Objects as Attributes for Scene Classification

Li-Jia Li*, Hao Su*, Yongwhan Lim, Li Fei-Fei

Computer Science Department, Stanford University

Abstract. Robust low-level image features have proven to be effective representations for a variety of high-level visual recognition tasks, such as object recognition and scene classification. But as the visual recognition tasks become more challenging, the semantic gap between low-level feature representation and the meaning of the scenes increases. In this paper, we propose to use objects as attributes of scenes for scene classification. We represent images by collecting their responses to a large number of object detectors, or "object filters". Such representation carries high-level semantic information rather than low-level image feature information, making it more suitable for high-level visual recognition tasks. Using very simple, off-the-shelf classifiers such as SVM, we show that this object-level image representation can be used effectively for high-level visual tasks such as scene classification. Our results are superior to reported state-of-the-art performance on a number of standard datasets.

1 Introduction

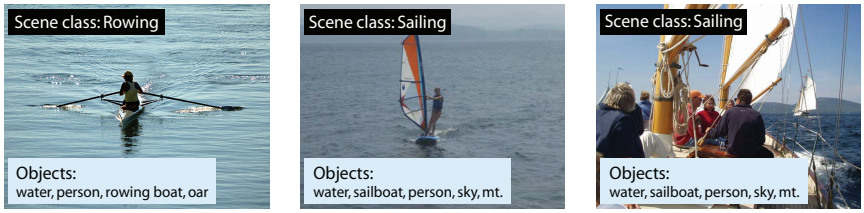


Fig. 1. Classifying complex, event scenes such as "sailing" and "rowing" involves high-level understanding of the images. The "sailing" image in the middle has similar low-level texture statistics as the "rowing" image on the left, and a significantly different texture distribution from the "sailing" image on the right. Humans, however, would classify the middle and the right images as belonging to the same event class ("sailing") based on the objects and the high-level meaning pertaining to the scenes.

Much of the recent progress in high-level vision has been made by designing robust image features for recognition tasks such as object and scene recognition. Nearly all of these features are based on some kind of low-level image

*indicates equal contributions.

properties (e.g. SIFT [29], filterbanks [14, 34], GIST [33], etc.). With the help of more and more sophisticated statistical models, these features have achieved good successes in high-level recognition tasks. Particularly exciting is the recent development of robust single object detectors ([5, 12]) based on these features. Using these as off-the-shelf detectors, researchers have developed algorithms to further incorporate context information such as scene layout [19], background class [32], and object co-occurrences [35] to achieve better detections of objects in scenes. But as the visual task becomes higher and higher level, the limitations of low-level features become more obvious. Take Fig. 1 as an example. A classification algorithm based mostly on texture statistics would easily confuse the left and middle scenes as the same class. Even introducing some contextual information such as background scene environment or overall layout would do little to differentiate the left "rowing" scene from the middle "sailing" scene. Our visual experiences and intuition suggest that a straightforward way of distinguishing many complex real-world scenes would be object-based – the presence of a sailboat is more indicative of a "sailing" scene rather than a "rowing" scene.

"Semantic gap" is a widely accepted notion to capture the discrepancy between image representations and image recognition goals. In general, the lower-level the feature is (think of raw pixel values), the more work a model has to do towards higher-level recognition goals, analogous to the concept of "potential energy" in physics. One way to close the semantic gap is by deploying increasingly sophisticated models, such as the probabilistic grammar model [42], compositional random fields [21], and graphical models [11, 38]. While these approaches are based on rigorous statistical formulation, good learning and inference are still extremely difficult. Most of the papers have shown promising results on only small scale datasets.

Attribute-based methods have shown promising potential in object recognition in recent few years. Its success in recognition is largely accredited to the introduction of "attribute", a high-level semantically meaningful representation. In attribute-based methods for object recognition, an object is represented by using visual attributes. For example, a polar bear can be described as white, fluffy object with paws. Such visual attributes summarize the low-level features into object parts and other properties, and then are used as the building blocks for recognizing the object.

Similarly, we hypothesize that an image representation based on objects would be very useful in high-level visual recognition tasks for scenes cluttered with objects. It provides complementary information to that of the low-level features.

