Interactive Discovery of Task-Specific Nameable Attributes

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

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

Human-nameable visual attributes are useful as mid-level features for recognition, yet the question of which attributes should be learned for a given task remains a critical one. We introduce an approach to discover a task-specific attribute vocabulary that is both human understandable and discriminative. 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 our approach's advantages on a spectrum of coarse- to fine-grained categorization tasks. Our work will appear at CVPR 2011, and here we also present an extension to discover localized attributes for the fine-grained task of recognizing the identity of faces.

Visual attributes [1–3] offer a useful mid-level representation that allows a recognition system to learn new categories from textual descriptions [2], compute meaningful descriptions of unfamiliar objects, or report on unusual aspects of some instance [3]. Despite attributes' apparent assets for recognition problems, a critical question remains unaddressed: which visual attributes should be learned for a given task at hand?

The choice of attributes becomes even more crucial in the setting of fine-grained visual recognition. On the one hand, zero-shot learning and other forms of transfer become more natural and arguably even necessary [10] in this setting where categories may be more similar than different, thus requiring attributes that are defined by human language *i.e*. are "nameable" (*e.g*. to learn a model for Megan Fox, it is useful to know that Megan Fox is like Angelina Jolie but younger). On the other 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 [2–7], or if we automatically mine text on the Web [8, 9], there is no guarantee that those attributes will faithfully capture the subtle differences in appearance ensuring that the categories be separable in the image feature space—a necessary condition if they are intended to serve as the mid-level cues for recognition. Thus, nameability and discriminativeness, unfortunately, appear to be at odds.

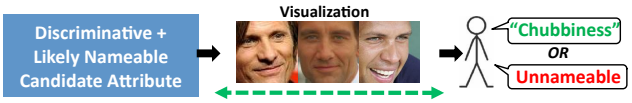

Figure 1. Interactively building a discriminative vocabulary of nameable attributes.

Our idea We aim to build a *discriminative* attribute *vocabulary* that is amenable to a given visual recognition task at some level of granularity, yet also serves as interpretable mid-level cues. We propose an interactive approach that prompts a (potentially non-expert) human-in-the-loop to provide names for attribute hypotheses it discovers. See Figure 1.

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 (see Figure 2 for examples). 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.

An overview of our interactive approach is as follows.[1] At each iteration, we actively determine an attribute hypothesis that helps discriminate among the most confused classes given the current collection of attributes. We then estimate the probability that the hypothesis is *nameable* by exploring a manifold structure in the space of nameable hyperplane separators. If it appears unnameable, we discard it and loop back to select the next potential attribute hypothesis. If it appears nameable, the system presents a visualization of the attribute to the annotator. If the annotator accepts and names it, we append this new attribute to our discovered vocabulary, retrain the higher-level classifier accordingly, and update our nameability model. If it is rejected, the system loops back to generate a new attribute hypothesis.

[1] and is depicted in the first figure of the attached draft poster.

Figure 2. Example visualizations of candidate attributes shown to subjects. Responses were "spotted" (top) and "smiling" (bottom).

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 to isolate each category from the other (potentially closely related) categories, and are also human understandable. These attributes can then be used for fine-grained or basic recognition, zero-shot learning, or describing novel images.

Results We evaluate our approach on a series of categorization tasks that range from coarse- to fine-grained: outdoor scene recognition, animal species recognition, and human identification. For each, we study how our method's discovered task-specific attributes fare compared to either a traditional approach that relies on purely discriminative features or one in which an expert simply enumerates the vocabulary. We highlight a couple key results of the full system here; our poster draft further analyzes components of the approach in isolation.

We use the Outdoor Scene Recognition [11] (OSR) data containing 8 basic categories, a subset of the Animals With Attributes dataset [2] containing 8 animal species (AWA), and a subset of the Public Figures Face dataset (PubFig) [6] containing 6 male celebrities. For OSR and AWA we discover attributes from global descriptions (Gist, color), while for PubFig we discover localized attributes from horizontal blocks extracted from each face region. For evaluation purposes, for each dataset, we collect human judgements on how confidently nameable each discriminative split in the feature-space is via Amazon Mechanical Turk (20 subjects per visual split, and a total of 25,000 responses).

A comparison to the *purely discriminative* baseline is shown in Figure 3. 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, resulting in better object/scene/face categorization accuracy for novel images. 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.

A comparison to a *purely descriptive* baseline that relies on a hand-generated list of attributes, as is typically done in previous work (e.g., [2, 3, 6, 7, 12]), is shown in Figure 4. At each iteration, we add one random attribute from this hand-generated list to the attribute vocabulary. We see that attributes discovered by our approach tend to be more discriminative than those designated a priori with a purely descriptive list of words.

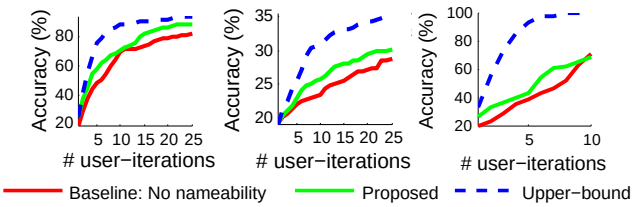

Figure 3. Comparison of our approach to a discriminative-only baseline. Left to right: AWA gist, OSR color, PubFig.

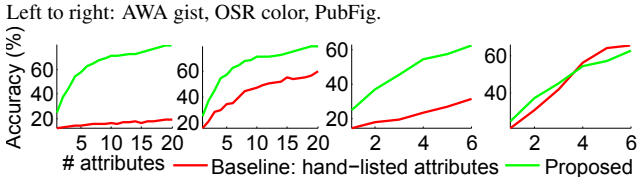

Figure 4. Comparison of our approach to a nameability-only baseline. Left to right: AWA color, AWA gist, OSR color, OSR gist.

We also use our discovered attributes to describe images and find that the resultant descriptions are qualitatively meaningful. For instance, we find that an image of a zebra, a previously unseen category, is described by the system as "black", "white" and "long-neck" (see poster).

Summary The proposed interactive approach discovers attributes that are both human understandable and discriminative—two characteristics needed for attributes to be truly useful, especially for fine-grained visual recognition. Results on a variety of datasets ranging from basic to fine-grained tasks indicate its clear advantages over today's common practices in attribute-based recognition.

References

[1] V. Ferrari and A. Zisserman. Learning Visual Attributes. *NIPS*, 2007.

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

[3] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by their Attributes. *CVPR*, 2009.

[4] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona and S. Belongie. Visual Recognition with Humans in the Loop. *ECCV*, 2010.

[5] J. Wang, K. Markert and M. Everingham. Learning Models for Object Recognition from Natural Language Descriptions. *BMVC*, 2009.

[6] N. Kumar, A. Berg, P. Belhumeur and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. *ICCV*, 2009.

[7] Y. Wang and G. Mori. A Discriminative Latent Model of Object Classes and Attributes. *ECCV*, 2010.

[8] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych and B. Schiele. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer. *CVPR*, 2010.

[9] T. L. Berg, A. C. Berg and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. *ECCV*, 2010.

[10] J. Wang, K. Markert and M. Everingham. Learning Models for Object Recognition from Natural Language Descriptions. *BMVC*, 2009.

[11] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *IJCV*, 2001

[12] O. Russakovsky and L. Fei-Fei. Attribute Learning in Large-scale Datasets. *Workshop on Parts and Attributes, ECCV*, 2010.