Involving humans to learn attributes

Semantic visual attributes (such as "metallic" or "smiling") have been used for a variety of tasks: as a low-dimensional representation for object recognition [1, 5, 9, 11, 2], as a textual representation to recognize unseen categories [1, 5, 7, 2, 6], as a supervision modality for active learning [4, 8], 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".

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.

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.

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, and shoes, 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.

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, achieves competitive attribute prediction accuracy compared to five methods in Table 1, desccribed below.

- 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

Preprint. Under review.

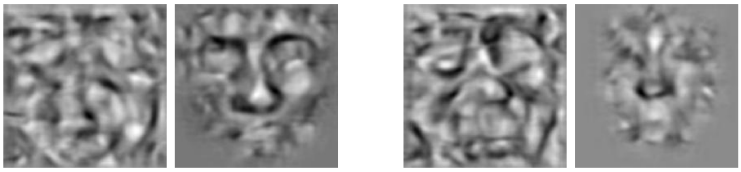
	WI	ST	MT (**ours**)	STP	MTP (**ours**)	DD	US	R	RE
feminine	**0.83**	0.80	0.60	0.78	0.62	0.68	0.63	0.78	0.82
formal	0.75	0.75	**0.81**	0.76	0.76	0.55	0.66	0.75	0.74
open	0.53	**0.58**	0.57	0.53	0.56	0.30	0.43	0.50	0.57
pointy	0.16	0.30	0.53	0.10	0.48	**0.55**	0.00	0.23	0.20
sporty	0.74	0.81	**0.82**	0.80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

Table 1: F-measure using bag-of-words SIFT features. WI = WHOLE IMAGE, ST = SINGLE TEMPLATE, MT = MULTIPLE TEMPLATES, STP = SINGLE TEMPLATE PREDICTED, MTP = MULTIPLE TEMPLATES PREDICTED, DD = DATA-DRIVEN, US = UNSUPERVISED SALIENCY, R = RANDOM, RE = RANDOM ENSEMBLE. Bold indicates best performer excluding ties.

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 [3][1] 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.

In addition to this quantitative result, we demonstrate an application showing how our method can be used to visualize attribute models. We use Vondrick et al.'s Hoggles [10], 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. 2. 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.



(a) (b)

Figure 2: Model visualizations for (a) the attribute "baby-faced", using whole image features (left) and our template masks (right), and (b) the attribute "big-nosed".

To conclude, 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.

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

References

[1] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing Objects by Their Attributes. In *CVPR*, 2009.

[2] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

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

[4] Adriana Kovashka, Sudheendra Vijayanarasimhan, and Kristen Grauman. Actively Selecting Annotations Among Objects and Attributes. In *ICCV*, 2011.

[5] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In *CVPR*, 2009.

[6] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[7] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011.

[8] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *ECCV*. Springer, 2012.

[9] Genevieve Patterson and James Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*, 2012.

[10] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013.

[11] Felix X Yu, Liangliang Cao, Rogerio Schmidt Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.