Large Scale Image Retrieval with Practical Spatial Weighting for Bag-of-Visual-Words

Fangyuan Wang[1,2], Hai Wang[1], Heping Li[1], and Shuwu Zhang[1]

[1] High-Tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] WaSu Media Group Co.Ltd, Hangzhou, China
{fangyuan.wang,hai.wang,heping.li,shuwu.zhang}@ia.ac.cn

Abstract. Most large scale image retrieval systems are based on Bag-of-Visual-Words (BoV). Typically, no spatial information about the visual words is used despite the ambiguity of visual words. To address this problem, we introduce a spatial weighting framework for BoV to encode spatial information inspired by Geometry-preserving Visual Phrases (GVP). We first interpret GVP method using this framework. We reveal that GVP gives too large spatial weighting when calculating L2-norm for images due to its implicit assumption of the independence of co-occurring GVPs. This makes GVP sensitive to images with small number of visual words. Then we propose an improved practial spatial weighting for BoV (PSW-BoV) to alleviate this effect while keep the efficiency. Experiments on Oxford 5K and MIR Flickr 1M show that PSW-BoV is robust to images with small number of visual words, and also improves the general retrieval accuracy.

Keywords: image retrieval, spatial weighting, bag-of-visual-words, geometry-preserving visual phrases.

1 Introduction

Large scale image retrieval is receiving more and more attentions owing to its great potential in application and importance of theory in research. The goal of an image retrieval system is to return the similar images in a ranked list for a query image.

In order to deal with large scale image dataset, most existing state-of-the-art image retrieval systems are based on bag-of-visual-words (BoV) model, which is firstly introduced as Video-Google in [3]. Numerous successful works have been proposed to improve the retrieval accuracy and efficiency based on this model. The vocabulary tree [4] and approximate nearest neighbor [5] increase the efficiency of building a large vocabulary, while soft matching [6] and hamming embedding [7] address the hard quantization problem of visual words. But, in most of these approaches, spatial information which is useful to alleviate the ambiguity of visual words is usually ignored. Several researches have been conducted to introduce spatial information into BoV model. The RANSAC [5] re-introduces spatial information in the post-processing step through geometry verification which is usually computationally expensive. Spatial

S. Li et al. (Eds.): MMM 2013, Part I, LNCS 7732, pp. 513–523, 2013.

Pyramid Matching [8] (SPM) encodes rigid spatial information by quantizing the image space and lacks the invariance to transformations. Spatial-bag-of-features [2] handle variances of SPM by changing the order of the histograms; the spatial histogram of each visual word is rearranged by starting from the position with the maximum frequency. But, the arrangement may not correspond to the true transformation. Geometry-Preserving Visual Phrases (GVP) [1] uses the co-occurring GVPs between images to encode both local and long-range spatial information. When calculating the number of co-occurring GVPs, it implicitly assumes all GVPs in an offset bin are independent, which makes this method sensitive to distracting images with small number of visual words. Some researchers also consider to encode spatial information through introducing spatial weighting for visual words, but their methods either need learning step[12] or are difficult to be facilitated by inverted files[13,14].

To address this problem, we introduce a spatial weighting framework for BoV inspired by GVP method. Using this framework, we reveal that GVP method is sensitive to images with small number of visual words. Further more to alleviate this effect, we propose a practical spatial weighting for BoV (PSW-BoV) which calculates the spatial weighting for visual words based on the following two principles:

(1) When the dependence of GVPs is not serious, which means the number of co-occurring visual words in an offset bin is not too big, we use similar spatial weighting for visual words as GVP method;

(2) When the dependence of GVPs is very serious, which means the number of co-occurring visual words in an offset bin is very big, we use a much smaller spatial weighting for visual words than GVP method;

Although PSW-BoV is quite simple, experiments on Oxford 5K[5] and MIR Flickr 1M datasets [11] demostrate that it can alleviate the sensitive effect of GVP to a large extent and significantly improve the general retrieval accuracy.

The rest of the paper is organized as follows: section 2 introduces the spatial weighting framework for BoV; in section 3, we interpret and analyze GVP using the spatial weighting framework; section 4 introduces PSW-BoV; section 5 is the comparative experiments; finally we draw conclusions in section 6.

