TOWARDS OPTIMAL OBJECT BANK FOR SCENE CLASSIFICATION

Lei Zhang[†] *, Shouzhi Xie*[†] *and Xiantong Zhen*[‡]

[†] College of Information and Communication Engineering, Harbin Engineering University, Harbin, PRC

[‡] Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, UK

ABSTRACT

High-level image representations have drawn increasing attention in visual recognition, e.g., scene classification, since the invention of the object bank (OB). The object bank represents an image as a response map of a large number of pre-trained object detectors and has achieved superior performances for visual recognition.

However, the object bank representation can be further improved by considering the distributions of the object across categories and the discriminative contributions to the image representation. In this paper, we propose an optimal object bank (OOB) by imposing weights on the detectors according to their discriminative abilities. Through extensive experiments on two benchmark datasets: UIUC-Sports dataset and 15-Scene dataset, we prove that the proposed OOB can significantly improve the original object bank and achieves state-of-the-art performances.

***Index Terms*—** Optimal object bank, discriminative coefficient, scene classification

1. INTRODUCTION

Mid-level image representations, e.g., the bag-of-word (BoW) model, have long dominated in visual recognition due to its simplicity and computational efficiency. However, they fail to capture enough semantic meanings of images and still pose gap between image representations and human visual perceptions. Recently, high-level image representations have attracted increasing interest in visual recognition, among which, the object bank (OB) [1], has been proposed and demonstrated to be more effective in image representation for scene classification. The advantages of the object bank lies in its ability to extract more semantic meanings of images offering a rich and high-level representation of image. The success of the object bank hinges on the detection of meaningful visual concepts. Fortunately, the availability of large-scale image datasets, e.g., LabelMe [2] and ImageNet [3] makes it possible to obtain object detectors for a wide range of visual concepts. In addition, Zipf's law [4] known in the natural language processing implies that only a small proportion of objects account for the majority of object instances. By selecting a proper object list, the main content of image can be sufficiently represented by the responses of the object filters with semantic concepts.

Actually, the object bank represents an image by the responses of pre-trained object filters. Due to the explicit detection of objects in images, the object bank provides an effective avenue to understand scene images. However, the object detectors have no knowledge of the distributions of objects in images and their responses are treated equally in the final image representation, which would suffers from being less discriminative for classification.

We argue that the object bank can be further improved by taking into account the distinct roles of object detectors in image representations. As a motivation, we propose an optimal object bank (OOB) for high-level scene representations, in which the object detectors are weighted according to their discriminability in the representations.

2. OBJECT BANK

2.1. Object detectors

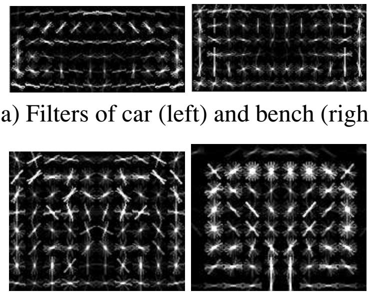
In the object bank representation [1], the latent SVM object detectors [5] for most of the blobby objects (tables, cars, humans, etc) and a texture classifier [6] for more texture and material based objects (sky, road, sand, etc) are used to build a bank of objects. According to the frequency of occurrences of objects in different datasets, 177 of the most frequent objects are selected. Shown in Fig. 1 are examples of filters from [1], which, as we can see, reflect the roughly the outlines of the objects.

2.2. High-level representations

Given an image G and a filter F in the object bank, the response of the filter at point (x, y) is the sum of the products of the filter coefficients and the corresponding neighborhood points in the area spanned by the filter mask, which can be formulated as:

$$\sum_{x',y'} F\left[x', y'\right] \cdot G\left[x + x', y + y'\right] \tag{1}$$

This work is supported by Young Teacher Support Plan by Heilongjiang Province and Harbin Engineering University in China (No.1155G17), and Fundamental Research Funds for the Central Universities in China.



(a) Filters of car (left) and bench (right)

(b) Filters of elephant (left) and tree (right)

Fig. 1. Illustration of filters.

