Invited paper to appear in the Proceedings of AAAI 2012, Sub-Area Spotlights Track for Best Papers.

Relative Attributes for Enhanced Human-Machine Communication

Devi Parikh[1], **Adriana Kovashka**[3], **Amar Parkash**[2], **and Kristen Grauman**[3]
[1] Toyota Technological Institute, Chicago
[2] Indraprastha Institute of Information Technology, Delhi
[3] University of Texas, Austin

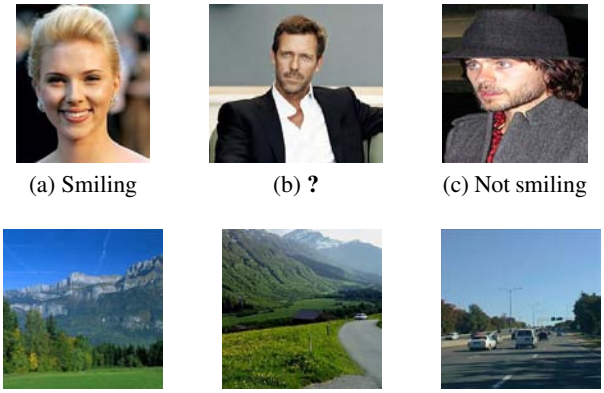
Abstract

We propose to model *relative attributes*[1] that capture the relationships between images and objects in terms of human-nameable visual properties. For example, the models can capture that animal A is 'furrier' than animal B, or image X is 'brighter' than image B. Given training data stating how object/scene categories relate according to different attributes, we learn a ranking function per attribute. The learned ranking functions predict the relative strength of each property in novel images. We show how these relative attribute predictions enable a variety of novel applications, including zero-shot learning from relative comparisons, automatic image description, image search with interactive feedback, and active learning of discriminative classifiers. We overview results demonstrating these applications with images of faces and natural scenes. Overall, we find that relative attributes enhance the precision of communication between humans and computer vision algorithms, providing the richer language needed to fluidly "teach" a system about visual concepts.

Introduction

Traditional visual recognition approaches map low-level image features directly to object category labels. Recent work proposes models using *visual attributes*. Attributes are properties observable in images that have human-designated names and are typically shared across object categories (e.g., 'striped', 'four-legged'). They are valuable as semantic cues in various problems. Researchers have shown their impact for strengthening facial verification (Kumar et al. 2009), object recognition (Wang, Markert, and Everingham 2009; Wang and Mori 2010; Branson et al. 2010), generating descriptions of unfamiliar objects (Farhadi et al. 2009), and to facilitate "zero-shot" transfer learning (Lampert, Nickisch, and Harmeling 2009).

Problem: Most existing work focuses wholly on attributes as binary predicates indicating the presence (or absence) of a certain property in an image. This may suffice for part-based attributes (e.g., 'has a head') and some binary properties (e.g., 'spotted'). However, for many attributes, not only is this binary setting restrictive, but it is also unnatural. For instance, it is not clear if in Figure 1(b) Hugh Laurie is smiling or not; people are likely to respond inconsistently in providing the presence or absence of the 'smiling' attribute for this image, or of the 'natural' attribute for Figure 1(e).



(a) Smiling (b) **?** (c) Not smiling

(d) Natural (e) **?** (f) Manmade

Figure 1: Binary attributes can be a restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on a variety of applications involving human-machine communication: zero-shot learning, generating image descriptions, improved image search, and active learning of classifiers.

Indeed, we observe that *relative* visual properties are a semantically rich way by which humans describe and compare objects in the world. They are necessary, for instance, to refine an identifying description ("the 'rounder' pillow"), or to situate with respect to reference objects ("'brighter' than a candle; 'dimmer' than a flashlight"). Furthermore, they have potential to enhance active and interactive learning—for instance, offering a better guide for a visual search ("find me similar shoes, but 'shinier'.").

[1]This paper is an invited follow-up to our previous paper (Parikh and Grauman 2011b). Here we provide an overview of a broad variety of applications of the proposed relative attributes.

Proposal: In this work, we propose to model *relative attributes*. As opposed to predicting the presence of an attribute, a relative attribute indicates the strength of an attribute in an image with respect to other images. For example, in Figure 1, while it is difficult to assign a meaningful value to the binary attribute 'smiling', we could all agree on the relative attribute, *i.e*. Hugh Laurie is smiling less than Scarlett Johansson (a), but more than Jared Leto (c). In addition to being more natural, we show how relative attributes offer a richer mode of communication, thus allowing access to more detailed human supervision (for recognition tasks) and guidance (for interactive tasks such as image search).

