# Discovering Shades of Attribute Meaning with the Crowd

Adriana Kovashka and Kristen Grauman

The University of Texas at Austin
Technical Report AI13-02
November 5, 2013

**Abstract.** To learn semantic attributes, existing methods typically train one discriminative model for each word in a vocabulary of nameable properties. This "one model per word" assumption is problematic: while a word might have a precise linguistic definition, it need not have a precise visual definition. We propose to discover *shades* of attribute meaning. Given an attribute name, we use crowdsourced image labels to discover the latent factors underlying how different annotators perceive the named concept. We show that structure in those latent factors helps reveal shades, that is, interpretations for the attribute shared by some group of annotators. Using these shades, we train classifiers to capture the primary (often subtle) variants of the attribute. The resulting models are both semantic and visually precise. By catering to users' interpretations, they improve attribute prediction accuracy on novel images.

## 1 Introduction

Attributes are semantic properties of objects and scenes. They can correspond to textures, materials, functional affordances, parts, moods, or other human-understandable aspects [1–5]. For instance, a scene can be *manmade*, or one shoe can be *more formal* than another. By injecting language into visual analysis, attributes broaden the visual recognition problem—from labeling images, to *describing* them. This linguistic interpretability opens up several interesting applications. For example, a user can search for an image by describing it [2, 6–9] train an object model by describing the category [3, 5], or help the system perform fine-grained recognition by naming the object's properties [10].

Typically one defines a vocabulary of attribute words relevant to the domain at hand (e.g., a vocabulary of facial characteristics for people search [2], textures and parts for animals [3, 11, 10], or clothing properties for shopping [12, 7]). Then one gathers labeled images depicting each attribute in the vocabulary, and trains a model to recognize each word [1–4, 6, 10, 13, 5, 7, 14].

The problem with this standard approach, however, is that there is often a gap between language and visual perception. In particular, *the words in an attribute vocabulary need not be visually precise.* An attribute word may connote multiple "shades" of meaning—whether due to polysemy, variable context-specific meanings, or differences in humans' perception. For instance, the attribute *open* can describe a door that's ajar, a fresh countryside scene, a peep-toe


**Fig. 1.** Our method uses the crowd to discover factors responsible for an attribute's presence, then learns predictive models based on those visual cues. For example, for the attribute *open*, our method will discover shades of *meaning*, e.g., peep-toed (*open* at toe) vs. slip-on (*open* at heel) vs. sandal-like (*open* at toe *and* heel), which are three visual definitions of openness. Since these shades are not coherent in terms of their global descriptors, they'd be difficult to discover using traditional image clustering.

high heel, or a backless clog.[1] Each shade is distinct and may require dramatically different visual cues to correctly capture. Thus, the standard approach of learning a single classifier for the attribute as a whole may break down.

Unfortunately, neither bottom-up attribute "discovery" nor relative attributes solve the problem. Unsupervised discovery methods detect clusters or splits in the low-level image descriptor space [15–20]. While they might discover finer-grained shades of *some* property, they need not be human-nameable (semantic). Furthermore, discovery methods are intrinsically biased by the choice of features. For example, the set of salient splits in color histogram space will be quite different than those discovered in a dense SIFT feature space. Similarly, unsupervised methods that cluster global descriptors have no way to intelligently focus on only localized regions of the image, yet an attribute may occupy an arbitrarily small part of an image.

Relative attributes [5] do not address the existence of shades, either. They represent whether an image has a property "more" or "less". The point in relative attributes is that people may agree best on *comparisons* or *strengths*, not binary labels. However, just like categorical attributes, *relative attributes assume that there is some single, common interpretation of the property shared consistently by all human viewers*—namely, that a single ordering of images from least to most [attribute] is possible. Thus, shades are relevant whether the attributes are modeled with classifiers (binary) or ranking functions (relative).

Our goal is to automatically discover the shades of an attribute. **An attribute "shade" is a visual interpretation of an attribute name that one or more people apply when judging whether that attribute is present in an image.**[2] See Figure 1.

Given a semantic attribute name, we want to discover its multiple visual interpretations and train a discriminative model for each one. Rather than attempt to manually enumerate the possible shades, we propose to learn them indirectly from the crowd. First we ask many annotators to label various images, reporting whether the attribute is present or not. Using their responses, we estimate

[1] Note multiple shades of an attribute may exist even within a specific object category (like shoes, in this example).

[2] Similarly, if learning relative attributes, a shade is an interpretation when judging whether that attribute is present more in image A or image B.

latent factors that represent the annotators in terms of the kinds of visual cues that they associate with the attribute. Then, clustering in the low-dimensional latent space, we identify the "schools of thought" (about how to interpret this attribute) underlying the discrete set of labels the annotators provided. Finally, we use the positive exemplars in each school to train a predictive model, which can then detect when the particular attribute shade is present in novel images.

The resulting models are both semantic and visually precise. By discovering the shades from the crowd's latent factors, we isolate the features corresponding to the perceived shades. This makes our method less susceptible to the more "obvious" splits in the feature space that an image clustering approach (including today's sophisticated discovery methods [15–20]) may find, which need not directly support the semantic attribute of interest. See Figure 1.

On two datasets, we find that not only are the discovered shades visually meaningful, they are also well-aligned with annotators' textual explanations of their labels. Most importantly, we show their practical utility to reliably estimate perceived attributes in novel images, which is crucial for any application relying on the descriptive nature of attributes (e.g., image search or zero-shot learning).

