

The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding

Genevieve Patterson · Chen Xu · Hang Su · James Hays

Received: 27 February 2013 / Accepted: 28 December 2013 / Published online: 18 January 2014
© Springer Science+Business Media New York 2014

Abstract In this paper we present the first large-scale scene attribute database. First, we perform crowdsourced human studies to find a taxonomy of 102 discriminative attributes. We discover attributes related to materials, surface properties, lighting, affordances, and spatial layout. Next, we build the “SUN attribute database” on top of the diverse SUN categorical database. We use crowdsourcing to annotate attributes for 14,340 images from 707 scene categories. We perform numerous experiments to study the interplay between scene attributes and scene categories. We train and evaluate attribute classifiers and then study the feasibility of attributes as an intermediate scene representation for scene classification, zero shot learning, automatic image captioning, semantic image search, and parsing natural images. We show that when used as features for these tasks, low dimensional scene attributes can compete with or improve on the state of the art performance. The experiments suggest that scene attributes are an effective low-dimensional feature for capturing high-level context and semantics in scenes.

Keywords Scene understanding · Crowdsourcing · Attributes · Image captioning · Scene parsing

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 (Oliva and Torralba 2001) or spatial pyramid (Lazebnik et al. 2006), 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 Fig. 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.¹ (3) Images often simultaneously exhibit characteristics of multiple distinct scene categories. (4) A categorical representation cannot generalize to types of scenes which were not seen during training.

An attribute-based representation of scenes would address these problems by expressing variation within a scene category. Scenes would have a multi-variate attribute representation instead of simply a binary category membership. Scene types not seen at training time could also be identified by a canonical set of scene attributes in a zero-shot learning manner.

In the past several years there has been interest in attribute-based representations of objects (Ferrari and Zisserman

G. Patterson (✉) · C. Xu · H. Su · J. Hays
Department of Computer Science, Brown University, 115 Waterman St., Providence, RI 02912, USA
e-mail: gen@cs.brown.edu

¹ Individual attribute presence might be ambiguous in certain scenes, just like category membership can be ambiguous. Scenes only have one category label, though, and the larger the number of categories the more ambiguous the membership is. However, with over one hundred scene attributes in our taxonomy, several attributes may be strongly present, offering a description of that scene that has more context than simply the scene category label. This also enables an attribute-based representation to make finer-grain distinctions about which which components or characteristics of the scene are ambiguous or obvious.



Fig. 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 (Parikh and Grauman 2011b)

2008; Farhadi et al. 2009; Lampert et al. 2009; Farhadi et al. 2010a; Endres et al. 2010; Berg et al. 2010; Russakovsky and Fei-Fei 2010; Su et al. 2010), faces (Kumar et al. 2009), and actions (Yao et al. 2011; Liu et al. 2011b) 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).

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 water or it might not) and inter-class relationships (e.g. both a canyon and a beach could have water). As stated by Ferrari and Zisserman (2008), “recognition of attributes can complement category-level recognition and therefore improve the degree to which machines perceive visual objects”.

A small set of scene attributes were explored in Oliva and Torralba’s seminal ‘gist’ paper (Oliva and Torralba 2001) and follow-up work (Oliva and Torralba 2002). 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 pre-

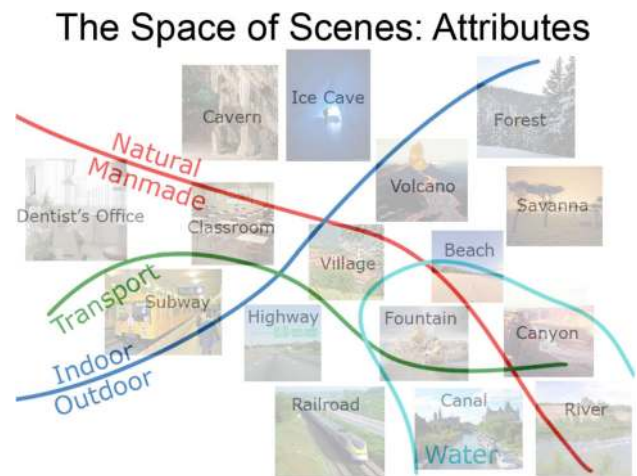


Fig. 2 Hypothetical space of scenes partitioned by attributes rather than categories. In reality, this space is much higher dimensional and there are not clean boundaries between attribute presence and absence

dicted based on the gist representation. Greene and Oliva show that these global scene attributes are predictive of human performance on a rapid basic-level scene categorization task. Greene and Oliva (2009) argue that global attributes of the type we examine here are important for human perception, saying, “rapid categorization of natural scenes may not be mediated primarily through objects and parts, but also through global properties of structure and affordance.”

Russakovsky and Fei-Fei identify the need to discover visual attributes that generalize between categories in Russakovsky and Fei-Fei (2010). Using a subset of the categories from ImageNet, Russakovsky and Fei-Fei show that attributes can both discriminate between unique examples of a category and allow sets of categories to be grouped by common attributes. In Russakovsky and Fei-Fei (2010) attributes were mined from the WordNet definitions of categories. The attribute discovery method described in this paper outlines how attributes can be identified directly by human users. In the end we discover a larger set of attributes, including attributes that would be either too common or too rare to be typically included in the definition of categories.

More recently, Parikh and Grauman (Parikh and Grauman 2011a) 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.

Lampert et al. demonstrate how attributes can be used to classify unseen categories (Lampert et al. 2009). Lampert et al. show that attribute classifiers can be learned independent of category, then later test images can be classified as part of an unseen category with the simple knowledge of the expected attributes of the unseen category. This opens the door for classification of new categories without using training examples to learn those unseen categories. We demonstrate in Sect. 6.1 the performance of our scene attributes for zero-shot learning by classifying test images from all of the categories in our dataset without training classifiers for those scene categories.

1.1 Paper Outline

This paper describes the creation and verification of our SUN attribute database in the spirit of analogous database creation efforts such as ImageNet (Deng et al. 2009), LabelMe (Russell et al. 2008), and Tiny Images (Torralba et al. 2008). First, we derive a taxonomy of more than 100 scene attributes from crowd-sourced experiments (Sect. 2). Next, we use crowd-sourcing to construct our attribute-labeled dataset on top of a significant subset of the SUN database (Xiao et al. 2010) spanning more than 700 categories and 14,000 images (Sect. 3). We visualize the distribution of scenes in attribute space (Sect. 4). The work in these sections largely appears in a previous publication (Patterson and Hays 2012). Section 5 contains significantly expanded work, and Sects. 6 and 7 are previously unpublished novel experiments.

In order to use scene attributes for vision tasks, we train and test classifiers for predicting attributes (Sect. 5). We demonstrate the output of these classifiers on novel images. Furthermore, in Sect. 6 we explore the use of scene attributes for scene classification and the zeroshot learning of scene categories. We compare how scene classifiers derived using scene attributes confuse scene categories to how human respondents confuse categories.

The final section of the paper experiments with using scene attributes for challenging scene understanding tasks. We use attributes as features in the pipelines for the tasks of scene parsing (Sect. 7.1) and automatic image captioning (Sect. 7.2). We also investigate image retrieval with image descriptors derived from attributes (Sect. 7.3).

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 interested in find-

ing 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. None-the-less, 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. There is one probe image for every category, selected for its canonical appearance. We want a set of images which is maximally diverse and representative of the space of scenes. For this reason the probe images are the images which human participants found to be most typical of 707 SUN dataset categories (Ehinger et al. 2011).

We first ask AMT workers to provide text descriptions of the individual probe images. 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 Oliva and Torralba (2001). 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. The attribute gathering task was repeated over 6000 times. From the thousands of raw discriminative attributes reported by participants we manually 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

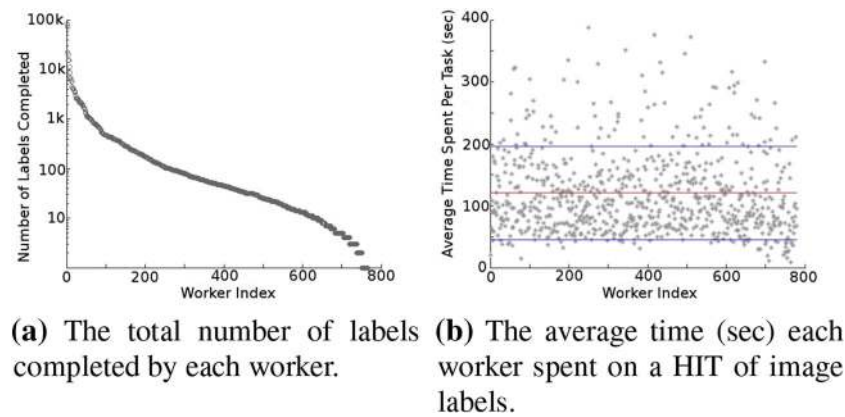


Fig. 6 These plots visualize our criteria for identifying suspicious workers to grade. **a** shows the heavy-tailed distribution of worker contributions to the database. The top workers spent hundreds of hours on our HITs. The *red line* in plot **b** demarcates the average work time

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 (Color figure online)

3.2 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 or less than one standard deviation away from the average for all workers is added to a list of workers to manually review. This way workers who are randomly labeling images and workers who may have been confused by the task are both caught. Workers who clicked images randomly finished faster than workers who considered each image on the HIT. We review by hand a 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.

3.3 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 [Chen and Dolan \(2011\)](#). In general, worker accuracy rose over time and we omit over one

million early annotations from the final dataset. Worker accuracy improved over time as the workers who did not follow instructions were culled from the pool of workers who were offered the opportunity to complete HITs.

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. We repeat the labeling process in order to obtain consensus as the presence of some of the scene attributes may be a subjective decision. No worker is allowed to label the same image for the same attribute more than once. The idea of finding and heavily utilizing good workers is in contrast to the “wisdom of the crowds” crowdsourcing strategy where consensus outweighs expertise. Our choice to utilize only workers who give higher quality labels is supported by recent research such as [Lasecki et al. 2011](#) where good workers were shown to be faster and more accurate than the average of many workers. Figure 6 shows the contributions of all workers to our database.

Figure 7 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. The population of labels in the dataset is not even (8%/92% positive/negative). 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’).