How can we learn relative properties? Whereas traditional supervised classification is appropriate to learn attributes that are intrinsically binary, it falls short when we want to represent visual properties that are nameable but not categorical. Our goal is instead to estimate the degree of that attribute's presence—which, importantly, differs from the probability of a binary classifier's prediction. To this end, we devise an approach that learns a *ranking function* for each attribute, given relative similarity constraints on pairs of examples (or more generally a partial ordering on some examples). The learned ranking function can estimate a *real-valued* rank for images indicating the relative strength of the attribute presence in them.

The proposed ranking approach accounts for a subtle but important difference between relative attributes and conceivable alternatives based on regression or multi-way classification. While such alternatives could also allow for a richer vocabulary, during training they could suffer from similar inconsistencies as binary attributes. For example, it is more difficult to define and perhaps more importantly, agree on, "With what strength is he smiling?" than "Is he smiling more than she is?". Thus, we expect the relative mode of supervision our approach permits to be more natural and consistent for human labelers.

Contributions: Our main contribution is the idea to learn relative visual attributes. We demonstrate the benefits of the enhanced human-machine communication offered by relative attributes on four novel applications: (1) zero-shot learning from relative comparisons, (2) image description in reference to example images or categories, (3) image search with relative relevance feedback, and (4) active training of discriminative classifiers using feedback. Through comparisons to several strong baselines using image datasets of scenes and faces, we show that relative attributes result in superior performance on all applications.

Related Work

We review related work on visual attributes, other uses of relative cues, and methods for learning comparisons.

Binary attributes: Learning attribute categories allows prediction of color or texture types (Ferrari and Zisserman 2007), and can also provide mid-level cues for object or face recognition (Lampert, Nickisch, and Harmeling 2009; Wang and Mori 2010; Kumar et al. 2009). Moreover, the semantics intrinsic to attributes enable zero-shot transfer (Lampert, Nickisch, and Harmeling 2009; Wang, Markert, and Everingham 2009; Russakovsky and Fei-Fei 2010), or description and part localization (Farhadi et al. 2009; Farhadi, Endres, and Hoiem 2010; Wang and Forsyth 2009). Rather than manually define attribute vocabularies, some work aims to discover attribute-related concepts on the Web (Rohrbach et al. 2010; Berg, Berg, and Shih 2010), extract them from existing knowledge sources (Wang, Markert, and Everingham 2009; Branson et al. 2010) or discover them interactively (Parikh and Grauman 2011a). In contrast to our approach, all such methods restrict the attributes to be categorical (and in fact, binary).

Relative information: Relative information has been explored in vision in a variety of ways. Stemming from the motivation of limited labeled data, Wang, Forsyth, and Hoiem use explicit similarity-based supervision such as "A serval is like a leopard" or "A zebra is similar to the cross-walk in texture" to share training instances for categories with limited or no training instances (2010). Unlike our approach, that method learns a model for each object category, and does not model attributes. In contrast, our attribute models are category-independent and transferrable, enabling relative descriptions between all classes. Moreover, whereas that technique captures similarity among object categories, ours models a general ordering of the images sorted by the strength of their attributes, as well as a joint space over multiple such relative attributes.

Kumar et al. explore comparative facial attributes such as "Lips like Barack Obama" for face verification (2009). These attributes, although comparative, are also modeled as binary classifiers and are similarity-based as opposed to an ordering. Gupta and Davis and Siddiquie and Gupta use prepositions and adjectives to relate objects to each other for more effective contextual modeling and active learning, respectively (2008; 2010). In contrast, our work involves relative modeling of attribute strengths for richer human-machine communication.

Learning to rank: Learning to rank has received extensive attention in the machine learning literature (Joachims 2002; Cao et al. 2007; Liu 2009), for information retrieval in general and image retrieval in particular (Jain and Varma 2011; Hu, Li, and Yu 2008). Given a query image, user preferences (often captured via click-data) are incorporated to learn a ranking function with the goal of retrieving more relevant images in the top search results. Learned distance metrics (e.g., (Frome et al. 2007)) can induce a ranking on images; however, this ranking is also specific to a query image, and typically intended for nearest-neighbor-based classifiers. Our work learns a ranking function on images based on constraints specifying the relative strength of attributes, and the resulting function is not relative to any other image in the dataset. Thus, unlike query-centric retrieval tasks, we can characterize individual images by the strength of the attributes present, which we show is valuable for recognition, search, and description applications.