In this paper, we introduce the concept of using object as attributes for scene representation. We describe complex real-world scenes by collecting the responses of many object detectors. Drawing an analogy to low-level image representation, instead of using image filters (and their alike) to represent local texture, we introduce *object filters* to characterize local image properties related to the presence/absence of objects. By using a large number of such object filters, our *object bank* representation of the image can provide rich information of

the scene that captures much of the high-level meaning (Sec. 3). As a proof of concept, we test this new image representation on a series of scene recognition tasks by using simple, off-the-shelf SVM classifiers. Our results show that our object-level image representation delivers superior scene recognition results to all reported state-of-the-art on a number of standard scene datasets (Sec.4).

2 Related work

A plethora of image feature detectors and descriptors have been developed for object recognition and image classification [33, 1, 22, 30, 29]. We particularly draw the analogy between our *object* filters and filter banks and the *texture* filters and filter banks [34, 14].

Object detection and recognition also entail a large body of literature [10, 3]. In this work, we mainly use the current state-of-the-art object detectors of Felzenszwalb et. al. [12], as well as the geometric context classifiers ("stuff" detectors) of Hoeim et. al. [18].

The idea of using many object detectors as the basic representation of images is analogous to work in the multi-media community on applying a large number of "semantic concepts" to video and image annotation [16] and semantic indexing [37]. In contrast to our work, in [16] and [37] each semantic concept is trained by using entire images or frames of videos. There is no sense of localized representation of meaningful object concepts in scenes. As a result, this approach is difficult to use for understanding cluttered images composed of many objects.

Finally, our approach is inspired by earlier approaches of attribute-based object recognition [24, 23, 13, 8]. Similar to the "semantic concepts" approaches, attributes classifiers in these works are also trained by using the entire images instead of those regions containing the visual attributes. Furthermore, these approaches focus on single object classification based on human-defined attributes. Our approach, however, investigates the contribution of objects to scene classification.

We evaluate the utility of our object bank representation on a number of scene classification tasks. It is outside of the scope of this paper to discuss in detail the large pool of related literature. To summarize, a number of papers have focused on using low-level features for image classification, such as GIST [33], filterbanks [14, 34, 27], and Bag of Words of local features [2, 11]. Hierarchical modeling of images or images and texts is popular for more complex scene classification [21, 38, 17].

3 The Object Bank Representation of Images

3.1 What is an Object Filter?

Given an image, an *object filter* response can be viewed as the response of a "generalized object convolution." In the simplest case, we take an object template (e.g., a picture of a face), and scan it across the image, resulting in a map

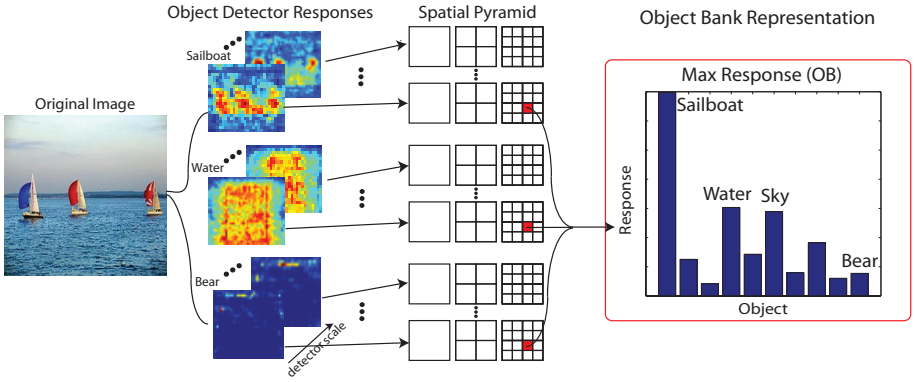


Fig. 2. (Best viewed in colors and magnification.) Illustration of the object filter representation. Given an input image, we first run a large number of object detectors at multiple scales. For each object at each scale, use a three-level spatial pyramid representation of the resulting object filter map, resulting in $\texttt{No.Objects} \times \texttt{No.Scales} \times (1^2+2^2+4^2)$ grids. An object filter descriptor of an image is a concatenation of features described in each of these grids. We compute the maximum response value of each object, resulting in a feature vector of `No.Objects` length for each grid.

of face filter responses. Thanks to the recent development of more robust object detectors, we are able to use more sophisticated methods than simple image templates as object filters.

