Learning Attributes from Human Gaze

Nils Murrugarra-Llerena Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{nineil, kovashka}@cs.pitt.edu

Abstract

While semantic visual attributes have been shown useful for a variety of tasks, many attributes are difficult to model computationally. One of the reasons for this difficulty is that it is not clear where in an image the attribute lives. We propose to tackle this problem by involving humans more directly in the process of learning an attribute model. We ask humans to examine a set of images to determine if a given attribute is present in them, and we record where they looked. We create gaze maps for each attribute, and use these gaze maps to improve attribute prediction models. For test images we do not have gaze maps available, so we predict them based on models learned from collected gaze maps for each attribute of interest. Compared to six baselines, we improve prediction accuracies on attributes of faces and shoes, and we show how our method might be adapted for scene images. We demonstrate additional uses of our gaze maps for visualization of attribute models and learning "schools of thought" between users in terms of their understanding of the attribute.

1. Introduction

Semantic visual attributes (such as "metallic" or "smiling") have been used for a variety of tasks: as a low-dimensional representation for object recognition [7, 26, 33, 53, 16], as a textual representation used to recognize previously unseen categories [7, 26, 31, 16, 27], as a supervision modality for active learning [24, 32], etc.

However, unlike object categories, attributes are not well-defined. To see why, consider the following thought experiment. If a person is asked to draw a "boot", the drawings of different people will likely not differ very much. But if a person is asked to draw what the attributes "formal" or "feminine" mean, drawings will vary. Drawings of a "forest" will likely all include a number of trees, but drawings of a "natural", "open-area", or "cluttered" scene will differ greatly among artists. Finally, if humans are asked to draw or even pick from a set of male actors an "attractive" or "masculine" person, responses will differ more than if they were asked to draw or select a "man".

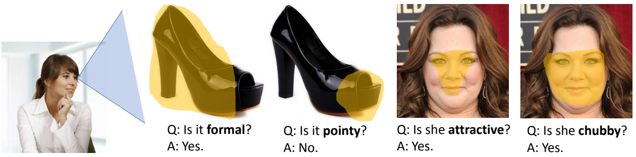


Figure 1: We learn the spatial support of attributes by asking humans to judge if an attribute is present in training images. We use this support to improve attribute prediction.

Since attributes are less well-defined, capturing them with computational models poses a different set of challenges than capturing object categories does. There is a disconnect between how humans and machines perceive attributes, and it negatively impacts tasks that involve communication between a human and a machine, since the machine may not understand what a human user has in mind when referring to a particular attribute. Since attributes are human-defined, the best way to deal with their ambiguity is by learning from humans what these attributes really mean.

We propose to learn attribute models using human gaze maps that show which part of an image contains the attribute. To obtain gaze maps for each attribute, we conduct human subject experiments where we ask viewers to examine images of faces, shoes, and scenes, and determine if a given attribute is present in the image or not. We use an inexpensive GazePoint eyetracking device which is simply placed in front of a monitor to track viewers' gaze, and record the locations in the image that had some number of fixations. We aggregate the gaze collected from multiple people on training images, to obtain an averaged gaze map per attribute that we use to extract features from both training and test images. We also experiment with learning a saliency model that *predicts* which pixels will be fixated. To capture the potential ambiguity and visual variation within each attribute, we cluster the positive images per attribute and their corresponding gaze locations, and obtain multiple gaze maps per attribute. We create one classifier per gaze map which only uses features from the region under non-zero gaze map values, for both training and testing.

The gaze maps that we learn from humans indicate the spatial support for an attribute in an image and allow us to better understand what the attribute means. We use gaze maps to identify regions that should be used to train attribute models. We show this achieves competitive attribute prediction accuracy compared to six methods, five of which are alternative ways to select relevant features. We also demonstrate additional applications showing how our method can be used to visualize attribute models, and how it can be employed to discover groups among users in terms of their understanding of attribute presence.

The main contribution of our work is a new method for learning attribute models, using inexpensive but rich data in the form of gaze. We show that our method successfully discovers the spatial support of attributes. Despite the close connection between attributes and human communication, gaze has never been used to learn attribute models before.

2. Related Work

Attributes. Semantic visual attributes are properties of the visual world, akin to adjectives [26, 7, 1, 33]. In this work, we focus on modeling attributes as *binary* categories [26, 7, 1, 33]. Attributes bring recognition closer to human-like intelligence, since they allow generalization in the form of zero-shot learning, i.e. learning to recognize previously unseen categories using a textual attribute-based description and prediction models for these attributes learned on other categories [26, 7, 31, 16]. Attributes have also been shown useful for actively learning object categories [32], scene recognition [33], and action recognition [27].

Attribute naming and ambiguity. [30, 33] propose how to develop an attribute vocabulary. [20, 21] show attributes can be subjective and there exist "schools of thought" in terms of how users use an attribute word. In other words, users can be grouped in terms of how they respond to questions about the presence or absence of attributes, and how they use the attribute name. Some work discovers non-semantic nameless attributes [53, 35, 38], but for tasks involving search and zero-shot learning we require attributes that have human-given names.

