To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

Interactively Building a Discriminative Vocabulary of Nameable Attributes

Devi Parikh
Toyota Technological Institute, Chicago (TTIC)
dparikh@ttic.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

Abstract

Human-nameable visual attributes offer many advantages when used as mid-level features for object recognition, but existing techniques to gather relevant attributes can be inefficient (costing substantial effort or expertise) and/or insufficient (descriptive properties need not be discriminative). We introduce an approach to define a vocabulary of attributes that is both human understandable and discriminative. The system takes object/scene-labeled images as input, and returns as output a set of attributes elicited from human annotators that distinguish the categories of interest. To ensure a compact vocabulary and efficient use of annotators' effort, we 1) show how to actively augment the vocabulary such that new attributes resolve inter-class confusions, and 2) propose a novel "nameability" manifold that prioritizes candidate attributes by their likelihood of being associated with a nameable property. We demonstrate the approach with multiple datasets, and show its clear advantages over baselines that lack a nameability model or rely on a list of expert-provided attributes.

1. Introduction

Visual attributes offer a useful intermediate representation between low-level image features and high-level categories, and are the subject of a growing body of work in the recognition literature [1–8]. Whereas traditional object detectors are built via supervised learning on image features, an attribute-based detector first predicts the presence of an array of visual properties (e.g., 'spotted', 'metallic', etc.), and then uses the outputs of those models as features to an object classification layer.

Attributes are attractive because they allow a recognition system to do much more than predict category labels. Attributes shared among objects facilitate *transfer*, such as zero-shot learning of a new category based on a specification of which attributes define the unseen object [2, 9]. In addition, the fact that attributes are defined by human language makes them useful to compute meaningful *descriptions* of unfamiliar objects (e.g., the system trained on dogs and cats cannot recognize a rabbit, but it can at least describe it as furry and white), or to report on *unusual aspects* of some instance (e.g., the pot is missing a handle).

Figure 1. Interactively building a vocabulary of nameable attributes.

Problem: Despite attributes' apparent assets for recognition problems, a critical question remains unaddressed: which visual attributes should be learned? Specifically, which attributes are relevant for a given object categorization task? Most existing work uses a list of attributes handgenerated by experts—whether from knowledge bases prepared by domain specialists [2, 6, 10], or by vision researchers with intuition for the properties needed [1, 5, 8]. Others leverage text on the Web to discover potentially relevant properties [3, 4]; while less expensive in terms of human effort, such sources can be narrow or biased in scope, and most "obvious" visual properties are unlikely to ever be mentioned in text published near images (e.g., it may be difficult to mine the Web to learn that highways and kitchen scenes have the 'manmade' attribute in common, or that TVs and microwave ovens are both 'rectangular').

Unfortunately, "nameability" and discriminativeness appear to be at odds. On the one hand, even if we can afford to ask domain experts to provide a list of attributes most descriptive of the objects we wish to categorize, there is no guarantee that those attributes will be sufficiently separable in the image feature space—a necessary condition if they are intended to serve as the mid-level cues for recognition. On the other hand, even though we have abundant machine learning tools to discover discriminative splits in image feature space that together carve out each object of interest, there is no guarantee that any such features will happen to correspond to human-nameable traits—a desirable condition if we are to leverage the transfer, description, and other attractive aspects mentioned above.

Goal and approach: We aim to build *discriminative* attribute *vocabularies* that are amenable to visual recognition tasks, yet also serve as interpretable mid-level cues. We pro-

pose an interactive approach (Figure 1) that prompts a (potentially non-expert) human-in-the-loop to provide names for attribute hypotheses it discovers. The system takes as input a set of training images with their associated category labels, as well as one or more visual feature spaces (Gist, color, etc.), and returns as output a set of attribute models that together can distinguish the categories of interest.

To visualize a candidate attribute for which the system seeks a name, a human is shown images sampled along the direction normal to some separating hyperplane in the feature space. Since many hypotheses will not correspond to something humans can visually identify and succinctly describe, a naive attribute discovery process—one that simply cycles through discriminative splits and asks the annotator to either name or reject them—is impractical. Instead, we design the approach to actively minimize the amount of meaningless inquiries presented to an annotator, so that human effort is mostly spent assigning meaning to divisions in feature space that actually have it, as opposed to discarding un-interpretable splits. We accomplish this with two key ideas: at each iteration, our approach 1) focuses on attribute hypotheses that complement the classification power of existing attributes collected thus far, and 2) predicts the nameability of each discriminative hypothesis and prioritizes those likely to be nameable. For the latter, we explore whether there exists some manifold structure in the space of nameable hyperplane separators.