## 2 Related Work

*Learning attributes* Attributes [1, 3, 4] are nameable visual properties that can aid both classification [3, 4, 10, 13, 2, 5, 14] and image search [2, 6, 7]. Whether categorical or relative, prior work assumes that each attribute word corresponds to one coherent visual property, and so trains one classifier [1–4, 6, 10, 13, 14] or one ranking function [5, 7] per attribute.

Since annotators may disagree about the attribute label for an image [4, 21, 14, 22], the norm is to take the majority vote label (and discard the image if votes are too split). Thus, prior work treats differences in attribute perception as noise. To our knowledge, the only exception is the transfer learning approach of [23], which trains *user-specific* models for personalized image search. In contrast, we discover schools of thought among the crowd, and our method produces a set of attribute shades capturing commonly perceived variations. These schools of thought are a valuable midpoint on the spectrum from purely consensus models to purely user-specific models, resulting in better accuracy for perceived attributes (cf. Sec. 3.4). Shades also have broader utility than [23], since they let us explicitly organize perceived properties.

*Distinction with relative attributes* We stress that relative attributes [5, 24], while avoiding the need for forced categorical judgments, still assume a single underlying visual property exists. They do not represent multiple interpretations. Relative attributes construct a *universal* model for "less brown" vs. "more brown". They don't address the issue that one person may say "image X is *browner* than Y", while another may say the opposite. Shades, on the other hand, are concerned with discovering *multiple* models for varying perceptions of brown, e.g., chocolate brown vs. goldish brown. The two goals are orthogonal. In fact, while we study categorical attributes, our exact same algorithm could be applied to discover shades of relative attributes; the label matrix in Sec. 3.2 would simply

record whether the person finds a first image to exhibit the attribute more or less than a second image.

*Defining attribute vocabularies* Most work defines the attribute vocabulary manually, or by eliciting discriminative properties from annotators [14, 25]. However, in some cases it is possible to generate it (semi-)automatically [11, 10, 12, 15, 26]. For animal species, field guides are a natural source of attribute names [11, 10]. Given their focus on concrete parts, such domains are less prone to shades. When suitable text sources are available—such as captioned images on web pages [12] or activity scripts [26]—one can mine for candidate attribute words. Since not all words will be visually detectable, the authors of [12, 27] show how to prune the vocabulary automatically. Rather than mined text, our shades use sparse crowd labels to capture latent interpretations of an attribute, which may not even be concisely describable with a keyword.

*Discovering non-semantic attributes* While the term "attribute" typically connotes a *semantic* property, some researchers also use the term to refer to discovered *non-semantic* features [16, 18, 20, 19]. The idea is to identify "splits" or clusters in the low-level image descriptor space, often subject to constraints that deter redundancy and promote discriminativeness for object recognition. However, being bottom-up, there is no guarantee the splits will correspond to a nameable property. Hence, unlike our shades, they are non-semantic and inapplicable to descriptive attribute tasks, like image search or zero-shot learning. One can attempt to assign names to discovered "attributes" after the fact [15, 17, 20], but the patterns that are even discoverable remain biased by the chosen low-level image feature space, as discussed above.

*Polysemy and images* A polysemous word has multiple "senses" or meanings. Some work bridging text and visual analysis aims to cluster Web images according to distinct senses [28–31]. However, the focus is on nouns/object categories, not descriptive properties. Typically the visual differences (or surrounding text context) are fairly stark (e.g., a river *bank* or financial *bank*). In contrast, attribute shades are often subtle differences in interpretation. Furthermore, unlike a truly polysemous word, for which one can enumerate the multiple dictionary definitions, attribute shades are often more difficult to definitively express in language. We show how to automatically infer them from trends in crowd labels.

*Aggregating crowd labels* Crowd input has been aggregated in novel ways for image clustering [32], image similarity [33], and object labeling [34]. In [34], modeling annotators' competence and bias makes it possible to discover their schools of thought, and subsequently undo their biases to produce more reliable ground truth. While that work aims to recover a single true label for each image, our goal is to discover the crowd's multiple interpretations of a label.

Matrix factorization is often used for matrix completion, to solve collaborative filtering problems (e.g., the Netflix challenge) by exploiting commonalities among users [35, 36]. Rather than impute missing labels, we propose to use the latent factors themselves to represent the interplay between language, human perception, and image examples. We show how to use the recovered schools of thought to build content-based attribute models.

| Attribute | Dictionary definition |
|---|---|
| Pointy | having a comparatively sharp point, or having numerous pointed parts |
| Open | having interspersed gaps, spaces, or intervals |
| Ornate | made in an intricate shape or decorated with complex patterns |
| Comfortable | providing physical comfort, ease and relaxation |
| Formal | designed for wear or use at elaborate ceremonial or social events |
| Brown | the color of, for example, chocolate and coffee |
| Fashionable | conforming to the current fashion; stylish; trendy; modern |
| To clutter | to make disorderly or hard to use by filling or covering with objects |
| To soothe | to bring comfort, composure, or relief |
| Open (area) | affording unobstructed passage or view |
| Modern | characteristic or expressive of recent times or the present; contemporary |
| Rustic | of, relating to, or typical of country life or country people |

**Table 1.** The 12 attribute definitions shown to annotators.

## 3 Approach

We first explain our crowdsourced label collection in Sec. 3.1. Then we describe how we recover the latent factors responsible for those labels (Sec. 3.2) and use them to discover attribute shades (Sec. 3.3). Finally, we exploit the discovered shades to improve attribute prediction by accounting for the users' varying interpretations (Sec. 3.4).

