on mechanical turk for multilingual data collection. In *The 3rd Human Computation Workshop (HCOMP)*, 2011. 4

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[4] K. A. Ehinger, J. Xiao, A. Torralba, and A. Oliva. Estimating scene typicality from human ratings and image features. In *33rd Annual Conference of the Cognitive Science Society*, 2011. 2

[5] I. Endres, A. Farhadi, D. Hoiem, and D. Forsyth. The Benefits and Challenges of Collecting Richer Object Annotations. In *ACVHL 2010 (in conjunction with CVPR).*, 2010. 1

[6] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 1

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1

[8] 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. 2

[9] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1

[10] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009. 1

[11] Lasecki, Murray, White, Miller, and Bigham. Real-time Crowd Control of Existing Interfaces. In *UIST*, 2011. 4

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 1

[13] J. Liu, B. Kuipers, and S. Savarese. Recognizing Human Actions by Attributes. In *CVPR*, 2011. 1

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 2, 3

[15] A. Oliva and A. Torralba. Scene-Centered Description from Spatial Envelope Properties. In *2nd Workshop on Biologically Motivated Computer Vision (BMCV)*, 2002. 2

[16] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *CVPR*, 2011. 2

[17] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011. 1

[18] O. Russakovsky and L. Fei-Fei. Attribute learning in largescale datasets. In *ECCV 2010 Workshop on Parts and Attributes*, 2010. 1

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a Database and Web-based Tool for Image Annotation. *IJCV*, 77(1-3), 2008. 2

[20] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *First IEEE Workshop on Internet Vision at CVPR 08*, 2008. 3

[21] Y. Su, M. Allan, and F. Jurie. Improving Object Classification using Semantic Attributes. In *BMVC*, 2010. 1

[22] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 2008. 2

[23] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2, 3, 4, 5

[24] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *ICCV*, 2011. 1