Ultimately, at the end of the semi-automatic learning process, we should have discovered something akin to the classic "20 questions" game—divisions that concisely carve up the relevant portions of feature space and are also human understandable. These attributes can then be used for recognition, zero-shot learning, or describing novel images.

Contributions: Our main contribution is to show how to efficiently designate a discriminative attribute vocabulary. Beyond the system itself, this work raises a number of interesting questions regarding iterative discovery of attributes, visualization of discriminative classifiers, and the statistics of those visual cues that humans identify with language. In particular, we are the first to show that the space of nameable attributes is structured, and can be used to predict the nameability of splits in a visual feature space.

2. Related Work

We overview work on attributes and interactive learning.

Learning a set of hand-listed attributes: A number of researchers have explored attribute models, demonstrating their applicability for color and texture naming [11], intermediate features for object recognition [2, 8], face verification [5], zero-shot learning [2, 6, 12], or description and part localization [1, 13]. Typically one gathers image exemplars containing or lacking an attribute of interest, and then trains a classifier to predict whether the property is present in a novel image; however, recent work shows the value in jointly training attribute and object category models [7, 8]. In contrast to our approach, all such methods manually define the attributes of interest *a priori*, and none attempts to model whether the desired attributes are predictable within the chosen visual feature space.

Mining online text and images to discover attributes: In reaction to the expense and/or expertise required to manually define attributes of interest, some recent work aims to discover attribute-related concepts on the Web. The authors of [3] discover semantic relatedness among categories and attributes using a variety of text sources (e.g., Wikipedia), while the "visualness" of adjectives or nouns appearing near image content is evaluated automatically using Web data in [4, 14]. While the Web can be a rich source of data, it can also be biased or lack information that is critical to the categorization task at hand. For example, while one can collect useful descriptions of handbags and shoes from shopping websites [4], it may be substantially harder to find text that adequately describes generic categories like offices, hallways, or roads. Furthermore, the attributes discovered via text-mining may not be separable in the visual feature space, and/or are likely to be generative as opposed to discriminative for the high-level object categories of interest. Hence, we propose to first discover visually discriminative features as potential attributes, and then determine their nameability.

Human-in-the-loop: Interactive systems bring a human into the loop to facilitate some target task. When training an object recognition system, active learning algorithms can focus annotation requests so as to quickly improve the object or context models (e.g., [15–17]). In a novel form of interactive classification, the system proposed in [10] recognizes a bird species with the help of a human expert; it prompts the human to answer questions pertaining to a visual attribute of the bird, which is actively selected from a list of expert-provided attributes. Like the above, we also wish to efficiently utilize human effort, but from the novel perspective of reducing the proportion of unnameable queries we pose to the user.

Nearly all previous work in active learning assumes that any query will be answerable by a human (hence the standard term "oracle"). However, as also observed in [18], irrelevant examples in an unlabeled pool can slow down active learning. Their approach avoids presenting such examples by training a second classifier to distinguish between relevant and irrelevant data points. Our motivation for ignoring unnameable attributes is related; however, our problem setting is quite different, as is our solution to learn a manifold that captures the structure among valid classifiers.

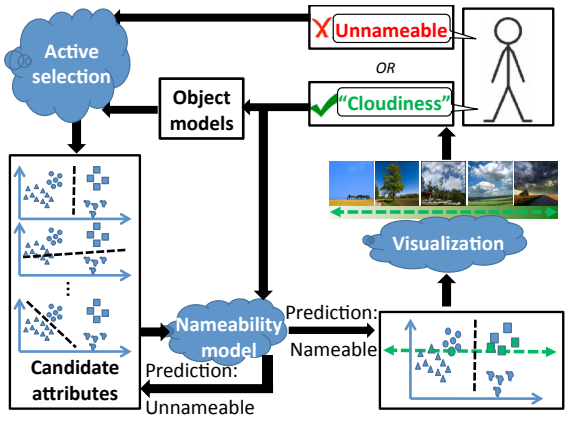

Figure 2. Overview of our approach.

3. Approach

First, we formally state the problem. We are given a set of n images $\{I_i\}$, along with their representations and associated class labels $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an arbitrary visual feature space, and $\mathcal{Y} = \{y_1, \ldots, y_K\}$ consists of K discrete classes. (Our experiments consider datasets where $\mathcal{Y}$ consists of scene categories or animal classes.) We wish to discover an intermediate representation of m attribute classifiers $\mathcal{A} = [a_1, \ldots, a_m]$ such that each binary attribute $a_j : \mathcal{X} \rightarrow \{0, 1\}$ is nameable, *i.e.* has a semantic word meaning associated with it, and together the outputs of $\mathcal{A}$ are discriminative, *i.e.* a classifier $h : \mathcal{A} \rightarrow \mathcal{Y}$ has high accuracy. Note that h is an instance of "direct attribute prediction", as defined in [2]. In our implementation, the attribute classifiers and h are all linear support vector machines (SVM).[1]