Learning and localizing attribute models. Some work studies specifically how to learn accurate attribute models. For example, [2, 47, 8, 39, 12, 52] model attributes jointly. [16] improve attribute predictions by decorrelating the use of features by different attribute models, and [9] improve attribute prediction accuracy by finding a feature representation that is invariant across categories. In the domain of *relative* attributes [31], which we do *not* study, [37] discover parts that improve relative attribute prediction accuracy. It is unclear whether the discovered parts capture the true meaning of attributes as humans perceive them, or simply exploit image features which are *correlated* [16] with the attribute of interest, but are *not* part of the human perception of the attribute.[1] Further, [37]'s method is not applicable to binary attributes and requires pre-trained facial landmark detectors. In recent work, [48] propose to discover the spatial extent of relative attributes, and apply their method to images of shoes and faces as we do. They find spatial extent by building "visual chains" that capture the commonalities between images which flow from ones having the attribute to a strong degree, to those that have it less. While we model attributes as binary properties (in contrast to [48]), and use human insight to learn where an attribute lives, [48] is the most related work to ours so we compare to it in Sec. 4.

Other recent work applies deep neural networks to predict attributes [39, 40, 6, 46, 8]. While deep nets can improve the discriminativity of attribute models, they do not exploit human supervision on the meaning or spatial support of attributes. Thus, progress in deep nets is *orthogonal* to the objective of our study. We show that *even when deep features are employed*, using gaze maps to determine the spatial support of attributes improves performance.

Using humans to select relevant regions. [44] pair two humans in an image-based guessing game, where the goal is for the first person to reveal such image regions that allow the second person to most quickly guess the category of the image. The revealed regions are then assumed to be the most relevant for the category of interest. [3, 4] propose a single-player guessing game called "Bubbles," where the player must reveal as few circular regions of an image as possible, in order to match that image to one of two categories with several examples shown. There are three important differences between our work and [44, 3, 4]: (1) These approaches are used to learn objects, not attributes, and attributes have much more ambiguous spatial support; (2) They require that a human should *click* on a relevant image region, which means that the user is consciously aware of what the relevant regions are, whereas in our approach a human uses her potentially subconscious intuition about what makes an image "natural", "formal", or "chubby"; and (3) Clicking or drawing requires a bit more effort (looking is easier than moving one's hand to use the mouse).

Our method can be seen as a form of annotator rationales [55, 5], which are annotations that humans provide to teach the classifier why a category is present. For example, the user can mark which regions of the face make a person "attractive". However, providing gaze maps by looking is much faster than drawing rationales (see Sec. 3.2).

Gaze and saliency. [29] use human gaze to reduce the effort required in obtaining data for object detectors. They build bounding boxes from locations in a photo where a user fixates when judging which of two categories is portrayed

[1] This is also true for attention networks [42, 15] as they are data-driven, not based on human intuition.

in the image. [54] argue that using gaze can improve object detection—bounding box predictions that do not align with fixations can be pruned. They also use a gaze-based feature to classify detections into true and false positives, but only show small gains in detection accuracy.

In addition to gaze, saliency examines where a viewer will fixate in an image [14, 34, 28, 11, 19, 10, 18, 13]. We use [19]'s method to predict gaze maps for novel images. No prior work uses gaze to learn attribute models.

3. Approach

We first describe our datasets (Sec. 3.1) and how we collect gaze data from human subjects (Sec. 3.2). In Sec. 3.3, we discuss how we compute one or multiple gaze templates per attribute, and in Sec. 3.4, we describe how we use the templates to restrict the range of an image from which an attribute model is learned. Finally, in Sec. 3.5, we show how we predict an individual gaze template for each test image.

Like [48], our method is designed for images which contain a single object, specifically faces and shoes. See Sec. 4.2 for a preliminary adaptation of our work for scenes.

3.1. Datasets

We use two attribute datasets: the **Faces** dataset of [25] (also known as PubFig), and the **Shoes** dataset of [22]. All images are of the same square size (200x200 pixels for faces and 280x280 for shoes). The attributes we use are: for **Shoes**, "feminine", "formal", "open", "pointy", and "sporty"; and for **Faces**, "Asian", "attractive", "baby-faced", "big-nosed", "chubby", "Indian", "masculine", and "youthful". Like [48], we consider a subset of all attributes, in order to focus the analysis towards attributes whose spatial support does not seem obvious, i.e. it could not be predicted from the attribute name alone. This allows insight into the meaning of some particularly ambiguous attributes (e.g. "formal", "feminine" and "attractive"). We also selected some attributes ("pointy" and "big-nosed") where we had a fairly confident estimate of where gaze locations would be. This allows us to qualitatively evaluate the collected gaze maps via their alignment with the expected gaze locations. The annotation cost per attribute is small, about 1 minute per image-attribute pair (see below).

We select 60 images total per attribute. In order to get representative examples of each attribute, we sample: (a) 30 instances where the attribute is definitely present, (b) 18 instances where it is definitely not present, and (c) 12 instances where it may or may not be present. For **Faces**, we use the provided SVM decision values to select images in these three categories. For **Shoes**, we use the ordering of ten shoe categories from most to least having each attribute, which we map to individual images using their class labels.

3.2. Gaze data collection

We employ a $495 GazePoint GP3 eye-tracker device[2] to collect gaze data from 14 participants. The 320x45x40mm eye-tracker is placed in front of a monitor, and the participants do *not* have to wear it, in contrast to older devices. Gaze data can also be collected via a webcam; see [50].