We graded equal numbers of positive and negative labels to understand if there was a disparity in accuracy between them. For both types of annotation, we find $\sim 93\%$ of labels to be reasonable, which means that we as experts would agree with the annotation.





















Attribute	Images given 0 votes	Images given 1 vote	Images given 2 votes	Images given 3 votes
Camping				
Diving				
Medical Activity				
Cluttered Space				
Fire				

Fig. 7 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. AMT workers are instructed to positively label an image if the functional attribute is likely

to occur in that image, not just if it is actually occurring. For material, surface property, or spatial envelope attributes, workers were instructed to positively label images only if the attribute is present

In the following sections, our experiments rely on the consensus of multiple annotators rather than individual annotations. This increases the accuracy of our labels. For each of our 102 attributes, we manually grade 5 scenes where the consensus was positive (2 or 3 votes) and likewise for negative (0 votes). We find that if 2 out of 3 annotations agree on a positive label, that label is reasonable $\sim 95\%$ of the time. Many attributes are very rare, and there would be a significant loss in the population of the rare attributes if consensus was defined as 3/3 positive labels. Allowing for 2/3 positive labels to be the consensus standard increases the population of rare attributes without degrading the quality of the 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 Fig. 8 we show all 14,340 of our scenes projected onto two dimensions by t-SNE dimensionality reduction (Van der Maaten and Hinton 2008). 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.

To better understand where images with different attributes live in attribute space, Fig. 9 illustrates where dataset images that contain different attributes live in this 2D version of the attribute feature space.

Figure 10 shows the distribution of images from 15 scene categories in attribute space. The particular scene categories were chosen to be close to those categories in the 15 scene benchmark (Lazebnik et al. 2006). In this low dimensional visualization, many of the categories have considerable overlap (e.g. bedroom with living room, street with highway, city with skyscraper). This is reasonable because these overlapping categories share affordances, materials, and layouts. With the full 102 dimensional attribute representation, these scenes could still be differentiated and we examine this task in Sect. 6.

5 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 SUN database. In Xiao et al. (2010) show that a combination of several scene descrip-

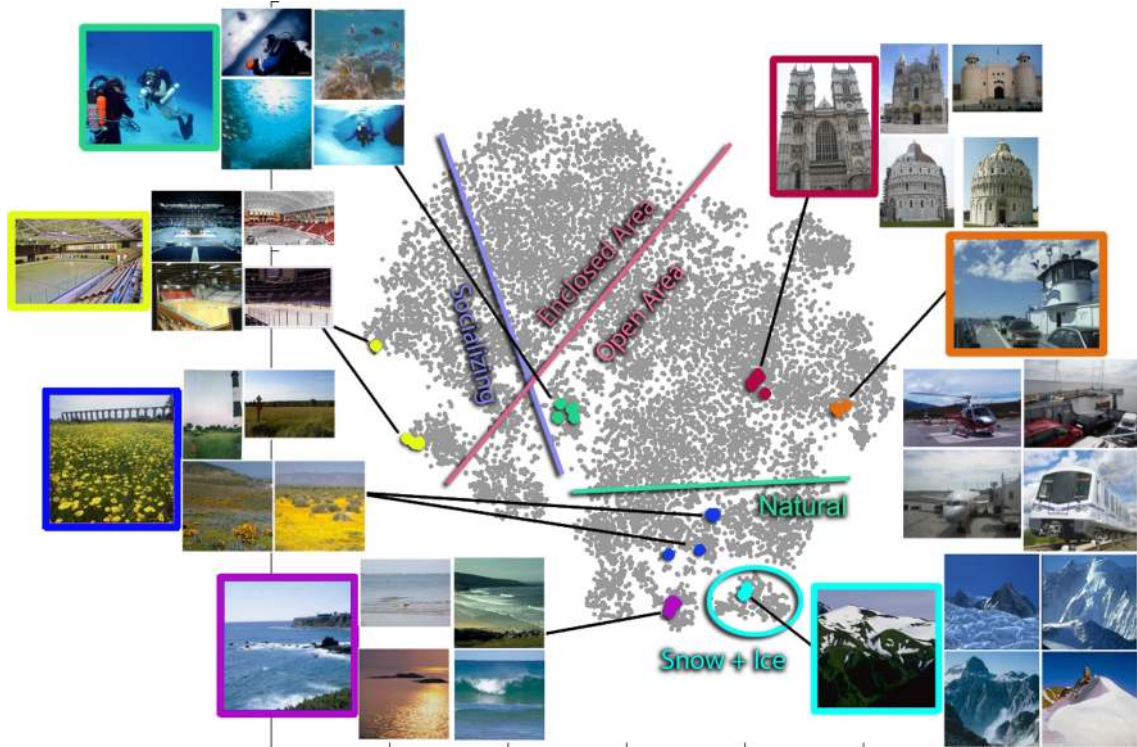


Fig. 8 2D visualization of the SUN Attribute dataset. Each image in the dataset is represented by the projection of its 102-dimensional attribute feature vector onto two dimensions using t-Distributed Stochastic Neighbor Embedding (Van der Maaten and Hinton 2008). There are groups of nearest neighbors, each designated by a color. Interestingly, while the nearest-neighbor scenes in attribute space are seman-

tically very similar, for most of these examples (underwater_ocean, abbey, coast, ice skating rink, field_wild, bistro, office) none of the nearest neighbors actually fall in the same SUN database category. The colored border lines delineate the approximate separation of images with and without the attribute associated with the border. Figure best viewed in color (Color figure online)

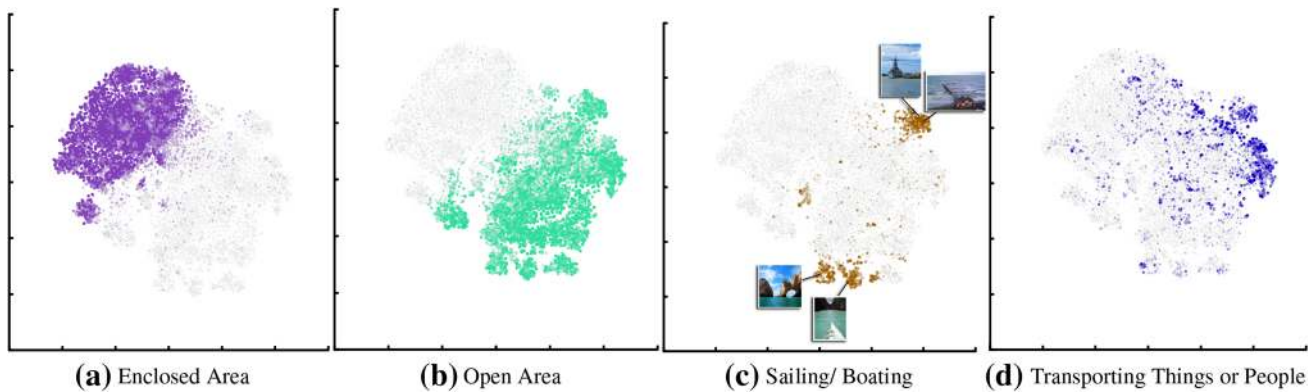


Fig. 9 Distributions of scenes with the given attribute. This set of reduced dimensionality plots highlights the populations of images with the listed attributes. Grey points are images that do not contain the given attribute. The boldness of the colored points is proportional to the amount of votes given for that attribute in an image, e.g. darkest colored points have 3 votes. ‘Enclosed area’ and ‘open area’ seem to

have a strong effect on the layout of scenes in “attribute space”. As one might hope, they generally occupy mutual exclusive areas. It is interesting to note that ‘sailing/boating’ occurs in two distinct regions which correspond to open water scenes and harbor scenes (Color figure online)

tors results in a significantly more powerful classifier than any individual feature. Accordingly, our classifiers use a combination of kernels generated from gist, HOG 2×2 , self-similarity, and geometric context color histogram features

[see (Xiao et al. 2010) for feature and kernel details]. These four features were chosen because they are each individually powerful and because they can describe distinct visual phenomena.

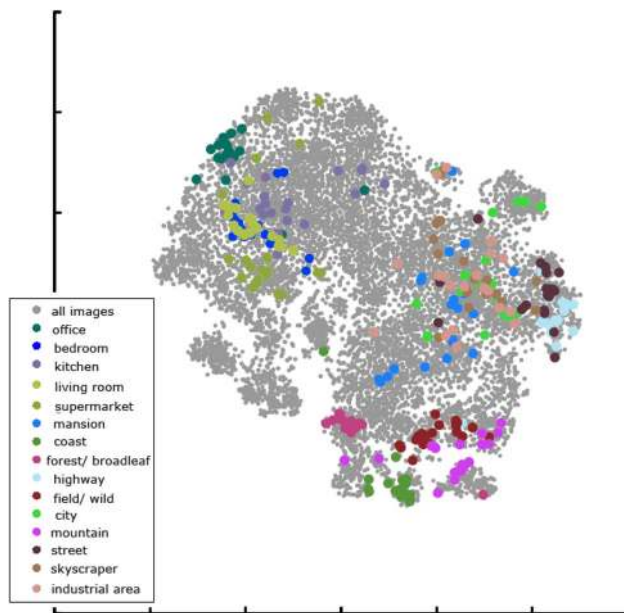


Fig. 10 Location of member images of 15 scene categories in attribute space. Figure best viewed in color (Color figure online)

5.1 How Hard is it to Recognize Attributes?

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 scene 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, i.e. consensus is established, and absent if it receives zero votes. As shown in Fig. 7, images with a 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. Correlation between attributes could make ‘multi-label’ classification methods advantageous, but we choose to predict attributes independently. For each attribute we wish to recognize, we first evaluate SVM classifiers trained independently on each of our four features. We report their performance in Fig. 11. To train a classifier which uses all features, we construct a combined kernel from a linear combination of individual feature kernels. 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 11 shows that the combined classifier outperforms any individual feature. Not all attributes are equally easy

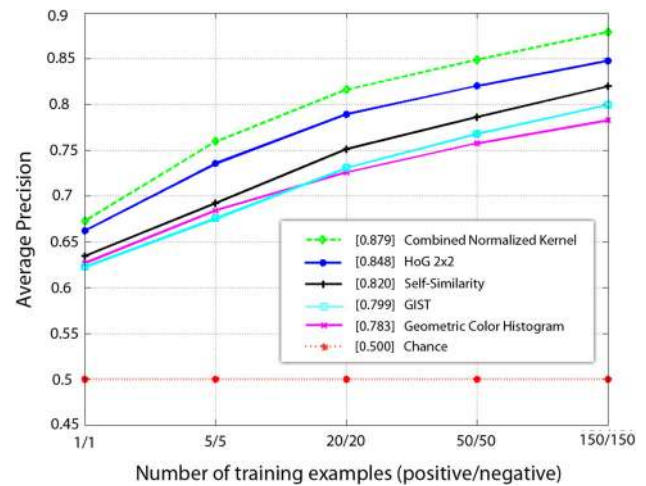


Fig. 11 Average Precision values averaged for all attributes. The combined feature classifier is more accurate than any individual feature classifier. Average Precision steadily increases with more training data

to recognize Fig. 12a plots the average precision for each attribute’s combined feature SVM. It is clear from Fig. 12a that certain attributes, especially some surface properties and spatial envelope attributes, are particularly difficult to recognize with our global image features.