Figure 2 shows an overview of our approach. At each iteration t, we actively determine an attribute hypothesis (a hyperplane in the visual feature space) that helps discriminate among classes that are most confused given the current collection of attributes $\mathcal{A}_t$. We then estimate the probability that the hypothesis is nameable, using a learned model of nameability that is continually augmented by any hypotheses accepted (*i.e.* named) by the human in the loop. If it appears unnameable, we discard it and loop back to select the next potential attribute hypothesis. If it appears nameable, the system creates a visualization of the attribute using a subset of training images, presents the images to the annotator, and requests an attribute label. The annotator may either accept and name the hypothesis, or reject it. If it is accepted, we append this new named attribute a_j to our discovered vocabulary, $\mathcal{A}_{t+1} = [\mathcal{A}_t, a_j]$, retrain the higher-level classifier h accordingly, and update our nameability model. If it is rejected, the system loops back to generate a new attribute hypothesis. Thus, only those attributes that are named by the user are added to the pool and can be used for recognition. The discovery loop terminates once human resources are exhausted, or when a desired number of named attributes have been collected.

[1] Throughout we refer to intermediate attribute classifiers as simply "attributes", and higher-level object/scene category models as "classifiers".

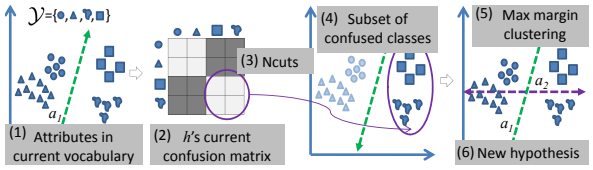

Figure 3. Sketch of the process to discover a discriminative attribute hypothesis. Refer to text for steps (1)-(6).

Thus, the key technical challenges are: (1) determining attribute hypotheses based on visual feature separability and current class confusions, (2) modeling the "nameability" of these hypotheses, and (3) selecting representative image exemplars that will prompt reliable human responses of attribute names, as we detail in the following subsections.

3.1. Discovering Attribute Hypotheses

Figure 3 illustrates the process for identifying candidate discriminative attributes. At each iteration t, we classify the training data into the different categories in $\mathcal{Y}$ using the attribute vocabulary $\mathcal{A}_t$ collected thus far (see (1) in Figure 3). We compute the confusion matrix among the categories (2), and view this as an adjacency matrix on a fully connected graph whose nodes correspond to the different categories. A strong edge between two categories indicates high confusion between them. We perform normalized cuts on this graph (3) to discover two or more clusters. Each cluster is a subset of classes that are most confused amongst themselves under the current vocabulary (4).

Having actively focused on this subset of classes, we use unsupervised iterative max-margin clustering [19] to generate a new attribute hypothesis that distinguishes them (5). This finds the maximum margin hyperplane separating the selected data into two groups. We "purify" the resulting clusters by assigning each training image to the cluster containing the most images from its class, and then train an SVM to discriminate between these two pure clusters. The resulting hyperplane is our candidate binary attribute (6).

If it is predicted to be nameable (see Sec. 3.2), it is presented to the annotator (see Sec. 3.3), and, if named, added to our attribute vocabulary: $\mathcal{A}_{t+1} = [\mathcal{A}_t, a_j]$. An accepted hypothesis SVM is applied to the entire input training dataset to infer their attribute values, $a_j(x_1), \ldots, a_j(x_n)$, which are needed to update h.

If instead it is predicted to be unnameable, or if the annotator were to reject this hypothesis, we generate a second hypothesis by focusing on the second cluster of classes. Similarly, subsequent hypotheses are determined by generating more confusion clusters with Ncuts, and focusing on them in decreasing order of their density (average confu-

sion). The system does not allow for the same attribute hypothesis to be generated more than once. If the confusion-based hypotheses are exhausted without finding a nameable or accepted hyperplane, the system finally resorts to max-margin clustering on random subsets of classes. If the initial attribute vocabulary is empty, we focus on all classes simultaneously for max-margin clustering in the first iteration.

Our approach to discover discriminative hypotheses is thus both active and myopic (greedy) in nature. Each hypothesis is intentionally chosen to introduce a discriminative binary property among a subset of two or more classes. However, note that since a given split considers only a subset of classes, the approach does *not* assume that all classes have a clean binary membership per attribute (as in [2]); a hypothesis can divide any class outside of the current subset (*i.e.*, some cow instances may be 'spotted' while others are 'unspotted'). In fact, with a soft margin SVM, even instances within a class in the current subset need not uniformly share the property.

