Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening

Hervé Jégou, Ondrej Chum

Negative evidences and co-occurences in image retrieval: the benefit of PCA and whitening

Hervé Jégou[1] Ondřej Chum[2]

[1] INRIA Rennes
[2]CMP, Department of Cybernetics, Faculty of EE, CTU in Prague

Abstract. The paper addresses large scale image retrieval with short vector representations. We study dimensionality reduction by Principal Component Analysis (PCA) and propose improvements to its different phases. We show and explicitly exploit relations between i) mean subtraction and the negative evidence, i.e., a visual word that is mutually missing in two descriptions being compared, and ii) the axis de-correlation and the co-occurrences phenomenon. Finally, we propose an effective way to alleviate the quantization artifacts through a joint dimensionality reduction of multiple vocabularies. The proposed techniques are simple, yet significantly and consistently improve over the state of the art on compact image representations. Complementary experiments in image classification show that the methods are generally applicable.

1 Introduction

This paper mainly addresses the problem of large-scale image search and object recognition, as considered by many papers in the literature [1–4]. More precisely, the task consists of finding images in a large image database that most closely resemble a query image based on their visual similarity. A majority of the papers rely on the bag-of-words (BOW) representation [1, 5, 2, 3] or its derivatives, e.g., [4]. These approaches are limited to search in a few million images on a single machine due to computational or memory constraints. In this paper, we will mainly focus on more scalable approaches in the spirit of recent work on compact images representations [6–8], where the image description is a short vector, which is subsequently encoded in compact codes using binarization [6, 9] or product quantization techniques [10]. The best performing methods in this context are those that produce the vector representing an image from local features [7, 8] such as the Fisher Vectors [11, 12, 7] or its non probabilistic version, namely the VLAD descriptor [8]. In contrast to global description techniques computed from the pixels [13, 6] in a more direct manner, these representations inherit, to some extent, the invariance properties (change in viewpoints, cropping, etc) of the local descriptors from which they are computed.

Methods generating a short code image representation commonly exploit the PCA [14] to perform the dimensionality reduction. It was observed [8] that the performance of BOW is even improved by PCA reduction. In the paper, we study this phenomenon. The PCA can be seen as a two step process (1) centering the

data, and (2) selecting a de-correlated (orthogonal) basis of a subspace minimizing the dimensionality reduction error. We show that each of the steps has a positive impact on the retrieval, and we provide interpretation of such a behavior. Based on the analysis, we propose simple yet effective techniques to further improve the quality of BOW and VLAD representations. First, we consider the role of *negative evidence*: given two BOW vectors, a visual word jointly missing in both vectors is information that should receive more importance in the similarity measurement. We show relation of the negative evidence to the centering of BOW vectors (mean subtraction). Secondly, both BOW and VLAD representations are further improved by exploiting the de-correlation of the descriptor entries. Two complementary approaches are proposed 1) whitening the vector space, thereby addressing the problem of co-occurrences; and 2) by considering multiple vocabularies with a joint dimensionality reduction. Multiple vocabularies have been considered by prior art, e.g., in the hierarchical k-means [2] or in the rank aggregation technique of [15]. In contrast to those, our method increases the search accuracy for *a fixed size* of the vector describing the image. When querying the indexing structure, the memory and computational complexities are the same as when considering a unique vocabulary.

Albeit simple, the proposed techniques consistently and significantly improve the state-of-the-art image search based on short vectors, as demonstrated by our results on four popular benchmarks. Finally, we will briefly show with experiments on the PASCAL VOC'07 benchmark that the better representation for retrieval also translates to better classification results: Our short vectors obtained from BOW and combined with a linear classifier significantly outperform a soft BOW combined with a Chi-square kernel.

The paper is organized as follows: After introducing the context in Section 2, Section 3 shows the role of co-missing visual words and Section 4 exploits whitening to address the issue arising with co-occurrence over-counting. Section 5 extends it to multiple vocabularies and compares with the state of the art.

2 Background and datasets

2.1 Image description Framework

Bag-of-words. As a baseline, we first consider the regular bag-of-words representation, as proposed by Sivic and Zisserman [1]. This representation extracts a global description vector from an image using the following procedure.

1. Covariant regions of interest are detected [16, 17] in the image and described by a local d-dimensional descriptor. We used the Hessian-Affine detector jointly with the SIFT descriptor [18].
2. The resulting descriptors are quantized using a so-called "visual vocabulary", which is learned using k-means algorithm, producing "visual words".
3. The histogram of occurrences of visual words (of size $D = k$) is computed and weighted using inverse document frequency (idf) terms.

