

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

Abstract

This paper shows how semantic attribute features can be used to improve object classification performance. 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 classifiers from representative images for 60 semantic attributes. We first assess the accuracy of the individual classifiers, and show that they can be used to predict semantic annotations for test images. We then use output from the set of trained classifiers to create a new low-dimensional image representation. Experiments on data from the PASCAL VOC challenge show that the semantic attribute features achieve an object classification performance close to that of high-dimensional bag-of-words features, and that using a combination of semantic attribute features and bag-of-words features gives a better classification performance than using either feature set alone.

1 Introduction

Object classification has been a central area in computer vision research in recent years. Current computer vision systems perform well at recognising specific objects, but have trouble recognising many object categories: while humans typically find matching categories easier, machine vision systems are better at finding exact matches than learning more general categories. Even specific objects may appear in different poses, under varied lighting conditions, in cluttered images where irrelevant objects may add confusion or occlude the object of interest, but to recognise object classes we also need to deal with the often large appearance variations between class members.

When humans learn about new object classes, we make use of our existing knowledge of visual categories. For example, if we see a new animal we can immediately apply to it previously-learnt concepts of ‘grey’, ‘head’, ‘hooves’, and ‘wings’, and use these to recognise the class in future. As well as colours and object parts, this kind of shared ‘semantic attribute’ might describe common scene types (*e.g.* road), common shapes (*e.g.* box) and common materials (*e.g.* wood). Semantic attributes can also be used by computer vision systems, to guide learning to focus on common features of real-world images.

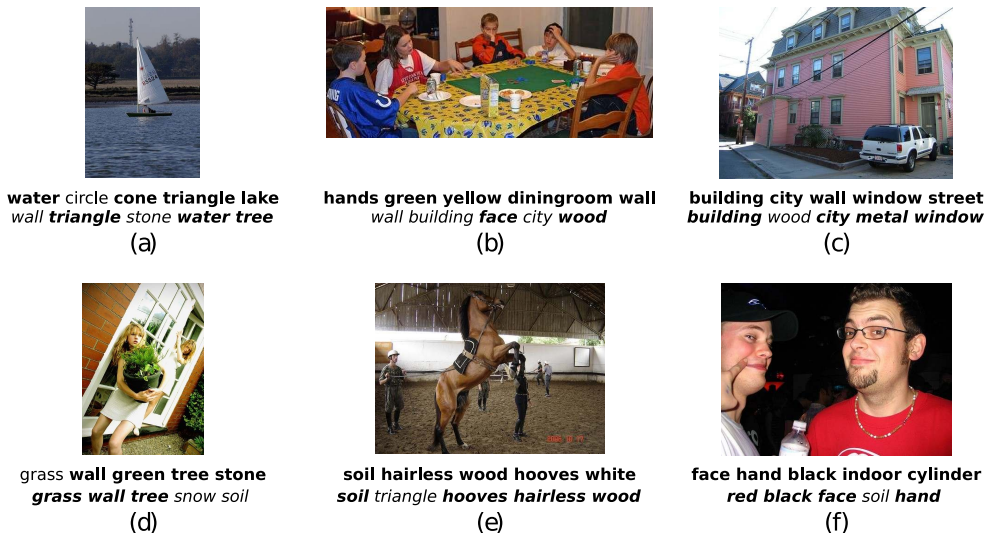


Figure 1: Semantic attributes predicted for example images: global attributes (first row), and local attributes (second row, oblique text). Predictions in bold text match the images’ manual annotations.

This paper shows how semantic attributes can be used to improve object classification. 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.

Section 2 discusses previous work related to the approach taken here. Section 3 describes the set of semantic attributes we use, and how we create semantic attribute classifiers. In our experiments we first evaluate the performance of the individual semantic attribute classifiers (Section 4.1), then look at using the semantic attribute features and bag-of-words features for object classification (Section 4.2). We also compare the semantic attribute features with other methods for image feature dimensionality reduction (Section 4.3). Section 5 discusses the results and suggests future directions.