Alternative strategies for computing initial hypotheses can be seamlessly incorporated in our system. In particular, to exploit multiple feature types and kernels simultaneously, one could instead use a multiple-kernel max-margin clustering framework (e.g., see [20]).

3.2. Predicting the Nameability of a Hypothesis

Many of the discriminative attribute hypotheses generated above may not correspond to properties of the images that humans can notice and name. A naive approach that cycles through all max-margin hyperplanes would waste annotator effort, since he/she would need to examine but then reject many candidates. To better utilize the human effort, we first predict the nameability of each hypothesis, and propose it only if it is likely to be nameable.

How can we possibly gauge "how nameable" the visual division implied by a hyperplane is? We speculate that there is *shared structure* among nameable visual attributes. In other words, those attributes in the given image descriptor space corresponding to truly nameable properties occupy only a portion of that space—a *manifold* of decision boundary directions. If so, that means we can prioritize the candidates according to their distance to this manifold.

Thus, we construct a low-dimensional *nameability manifold* using instances of SVM hyperplane parameters (weight vector and bias) that correspond to truly nameable attributes. Specifically, we model the nameable attributes with a mixture of probabilistic principal component analyzers (MPPCA) [21]. Given a novel attribute hypothesis, we compute the probability it belongs to the nameability manifold, that is, the probability the (projected) parameters would be generated by the mixture of subspaces. Since an attribute hypothesis is proposed to the annotator only if it is deemed nameable, we simply threshold (learned via cross-validation) this probability to make a hard decision.

To learn the manifold, the user responses ("accept" or "reject") for the proposed attributes are gradually collected at each iteration. At the first iteration, the nameability-space is not populated and hence the manifold is not yet learned.[2] When the system predicts nameability based on responses obtained thus far, it can adapt to the user, including his/her preferences for naming attributes. Alternatively, if a generic set of nameable attributes is available as input to the system (possibly on a disjoint set of images), they can be used to populate the space at the onset. In either case, the annotator is free to provide any name that he/she finds meaningful and relevant for the intended application. We stress that discovering and exploiting this manifold structure does *not* require that the attribute assigned to a split be agreed upon by different annotators; it is the nameability—not the precise name—that we intend to capture.

One potential concern of populating the nameability manifold at the same time the system collects attributes is the possible lack of exploration, *i.e.*, the manifold might bias future hypotheses to be similar to existing ones. However, our use of multiple lower-dimensional subspaces (which enable interpolation) accompanied by our active selection of discriminative attributes that complement existing ones (which encourages diversity) both counter this concern. Furthermore, one can control the manifold's selectivity based on the probability estimates.

Other empirically discovered statistics of natural images, such as the distribution of derivative filter outputs [22], have been exploited as powerful priors for low-level tasks ranging from computing intrinsic images [23] or image dehazing [24]. Our idea for a nameability manifold is similarly intended to constrain a solution space, albeit using the much more abstract notion of the statistics of visual features that humans identify with language. The proposed view of nameability is also supported by the multi-task learning work of [25], in which the authors observe that not all predictors are equally "good", and show how to learn smooth function classes by considering multiple prediction problems simultaneously.

3.3. Visualizing an Attribute

In order to display the attribute hypothesis to the annotator, we wish to convey the difference in the images that lie on either side of the hyperplane, while ensuring that within the constraints of finite data, we show *only* the changes induced along the direction orthogonal to that hyperplane. To do this, we first consider the range from the hyperplane within which 95% of the training data falls, in order to disregard potential outlier instances. We divide this range into 15 equidistant bins, and select three images per bin that are

[2] In our experiments, we bypass the nameability prediction step until we have at least three nameable attribute instances.

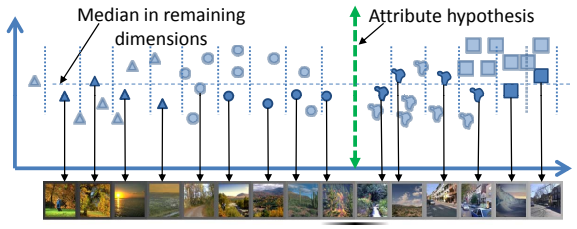

Figure 4. Sketch of our approach for sampling images to visualize an attribute hypothesis. (Only one image per distance bin is displayed here.)

closest to the median along all other dimensions, yielding a 3 $\times$ 15 collage.[3] Note that while it is straightforward to select the orthogonal direction for a linear hyperplane, one can also derive the "discriminative direction" for certain non-linear kernels as well [26], making our approach extendible to non-linear attribute predictors.