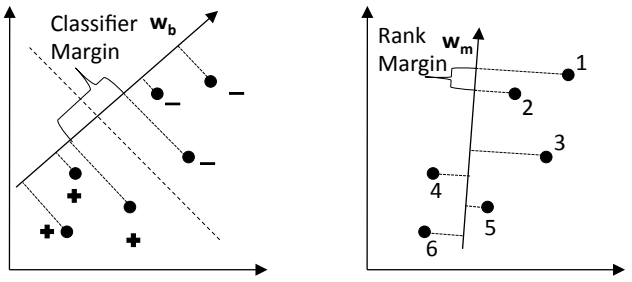


Figure 2: Distinction between learning a wide-margin ranking function (right) that enforces the desired ordering on training points (1-6) , and a wide-margin binary classifier (left) that only separates the two classes ($+$ and $-$), and does not necessarily preserve a desired ordering on the points.

Learning Relative Attributes

We first present our approach for learning relative attributes, and then demonstrate how relative attributes enable several novel applications.

We are given a set of training images $I = \{i\}$ represented in $\mathbb{R}^n$ by feature vectors $\{\boldsymbol{x_i}\}$ and a set of M attributes $A = \{a_m\}$. In addition, for each attribute a_m, we are given a set of ordered pairs of images $O_m = \{(i,j)\}$ and a set of unordered pairs $S_m = \{(i,j)\}$ such that $(i,j) \in O_m \implies i \succ j$, *i.e.* image i has a stronger presence of attribute a_m than j, and $(i,j) \in S_m \implies i \sim j$, *i.e.* i and j have similar relative strengths of a_m. We note that O_m and S_m can be deduced from any partial ordering of the images I in the training data with respect to strength of a_m. Either O_m or S_m, but not both, can be empty.

We adapt the formulation proposed in (Joachims 2002) to learn M ranking functions $r_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i}$. We use a quadratic loss function together with similarity constraints, leading to the following optimization problem:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \quad (1)$$

$$s.t. \quad \boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}; \forall (i,j) \in O_m \quad (2)$$

$$|\boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}; \forall (i,j) \in S_m \quad (3)$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0, \quad (4)$$

where C is the trade-off constant between maximizing the margin and satisfying the pairwise relative constraints. We solve the above primal problem using Newton's method (Chapelle 2007). While we use a linear ranking function in our experiments, the above formulation can be easily extended to kernels.

This learning-to-rank formulation learns a function that explicitly enforces a desired ordering on the training images; the margin is the distance between the closest two projections within all desired (training) rankings. In contrast, if one were to train a binary classifier, only the margin between the nearest binary-labeled examples would be enforced; ordering among examples beyond those defining the margin is arbitrary. See Figure 2. Our experiments confirm this distinction does indeed matter in practice, as our learned ranking function is more effective at capturing the relative strengths of the attributes than the score of a binary classifier (*i.e.*, the magnitude of the SVM decision function). In addition, training with *comparisons* (image i is similar to j in terms of attribute a_m, or i exhibits a_m less than j) is well-suited to the task at hand. Attribute strengths are arguably more natural to express in relative terms, as opposed to requiring absolute judgments in isolation (*i.e.*, i represents a_m with degree 10).

In the following, we introduce four novel tasks enabled by the learned relative attributes: (1) zero-shot learning with relative relationships, (2) generating image descriptions, (3) using relative relevance feedback in image search, and (4) providing relative feedback to a classifier.

For each task, we show example results on two datasets: (1) the Outdoor Scene Recognition (OSR) Dataset (Oliva and Torralba 2001) containing 2,688 images from 8 categories (such as highway, mountain, forest, etc.), which span 6 attributes (such as 'natural', 'open', 'perspective', etc.), and (2) a subset of the Public Figure Face (PubFig) Database (Kumar et al. 2009) containing 772 images from 8 identities, which span 11 facial attributes (such as 'chubby', 'smiling', 'masculine looking', etc.) As image features, we use the texture-based gist (Oliva and Torralba 2001) and color histograms. To train the relative attributes, we specify how the categories relate in terms of their datasets' respective attributes (e.g., forests are more 'natural' than highways; Scarlett is 'younger' than Clive; see (Parikh and Grauman 2011b) for details).