We point out here that we use the word "object" in its very general form – while cars and dogs are objects, so are sky and water. Our image representation is agnostic to any specific type of object detector; we take the "outsourcing" approach and assume the availability of these detectors. In this paper, we use the latent SVM object detectors [12] for most of the blobby objects such as tables, cars, humans, etc, and a texture classifier by Hoiem [18] for more texture- and material-based objects such as sky, road, sand, etc.

Fig. 2 illustrates the general setup for obtaining the object bank image representation. A large number of object detectors are run across an image at different scales. For each scale and each detector, we obtain an initial response map of the image by using the state-of-the-art object detectors [12, 18]. In this paper, we use 200 objects detectors at 12 detection scales and 3 spatial pyramid levels (L=0,1,2) [26].

We now compare the object bank image representation to two popular low-level image representations: GIST [33] and the Spatial Pyramid (SPM) representation of SIFT [26], illustrated by Fig. 3. It is interesting to observe that images with very similar low-level statistics might carry very different meanings (e.g., mountain and city street). The GIST and SIFT-SPM representations often show similar distributions for such images. In contrast, the object bank representation can easily distinguish such scenes due to the semantic information provided by the object filter responses.

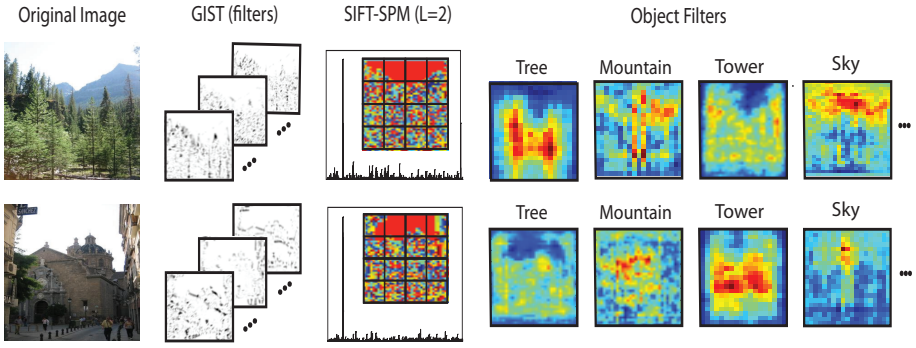


Fig. 3. (Best viewed in colors and magnification.) Comparison of the object bank representation with two low-level feature representations, GIST and SIFT-SPM of two types of images, mountain vs. city street. For each input image, we first show the selected filter responses in the GIST representation [33]. Then we show a histogram of the SPM representation of SIFT patches [26] at level 2 of the SPM representation where the codeword map is also shown as a histogram. Finally, we show a selected number of object filter responses.

Object bank representation is built upon image responses to object detectors that have been pre-trained with a large number of images. These detectors might not yet be perfect, but are designed to capture much of the variations of their respective objects. The result is a stable representation of scenes that has already encoded much of the prior information of the visual space. We show in an analysis experiment of a scene classification task that even when there is only a very small number of scene training examples, our object bank representation can achieve reasonable recognition results on two challenging scene datasets, significantly outperforming low-level features (Fig.4 Left).

In a similar vein, we vary the number of object filters in the image representation to investigate its effect on a scene classification task (Fig.4 Right). Our experiment shows that performance plateaus after applying a relatively small number of object filters (a couple of dozen). This result suggests that our object bank representation captures rich image information even with a modest number of object detectors. It is also good news for a practical system that would use the object bank representation: one does not need an extremely large number of object filters to start achieving reasonable results.

We emphasize that object filter bank features are not meant to replace low-level image features. Instead, we observe (and will show in Fig. 7(c)) that they offer important complementary information of the images. While this paper itself is focused on the utility of the object filter bank features, we envision that future algorithms and systems would combine both low-level features as well as these high-level features for visual recognition tasks.