Figure 4 sketches this procedure, and Figure 5 shows several real examples. The shaded borders around each image indicate its distance from the hyperplane (dark = far on one side, bright = far on the other). The fading horizontal bars above and below the collage indicate where the images transition from one side of the hyperplane to the other.

Potential concerns are the inherent lossiness of our visualization (each collage is comprised of only a sample of the training data), and whether annotators are even capable of producing a property given such a sample. By design our sampling strategy isolates the property in the feature space we wish to analyze. Furthermore, previous studies with human subjects show that people are able to reliably organize a batch of images along discovered properties of their own choosing, such as openness, roughness, etc. [27]. Finally, in practice we can control for nameability uncertainty by soliciting redundant annotator responses and asking each annotator to rate the property from subtle to obvious (see Section 4). Again, our goal is to build a vocabulary of discriminative yet human-understandable attributes, and this does not require that humans agree on the name itself.

4. Experimental Setup

Datasets: We experiment with two datasets of 8 categories each: Outdoor Scene Recognition [27] (**OSR**): coast, forrest, highway, inside-city, mountain, open-country, street, tall-building; and a subset of the Animals With Attributes dataset [2] (**AWA**): elephant, giant-panda, giraffe, gorilla, leopard, lion, polar-bear, sheep. We purposely select animals that have some shared properties, but also are distinct; we did not try any other subset. These datasets assume a single class of interest is present per image; to generalize to multi-object images one would need to incorporate segmentation or sliding windows. We resize the AWA images to 256 $\times$ 256, and randomly select 200 images per class for training, and 100 for testing. For both datasets, we extract 512-D Gist descriptors [27] and 45-D global LAB color-histograms (15 bins per channel). Of course, other visual features can easily be used in our system. We perform LDA on each feature type to concentrate its discriminative power in fewer dimensions, yielding 7-D Gist and color as our input feature spaces $\mathcal{X}$ (classical LDA results in a $C-1$ dimensional space, where C is the number of classes). We build the MPPCA model using 5 components and 3-dimensional subspaces.

Offline collection: To perform thorough quantitative experiments with the proposed system, we want to simulate the human-in-the-loop using real annotator data, *i.e.*, still using user input, but without having to run our system online/live. Looking at our system in Figure 2 closely, we see that the only input from a user that affects subsequent iterations of the system is the user's decision to accept or reject a proposed hypotheses. This means that if we can generate an exhaustive set of all possible attribute hypotheses our system would ever generate, we can collect nameability responses ('accept' or 'reject') from subjects offline, and then draw on these responses during experiments that simulate the interactive system. Note that results from this setup will be *equivalent* to running the system live; we'll have simply collected more annotations than may actually be used. Given our approach to generate hypotheses (see Section 3.1), we can generate an exhaustive list of candidates using all possible subsets of the classes. For the 8-class datasets, this amounts to $\sum_{i=2}^{8}\binom{8}{i} = 247$ possible attribute hypotheses, for each dataset in each feature space.[4] When running the full system, we terminate after a fixed number of iterations.

Annotators: We collect nameability data via Amazon Mechanical Turk (MTurk). We show every attribute hypothesis visualization to 20 subjects. The subjects are asked to name the property of the images that changes from left to right, and indicate whether it is increasing (or simply present) or decreasing (absent) from left to right. Subjects are told that properties include characteristics such as color or layout or general feel, but should not be names of objects, scenes, or animals. We also ask subjects how obvious they think the property is, on a scale of 1 ("very subtle") to 4 ("obvious"). We use this obviousness rating to determine which attributes are nameable; an attribute hypothesis is nameable if the average score over 20 subjects is greater than 3. Note we use the wording "very subtle" rather than "unnameable" in the interface, since we want workers to attempt to find the pattern in every instance, rather than de-

[3] We experimented with a few variants, including cluster analysis in the remaining dimensions to ensure sampling from dense regions. The resulting visualizations were qualitatively very similar.

[4] Our 8-class datasets keep this number manageable. However, we stress that **the enumeration is only for experimental convenience**; with a live human-in-the-loop, it would not be done. Our system can certainly be used for datasets with many more classes.

(a) OSR gist: nameable, "congested"

(b) OSR color: nameable, "green"

(c) AWA gist: nameable, "spotted"

(d) AWA gist: unnameable

Figure 5. Example visualizations of attribute hypotheses created by our algorithm, and some responses received on Mechanical Turk. (d) shows a discriminative candidate attribute (elephant, lion, polar-bear, sheep on one side; gorilla, giraffe, giant-panda on the other) that subjects found to be unnameable.

