Improving scene attribute recognition using web-scale object detectors

Frederick Tung and James J. Little

Department of Computer Science, University of British Columbia
{ftung,little}@cs.ubc.ca

Abstract. In recent years there has been growing interest in describing scenes using semantic attributes. While traditionally scenes have been analyzed using *global* image features such as Gist, recent studies suggest that humans often describe scenes in ways that are naturally characterized by *localized* image evidence. In particular, humans often describe scenes by their functions or affordances, which are largely suggested by the objects in the scene. In this paper, we show that the aggregated responses of a large collection of modern object detectors trained at the web scale can be used to derive effective high-level features for scene attribute recognition. On the SUN Attribute benchmark, these detector-based features obtain 2-9% higher average precision than traditional Gist, HOG2x2, self-similarity, and geometric context color histograms.

1 Introduction

In recent years, there has been increasing interest in describing objects and scenes using semantic attributes [2], [3], [4], [7], [9]. Attributes are particularly appropriate for characterizing scenes because scene categories can exhibit wide intra-class variation, the scene space is continuous (i.e. there are smooth transitions between categories), and a single image can contain multiple scene categories [8].

Traditionally, scenes have been analyzed using global image features such as Gist [6] or HOG visual words [5], [10]. While global image features can be used to recognize many types of scene attributes (e.g. degree of naturalness or openness [6]), a large number of interesting scene attributes are not well captured by global features. In particular, many scene attributes in the crowd-sourced SUN Attribute database are more naturally characterized by *localized* image evidence. For example, Patterson and Hays [8] found that humans often describe scenes based on their functions or affordances, such as "camping" or "studying/learning". A strong cue for "camping" (defined as "either an actual camp site, or scene in wilderness suitable enough for humans to make a tent and/or sleep" [8]) would be the presence of a tent.

The core contribution of our paper can be summarized as follows. Motivated by recent findings that humans often describe scenes by their functions or affordances, which are largely suggested by the objects in the scene, we propose leveraging open, web-scale object detector responses to improve scene attribute recognition. We demonstrate the effectiveness of this simple idea on the standard scene attribute benchmark [8].

2 Method

We explore the hypothesis that since many scene attributes are characterized by localized image evidence, recognition of scene attributes can be improved using a large collection of object detectors trained at the web scale.

Inspired by a recent development in text understanding, Chen et al. [1] recently introduced the Never Ending Image Learner (NEIL). NEIL is an iterative, semi-supervised algorithm that learns objects and their relationships from downloaded web images. In each iteration, only the most confident detections and relationships are added to the knowledge base, a strategy the authors refer to as "macro-vision". Detectors are based on color HOG features. In this work, we use the most recent collection of object detectors released by NEIL (as of June 2014, this was the Dec. 2013 version), which consists of 8685 detectors spanning 1190 unique object categories.

Let $f_c(I, w)$ denote the response of an object detector of category c evaluated in window w of an image I. We form a feature descriptor F for image I that concatenates the responses of all object detectors in the collection, max-pooled over the image windows:

$$F[c] = \max_w f_c(I, w), \quad c = 1, ..., D \tag{1}$$

where $F[c]$ denotes the c^{th} component of the feature vector $F \in \mathbf{R}^D$, and D is the number of object detectors in the collection.

In high dimensional feature spaces, the partial order statistics of a feature descriptor are often more robust for classification and retrieval tasks than the descriptor's precise numeric values [11]. Following [11], we capture partial order statistics by taking random subsets of the feature descriptor's dimensions. Specifically, we derive an ordinal feature representation of the detector response vector F as follows. Given a subset size $K \ll D$, generate m random ordered subsets of size K of the dimensions in F (in general, larger m captures more ordinal relationships but increases memory requirements as the transformed features are larger). That is, each ordered subset θ consists of K unique indices from 1 to D: $\theta \in \{1...D\}^K$. Denote by Θ the matrix formed by stacking all m ordered subsets: $\Theta \in \{1...D\}^{m \times K}$. Form an intermediate matrix $Z \in \mathbf{R}^{m \times K}$ by looking up the entries in F corresponding to the indices in Θ:

$$Z[i, j] = F[\Theta[i, j]] \tag{2}$$

where i and j are row and column identifiers. Next, let $\tilde{x} \in \mathbf{N}^m$ collect the indices of the largest elements of each row in Z:

$$\tilde{x}[i] = \arg\max_j Z[i, j] \tag{3}$$

Each entry in $\tilde{x}$ encodes the maximum rank information for the corresponding random ordered subset θ. The final ordinal feature representation x is a binary output encoding of $\tilde{x}$ in which each scalar in $\tilde{x}$ is translated into a binary indicator vector of length K. Hence, $x \in \{0, 1\}^{m \times K}$. An example of the ordinal feature transform is illustrated in Figure 1.

$$F = \begin{bmatrix} -0.3 & 0.6 & 0.1 & -0.7 & -0.4 & 0.2 \end{bmatrix}$$
$$K = 2, \quad m = 4:$$
$$\Theta = \begin{bmatrix} 1 & 4 \\ 2 & 6 \\ 5 & 3 \\ 4 & 6 \end{bmatrix} \quad Z = \begin{bmatrix} -0.3 & -0.7 \\ 0.6 & 0.2 \\ -0.4 & 0.1 \\ -0.7 & 0.2 \end{bmatrix} \quad \tilde{x} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$
$$x = \begin{bmatrix} 1 & 0 & | & 1 & 0 & | & 0 & 1 & | & 0 & 1 \end{bmatrix}$$

Fig. 1. Toy example of ordinal feature transform. An ordinal feature representation x is derived from a vector of max-pooled object detector responses F using partial order statistics from random ordered subsets of the dimensions of F.