Before moving onto the next section, we make the observation that there is an increasing trend in computer vision and multi-media communities to use

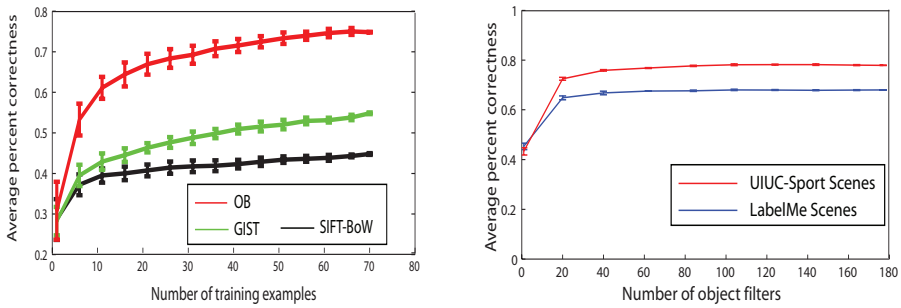


Fig. 4. (Best viewed in colors and magnification.) Left: Comparison of scene classification performance vs. the number of training examples among different features: GIST, SIFT-BoW and OB. Right: Scene classification performance vs. the number of object filters. 10 round of randomized sampling is performed to choose the object filters among 177 object detectors (We excluded object detectors obtained from LabelMe training data to avoid possible contamination with the testing images in LabelMe Scene. More details of the datasets are described in Sec. 4.). Standard deviation is plotted as the error bar at each sample point.

stand-alone detectors as the training data becomes more freely available and also more heterogeneous (Flickr, Google, Facebook, etc.). Just like most of today's object recognition systems are built upon interest point detectors and descriptors instead of raw pixels, we hypothesize that solving more complex and large scale image understanding problems will depend more and more on basic object detectors in addition to local image features. Object bank representation largely decouples the development of individual object detectors and higher level visual recognition tasks, making it possible for us to make progress now on higher level visual recognition tasks.

3.2 What are the Objects Filters for the Object Filter Bank?

So what are the "objects" to use in these object filters? And how many? An intuitive answer to this question is to use all possible objects. As the detectors become more robust, and especially with the emergence of large-scale datasets such as LabelMe [36] and ImageNet [6], this goal has become more reachable.

But time is not fully ripe yet to consider using all objects in, say, the LabelMe dataset. Not enough research has yet gone into building robust object detector for tens of thousands of generic objects. As we increase the number of objects, the issue of semantic hierarchy becomes more prominent. Not much is understood about what it means to detect a mammal and a dog simultaneously. And even more importantly, not all objects are of equal importance and prominence in natural images. As Fig. 5 shows, the distribution of objects follows Zipf's Law, which implies that a small proportion of object classes account for the majority of object instances. Hauptmann and colleagues have postulated that using 3000-4000 concepts should suffice to satisfactory annotate most of the video data [16].

For this paper, we will choose a few hundred most useful (or popular) objects in images[1]. An important practical consideration for our study is to ensure the availability of enough training images for each object detectors. We therefore focus our attention on obtaining the objects from popular image datasets such as ESP [41], LabelMe [36], ImageNet [6]. We also consider image search engines such as Google image search, Ask.com image search and Bing image search and the Flickr! online photo sharing community. After ranking the objects according to their frequencies in each of these image data sources, we take the intersection set of the most frequent 1000 objects, resulting in 200 objects, where the identities and semantic relations of some of them are illustrated in Fig. 6. To train each of the 200 object detectors, we use 100∼200 images and their object bounding box information from the LabelMe [36] (86 objects) and ImageNet [6] datasets (177 objects). We use a subset of LabelMe scene dataset to evaluate the object detector performance. Final object detectors are selected based on their performance on the validation set from LabelMe.

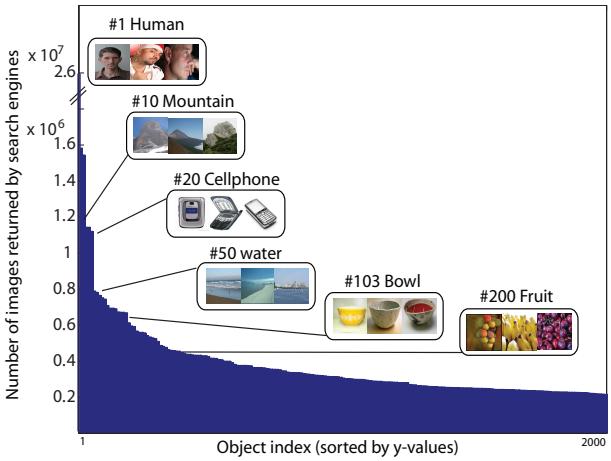


Fig. 5. (Best viewed in colors and magnification.)The frequency (or popularity) of objects in the world follows Zipf's law trend: a small proportion of objects occurs much more frequently than the majority. While there are many ways of measuring this, e.g., by ranking object names in popular corpora such as the American National Corpora [20] and British National Corpus [7], we have taken a web-based approach by counting the number of downloadable images corresponding to object classes in WordNet on popular search engines such as Google, Ask.com and Bing. We show here the distribution of the top 2000 objects.