2 Related work

The most popular current approach for object classification is to extract local regions from images, assign them to clusters, and use the count distribution across clusters to describe the images as input to a general classification method (see [16]). Some other approaches try to locate the object within the image, and to take into account the spatial relationship of the image regions which trigger potential matches (e.g. [3, 4, 15]), but these methods tend to have high computational complexity, making it feasible to learn about only a small number of object parts, and so limiting their descriptive power. The ‘bag of words’ approach instead uses the histogram of counts across the whole image to predict whether or not the

image contains the class of interest. While the image background may create confusion in recognising object classes, the background can also provide useful cues to aid recognition, since many objects tend to occur in particular types of scene [8, 11]. Simple bag-of-words methods have shown impressive performance for object classification when used with large numbers of region descriptor types and optimised parameter settings [9].

The local features used in bag-of-words methods typically lack any clear semantic meaning. The individual features do not usually have strong discriminative power, and the methods perform best when provided with very high-dimensional image descriptors which allow the discovery of significant differences between different classes' feature distributions. Learning features with more specific meanings might help improve classification performance. Some previous work has looked at explicitly learning semantically meaningful features. For example, van de Weijer et al. [12] learnt to map from image colour to the colour names people use to describe objects. Ferrari and Zisserman [5] also learnt simple texture attributes such as 'stripes' and 'dots'.

Recent work has embraced more complex attributes. Vogel and Schiele [13] used attributes describing scene, material, and shape to retrieve images of coasts, rivers/lakes, forests, plains, mountains, and sky/clouds. Quattoni et al. [10] proposed learning a set of classifiers to predict the presence or absence of content words in the captions associated with images in auxiliary training data. They then use SVD on the learnt classifier parameters to create a new image representation which reflects the semantic content of images. Farhadi et al. [2] used a set of semantic attributes such as 'hairy' and 'four-legged' to identify familiar objects, and to describe unfamiliar objects when an image and bounding box annotation is provided. Lampert et al. [6] showed that high-level descriptions in terms of semantic attributes can be used to recognise object classes without any example images, once semantic attribute classifiers are trained from other classes' data.

Like [2] and [6], we will use diverse semantic attributes with explicit meanings to describe the visual content of images. As in that work, we will learn classifiers for the semantic attributes then predict them from the low-level features of new images. Notable differences include that our set of semantic attributes is extended to include some relating to the overall scene, *e.g.* indoor/outdoor, 'sky' and 'street', providing contextual information which is important for object classification, and that we do not use any manual semantic attribute labels for images from the data set we test on, instead collecting a separate set of representative images. We hope that the semantic attribute classifiers learnt in this way are more general than those learnt on a specific data set. In our experiments below we evaluate the performance of semantic attribute features on the large data set used in the 2007 PASCAL VOC challenge [1], obtaining state-of-the-art results; we believe this is the first time that the performance of semantic attribute features has been tested on a standard object classification benchmark.

3 Semantic attributes

In this paper we will use five groups of semantic attributes to describe images: 'scene', 'colour', 'part', 'shape', and 'material'. Figure 2 lists the semantic attributes, grouped by type, along with some representative images for some attributes. 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. While the previous work mentioned above used only attributes of the objects themselves, we include attributes describing the overall image, since contextual information often helps in recognising objects.































scene		colour		part		shape		material	
lake	road	green	red	face	window	box	cylinder	leather	wood
									
									
									
+		+		+		+		+	
indoor	living room	black		flippers		circle		ceramics	metal
outdoor	mountain	blue		head		cone		cloth	paper
city	ocean	grey		wings		oval		feather	plastic
landscape	river	orange		door		pyramid		fur	rubber
beach	sky	white		hands		triangle		glass	stone
building	snow	yellow		hooves				hairless	water
bedroom	soil			screen					
desert	street			wheel					
dining room	tree								
forest	wall								
grass									

Figure 2: Semantic attributes used, grouped by type, including representative images for some attributes.