Figure 12a evaluates attribute recognition with fixed proportions of positive and negative examples. However, some attributes are vastly more popular than others in the real world. To evaluate attribute recognition under more realistic conditions, and to make use of as much training data as the SUN attribute database affords us, we train classifiers on 90 % of the dataset and test on the remaining 10 %. This means that some attributes (e.g. ‘natural’ will have thousands of positive examples, and others e.g. ‘smoke’ will have barely 100). Likewise, chance is different for each attribute because the test sets are similarly skewed. The train and test instances for each attribute vary slightly because some images have confident labels for certain attributes and ambiguous labels for others and again we only use scenes with confident ground truth labels for each particular attribute classifier. Figure 12b shows the AP scores for these large scale classifiers.

For these classifiers, we averaged the kernels from each feature. With the larger training set, there was no observed benefit to weighting individual feature kernels differently. Figure 12b demonstrates how more popular attributes are easier to recognize, as expected. Overall, the average AP scores for different types of attributes are similar—functions/affordances (AP 0.44), materials (AP 0.51), surface properties (AP 0.50), and spatial envelope (AP 0.62).

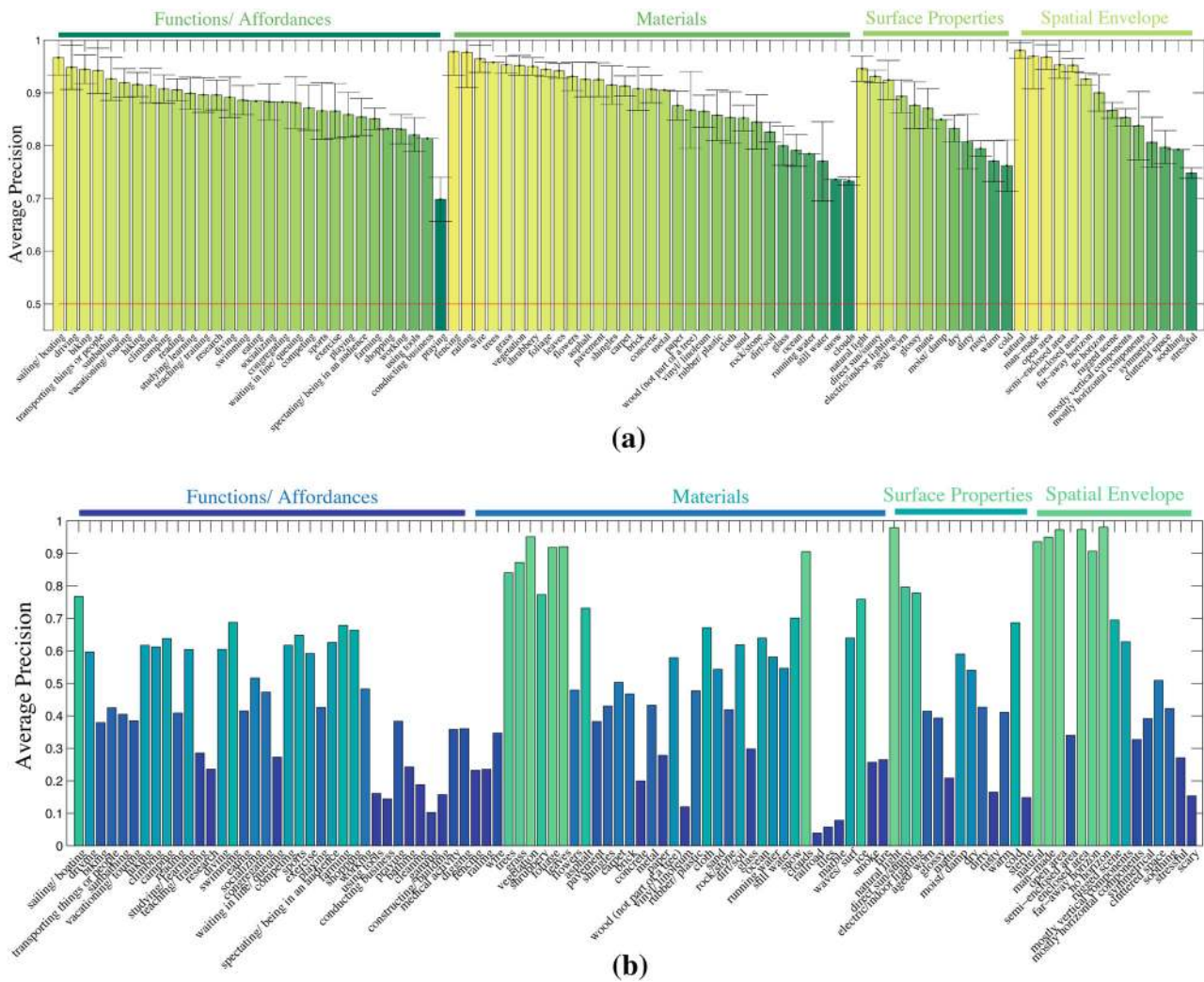


Fig. 12 **a** 300 training/50 test examples; training and testing sets have a balance positive to negative example ratio. The AP of chance selection is marked by the red line. AP scores are often high even when the visual manifestation of such attributes are subtle. This plot show that it is possible to recognize global scene attributes. Attributes that occur fewer than 350 times in the dataset were not included in this plot. **b** 90 % of the

dataset for training/10 % for test. All of the scene attributes are included in this plot. Chance is different for every attribute as they appear with variable frequency in nature. Note that the most difficult to recognize attributes are also the rarest. Many attributes that are not strongly visual such as ‘studying’, ‘spectating’, or ‘farming’ are nonetheless relatively easy to recognize. Average precision for attribute classifiers

The classifiers used for Fig. 12b and the code used to generate them are publicly available.³ The attribute classifiers trained on 90 % of the SUN Attribute dataset are employed in all further experiments in this paper.

5.2 Attribute Classifiers in the Wild

We show qualitative results of our attribute classifiers in Fig. 13. 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.

³ SUN Attribute Classifiers along with the full SUN Attribute dataset and associated code are available at www.cs.brown.edu/~gen/sunattributes.html.

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 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 Fig. 13 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 Fig. 13. Figure 14 shows several cases where the most confidently detected attributes are incorrect.





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

Fig. 13 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 (Color figure online)

In earlier attribute work where the attributes were discovered on smaller datasets, attributes had the problem of being strongly correlated with each other (Farhadi et al. 2009). This is less of an issue with the SUN Attribute dataset because the dataset is larger and attributes are observed in many different contexts. For instance, attributes such as “golf” and “grass” are correlated with each other, as they should be. But the correlation is not so high that a “golf” classifier can simply learn the “grass” visual concept, because the dataset contains thousands of training examples where “grass” is present but “golf” is not possible. However, some of our attributes, specifically those related to vegetation, do seem overly correlated with each other because the concepts are not semantically distinct enough.

Figure 15 shows the most confident classifications in our test set for various attributes. Many of the false positives,

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

Fig. 14 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’

highlighted in red, are reasonable from a visual similarity point of view. ‘Cold’, ‘moist/damp’, and ‘eating’ all have false positives that could be reasonably considered to be confusing. ‘Stressful’ and ‘vacationing’ have false positives that could be subjectively judged to be correct—a crowded subway car could be stressful, and the New Mexico desert could be a lovely vacation spot.

5.3 Correlation of Attributes and Scene Categories

To better understand the relationships between categories and attributes, Table 1 lists a number of examples from the SUN 397 categories with the attribute that is most strongly correlated with each category.

The correlation between the scene category and the attribute feature of an input image is calculated using Pearson’s correlation. We calculate correlation between the predicted attribute feature vectors for 50 examples from each of the SUN 397 categories and a feature vectors that indicate the category membership of the example images.



Fig. 15 Top 5 most confident detections in test set. For each attribute, the top five detections from the test set are shown. Images boxed in *green* are true positives, and *red* are false positives. Examples of false positives, such as the ‘praying’ examples, show how attributes are identified in images that arguably contain the attribute, but human annotators disagreed about the attribute’s presence; in this case the false positives were a sacristy, which is a room for the storage of religious items, and a cathedral pictured at a distance. The false positive for ‘glass’ also contain glass, although photographed under glancing illumination, which may have caused the human annotators to mislabel it. For several of the examples, all of the top 5 detections are true positives. The detections for ‘brick’, ‘metal’, and ‘competing’ demonstrate the ability of attribute classifiers to recognize the presence of attributes in scenes that are quite visually dissimilar. For ‘brick’ and ‘metal’ even the kinds of bricks and metals shown are differ greatly in type, age, and use case. Figure best viewed in color (Color figure online)

Table 1 has many interesting examples where an attribute is strongly correlated with visually dissimilar but semantically related categories, such as ‘praying’ for both the indoor and outdoor church categories. Even attributes that are quite abstract concepts, such as ‘socializing’ and ‘stressful’, are the most strongly correlated attributes for ‘pub/indoor’ and ‘cockpit’, respectfully. Scene attributes capture information that is intrinsic to the nature of scenes and how humans interact with them.

6 Predicting Scene Categories from Attributes

In this section we measure how well we can predict scene category from 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.