Our experiment begins with a screening phase in which we show ten images to each participant and ask him/her to look at a fixed region in the image that is marked by a red square, or to look at e.g. the nose or right eye for faces. If the fixated pixel locations lie within the marked region, the participant moves on to the data collection session. The latter consists of 200 images organized in four sub-sessions. In order to increase the participants' performance, we allow a five-minute break between sub-sessions. We ask the viewer whether a particular attribute is present in a particular image which we then show him/her. The participant has two seconds to look at the image and answer. His/her gaze locations and answers are recorded. We obtain 2.5 gaze maps on average, for each image-attribute question.

Of the 200 images, 20 are used for validation. If the gaze fixations on some validation image are not where they should be, we discard data from the annotator that follows that validation image and precedes the next one.

Each experiment took one hour, for a total of 14 hours of human time. Thus, obtaining the gaze maps for each of our 13 attributes took a short amount of time, *about one hour per attribute* or one minute per image-attribute pair. Our collected gaze data is available on our website[3]. Note that viewing an image is faster than drawing a rationale (45 seconds), so we save time and money compared to [5].

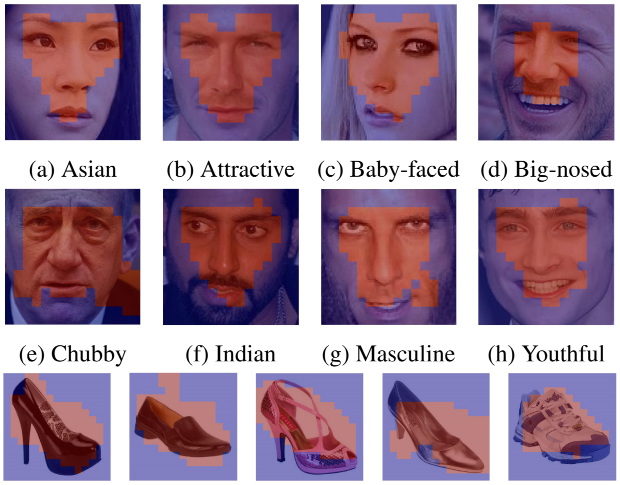
In contrast to our approach, some saliency work [18, 13] approximates gaze with mouse clicks, but as argued in relation to region selection methods (Sec. 2), clicks require conscious awareness of what makes an image "formal" or "baby-faced", which need not be true for attributes.

3.3. Generating gaze map templates

The gaze data and labels are collected jointly but aggregated *separately* for each attribute. The format of a recorded gaze map is an array of coordinates (x, y) of the image being viewed. We convert this to a map with the same size as the image, with a value of 1 or 0 per pixel denoting whether the pixel was fixated or not. First, the gaze maps across all images that correspond to positive attribute labels are OR-ed (the maximum value is taken per pixel) and divided by the maximum value in the map. Thus we arrive at a gaze map gm_m for the attribute m with values in the range $[0, 1]$. Second, a binary template bt_m is created using a threshold of $t = 0.1$ on gm_m. All locations greater than t are marked as

[2] http://www.gazept.com/product/gazepoint-gp3-eye-tracker/

[3] http://www.cs.pitt.edu/∼nineil/gaze_proj/



(a) Asian (b) Attractive (c) Baby-faced (d) Big-nosed

(e) Chubby (f) Indian (g) Masculine (h) Youthful

(i) Feminine (j) Formal (k) Open (l) Pointy (m) Sporty

Figure 2: Grid templates for the face (top two rows) and shoe attributes. Best viewed in color.

1 in bt_m and the rest as 0. Third, we apply a 15x15 grid over the binary template to get a grid template gt_m. The process starts with a grid template filled with all 0 values. Then if a pixel with value 1 of bt_m falls inside some grid cell of gt_m, this cell is turned on (all pixels in that cell are replaced with 1). Some examples of the generated templates are shown in Fig. 2. Red regions are cells with value 1, while blue regions are cells with value 0.

To get templates that capture the subtle variations of how an attribute might appear [21] and also separate different types of objects, a clustering is performed over the images labeled as positive by our human participants. For example, boots can be in one group and high-heels in another. We use K-means with $k = 5$.[4] After the clustering procedure, we repeat the grid template generation, but now separately for each of the five clusters. Thus, we obtain five grid templates per attribute. Each attribute classifier can then specialize to a very concrete appearance, which might make learning a reliable model easier than learning an overall single-template model.

Examples of the five templates for the attribute "open" are shown in Fig. 3. We observe that each template captures a different meaning of "openness", e.g. open at the back (first, second and third image), front (fifth), or throughout (fourth). We also show multiple templates for the attribute "chubby" on the same image, for easier comparison. We quantitatively compare using one versus five grid templates in Tab. 1 and 3, and show additional qualitative results in our supplemental material.

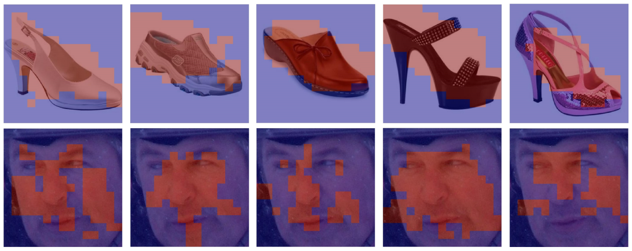


Figure 3: Grid templates for each positive cluster for the attributes "open" (top) and "chubby" (bottom). At the top, we show multiple templates capturing the nuances of "openness". At the bottom, we show how multiple templates for "chubby" look on the same image. Best viewed in color.