Applications

We now introduce our approach to incorporate relative attributes for four different applications.

Zero-Shot Learning From Relationships

Consider N categories of interest. For example, each category may be an object class, or a type of scene. During training, S of these categories are 'seen' categories for which training images are provided, while the remaining $U = N - S$ categories are 'unseen', for which no training images are provided. A set of M relative attributes are pretrained. The U unseen categories are described relative to one or two seen categories for a subset of the attributes, *i.e.*, unseen class $c_j^{(u)}$ can be described as $c_i^{(s)} \succ c_j^{(u)} \succ c_k^{(s)}$ for attribute a_m, or $c_i^{(s)} \succ c_j^{(u)}$, or $c_j^{(u)} \succ c_k^{(s)}$, where $c_i^{(s)}$ and $c_k^{(s)}$ are seen categories. For example, one could describe the unseen bear class as 'furrier' than giraffes and 'bigger' than dogs. During testing, a novel image is to be classified into any of the N categories.

Predicting the real-valued rank of all images in the training dataset I using the M relative attributes allows us to transform $\boldsymbol{x_i} \in \mathbb{R}^n \rightarrow \tilde{\boldsymbol{x}}_{\boldsymbol{i}} \in \mathbb{R}^M$, such that each image i is now represented as an M-dimensional vector $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ indicating its rank score for all M attributes. We now build a generative model for each of the S seen categories in $\mathbb{R}^M$. We use a Gaussian distribution, and estimate the mean $\boldsymbol{\mu}_i^{(s)} \in \mathbb{R}^M$ and $M \times M$ covariance matrix $\Sigma_i^{(s)}$ from the

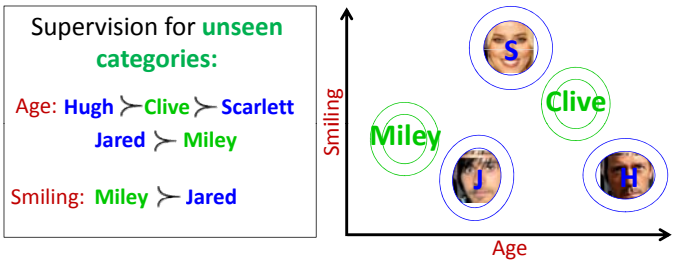


Figure 3: Relative attributes allow the system to learn novel categories using statements about how they relate to previously seen categories.

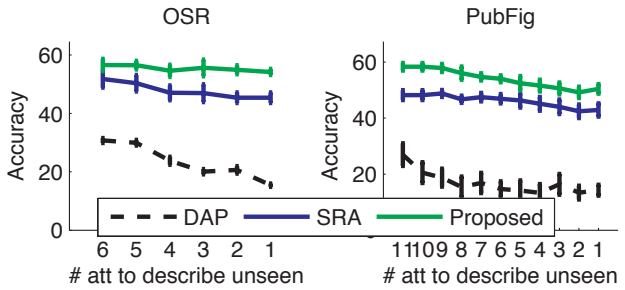


Figure 4: Zero-shot learning performance as fewer attributes are used to describe the unseen categories.

ranking-scores of the training images from class $c_i^{(s)}$, so we have $c_i^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(s)}, \Sigma_i^{(s)})$, for $i = 1, \ldots, S$.

The parameters of the generative model corresponding to each of the U unseen categories are selected under the guidance of the input relative descriptions. See Figure 3. In particular, given an unseen category $c^{(u)}$, we employ the following: If $c_j^{(u)}$ is described as $c_i^{(s)} \succ c_j^{(u)} \succ c_k^{(s)}$, where $c_i^{(s)}$ and $c_k^{(s)}$ are seen categories, then we set the m-th component of the mean $\boldsymbol{\mu}_{jm}^{(u)}$ to $\frac{1}{2}(\boldsymbol{\mu}_{im}^{(s)} + \boldsymbol{\mu}_{km}^{(s)})$. If $c_j^{(u)}$ is described as $c_i^{(s)} \succ c_j^{(u)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to $\boldsymbol{\mu}_{im}^{(s)} - d_m$, where d_m is the average distance between the sorted mean ranking-scores $\boldsymbol{\mu}_{im}^{(s)}$'s of seen classes for attribute a_m. It is reasonable to expect the unseen class to be as far from the specified seen class as other seen classes tend to be from each other. Similarly, if $c_j^{(u)}$ is described as $c_j^{(u)} \succ c_k^{(s)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to $\boldsymbol{\mu}_{im}^{(s)} + d_m$. If a_m is not used to describe $c_j^{(u)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to be the mean across all training image ranks for a_m and the m-th diagonal entry of $\Sigma_j^{(u)}$ to be the variance of the same. In the first three cases, we simply set $\Sigma_j^{(u)} = \frac{1}{S}\sum_{i=1}^{S} \Sigma_i^{(s)}$.