Moving the center point (x, y) to go through all the pixels in the image, responses for all the pixels are obtained. Since each filter can reflect the outline of an object, the sum operation in Eq. (1) essentially calculates the similarity between the object in the filter and a patch around a pixel in the image. If normalized, the maximum value can be viewed as the probability of the object occurring in the image.

An image is finally represented by a feature vector concatenating the maximum responses of all the filters. If the scales and the location information by spatial pyramid are further considered, the feature vector could be in a high-dimensional space. In our experiments, we follow the setting as [1], in which 177 object filters (with front and profile models), 6 scales and 3-level spatial pyramid $(1+4+16)$ are employed. The final feature vector is of $177 \times 2 \times 6 \times 21 = 44604$ dimensions. If no location information is used, the dimensionality will be $177 \times 2 \times 6 = 2124$.

3. OPTIMAL OBJECT BANK

With each image represented by a high-dimensional vector of filter responses, the whole training set can be viewed as the matrix as shown in Fig. 2. Each column is the image feature vector, while each row (dimension) spans different images across different scene classes. Note that an object with different scales and angles will occur in multiple dimensions of the feature vector.

In the original object bank representation, each vector is treated as the final feature and is feed to a classifier, e.g., SVM, for classification, which unfortunately ignores the information of the occurrences of objects in images from different scene classes and the whole dataset, resulting to be less discriminative. We aim to consider the distributions of the object detectors across all images in the dataset and weight them according to their discriminability.

In Fig. 3, we plot the distributions of the object ‘elephant’ and ‘tree’. As can be seen, both of them follow the Gaussian distributions with different means and variances.

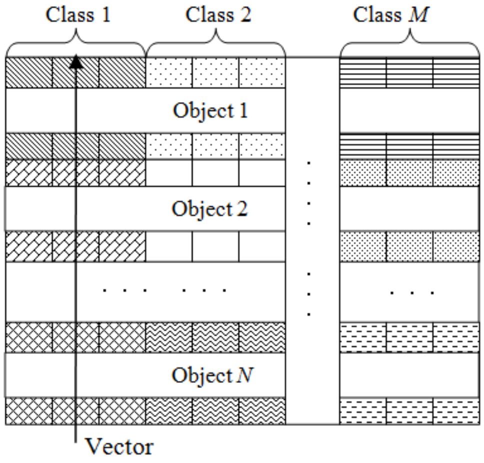


Fig. 2. Illustration of feature vectors of samples in the training set.

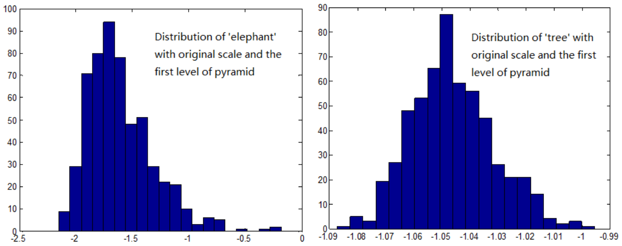


Fig. 3. Distributions of the objects: ‘elephant’ and ‘tree’ in the training set.

3.1. Unsupervised learning of coefficients

Assuming here that each dimension in the feature vector obey the Gaussian distribution as shown in Fig.3, we normalize each dimension to standard Gaussian distribution.

Let $G_a = [r_{a1}, r_{a2}, ..., r_{aN}]$ be the vector of image a, and there are A images in training dataset. The mean m_n and standard deviation s_n for n-th dimension can be obtained by

$$m_n = \frac{1}{A}\sum_{a=1}^{A} r_{an} \tag{2}$$

$$s_n = \sqrt{\frac{1}{A-1}\sum_{a=1}^{A}(r_{an} - m_n)^2} \tag{3}$$

where each dimension can be viewed as an object. Then the final feature vector of image a becomes $\hat{G_a}$ as $[\hat{r}_{a1}, \hat{r}_{a2}, ..., \hat{r}_{aN}]$ with

$$\hat{r}_{an} = \frac{r_{an} - m_n}{s_n} \tag{4}$$

The learned m_n and s_n from the training set are applied to the corresponding dimensions of the feature vectors in the test set. Without *priori knowledge*, e.g., the class labels of images, each dimension (object) is rectified according to the distribution of this object among the whole dataset in Eq.(4).