2 Spatial Weighting Framework for BoV

BoV typically represents an image I_i as a vector $V(I_i)$, with one component for each visual word in the vocabulary. The j^{th} component $v_j(I_i)$ in the vector is the weight of the word *j*: the *tf-idf* weighting scheme [3] is usually used, which can be calculated using the following formular:

$$v_j(I_i) = \frac{n_{jI_i}}{n_{I_i}} \cdot \log(\frac{N}{n_j}) \tag{1}$$

where, n_{jI_i} is the number of word *j* in image I_i, n_{I_i} is the total number of words in image I_i, n_j is the number of images that contain word *j* and *N* is the total number of images in the whole dataset. The similarity of two images I_i and $I_{i'}$ is usually defined

as the *cosine* similarity of the two vectors: $< V(I_i), V(I_{i'}) > / \| V(I_i) \| \cdot \| V(I_{i'}) \|$. With large vocabularies, BoV representation is very sparse and inverted files can be used to facilitate the searching.

Typical BoV model just ignores the spatial information of visual words. An instinctive method is to mimick the *tf-idf* weighting to consider the spatial weighting for visual words. Suppose we have already got the spatial weighting $\alpha_j(I_i)$ for word j in image I_i, then the weighting component $v_j(I_i)$ changes to:

$$v_j(I_i) = \alpha_j(I_i) \cdot \frac{n_{jI_i}}{n_{I_i}} \cdot \log(\frac{N}{n_j}) \tag{2}$$

This can be regarded as a framework because we can use different methods to calculate the spatial weighting for visual word.

3 Spatial Weighting Interpretation of GVP

3.1 Interpretation

According to [1], a geometry-preserving visual phrase (GVP) of length k is defined as k visual words in a certain spatial layout. To tolerate shape deformation, the image space is quantized into bins. Each image is represented as a vector of GVPs. Similar to BoV model, the vector representation $V^k(I)$ is defined as the histogram of GVP of length k (k-GVP), with the i^{th} component representing the *tf-idf* weighting of phrases p_i. But, this kind of vector can be extremely long even when k=2 while a large vocabulary is used. However, if ignores the *idf* weights, the dot product of such vectors of two images equals the total number of co-occurring GVPs in these images, the L2-norm of a vector can be calculated by counting the co-occurring GVPs with itself, since $\| V^k(I) \| = \sqrt{V^k(I) \cdot V^k(I)}$. Then the *cosine* similarity can be calculated to measure the similarity between images as follows:

$$sim(I, I') = \frac{< V^k(I), V^k(I') >}{\| V^k(I) \| \cdot \| V^k(I') \|} = \frac{\sum_{m_{I,I'} \geq k} \left\{ \sum_{i=1}^{m_{I,I'}} \binom{m_{I,I'}}{k} \right\}}{\left(\sum_{m_{I,I} \geq k} \left\{ \sum_{i=1}^{m_{I,I}} \binom{m_{I,I}}{k} \right\} \right)^{1/2} \cdot \left(\sum_{m_{I',I'} \geq k} \left\{ \sum_{i=1}^{m_{I',I'}} \binom{m_{I',I'}}{k} \right\} \right)^{1/2}} \tag{3}$$

where, I, I' represents two images, $m_{I,I'}$, $m_{I,I}$ $m_{I',I'}$ is the co-occurring visual word number in an offset space bin between images I and I', I and itself, I' and itself respectively, $\binom{m_{I,I'}}{k}$, $\binom{m_{I,I}}{k}$, $\binom{m_{I',I'}}{k}$ is the number of co-occurring k-GVP in corresponding offset bin respectively.

If considering the *idf* weights of GVPs, the final similarity can be calculated as follows:

$$sim(I,I') = \frac{\sum_{m_{I,I'} \geq k} \left\{ \sum_{i=1}^{m_{I,I'}} \binom{m_{I,I'} - 1}{k-1} \cdot idf^2(w_i) \right\}}{\left(\sum_{m_{I,I} \geq k} \left\{ \sum_{i=1}^{m_{I,I}} \binom{m_{I,I} - 1}{k-1} idf^2(w_i) \right\} \right)^{1/2} \cdot \left(\sum_{m_{I',I'} \geq k} \left\{ \sum_{i=1}^{m_{I',I'}} \binom{m_{I',I'} - 1}{k-1} \cdot idf^2(w_i) \right\} \right)^{1/2}} \quad (4)$$

where, w_i is a visual word, $idf(w_i)$ is the *idf* weight of the visual word.