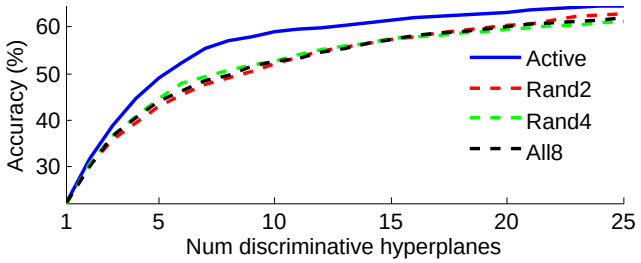

Figure 6. Our approach to actively discover discriminative attribute hypotheses outperforms several baselines.

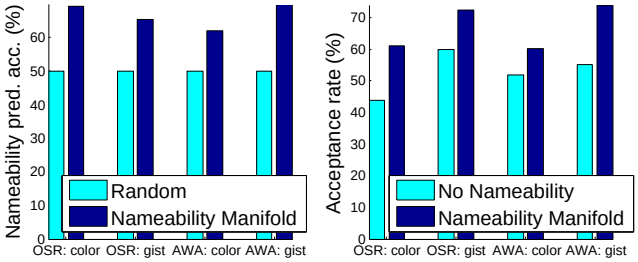

Figure 7. By predicting the nameability of attribute hypotheses, we focus annotator effort on properties likely to be nameable.

fault to immediately selecting "unnameable".

On average, responses were collected from 162 unique subjects for each dataset. Over all datasets, 4% of responses were rated "very subtle", 19% were "somewhat subtle", 44 were "somewhat obvious", and 32% were "obvious". Finally, only 3% of the images received obviousness scores with a standard deviation larger than 0.25 across subjects, indicating a strong agreement among subjects about the nameability of attributes. Figure 5 shows some example displays and names. Detailed instructions given to subjects, more attribute visualizations, and names collected can be found on the authors' webpages.

5. Results

Our results validate the components of our approach, and show its advantages over traditional methods for gathering attribute lists.

Active discriminative hypothesis generation. We first evaluate our approach to actively generate discriminative attributes in isolation from the rest of the system. Figure 6 compares our approach (Active) to baselines that randomly select two (Rand2), four (Rand4), or all classes (All8) on which to focus at each iteration. For compactness, we average results across all dataset/feature-space combinations, for 10 trials each; similar trends occur for individual cases. The curves clearly show that our approach identifies more discriminative features, successfully focusing the learner's attention on resolving remaining inter-class confusions.

Structure in nameability space. Next we evaluate if the space of nameable visual splits is structured, and whether that structure allows us to predict the nameability of a novel hypothesis. We fit an MPPCA model to each dataset and feature type using the "obvious" attributes identified by the human subjects, and then test the models with leave-one-out cross validation on both nameable and unnameable attribute instances. Figure 7 shows the results. Our approach predicts the nameability of novel attributes significantly better than chance for all four datasets and feature types (left chart). This confirms that the structure does exist. Moreover, we see that if we were to make attribute proposals based on these predictions, the user is more likely to accept our system's proposal, as compared to an approach that forgoes any nameability analysis (right chart). This indicates our approach can better utilize human annotation effort.

Discriminative and descriptive attributes. We now validate our entire interactive system, and compare to two baselines: one that searches for discriminative attributes, and one that uses a human-defined list of descriptive attributes. Note that these baselines represent traditional approaches for feature generation, and are also the two aspects that our algorithm is intended to balance.

First we compare our approach to the *purely discriminative* baseline, which proposes the most discriminative attribute hypothesis at each iteration without analyzing its nameability. For both our method and this "No nameability" baseline, if at any iteration a proposed attribute is in fact unnameable, it is not added to the vocabulary.