4. The resulting vector is subsequently normalized. As proposed in [1], we adopt the L2 normalization.

VLAD. The vector of locally aggregated descriptors [8] is a simplification of the Fisher vector [11]. This representation departs from BOW only in the Step 3: instead of producing the histogram of occurrences, VLAD accumulates, in the output vector of size $D = k \times d$, the difference between the descriptors and their respective centroids.

Power-law normalization. Both the VLAD and Fisher vector representations are improved [19] by using the so-called *power-law* normalization [7]. This simple method post-processes the output image vector $v = (v_1, ...v_D)$ as $v_i := |v_i|^\beta \times \text{sign}(v_i)$, with $0 \leq \beta < 1$ a fixed constant. The updated vector v is L2-normalized in turn. The impact of this post-processing is argued [19] to reduce the impact of multiple matches and visual bursts [20]. This variant will be considered for $\beta = 0.5$ in the following, denoted by SSR (signed square rooting).

2.2 Efficient PCA

The BOW and VLAD vectors are high dimensional. For instance, typical values of D for BOW ranges from one thousand to one million components, while VLAD vectors are $k \times d$-dimensional, d being the dimensionality of the local descriptor. This means $D = 65,536$ for the typical parameters $d = 128$ and $k = 512$. It is therefore not efficient or even feasible to perform the PCA using the covariance matrix method. However, we only need the first D' first eigenvectors and eigenvalues in Equation 5. By limiting the learning set $\boldsymbol{Y}$ to a reasonable number of vectors (we used the learning image sets introduced in Section 2), one can use the dual gram method (see, e.g., Paragraph 12.1.4 in [14]) to learn the matrix $\boldsymbol{P}$ and eigenvalues λ_1 to $\lambda_{D'}$. This amounts to computing the $n \times n$ gram matrix $\boldsymbol{Y}^\top \boldsymbol{Y}$ instead of the $D \times D$ covariance matrix $\boldsymbol{C}$, and to exploiting the analytical relationship between the eigen-decomposition of these two matrices. The eigenvalue decomposition is performed using the Arnoldi algorithm, which computes the D' desired eigenvectors, i.e., those associated with the largest eigenvalues, using an iterative procedure.

2.3 Datasets

The interest of the proposed techniques is evaluated on a number of popular datasets widely used in the literature.

Oxford5k [21] and Paris6k [22]: These datasets are collections of images from Flickr. The 55 queries correspond to 11 distinct buildings, given by bounding boxes in 55 images from the set. The task is to retrieve all corresponding buildings. The performance is measured by mean average precision (mAP), as defined in [21].

Holidays (+Flickr1M): This dataset contains personal Holiday photos provided by INRIA [4]. The dataset itself contains 1491 images. A subset of 500

images serves as queries. Each query is compared to the other 1490 images in a leave-one-out fashion. To evaluate the performance on a large scale, a distractor dataset of 1 million images downloaded from Flickr is also provided. As for Oxford5k, the performance is measured by mAP.

University of Kentucky benchmark (UKB). This image set contains 10200 images, corresponding to 2550 distinct objects and scenes (4 images per group). Each image is compared to all the others. The usual performance score is the mean number of images ranked in the first 4 positions.

For the Oxford5k and Paris datasets, we have used to the detector and descriptor used in [23], while the descriptors available online have been used for the other datasets.

Dataset for learning stages: We use an independent dataset (no intersection with the test set) to learn the visual vocabularies and for the other learning stages involved in our technique. When evaluating on Holidays, Holidays+Flickr1M and UKB, the independent dataset consists of 10000 images from Flickr. Paris6k is used to learn the meta-data associated with the evaluation on Oxford5k. Note that the *idf* terms do not involve any learning stage and are applied on-the-fly, based on the indexed dataset statistics.

3 Exploiting evidences from co-missing words

Being produced as a weighted histogram of occurrences of visual words, the regular BOW representation contains only non-negative values. Let consider the cosine measure for similarity $s(u, v)$ between BOW vectors u and v, i.e.,

$$s(u, v) = \frac{1}{\|u\| \, . \, \|v\|} \sum_{i}^{k} u_i \, v_i. \tag{1}$$

If $u_i = 0$, the individual contribution of the visual word with index i is the same if v_i is equal or greater than 0. The difference between these cases is only taken into account by the normalization factor $||v||$. This under-estimates the importance of jointly zero components, which give some limited yet important evidence on visual similarity.