$\binom{m_{I,I'} - 1}{k-1}$, $\binom{m_{I,I} - 1}{k-1}$ and $\binom{m_{I',I'} - 1}{k-1}$ in equation (4) can be regarded as the spatial weighting for the visual words as formular (2), so GVP method is essentially equivalent to a spatial weighting method for BoV.

3.2 Analysis

As shown in formular (3), GVP method directly uses the combination number $\binom{m}{k}$ of all visual words in an offset bin as the co-occurring GVPs number. Obviously, they assume that all co-occurring GVPs in the same offset bin are totally independent. However, this assumption is not true as illustrated below.

Suppose both image I and I' contain visual words A, B and C, the calculation of co-occurring visual words can be shown in Fig. 1.

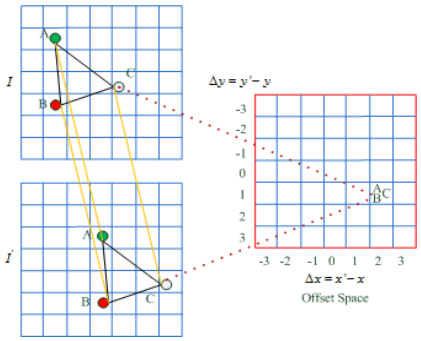

Fig. 1. Illustrative example for co-occurring GVPs. Different alphabets (A, B, C) represents different visual word.

As shown in Fig.1, the co-occurring GVPs are (AB), (BC), (AC). However, (AB) and (BC) share the same visual word B; (AB) and (AC) share the same visual word A; (BC) and (AC) share the same visual word C. This means different GVP may share the same visual word. Because all the visual words in (AC) are contained in (AB) and (BC), the spatial information encoded in (AC) is partially encoded in (AB) and (BC), vice versa. So they are also not spatially independent.

Therefore, the independence assumption of GVP is incorrect. This means the real number of GVPs should be less than the combination number of visual words.

Based on these analysis, we can infer that the spatial weighting of GVP (as shown in formular (4)) is prone to be bigger than ideal weighting (considered the dependence of GVPs), and the bias will increase as the number of co-occurring words increases.

In most cases of calculating co-occurring GVPs, due to the quantization effect of visual words (especially for a large vocabulary), the number of co-occurring visual words in an offset bin is not very big even for similar images. This means the dependence of GVPs is not very serious, so usually GVP method is still very effective.

However, GVP method is sensitive to distacting images with small number of visual words due to the independence assumption. Fig.2. is a real case example of the sensitivity effect of GVP.

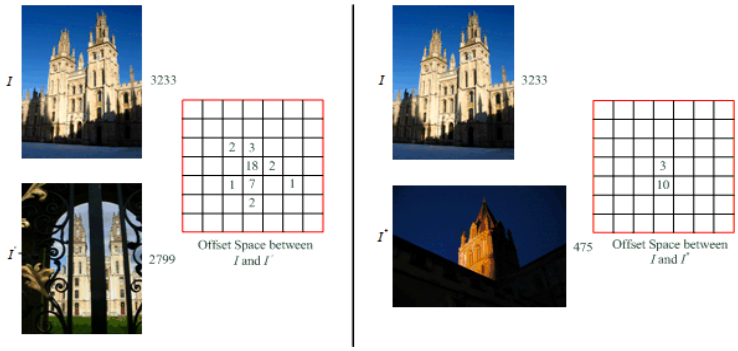

Fig. 2. Illustrative example of sensitivity of GVP to images with small number of visual words: image I, I',I^* has 3233,2799,475 visual words respectively, the numbers in bins are the number of co-occurring visual words. GVP method ranks I' and I^* incorrectly.

The reason can be explained as follows. As we known, the *consine* similarity needs to be normalized. When calculating the L2-norm for I', because the co-occurring visual words are generated with itself, the co-occurring number in the central offset bin is usually bigger than its total visual words number 2799 due to multiple times occurrence in I' of some words. If k=2, the number of GVP is more than 3915810 which is too large that the independece assumption is not reasonable any more. Therefore the spatial weighting ($m_{I',I'}-1$=2799-1=2798) is too large. The case for I^* is similar, its spatial weigthing is 474. However, the bias of the spatial weighting of I' is much bigger than that of I^* according to the above analysis (if I^* has similar number of visual words with I', the bias effect can be roughly cancelled out). So after normalized using the biased L2-norms, the biased similarity of I' and I becomes smaller than that of $I^*$and I.

