'OBJECT BANK'-BASED SCENE CLASSIFICATION

A. F. de Araujo [1], *P. Weinzaepfel* [2], *P. Perez* [3], *C. Diot* [3]

afaraujo@stanford.edu , patrick.perez@technicolor.com
[1]Stanford University
[2]Ecole Normale Superieure Cachan-Bretagne
[3]Technicolor

ABSTRACT

We discuss the effectiveness of the Object Bank scene classification method. Experiments show that the object detection interpretation of the method is not the key to the performance of this approach: a) at least half of the most important object filters for the classification task are not present in the datasets, b) random filters with similar dimension achieve nearly equal performance (i.e., state-of-the-art performance), c) when using comparable number of dimensions with respect to the Bag-of-Visual-Words method, performance also drops to a comparable level. We understand that the high performance of the technique is mainly based on its use of high-dimensional vectors of local scale-space features.

1. INTRODUCTION

Scene classification is a classic research problem in computer vision, and many recent efforts are found in the literature. However, the achievable performance of state-of-the-art systems is still far from satisfactory. Probably the most interesting example can be drawn from the video semantic indexing community (which basically employs scene classification techniques in still frames of each video shot), with the TRECVID annual evaluation, in which the top performing team achieved no more than 0.09 in terms of Mean Average Precision in 2010 [1]. Other common benchmarks, such as the 'MIT Indoor' [2] dataset, have state-of-the-art schemes achieving less than 40% classification accuracy [3].

As a consequence, many new ideas have been introduced in the past decade. A recent technique that seemed promising, namely the 'Object Bank' (OB) approach, was introduced both in [3] and [4]. Its key aspect is the (supposedly) use of the information of the presence/absence of a list of objects at predefined scales and spatial positions to identify the class of the scene.

In this work, we perform experiments in order to understand and analyze the 'Object Bank' technique for scene classification. The contributions are:

- we show that there is very little semantic coherence in terms of object detection. Our experiments show that, for two common datasets, half of the 10 most relevant objects (as measured by feature selection methods) for the classification task aren't even present in the datasets;

- we infer that the object filters act simply by means of projecting the data into a very high dimensional space, and not by detecting objects, as it would be expected. When using random filters of a similar dimension, nearly equal performance is obtained in common benchmark datasets. When reducing the number of dimensions, a comparable performance is obtained with respect to the Bag-of-Words method.

The rest of this work is organized as follows. In Section 2, we introduce the basic functioning of the OB technique. In Section 3, we explain the accomplished experiments and results. We conclude with discussion in Section 4.

2. OVERVIEW OF THE OBJECT BANK APPROACH

The key idea of the OB approach is the use of responses to "object filters" (based on trained classifiers in HOG feature space) to construct a feature vector (which will, then, be associated to a machine learning method to perform classification). Intuitively, the arrangement and presence/absence of a set of objects help identify the type of scene under evaluation.

With that purpose, the method calculates responses at a predefined number of scales and locations (levels of the spatial pyramid) to construct the feature vector. Each combination of object filter, scale and pyramid level gives a 'response map', which is basically a heat map indicating the strength of the response when the filter is placed at each position. For each of those, the method finds the maximum and inserts its value into the feature vector. As a result, for O object filters, S scales and L pyramid levels, a feature vector of dimension $O \times S \times L$ is constructed (in other words, each image is represented by a $(O \times S \times L)$-dimensional vector).

The object filters are carefully designed to provide the best possible detection. By default, OB uses the state-of-the-art object detection method introduced by [5]. It is interesting to notice that the design of object filters involves extremely high computational cost, with complex machine learning methods. The default object filters of OB's implementation were trained based on the Imagenet dataset [6].

The employed object detection method is a part-based model, i.e., it detects objects based on a score resulting from the positioning of the various parts of the object. Two variants of OB's implementation exist: one with part-based models, and a 'partless' one, which uses only one part (the root) to detect an object. As stated in the release notes of the implementation [7], the 'partless' version of the method "is available to speed up the feature extraction while maintaining even higher classification result" — so, we choose to employ this one.

3. EXPERIMENTS AND RESULTS

After introducing the basic architecture of OB in the previous Section, we now describe the realized experiments and the obtained results. But, first, we motivate this analysis with a simple example of results obtained from three sample images, given in Figure 1. These images, along with the extracted OB features, are given as examples in [7]. It is clear that filters with higher responses correspond to objects that are not present in the given images.

We proceed, then, to a more detailed analysis of the performance of this technique, using the UIUC Sports [8] and MIT Indoor [2] datasets. The UIUC Sports is composed of 8 classes of sports events images, and the MIT Indoor contains 67 classes of indoor images.