3.4. Learning attribute models using gaze templates

We consider two approaches: SINGLE TEMPLATE (ST) and MULTIPLE TEMPLATES (MT). For SINGLE TEMPLATE, the parts of images involved in training and testing are multiplied by the grid template values, which results in image pixels under a 0 value being removed and keeping other pixels the same. We then extract both local and global features from the remaining part of the image, and train a classifier corresponding to the template using these features. At test time, we apply the template to each image, extract features from the 1-valued part, and apply the classifier. For MULTIPLE TEMPLATES, we train five different classifiers (one per cluster), each corresponding to one grid template. We classify a new image as positive if at least one of the five classifiers predicts it contains the attribute.

Comparison to rationales. To test the effectiveness of our gaze template construction, we also tried implementing our gaze templates as rationales [55, 5]. In this work, the authors seek not only labels from their annotators (*e.g.* this person is attractive, and that person is not), but also ask annotators to mark with a polygon the region in the image that determined their positive/negative response. Our gaze templates resemble attributes since they indicate which region a human looked at to determine if an attribute is present. We implement gaze as a form of rationales as follows. If we have a positive image x_i and a template region within it r_i, we construct an artificial training example $x_i - r_i$ that excludes r_i, and then generate an additional constraint in the SVM formulation that enforces that x_i examples should receive a higher score than $(x_i - r_i)$ examples. This resulted in inferior results, thus confirming our choice of how to incorporate the gaze templates into attribute learning.

3.5. Learning attribute models with gaze prediction

So far we have used a single gaze template (or five templates) for each attribute, and applied it to all images. Rather than using a fixed template, one can also *learn* what a gaze

[4] We did not tune this parameter but also found the performance of our algorithm *not* to be sensitive to its choice. One can pick K using the silhouette coefficient [36] or a validation set.

map would look like for a novel test image. We construct a model following Judd's simple method [19], by inputting (1) our training gaze templates, from which 0/1 gaze labels are extracted per pixel, and (2) per-pixel image features (the same feature set as in [19] including color, intensity, orientation, etc; but excluding person and car detections). This saliency model learns an SVM which predicts whether each pixel will be fixated or not, using the per-pixel features. We learn a separate saliency model for each attribute.

```
Data: Training grid templates $templates_{train,m}$ for
      attribute $m$; test image $i$
Result: Template for the test image $template_i$, to be
        used for feature extraction
Train a saliency model using $templates_{train,m}$;
Apply saliency model to $i$ to predict gaze map $gm_m^i$;
for $u \in \{0.1, 0.2, \ldots, 0.9\}$ do
    $r \leftarrow$ Threshold $gm_m^i$ at $u$;
    $score_u \leftarrow$ similarity of $r$ and $templates_{train,m}$
end
$fu \leftarrow$ Set the final threshold to $\arg\max_u(score_u)$;
$template_i \leftarrow$ Apply threshold $fu$ to gaze map $gm_m^i$
```

Algorithm 1: Predicting a gaze template using saliency.

For each attribute, as outlined in Alg. 1, we first learn a saliency model. Then we predict a real-valued saliency score for each pixel in each test image. Finally, we convert this real-valued saliency map to a binary template. To generate the latter, we consider thresholds u between 0.1 and 0.9. To score each u, we apply it to the predicted gaze template for our test image to obtain a binary test template. We compute the similarity between that test template and the training binary templates (Sec. 3.3), as the intersection over union of the 1-valued regions. Finally, we fix our choice of the threshold u to the one with the highest similarity score.

Once we have the binary grid template for the test image, we can extract features from it as in Sec. 3.4, only from the area predicted to have fixations on it. However, the size of the gaze template on test images is no longer guaranteed to be the same as the size of the template on training images, so we have a feature dimensionality mismatch. Thus, we opt for a bag-of-visual-words representation over dense SIFT features (from the part of the image under positive template values in the train/test images) and a vocabulary of 1000 visual words. Then, we build a new classifier using the templates on the training data as discussed above, and apply this model to the features extracted from our new *predicted* grid template. We call this approach SINGLE TEMPLATE PREDICTED (STP) or MULTIPLE TEMPLATES PREDICTED (MTP), depending on whether a single or multiple templates were used per attribute at training time. The names denote that at test time, we use a predicted template.

4. Results

In this section, we present a comparison (Sec. 4.1) of our approach against six different baselines on the task of attribute prediction, five of which are alternative methods to select relevant regions in the image from which to extract features. We also include two additional applications: using gaze templates to visualize attribute models (Sec. 4.3), and discovering "schools of thought" among annotators which denote how they perceive attribute presence (Sec. 4.4). We primarily test our approach on the **Faces** and **Shoes** datasets, but in Sec. 4.2, we show an adaptation of our approach for scene attributes.

4.1. Attribute prediction

We build attribute prediction models using both standard vision features and features extracted from convolutional neural networks (CNNs). We use HOG+GIST concatenated, the *fc6* layer of CaffeNet [17], and dense SIFT bag-of-words extracted in stride of 10 pixels at a single scale of 8 pixels. Following [41], we use CaffeNet's *fc6* since *fc7* and *fc8* may be capturing full objects and not be very useful for learning attributes.