### 3.1 Collecting crowd labels per attribute

We use two datasets: Shoes [12] and SUN Attributes [14]. We use 2559 and 2086 total images from Shoes and SUN, respectively. While attribute labels are available for both, our method needs to record which annotator labeled which image. Therefore, we run our own crowdsourced label collection.

To focus our study on plausibly "shaded" words, we select 12 attributes that can be defined concisely in language, yet may vary in their visual instantiations. This helps ensure that variance in the annotators' labels stems from the attribute's visual sub-meanings, as opposed to external factors like the annotator's personal taste. The 12 attributes are: "pointy", "open", "ornate", "comfortable", "formal", "fashionable", "brown" (for Shoes); and "cluttered", "soothing", "open area", "modern", "rustic" (for SUN). We sample $N = 250$ to 1000 images per attribute. To get representative images spanning the dataset, we cluster all images using $K$-means, then sample ones near the cluster centers.

We build a Mechanical Turk interface to gather the labels. Workers are shown definitions of the attributes from a web dictionary (see Table 1), but no example images. Thus, they all receive the same linguistic definition, but they are not prompted with any particular *visual* definition. Then, given an image, the worker must state whether it does or does not possess a specified attribute. Additionally, for a random set of 5 images, the worker must explain his label in free-form text, and state which image most has the attribute, and why. These questions both slow the worker down, helping quality control, and also provide valuable ground truth data for evaluation, as we will explain in Sec. 4.2.

Our latent factor model (defined next) can accommodate imbalanced and sparse labels. This is good, because in realistic scenarios, labels may not originate from concentrated one-time labeling efforts (like ours), but rather as a side product of another task—such as click data in image search. In such a case, the images that one user labels will not entirely overlap with those that another user labels. Furthermore, each user will label few examples. To mimic this scenario, we gather labels in a sparse fashion. Each worker labels 50 randomly chosen images, per attribute. To help ensure self-consistency in the labels, we exclude workers who fail to consistently answer 3 repeated questions sprinkled among the 50. This yields annotations from 195 workers per attribute on average.

While multiple workers may label the same image, we stress their labels are *not* aggregated to create a majority vote "ground truth". The main premise of our work is that attribute names can be visually imprecise and so admit multiple interpretations. The same attribute word can have different meanings to different people, even if they all know the same linguistic definition of the word. (Contrast this with object category names, which are relatively precise.) Thus, rather than discard label discrepancies as noise, we use them to discover shades.

### 3.2 Recovering latent factors for attribute labels

Now we use the label data to discover latent factors, which are needed to recover the shades of meaning. Note that we learn factors for each attribute independently, so all variables below are attribute-specific. From the above data collection, we retain each worker's ID, the indices of images he labeled, and how he labeled them. Let $M$ denote the number of unique annotators, and let $N$ denote the number of images seen by at least one annotator. Let $\mathbf{L}$ be the $M \times N$ label matrix, where $L_{ij} \in \{0, 1, ?\}$ is a binary attribute label for image $j$ by annotator $i$. A ? denotes an unlabeled example. The matrix is only partially observed, as on average only 20% of the possible image-worker pairs are labeled.

We suppose there is a small number $D$ of unobserved factors that influence the annotators' labels. This reflects that their decisions are driven by some mid-level visual cues. For example, when deciding whether a shoe looks "ornate", the latent factors might include presence of buckles, amount of patterned textures, material type, color, and heel height; when deciding whether a scene looks "modern", they might include color, object composition, and materials.

Assuming a linear factor model, the label matrix $\mathbf{L}$ can be factored as the product of an $M \times D$ annotator latent factor matrix $\mathbf{A}^T$ and a $D \times N$ image latent factor matrix $\mathbf{I}$: $\mathbf{L} = \mathbf{A}^T\mathbf{I}$. A number of existing methods can be used to factor this partially observed matrix, by finding the best rank-$D$ approximation under some loss function [37, 35, 36]. We use a probabilistic matrix factorization algorithm (PMF) [37, 35], due to its efficiency for large, sparse matrices. Briefly, it works as follows. PMF takes a probabilistic approach to recover the two low-rank matrices. The likelihood distribution for the observed labels is

$$p(\mathbf{L}|\mathbf{A}, \mathbf{I}, \sigma^2) = \prod_{i=1}^{M} \prod_{j=1}^{N} \left[\mathcal{N}(L_{ij}|A_i^T I_j, \sigma^2)\right]^{\ell_{ij}}, \quad (1)$$

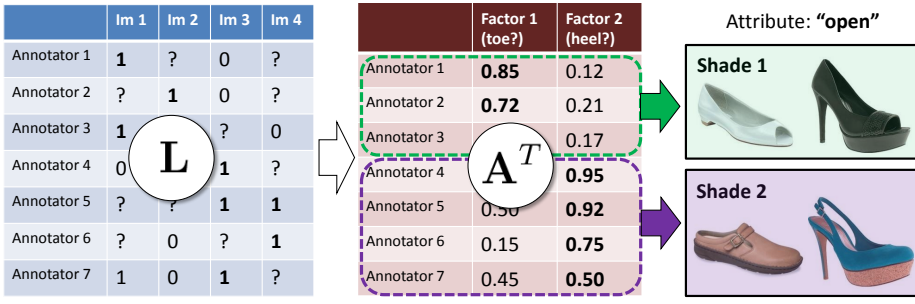


**Fig. 2.** Given a partially observed attribute-specific label matrix (left), we recover its latent factors and their influence on each annotator (middle). We discover shades by clustering in this space (right).