For both datasets, we follow the approach of [3]:

- UIUC Sports: for each of the 8 classes, we pick randomly 70 images for training and 60 for testing. 10 sets of training and testing images are generated.
- MIT Indoor: for each of the 67 classes, we pick randomly 80 images for training and 20 for testing. 10 sets of training and testing images are generated.

We use the 'partless' variant of OB in all the experiments in this work. Also, we use the default implementation [7] with 177 object filters (each with two models), 6 scales and 21 spatial locations (levels of the spatial pyramid).

3.1. Feature selection applied to OB components

It is natural to wonder which components of the high-dimensional OB feature vector are more relevant for a given classification task. Furthermore, from the sample examples given in the beginning of this Section, it becomes clear that this is a question of interest.

We employ two feature selection methods: 'Information Gain' (IG) and 'Chi Square' (χ^2). These are calculated for the classification problem (i.e., the labels) with respect to each feature, one at a time. IG (also known as the mutual information) gives the removed uncertainty from the problem by using a certain feature. χ^2 is used as a means to assess the level of independence between the labels and each feature. For both approaches, the higher the score, the more relevant the feature is for the classification problem. For more details, we refer the reader to [9].

For each set of images, we calculate the score of each feature (each among the 44604-dimensional vector). A final score for each feature and each dataset is given by the average of the 10 scores (one for each set of images). Table 1 gives the 10 top-scoring object classes for both datasets, along with their average IG and χ^2 scores.

It is clear that there is little semantic coherence between the top-scoring features and the types of images in the dataset. For example, for the UIUC Sports dataset, among the objects cited in Table 1, only 4 (sailboat, sail, seashore and beach) are really present in this dataset. In the MIT Indoor case, this happens for 5 of the listed objects (bench, curtain, keyboard, people, window). In both cases, unexpected objects rank high: lion, oxygen mask, gravel, microwave, basketball court, among others.

3.2. Comparison of object filters with random filters

Now, we compare the results of the OB approach with a very simple alternative approach: using random filters, instead of object filters. The objective is to infer how important the specification of the filters is.

Initially, we generate random 10x10x31 matrices (which is similar to the dimensions of the object filters) using a Gaussian distribution — each of the components is independently generated according to $N(0, 1)$. 177 filters are generated, the same number which is used in the standard OB implementation [7].

We use both approaches to predict the class labels for both datasets. The classification method comprises the following steps: 1) extraction of the features of the images in the set, 2) construction of a classifier (model) using the training set only, 3) prediction of the class of each test image using the constructed model. We follow the approach in [3] and employ a linear one-against-one Support Vector Machine (SVM) classifier, using libSVM [10].

The employed performance measure is the classification accuracy: simply the proportion of correct classifications of the test set, as in [3] and [4]. We perform this experiment for each of the 10 image sets, and calculate the final score for each method as the average of these 10 runs. These results are given in Table 2.

For the UIUC Sports dataset, we obtain on average 78.29% and 76.54% using OB features and random features, respectively. For the MIT Indoor dataset, the results of 37.38% and 33.64% are obtained with OB features and random features, respectively. Both are superior to the results of alternative methods presented in [3].

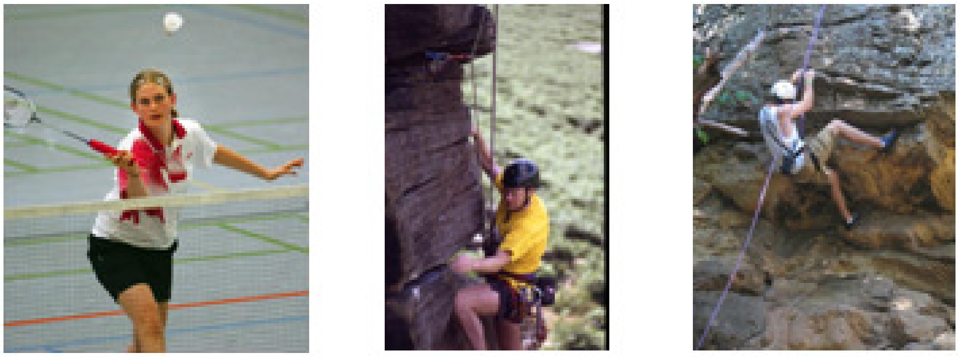
We observe that the methods present a similar level of performance, for both datasets. For the UIUC Sports dataset, the difference is of less than 2 percentage points, and for the MIT Indoor dataset it is less than 4 percentage points.