Given a test image i, we compute $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^M$ indicating the relative attribute ranking-scores for the image. It is then assigned to the seen *or* unseen category that assigns it the highest likelihood:

$$c^* = \underset{j \in \{1,\ldots,N\}}{\operatorname{argmax}} \; P\left(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{\mu}_j, \Sigma_j\right). \qquad (5)$$

From a Bayesian perspective, our approach to setting the parameters of the unseen categories' generative models can be considered to be priors transferred from the knowledge of the models for the seen categories. Under reasonable priors, the choice of mean and covariances correspond to the minimum mean-squared error and maximum likelihood estimates. Related formulations of transfer through parameter sharing have been studied in (Fei-Fei, Fergus, and Perona 2003) and (Stark, Goesele, and Schiele 2009) for learning shape-based object models with few training images, though no prior models consider transferring knowledge based on relative comparisons, as we do here.

We note that if one or more images from the unseen categories were subsequently to become available, our estimated parameters could easily be updated in light of the additional evidence. Furthermore, our general approach could potentially support more specific supervision about the relative relationships, should it be available (e.g., bears (unseen) are significantly more furry than cows (seen)).

Whereas the status quo for training object recognition systems is to rely on labeled image exemplars alone, the proposed approach seamlessly integrates top-down human knowledge about how the object categories *relate* in semantic terms. Our zero-shot learning setting is more general than the model in (Lampert, Nickisch, and Harmeling 2009), in that the supervisor may not only associate attributes with categories, but also express how the categories relate along any number of the attributes. We expect this richer representation to allow better divisions between both the unseen and seen categories, as we demonstrate in the experiments.

We compare our zero-shot approach to two baselines: a Direct Attribute Prediction (DAP) model (Lampert, Nickisch, and Harmeling 2009), which uses binary attribute descriptions for all categories, and a "score-based relative attribute" (SRA) model, which follows our method except it replaces rank values with binary classifier output scores. It is a stronger baseline than DAP, as it has the same benefits of the generative modeling of seen classes and relative descriptions of unseen classes as our approach.

Figure 4 shows the per-class recognition accuracy on the OSR and PubFig datasets as we decrease the number of attributes used to describe the unseen category during training. Note that the number of attributes used to describe the *seen* categories during training remains the same. The accuracy of all methods degrades; however, the approaches using relative attributes (SRA and ours) decay gracefully, whereas DAP suffers more dramatically. This illustrates how each attribute conveys stronger distinguishing power when used relatively. Our improved performance over SRA demonstrates the benefits of learning a ranking function to model relative attributes as opposed to continuous scores predicted by a binary classifier.

For implementation details and results analyzing the performance of the proposed approach when varying the number of unseen categories as well as the amount and quality of supervision, please see (Parikh and Grauman 2011b).

Describing Images in Relative Terms

As a second task, we employ relative attributes to automatically generate textual image descriptions. The goal is to be

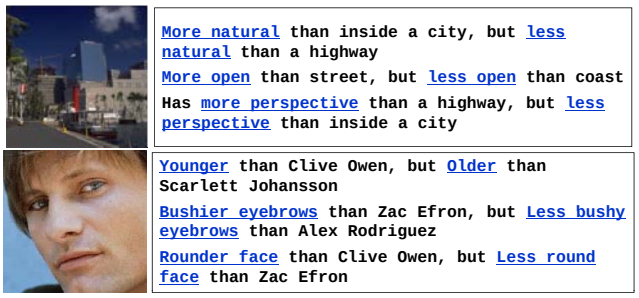


Figure 5: Example descriptions generated by our method for two images. Relative attributes offer more precise descriptions than are possible with categorical properties—which for these examples would only state that the scene "is not natural; is not open; has perspective" (top), or the face "is not young; has bushy eyebrows; is round".