Our training data consists of the images chosen for the gaze data collection experiments (Sec. 3.1), for a total of 300 for shoes and 480 for faces. The training labels are those provided by our human subject annotators. We perform a majority vote over the labels in case the annotators who labeled an image disagree over its label. We might have more positive images for an attribute than we have negatives, so we set the SVM classifier penalty on the negative class to the ratio of positive images to negative images. We use a linear SVM, and employ a validation set to determine the best value of the SVM cost C in the range [0.1, 1, 10, 100], separately for each attribute.

The test data consists of 341 images from **Shoes** and 660 from **Faces**. The test labels are those that came with the dataset. We pool together positive and negative test data for different attributes, so we often have significantly more negatives than positives for any given attribute. Thus, we use the F-measure because it more precisely captures accuracy when the data distribution is imbalanced.

Our proposed techniques for computing the spatial support of an attribute and extracting features accordingly, MULTIPLE TEMPLATES and MULTIPLE TEMPLATES PREDICTED, as well as their simplified versions SINGLE TEMPLATE and SINGLE TEMPLATE PREDICTED, were compared with the following baselines:

- using the whole image for both training and testing (WHOLE IMAGE);
- DATA-DRIVEN, a baseline which selects features using an L1-regularizer over features extracted on a

grid, then sets grid template cells on/off depending on whether at least one feature in that grid cell received a non-zero weight from the regularizer (note we do this only for localizable features);

- UNSUPERVISED SALIENCY, a baseline which predicts standard saliency using a state-of-the-art method [18][5] but without training on our attribute-specific gaze data, and the resulting saliency map is then used to compute a template mask;
- RANDOM, a baseline which generates a random template over a 15x15 grid, where the number of 1-valued cells is equal to the number of 1-valued cells in the corresponding SINGLE TEMPLATE template; and
- an ensemble of random template classifiers (RANDOM ENSEMBLE), which is the random counterpart to the ensemble used by MULTIPLE TEMPLATES.

Finally, we compare our method to the SPATIAL EXTENT (SE) method of Xiao and Lee [48] which discovers the spatial extent of *relative* attributes. While we do not study relative attributes, this is the work that is most relevant to our approach, thus prompting the comparison. [48] form "visual chains" from which they then build heatmaps showing which regions in an image are most responsible for attribute strength. We are only able to perform a comparison for attributes that have relative annotations on our datasets, which we take from [23, 31]. We use these heatmaps as saliency predictions, which in turn are used to mask the image and perform feature selection and attribute prediction (with the SVM cost C chosen on a validation set). We use dense SIFT and bag-of-words as for our SINGLE TEMPLATE PREDICTED.

In Tables 1 and 2, we show results for SINGLE TEMPLATE and MULTIPLE TEMPLATES, for HOG+GIST and *fc6*, respectively. In all tables, "total avg" is the mean over the two per-attribute "avg" values above (for shoe and face attributes, respectively). Our MT performs better than the other approaches. In Tab. 1, MT improves the performance on shoes by 6 points or 10% (=0.66/0.60-1) relative to the second-best method, and on faces, it improves performance by 3 points or 7%. In Tab. 2, our method improves performance by 2% on shoes and 7% on faces. Our MT approach captures the different meanings that an attribute can have and its possible locations. In contrast, ST imposes a fixed template and ignores possible shades of meaning and distinctions between the images viewed.

	WI	ST	MT (ours)	DD	US	R	RE
feminine	**0.80**	0.78	0.71	0.74	0.79	0.74	0.75
formal	0.78	**0.81**	0.80	0.79	0.77	0.77	0.77
open	0.52	0.53	**0.57**	0.45	0.55	0.51	0.51
pointy	0.17	0.17	**0.46**	0.00	0.10	0.14	0.10
sporty	0.74	0.70	**0.76**	0.72	0.71	0.72	0.72
avg	0.60	0.60	**0.66**	0.54	0.58	0.58	0.57
Asian	0.24	**0.33**	0.30	0.22	0.25	0.21	0.21
attractive	0.71	0.74	**0.81**	0.71	0.73	0.75	0.75
baby-faced	0.03	0.06	0.04	0.06	0.06	0.06	0.06
big-nosed	0.47	0.35	**0.52**	0.41	0.39	0.40	0.31
chubby	0.46	0.46	0.43	0.38	0.39	0.43	0.44
Indian	0.24	0.21	0.22	0.18	0.24	0.25	**0.27**
masculine	0.69	0.71	**0.77**	0.69	0.71	0.73	0.75
youthful	0.69	0.65	**0.7**	0.68	0.67	0.68	0.68
avg	0.44	0.44	**0.47**	0.42	0.43	0.44	0.43
total avg	0.52	0.52	**0.57**	0.48	0.51	0.51	0.50

Table 1: F-measure using HOG+GIST features. WI = WHOLE IMAGE, ST = SINGLE TEMPLATE, MT = MULTIPLE TEMPLATES, DD = DATA-DRIVEN, US = UNSUPERVISED SALIENCY, R = RANDOM, RE = RANDOM ENSEMBLE. Bold indicates best performer excluding ties.