Figure 8 shows the results, averaged over 10 trials. We

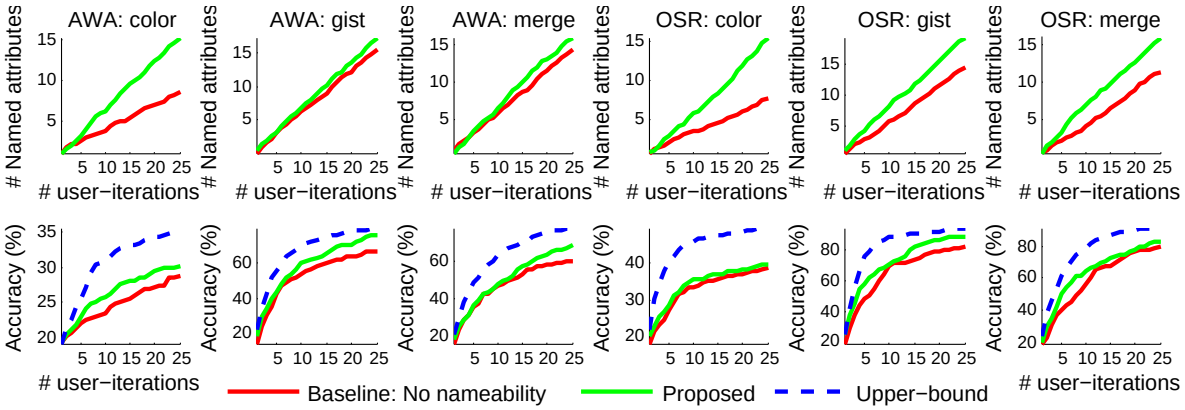

Figure 8. Our approach accumulates more descriptive/named attributes with the same amount of human effort as compared to a discriminative baseline that does not model nameability (top row), while at the same time maintaining better object/scene categorization accuracy for novel images (bottom row).

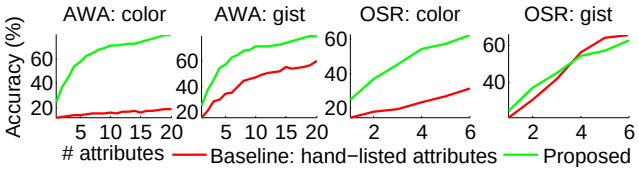

Figure 9. Attributes discovered by our approach tend to be more discriminative than those designated a priori with a purely descriptive list of words. Note, both methods shown produce an equal number of named attributes.

see that our approach yields more named attributes as a function of annotator effort (top row), while it also maintains strong performance on the novel test set due to those attributes' separability in the visual feature space (bottom row). Note that our method's stronger accuracy is not simply a given: we risk eliminating discriminative attributes during nameability analysis, whereas the baseline can select very discriminative attributes that happen to be nameable. For reference, we also show the "upper bound" on accuracy attainable if one were to simply use all actively discovered hyperplanes, whether named or un-named (dotted curves). This reflects the compromise an attribute-based recognition system makes in exchange for added descriptive power.

In addition to discovering attributes within a single feature space (i.e., Gist or color), the results also show that our system can seamlessly discover attributes from multiple feature spaces at once (see plots titled 'merge' in Figure 8). To implement this variant, we generate hypotheses by alternating among each feature space until we find one that our system predicts to be nameable. (Whether merged or separate, a nameability manifold is learned per feature type.)

Next we compare to a *purely descriptive* baseline that relies on a hand-generated list of attributes, as is typically done in previous work (e.g., [1, 2, 5, 8, 12]). At each iteration, we add one random attribute from this hand-generated list to the attribute vocabulary, and report results for 10 total trials. We draw on existing data for an objective and realistic source of the manually provided attributes.

For the AWA dataset, we use the attributes composed by cognitive scientists that are used by the dataset creators [2] (we take the 54 attributes in the list that are not constant across our 8 classes). The first two plots in Figure 9 show the results. For both feature spaces, the attributes discovered by our approach are more discriminative, leading to significantly more accurate predictions on the test set. This result highlights the advantage of considering separability in the visual feature space when defining nameable attributes, as is done automatically by our approach.

For the OSR dataset, we use the attributes listed in [27] (natural, open, perspective, size, diagonal plane, and depth), which were the properties they found their subjects used to organize the images. We asked a computer vision expert (outside from the authors of this paper) to assign class memberships for each of these attributes. The last two plots in Figure 9 show the results. Interestingly, while our approach discovers more discriminative attributes in the color space, the manually defined attributes perform better in the Gist space. This is expected, since Gist was explicitly designed to capture precisely these properties (see [27]). On both AWA and OSR, the trends substantiate our earlier claim that hand-generated lists of attributes may not be discriminative unless a precise match (via expert involvement) has been ensured between the attributes and the visual feature space. Our system allows users to discover discriminative as well as descriptive attributes in the most general setting, with minimal annotator effort.

Describing categories. A key characteristic of our approach is that it adapts to the user, who may provide any words for attributes that he/she finds relevant. Upon examining the nameable attributes in our MTurk collection, we find more than 50% of the attributes are shared across at least three of the eight categories in all datasets. A random sampling of the unique responses obtained from MTurk workers are as follows: **OSR color:** coastal, warmth, sharpness, brightness, slope, outdoor, snow, artificial, vegetative, seasons, natural, architecture, gray, evergreen, civilization; **OSR gist:** directional, rocks, serene, rural, paved, cold, educational establishments, brown, wildlife encounters, ge-

Figure 10. Descriptions of familiar and unseen (*e.g.* Zebra) categories generated by our discovered attributes (-G, -C = Gist, Color).

ometric, deep, aquatic life, close up, rocky; **AWA color:** dotted, large and black, attacking, mono-color, furry white, long-neck, lonely, human-like, gloomy, orange, smooth, hair; **AWA gist:** bright, big, thick fur, short-neck, hard, long-leg, fearful, white and black, rough, sad, endangered, climbs, unfriendly. The wide variety of responses our system can elicit is evident. On average, $\sim$ 12 out of 20 subjects provided the same word to describe the 50 most nameable (obvious) attributes, while only $\sim$ 6 subjects agreed on the word for the 50 "most subtle" attributes.

