EXTENDED-BAG-OF-FEATURES FOR TRANSLATION, ROTATION, AND SCALE-INVARIANT IMAGE RETRIEVAL

Chia-Yin Tsai[†] *Ting-Chu Lin*[‡] *Chia-Po Wei*[⋆] *Yu-Chiang Frank Wang*[⋆]

[†] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA
[‡] Department of Computer Science, Columbia University, New York, USA
[⋆] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

While bag-of-features (BOF) models have been widely applied for addressing image retrieval problems, the resulting performance is typically limited due to its disregard of spatial information of local image descriptors (and the associated visual words). In this paper, we present a novel spatial pooling scheme, called extended bag-of-features (EBOF), for solving the above task. Besides improving image representation capability, the incorporation of the our EBOF model with a proposed circular-correlation based similarity measure allows us to perform translation, rotation, and scale-invariant image retrieval. We conduct experiments on two benchmark image datasets, and the performance confirms the effectiveness and robustness of our proposed approach.

Index Terms— Image retrieval, bag-of-features

1. INTRODUCTION

The amount of online image data is exploding in the past decade due to the rapid growth of Internet users. Since most of such data are not properly tagged when uploading, how to search or retrieve the images of interest is still a very challenging task. This is the reason why content-based image retrieval (CBIR) attracts the attention of researchers in related fields. The use of image descriptors like SIFT [1] are popular in terms of describing the visual appearances of images. Based on the extracted SIFT descriptors, the use of the bag-of-features (BOF) model [2] provides a robust image representation, which is a histogram indicating the numbers of occurrences of each learned visual word.

Although the use of BOF models has been shown to be very effective [2, 3, 4], it discards the spatial information of the visual words (or the associated image descriptors) when describing each image. To address this problem, Lazebnik *et al.* [5] proposed a spatial pyramid matching (SPM) and characterized each image by concatenating multiple BOF models at different positions and scales. Recently, Cao *et al.* [6] chose to pool the local image descriptors from each image in a particular spatial order. Instead of explicitly dividing an image into different regions for pooling, the co-occurrence of visual words were also utilized to improve the image retrieval or categorization tasks [7, 8].

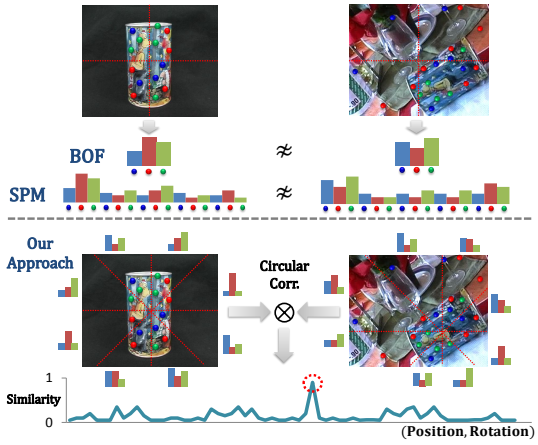


Fig. 1. Advantages of our proposed spatial pooling scheme for translation, rotation, and scale-invariant image retrieval.

In this paper, we present a novel pooling scheme for BOF, named *extended bag-of-features* (EBOF). While the goal of EBOF is to better represent an image by preserving the spatial information of visual words, the integration of EBOF with our proposed circular-correlation based algorithm further allows us to perform translation, rotation, and scale-invariant image retrieval. It is worth noting that, when performing image retrieval, our method does not need to assume self-similarity or to calculate the co-occurrences of visual words explicitly. Later in our experiments, we will verify the effectiveness and robustness of our proposed method.

2. OUR PROPOSED METHOD

2.1. A Brief Review of BOF, SPM, and SBOF

To represent an image, the bag-of-features (BOF) model [2] quantizes image descriptors such as SIFT [1] into distinct visual words. As a histogram-based representation, each attribute of BOF indicate the number of occurrences of each word in an image. While BOF has been applied to image retrieval or classification, it discards the spatial information of

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

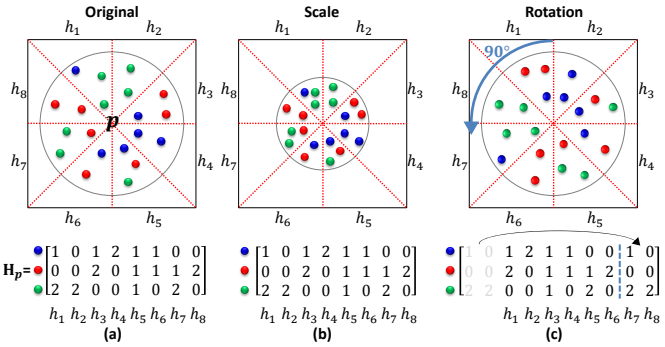


Fig. 2. An example of our extended bag-of-features (EBOF) model $\mathbf{H}_p$. (a) Original image with EBOF centered at p, (b) a scaled version of (a), and (c) a rotated version of (a). Note that each colored point denotes a local image descriptor with a corresponding visual word.

visual words and thus limits the representation capability.