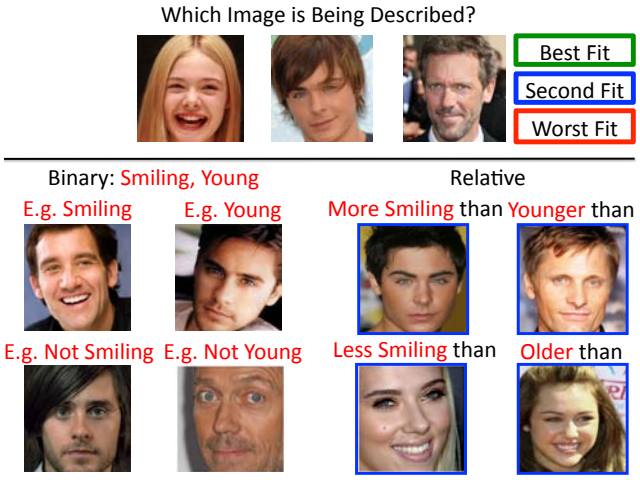


Figure 6: Illustration of task in human subject study to evaluate how informative the proposed relative descriptions are compared to the baseline binary descriptions.

able to relate any new example to other images according to different properties—whether its class happens to be familiar or not.

During training, we are given a set of training images $I = \{i\}$, each represented by a feature vector $\boldsymbol{x_i} \in \mathbb{R}^n$ and a pre-trained set of M attributes $A = \{a_m\}$. We evaluate these attributes on all training images in I. Given a novel image j to be described, we evaluate all learned ranking functions $r_m(\boldsymbol{x_j})$. For each attribute a_m, we identify two reference images i and k from I that will be used to describe j via relative attributes. To avoid generating an overly precise description, we wish to select i and k such that they are not very similar to j in terms of attribute strength. However, to avoid trivial descriptions, they must not be too far from j, either. Hence, we select i and k such that $i \succ j$ and $j \succ k$ in strength of a_m, and $\frac{1}{8}^{th}$ of the images in I lie between i and j, as well as between j and k. In the case of boundary conditions where no such i or k exist, i is chosen to be the image in I with the least strength of a_m, and k is set to the image in I with the highest strength of a_m. The image j can then be described in terms of all or a subset of the M attributes, relative to any identified pairs (i, k). While more elaborate analysis of the dataset distribution—and even psychophysics knowledge of the sensitivity of humans to change in different attributes—could make the selection of reference images more effective, we employ this straightforward technique as a proof-of-concept and leave such analysis for future work.

In addition to describing an image relative to other images, our approach can also be used to generate purely textual descriptions by describing an image relative to categories. See Figure 5. Here our method selects the categories to compare to such that at least 50% of the images in the category have an attribute strength larger than (less than) that computed for the image to be described. We can qualitatively see that the relative descriptions are more precise and informative than the binary ones given in the caption, which are generated using the outputs from binary classifiers.

To quantify the quality of automatically generated descriptions, we perform a human subject study that pits the binary attribute baseline against our relative approach. Our method reports the properties predicted relative to reference images, while the baseline reports the predicted presence/absence of attributes only. The human subject must guess which image led to the auto-generated descriptions. To our knowledge, these are the first results to quantify how well algorithm-generated attribute descriptions can communicate to humans.

We recruited 18 subjects, only some familiar with vision. We randomly selected 20 PubFig and 10 OSR images. For each of the 30 test cases, we present the subject a description using three randomly selected attributes plus a multiple-choice set of three images, one of which is correct. The subject is asked to rank their guesses for which fits the description best. See Figure 6. To avoid bias, we divided the subjects into two groups; each group saw either the binary or the relative attributes, but not both. Further, we display reference images for either group's task, to help subjects understand the attribute meanings.

We find that subjects are significantly more likely to identify the correct image using our method's description—48% vs. 34% in the first choice. This supports our claim that relative attributes can better capture the "concept" of the image in a manner that is understandable to humans, and reinforces their real promise for improved human-machine communication. More results can be found on the authors' websites.

Relative Relevance Feedback in Image Search

As a third task, we propose using relative attributes to express feedback during an interactive image search (Kovashka, Parikh, and Grauman 2012). In this scenario, a user can envision image content of interest, but needs the system's help to find it—for example, a graphic designer might seek a particular kind of illustration, or a shopper may envision a product to be found online. Using keywords or even image-based search as a starting point often does not return satisfactory results. Hence, feedback plays an important role. The most common form is *binary relevance feedback*,

Figure 7: Our model allows users to give relative attribute feedback on reference images to refine their image search.

where the user identifies some images that are relevant and some that are not (Rui et al. 1998).