4 Practical Spatial Weighting for BoV

In order to alleviate the sensitive effect of GVP, the best method is to apply independence analysis for the co-occurring GVPs. But, the co-occurring GVPs are generated in the searching step, which means too much analysis will dramatically affect the efficiency of retrieval. Thus we propose a practical spatial weighting for BoV (PSW-BoV) to handle this issue which do not need extra analysis.

4.1 Practical Spatial Weigthing Scheme

As discussed in section 3, in most of cases the number of co-occurring GVPs is not very big, while the big number of co-occurring GVPs only occurs in the central offset bin when calculating the L2-norm for an image, in which the independence assumption is not acceptable. Based on these analyses, the practial spatial weighting scheme can be described as follows:

(1) When calculating the inner product of the co-occurring visual words between images, we use the same spatial weighting $\binom{m-1}{k-1}$ as GVP method;

(2) When calculating L2-norm of each image, we use the same spatial weighting $\binom{m-1}{k-1}$ as GVP for the visual words in the bins whose total co-ocurring visual words number is small, while use a small number α as the spatial weighting for the visual words which co-occurrs in the central offset bin;

Besides, in order to make sure the final similarity lies between 0 and 1, we use the spatial weighting of co-occurring visual words between query image and target image to re-adjust the L2-norm of each image in the searching step.

The final similarity with the spatial weighting scheme can be formulated as follows:

$$sim(I,I') = \frac{< V'^k(I), V'^k(I') >}{\| V'^k(I) \| \cdot \| V'^k(I') \|} \tag{5}$$

where,

$$< V'^k(I), V'^k(I') >= \sum_{m_{I,I'} \geq k} \left\{ \sum_{i=1}^{m_{I,I'}} \binom{m_{I,I'}-1}{k-1} \cdot idf^2(w_i) \right\} \tag{6}$$

$$\| V'^k(I) \| = \left(\| V^{*k}(I) \|^2 + < V'^k(I), V'^k(I') > - \sum_{m_{I,I'} \geq k} \left\{ \sum_{i=1}^{m_{I,I'}} \alpha \cdot idf^2(w_i) \right\} \right)^{\frac{1}{2}} \tag{7}$$

where,

$$\| V^{*k}(I) \| = \left(\sum_{k \leq m_{I,I} < n_I} \left\{ \sum_{i=1}^{m_{I,I}} \binom{m_{I,I}-1}{k-1} \cdot idf^2(w_i) \right\} + \sum_{m_{I,I} \geq n_I} \left\{ \sum_{i=1}^{m_{I,I}} \alpha \cdot idf^2(w_i) \right\} \right)^{\frac{1}{2}} \tag{8}$$

$$\| V'^k(I') \| = \left(\| V^{*k}(I') \|^2 + < V'^k(I), V'^k(I') > - \sum_{m_{I,I'} \geq k} \left\{ \sum_{i=1}^{m_{I,I'}} \alpha \cdot idf^2(w_i) \right\} \right)^{\frac{1}{2}} \tag{9}$$

where,

$$\| V^{*k}(I') \| = \left(\sum_{k \leq m_{I',I'} < n_{I'}} \left\{ \sum_{i=1}^{m_{I',I'}} \binom{m_{I',I'}-1}{k-1} \cdot idf^2(w_i) \right\} + \sum_{m_{I',I'} \geq n_{I'}} \left\{ \sum_{i=1}^{m_{I',I'}} \alpha \cdot idf^2(w_i) \right\} \right)^{\frac{1}{2}} \tag{10}$$

Formular (5) is the final similarity, its numerator is calculated by formular (6) which corresponds to principle (1), its denominator is calculated by formular (7) and (9), which are the final L2-norms of I and I' re-adjusted based on the preliminary

L2-norms calculated by formular (8) and formular (10) respectively. k is moved from GVP, here we also only consider the visual words in the bins whose total co-occurring word number is $\geq k$. n_I and $n_{I'}$ are the total visual word numbers in image I and I'.