Attributes allow for the exploration of scenes using information that is complementary to the category labels of those scenes. Because attributes are powerful descriptors of scenes, they can also be used as a feature to predict scene categories. To the best of our knowledge these experiments are the first to explore the use of attributes as features for scene classification. As with objects (Lampert et al. 2009), attributes also offer the opportunity to learn new scene categories without using any training examples for the new categories. This “zero-shot” learning for scenes will also be explored.

We evaluate the task of classifying all 397 categories of the SUN 397 dataset (Xiao et al. 2010) using our 102 attribute classifiers as an intermediate representation. We compare this to scene recognition using recent low-level features. We also compare to classifiers trained with ground-truth attributes to derive a scene classification upper bound for our attribute taxonomy. Finally, we evaluate zero-shot learning scenarios where an oracle provides attribute distributions for all categories and we use our classifiers to estimate attributes for query instances.

One hundred binary attributes alone could potentially predict membership in 397 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.

Table 1 Most correlated attributes

Category	Most corr. attribute	Pearson's corr. coeff.
Airport terminal	Socializing	0.051
Art studio	Cluttered space	0.039
Assembly line	Working	0.055
Athletic field/outdoor	Playing	0.116
Auditorium	Spectating	0.096
Ball pit	Rubber/plastic	0.149
Baseball field	Sports	0.088
Basilica	Praying	0.101
Basketball court/outdoor	Exercise	0.074
Bathroom	Cleaning	0.092
Bayou	Still water	0.092
Bedroom	Carpet	0.054
Biology laboratory	Research	0.053
Bistro/indoor	Eating	0.055
Bookstore	Shopping	0.079
Bowling alley	Competing	0.055
Boxing ring	Spectating	0.049
Campsite	Camping	0.053
Canal/natural	Still water	0.080
Canal/urban	Sailing/boating	0.038
Canyon	Rugged scene	0.110
Car interior/backseat	Matte	0.079
Car interior/frontseat	Matte	0.098
Casino/indoor	Gaming	0.070
Catacomb	Digging	0.081
Chemistry lab	Research	0.067
Chicken coop/indoor	Dirty	0.039
Chicken coop/outdoor	Fencing	0.045
Cathedral/indoor	Praying	0.148
Church/outdoor	Praying	0.088
Classroom	Studying/learning	0.070
Clothing store	Cloth	0.063
Cockpit	Stressful	0.048
Construction site	Constructing/building	0.041
Corn field	Farming	0.111
Cottage garden	Flowers	0.106
Dentists office	Medical activity	0.070
Dining room	Eating	0.064
Electrical substation	Wire	0.054
Factory/indoor	Working	0.047
Fastfood restaurant	Waiting in line	0.057
Fire escape	Railing	0.051
Forest path	Hiking	0.111
Forest road	Foliage	0.095
Fountain	Running water	0.041
Ice skating rink/indoor	Sports	0.058
Ice skating rink/outdoor	Cold	0.065

Table 1 continued

Category	Most corr. attribute	Pearson's corr. coeff.
Iceberg	Ocean	0.148
Lecture room	Studying/learning	0.080
Mosque/indoor	Cloth	0.060
Mosque/outdoor	Praying	0.066
Operating room	Sterile	0.058
Palace	Vacationing	0.045
Poolroom/establishment	Gaming	0.068
Poolroom/home	Gaming	0.075
Power plant/outdoor	Smoke	0.074
Pub/indoor	Socializing	0.065
Restaurant	Eating	0.088
Restaurant kitchen	Working	0.058
Stadium/football	Spectating	0.132
Subway station/platform	Railroad	0.052
Underwater/coral reef	Diving	0.165
Volcano	Fire	0.122
Wheat field	Farming	0.133

A sampling of scene categories from the SUN 397 dataset listed with their most correlated attribute

6.1 Scene Classification

6.1.1 Attributes as Features for Scene Classification

Although our attributes were discovered in order to understand natural scenes more deeply than by simply knowing their scene categories, scene classification remains a challenging and interesting task. As a scene classification baseline, we train one-vs-all non-linear SVMs with the same low level features used to predict attributes. Figure 16 compares this with various classifiers which instead operate on attributes as an intermediate representation.

The simplest way to use scene attributes as an intermediate representation is to run our attribute classifiers on the scene classification training instances and train one-vs-all SVMs in the resulting 102 dimensional space. This “predicted attribute feature” performs better than three of the low-level features, but worse than the HoG 2×2 feature.⁴

It is important to note that the low-level features live in spaces that may have thousands of dimensions, while the

⁴ The images in the SUN Attribute dataset were originally taken from the whole SUN dataset, which includes more than 900 scene categories. Thus, some portion of the SUN Attribute images also appear in the SUN 397 dataset, which is also a subset of the full SUN dataset. The scene classifiers using low-level and predicted attribute features were trained and tested on the SUN397 dataset minus any overlapping images from the SUN Attribute dataset to avoid testing scene classification on the same images used to train attribute classifiers.

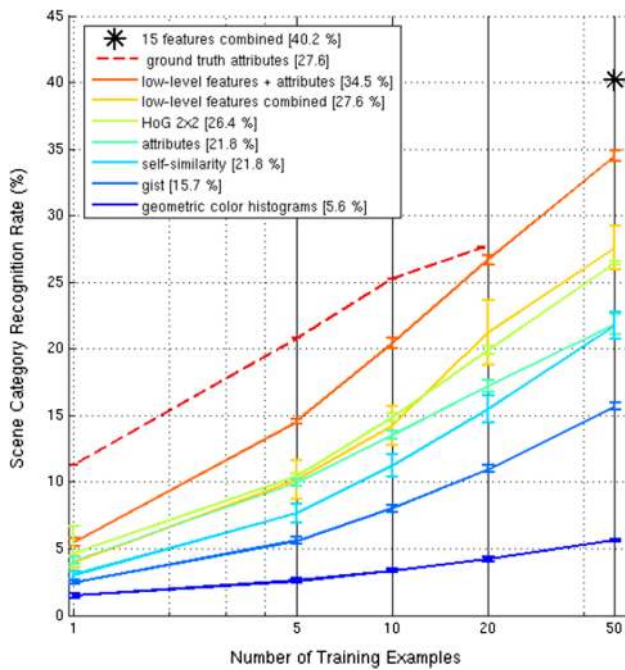


Fig. 16 Scene category recognition rate versus number of training examples. Classification tested on the SUN 397 dataset (Xiao et al. 2010). Images that occur in both the SUN 397 and SUN attribute datasets were omitted from the training and test sets of the above classifiers. Each trend line plots the scene classification accuracy of the associated feature. All predicted features use the same test/train sets, and results averaged over several random test/train splits. When combined with the 4 low-level features originally used in the attribute classifiers, the ‘attributes’ feature clearly improves performance over a scene classifier that only uses low-level features. This further supports our claim that attributes are encoding important contextual knowledge. Classification accuracy using 15 different low-level features (the same features used in Xiao et al.) plus attribute features at 50 training examples is 40.22 %, slightly beating the 38.0 % accuracy reported in Xiao et al. (2010). The ground truth attribute feature is trained and tested on 10 random splits of the SUN Attribute dataset. Thus the number of test examples available for the ground truth feature are $(20 - n_{train})$, where n_{train} is the number of training examples. As the number of training examples increases, the ground truth feature trend line is less representative of actual performance as the test set is increasingly small. Using ground truth attributes as a feature gives an upper bound on what attribute features could possibly contribute to scene classification. Figure best viewed in color (Color figure online)

attribute feature is only 102-dimensional. Partly for this reason, the attribute-based scene classifier seems to benefit less from additional training data than the low level features. This makes sense, because lower dimensional features have limited expressive capacity and because the attribute distribution for a given category isn’t expected to be especially complex (this is, in fact, a motivation for zero-shot learning or easy knowledge transfer between observed and unobserved categories).

The performance of a scene classifier that uses 15 canonical low-level features plus attributes is 40.22 %. The 15 features used were HoG 2×2 , geometric texon histograms,

self-similarity measure, dense SIFT, local binary patterns, texon histograms, gist, the first nearest neighbor, LBP HF feature, sparse SIFT histograms, geometric color histograms, color histograms, geometric classification map, straight line histograms, and tiny image feature Xiao et al. (2010). This improves on the 38 % highest accuracy reported in Xiao et al., which uses these 15 features combined without attributes. Scene classification with attributes falls short of the more recent features suggested by Sanchez et al. which achieve 47 % average accuracy (Sanchez et al. 2013). The performances of scene classifiers trained on each low-level feature and attributes separately are shown in Fig. 16.

It is also important to remember that attribute classification itself is a difficult task. If it were possible to perfectly predict attributes, scene classification performance would jump dramatically. We estimate an upper bound for scene classification with our attribute taxonomy by training and testing on the ground truth attribute annotations for each scene category. As shown in Fig. 16, such a classifier outperforms the best low-level feature by a huge margin 25 versus 15% accuracy with 10 training examples per category.⁵

6.1.2 Learning to Recognize Scenes Without Visual Examples

In zero-shot learning, a classifier is presented (by some oracle) a ground truth distribution of attributes for a given category rather than any visual examples. Test images are classified as the category whose oracle-annotated feature vector is the nearest neighbor in feature space to the test images’ features.

Canonical definitions of zero-shot learning use an intermediate feature space to generalize important concepts shared by categories (Lampert et al. 2009; Palatucci et al. 2009). Lampert et al. uses an attribute representation to enable knowledge transfer between seen and unseen categories, and Palatucci et al. uses phonemes. In these zero-shot learning scenarios, it is prohibitively difficult or expensive to collect low-level feature examples of an exhaustive set of categories. The use of oracle features for those unseen categories is a way to identify them without collecting enough examples to train a classifier.

The goal of zero-shot learning is to learn a classifier $f : \mathcal{X} \rightarrow \mathcal{Z}$ for a label set \mathcal{Z} , where some categories in \mathcal{Z} were not seen during training. This is accomplished by learning two transfer functions, $g : \mathcal{X} \rightarrow \mathcal{A}$ and $h : \mathcal{A} \rightarrow \mathcal{Z}$. The set \mathcal{A} is an intermediate feature space like attributes or phonemes. Some oracle provides the labels for the unseen categories in \mathcal{Z} using the feature space of \mathcal{A} . In traditional

⁵ Because ground truth attributes were collected on the SUN Attribute set of images, the classifiers using the ground truth attributes directly as features were trained and tested on the SUN Attribute dataset.