Figure 10 shows descriptions our system automatically generates for the categories of interest, as well as previously unseen categories. To select attributes to display, we require that they be positive for more than 90% of the images in the category. The text size corresponds to the portion of the 20 subjects that provided that particular word for the attribute.

6. Conclusions and Future Work

We introduced an interactive approach to discover a vocabulary of attributes that is both human understandable and discriminative—two characteristics that are required for attributes to be truly useful. Our approach actively identifies attribute candidates that not only resolve current confusions among classes, but are also likely to be human nameable. Results on multiple of datasets indicate its clear advantages over today's common practices in attribute-based recognition. Our novel nameability manifold bridges the gap between visual features and human language, a powerful concept potentially useful in a wide range of scenarios. Future work involves investigating a universal nameability space for standard visual features, considering non-binary or localized attributes, and alternative visualization techniques.

Acknowledgements: This research is supported in part by the Luce Foundation and NSF IIS-1065390.

References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by their Attributes. *CVPR*, 2009.
[2] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. *CVPR*, 2009.
[3] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych and B. Schiele. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer. *CVPR*, 2010.
[4] T. L. Berg, A. C. Berg and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. *ECCV*, 2010.
[5] N. Kumar, A. Berg, P. Belhumeur and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. *ICCV*, 2009.
[6] J. Wang, K. Markert and M. Everingham. Learning Models for Object Recognition from Natural Language Descriptions. *BMVC*, 2009.
[7] G. Wang and D. Forsyth. Joint Learning of Visual Attributes, Object Classes and Visual Saliency. *ICCV*, 2009.
[8] Y. Wang and G. Mori. A Discriminative Latent Model of Object Classes and Attributes. *ECCV*, 2010.
[9] M. Palatucci, D. Pomerleau, G. Hinton and T. Mitchell. Zero-Shot Learning with Semantic Output Codes. *NIPS*, 2009.
[10] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona and S. Belongie. Visual Recognition with Humans in the Loop. *ECCV*, 2010.
[11] V. Ferrari and A. Zisserman. Learning Visual Attributes. *NIPS*, 2007.
[12] O. Russakovsky and L. Fei-Fei. Attribute Learning in Large-scale Datasets. *Workshop on Parts and Attributes, ECCV*, 2010.
[13] A. Farhadi, I. Endres and D. Hoiem. Attribute-centric Recognition for Cross-category Generalization. *CVPR*, 2010.
[14] K. Yanai and K. Barnard. Image Region Entropy: A Measure of "Visualness" of Web Images Associated with One Concept. *ACM Multimedia*, 2005.
[15] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. *CVPR*, 2010.
[16] S. Vijayanarasimhan and K. Grauman. Multi-Level Active Prediction of Useful Image Annotations for Recognition. *NIPS*, 2008.
[17] A. Joshi, F. Porikli, and N. Papanikolopoulos. Breaking the Interactive Bottleneck in Multi-class Classification with Active Selection and Binary Feedback. *CVPR*, 2010.
[18] D. Mazzoni, K. L. Wagstaff and M. Burl. Active Learning with Irrelevant Examples. *ECML*, 2006.
[19] K. Zhang, I. W. Tsang and J. Kwok. Maximum Margin Clustering Made Practical. *ICML*, 2007.
[20] B. Zhao, J. Kwok, F. Wang and C. Zhang. Unsupervised Maximum Margin Feature Selection with Manifold Regularization. *CVPR*, 2009.
[21] M. Tipping and C. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11:443–482, 1999.
[22] B. A. Olshausen and D. J. Field. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381:607-609, 1996.
[23] Y. Weiss. Deriving Intrinsic Images from Image Sequences. *ICCV*, 2001.
[24] K. He, J. Sun and X. Tang. Single Image Haze Removal Using Dark Channel Prior. *CVPR*, 2009.
[25] R. Ando and T. Zhang. A High-Performance Semi-Supervised Learning Method for Text Chunking. *ACL*, 2005.
[26] P. Golland. Discriminative Direction for Kernel Classifiers. *NIPS*, 2001.
[27] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *IJCV*, 2001