4.2 Searching with Practical Spatial Weighting

The practical spatial weighting scheme can be integrated into the searching step with inverted file which keeps one entry for each word occurrence with the image ID and word location [1]. For each image in database, we keep M bins to calculate the co-occurring words with query image, and another M bins to accumulate the summation of the *idf* weights, where M is the number of possible offsets.

(1) Initialize the two M bins for each image in the database to 0.

(2) For each word w in query image I , retrieve the image IDs and locations of the occurrences of w through the inverted files. For each retrieved word occurrence w' in image I_j , calculate the offset w and w', increment the corresponding offset bin of image I_j and accumlate the *idf* weighting in the offset bin.

$$N_{I_j,\Delta(x_w,x_{w'}),\Delta(y_w,y_{w'})} += 1 \tag{11}$$

$$D_{I_j,\Delta(x_w,x_{w'}),\Delta(y_w,y_{w'})} += idf^2(w) \tag{12}$$

where, N_{I_j} and D_{I_j} are the co-occurring word number matrix and *idf* summation matrix for image I_j.

(3) For each image I_j , traverse each bin m, calculate the scores as follows

$$S_{I_j} = \sum_{N_{I_j,m} \geq k} \binom{m-1}{k-1} \cdot D_{I_j,m} \tag{13}$$

$$S'_{I_j} = \sum_{N_{I_j,m} \geq k} \alpha \cdot D_{I_j,m} \tag{14}$$

where, S_{I_j} is corresponding to forumlar (6), S'_{I_j} is corresponding to the last part in formular (7) and (9).

(4) Suppose we pre-calculated the preliminary L2-norm $\| V^{*k}(I) \|$ and $\| V^{*k}(I_j) \|$, obtain the final score $\hat{S}_{I_j}$ by normalizing S_{I_j} as follows.

$$\hat{S}_{I_j} = \frac{S_{I_j}}{(\| V^{*k}(I) \|^2 + S_{I_j} - S'_{I_j})^{1/2} (\| V^{*k}(I_j) \|^2 + S_{I_j} - S'_{I_j})^{1/2}} \tag{15}$$

For $\| V^{*k}(I_j) \|$, it can be calculated usig similar steps. The difference is that in step (3), formular (13) changes to:

$$S_{I_j} = \sum_{k \leq N_{I_j,m} < n_{I_j}} \binom{m-1}{k-1} \cdot D_{I_j,m} + \sum_{n_{I_j} \leq N_{I_j,m}} \alpha \cdot D_{I_j,m} \tag{16}$$

Then,

$$\| V^{*k}(I_j) \| = \sqrt{S_{I_j}} \tag{17}$$

5 Experiments

5.1 Datasets and Evaluation Measure

Oxford 5K dataset is first introduced in [5] and has become a widely used evaluation benchmark. It contains 5062 images with more than 16M features. It also provides 55 test queries of 11 different Oxford landmarks with their ground thruth retrieval results.

MIR Flickr 1M dataset is provided by the ACM MIR Committee [11]. It contains roughly 1000,000 images retrieved from Flickr. Similar to [1,2,5], we add this dataset as distractors to the Oxford 5K dataset to test the scalability of our approach (this dataset is similar to that used in [5], the resolution of images for both datasets is roughly 500×333).

As in [1,2,5], we use the mean average precision (mAP) to evaluate the performance of all experiments.

5.2 Experimental Setting and Baseline

In order to be comparable with other methods, we use source descriptors (SIFT[9] on hessian affine regions[10]) and 1M vocabulary provided in [5], and the same AKM[5] method to train the other size vocabularies (50K, 100K, 250K, 500K). For MIR Flickr 1M dataset, we draw the same type descriptors using the tool available in [15] and the fastANN[5] to assign the visual word IDs. We use the same inverted file structure introduced in [1] to facilitate the searching and the same parameter setting as in [1].

We mainly consider BoV and GVP as our baseline. We also compare favorably with BOV+RANSAC and SBoF cited their reported results. We implemented BoV, GVP according to [1,5] respectively, we don't directly cite their reported results, because the trained vocabularies are different even they are got using the same tool.

5.3 Experimental Results

Firstly, we examine the effect of using different α as spatial weighting in formular (7), (8), (9),(10),(14) and (16). We verify on all the size of vocabularies. Table 1 shows the mAP scores using different α .When $\alpha = 0.8$, the mAP score on 500K and 1M vocabulary are larger than other coresponding values. When $\alpha > 0.8$, the mAP scores on all size vocabularies decrease as α increases. Therefore we set $\alpha = 0.8$ for the following experiments.