where $A_i$ and $I_j$ denote columns of $\mathbf{A}$ and $\mathbf{I}$, respectively, and $\ell_{ij} = 1$ if we received a label on image $j$ by annotator $i$, and $\ell_{ij} = 0$ otherwise. $\mathcal{N}(x|\mu, \sigma^2)$ denotes a Gaussian distribution with mean $\mu$ and standard deviation $\sigma^2$. The priors over the latent factors are spherical Gaussians, $p(\mathbf{A}|\sigma_A^2) = \prod_{i=1}^{M} \mathcal{N}(A_i|0, \sigma_A^2 \mathbb{I})$ and $p(\mathbf{I}|\sigma_I^2) = \prod_{j=1}^{N} \mathcal{N}(I_j|0, \sigma_I^2 \mathbb{I})$.

We seek the latent features that maximize the log-posterior:

$$\mathbf{A}^*, \mathbf{I}^* = \arg\max_{\mathbf{A},\mathbf{I}} \ln p(\mathbf{A}, \mathbf{I}|\mathbf{L}, \sigma^2, \sigma_A^2, \sigma_I^2). \quad (2)$$

Obtaining the MAP factors amounts to minimizing an SSD objective function with quadratic regularization terms using gradient descent [37]; this yields a probabilistic extension of what would be standard SVD in the case of fully observed labels. Upgrading to a full Bayesian treatment [35], we put priors on the hyperparameters $\sigma^2$, $\sigma_A^2$, $\sigma_I^2$ and obtain a predictive distribution for the latent factors, using MCMC to sample the latent feature matrices in parallel. This reduces overfitting and saves parameter tuning. See [35] for details.

### 3.3 Discovering shades of meaning

In collaborative filtering, the goal of such a factorization is to impute missing labels (e.g., to predict how a user will rate an unseen movie, $L_{ij} \approx \langle A_i, I_j \rangle$). While missing labels could similarly be estimated for our data, our goal is different. We aim to discover attribute shades of interpretation and generate predictive visual models for them.

To this end, we first represent each annotator in terms of his association with each discovered factor. The "latent feature vector" for annotator $i$ is $A_i \in \Re^D$, the $i$-th column of $\mathbf{A}$. It represents how much each of the $D$ factors influences that annotator when he decides if the named attribute is present. Likewise, the latent feature for image $j$ is $I_j \in \Re^D$, the $j$-th column of $\mathbf{I}$, and represents how much each of the $D$ factors is visible in the image.

Figure 2 illustrates with a cartoon example. As seen on the left, annotators did not label all images for the attribute "open". Some tended to label images

1 and 2 as having the attribute, whereas others tended to label 3 and 4 as positive. After factoring the label matrix, suppose we discover $D = 2$ latent factors. Though nameless, they align with semantic visual cues; suppose here they are "toe is open" and "heel is open". Each annotator's feature $A_i$ encodes how important those two factors were for his label decision. In this hypothetical example, we see the first three annotators labeled images 1 and 2 as open due to factor 1, whereas the others focused on factor 2 in other images.

We pose shade discovery as a grouping problem in the space of these latent features.[3] While other clustering algorithms could be used, we apply $K$-means to the columns of $\mathbf{A}$ to obtain clusters $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$.[4] Each cluster is a shade. Annotators in the same cluster display similar labeling behavior, meaning they interpret similar combinations of mid-level visual cues as salient for the attribute at hand. For example, in Fig. 2, the two dominant shades reflect which part of the shoe the annotator focused on to judge openness—toe or heel. (Of course, for real data, there will be $D > 2$ factors, and shades will combine many such factors.)

Recall that shade discovery is done on a per-attribute basis. Depending on the visual precision of the word, some attributes may have only one shade; others may have many. To automatically select $K$ based on the structure of the data, we use the silhouette coefficient [38]. It quantifies how tightly grouped all the latent features in a cluster are, averaged across all clusters, in terms of the average distance of an instance to all other data in the same cluster versus its distance to a neighboring cluster.

### 3.4 Using shades to predict perceived attributes

A key valuable application of shades is to improve attribute prediction accuracy, generalizing what the system discovered to novel images. Prior work uses one of two extremes for attribute prediction—either 1) a *consensus* classifier: a single model trained with majority vote labeled examples (e.g., [2–4, 6, 14]), or 2) a *user-specific* classifier trained by adapting that majority vote model to satisfy an individual user's training labels [23].

We propose a solution in between these two extremes. With shades, we can account for the fact that people perceive an attribute differently, yet avoid specializing predictions down to the level of each individual user. The idea is to tailor an attribute classifier according to the user's "school of thought", i.e., the shade to which he subscribes.

To this end, we train shade-specific classifiers that adapt the consensus model. Each shade $\mathcal{S}_k$ is represented by the total pool of images that its annotators labeled as positive. Several annotators in the cluster may have labeled the same image, and their labels need not agree. Thus, we perform majority vote (over

[3] While we can cluster either annotators or images to identify shades, we choose annotators in order to facilitate ground truth evaluation in Sec. 4.