[1] This criterion prevents us from using the Caltech101/256 datasets to train our object detectors [9, 15] where the objects are chosen without any particular considerations of their relevance to daily life pictures.

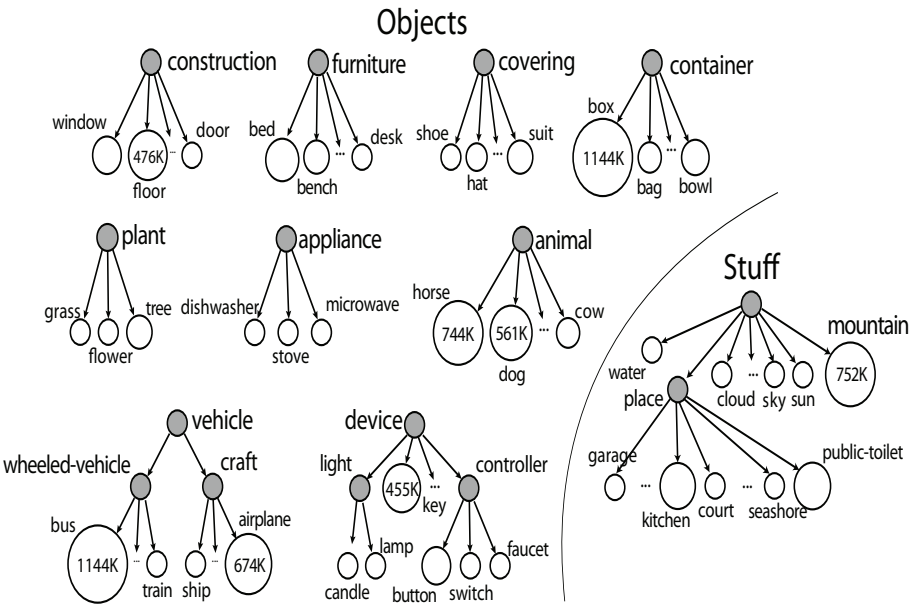


Fig. 6. Rough grouping of the chosen object filters based loosely on the WordNet hierarchy [31]. The size of each unshaded node corresponds to the number of images returned by the search.

To our knowledge, no previous work has applied more than a handful of object detectors in scene recognition tasks [4]. But our initial object filter bank of 200 object detectors is still of modest size. We show in Sec. 4 that even with this relatively small number of objects we achieve excellent recognition results (e.g., Fig. 4 Right). A future study for our work is to increase the number of object detectors to possibly thousands.

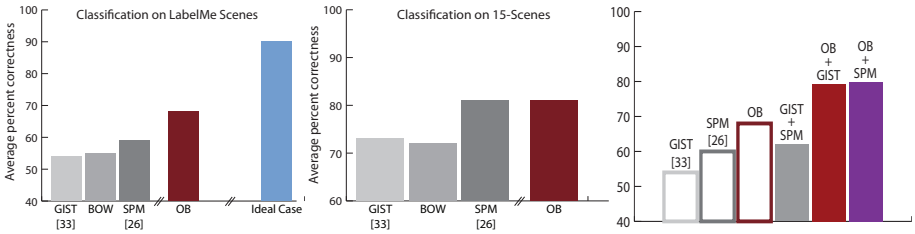
4 Using Object Filter Bank for Scene Recognition

The object bank representation can be useful for many high-level visual recognition tasks, especially where the images contain many objects. In this paper, we focus on the general problem of scene classification to illustrate the usefulness of the object filter bank. Scene classification involves many degrees of abstraction [39]. Here we take a loose definition from the psychology literature, and consider the simple *basic-level* scene recognition task to be classifying generic places such as kitchen vs. mountain vs. highway. We further investigate the performance of our object bank representation on higher level tasks like activity and event recognition as *super-ordinate* scene recognition.

4.1 Basic-level Scene Classification: 15-Scene and LabelMe Scene

We first use two basic-level scene classification datasets to evaluate the utility of the object bank representation: the 15-Scene classes [26] and a LabelMe 9-class scene dataset[2].