	WI	ST	MT (ours)	US	R	RE
feminine	**0.77**	0.73	0.66	0.70	0.69	0.74
formal	**0.63**	0.57	0.61	0.58	0.59	0.58
open	0.51	0.51	0.51	0.49	0.47	**0.53**
pointy	0.19	0.18	**0.38**	0.17	0.18	0.13
sporty	**0.82**	0.78	0.79	0.77	0.67	0.69
avg	0.58	0.55	**0.59**	0.54	0.52	0.53
Asian	0.25	**0.30**	0.22	0.26	0.21	0.24
attractive	0.72	0.73	**0.81**	0.77	0.71	0.73
baby-faced	0.08	**0.12**	0.09	0.10	0.09	0.09
big-nosed	0.46	0.44	**0.67**	0.44	0.40	0.31
chubby	**0.42**	0.37	0.41	0.35	0.34	0.32
Indian	**0.28**	0.13	0.27	0.22	0.16	0.13
masculine	0.7	0.67	0.71	0.66	0.69	**0.73**
youthful	0.65	0.60	**0.68**	0.58	0.61	0.64
avg	0.45	0.42	**0.48**	0.42	0.40	0.40
total avg	0.51	0.49	**0.54**	0.48	0.46	0.47

Table 2: F-measure using *fc6*. See legend in Tab. 1.

In Tab. 3, we examine the performance of SINGLE TEMPLATE PREDICTED and MULTIPLE TEMPLATES PREDICTED. We observe that *predicting* the gaze map, as opposed to using a fixed map, only helps to improve the performance of the proposed feature selection approach on a few attributes ("formal", "Asian" and "masculine" for STP vs ST, and "feminine" and "baby-faced" for MTP vs MT). This may be because for our face and shoe data, the object of interest is fairly well-centered (although faces can be rotated to some degree). We show some unthresholded predicted gaze maps in Fig. 4. Note how our raw gaze maps correctly detect cheeks as salient for "chubbiness", and shoe

[5] We used the authors' online demo to compute saliency on our images, as code was not available.

	WI	ST	MT (ours)	STP	MTP (ours)	DD	US	SE	R	RE
feminine	**0.83**	0.80	0.60	0.78	0.62	0.68	0.63	0.79	0.78	0.82
formal	0.75	0.75	**0.81**	0.76	0.76	0.55	0.66	0.78	0.75	0.74
open	0.53	0.58	0.57	0.53	0.56	0.30	0.43	**0.59**	0.50	0.57
pointy	0.16	0.30	0.53	0.10	0.48	0.55	0.00	**0.56**	0.23	0.20
sporty	0.74	0.81	**0.82**	0.80	0.77	0.54	0.66	0.72	0.70	0.72
avg	0.60	0.65	0.67	0.59	0.64	0.52	0.48	**0.69**	0.59	0.61
Asian	0.22	0.28	**0.32**	0.30	0.26	0.24	0.29	N/A	0.23	0.24
attractive	0.61	0.80	**0.84**	0.80	0.82	0.69	**0.84**	N/A	0.76	0.77
baby-faced	0.06	0.11	0.07	0.06	0.10	0.09	0.06	N/A	0.08	**0.22**
big-nosed	**0.64**	0.33	0.43	0.27	0.40	0.41	0.32	N/A	0.27	0.15
chubby	0.36	0.34	**0.40**	0.30	0.36	0.24	0.24	0.32	0.27	0.29
Indian	**0.25**	0.15	0.24	0.12	0.18	0.12	0.20	N/A	0.16	0.08
masculine	0.68	0.68	0.78	0.71	0.70	0.63	**0.80**	0.71	0.69	0.72
youthful	0.65	0.62	0.66	0.58	0.63	0.53	0.60	**0.69**	0.61	0.60
avg	0.43	0.41	**0.47**	0.39	0.43	0.37	0.42	N/A	0.38	0.38
total avg	0.52	0.53	**0.57**	0.49	0.53	0.45	0.45	N/A	0.49	0.50

Table 3: F-measure using gaze maps predicted using the saliency method of [19]. STP = SINGLE TEMPLATE PREDICTED, MTP = MULTIPLE TEMPLATES PREDICTED, SE = SPATIAL EXTENT. Other abbreviations are as before.

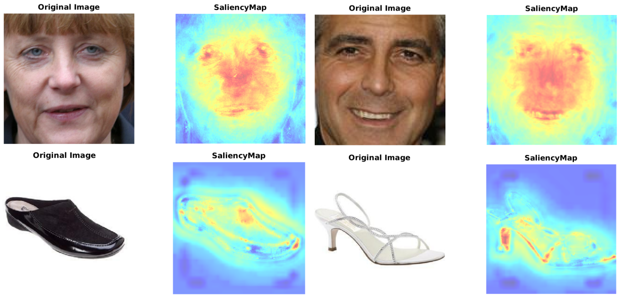


Figure 4: Representative predicted templates for "chubby" and "pointy". Red = most, blue = least salient.

toes and heels as salient for "pointiness".

As before, our best results are achieved by using multiple templates. The MT method outperforms the standard way of learning attributes, namely WI, by 10% on average.