-	UIUC Sports				MIT Indoor			
Rank	Object	IG score	Object	χ^2 score	Object	IG score	Object	χ^2 score
1	sailboat	0.5042	sailboat	514.3	gravel	0.2288	gravel	1550.4
2	sail	0.4797	sail	466.2	bench	0.2211	bench	1500.5
3	lion	0.4262	oxygen mask	330.7	soil	0.2149	soil	1457.1
4	oxygen mask	0.4168	lion	328.2	basketball court	0.2022	basketball court	1381.4
5	bouquet	0.4090	microwave	319.8	curtain	0.1972	curtain	1363.7
6	groom	0.3966	bouquet	313.4	beach	0.1947	keyboard	1359.7
7	attire	0.3966	attire	313.2	bus	0.1927	beach	1333.9
8	seashore	0.3929	seashore	302.0	keyboard	0.1897	bus	1324.2
9	monkey	0.3807	airplane	299.7	cloud	0.1865	window	1314.3
10	rabbit	0.3769	beach	297.8	people	0.1865	people	1287.4

Table 1. Results of the feature selection algorithms Information Gain and χ^2 for the Object Bank features of the datasets UIUC Sports and MIT Indoor. From the top-10 ranking objects of the UIUC Sports dataset, only 4 are objects that are in fact present in it. For the MIT Indoor dataset, this is the case for only 5 of the top-ranked objects.

-	UIUC Sports		MIT Indoor	
Run	Object Bank	Random filters	Object Bank	Random filters
1	80.21%	79.17%	37.62%	34.10%
2	76.04%	72.50%	37.62%	34.10%
3	78.13%	76.04%	37.09%	31.79%
4	76.46%	75.42%	37.76%	34.03%
5	78.54%	75.00%	36.79%	33.51%
6	78.75%	77.92%	35.30%	32.24%
7	78.96%	77.71%	36.64%	32.46%
8	77.50%	78.54%	38.58%	34.85%
9	79.38%	78.13%	38.06%	35.00%
10	78.96%	75.00%	38.36%	34.33%
Mean	78.29%	76.54%	37.38%	33.64%

Table 2. Results of the classification experiments using OB and random filters.



Rank			
1	Beach	Coral	Beach
2	Helmet	Fridge	Carpet
3	Seashore	Keyboard	Keyboard
4	Vase	Soil	Coral
5	Basketball	Fence	Seashore

Fig. 1. Top ranked object filter responses for some sample images provided in [7]. None of the top-5 objects are present in the three sample images.

3.3. Reducing the dimension of the feature vector

Now, we experiment with reducing the number of dimensions employed in the feature vector: we vary the number of filters, scales and levels of the spatial pyramid. As commented previously, 177 filters (each with two models, total of 354 filters), 6 scales and 21 levels are employed in the standard OB implementation.

We employ the UIUC Sports dataset with random filters, which allows for choosing randomly the selection of filters (using OB filters would imply establishing a criterion on what objects to select). We perform classification for 10 randomly selected sets of images (exactly the same which were used in the previous Subsection) and report the mean as the final classification accuracy.

The results are presented in Figure 2. It is clear that there is no significant performance reduction when reducing the dimension of the vector from 44604 to 3000 (a performance decrease from 76.54% to 75.14%, but using only 6.7% of the dimensions). It is also interesting to notice that the configuration with 10620 dimensions performs on average even better than the standard one. Furthermore, when the feature vector dimension drops to 1000, the accuracy drops to 71.48%, which is close to the reported Bag-of-Words [12] performance on [3] (the exact value is not reported, but it is approximately 69%), which in most cases uses about 1000 dimensions.

4. DISCUSSION

In this work, we performed experiments in order to understand the good performance observed by the Object Bank method for scene classification. It is clear that the interpretation that the method allows scene classification by using information on the presence/absence of several objects is not consistent. In Subsection 3.1, we showed that at least half of the ‘object filters’ which provide the most relevant features for the classification task do not correspond to objects in the dataset under evaluation.

Moreover, in Subsection 3.2, we showed that even employing randomly selected filters (which imply no computational cost in the training stage) the performance is similar to the ones obtained by the use of costly object filters. In other words, state-of-the-art performance for both UIUC Sports and MIT Indoor datasets may be obtained using random filters instead of carefully designed object filters, if we are to work with such a high-dimensional space (dimensionality usually much greater than the number of training examples, and much greater than the number of classes - as in the experiments of [3] and [4] and reproduced in this work).