3 Experiments

We conducted experiments using the SUN Attribute database [8], which contains over 14,000 scenes spanning over 700 categories. Attribute annotations are sourced for each image individually instead of assumed the same for all images of the same scene category. The 102 crowd-sourced scene attributes span materials, surface properties, functions or affordances, and spatial envelope [6] properties. Given a semantic attribute of interest, we train a linear SVM, setting the regularization parameter C by five-fold cross-validation. For a direct comparison with the SUN Attribute database baseline that uses global image features, we train each attribute classifier independently and do not take advantage of possible correlation cues.

Table 1 summarizes experimental results using the standard train and test splits of the benchmark. We report results using both the ordinally transformed features x and the original max-pooled detector responses F. To our surprise, we found that detector-based features perform better than any individual global image feature baselined in the SUN Attribute database, including Gist (+7%), HOG2x2 (+2%), self-similarity (+5%), and geometric context color histograms (+9%). Applying the ordinal transform ($K = 2$ and $m = 64$k) provides only a small improvement in accuracy.

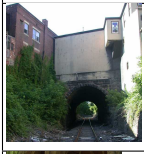
To demonstrate that the information captured by the object detectors is *complementary* to that captured by traditional global image features, we also combined the two types of image evidence by a simple average. As Table 1 shows, combining both types of image evidence enables higher accuracy than either type by itself. In addition to the motivating function or affordance based attributes, we found that detector-based features broadly outperform traditional global features in recognizing scene attributes related to materials, surface properties, and spatial envelope properties.

Table 2 shows sample qualitative results. Each row shows an image from the SUN Attribute database, the most confident attributes predicted using the linear SVMs trained on ordinally transformed detector response features, and the ground truth annotation.

Table 1. Average precision (AP) results on the SUN Attribute benchmark, standard training and testing splits

Individual features	
Global image features [8]	
Geometric context color histogram	0.783
Gist	0.799
Self-similarity	0.820
HOG2x2	0.848
Web-scale object detector responses F	**0.864**
Web-scale object detector responses with ordinal transform x	**0.868**
Combination of multiple features	
Combined normalized kernel of global image features (Geo. + Gist + SS + HOG) [8]	0.879
Combined normalized kernel of global image features (Geo. + Gist + SS + HOG) + Web-scale object detector responses with ordinal transform	**0.888**

Table 2. Example attribute predictions using web-scale object detector responses. Predicted attributes that match the ground truth are highlighted in green.

Query image	Most confident attributes (in order of confidence)	Ground truth attributes (unordered)
	vegetation, grass, foliage, leaves, playing, farming, shrubbery, camping, trees	camping, trees, grass, vegetation, shrubbery, foliage, leaves, natural, open area
	open area, far-away horizon, natural light, dirt/soil, rugged scene, sand, dry, concrete, dirty	driving, biking, transporting things or people, camping, fencing, natural light, dry, man-made, open area, far-away horizon
	glass, no horizon, natural light, leaves, foliage, man-made, flowers, shrubbery, semi-enclosed area, vegetation	vegetation, foliage, leaves, brick, glass, natural light, man-made, open area, semi-enclosed area, no horizon
	no horizon, praying, symmetrical, mostly vertical components	vacationing/ touring, man-made, enclosed area, no horizon
	enclosed area, carpet, no horizon, eating, paper, studying/learning, electric/indoor lighting, vinyl/linoleum, working, reading	studying/learning, eating, playing, carpet, tiles, rubber/plastic, enclosed area, no horizon, cluttered space, soothing

References

1. Chen, X., Shrivastava, A., Gupta, A.: NEIL: extracting visual knowledge from web data. In: Proc. IEEE International Conference on Computer Vision (2013)
2. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785 (2009)
3. Kovashka, A., Parikh, D., Grauman, K.: WhittleSearch: image search with relative attribute feedback. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2012)
4. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958 (2009)
5. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing via label transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(12), 2368–2382 (2011)
6. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
7. Parikh, D., Grauman, K.: Relative attributes. In: Proc. IEEE International Conference on Computer Vision. pp. 503–510 (2011)
8. Patterson, G., Hays, J.: SUN Attribute database: discovering, annotating, and recognizing scene attributes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2751–2758 (2012)
9. Sadovnik, A., Gallagher, A., Parikh, D., Chen, T.: Spoken attributes: mixing binary and relative attributes to say the right thing. In: Proc. IEEE International Conference on Computer Vision (2013)
10. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (2010)
11. Yagnik, J., Strelow, D., Ross, D.A., Lin, R.: The power of comparative reasoning. In: Proc. IEEE International Conference on Computer Vision. pp. 2431–2438 (2011)