We, therefore, propose a simple way to better take into account this case in BOW vectors. Instead of measuring the angle between points u and v from the origin, we consider an angle between those two points measured at different point m. A good choice for m is a fraction of the mean bag-of-words vector $m = \alpha \cdot \bar{v}$. The novel cosine similarity is computed by Equation (1) on transformed vectors by

$$v := v - \alpha . \bar{v}. \tag{2}$$

The value $\alpha = 1$ corresponds to the case where the mean of the vector (produced from a learning set) is subtracted. Applying such a transformation, the cosine similarity gives a positive contribution for a particular visual word if it is absent

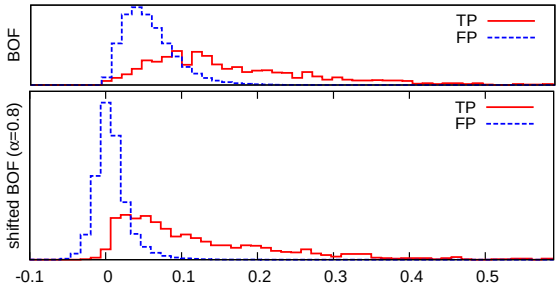

Fig. 1. Empirical distribution (Holidays dataset) of the similarities for true (TP) and false positives (FP) before (regular BOW) and after the proposed correction (shifted BOW). Observe the better separation of true and false positives, which are centered on zero with the shifted BOW.

(more precisely, if it appeared less than on average) in both the compared images. As a result, the similarity between bag-of-words is improved, as depicted in Figure 1.

Figure 2 shows the impact of this correction as a function of α when measuring the performance on the Holidays and Oxford5K benchmarks. As one can see, the proposed update gives some significant improvement, in particular for smaller vocabulary sizes, and this for a negligible computing and memory cost, as explained below.

Integration within the inverted file system. The inverted file structure allows for efficient evaluation of the cosine distance for sparse vectors by evaluating only non-negative elements of the product in Equation (1). A naive subtraction of a non-sparse vector m from all sparse vectors in the database has a severe negative effect on both the efficiency of the retrieval and the memory footprint. It is however possible to compute the new similarity measure using the same inverted file structure as for evaluating (1). The cosine distance after the subtraction is expressed as

$$s(u,v) = \frac{1}{\|u-m\| \,.\, \|v-m\|} \sum_{i}^{k} (u_i - m_i)(v_i - m_i). \tag{3}$$

For each document v in a database, the normalization factor $||v-m||$ is query-independent and therefore pre-computed. Re-writing the similarity as

$$\sum_{i}^{k} (u_i - m_i)(v_i - m_i) = u^\top v - v^\top m - u^\top m + \|m\|^2 , \tag{4}$$

where the dot product $u^\top v$ is efficiently computed using the original inverted file structure as in (1). The term $v^\top m$ is independent of the query. It is therefore computed and stored when adding a BOW vector to the index. The term $u^\top m$ only depends on the query (computed once per query) and $\|m\|^2$ is a constant. Therefore, although the rest of this paper mainly considers short vector

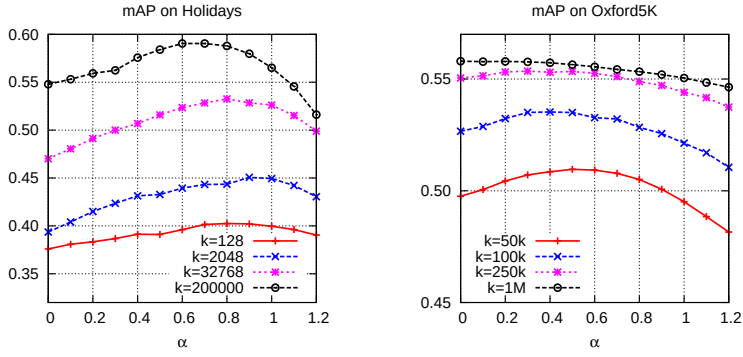

Fig. 2. BOW for Holidays and Oxford5K: mAP performance as a function of α, for different vocabulary sizes k. The optimum values outperform the state-of-the-art for pure-BOW approaches.

representations, we must mention that our shifting approach is effective when considering a regular inverted file implementation [1–3], at negligible memory and computational costs.

Discussion. From Fig. 2 it can be seen that the impact of the negative evidence is higher for small vocabularies and diminishes for large vocabularies. This has an intuitive explanation. For large (fine) vocabularies the co-missing visual words become more common event that carries less evidence, while the presence of the same visual words provides a strong evidence. For smaller vocabularies, the relative weights of the positive and negative evidence is changing. This can be observed in Fig. 2 (especially the plot for Oxford 5K dataset), where the optimal value of α (the higher value the higher weight on the negative evidence) is shifting to the left with increasing size of the vocabulary.

4 Co-occurrence over-counting: the benefit of whitening

An efficient way to obtain a shorter image vector representation consists of applying principal component analysis (PCA) dimensionality reduction directly on the BOW (or VLAD) vector [8]. This, first, performs the implicit centering of the data, therefore taking into account the co-missing visual words and thereby improving the similarity measurement. Second, by concentrating the vector energy of the first components, the similarity between reduced vectors provides a reasonable approximation of the similarity before the projection. We adopt this method to produce short vectors from BOW and VLAD representations.