The value of best α is pretty small (much smaller than the visual word number in an image), this is quite reasonable because it means for most of visual words in an image it's no necessary to consider the spatial weigthing for them, and the reason it's smaller than 1 may due to the spatial weightings for other visual words have already been prone to be bigger than the ideal values.

Table 1. The effect of parameter changes under all size vocabularies on Oxford 5K dataset

α	50K	100K	250K	500K	1M
0.5	0.571	0.595	0.634	0.643	0.662
0.6	0.571	0.595	0.634	0.643	0.663
0.7	0.571	0.596	0.635	0.644	0.663
0.8	0.571	0.596	0.635	**0.645**	**0.663**
0.9	0.571	0.597	0.635	0.644	0.662
1	**0.571**	**0.597**	**0.636**	0.644	0.662
2	0.570	0.597	0.635	0.644	0.661
3	0.570	0.597	0.635	0.644	0.660
4	0.569	0.596	0.634	0.643	0.659
5	0.569	0.595	0.633	0.642	0.659

Secondly, we compare our approach with other methods under different size vocabularies on Oxford 5K dataset. Table 2 is the performance of different methods (The performance of BoV+RANSAC and SboF are cited from [2,5] respectively).

Table 2. Comparison of the performance of PSW-BoV with other methods using different size vocabularies on Oxford 5K

Vocab.	BoV	BoV+ RANSAC	SboF	GVP	PSW-BoV
50K	0.486	0.569	0.523	0.551	**0.571**
100K	0.529	0.595	0.571	0.585	**0.596**
250K	0.574	0.633	^	0.627	**0.635**
500K	0.604	0.643	0.644	0.636	**0.645**
1M	0.617	0.645	0.651	0.654	**0.663**

Table 2 shows that our approach can outperform other methods on all size of vocabularies. Our approach significantly outperforms BoV method, more significant improvement is made on smaller vocabulary, because the visual words are more ambiguous. Compared with GVP, our approach further improves the mAP due to alleviate the sensitivity to distracting images with small number of visual words.

Thirdly, we verify our anlysis that GVP method is sensitive to distracting images with small number of visual words. We construct two distracting image datasets 65K_small and 65K_large from MIR Flickr 1M dataset. Where, 65K_small is composed by 65090 images where each image has less than 400 visual words; 65K_large is composed by 65090 images which are randomly selected from the images that have more than 1500 visual words. All experiments here are conducted on 1M vocabulary.The results are shown in Fig. 3.

The result in Fig. 3 shows that the difference of BoV on the two datasets is roughly 5.8%, the difference of GVP is roughly 6.3%, and the difference of PSW-BoV is roughly 0.6%. This promises the analysis that GVP is sensitive to the images with small number of visual words, while PSW-BOV alleviates this effect quite well.

#Images	BoV	GVP	PSW-BoV
65K_small+5K	0.534	0.556	0.638
65K_large+5K	0.592	0.619	0.644

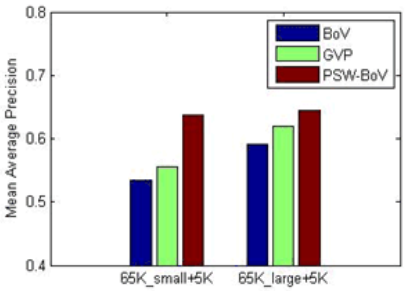

Fig. 3. The mAP scores on Oxford 5K+65K _small and Oxford 5K+65K_large datasets for BoV, GVP and PSW-BoV

Fourthly, we examine the performance for PSW-BoV on different size large scale datasets (100K, 200K, 500K, 1M, where the small datasets are randomly constructed from 1M dataset). The results are shown in Fig. 4.