Instead, we will use relative attributes as a mode of feedback where a user directly describes how high-level properties of exemplar images should be adjusted in order to more closely match his/her envisioned target images. For example, when conducting a query on a shopping website, the user might state: "I want shoes like these, but *more formal*." Or, when searching for a suspect in a database, a witness might state: "He looked similar to this guy, but with a *broader nose*." See Figure 7.

Offline, we first learn a set of M relative attributes. At query time, the system presents an initial set of reference images, and the user selects among them to provide relative attribute feedback. Using the resulting constraints in the multi-dimensional attribute space, we update the system's relevance function. Each image in the pool is given a relevance score corresponding to the number of user specified constraints it satisfies. The system then displays the top-ranked set to the user. This procedure iterates using the accumulated constraints until the top ranked images are acceptably close to the user's target. We call the approach *WhittleSearch*, since it allows users to "whittle away" irrelevant portions of the visual feature space via precise, intuitive statements of their attribute preferences.

In experiments, we analyze how the proposed relative feedback enhances image search compared to classic binary feedback, where an SVM classifier is trained to separate relevant from irrelevant images as marked by the user. For each query we select a random *target image* and score how well the search results match that target after feedback. We evaluate the correlation (using use Normalized Discounted Cumulative Gain at top K, or NDCG@K) between the full ranking computed by our approach and a ground truth ranking that reflects the perceived relevance of all images in the pool. The metric captures not only where the target itself ranks, but also how similar to the target the other top-ranked images are. We form the ground truth relevance ranking by sorting all images by their distance to the given target using a learned distance that mimics human perception of similarity; see (Kovashka, Parikh, and Grauman 2012) for details.

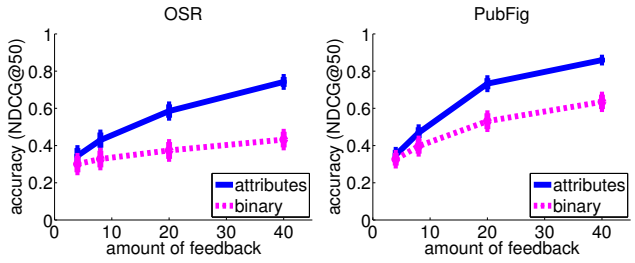


Figure 8: The proposed attribute feedback yields faster gains per unit of feedback compared to a traditional binary relevance feedback model.

Figure 8 shows the rank correlation results for 100 queries. These curves show the quality of all top-ranked results as a function of the amount of feedback given in a single iteration. For both datasets, both methods clearly improve with more feedback. However, the precision enabled by our attribute feedback yields a greater "bang for the buck"—higher accuracy for fewer feedback constraints. We present further results using feedback from real human subjects, and analyzing the impact of iterations and choice of reference images in (Kovashka, Parikh, and Grauman 2012).

Relative Feedback to Discriminative Classifiers

Finally, we present an active learning scenario where a classifier incrementally collects labels for unlabeled images in a dataset. At each iteration, the classifier conveys its predicted label for the image to the supervisor. The supervisor confirms or rejects this prediction. Moreover, if rejected, the supervisor provides feedback using relative attributes as to why the classifier's prediction is incorrect. *E.g.* if a classifier classifies a coast image as forest, the supervisor may say "this image is too open to be a forest image". Hence, all images more open than this image are very unlikely to be forest images. See Figure 9. The classifier uses the predicted relative attribute values of all images in the unlabeled pool, identifies the ones more open than the query image, and uses them as negative examples for the forest category. Hence, the supervisor's response to one image is transferred to many unlabeled images, in turn accelerating the training of the classifier.

Similar to zero-shot learning, the relative attributes based feedback alleviates the labeling burden of the supervisor, while still allowing for discriminative learning of the classifier in any feature space. Hence, this application allows for a marriage between powerful discriminative learning and direct injection of domain knowledge on the part of the supervisor (using means beyond labeled data).