[4] Preliminary tests with Bayesian non-parametric clustering showed inferior results. An alternative would be to impute missing labels and group with EM, but clustering in the compact latent space is preferable when labels are very sparse.

just the annotators in $\mathcal{S}_k$) to decide whether an image is positive or negative for the shade. For both our shade models and the consensus model, we discard labels where fewer than 90% of users agree. We use the images to train a discriminative classifier, using the Adapt-SVM objective [39] to regularize its parameters to be similar to those of the consensus model. (See Sec. 4.1 for details.) Then we apply the adapted shade model for the cluster to which a user belongs to predict the presence/absence of the attribute in novel images. Thus, the predictions are automatically tailored to that user's perception of the property.

To recap, shades offer an important midpoint on the spectrum discussed above. Compared to the standard consensus approach, we account for distinct perceived shades. Compared to user-specific models [23], the advantages are twofold. First, each model typically leverages more training data than a single user provides. This lets us effectively "borrow" labeled instances from the user's neighbors in the crowd. Second, we leverage the robustness of the intra-shade majority vote. This helps reduce noise in an individual user's labeling. The results in Sec. 4.1 reveal the impact of these advantages in practice.

Note, a user must provide at least some attribute labels to benefit from the shade models, since we need to know which shade to apply. For users who contributed to the label matrix $\mathbf{L}$ this is straightforward. For users adding labels later, we could either re-factor $\mathbf{L}$, or use a folding-in heuristic [40] (not attempted in our experiments).

### 3.5 Discussion

The key thing to note about the shade classifiers is how their positive labeled exemplars came about. Images within a shade can be visually diverse from the point of view of typical global image descriptors, since annotators attuned to that shade's latent factors could have focused on arbitrarily small parts of the images, or arbitrary subsets of feature modalities (color, shape, texture). An approach that attempts to discover shades based on image clustering—or non-semantic attribute discovery [15–20]—will be hard pressed to group images according to these perceived, possibly subtle, cues. Our insight is to leverage patterns among the crowd labels to partition the images *semantically*. Then, even though the training images may be visually diverse, standard discriminative learning methods let us isolate the informative features. Essentially, we avoid biasing the shades to a particular low-level descriptor space, since their training images are determined independent of the descriptors.

One might wonder: why not just manually enumerate the attribute shades with words? Our approach has multiple advantages over that strategy, beyond being automatic. For polysemous *nouns*, the visual definitions are enumerable—check the dictionary. In contrast, it can be difficult to put an attribute's distinct visual instantiations in words. Furthermore, the words annotators typically provide are concrete instances of the shade, which need not comprehensively define the shade. For example, in our data collection, when asked to explain why an image is "ornamented", an annotator might comment on the "buckle" or "bow"; yet the latent shade of "ornamented" underlying many users' labels is more abstract. It encompasses combinations of such concrete mid-level cues. In short, we

| Attribute | Present? | Explanation |
|---|---|---|
| Open | Yes | This shoe is somewhat open because the heel of the foot is exposed rather than covered. |
| Comfortable | No | Heels are too thin and high. May cause heel/back pain. |
| Cluttered | No | All of the objects in this room appear to be in an orderly fashion, with nothing out of place, leading me to feel that this space is not cluttered. |
| Soothing | Yes | It looks like a cottage someone would go to for relaxation. The pretty plant life in the background are also soothing because they depict a serene type of nature. |

**Table 2.** Example label explanations that annotators provided.

find that people are good at naming examples, but less good at characterizing an entire shade in words. Our method fills that gap, using structure in the labels to identify shades.

## 4 Results

We first demonstrate shades' key utility for improving attribute prediction (Sec. 4.1). We then quantitatively analyze the purity of the discovered shades (Sec. 4.2). We offer comparisons to existing techniques, including both standard consensus attributes as well as state-of-the-art methods for attribute discovery [18] and personalized attributes [23]. Finally, we analyze the shades qualitatively (Sec. 4.3) to visualize what is discovered.

*Implementation details* We use provided image descriptors for all methods: concatenated GIST and Lab color histograms for Shoes, and GIST, color, HOG, and self-similarity histograms for SUN. See [5, 14] for details. We use the Bayesian PMF implementation of [36]. We fix $D = 50$, then use the default parameter settings. For $N = 1000$ and $M = 195$, MCMC with 500 samples takes about 21 minutes. We cross-validate all classifier parameters. We set $K$ automatically per attribute based on the optimal silhouette coefficient within $K = \{2, \ldots, 15\}$. Typically values of $K \approx 7$ are chosen by the algorithm.

As noted in Sec. 3.1, during data collection annotators must explain their attribute labels. Specifically, we ask, "Please explain your response. What part or aspect of the image do you associate with the attribute [attribute name]? What part or aspect of the image led you to say that the attribute [attribute name] is present or not present?" Table 2 shows a sample of their responses (and see Supp for additional examples). We draw on their explanations below to aid our quantitative evaluation, but they are never seen by our method.