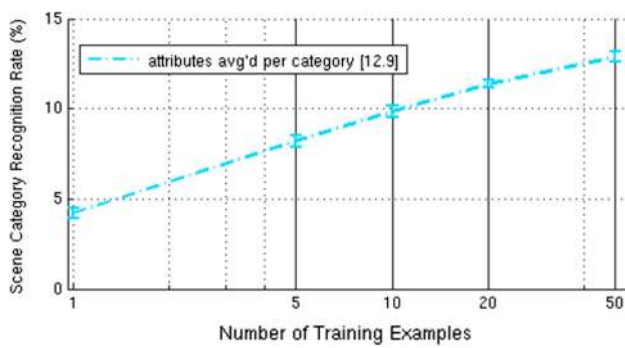


Fig. 17 Scene category recognition without visual examples. The ‘attributes averaged per category’ feature is calculated by averaging the predicted attribute features of all of the training instances of a given scene category in the SUN 397 dataset. Test instances are evaluated by selecting the nearest neighbor scene category feature, and taking that scene category’s label

zero-shot learning experiments, instances from the unseen categories in \mathcal{Z} are not used to learn the transfer function $g : \mathcal{X} \rightarrow \mathcal{A}$. This makes sense if obtaining examples of the unseen categories is difficult as in Lampert et al. (2009), Palatucci et al. (2009).

Because we already had a nearly exhaustive set of scene categories in the SUN Attribute dataset, the attribute classifiers were trained using images that belonged to categories that were held out during the “zero-shot” testing of the transfer function $h : \mathcal{A} \rightarrow \mathcal{Z}$. In our “zero-shot” experiment, all of the possible scene category labels in \mathcal{Z} were held out. The experiments conducted using scene attributes as features in this subsection are an expanded version of traditional zero-shot learning, and we have maintained that term to support the demonstration of how a scene category can be identified by its typical attributes only, without any visual examples of the category. The entire “zero-shot” classification pipeline in this section never involved showing the classifier a visual training example of any scene category. The classifier gets an oracle feature listing the typical attributes of each of the 397 categories.

Our goal is to show that given some reasonable estimate of a scene’s attributes it is possible to estimate the scene category without using the low-level features to classify the query image. Scene attributes are correlated with scene categories, and query scenes can be successfully classified if only their attributes are known. In this sense our experiment is similar to, but more stringent than canonical knowledge transfer experiments such as in Rohrbach et al. because the scene category labels were not used to help learn the mapping from pixel-features to attributes (Rohrbach et al. 2011).

Despite the low number of training examples (397, one oracle feature per category, for zero-shot features vs. $n \times 397$ for pixel-level features), the zero-shot classifier shown in Fig. 17 performs about as well as the gist descriptor. It does, however, perform significantly worse than the attribute-

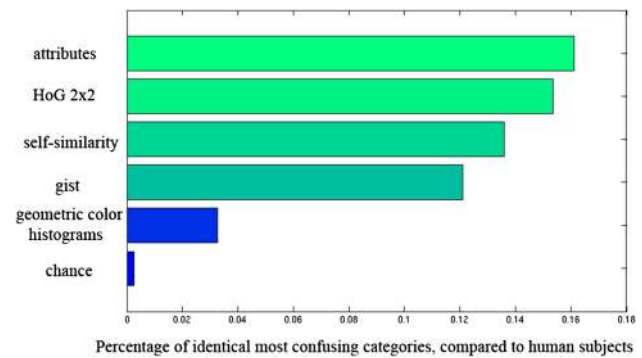


Fig. 18 Comparison to human confusions. Using the human scene classification confusions from Xiao et al. (2010), we report how often the large incorrect (i.e. off-diagonal) confusion is the same for a given feature and the human classifiers

based classifier trained on n examples of predicted attributes shown in Fig. 16. Averaging the attributes into a single “characteristic attribute vector” for each category is quite lossy. In some ways, this supports the argument that there is significant and interesting intra-category variation of scene attributes.

6.2 Predicting Human Confusions

Scene classification is a challenging task, even for humans. In the previous sections, we show that attributes do not always out-perform low-level features at scene classification. Figure 18 shows the performance of several features at another challenging task—predicting human confusions for scene classification on the SUN 397 dataset. At this task, attributes perform slightly better than any other feature.

We compare the confusions between features and humans using the scene classification confusion matrices for each feature. The human classification confusion for the SUN 397 dataset is reported in Xiao et al. (2010). We determined that a feature classifier and the humans had the same confusion if the largest off-diagonal elements of the corresponding rows of their confusion matrices were the same, e.g. both the attribute classifier and the human respondents confused ‘bayou’ for ‘swamp’.

In Xiao et al. (2010), the low-level features that performed the best for scene classification also performed the best at predicting human confusions. Here we demonstrate that although predicted attributes do not perform as well as HoG 2×2 features at scene classification, they are indeed better at predicting human confusions.

This result supports the conclusions of Greene and Oliva (2009). Attributes, which capture global image concepts like structure and affordance, may be closer to the representations humans use to do rapid categorization of scenes than low-level image features by themselves.

7 Image Retrieval Applications with Attribute-Based Representations

Thus far we have introduced the SUN Attribute dataset and we have measured how well attributes can be recognized based on typical global image features. We have also used predicted attributes as an intermediate representation for scene classification and zero shot learning. An appealing property of attributes as an intermediate, low-dimensional representation is that distances in attribute space tend to be more meaningful than distances in the high-dimensional, low-level feature spaces from which those attributes were predicted. For example, with the SUN 397 database, the best combination of features gets 38 % scene classification accuracy with a non-linear SVM, but those same features get only 13 % accuracy with a nearest neighbor classifier. Strong supervision is required to harness global image descriptors for recognition tasks and this is a problem for the numerous highly data-driven recognition tasks which rely on unsupervised image retrieval (or “scene matching”) is at the start of the pipeline. In this section we demonstrate three tasks in which we have replaced (or augmented) typical global image features with attribute-based representations (1) scene parsing (2) image captioning and (3) text-based image retrieval. In all of these applications, the attribute representation for image retrieval is the 102 dimensional real-valued confidences from the classifiers in Sect. 5.

7.1 Scene Parsing with Attribute-Based Scene Retrieval

Scene parsing is the task of segmenting and recognizing all objects and surfaces in an image. Categorical labels can be assigned to either each pixel or each region (e.g. superpixel) of the input image, giving a thorough interpretation of the scene content. Most methods proposed for this problem require a generative or discriminative model to be trained for each category, and thus only work with a handful of pre-defined categories (Gould et al. 2009; He et al. 2004; Hoiem et al. 2007; Ladický et al. 2010; Malisiewicz and Efros 2008; Rabinovich et al. 2007; Shotton et al. 2008, 2006; Socher et al. 2011). The training process can be very time-consuming and must be done in advance. Even worse, the entire training has to be repeated whenever new training images or class labels are added to the dataset. Recently, several nonparametric, data-driven approaches have been proposed for the scene parsing problem (Liu et al. 2011; Tighe and Lazebnik 2013; Eigen and Fergus 2012). These approaches require no training in advance. They can easily scale to hundreds of categories and have the potential to work with Internet-scale, continuously growing datasets like LabelMe (Russell et al. 2008).

In this section we show how well we can improve data-driven scene parsing by adopting scene attributes. Tighe

and Lazebnik investigate nonparametric, data-driven scene parsing and achieve state-of-the-art performance (Tighe and Lazebnik 2013). We follow their system pipeline and show that by simply adding scene attributes as one of the features used for scene representation we can achieve significant performance gains.

7.1.1 System Pipeline

The first step in parsing a query image is to find a retrieval set of images similar to the query image. The purpose of finding this subset of training images is to expedite the parsing system and at the same time throw away irrelevant information which otherwise can be confusing. In Tighe and Lazebnik (2013), three types of global image features are used in this step: gist, spatial pyramid, and color histogram. For each feature type, Tighe and Lazebnik sort all the training images in increasing order of Euclidean distance from the query image. They take the minimum rank across all feature types for each training image and then sort the minimum ranks in increasing order to get a ranking among the training images for the query image. The top ranking K images are used as the retrieval set.

After building the retrieval set, the query image and the retrieval set images are segmented into superpixels. Each superpixel is then described using 20 different features. A detailed list of these features can be found in Table 1 in Tighe and Lazebnik (2013). For each superpixel in the query image, nearest-neighbor superpixels in the retrieval set are found according to the 20 features for that superpixel. A likelihood score is then computed for each class based on the nearest-neighbor matches.

In the last step, we can simply assign the class with the highest likelihood score to each superpixel in the query image, or use Markov Random Field (MRF) framework to further incorporate pairwise co-occurrence information learned from training dataset. As in Eigen and Fergus (2012), we report the performance without using the MRF layer in this paper so differences in local classification performance can be observed more clearly.

7.1.2 Scene Attributes as Global Features

Our goal in investigating scene parsing is to characterize how well our scene attributes work as a scene representation for image retrieval. Thus, we keep the system in Tighe and Lazebnik (2013) unchanged except for using scene attributes as global image features, either by themselves or in conjunction with the low-level features used in the original system.