We conjecture that the main reason for the high performance of the method is its very high dimensionality: with the default implementation, a 44604-dimensional vector is the result for each image. Typical image classification methods, us-

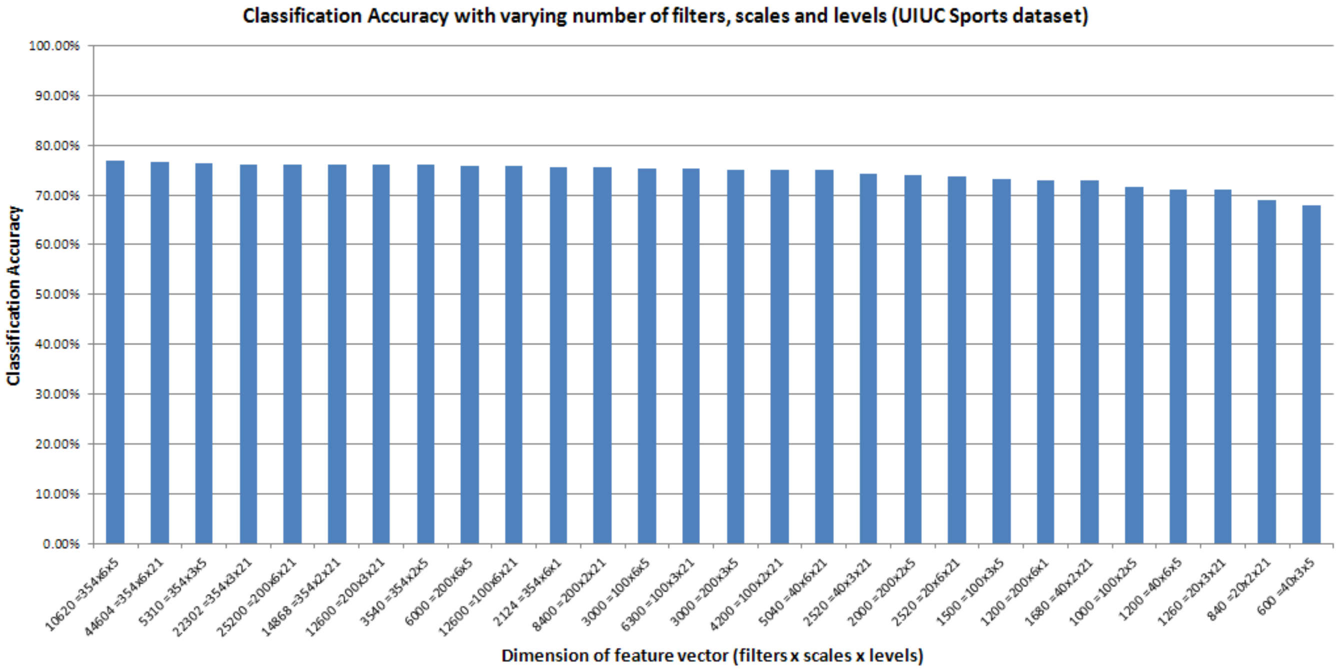


Fig. 2. Classification accuracy with varying number of filters, scales and levels, using random filters. The horizontal axis shows also the total feature dimension.

ing GIST [11], or Bag-of-Words [12], which were employed for comparisons in [3] and [4], usually present at most a couple of thousands of dimensions. Our final experiment, in Subsection 3.3, shows that when the dimensionality is roughly the same as a simple Bag-of-Words [12] scheme, the classification accuracy is very similar with respect to this method.

5. REFERENCES

[1] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, E. Gavves, D. Odijk, M. de Rijke, Th. Gevers, M. Worring, D.C. Koelma, A.W.M. Smeulders "The MediaMill TRECVID 2010 Semantic Video Search Engine," *TRECVID 2010*

[2] A. Quattoni, A. Torralba "Recognizing Indoor Scenes," *IEEE Conference in Computer Vision and Pattern Recognition*, 2009.

[3] L-J. Li, H. Su, E. P. Xing, L. Fei-Fei "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," *Neural Information Processing Systems (NIPS)*, Vancouver, 2010.

[4] L-J. Li, H. Su, Y. Lim, L. Fei-Fei "Objects as Attributes for Scene Classification," *European Conference of Computer Vision (ECCV)*, 2010.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, September 2010.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei "ImageNet: A Large-Scale Hierarchical Image Database ," *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] L-J. Li, H. Su, Y. Lim, R. Cosgriff, D. Goodwin, L. Fei-Fei "Object Bank," *http://vision.stanford.edu/projects/objectbank*, accessed on 09/26/2011.

[8] L-J. Li, L. Fei-Fei "What, where and who? Classifying event by scene and object recognition," *IEEE International Conference in Computer Vision*, 2007.

[9] Z. Zhao, S. Sharma, A. Anand, F. Morstatter, S. Alelyani, H. Liu "Advancing Feature Selection Research," *ASU Feature Selection Repository*, 2010.

[10] C.-C. Chang, C-J. Lin "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[11] A. Oliva, A. Torralba "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, Vol. 42, No. 3, 2001.

[12] G. Csurka, C. Bray, C. Dance, L. Fan "Visual Categorization with Bags of Keypoints," *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.