### 4.1 Accuracy of perceived shade predictions

We begin with a key result demonstrating how well shades capture perceived attributes. We apply the shades as described in Sec. 3.4 to predict user-specific labels. We compare to three methods: (1) Majority vote, which is the standard consensus approach [1–4, 6, 10, 13, 14], (2) User-specific, which trains one attribute classifier per user using only his labeled images, and (3) User-adaptive,

| Attribute | Shades | Majority | User-spec. | Adapt [23] |
|---|---|---|---|---|
| Pointy | **76.3** (0.3) | 74.0 (0.4) | 67.8 (0.2) | 74.8 (0.3) |
| Open | **74.6** (0.4) | 66.5 (0.5) | 65.8 (0.2) | 71.6 (0.3) |
| Ornate | **62.8** (0.7) | 56.4 (1.1) | 59.6 (0.5) | 61.1 (0.6) |
| Comfort. | **77.3** (0.6) | 75.0 (0.7) | 68.7 (0.5) | 75.5 (0.6) |
| Formal | **78.8** (0.5) | 76.2 (0.7) | 69.6 (0.4) | 77.1 (0.4) |
| Brown | **70.9** (1.0) | 69.5 (1.2) | 61.9 (0.5) | 68.5 (0.9) |
| Fashion. | **62.2** (0.9) | 58.5 (1.4) | 60.5 (1.3) | 62.0 (1.4) |
| Cluttered | **64.5** (0.3) | 60.5 (0.5) | 58.8 (0.2) | 63.1 (0.4) |
| Soothing | **62.5** (0.4) | 61.0 (0.5) | 55.2 (0.2) | 61.5 (0.4) |
| Open area | **64.6** (0.6) | 62.9 (1.0) | 57.9 (0.4) | 63.5 (0.5) |
| Modern | **57.3** (0.8) | 51.2 (0.9) | 56.2 (0.7) | 56.2 (1.1) |
| Rustic | **67.4** (0.6) | 66.7 (0.5) | 63.4 (0.5) | 67.0 (0.5) |

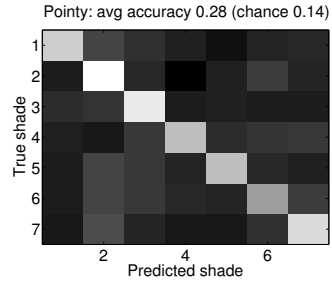


**Fig. 3.** (a) Accuracy of predicting perceived attributes, with standard error in parens. (b) Confusion matrix for multi-way shade classification, for the attribute "pointy".

the transfer method of [23], which adapts the majority vote model with the same user-specific labeled data. All methods use linear SVMs (for consistency with [23][5]). Our method selects $K$ automatically per attribute, yielding values between 5 and 10. We run 30 trials, sampling 20% of the available labels to obtain on average 10 labels per user (representing what a user might reasonably contribute to train the system).

Figure 3 (a) shows the results. Our method outperforms all other methods. It is more reliable than MAJORITY VOTE, which is the status quo attribute learning approach. For "open", we achieve an 8 point gain over MAJORITY VOTE and USER-SPECIFIC, which indicates both how different user perceptions of this attribute are, as well as how useful it is to rely on schools rather than individual users. We also outperform [23], while requiring the exact same labeling effort. While their method learns personalized models, shades leverage *common perceptions* and thereby avoid overfitting to a user's few labeled instances. We also tested models for shades obtained by the baselines defined below in Sec. 4.2, but they were not competitive with our result.

While Fig. 3 (a) measures binary attribute classification, our method can also perform multi-way shade classification. Here we cluster in the latent feature space of the images $I_j$, and again automatically select $K$. Fig. 3 (b) shows a representative resulting confusion matrix for the attribute "pointy" (see Supp for all matrices). Our average multi-way accuracy over all attributes is 0.28, much better than chance (0.15 on average). This result indicates that the discovered shades per attribute are indeed distinct and detectable. Note, the other baselines are not relevant for this task, since they do not group images into multiple sub-attributes.

These results clearly demonstrate the utility of shades. For all attributes, mapping a person's use of an attribute to a shade allows us to *predict attribute presence more accurately*. This is achieved at no additional expense for each user. As a result, applications demanding descriptive attributes (e.g., image search, zero-shot learning, etc.) benefit from the more accurate representation.

[5] Results are similar for our method and the other baselines using non-linear RBF kernels, but the authors of [23] did not have a kernelized implementation available.

### 4.2 Quantifying the accuracy of shade formation

To further quantify how accurately our shades capture perceived interpretations, we next score how *coherent* the textual explanations (cf. Table 2) are among annotators in the same shade. Whereas random clusters would group diverse ground truth explanations together, good shades should align with coherent explanations. We stress that these explanations are never seen by our algorithm; they are for evaluation purposes only.

To measure coherency, we first perform probabilistic Latent Semantic Analysis (pLSA) [40] on the Porter-stemmed textual descriptions. We treat each description for which $L_{ij} = 1$ as a document and discover $T = 200$ topics with pLSA. Then we map each explanation to its distribution of topics (a vector of $T$ weights). This representation accounts for word meaning, not just word occurrences (e.g., "image" and "picture" will be treated as synonyms by pLSA). Next, we average the topics for all positive descriptions originating from annotators in a given shade, yielding one topic vector per shade. Finally, we score the quality of a shade by its topic distribution entropy. Low entropy is better, as it indicates that the shade corresponds to a more coherent set of descriptions.

We compare to two methods: (1) ATTRIBUTE DISCOVERY: a state-of-the-art non-semantic attribute discovery method [18], and (2) IMAGE CLUSTERS: an image clustering approach inspired by [29]. The first baseline discovers splits in the feature space that are discriminative for object categories. We use the code kindly provided by the authors; we train it with the 10 Shoe and 611 SUN categories in the training images used by our method. We then cluster images in the discovered attribute space. (We also tried using [18] with the human-labeled attributes as "categories", but it performed significantly worse.) The second baseline is inspired by prior work for discovering word "senses" [29]; it clusters the image descriptors for all images labeled positive by at least one annotator. For both, to map an image cluster to ground truth descriptions, we look at the bag of images each annotator labeled as positive, find the image cluster to which the largest portion of the bag belongs, and assign it to be this user's shade ID. All methods use $K$-means and remove clusters with fewer than 10 members, which tend to be too sparse to form a meaningful shade.