#Images	BoV	GVP	PSW-BoV
100K+5K	0.561	0.602	0.637
200K+5K	0.540	0.586	0.628
500K+5K	0.509	0.554	0.607
1M+5K	0.475	0.523	0.586

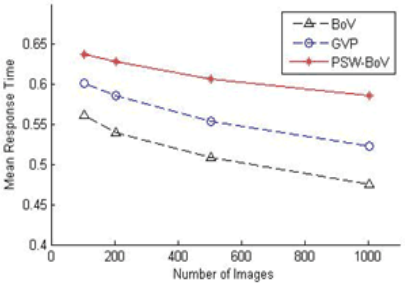

Fig. 4. The mAP scores on Oxford 5K+100K, Oxford 5K+ 200K, Oxford 5K+500K, Oxford 5K+1M datasets for BoV, GVP and PSW-BoV

Fig. 4 shows that PSW-BoV has a good scalability that can consistently improve the accuracy on different number of distracting images. The PSW-BoV method outperforms the BoV method roughly by 11.1%, GVP roughly by 6.3% on 1M dataset.

Finally, we report the efficiency of our approach. As PSW-BoV is improved from GVP, and uses different spatial weighting scheme, so the time efficency of PSW-BoV is smilar to GVP. In our experiments, a typical query on 200K+5k dataset consumes roughly 0.5s (CPU time in searching step, our CPU is 3.2GHz, main memory is 4G).

In our experiments, we do not directly compare with the existing spatial weighting methods introduced in [12,13,14]. According to their descriptions, they either need traning step or not suitable for large scale datasets. Our approach does not need training to get spatial weighting and can be facilated by inverted file for large datasets.

6 Conclusions

We first introduced a universal spatial weighting framework for BoV model. Then through analyzing GVP method using this framework, we reveal that GVP is sensitive

to images with small number of visual words due to its implicit assumption of the independence of co-occurring GVPs. Finally to alleviate the sensitive effect of GVP, we proposed a practical spatial weighting for BoV (PSW-BoV) to encode more appropriate spatial information by considering the dependence influence while keep the efficiency.Experiments on Oxford 5K and MIR Flickr 1M datasets show that PSW-BoV can allevaite the sensitive effect of GVP to a large extent and further improve the general retrieval accuracy.

Acknowledgements. This wok is supported by National Key Technology R&D Program of China under Grant 2011BAH16B01, 2011BAH16B02 and the Cloud Computing Demonstration Project of National Development and Reform Commission.

References

1. Zhang, Z., Jia, Z., Chen, T.: Image Retrieval with Geometry-Preserving Visual Phrases. In: CVPR 2011, pp. 809–816. IEEE Computer Society, Colorado Spings (2011)
2. Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L.: Spatial Bag-of-Features. In: CVPR 2010, pp. 3352–3359. IEEE Computer Society, San Francisco (2010)
3. Sivic, J., Zisserman, A.: Video google: A Text Retrieval Approach to Object Matching in Videos. In: ICCV 2003, pp. 1470–1477. IEEE Computer Society, Nice (2003)
4. Nister, D., Stewenius, H.: Scalable Recognition with A Vocabulary Tree. In: CVPR 2006, pp. 2161–2168. IEEE Computer Society, New York (2006)
5. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object Retrieval with Large Vocabularies and Fast Spatial Matching. In: CVPR 2007, pp. 1–8. IEEE Computer Society, Minneapolis (2007)
6. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In: CVPR 2008, pp. 1–8. IEEE Computer Society, Anchorage (2008)
7. Jegou, H., Douze, M., Schmid, C.: Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: CVPR 2006, pp. 2169–2178. IEEE Computer Society, New York (2006)
9. Mikolajczyk, K., Schmid, C.: Scale & Affine Invariant Interest Point Detectors. International Journal of Computer Vision 60, 63–86 (2004)
10. Lowe, D.: Distinctive Image Features From Scale-Invariant Interest Point Detectors. International Journal of Computer Vision 60, 91–110 (2004)
11. Huiskes, M., Thomee, B., Lew, M.: New Trends and Ideas in Visual Concept Detection. In: ACM MIR 2010, pp. 527–536. ACM, Pennsylvania (2010)
12. Marszalek, M., Schmid, C.: Spatial Weighting for Bag-of-Features. In: CVPR 2006, pp. 2118–2125. IEEE Computer Society, New York (2006)
13. Chen, X., Hu, X., Shen, X.: Spatial Weighting for Bag-of-Visual-Words and Its Application in Content-Based Image Retrieval. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 867–874. Springer, Heidelberg (2009)
14. Martinet, J., Urruty, T., Djeraba, C.: A new spatial weighting scheme for bag-of-visual-words. In: CBMI 2010, Grenoble, France, pp. 1–6 (2010)
15. `http://www.robots.ox.ac.uk/~vgg/software/`