For the 15-Scene classes, we follow the experimental setting in [26] by using 100 images in each class for training, and the rest for testing. For the LabelMe Scene dataset, 50 randomly drawn images from each scene classes are used for training and 50 for testing. In Fig. 7, we show the average performance of multi-way classification results. We compare our object bank features to SIFT-BoW, GIST, and SPM. For all image representations, we use a simple liblinear SVM classifier.



(a) LabelMe (b) 15-Scene (c) Combining features

Fig. 7. Comparison of classification performance of different features on LabelMe scene (a) and 15 scene (b) datasets. In the LabelMe dataset, we also show an "ideal" experiment where we use the human ground-truth object identities to predict the labels of the scene classes. (c) Classification performances of different combinations of image representations on the LabelMe scene dataset.

Our object bank representation significantly outperforms the other features in the LabelMe Scene dataset, and is on par with the current state-of-the-art results in the 15-Scene dataset [26]. It is worth noticing that in our experiment, we have used a much simpler classifier (linear SVM) compared to the SPM results in [26]. Our results show that when using object-level features, our image representation carries rich enough semantic information for understanding various types of real-world scenes.

In Fig. 7(c), we further investigate the complementarity of low- and high-level image representations for scene classification tasks. We apply a multiple kernel learning algorithm [40] to different combinations of image representations. We show that the combination of object filter and GIST or SIFT features significantly boosts the performance over each individual type of feature. On the

[2] From 100 popular scene names, we obtained 9 classes from the LabelMe dataset in which there are more than 100 images: beach, mountain, bathroom, church, garage, office, sail, street, and forest. The maximum number of images in those classes is 1000.

other hand, the combination of SIFT and GIST features does not improve the classification performance, suggesting that these two low-level representations are largely similar to each other, offering little complementary information for further discrimination of the scenes.

4.2 Super-ordinate Level, Complex Scenes: UIUC-Sports

We further consider a higher-level scene recognition task involving activities and events. Here we use UIUC-Sports dataset. The images in this datasets are highly cluttered by objects. Further, referring back to Fig. 1, images of activities such as sailing and rowing have very similar background and thus almost indistinguishable image statistics, but they do differ in the types of object present.

For the UIUC-Sports dataset, we follow the experiment setting of [28], and train on 70 images and test on 60 images.

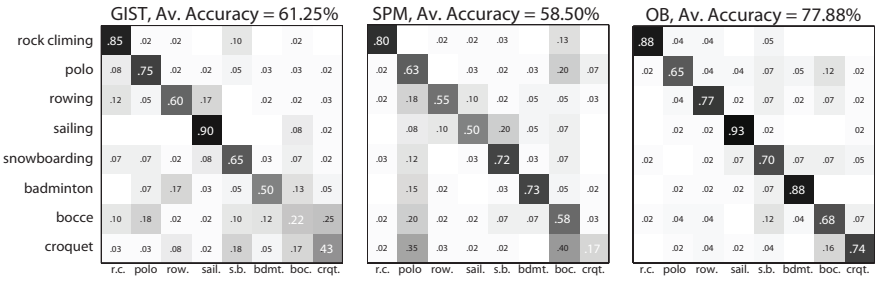


Fig. 8. Confusion matrices of three different image feature representations on the UIUC-Sports scene dataset: GIST [33], SIFT-SPM [26] and OB. The average accuracy for each method is the mean of the diagonal entries of each confusion table.

We show the performance of different feature representations on the UIUC-Sports dataset in Fig. 8. Our object bank representation shows significantly superior results to the low-level image features, where the OB representation shows $15\% \sim 20\%$ improvement over the GIST and SPM features. Our result (77.88%) is also substantially higher than the reported state-of-the-art performance of 73.40%. It is worth noting that the state-of-the-art performance is obtained by using a fully supervised algorithm where all object outlines and identities are given during the training stage [28], whereas we only have the class label of each training image. Upon closer examination of the confusion patterns among different scene classes, we can see that a more semantic-level image representation overcomes some of the usual confusion caused by low-level features, such as the confusion between sailing images and rowing images (see the confusion tables for GIST and SPM). When incorporating object-level information, sailing class and rowing class are no longer confusing.