The attribute classifiers were trained from a data set of images obtained from Google image search. For each attribute we downloaded a set of image results from Google and manually rejected false positives, to leave about 400 true positive examples as representative images for the attribute. For most attributes negative examples were sampled from all the other attributes' representative images, while in the pairs 'indoor'/'outdoor' and 'city'/'landscape' two attributes provided positive and negative examples for a single classifier.

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. However, in real-world images, the object of interest often occupies only a small area of an image. In this case, it may be better to predict semantic attributes from image patches rather than from the whole image. In addition to global features, we will use local features obtained by calculating the average response of an attribute classifier for 100 patches sampled from an image for randomly selected positions and scales (with scale from 100 to 200 pixels). The local features are intended to help recognise attributes which only occur in localised areas of an image, so we do not use local attribute features for the attributes 'indoor'/'outdoor', 'city'/'landscape', 'bedroom', 'dining room' and 'living room' because these attributes describe whole images. 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 below each attribute classifier is a linear SVM, with each real-valued SVM score providing one dimension in the overall semantic descriptor. We use non-negative SVM scores obtained by fitting a sigmoid function to the original SVM decision value to estimate the attribute probability. The SVM parameter C is set to 10, as determined by five-fold cross-validation. Four image feature types are used as input to the SVMs: SIFT, texton filterbank, LAB and Canny edge detection. Specifically, SIFT features are computed for 2000 image patches with randomly selected positions and scales (with scales from 16 to 64 pixels), and are quantised to 1024 k -means centres. Texton and LAB features are computed for each pixel, and quantised to 256 and 128 k -means centres respectively, while Canny

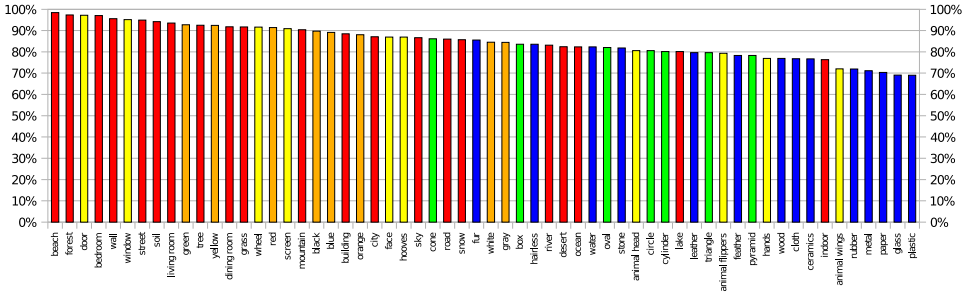


Figure 3: Accuracy achieved by SVM classifiers for individual semantic attributes. The colours show the groups of attributes: **red** scene, **orange** colour, **yellow** part, **green** shape, **blue** material.



screen wall window city box
wall screen city door indoor
 (a)



door wall window city building
wall door building city wood
 (b)



street wheel wall city road
wall city door building wood
 (c)



bedroom indoor screen window city
wall screen city door white
 (d)



bedroom wall city indoor window
wall city screen metal window
 (e)



green water grass tree stone
wall grass soil tree building
 (f)

Figure 4: 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.

edge features are 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]. 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.

4 Experiments

In this section we first evaluate the performance of the semantic attribute classifiers and show how they can be used to predict attributes for new images (Section 4.1), then look at using the semantic attribute features and bag-of-words features for object classification (Section 4.2). Finally we compare the semantic attribute features with other methods for image feature dimensionality reduction (Section 4.3).