The dataset we use for this experiment is the SIFT-Flow dataset (Liu et al. 2011). It is composed of 2,688 annotated images from LabelMe and has 33 semantic labels. Since the class frequencies are highly unbalanced, we report both per-pixel classification rate and per-class rate, which is the aver-

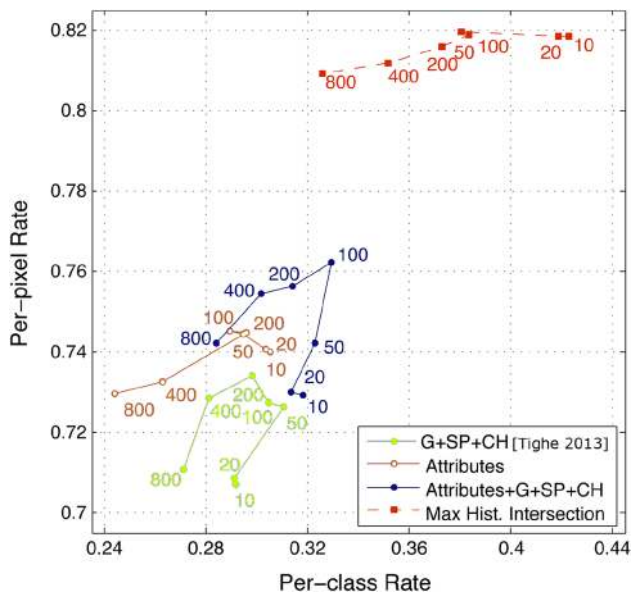


Fig. 19 Evaluation of using our scene attributes as a global feature for scene parsing on the SIFT-Flow dataset. The x -axis shows mean per-class classification rate and the y -axis shows per-pixel classification rate. The plots show the impact of using different retrieval set sizes K ranging from 10 to 800. The closer a line gets to the top-right corner of the space (regardless of the value of K), the better the retrieval method. The *blue* plot shows the result of using gist (G), spatial pyramid (SP), and color histogram (CH) together as scene descriptors for finding retrieval sets (Tighe and Lazebnik 2013). Using scene attributes alone improves the per-pixel rates while the per-class rates are similar. Using scene attributes together with the previous three features increases both the per-pixel rates and the per-class rates. Maximum histogram intersection is the upper bound we get by finding retrieval set using ground-truth labels of the query image (Color figure online)

age of the per-pixel rates over all classes. We also report the performance of an “optimal retrieval set”, which uses ground-truth class labels instead of global features to find similar scenes for the query image. This retrieval set is called Maximum Histogram Intersection. It is found by ranking training images according to the class histogram intersections they have with the query image. This optimal retrieval set is meant to be a performance upper bound and should provide an insight into how much room for improvement there is in the image retrieval step.

Figure 19 shows the performance comparison among different global features. As we can see from the result, using only scene attributes as global features we get higher per-pixel rates than (Tighe and Lazebnik 2013), which uses three global features (G+SP+CH), while getting similar per-class rates. When combining our scene attributes with those three global features (Attributes+G+SP+CH), both the per-pixel rates and the per-class rates increase significantly [73.4, 29.8 % ($K = 200$) vs. 76.2, 33.0 % ($K = 100$)]. Considering the compact size of our scene attributes, 102 dimensions compared with the 5184-dimension G+SP+CH, this result supports the hypothesis that scene attributes are a compact

and useful high-level scene representation. It is also worth noting that adding more features beyond this point does not necessarily improve the performance because all features, including weak ones, contribute equally to the found retrieval sets. For instance, by using all 12 features from (Xiao et al. 2010) together with the scene attributes, the per-pixel rate and the per-class rate drop to 74.6 and 30.4 % respectively ($K = 100$).

7.2 Data-Driven Image Captioning with Attribute-Based Scene Retrieval

In the previous subsection we showed that attribute-based image retrieval can lead to improved scene parsing performance. Here we take an analogous approach—modifying the image retrieval stage of data-driven pipeline—for the task of image captioning. There has been significant recent interest in generating natural language descriptions of photographs (Kulkarni et al. 2013; Farhadi et al. 2010b). These techniques are typically quite complex: they recognize various visual concepts such as objects, materials, scene types, and the spatial relationship among these entities, and then generate plausible natural language sentences based on this scene understanding. However, the “Im2text” (Ordonez et al. 2011) method offers a simple data-driven alternative. Instead of trying to achieve deep scene understanding and then link visual entities to natural language entities, Im2text simply tries to find a similar scene in a large database and then “steals” the existing caption in its entirety. Because the success of Im2text depends entirely on its ability to find similar scenes (which hopefully have similar captions), we use it to evaluate attribute-based scene representations.

In its simplest form, Im2text uses the gist and tiny image descriptors (Torralba et al. 2008) as the global features for image retrieval. On top of this baseline, Im2text also examines “content matching” in which the retrieved scenes are re-ranked according to overlap of recognized objects, “stuff”, people, and scene types. In Table 2, top we compare to the “baseline” Im2text by replacing the high dimensional global image features with our 102 dimensional predicted attributes. As in Im2text, the experiments are carried out by retrieving scenes from the SBU Captioned Photo Dataset⁶ which contains 1 million Flickr images with captions. Image captioning performance is quantified as the similarity between a ground truth caption and a predicted caption according to BLEU score (Papineni et al. 2002). As used in Im2text, the BLEU score is the proportion of words in the captions which are the same. This corresponds to the “unigram precision” BLEU score which does not consider the ordering of words. Stricter forms of BLEU also measure “ n -gram precision” which considers how many word sequences of length n are

⁶ <http://dsl1.cewit.stonybrook.edu/vicente/py/website/search>

Table 2 Global matching BLEU score comparison between baseline features and attributes on 10K, 100K and 1M dataset, 10K*, 100K* and 1M* are the dataset results with caption preprocessing

	10K	100K	1M
Gist + tiny image	0.0869 ± 0.002	0.0999 ± 0.009	0.1094 ± 0.0047
Attributes	0.0934 ± 0.01	0.1058 ± 0.015	0.1140 ± 0.0199
Chance	0.086		
	10K*	100K*	1M*
Gist + tiny image	0.02 ± 0.006	0.0255 ± 0.0079	0.0398 ± 0.0122
Attributes	0.0298 ± 0.0052	0.0366 ± 0.0132	0.0551 ± 0.0258
Chance	0.0144		

Removing stop words, punctuations, stemming, all lower case

shared between sentences. The ambiguity of the BLEU score has been criticized in the literature, and despite its technical shortcomings we use it here in order to compare to previous methods. By using attributes instead of the baseline Im2text image features we see small gains in captioning performance – from average BLEU of 0.109 with gist and tiny images to 0.114 with predicted attributes. However, the content matching approach in Im2text which re-ranks similar scenes based on deeper scene understanding still exceeds both global matching schemes with an average BLEU of 0.126.

In the course of the experiments, we noticed that “chance” performance, in which a random caption is taken from the database for each query, was surprisingly competitive with the retrieval methods (average BLEU of .086). We believe this is because the unigram BLEU score rewards matched words such as “the” and “a” just as much as more content descriptive terms such as “forest” and “baby” and this obscures the differences between the retrieval methods. In Table 2, bottom we measure performance when we perform three operations to try and make the caption evaluation more rigorous: (1) stemming captions to root words, e.g. “run”, “ran”, “running” and “runs” are stemmed to “run”. (2) converting all words to lower case and (3) removing frequent “stop words” such as articles and prepositions. While steps 1 and 2 make it easier for captions to match under the BLEU criteria, step 3 dramatically decreases performance. Chance performance drops by a factor of 6 to .014. The difference between attributes and the baseline global image features is more pronounced under this scheme -0.055 versus 0.040 , respectively. These numbers are quite low in absolute terms because the captions in the Im2text database are exceedingly diverse, even for very similar scenes. Figure 20 shows example retrieval results where scene attributes lead to better matching scenes than the baseline features (gist + tiny images). For these examples, the captions obtained using attributes get higher BLEU scores than the captions from the baseline features.

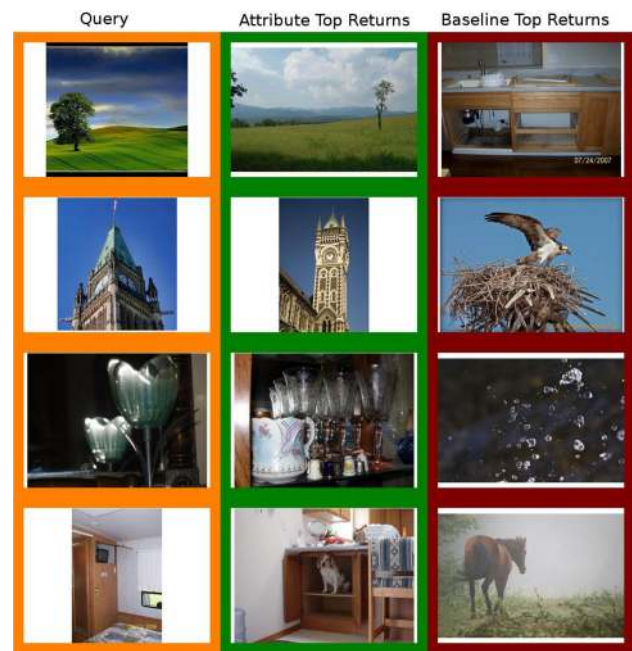


Fig. 20 Attribute search versus Im2Text baseline. Example image retrieval results that show how scene attributes can provide more relevant results than the Im2Text baseline

7.3 Text-Based Image Retrieval with Scene Attributes

In the previous subsection we address generating captions from images and here we address the inverse task—retrieving images most relevant to a text query. There has been recent work on directly using attributes to search image collections (Kumar et al. 2011; Siddiquie et al. 2011; Kovashka et al. 2012; Scheirer et al. 2012) but here we investigate the longstanding problem of query-by-text, as in typical search engines. Therefore we first focus on learning a mapping from text keywords to attributes and then perform image retrieval in the 102 dimensional predicted attribute space. While the observed correlations between attributes and keywords are interesting and the retrieval results are promising, we do not claim that this application represents the state of the art in text-based image retrieval. Instead we offer a qualitative comparison to the most common image retrieval baseline tf-idf (term frequency-inverse document frequency) weighted comparisons of query keywords to captions.

7.3.1 Attribute and Word Correlation

To link keyword queries to our scene attribute representation, we measure the correlation of individual scene attributes and words with a method inspired the co-occurrence model of Hironobu et al. (1999). Hironobu et al. (1999) counts the number of co-occurrences of image patch features and caption keywords to find correlations between keywords and features. We discover the correspondence of attributes and key-

Table 3 Examples of top correlated words for attributes

Attr.	Sail./boat.	Driving	Eating	Railroad	Camping
Top 20 correlated words	Cruis	Sand	Bar	Moon	Grass
	Harbor	Road	Cabinet	Railwai	Pastur
	Ocean	Sidewalk	Desk	Lit	Field
	Sail	Lane	Kitchen	Exposur	Forest
	Swim	Dune	Oven	Harbour	Landscap
	Boat	Highwai	Tv	Track	Fallen
	Dock	Moon	Een	Southern	Lone
	Sunset	Traffic	Shelf	Train	Hidden
	Sky	Canyon	Breakfast	Mother	Hill
	Airplan	Track	Dine	Star	Flow
	Beach	Wind	Tabl	Light	Stream
	Sea	Order	Ceil	Tank	Canyon
	Coast	Cross	Candl	Traffic	Oak
	Wave	Bridg	Lit	Night	Trail
	Ski	Cabl	Sunris	Glow	Distanc
	Clear	Ga	Chocol	Pass	Road
	Lake	Drive	Second	Shadow	Camp
	Ship	Fallen	Room	Salt	Creek
	Moon	Colorado	Bathroom	Site	Grow
	Sunris	Toward	Cherri	Wing	Wind

Words are stemmed

words by counting the number of times that a given attribute and keyword appear in the same image.