To address the above problem, spatial pyramid matching (SPM) [5] extends BOF by partitioning an image into several grids at different scales. It pools the BOF models from each grid and concatenates them as a final feature vector. Although the spatial order of the visual words is preserved by SPM, it cannot be easily extended to retrieval or classification problems in which the object of interest exhibits translation, rotation, or scale variations in an image.

Recently proposed in [6], spatial-bag-of-features (SBOF) pools BOF models for each visual word from different designated regions within an image, so that translation, rotation, and scale-invariance can be possibly achieved. Since SBOF only preserves the spatial information of *each* word when deriving their feature representation, their disregard of visual word co-occurrences during their pooling process would limit their performance (as verified later by our experiments).

2.2. Extended Bag-of-Features

Unlike SPM which pools and concatenates BOF models from different grids of an image as an one-dimensional feature vector, we choose to uniformly divide an image into L fan-shaped sub-images (centered at p), as shown in Figure 2(a). For a codebook with K codewords, we calculate our extended bag-of-features (EBOF) model at center p of an image as

$$\mathbf{H}_p = [\mathbf{h}_{\{p,1\}}, \mathbf{h}_{\{p,2\}}, ..., \mathbf{h}_{\{p,L\}}], \tag{1}$$

where $\mathbf{h}_{\{p,i\}} \in \mathbf{R}^{K\times 1}$ is the BOF of the ith sub-image, and $\mathbf{H}_p$ is of size $K \times L$. Once this EBOF is constructed, we apply a 2D Gaussian weighting function (centered at p) to suppress the contributions of visual words farther away from p. In our work, we set the standard deviations of both dimensions of this Gaussian function as half of the longer length of the image. Finally, we normalize this calculated EBOF by $\mathbf{H}_p/\|\mathbf{H}_p\|_1$ for later correlation and retrieval purposes.

Comparing Figures 2(a) and (b), we see that a scale change will not affect the EBOF model, and thus scale invariance can be achieved. As for rotation variations as shown in Figure 2(c), the resulting EBOF will be a shifted version (in column) of that of the original image. In addition to scale and rotation changes, we also need to deal with translation variations. In our work, we consider that the object of interest is located at the center of the query image Q when calculating its EBOF $\mathbf{H}^Q$ as the image feature. Thus, the subscript p is ignored in $\mathbf{H}^Q$ for simplicity. For the target images to be retrieved, we uniformly divide each image I into $5 \times 5 = 25$ grids, and use the center p of each grid to extract the EBOF model for deriving different $\mathbf{H}_p^I$ (see discussions in Section 2.3.2 for this choice).

Once the EBOF models are extracted from both query and target images, we perform image retrieval based on the maximum similarity score between $\mathbf{H}^Q$ and each $\mathbf{H}_p^I$ for translation, rotation, and scale invariance, which will be detailed in the next subsection.

2.3. Image Retrieval with EBOF

2.3.1. Circular-correlation based image retrieval

We now discuss how we utilize the proposed EBOF model in (1) for addressing the retrieval task. Given a query image Q and a target image I in the database, we need to determine the similarity score between their EBOF models $\mathbf{H}^Q$ and $\mathbf{H}_p^I$. Recall that we only construct one EBOF for the query (centered at the query Q), and we have 25 EBOFs for I at different centers. We now determine $\mathbf{S}_p^{\{Q,I\}} = (\mathbf{H}^Q \otimes \mathbf{H}_p^I)$ as a K-by-L correlation matrix, and each row $\mathbf{r}_k$ of $\mathbf{S}_p^{\{Q,I\}}$ is calculated by

$$\mathbf{r}_k[l] = \sum_{m=1}^{L} \mathbf{H}^Q[k,m]\, \mathbf{H}_p^I[k, \text{mod}(l+m-1, L)], \tag{2}$$