In terms of region selection baselines, the RANDOM and RANDOM ENSEMBLE baselines perform somewhat worse than WHOLE IMAGE. The SINGLE TEMPLATE method performs similar to WHOLE IMAGE (slightly better or worse, depending on the feature type). In contrast, our MULTIPLE TEMPLATES perform much better. This indicates that capturing the meaning of an attribute does indeed lie in determining where the attribute lives, by also accounting for different participants' interpretations. The DATA-DRIVEN baseline performs weaker than the random baselines and our method, indicating the need for rich human supervision. The UNSUPERVISED SALIENCY baseline outperforms our method in a few cases (e.g. "feminine"), but overall performs similarly to RANDOM ENSEMBLE and weaker than our multiple template methods. Thus, attribute information is required to learn accurate gaze templates.

The results of [48] (SPATIAL EXTENT) are better than MT for four of the eight attributes available to test for SE, but the average over the eight attributes is almost the same (ours is slightly higher). However, for each attribute, SE required *38 hours* to run on average, on 2.6GHz Xeon processor with 256GB RAM. In contrast, our method only requires the time to capture the gaze maps, i.e. about *one hour*. In Fig. 6 (a), we compare MT with different configurations of SE that take a different amount of time to compute. (The results in Tab. 3 used the original most expensive setting.) Overall our method has similar or better performance than the different runs of SE, but it requires much less time.

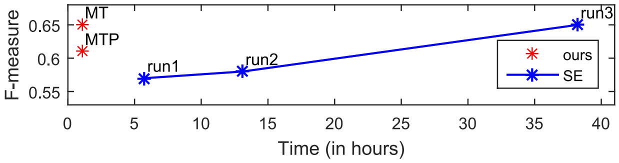


Figure 5: Time comparison of our MT and MTP with SE. On the y-axis is the average F-measure over the attributes tested. Run1, run2, and run3 use different parameter configurations for SE (each one requiring more processing time). Our MT is *more accurate* than the cheaper SE versions and *as accurate* as the most expensive one.

4.2. Adaptation for scene attributes

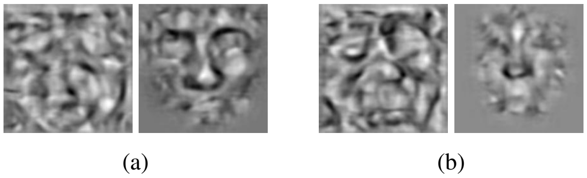
Similar to [48], the method most relevant to our work, we have so far only attempted our method on faces and shoes. Given our encouraging performance, we also tested it on ten scene attributes [33] (see Tab. 4 for the list), using 60 images per attribute for training and 700 for testing.

A direct application of our MT and MTP performed weaker or similar to WI, likely because scene images contain more than one object. Thus, we adapted our method for this dataset, using five seconds of gaze data. The intuition for our adapted method is as follows: For the attributes "natural" and "sailing", people might look at e.g. trees and water, respectively. Thus, we can use *objects* as cues for where people will look. Such an approach computes location-invariant masks that depend on *what* is portrayed, not *where* it is portrayed.

Our approach consist of three steps: learning an object detector, modeling attributes via objects, and predicting attributes on test images. We fine-tuned the VGG16 network [43] with object annotations from SUN [49] on images not contained in our gaze experiments or test set. We trained three CNNs grouping the objects with similar bounding box size. To learn attributes, we first ran the object detector on our training images. For a given attribute, we counted how many objects intersect with its gaze fixations. Next, we normalized these values and compiled a list of the five most frequently fixated, hence *most relevant* categories for each

Attribute	Relevant objects
climbing	mountain, sky, tree, trees, building
open area	sky, trees, grass, road, tree
cold	tree, building, mountain, sky, trees
soothing	trees, sky, wall, floor, tree
competing	wall, floor, grass, trees, tree
sunny	sky, tree, building, grass, trees
driving	sky, road, tree, trees, building
swimming	tree, trees, water, sky, building
natural	trees, tree, grass, sky, mountain
vegetation	tree, trees, sky, grass, road

Table 4: The objects most often fixated per scene attribute.



(a) (b)

Figure 6: Model visualizations for (a) the attribute "baby-faced", using whole image features (left) and our template masks (right), and (b) the attribute "big-nosed".

attribute. At test time, if at least one of these is present, we predict the attribute is present as well.

This simple approach achieves an average F-measure of 0.37, compared to 0.33, 0.34 and 0.45 for WI with HOG+GIST, dense SIFT, and *fc6*, respectively. It outperforms *fc6* on the attributes "driving" and "open area". A more elaborate approach which extracts *fc6* features on a grid and masks out cells of the grid based on overlap with relevant objects, achieves 0.40.

The objects selected per attribute are shown in Tab. 4. We observe that for "natural", the fixated objects are trees, grass, sky and mountains; for "driving", one of the objects is road, for "swimming" water, and for "climbing" mountains and buildings. This result confirms our intuition that scene attributes can be recognized by detecting relevant objects associated with the attributes through gaze. In our future work, we will formulate this intuition such that it allows us to outperform whole-image *fc6* features on more attributes.

4.3. Visualizing attribute models

We conclude with two applications of our method. First, our gaze templates can be employed to visualize attribute classifiers. We use Vondrick et al.'s Hoggles [45], a method used for object model visualization, and apply it to attribute visualization, on (1) models learned from the whole image, and (2) models learned from the regions chosen by our templates. We show examples in Fig. 6. Using the templates produces more meaningful visualizations than using the whole image. For example, for the attribute "baby-faced", our visualization shows a smooth face-like image that highlights the form of the nose and the cheeks, and for "big-nosed", we see a focus on the nose.