We use the 10,000 images and captions from Im2text dataset as our training set. We only consider the 1,000 most common words in the Im2text dataset as keywords. Let n be the size of image dataset. We create an n -long vector W_i , for each word w_i and an n -long vector A_j for each attribute a_j . The k th element of W_i indicates if the word w_i exists in the caption of the k th example in the dataset. Similarly the k th element of A_j indicates if the attribute a_j exists in the image of the k th example in the dataset. For these experiments, an attribute exists in an image if the SVM classifier's confidence is above -0.75 . This threshold is set fairly low so that the attribute detections are not overly sparse.

We use a binary-idf, binary-inverse document frequency, style weighting for word vectors and tf-idf, term frequency-inverse document frequency, style weighting for attribute vectors. In detail, if w_i exists in the caption of the k th example, the weight of the k th element in W_i is set to be $1/f_w$, where f_w is the inverse document frequency of w_i ; otherwise the weight is zero. Similarly, if a_j exists in the image of the k th example, the weight of the k th element in A_j is set to be $conf/f_a$, where $conf$ is the scene attribute confidence score passed through a sigmoid, and f_a is the inverse document frequency of a_j ; otherwise the weight is zero. Finally, the

Table 4 Examples of top correlated attributes for words

Words	Kitchen	Mountain
Top 10 correlated attr.	Tiles	Far-away horizon
	Enclosed area	Hiking
	Cleaning	Camping
	Reading	Natural
	Wood (not part of a tree)	Foliage
	Glossy	Vegetation
	Electric/indoor lighting	Trees
	Glass	Rugged scene
	Eating	Shrubbery
	Studying/learning	Leaves
Words	Beach	Dress
Top 10 correlated attr.	Ocean	Cloth
	Far-away horizon	Medical activity
	Sand	Enclosed area
	Waves/surf	Paper
	Sunbathing	No horizon
	Sailing/boating	Sterile
	Diving	Research
	Swimming	Electric/indoor lighting
	Still water	Stressful
	Open area	Man-made

correlation between word w_i and attribute a_j is simply the inner product of W_i and A_j ; $C_{ij} = W_i * A_j$.

Binary-idf is very similar to tf-idf. The difference between the two is that for binary a word is counted only one for the document it appears in. The number of times the word appears in the document is not considered. If a word appears in one document, the binary-idf value will be $1/(\text{inverse document frequency})$, otherwise the value is zero. Binary-idf can suppress some words that have little semantic meaning in terms of our scene attributes but appear often in the document. For example, the words “nice”, “like” may appear more times than “sky” does in one document, but they are less informative and less related to our scene attributes.

Table 3 shows top correlated words for attributes and Table 4 shows top correlated attributes for words. We find that attributes predicted by our classifiers have high correlation with text words. The correlations tend to be quite reasonable, e.g. for the attribute ‘sailing/boating’ the most correlated keywords are ‘cruise’, ‘harbor’, ‘ocean’, ‘sail’, ‘swim’, etc. Note some words are transformed because of stemming.




































Attribute Top Returns					
sky	 Great egret against the glorious blue sky.	 That grove of trees on the right half of the horizon is the forest around Wendi's parent's home.	 Great cloud formation over the mountains in Rocky Mountain National Park. 2008	 Little pine tree in the big heath.	 A battle in the sky. Stormy black rain clouds versus the dry hot white desert clouds.
dark sky	 flying birds with the beautiful overcast sky colored by the sunrise	 I just loved the pink of this boat against the blue of the sky and the clouds.	 Taken in Mount. Bro-mo some time ago, when the sky is so blue and the cloud is so great.	 castle in the clouds	 The sun breaks through a cloud and illuminates a ship near the Golden Gate Bridge.
flower	 A stalk of brilliant yellow and orange flowers in the mountains of Oaxaca, Mexico.	 A little pink flower showing its beauty - Canon S5IS in Manual Mode	 I was standing over a railing, cursing the fact I didn't bring along my telephoto lense. Or a bottle of water.	 Little pink in the flower	 This was all over my pine tree it was really pretty.
red	 I love Dimples1967 little green haired girl dressed up in that awesome clown outfit!! Those boots were amazing!	 red rose in sunlight v	 Big new flower on new plant in the backyard.	 olive oil in a dish on a piece of glass on trestles, coloured card on the floor lit by a halogen desk lamp...enjoy	 Large dog at bar in Grand Lake
mountain	 Cloud curtain over Table Mountain	 Incredibly blue sky over Montserrat	 The lighthouse is on the rock in the distance, the fog is creeping up the shore and obscuring most of the rock.	 Still little water in the river	 A beautiful view from the microwave tower site on Grey Mountain in Whitehorse on June 30, 2006.
snow mountain	 in the airplane flying to chicago, 7/07	 Mt. Rainier right before sunrise (5:30am). Note the blue sky that early in the morning.	 looking out over the entrance to ruby bowl on blackcomb mountain	 Drinking water after the spruce trap field on the traverse over Boundary to Iroquois from Algenquin	 A picture of a mountain during a train ride, taken by my sister Elizabeth, somewhere near Anchorage.
night	 an unlit candle in a frosty glass	 Yvette makes her own heart. The red line in the back is a car that went past. Cool!	 High contrast strong orange rose against a black background.	 we had to ask for beer in plastic cups at strike! bowling bar in Melbourne	 gho0o0o0oost driver! I was packing the car for move in day tomorrow(!) and noticed this in my mirror

Fig. 21 Attribute based image retrieval results on 1M Im2text dataset. Although the captions are shown here for completeness, our text-based image retrieval method did not see them at query time, whereas the TF-

IDF method (Fig. 22) uses the captions exclusively. These search terms were selected for variety and breadth. We believe they are representative of results from our attribute based image retrieval pipeline generally

7.3.2 Word-to-Attribute Correlation Applied to Image Retrieval

Now that we can relate keywords to attributes, we apply the word-to-attribute correlation scores to the image retrieval task. Our text-based image retrieval approach is content-

based because it does not rely on text metadata (captions) for the database images at query time. For example, if user inputs a text query “sky”, we hope to map that text to its corresponding visual features (via attributes), such as blue background, clustered clouds and horizon line, and retrieve images which contain those visual features.

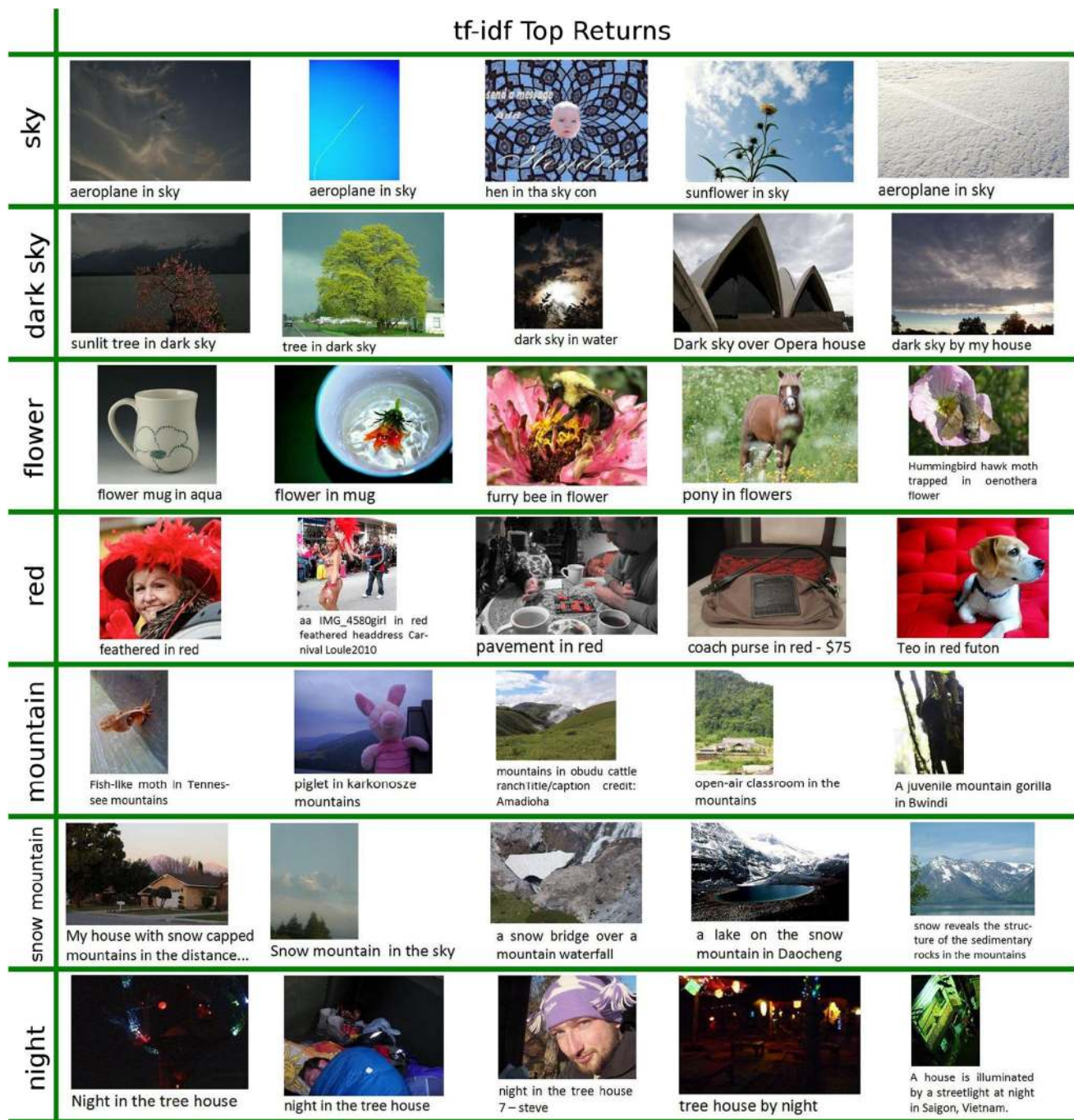


Fig. 22 Tf-idf based image retrieval results on 1M Im2text dataset

Given the query text, we break the text into words. Let T_{query} be the vector of query word indices. These indices are the positions of the query words in the list of 1,000 most common caption words. We use T_{query} and word-attribute correlation we have obtained to create a “target” scene attributes representation. Each word w_i has a vector of correlations $C_i = \langle c_{i,1}, \dots, c_{i,j}, \dots, c_{i,102} \rangle$, where each element $c_{i,j}$ is the correlation of word w_i and attribute a_j . The target scene

attribute representation is defined as the average of correlation vectors of the words in the query,

$$F_{target} = \frac{1}{N} \sum_{k=1}^N C_{T_{query,k}} \quad (1)$$

where N is the length of T_{query} , and $T_{query,k}$ is the k th element of T_{query} , the index of the k th query word in common words

list. We consider the same word multiple times if it appears multiple times in the caption.