3.2. Relation to *tf-idf*

Looking deeply into the normalization operation, we can find that it shares similar ideas with the 'term frequency-inverse document frequency', *tf-idf*, which has been successfully applied to video retrieval in [7].

$$term_{id} = tf_{id} \lg \frac{N}{n_i} \tag{5}$$

where tf_{id} is the term i frequency (*tf*) in document d and $\lg \frac{N}{n_i}$ is inverse document frequency (*idf*), which is in inverse proportion to the number of occurrences of term i in the whole database. In Eq. (5), the most important thing lies in the fact that $term_{id}$ depends not only on the document (image) itself, but also on the distribution of this term among all documents (images), where a higher n_i will produce a lower $term_{id}$. This is actually in line with the idea in Eq. (4). Denominator in Eq. (4) makes the width of the Gaussian density function of each dimension equal (unit variance), and denominator (the distance to the center of the density function) can reflect occurrence frequency of the object. Then similar to *tf-idf* in Eq. (5), if the value in one dimension is closer to the corresponding distribution center, it means the object occurs in more images and therefore this dimension will play less effect.

3.3. Discriminative weighting

In order to further improve the discriminability of the final image representations, we consider the roles of the filters on distinguishing images from different classes by taking into account their discriminabilities between scene classes. Inspired by the linear discriminative analysis (LDA) which finds a linear projection by the fisher criterion: minimizing the within-class $S_W(\hat{r}_n)$ and maximizing the between-class scatter $S_B(\hat{r}_n)$, we use the fisher criterion to weight the filters. The weight (discriminative coefficient) for the n-th dimension can be obtained by:

$$w_n = \frac{S_B(\hat{r}_n)}{S_W(\hat{r}_n)} \tag{6}$$

where

$$S_W(\hat{r}_n) = \sum_{c=1}^{M} \sum_{a}^{n_c} (\hat{r}_{an} - \hat{m}_{cn})^2 \tag{7}$$

$$S_B(\hat{r}_n) = \sum_{c=1}^{M} n_c \times (\hat{m}_{cn} - \hat{m}_n)^2 \tag{8}$$

$\hat{m}_{cn}$ is the n-th dimension of mean vector of class c after the normalization as in Eq. (4) and $\hat{m}_n$ is the corresponding n-th dimension of mean vector of all training data after the normalization. n_c is the number of images in class c.

15-Scene			
bedroom	inside of city	industry	kitchen
mountain	living room	highway	suburb
coast	open country	store	office
street	tall building	forest	
UIUC-Sports			
rowing	snow boarding	badminton	polo
sailing	rock climbing	croquet	bocce

Table 1. Scene categories in 15-Scene and UIUC-Sports

Then if the n-th dimension has a larger between-class scatter and a smaller within-class scatter, it will be more discriminative in the representations and more useful for the final classification, and therefore should be given more weights. The weighting can be simply done by:

$$\check{r}_{an} = w_n \times \hat{r}_{an} \tag{9}$$

After the weighting, we obtain the final optimal object bank (OOB) representation with each dimension $\check{r}_{an}$ carrying more discriminative information.

4. EXPERIMENTS

To validate the effectiveness of the proposed optimal object bank for image representation and classification, we have conducted extensive experiments on two benchmark scene datasets, i.e., the 15-Scene and UIUC-Sports datasets shown in Table 1.

4.1. Experiments setting

We follow the experimental settings in [1]. For the **15-Scene** dataset [8], 100 images are randomly selected as training data and the rest for testing. For the **UIUC-Sports** dataset [9], 70 images are randomly drawn for training and 60 for testing. A linear SVM classifier [10] is employed for the final scene classification. We have experimented on both the feature vectors with (44604 dimensions) and without location information (2124 dimensions).