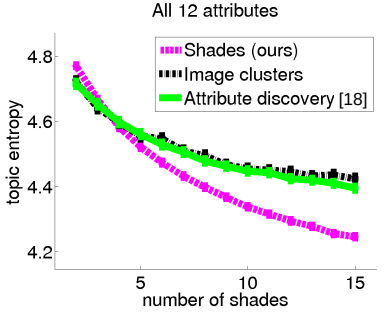


**Fig. 4.** Quality of discovered shades (low entropy is good)

Figure 4 shows the results. We plot topic entropy (and standard error) as a function of the number of shades $K$, over all attributes and 30 runs. Our shades are much more coherent overall. Clearly, image clustering falls short. The non-semantic attribute discovery method [18], while stronger than clustering, does not capture the shades of meaning since it lacks human input on the attribute interpretation. When $K = 2$, the baselines have lower entropy than our shades, showing that very coarse groups are sufficiently


**Fig. 5.** Top words and images for two shades per attribute (top and bottom for each attribute). Best viewed on PDF or in color. See text for description.

found with image clustering; however, these clusters are too coarse according to the silhouette coefficient model selection, which selects $K = 5$ to $K = 10$ shades as the optimal setting. This result indicates that the shades we have discovered are meaningful and accurately capture the varied attribute meanings that human users employ.

### 4.3 Visualizing attribute shades of meaning

Finally, we provide qualitative results. Figure 5 visualizes two shades each, for nine of the attributes. The images are those most frequently labeled as positive by annotators in a shade $\mathcal{S}_k$. The (stemmed) words are those that appear most frequently in the annotator explanations (cf. Table 2) for that shade, after we remove words that overlap between the two. Font size reflects relative frequency. To aid readability, we also outline words that stand out as good representatives of the shade.

We see the shades capture nuanced visual sub-definitions of the attribute words. For example, for the attribute "brown", one shade covers chocolate-colored shoes (top shade), while another is lighter and more gold (bottom shade). For "ornate", one shade focuses on straps/buckles (top), while another focuses on texture/print/patterns (bottom). For "comfortable", one shade emphasizes a low arch (top), while the other requires soft materials (bottom). For "open", one shade includes open-heeled shoes, while another includes sandals which are open at the front *and* back. In SUN, the "open areas" attribute can be either outside (top) or inside (bottom). For "soothing", one shade emphasizes scenes conducive to relaxing activities, while another focuses on aesthetics of the scene.

As discussed above, an important feature of our method is its ability to perform discovery independent of a particular image descriptor. To illustrate

this, we next use the shades' visual classifiers to examine their most informative *localized* features. We use $L_1$ regularization when training one-vs.-rest logistic regression classifiers for each shade, in order to isolate a sparse set of features most discriminative for that shade. For each $70 \times 70$ grid cell of the image, we sum the magnitude of the classifier weights for its features. Then we multiply those weights with the pixel intensities in order to visualize the relative impact of each portion of the image.

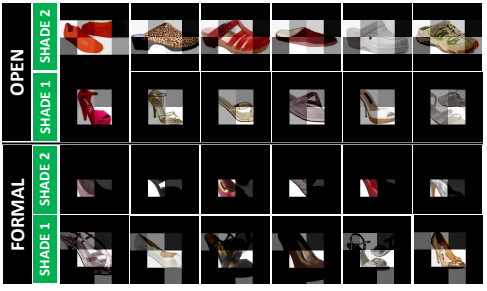


**Fig. 6.** Image regions highlighted according to the importance of the localized features for learning the shades.

Figure 6 shows example results. Brighter cells indicate regions more discriminative for that shade. For "open", we see one shade emphasizes openness at the back, and another openness at the toe. For "formal", the top shade emphasizes the arch of the shoe, while the bottom one emphasizes the toes. Such examples illustrate how our method isolates visual properties that support a shade, yet would not be tightly grouped if simply clustering global descriptors.

Of course, learning discriminative spatially localized features is nothing new; our point is that shades are what enable the training image groups that make this discriminative selection feasible. Furthermore, recent work using crowds to isolate informative spatial regions [41, 42] has a different purpose (fine-grained image classification) and takes an entirely different approach (explicitly asking labelers to outline the regions needed to make their label decisions).

## 5 Conclusion

Our work addresses the gap between how people *describe* attributes and how they *perceive* them visually. We show how to discover people's shared biases in perception, then exploit them with visual classifiers that can generalize to new images. Our approach to discover attribute shades brings together language, crowdsourcing, human perception, and visual representations in a new way.

The learned shades successfully tailor attribute predictions to cater to a user's "school of thought", boosting the accuracy of detecting perceived attributes. In systematic experiments, we quantify the impact of shades, both compared to standard paradigms and multiple state-of-the-art methods. The visualized shades show great promise to separate the (sub-)attributes involved in a person's use of an attribute vocabulary during search or organization of image content.

In future work, we will investigate ways to model dependencies between multiple attributes and their shades, and to predict a person's preferred shade based on a minimal set of label requests.

