Attributes make sense on segmented objects*

Zhenyang Li, Efstratios Gavves, Thomas Mensink, and Cees G. M. Snoek

ISLA, Informatics Institute, University of Amsterdam, The Netherlands

Abstract. In this paper we aim for object classification *and* segmentation by attributes. Where existing work considers attributes either for the global image or for the parts of the object, we propose, as our first novelty, to learn and extract attributes on segments containing the entire object. Object-level attributes suffer less from accidental content around the object and accidental image conditions. As our second novelty, we propose joint learning for simultaneous object classification and segment proposal ranking, solely on the basis of attributes. This naturally brings us to our third novelty: object-level attributes for zero-shot, where we use attribute descriptions of unseen classes for localizing their instances in new images and classifying them accordingly. Results on the Caltech UCSD Birds, Leeds Butterflies, and an a-Pascal subset demonstrate that *i)* extracting attributes on oracle object-level brings substantial benefits *ii)* our joint learning model leads to accurate attribute-based classification and segmentation, approaching the oracle results and *iii)* object-level attributes also allow for zero-shot classification and segmentation. We conclude that attributes make sense on segmented objects.

1 Introduction

The goal of this paper is object classification *and* segmentation using attributes. Representing an image by attributes [7, 8, 10] like *big ear*, *trunk*, and *gray color* is appealing when examples are rare or non-existent, feature encodings are non-discriminative, or a semantic interpretation of the representation is desired. Consequently, attributes are a promising solution for many current challenges in computer vision [4, 9]. Different from existing work, which computes object attributes either on the entire image [1, 10, 12] or on parts of the object [3, 5, 6, 8], we propose to learn the best possible segment that contains the entire object and compute all attributes on this segment. Inspired by Akata *et al.* [1], who adapt the model of [15] and propose attribute embedding learning for supervised and zero-shot object classification, we also optimize attribute learning for object classification, including the challenging zero-shot setting. However, we observe that attributes most often reflect object level properties, *e.g.* that an antelope has a *pointy snout*. Hence, reasoning these attributes *pointy snout* on the object segments instead of the whole images is more intuitive and accurate, see Fig. 1.

* The full paper is accepted for the European Conference on Computer Vision, 2014.

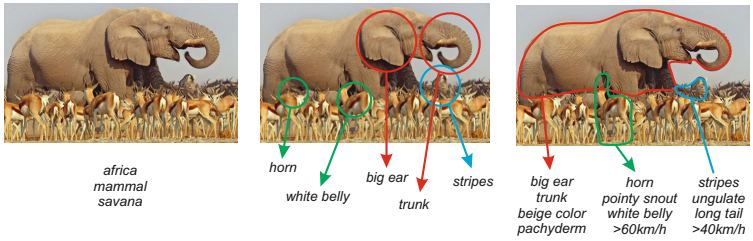

Fig. 1. Attributes make sense on segmented objects. Illustration of different level of attributes: (a) Considering the full image, one can only expect to describe generic attributes that apply to the whole scene. (b) When localizing attributes, one faces the problem that not all attributes can be localized. Partial occlusions, small scales, or uncommon viewpoints might reduce the visibility of a particular attribute. (c) Object-level attributes are constrained on segmented objects, allowing for description of object specific attributes, and furthermore helping to suppress irrelevant background signal.

2 Object-level Attributes

Given an image x, our classification function f is defined as follows:

$$f(x) = \arg\max_{y \in \mathcal{Y}} \max_{z \in Z(x)} F(z, y), \tag{1}$$

where z is a latent variable, $Z(x)$ indicates a set of segment proposals for image x, and $F(z, y)$ is a compatibility function between segment z and label y. This function returns the label with best score over all segments. We do not assume the object bounding box or object segmentation is known at prediction time.

Attribute embedding. We follow the label and attribute embedding approaches from [1,15], where each class label y is embedded in the m-dimensional space of attributes by $\phi(y) \in \mathbb{R}^m$. While [1, 15] embed the full image features, we embed the visual features of a segment z only with $\theta(z) \in \mathbb{R}^d$. In this work we use Fisher vector [13] for this visual embedding. The compatibility function $F(z, y)$ is defined as:

$$F(z, y; W, \phi) = \theta(z)' W \phi(y), \tag{2}$$

where $W \in \mathbb{R}^{d \times m}$ is the model parameter matrix to learn. We stack the attribute embeddings of each class $\phi(y)$ into an embedding matrix Φ for all classes.

We assume there is a collection of training images $\{(x_i, y_i, z_i)\}_{i=1}^N$, in which each image x_i has a ground truth label y_i and a ground truth object segment z_i. There exists a mapping from attributes to classes $\Phi^{\mathcal{A}}$, which defines the relevant attributes for each class. For learning we employ structured risk minimization, using a ranking objective built upon [1,2,15]. The loss function of a ground-truth image/label/segment triplet (x_i, y_i, z_i) for a label y, is defined as:

$$\ell(y, z_i, y_i, x_i) = \max_{z \in Z(x_i)} \Delta(z, y, y_i, z_i) + F(z, y) - F(z_i, y_i). \tag{3}$$

The Δ function, which determines the margin, is defined as:

$$\Delta(z, y, z_i, y_i) = \begin{cases} 1 - O(z, z_i) & \text{if } y = y_i, \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

where $O(z, z_i)$ is the intersection over union (IoU) between two segments. Similar to [1,15], we define the empirical risk $R(W, \Phi)$ as a weighted ranking loss over ℓ.