4.2. Results on 15-Scene

The results on the 15-Scene dataset are shown in Table 2. We also report the results with only normalized distributions (ND). The proposed optimal object bank (OOB) outperforms the original object bank (OB), especially when no location information is used. Interestingly, the simple normalizing distribution (ND in Table 2 and Table 3) is even better than both

	with location information	without location information
OB	82.03%	78.58%
ND	83.17%	83.23%
OOB	82.52%	**83.52%**

Table 2. Performance comparison on 15-scene dataset.

	with location information	without location information
OB	77.5%	73.60%
ND	79.77%	80.21%
OOB	**82.33%**	81.63%

Table 3. Performance comparison on UIUC-Sports dataset.

OB and OOB when location information is included in the features. Our best result is 83.52%, which increases the original object bank by 1.5%.

The comparison of our results with the state-of-the-art methods is shown in Table 4. Our OOB significantly outperforms the recent methods proposed in [11, 12], which combined co-occurrence and locality with spatial information respectively.

4.3. Results on UIUC-Sports

Results on UIUC-Sports dataset are reported in Table 3. The proposed OOB significantly improves the original OB, from 77.5% to 82.33% with location information and from 73.60% to 81.63% without location information. UIUC-Sports is regarded as a challenging dataset due to the complexity of scenes in the dataset. However, the superiority of OOB over OB is even more significant, which proves the effectiveness of the proposed OOB.

We have also compared OOB with the state-of-the-art methods in Table 4. Again OOB has achieved much better results than those of recently proposed methods in [13, 14].

Note that on both 15-Scene and UIUC-Sports, our OOB produces remarkable results with and without the location information. On 15-Scene, OOB without location information can even outperform that with local information, which demonstrates the robustness of OOB.

	OOB	State-of-the-Art	
15-Scene	**83.52%**	82.51% [11]	82.67% [12]
UIUC-Sports	**82.33%**	79.37% [13]	65.00% [14]

Table 4. Performance comparison of the proposed OOB with state-of-the-art.

5. CONCLUSION

In this paper, we have proposed an optimal object bank (OOB) for scene classification. By considering the distributions of the objects across scene classes and their discriminabilities, OOB can provide more effective high-level representations than the original object bank (OB).

Extensive experiments on two benchmark 15-Scene and UIUC-Sports datasets have demonstrated that the proposed OOB can significantly improve the original OB obtaining state-of-the-art performances and proven the effectiveness of the proposed OOB for scene classification.

6. REFERENCES

[1] L.J. Li, H. Su, E.P. Xing, and L. Fei-Fei, “Object bank: A high-level image representation for scene classification and semantic feature sparsification,” *NIPS*, vol. 24, 2010.

[2] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, “Labelme: a database and web-based tool for image annotation,” *IJCV*, vol. 77, no. 1, pp. 157–173, 2008.

[3] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.

[4] R. Edwards and L. Collins, “Lexical frequency profiles and zipf’s law,” *Language Learning*, vol. 61, no. 1, pp. 1–30, 2011.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[6] D. Hoiem, A.A. Efros, and M. Hebert, “Automatic photo pop-up,” in *ACM TOG*, 2005, vol. 24, pp. 577–584.

[7] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *ICCV*, 2003, pp. 1470–1477.

[8] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, 2006, vol. 2, pp. 2169–2178.

[9] L.J. Li and L. Fei-Fei, “What, where and who? classifying events by scene and object recognition,” in *ICCV*, 2007, pp. 1–8.

[10] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A library for support vector machines,” *ACM TIST*, vol. 2, pp. 27:1–27:27, 2011.

[11] Y. Yang and S. Newsam, “Spatial pyramid co-occurrence for image classification,” in *ICCV*, 2011, pp. 1465–1472.

[12] A. Shabou and H. LeBorgne, “Locality-constrained and spatially regularized coding for scene categorization,” in *CVPR*, 2012, pp. 3618–3625.

[13] S. Gao, L.T. Chia, and I.W.H. Tsang, “Multi-layer group sparse codingłfor concurrent image classification and annotation,” in *CVPR*, 2011, pp. 2809–2816.

[14] Z. Niu, G. Hua, X. Gao, and Q. Tian, “Spatial-disclda for visual recognition,” in *CVPR*, 2011, pp. 1769–1776.