where $l = 1, 2, ..., L$ denotes the number of rotation angles. From (2), one can see that we perform circular correlation between the kth rows of the EBOF models $\mathbf{H}^Q$ and $\mathbf{H}_p^I$, and thus the resulting vector $\mathbf{r}_k$ indicates the similarity of the kth visual word between these two images across different rotation angles. Once all rows of $\mathbf{S}_p^{\{Q,I\}}$ are obtained, we have each column of $\mathbf{S}_p^{\{Q,I\}}$ as the correlation response (i.e., similarity) between the BOF models between images Q and I at a specific rotation angle. As a result, we have $\mathbf{S}_p^{\{Q,I\}} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_L] = [\mathbf{r}_1^T; \mathbf{r}_2^T; ...; \mathbf{r}_K^T]$, where $\mathbf{s}_l \in \mathbf{R}^{K\times 1}$ and $\mathbf{r}_k \in \mathbf{R}^{L\times 1}$. As depicted in Figure 3, each column $\mathbf{s}_l$ represents the correlation between Q and I at a particular angle, while each row $\mathbf{r}_k$ denotes the correlation response of a particular visual word across different rotation angles.

To assess which rotation angle is most likely to be the match between Q and the image I, we apply the cosine similarity as the metric for determining the normalized similarity score between each column of $\mathbf{S}_p^{\{Q,I\}}$ and the autocorrelation output vector of the query Q. Note that the autocorrelation output vector of Q is calculated as $\mathbf{a} = diag(\mathbf{H}^Q \cdot (\mathbf{H}^Q)^T)$, in which each entry indicates the energy of the BOF model for

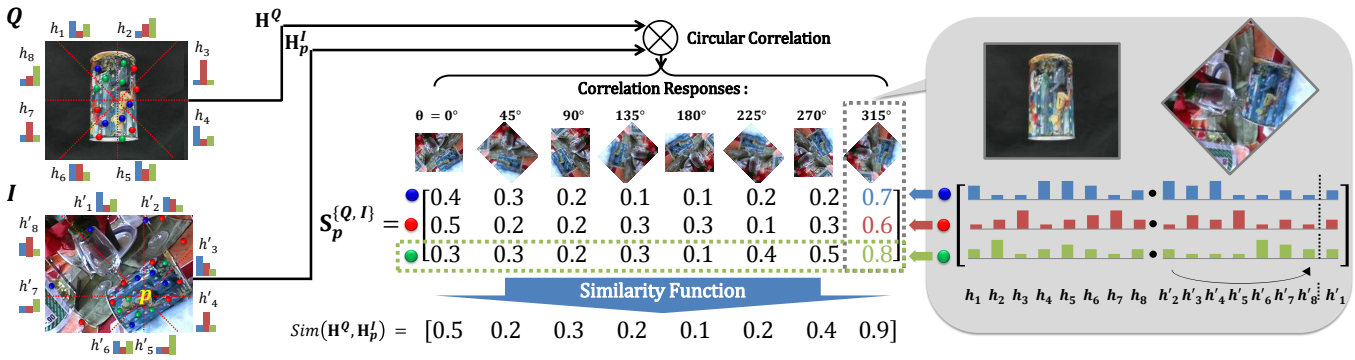


Fig. 3. Illustration of image retrieval using our proposed EBOF models. Note that each row in $\mathbf{S}_p^{\{Q,I\}}$ indicates the correlation response of a visual word between images Q and I across different rotation angles, while each column represents their correlation at a specific rotation angle. $Sim(\mathbf{H}^Q, \mathbf{H}_p^I)$ denotes the normalized similarity between Q and I at a particular center p.

the corresponding sub-image. As depicted in Figure 3, this normalized similarity $Sim(\mathbf{H}^Q, \mathbf{H}_p^I)$ between images Q and I across L different rotation angles is calculated as:

$$Sim(\mathbf{H}^Q, \mathbf{H}_p^I) = [\cos(\mathbf{a}, \mathbf{s}_1), \cos(\mathbf{a}, \mathbf{s}_2), ..., \cos(\mathbf{a}, \mathbf{s}_L)]. \quad (3)$$

By identifying the largest value in $Sim(\mathbf{H}^Q, \mathbf{H}_p^I)$, the rotation angle at which Q and I are most similar to each other can be determined. We then repeat the above correlation process for $\mathbf{H}_p^I$ at different centers p for translation invariance. The maximum output across different $Sim(\mathbf{H}^Q, \mathbf{H}_p^I)$ is the final similarity score for retrieval, i.e., $Score(Q, I) = \max_{p=1}^{P} \{\max\{Sim(\mathbf{H}^Q, \mathbf{H}_p^I)\}\}$.