## References


1. Ferrari, V., Zisserman, A.: Learning Visual Attributes. In: NIPS. (2007)
2. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Describable Visual Attributes for Face Verification and Image Search. PAMI (2011)
3. Lampert, C., Nickisch, H., Harmeling, S.: Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In: CVPR. (2009)
4. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing Objects by Their Attributes. In: CVPR. (2009)
5. Parikh, D., Grauman, K.: Relative Attributes. In: ICCV. (2011)
6. Vaquero, D., Feris, R., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-based People Search in Surveillance Environments. In: WACV. (2009)
7. Kovashka, A., Parikh, D., Grauman, K.: WhittleSearch: Image Search with Relative Attribute Feedback. In: CVPR. (2012)
8. Siddiquie, B., Feris, R., Davis, L.: Image Ranking and Retrieval Based on Multi-Attribute Queries. In: CVPR. (2011)
9. Scheirer, W., Kumar, N., Belhumeur, P., Boult, T.: Multi-Attribute Spaces: Calibration for Attribute Fusion and Similarity Search. In: CVPR. (2012)
10. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual Recognition with Humans in the Loop. In: ECCV. (2010)
11. Wang, J., Markert, K., Everingham, M.: Learning Models for Object Recognition from Natural Language Descriptions. In: BMVC. (2009)
12. Berg, T.L., Berg, A.C., Shih, J.: Automatic Attribute Discovery and Characterization from Noisy Web Data. In: ECCV. (2010)
13. Wang, Y., Mori, G.: A Discriminative Latent Model of Object Classes and Attributes. In: ECCV. (2010)
14. Patterson, G., Hays, J.: SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In: CVPR. (2012)
15. Parikh, D., Grauman, K.: Interactively Building a Discriminative Vocabulary of Nameable Attributes. In: CVPR. (2011)
16. Mahajan, D., Sellamanickam, S., Nair, V.: A Joint Learning Framework for Attribute Models and Object Descriptions. In: ICCV. (2011)
17. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering Localized Attributes for Fine-grained Recognition. In: CVPR. (2012)
18. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute Discovery via Predictable Discriminative Binary Codes. In: ECCV. (2012)
19. Sharmanska, V., Quadrianto, N., Lampert, C.: Augmented Attribute Representations. In: ECCV. (2012)
20. Yu, F., Cao, L., Feris, R., Smith, J., Chang, S.F.: Designing Category-Level Attributes for Discriminative Visual Recognition. In: CVPR. (2013)
21. Endres, I., Farhadi, A., Hoiem, D., Forsyth, D.: Benefits and Challenges of Collecting Richer Object Annotations. In: ACVHL. (2010)
22. Curran, W., Moore, T., Kulesza, T., Wong, W.K., Todorovic, S., Stumpf, S., White, R., Burnett, M.: Towards Recognizing "Cool": Can End Users Help Computer Vision Recognize Subjective Attributes or Objects in Images? In: IUI. (2012)
23. Kovashka, A., Grauman, K.: Attribute Adaptation for Personalized Image Search. In: ICCV. (2013)
24. Shrivastava, A., Singh, S., Gupta, A.: Constrained Semi-Supervised Learning using Attributes and Comparative Attributes. In: ECCV. (2012)

25. Maji, S.: Discovering a Lexicon of Parts and Attributes. In: ECCV Workshop on Parts and Attributes. (2012)
26. Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., Schiele, B.: Script Data for Attribute-based Recognition of Composite Activities. In: ECCV. (2012)
27. Barnard, K., Yanai, K.: Mutual Information of Words and Pictures. In: Information Theory and Applications Inaugural Workshop. (2006)
28. Barnard, K., Yanai, K., Johnson, M., Gabbur, P.: Cross Modal Disambiguation. Toward Category-Level Object Recognition (2006)
29. Loeff, N., Alm, C., Forsyth, D.: Discriminating Image Senses by Clustering with Multimodal Features. In: ACL. (2006)
30. Saenko, K., Darrell, T.: Unsupervised Learning of Visual Sense Models for Polysemous Words. In: NIPS. (2008)
31. Berg, T.L., Forsyth, D.A.: Animals on the Web. In: CVPR. (2006)
32. Gomes, R., Welinder, P., Krause, A., Perona, P.: Crowdclustering. In: NIPS. (2011)
33. Tamuz, O., Liu, C., Belongie, S., Shamir, O., Kalai, A.: Adaptively Learning the Crowd Kernel. In: ICML. (2011)
34. Welinder, P., Branson, S., Belongie, S., Perona, P.: The Multidimensional Wisdom of Crowds. In: NIPS. (2010)
35. Salakhutdinov, R., Mnih, A.: Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. In: ICML. (2008)
36. Xiong, L., Chen, X., Huang, T.K., Schneider, J., Garbonell, J.: Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization. In: SDM. (2010)
37. Salakhutdinov, R., Mnih, A.: Probabilistic Matrix Factorization. In: NIPS. (2007)
38. Rousseeuw, P.: Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics **20** (1987) 53–65
39. Yang, J., Yan, R., Hauptmann, A.G.: Adapting SVM Classifiers to Data with Shifted Distributions. In: ICDM Workshops. (2007)
40. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: UAI. (1999)
41. Donahue, J., Grauman, K.: Annotator Rationales for Visual Recognition. In: ICCV. (2011)
42. Deng, J., Krause, J., Fei-Fei, L.: Fine-Grained Crowdsourcing for Fine-Grained Recognition. In: CVPR. (2013)