We gather attribute-based feedback from real users on Mechanical Turk, and compare the proposed approach to a baseline that does not allow for attributes-based feedback. Figure 10 shows the results. On the x-axis are the number of iterations in the active learning process. We see that the relative attributes feedback leads to a more effective classifier with the same number of user iterations. More results on a variety of scenarios and feature spaces can be found

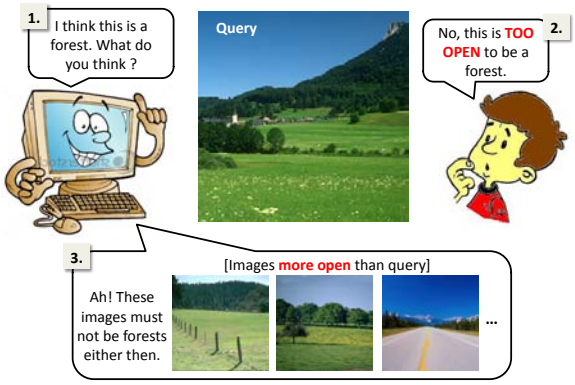


Figure 9: Using relative feedback to refine classifiers.

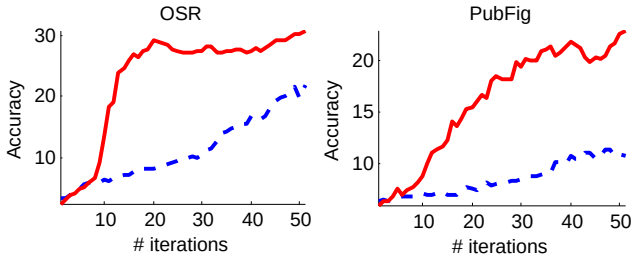


Figure 10: Our approach using relative attributes for active classifier feedback (solid) outperforms the baseline (dashed) for the same number of user iterations.

in (Parkash and Parikh 2012).

Conclusion

We introduced relative attributes, which allow for a richer language of supervision and description than the commonly used categorical (binary) attributes. We presented four applications: zero-shot learning based on relationships, describing images relative to other images or categories, user guidance in image search, and supervisor feedback to classifiers. Through experiments we have clearly demonstrated the advantages of our idea. Overall, our work indicates the importance of moving beyond category labels for visual recognition applications, and the promise of expanding human-machine communication by using semantic human-understandable visual terms.

Acknowledgments This research is supported in part by ONR YIP and NSF IIS-1065390 (K.G. and A.K.).

References

Berg, T.; Berg, A.; and Shih, J. 2010. Automatic attribute discovery and characterization from noisy web data. In *ECCV*.

Branson, S.; Wah, C.; Babenko, B.; Schroff, F.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. In *ECCV*.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*.

Chapelle, O. 2007. Training a support vector machine in the primal. *Neural Computation*.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*.

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

Fei-Fei, L.; Fergus, R.; and Perona, P. 2003. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*.

Ferrari, V., and Zisserman, A. 2007. Learning visual attributes. In *NIPS*.

Frome, A.; Sha, F.; Singer, Y.; and Malik, J. 2007. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*.

Gupta, A., and Davis, L. S. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*.

Hu, Y.; Li, M.; and Yu, N. 2008. Multiple-instance ranking: Learning to rank images for image retrieval. In *CVPR*.

Jain, V., and Varma, M. 2011. Learning to re-rank: query-dependent image re-ranking using click data. In *WWW*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD*.

Kovashka, A.; Parikh, D.; and Grauman, K. 2012. WhittleSearch: Image search with relative attribute feedback. In *CVPR*.

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

Lampert, C.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Liu, T.-Y. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*.

Parikh, D., and Grauman, K. 2011a. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*.

Parikh, D., and Grauman, K. 2011b. Relative attributes. In *ICCV*.

Parkash, A., and Parikh, D. 2012. Attributes for classifier feedback. Technical Report 2012-1, TTIC.

Rohrbach, M.; Stark, M.; Szarvas, G.; Gurevych, I.; and Schiele, B. 2010. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*.

Rui, Y.; Huang, T. S.; Ortega, M.; and Mehrotra, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans on Circuits and Video Technology*.

Russakovsky, O., and Fei-Fei, L. 2010. Attribute learning in large-scale datasets. In *Intl Workshop on Parts and Attributes, ECCV*.

Siddiquie, B., and Gupta, A. 2010. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*.

Stark, M.; Goesele, M.; and Schiele, B. 2009. A shape-based object class model for knowledge transfer. In *ICCV*.

Wang, G., and Forsyth, D. 2009. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*.

Wang, Y., and Mori, G. 2010. A discriminative latent model of object classes and attributes. In *ECCV*.

Wang, G.; Forsyth, D.; and Hoiem, D. 2010. Comparative object similarity for improved recognition with few or no examples. In *CVPR*.

Wang, J.; Markert, K.; and Everingham, M. 2009. Learning models for object recognition from natural language descriptions. In *BMVC*.