2.3.2. *Translation, rotation, and scale invariance*

To deal with translation variations when performing image retrieval, we consider that the object of interest is presented at the center of the query image Q without loss of generality. Thus, only one EBOF model $\mathbf{H}^Q$ is constructed (i.e., the one centered at Q). As for the image I in the database to be retrieved, we uniformly divide I into $5 \times 5 = 25$ grids and consider p as the centers of each grid when extracting the corresponding EBOF models. The EBOF models at 25 different locations in I are calculated for representing this image. We perform the above circular-correlation based procedure and consider the maximum normalized similarity output across 25 different $Sim(\mathbf{H}^Q, \mathbf{H}_p^I)$ as the final retrieval score. If p is located at/near the center of the object of interest in I, the corresponding EBOF model at a particular rotation angle would produce the highest similarity score. This is how translation-invariant image retrieval is achieved.

To verify the above setting is sufficient for translation-invariant retrieval performance, Figure 4 plots the mean average precision (MAP) scores of the ETHZ Toys Dataset [9] using different numbers of grids (from 1×1 up to 9×9). From this figure, it can be seen that the use of $5 \times 5 = 25$ grids is sufficient for producing improved retrieval results (compared

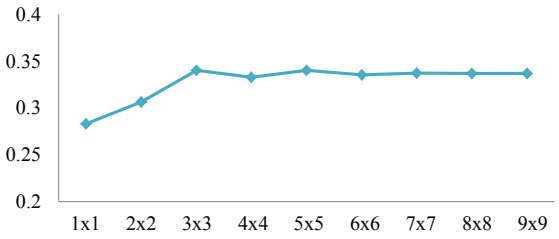


Fig. 4. MAP of the ETHZ Toys Dataset with different numbers of grids of an image (from $1 \times 1 = 1$ up to $9 \times 9 = 81$) for translation invariance.

to 1×1 without shift invariance), and uses of larger numbers of grids are not necessary. This because that our retrieval algorithm is based on the *maximum* correlation score. Thus, our choice is preferable for producing satisfactory translation-invariant results.

As discussed earlier in Section 2.2, our proposed EBOF model is robust to *scale* variations when describing an image. Since *rotation* variations would produce shifted EBOF models $\mathbf{H}_p$ in columns, we calculate the similarity between the resulting EBOF models for rotation invariance. By identifying the rotation angle of I which results in the associated rotated/shifted version to be most similar to Q, rotation-invariant image retrieval can be achieved. Similar to the above tests/verifcations for shift invariance, we also vary the number L of fan-shaped sub-images and evaluate the associated performance of rotation invariance on the ETHZ dataset. We also observed that L from 6 to 10 achieved comparable improved results as those with smaller L values. Therefore, our choice of $L = 8$ is sufficient for producing rotation-invariant results.

3. EXPERIMENTS

3.1. Datasets

We first consider the Oxford 5K dataset [10], which contains 5026 images of the landmarks in Oxford. Each image of Oxford 5K contains around 3000 SIFT interest points, and the

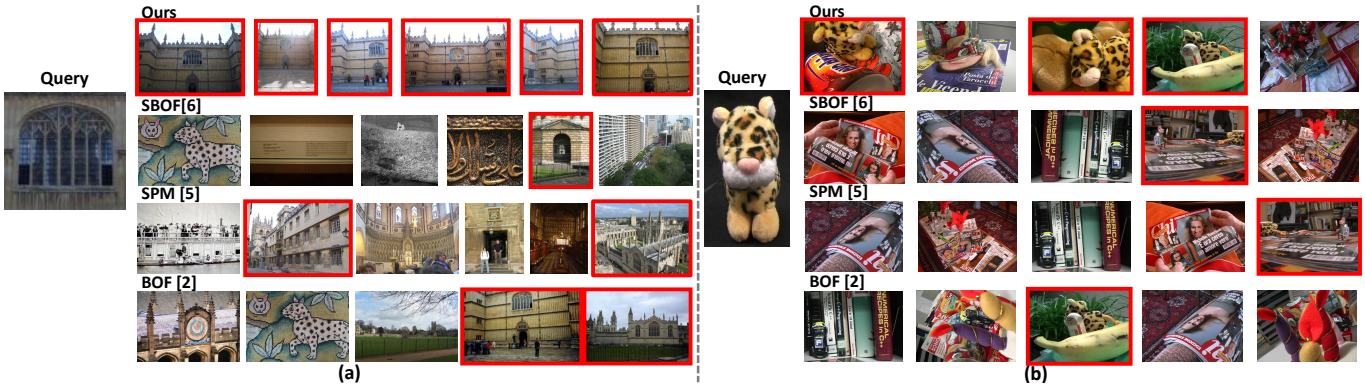


Fig. 5. Example retrieval results on (a) Oxford 5K and (b) ETHZ Toys datasets. Each row shows top retrieved outputs produced by different methods, and the relevant ones are circled in red.

longer dimension of these images is about 1024 pixel. This dataset provides 55 queries and the ground truth for all images to be retrieved. We resize each query image so that its longer side is 500 pixels. For computation efficiency, we set codebook size as $K = 1000$.