4.1 Semantic attribute prediction

Figure 3 shows the accuracy achieved by our individual semantic attribute classifiers. Five-fold cross-validation was used to compute the accuracy of the SVM classifiers on the database of representative images. The negative examples for each attribute were sampled to balance the positive examples, so making the same prediction for every image would give 50% accuracy. Most of the classifiers achieve more than 80% accuracy; the lowest accuracies are seen on the ‘material’ attributes, while on average the ‘scene’ attribute classifiers perform best.

We will use these attribute classifiers to make soft predictions of which attributes apply to an image, and use those predictions as features for object classification. Figures 1 and 4 show attribute predictions for some test images. The five strongest global (first row) and local (second row, oblique text) attribute predictions are listed for each image, with the predictions which matched manual annotations shown in bold. In many cases where the prediction does not match the manual annotation, it is possible to understand why the attribute classifier gave too high a score. For example, the train windows in Figure 4(a) look like screens, and no sky is visible to show that it is an outdoor scene, and the window shutters in Figure 4(b) look similar to doors.

4.2 Combining semantic attributes and bag-of-words features

We tested our semantic attributes as features for object classification on the large data set used in the 2007 PASCAL VOC challenge [1]. (This is the last PASCAL VOC challenge for which the test data annotations are publicly available.) The data set contains 9963 images, with 24640 objects annotated for 20 object classes. The images were collected from users’ uploads to the Flickr website, so the objects appear in cluttered scenes, with a high degree of variation in viewing, illumination and object appearance. For the challenge’s classification task, we must determine whether or not each image contains each object class of interest. Performance is measured by calculating the average precision for each class, *i.e.* the area under the precision/recall curve.

Table 1 shows the average precision achieved on each class by a chi-squared kernel SVM using the semantic attribute features, a chi-squared kernel SVM using the original bag-of-words features, and three methods of combining them, along with the best results for each class submitted to the 2007 PASCAL VOC challenge and more recent results obtained using locality-constrained linear coding [14]. In the original challenge, the best results for every class except ‘sofa’ were given by using a genetic algorithm to optimise the parameters of a generalised radial basis function kernel for an SVM, learning weights over a large number of low-level feature channels [9]. We tested three methods of combining the scores of SVM classifiers trained separately using bag-of-words image features and semantic attribute features derived from them: taking a weighted sum, their product, or the maximum. Taking a weighted sum performed best, with mean average precision 61.4%. The weight used was

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	MAP
VOC [1]	77.5	63.6	56.1	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.9	79.2	53.2	59.4
LLC [14]	74.8	65.2	50.7	70.9	28.7	68.8	78.5	61.7	54.3	48.6	51.8	44.1	76.6	66.9	83.5	30.8	44.6	53.4	78.2	53.5	59.3
semantic	70.5	53.7	55.6	61.6	31.4	50.2	71.6	53.6	57.3	32.7	51.1	43.6	65.6	59.3	85.5	31.4	30.3	52.9	74.1	54.4	54.3
BoW	71.8	63.1	53.6	67.2	29.5	60.2	77.5	59.2	56.9	41.6	55.2	41.6	75.2	63.8	85.2	30.0	44.3	52.4	75.3	54.5	57.9
sum	76.2	66.4	59.2	70.3	35.4	63.6	79.4	62.4	59.5	47.9	58.8	44.9	78.3	67.4	87.9	32.9	46.9	53.8	78.6	58.9	61.4
product	76.3	65.7	58.8	71.8	31.4	64.3	79.5	61.2	58.8	44.1	59.7	45.7	76.4	65.4	87.9	38.1	46.8	54.5	77.4	58.7	61.1
max	74.7	65.3	58.2	69.5	32.9	63.2	80.0	60.4	55.4	47.9	56.2	46.5	77.6	65.0	86.7	34.5	40.8	52.0	77.2	56.2	60.0

Table 1: Average precision on the PASCAL VOC challenge 2007 data [1] of the best results from the challenge, and of Locality-constrained Linear Coding [14], compared with our results using bag-of-words features or semantic features alone, and using the sum/product/maximum of the bag-of-words and semantic feature scores. Classes: (1) aeroplane, (2) bicycle, (3) bird, (4) boat, (5) bottle, (6) bus, (7) car, (8) cat, (9) chair, (10) cow, (11) diningtable, (12) dog, (13) horse, (14) motorbike, (15) person, (16) pottedplant, (17) sheep, (18) sofa, (19) train, (20) tvmonitor.