Fully supervised learning. In the fully supervised case, where we have visual examples from all classes, we minimize the following regularized objective:

$$\min_{W,\Phi} \frac{\lambda}{2}||W||^2 + \frac{\mu}{2}||\Phi - \Phi^{\mathcal{A}}||^2 + R(W, \Phi), \tag{5}$$

where λ and μ are trade-off parameters between the empirical risk and the regularization. Regularizing towards the pre-defined class-to-attribute encoding $(\Phi - \Phi^{\mathcal{A}})$ allows us to exploit this high-level semantic prior. This could be particularly beneficial when just a few examples per class are available.

Zero-shot learning. In the setting of zero-shot classification, visual training examples are given only for a subset of the classes, while evaluation is performed on a disjoint set of the classes. In this case the attribute embedding is fixed to the existing mapping $\Phi = \Phi^{\mathcal{A}}$, and Eq. 5 reduces to:

$$\min_{W} \frac{\lambda}{2}||W||^2 + R(W, \Phi^{\mathcal{A}}). \tag{6}$$

To efficiently sovle the problem of maximization over latent segments in Eq. 1 and Eq. 3, we make use of the codemaps framework [11].

3 Experiments

We conduct our main experiments on three datasets: the Caltech UCSD Birds 2011 dataset, the Leeds Butterfly dataset and a subset of the a-Pascal [7] dataset (a-Pascal++). For visual features we use the Fisher vector [13] with different GMM codebook size k, computed on dense RGB-SIFT [14] extracted every 2 pixels and at multiple scales, and projected to 80 dimensions using PCA. We use two measures for evaluation: the *mean class accuracy* (MCA), where for each class the top-1 accuracy is computed and averaged over all classes, and the *mean class accuracy over correctly segmented objects* (MSO). MSO is computed similar to MCA, except that a prediction is considered correct only if both the label is correct and the overlap of the latent segment with the ground-truth segmentation meets the Pascal VOC criterion (IoU greater than 50%).

Object-level attributes on latent segments. In this experiment we compare a full-image feature embedding to using an embedding of an oracle provided bounding box or segment. We train the model with the ALE framework [1]. We also evaluate the ability of our approach inferring the object segment as a latent variable in the model and to classify the segmented objects using attributes. We

Dataset	Codebook	Entire image	Oracle bbox	Oracle segment	Object-level attributes	
		MCA	MCA	MCA	MCA	MSO
CUB-2011	$k = 16$	13.8	25.8	43.9	35.2	29.9
	$k = 256$	21.4	36.4	52.9	39.2	35.5
Butterflies	$k = 16$	83.8	96.9	99.1	96.4	95.5
a-Pascal++	$k = 256$	30.6	33.6	40.2	35.0	24.7

Table 1. Object-level attributes on latent segments. We compare the performance of ALE [1] using full-image embeddings with oracle bounding box/segment embeddings, and our proposed learning object-level attributes on latent segments.

Dataset	Codebook	Entire image	Object-level attributes	
		MCA	MCA	MSO
CUB-2011	$k = 16$	11.3	15.7	12.4

Table 2. Object-level attributes for zero-shot classification on CUB-2011. Learning attributes on latent segments is able to not only improve the zero-shot classification, but also return the segmentations of objects that belong to unseen classes.

present the aggregated results in Table 1. We observe that by using oracle object segments we obtain large increase accuracy over using full images, as well as oracle bounding boxes. These numbers serve as an upper bound of the classification accuracy that we may obtain by using latent segments. Our approach learning attributes on latent segments improves the accuracy over full-image results of [1] by around 4-21%. The quality of our inferred segmentations is quite good, since the MSO is resonablely close to MCA. Moreover, we observe that for a larger codebook the discrepancy between accurate prediction and accurate prediction with accurate segmentation is smaller.

Comparison with part-localized attributes. To compare our approach with a recent part-localized attribute model, we also conduct an experiment on a subset of CUB-2011: five categories consisting of different species of warblers. We follow the same experimental protocol as [6]. Our model of learning object-level attributes on latent segments scores 65.8% accuracy using a codebook of GMM k=16, and using full image embedding scores 42.2%, while the localized attribute model [6] reports ∼55%.

Object-level attributes for zero-shot. In this experiment we perform zero-shot learning, which allows for simultaneous classification and segmentation of the object of interest. We experiment on the CUB-2011 dataset, using the same 150 train classes and 50 test classes as in [1]. We present the numberical results in Table 2. It shows that we improve the zero-shot classification accuracy, while returning the segmentations of objects that belong to classes we have not seen before.

We conclude that our joint learning with object-level attributes leads to accurate classification and segmentation. It also improves zero-shot classification, allowing object segmentation for unseen classes.

References

1. Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.
2. M. Blaschko and C. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.
3. L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.
4. S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.
5. J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.
6. K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.
7. A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.
8. V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.
9. A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *ICCV*, 2013.
10. C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based transfer learning for object categorization with zero/one training example. *TPAMI*, 2013.
11. Z. Li, E. Gavves, K. van de Sande, C. Snoek, and A. Smeulders. Codemaps segment, classify and search objects locally. In *ICCV*, 2013.
12. D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.
13. J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.
14. K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 2010.
15. J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.