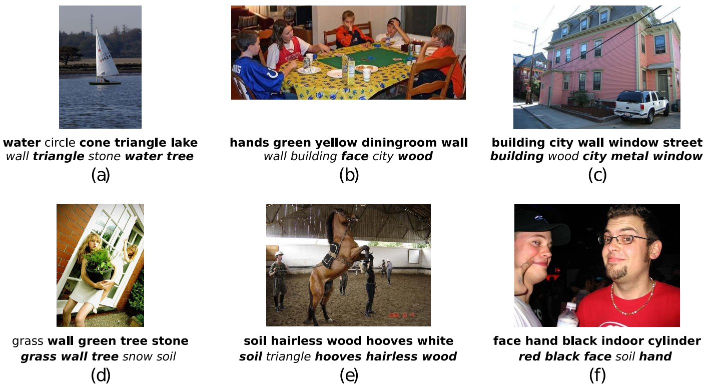
Improving object classification using semantic attributes

Yu Su
http://users.info.unicaen.fr/~ysu/

Moray Allan
http://users.info.unicaen.fr/~mallan/

Frédéric Jurie
http://users.info.unicaen.fr/~jurie/

GREYC
Université de Caen
14032 Caen Cedex
France



water circle **cone triangle lake**
wall ***triangle*** *stone* ***water tree***
(a)

hands green yellow diningroom wall
wall building ***face*** *city* ***wood***
(b)

building city wall window street
building *wood* ***city metal window***
(c)

grass **wall green tree stone**
grass wall tree *snow soil*
(d)

soil hairless wood hooves white
soil *triangle* ***hooves hairless wood***
(e)

face hand black indoor cylinder
red black face *soil* ***hand***
(f)

Figure 1: Semantic attributes predicted for example images: global attributes (first row), and local attributes (second row, oblique text). Predictions in bold text match the images' manual annotations.

1 Overview

This paper shows how semantic attributes can be used to improve object classification. The semantic attributes used fall into five groups: scene (*e.g.* 'road'), colour (*e.g.* 'green'), part (*e.g.* 'face'), shape (*e.g.* 'box'), and material (*e.g.* 'wood'). We train a set of classifiers for individual semantic attributes, and use them to make predictions on new images (Figure 1). We can then use the scores from the set of classifiers as a low-dimensional image representation. The object classification performance of the semantic attribute features alone is close to that of a much higher-dimensional bag-of-words image representation, while using the semantic attributes together with the bag-of-words features consistently improves performance.

2 Method

In this paper we will use 60 manually defined semantic attributes to describe images. We will learn a set of independent classifiers for these attributes, and use them to construct semantic image descriptors which can be used in object classification. For each attribute we downloaded a set of representative image results from Google and manually rejected false positives.

The attribute classifiers produce two types of feature, global and local. The global features are the result of running attribute classifiers on a whole image. To obtain the local features we sample 100 patches from an image, and run the attribute classifier on each. The result from each patch is averaged to produce a final feature. The local features are intended to help recognise attributes which only occur in localised areas of an image. We do not use local attribute features for the attributes 'indoor'/'outdoor', 'city'/'landscape', 'bedroom', 'dining room' or 'living room'. By concatenating the classifier output from 60 global semantic attributes and 55 local semantic attributes we obtain a 115-dimensional semantic attribute feature descriptor.

In our experiments each attribute classifier is a linear SVM, with each real-valued and non-negative SVM score providing one dimension in the overall semantic descriptor. Four image feature types are used as input to the SVMs: SIFT, texton filterbank, LAB and Canny edge detection. SIFT, texton, and LAB features were quantised to 1024, 256, and 128 k-means centres respectively, while Canny edge features were quantised to 8 orientation bins. Combining these features gives a 1416-dimensional feature descriptor. When using these features directly as a bag-of-words image representation, we additionally use spatial pyramid matching, as in [2].

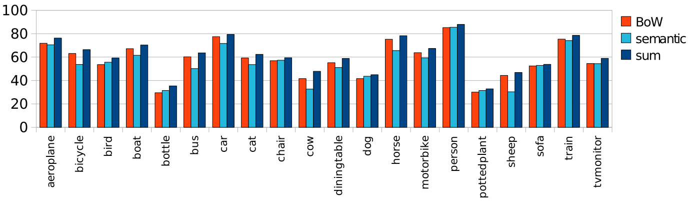


Figure 2: Average precision achieved using bag of words features, semantic attribute features, and both together, on the PASCAL VOC 2007 object class data.

Using a three level pyramid, 1 × 1, 2 × 2, 3 × 1, gives final bag-of-words features with a dimensionality eight times as high, but for semantic attribute prediction we keep the original single-level features as input to the classifiers to reduce the computational cost. Before feature computation, the images were scaled to be at most 300 × 300 pixels, with their original aspect ratios maintained.

3 Results

We tested our semantic attributes as features for object classification on the large data set used in the 2007 PASCAL VOC challenge [1]. Figure 2 compares the performance which is measured by average precision (AP) achieved by classifiers (SVMs with chi-square kernel) using bag-of-words features, the semantic attribute features, and a weighted sum combination of the two classifiers. On the majority of classes the classifier based on semantic attribute features performs worse than that using the original bag-of-words features, but it is impressive that this low-dimensional representation of the bag-of-words features performs almost as well on average, and outperforms the bag-of-words features on seven classes ('bird', 'bottle', 'chair', 'dog', 'person', 'potted plant', 'sofa'). The mean AP of all classes achieved by bag-of-words features and semantic features are 54.3% and 57.9% respectively. The performance on every class is increased by combining the two feature types rather than using either one alone. The mean AP achieved by the combined features is 61.4%, compared with two best results on 2007 PASCAL VOC challenge: 59.4% (VOC [1]) and 59.3% (LLC [3]).

Since our semantic attribute features are obtained from bag-of-words features by linear SVMs, they can be viewed as a dimensionality reduction method. Thus, as a comparison, we also give the performance of three other dimensionality reduction methods. The first method simply varies the number of visual words, the second method uses principal component analysis to project the larger bag-of-words feature set used previously into a lower-dimensional space, while the third method selects random directions in the high-dimensional space and projects onto those. The mean average precisions achieved by these features (with the same dimension as semantic features) are 35.2%, 39.0%, 32.3% respectively, which are much lower than the 54.3% achieved by semantic attribute features.

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 results. http://www.pascal-network.org/challenges/VOC/voc2007/.

[2] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.