4.4. Using gaze to find schools of thought

Kovashka and Grauman [21] show there exist "schools of thought" (groupings) of users in terms of their judgments about attribute presence. They discover these groupings and use them to build accurate attribute sub-models, each of which captures an attribute variation (e.g. open at the toe as opposed to at the heel). The goal is to disambiguate attributes and create clean attribute models. First, they build a "generic" model (by pooling labels from many annotators). They discover schools using the users' labels, by clustering in a latent space representation for each user, computed using matrix factorization on the annotators' sparse labels. Then they use domain adaptation techniques to adapt this "generic" model towards sparse labeled data from each school. At test time, they apply the user's group's model to predict the labels on a sample from that user. We follow the same approach, but employ gaze to discover the schools.

We factorize an (annotator, image) table where the entry for annotator i and image j is the cluster membership of image j, computed by clustering images using their gaze maps on positive and negative annotations separately. Thus, for each user, we capture what type of gaze maps they provide, using the intuition that how a user perceives an attribute affects where he/she looks. On our data, the original method of [21] achieves 0.37, and our gaze-based discovery achieves 0.40. Our method is particularly useful for the attributes "big-nosed" (0.41 vs 0.29 for [21]), "masculine" (0.40 vs 0.35), "feminine" (0.43 vs 0.36), "open" (0.58 vs 0.52), and "pointy" (0.43 vs 0.36), most of which are fairly subjective.[6] This indicates using gaze is very informative for disambiguating attributes, the original goal of [21].

5. Conclusion and Future Work

We showed an approach for learning more accurate attribute prediction models by using supervision from humans in the form of gaze locations. These locations indicate where in the image space a given attribute "lives". We demonstrate that on a set of face and shoe attributes, our method improves performance compared to six baselines including alternative methods for selecting relevant image regions. This indicates that human gaze is an effective cue for learning attribute models. We also show applications of gaze for attribute visualization and finding users who perceive an attribute in similar fashion.

In future work, we will explore learning from *sequences* of gaze locations, as in work on scanpaths [51]. Modeling how human gaze moves over the image might provide more information than modeling gaze statically. We will also explore modeling the commonalities between gaze maps for the same attribute, and the distinctions between maps for different attributes, using convolutional neural networks.

[6] See our supplemental file for the full results.

References

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

[2] L. Chen, Q. Zhang, and B. Li. Predicting multiple attributes via relative multi-task learning. In *CVPR*, 2014.

[3] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[4] J. Deng, J. Krause, M. Stark, and L. Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *TPAMI*, 38(4):666–676, 2016.

[5] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *ICCV*, 2011.

[6] V. Escorcia, J. C. Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *CVPR*, 2015.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by Their Attributes. In *CVPR*, 2009.

[8] D. F. Fouhey, A. Gupta, and A. Zisserman. 3D shape attributes. In *CVPR*, 2016.

[9] C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, 2016.

[10] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *TPAMI*, 34(10):1915–1926, 2012.

[11] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *ICCV*, 2007.

[12] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *CVPR*, 2015.

[13] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015.

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, (11):1254–1259, 1998.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[16] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[18] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *ICCV*, 2015.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[20] A. Kovashka and K. Grauman. Attribute Adaptation for Personalized Image Search. In *ICCV*, 2013.

[21] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *IJCV*, 114(1):56–73, 2015.

[22] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image Search with Relative Attribute Feedback. In *CVPR*, 2012.

[23] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *IJCV*, 115(2):185–210, 2015.

[24] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively Selecting Annotations Among Objects and Attributes. In *ICCV*, 2011.

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

[26] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In *CVPR*, 2009.

[27] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[28] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, 2006.

[29] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV*. Springer, 2014.

[30] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *CVPR*, 2011.

[31] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[32] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*. Springer, 2012.

[33] G. Patterson and J. Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*, 2012.

[34] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463, 2002.

[35] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. Springer, 2012.

[36] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[37] R. N. Sandeep, Y. Verma, and C. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014.

[38] G. Schwartz and K. Nishino. Automatically discovering local visual material attributes. In *CVPR*, 2015.

[39] S. Shankar, V. K. Garg, and R. Cipolla. Deep-carving: Discovering visual attributes by carving deep neural nets. In *CVPR*, 2015.

[40] J. Shao, K. Kang, C. C. Loy, and X. Wang. Deeply learned attributes for crowded scene understanding. In *CVPR*, 2015.

[41] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*, 2014.

[42] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[44] L. Von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *CHI*, 2006.

[45] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013.

[46] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, 2016.

[47] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*, 2013.

[48] F. Xiao and Y. J. Lee. Discovering the spatial extent of relative attributes. In *ICCV*, 2015.

[49] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[50] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[51] A. Yarbus. Eye movements and vision. 1967. *New York*, 1967.

[52] F. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012.

[53] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[54] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, 2013.

[55] O. Zaidan, J. Eisner, and C. D. Piatko. Using" annotator rationales" to improve machine learning for text categorization. In *HLT-NAACL*, pages 260–267. Citeseer, 2007.