5 Discussions

As we try to tackle higher level visual recognition problems, we show that more semantic level image representation such as the object filter bank can capture important information in a picture without evoking highly elaborate statistical models to build up the features and concepts from pixels or low-level features. We emphasize that low-level image texture-based features are still extremely useful in recognition task. The object filter bank features offer a complementary set of information. When either used alone or in combination with low-level features, these features yield very promising results in scene classification. But a number of issues still need to be addressed and improved.

One important consideration is computation. Training hundreds and thousands of object filters could be expensive. In this work, we have taken the "outsource" philosophy, where we assume that reasonable object detection algorithms are available for us to use. Obtaining an increasingly large number of trained detections is becoming more and more achievable with the emergence of large-scale datasets which provide available data and computing services such as cloud computing and large scale computing grid. But there is still an "object filtering" step in image representation. In this paper, we have used the naive scanning window approach. For each image, the total time for extracting the object filter feature for all 200 object filters is $\sim$1 minute on a modern CPU. Efficient algorithms such as robust branch and bound scheme proposed by Lampert et. al. [25] can further speed up the computation time.

As we mentioned in the introduction, a wave of recent work has shown the importance of context in visual recognition tasks. Roughly speaking, context can be grouped into semantic (or probability) context related to accounting for co-occurrences of objects and stuff, as well as geometric context related to the layout of scenes and constraints of camera(s). Our object bank representation implicitly encodes the co-occurrences context by concatenating the response maps of different detectors. In this paper, we have only used very weak spatial information in the image representation through the spatial pyramid representation. In the future, we will investigate how to incorporate more explicit or robust geometrical context of scenes into the representation, such as depth and layout information.

References

1. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 509–522, 2002.
2. A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. *Proc. ECCV*, 4:517–530, 2006.
3. L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. *ICCV*, 2009.
4. D. Ramanan C. Desai and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

5. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 886, 2005.
6. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
7. B. Edition and BNC Sampler. British National Corpus.
8. A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009.
9. L. Fei-Fei, R. Fergus, and P. Perona. One-Shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.
10. L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. Short Course CVPR: http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html, 2007.
11. L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *Computer Vision and Pattern Recognition*, 2005.
12. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *Journal of Artificial Intelligence Research*, 29, 2007.
13. V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS*, 2007.
14. W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.
15. G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. 2007.
16. A. Hauptmann, R. Yan, W. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958, 2007.
17. G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. *Proceedings of Neural Information Processing Systems. Vancouver, Canada: NIPS*, 8, 2008.
18. D. Hoiem, A.A. Efros, and M. Hebert. Automatic photo pop-up. *Proceedings of ACM SIGGRAPH 2005*, 24(3):577–584, 2005.
19. D. Hoiem, A.A. Efros, and M. Hebert. Putting Objects in Perspective. *CVPR*, 2006.
20. N. Ide and C. Macleod. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, pages 274–280. Citeseer, 2001.
21. Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *CVPR*, 2006.
22. T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.
23. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. *ICCV*, 2009.
24. C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, 2009.
25. C.H. Lampert, M.B. Blaschko, T. Hofmann, and S. Zurich. Beyond sliding windows: Object localization by efficient subwindow search. In *Proc. of CVPR*, volume 1, page 3, 2008.
26. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006.
27. T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.

28. L-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proc. ICCV*, 2007.
29. D. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, 1999.
30. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, volume 1, pages 128–142, 2002.
31. G.A. Miller. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM*, 1995.
32. K. Murphy, A. Torralba, and W.T. Freeman. Using the forest to see the trees:a graphical model relating features, objects and scenes. In *NIPS (Neural Info. Processing Systems)*, 2004.
33. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. Journal of Computer Vision.*, 42, 2001.
34. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern Analysis and machine intelligence*, 12(7):629–639, 1990.
35. A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE International Conference on Computer Vision*, 2007.
36. B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. 2005.
37. J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo*, pages 445–448, Washington, DC, USA, 2003. IEEE Computer Society.
38. E. Sudderth, A. Torralba, W.T. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. International Conference on Computer Vision*, 2005.
39. B. Tversky and K. Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15(1):121–149, 1983.
40. A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple Kernels for Object Detection. 2009.
41. L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.
42. L. Zhu, Y. Chen, and A. Yuille. Unsupervised learning of a probabilistic grammar for object detection and parsing. *Advances in neural information processing systems*, 19:1617, 2007.