Since the landmarks in the Oxford 5K dataset typically do not exhibit significant rotation variations, we further consider the ETHZ Toys dataset [9], which contains 40 query images for 9 different objects and a total of 23 images to be retrieved. The test images are heavily cluttered, so the toy objects might be partially occluded in addition to translation, rotation, or scale variations which make the retrieval task more challenging. In our experiments, we resize the query image so that the longer side is 100-pixel wide, and we also the codebook size $K = 1000$.

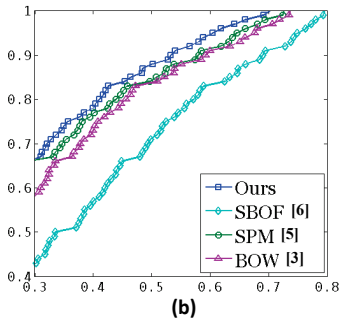
3.2. Discussions

We compare our method with three BOF-based approaches: the standard BOF [3], SPM [5], and SBOF [6]. For SPM, we divide each image into 2×2 grids and thus a total of $1 + 2 \times 2 = 5$ BOF will be concatenated as features. As for SBOF, we consider the number of fan-shaped sub-images as $L = 8$ (as we do). The number of angles for performing linear projections is 4 for SBOF, and we also consider the same 25 centers p for its circular projection. We use the same codebook with size $K = 1000$ for all approaches to be evaluated. It is worth noting that we do not perform feature selection for SBOF (as [6] did). This is because we assume that no labeled training data is available when performing retrieval (which is practical for real-world scenarios).

When performing retrieval, we consider the Euclidean distance as the similarity metric for BOF and SPM models. As for SBOF, we apply cosine similarity as suggested in [6]. Example retrieval results for the two datasets are shown in Figure 5. From the table shown in Figure 6(a), it is clear that we achieved the highest mean average precision (MAP) scores for both datasets. To better visualize the differences, we further plot the Receiver Operating Characteristic (ROC) curves for the Oxford 5K dataset in Figure 6(b), which shows

Methods	Oxford 5K	ETHZ Toys
BOF [3]	0.055	0.232
SPM [5]	0.146	0.243
SBOF [6]	0.030	0.234
Ours	**0.167**	**0.333**

(a)



(b)

Fig. 6. Performance comparisons. (a) MAP scores for Oxford 5K and ETHZ Toys datasets, (b) ROC for the Oxford 5K dataset.

that our method outperformed other approaches. We note that, since only a codebook with 1000 words was considered, the reported MAP values were not comparable as those using 1M words in [6]. However, it is clear that our approach was able to produce better retrieval results and achieved improved MAP scores when comparing to BOF-based methods with the same codebook. We also note that, the runtime estimate of our method is around 0.35 seconds per image on a PC with Intel Core 2 Duo CPU 2.66 GHz and 4G RAM (programmed in Matlab). From the above empirical results, the effectiveness of our proposed image retrieval framework can be verified, and our meethod is shown to be preferable when translation, rotation, and scale variations are presented.

4. CONCLUSION

We proposed an extended bag-of-features (EBOF) model for image retrieval. Our EBOF is able to exploit the spatial information of visual words presented in images. Together with a circular-correlation based similarity measure, the use of EBOF has been shown to achieve translation, rotation, and scale-invariant image retrieval. Unlike prior retrieval works, our approach does not require assumption of self-similarity or the calculation of visual word co-occurrences. Experiments on two benchmark datasets verified the effectiveness and robustness of our proposed method.

5. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *Int. J. Computer Vision*, 2004.

[2] G. Csurka, C. R. Dance, L. Fan, J. Williamowski, and C. Bray, "Visual categoriztion with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[3] J. Yang, T.-G. Jiang, A. G. Hauptmann, and G.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 2007.

[4] D. Li, L. Yang, X.-S. Hua, and H.-J. Zhang, "Large-scale robust visual codebook construction," in *ACM Multimedia*, 2010.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE CVPR*, 2006.

[6] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang., "Spatial-bag-of-features," in *IEEE CVPR*, 2010.

[7] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *IEEE CVPR*, 2011.

[8] C.-F. Chen and Y.-C. F. Wang., "Exploring self-similarity of bag-of-features for image classification," in *ACM Multimedia*, 2011.

[9] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Simultaneous object recognition and segmentation from single or multiple model views," *Int. J. Computer Vision*, 2006.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE CVPR*, 2007.