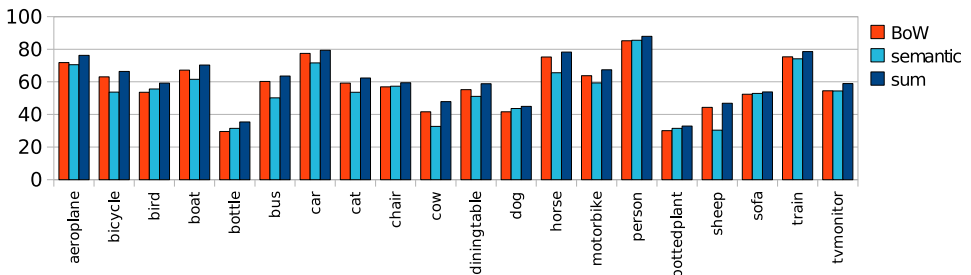


Figure 5: Average precision achieved using bag of words features, semantic attribute features, and both together, on the PASCAL VOC 2007 object class data.

learnt by cross-validation, but taking an unweighted sum of the two scores led to a mean average precision of 61.2%, only slightly lower.

Figure 5 compares the performance achieved by classifiers 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 performance on every class is increased by combining the two feature types rather than using either one alone.

BoW +	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	MAP
scene	73.5	63.4	53.7	67.5	33.5	60.7	77.3	60.7	56.4	47.5	56.5	40.5	76.8	66.3	86.6	38.4	45.3	52.8	75.7	58.2	59.6
colour	74.6	63.8	53.5	67.7	32.8	58.5	77.0	59.9	56.9	48.5	54.3	42.5	75.8	64.5	86.0	33.4	42.3	50.8	75.4	56.0	58.7
part	74.7	64.5	52.3	68.0	32.3	61.5	78.4	60.0	58.4	48.4	55.6	43.3	76.0	63.8	87.3	33.3	44.0	49.9	75.1	58.0	59.2
shape	73.0	64.5	54.5	66.6	32.7	58.0	77.3	60.6	56.2	48.7	53.8	42.1	76.8	64.4	86.1	37.5	42.3	51.6	74.2	56.6	58.9
material	74.5	64.8	55.6	67.9	33.4	59.9	77.6	60.5	58.9	49.2	54.9	40.9	75.4	64.5	86.6	33.4	40.8	51.7	75.2	56.9	59.1

Table 2: Average precision achieved using the bag-of-words features combined with subsets of the semantic attribute features. See the caption to Figure 1 for the list of class numbers.

	$d = 120$	$d = 200$	$d = 300$	$d = 400$	$d = 500$
BoW	35.2	38.4	40.5	42.4	43.4
PCA	39.0	41.2	43.7	45.4	45.6
random	32.3	36.0	37.8	39.5	40.1

Table 3: Mean average precision achieved on the PASCAL VOC challenge 2007 data [1] by low-dimensional bag-of-words features, principal component analysis dimensionality reduction of high-dimensional features, and random low-dimensional projections, according to feature dimensionality d .

Table 2 shows the average precision achieved on each class when subsets of the semantic attribute features are used in combination with the bag-of-words features. A weighted sum was used to combine the outputs of classifiers trained for the groups of semantic attribute features and for the bag-of-words features. Every combination performs better than using the bag-of-word features alone, and worse than including all the semantic attribute features, showing that all the groups of attributes provide useful information for object classification. The least benefit is given by the colour attributes, while the most benefit comes from the scene attributes.