We then learn multi-linear regressions to map target scene attributes to predicted scene attributes, which are the output feature vectors of attributes classifiers. In the training dataset, for each image, we know both its target attributes and predicted attributes. We then learn the regression to map from those attributes in the target representation to a_j in the predicted representation. Finally, we search for the nearest neighbors of the query's predicted attribute representation in the test dataset.

7.3.3 Experiments

For testing we search 90,000 captioned images from the Im2text dataset. We compare our method to tf-idf based retrieval because it is a widely used baseline method for text-based image retrieval. Figures 21 and 22 show the results of both methods separately. From the results, we can see that attribute-based image retrieval gives very promising search results. For most search results returned by attribute-based method, the target specified by the query text are the dominant visual concept in the retrieved images. However, that is not the case for tf-idf based method. For example, for the “flower” query, the five images returned by the attribute-based method all depict flowers, while the dominant objects in images returned by the tf-idf based method contain a mug, pony and bee. Our method also has some success with multiple keyword queries. For the queries “snow mountain” and “dark sky” most of the retrieved scenes have the correct semantics. These qualitative results again reinforce the idea that the 102 dimensional scene attributes are surprisingly expressive given their compact size and that our attribute classifiers are reliable enough to support other image understanding applications.

8 Discussion

In this paper, we use crowdsourcing to generate a taxonomy of scene attributes and then annotate more than ten thousand images with individual attribute labels. In order to promote the trustworthy responses from the Mechanical Turk users we employ several simple yet effective techniques for quality control. We explore the space of our discovered scene attributes, revealing the interplay between attributes and scene categories. We measure how well our scene attributes can be recognized and how well predicted attributes work as an intermediate representation for zero shot learning and image retrieval tasks.

8.1 Future Work

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 be the optimal descriptors for 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. Classification 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.

Acknowledgments 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.

Appendix: Scene Attributes

See Appendix Table 5.

Table 5 Complete list of discovered scene attributes

Scene attributes	
Functions/affordances	
Sailing/boating	Driving
Biking	Transporting things or people
Sunbathing	Vacationing/touring
Hiking	Climbing
Camping	Reading
Studying/learning	Teaching/training
Research	Diving
Swimming	Bathing
Eating	Cleaning
Socializing	Congregating
Waiting in line/queuing	Competing
Sports	Exercise
Playing	Gaming
Spectating/being in an audience	Farming
Constructing/building	Shopping
Medical activity	Working
Using tools	Digging
Conducting business	Praying

Table 5 continued

Materials	
Fencing	Railing
Wire	Railroad
Trees	Grass
Vegetation	Shrubbery
Foliage	Leaves
Flowers	Asphalt
Pavement	Shingles
Carpet	Brick
Tiles	Concrete
Metal	Paper
Wood (not part of a tree)	Vinyl/linoleum
Pubber/plastic	Cloth
Sand	Rock/stone
Dirt/soil	Marble
Glass	Waves/surf
Ocean	Running water
Still water	Ice
Snow	Clouds
Smoke	Fire
Surface properties/lighting	
Natural light	Direct sun/sunny
Electric/indoor lighting	Aged/worn
Glossy	Matte
Sterile	Moist/damp
Dry	Dirty
Rusty	Warm
Cold	
Spatial envelope	
Natural	Man-made
Open area	Semi-enclosed area
Enclosed area	Far-away horizon
No horizon	Rugged scene
Mostly vertical components	Mostly horizontal components
Symmetrical	Cluttered space
Scary	Soothing
Stressful	

References

- Berg, T., Berg, a., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. *ECCV*, 6311, 663–676.
- Chen, D., & Dolan, W. (2011). Building a persistent workforce on mechanical turk for multilingual data collection. In *The 3rd human computation workshop (HCOMP)*.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR*.
- Ehinger, K.A., Xiao, J., Torralba, A., & Oliva, A. (2011). Estimating scene typicality from human ratings and image features. In *33rd annual conference of the cognitive science society*.
- Eigen, D., & Fergus, R. (2012). Nonparametric image parsing using adaptive neighbor sets. In *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on* (pp. 2799–2806). doi:10.1109/CVPR.2012.6248004.
- Endres, I., Farhadi, A., Hoiem, D., & Forsyth, D. (2010). The benefits and challenges of collecting richer object annotations. In *ACVHL 2010 (in conjunction with CVPR)*.
- Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *CVPR*.
- Farhadi, A., Endres, I., & Hoiem, D. (2010a). Attribute-centric recognition for cross-category generalization. In *CVPR*.
- Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, I., C., Hockenmaier, J., & Forsyth, D.A. (2010b). Every picture tells a story: Generating sentences from images. In *Proc ECCV*.
- Ferrari, V., & Zisserman, A. (2008). Learning visual attributes. *NIPS 2007*
- Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *Computer vision, 2009 IEEE 12th International Conference on* (pp. 1–8). doi:10.1109/ICCV.2009.5459211.
- Greene, M., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137–176.
- He, X., Zemel, R., & Carreira-Perpinan, M. (2004). Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on* (Vol. 2, pp. II-695–II-702). doi:10.1109/CVPR.2004.1315232.
- Hironobu, Y.M., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Boltzmann machines, neural networks*, (pp. 405409).
- Hoiem, D., Efros, A. A., & Hebert, M. (2007). Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1), 151–172.
- Kovashka, A., Parikh, D., & Grauman, K. (2012). Whittlesearch: Image search with relative attribute feedback. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
- Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., & Berg, T.L. (2013). Babytalk: Understanding and generating simple image descriptions. In *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*.
- Kumar, N., Berg, A., Belhumeur, P., & Nayar, S. (2009). Attribute and simile classifiers for face verification. In *ICCV*.
- Kumar, N., Berg, A.C., Belhumeur, P.N., & Nayar, S.K. (2011). Describable visual attributes for face verification and image search. In *IEEE transactions on pattern analysis and machine intelligence (PAMI)*.
- Ladicky, L., Sturgess, P., Alahari, K., Russell, C., & Torr, P.H. (2010). What, where and how many? combining object detectors and crfs. In *Computer vision-ECCV 2010, Springer* (pp. 424–437).
- Lampert, C.H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.
- Lasecki, M., White, M., & Bigham, K. (2011). Real-time crowd control of existing interfaces. In *UIST*.
- Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.
- Liu, C., Yuen, J., & Torralba, A. (2011a). Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12), 2368–2382.
- Liu, J., Kuipers, B., & Savarese, S. (2011b). Recognizing human actions by Attributes. In *CVPR*.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579–2605), 85.
- Malisiewicz, T., & Efros, A.A. (2008). Recognition by association via learning per-exemplar distances. In *Computer vision and pattern*

- recognition, 2008. *CVPR 2008. IEEE Conference on, IEEE* (pp. 1–8).
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3), 145–175.
- Oliva, A., & Torralba, A. (2002). Scene-centered description from spatial envelope properties. In *2nd Workshop on biologically motivated computer vision (BMCV)*.
- Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural information processing systems (NIPS)*.
- Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in neural information processing systems* (pp. 1410–1418).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics, association for computational linguistics, Stroudsburg, PA, USA, ACL* (pp. 311–318). doi:10.3115/1073083.1073135.
- Parikh, D., & Grauman, K. (2011a). Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*.
- Parikh, D., & Grauman, K. (2011b). Relative attributes. In *CCV*.
- Patterson, G., & Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of the 25th conference on computer vision and pattern recognition (CVPR)*.
- Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on* (pp. 1–8). doi:10.1109/ICCV.2007.4408986.
- Rohrbach, M., Stark, M., & Schiele, B. (2011). Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on IEEE* (pp. 1641–1648).
- Russakovsky, O., & Fei-Fei, L. (2010). Attribute learning in largescale datasets. In *ECCV 2010 workshop on parts and attributes*.
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1), 157–173.
- Sanchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3), 222–245.
- Scheirer, W. J., Kumar, N., Belhumeur, P. N., & Boult, T. E. (2012). Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
- Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the 9th European conference on computer vision*. Berlin: Springer, ECCV'06 (pp. 1–15). doi:10.1007/11744023_1.
- Shotton, J., Johnson, M., & Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). doi:10.1109/CVPR.2008.4587503.
- Siddiquie, B., Feris, R. S., & Davis, L. S. (2011). Image ranking and retrieval based on multi-attribute queries. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
- Socher, R., Lin, C. C., Ng, A. Y., & Manning, C. D. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th international conference on machine learning (ICML) Vol. 2*, p. 7.
- Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. In *First IEEE workshop on internet vision at CVPR 08*.
- Su, Y., Allan, M., & Jurie, F. (2010). Improving object classification using semantic attributes. In *BMVC*.
- Tighe, J., & Lazebnik, S. (2013). Superparsing. *International Journal of Computer Vision*, 101, 329–349. doi:10.1007/s11263-012-0574-z.
- Torralba, A., Fergus, R., & Freeman, W. T. (2008a). 80 Million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 1958–1970.
- Torralba, A., Fergus, R., & Freeman, W. T. (2008b). 80 Million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8), 1371–1384.
- Xiao, J., Hays, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*.
- Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., & Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. In *ICCV*.