4.3 Semantic attributes and dimensionality reduction

Since our semantic attribute features are obtained from bag-of-words features by linear SVMs, they can be viewed as a dimensionality reduction method. Table 3 and Figure 6 show the performance achieved on the PASCAL VOC 2007 data when three other dimensionality reduction methods are used to produce image descriptors of varying dimensionality from the original bag-of-words features. The first method simply varies the number of visual words produced when the low-level features are quantised by k -means clustering, 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 (we report the average performance from 10 runs with different sets of random directions). In each case a linear SVM is learnt in the low-dimensional space. The mean average precisions achieved using these features are much lower than the 54.3% from the SVM using semantic attribute features. The chi-squared kernel which was used with the histograms of semantic attribute features, cannot be used with these alternative feature sets, which are not histograms and may include negative values; using a linear kernel with the 115-dimensional semantic attribute features gives a lower mean average precision of 46.8%, but this is still better than the results with these alternative reduced-dimensionality features.

5 Discussion

In this paper we showed that semantic attribute features can be used to improve object classification performance. We trained classifiers from representative images for 60 semantic attributes, and used the classifier output to create a new low-dimensional image representation. Tested on a standard data set, the semantic attribute features by themselves achieved an object classification performance close to that of high-dimensional bag-of-words features, and combining the two feature types improved the performance on every class.

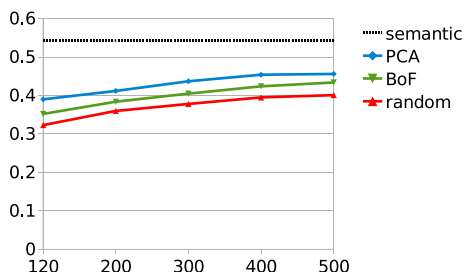


Figure 6: Mean average precision achieved on the PASCAL VOC challenge 2007 data [1] by low-dimensional bag-of-words features, principal component analysis dimensionality reduction of high-dimensional features, and random low-dimensional projections, according to feature dimensionality. The semantic attribute features have 115 dimensions.

We obtained the semantic attribute predictions from the high-dimensional bag-of-words features using linear SVMs. This linear projection may not be optimal, and we expect to be able to enhance the accuracy of the individual semantic attribute predictions by using alternative classifier types.

Experiments with ordinary bag-of-words features have shown increased performance by using a spatial pyramid to partition the image into increasingly fine regions, computing histograms for each region at each level of the pyramid, and concatenating all the pyramids to provide the final image representation [7]. It is likely that spatial pyramid matching could also enhance the performance of our semantic attribute features, at the cost of an increased feature dimensionality. Our semantic attribute predictions could also be used for object localisation. Even in cases where none of the learnt attributes apply to an object of interest, they could help recognise its context, and also help separate the object from background clutter.

We collected images for each attribute from the results of a web search, then manually rejected irrelevant images. It would also be possible to train classifiers directly from the noisy data, without any manual annotation, at the cost of degrading the classifier accuracy. This approach would become more compelling if larger numbers of attributes are used in future work.

This research was funded by the ANR-R2I and OSEO-QUAERO projects.

References

- [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] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1), 2007.

-
- [4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- [5] V. Ferrari and A. Zisserman. Learning visual attributes. *Advances in Neural Information Processing Systems*, 2008.
- [6] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [8] J. Lim, P. Arbeláez, C. Gu, and J. Malik. Context by region ancestry. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009.
- [9] Marcin Marszałek, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning object representations for visual object class recognition, October 2007. Visual Recognition Challenge workshop, at ICCV.
- [10] A. Quattoni, M. Collins, and T. Darrell. Learning Visual Representations using Images with Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [11] J. Uijlings, A. Smeulders, and R. Scha. What is the spatial extent of an object? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [12] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [13] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. *Proceedings of the International Conference on Image and Video Retrieval*, 2004.
- [14] 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.
- [15] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 2000.
- [16] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2), 2007.