However, it is worth noticing that an important phenomenon is ignored by such a blind dimensionality reduction, namely the problem of co-occurrences. Chum et al. [24] notice that co-occurrences lead to over-count some visual patterns when comparing two image vector representations. The detector may also introduce some artificial visual word co-occurrences, for instance when an image

region is described multiple times for different orientations [25], producing two different but strongly co-occurring descriptors.

Let consider the learning set of image global descriptors (BOW or VLAD), centered according to the mean, and represented by a matrix $\boldsymbol{Y} = [Y_1|\dots|Y_n]$. The D-dimensional covariance matrix is estimated as $\boldsymbol{C} = \boldsymbol{Y} \times \boldsymbol{Y}^\top$. The visual word co-occurrences are captured in this matrix, generating strong responses out of the diagonal and favoring the emergence of an eigenvector associated with a large eigenvalue comprising those values. An efficient way to limit the impact of co-occurrences therefore consists in whitening the data, as done in independent component analysis [26], and as implicitly performed by the Mahalanobis distance.

In our case, this whitening operation is performed jointly with the dimensionality reduction from D to D' components: A given image descriptor X (BOW or VLAD) is first PCA-projected and truncated, and subsequently whitened and re-normalized to a new vector $\hat{X}$ that is our short vector image representation. It is therefore obtained as follows:

$$\hat{X} = \frac{\mathrm{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_{D'}^{-\frac{1}{2}})\boldsymbol{P}^\top X}{\left\|\mathrm{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_{D'}^{-\frac{1}{2}})\boldsymbol{P}^\top X\right\|}, \tag{5}$$

where the $D \times D'$ matrix $\boldsymbol{P}$ is formed by the largest eigenvectors of the covariance matrix $\boldsymbol{C}$, and where λ_i is the eigenvalue associated with the i^{th} largest eigenvector. Comparing two vectors obtained after this dimensionality reduction with the Euclidean distance is therefore similar to using a Mahalanobis distance, but differs from it in that the vectors are truncated and re-normalized. The comparison is efficiently performed by comparing the reduced vectors using the Cosine similarity. The re-normalization step turns out to be critical for a better comparison metric (up to 10% of mAP of difference on the Holidays dataset).

Impact on performance. For the sake of consistency, the vector dimensionality is reduced to *D'=128 dimensions* in all the experiments presented in this paper. Figure 3 gives the impact of the dimensionality reduction, of the SSR component-wise normalization, and of our whitening technique, which is shown to provide a large improvement over the BOW baseline.

Remarks:

- The *idf* weighting terms can not be longer applied on-the-fly with the dimensionality reduction, and are therefore be learned on the independent dataset.
- As a side effect of the dimensionality reduction, two ambiguous visual words i and j generate a higher value than for other tuples in the covariance matrix, which favors the projection of these visual words to the same component in the projected vector. This phenomenon can be observed when reconstructing the BOW vector from its PCA projection: The component of the other visual word is "hallucinated".

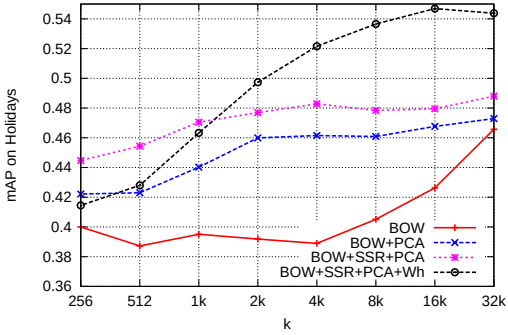

Fig. 3. Impact of the different steps on search accuracy (Holidays) for BOW vectors, as a function of vocabulary size k. The proposed whitening step is denoted by Wh.

– For large values of D', the whitening stage negatively impacts the performance by magnifying the noise on the low-energy components. This issue is addressed by using a robust PCA/whitening method.

5 Joint de-correlation of multiple vocabularies

It is well known that the quantization effect have significant impact on the retrieval quality. Different approaches were suggested to overcome the problem, ranging from hierarchical quantization [2], soft assignment [22], to Hamming embedding [4]. We show that the quantization effects are alleviated by multiple quantization. However, straightforward concatenation of the BOW representations not only linearly increases the memory requirements, but improve only marginally the retrieval results, see Fig. 4 or [15]. The different BOW representations are strongly correlated. We show that the PCA removes the correlation, while preserving the additional information from the different quantizations. Results outperforming the state of the art for short image representation are achieved.

5.1 Related work on multiple vocabularies

Some prior art has proposed to use multiple vocabularies to improve the quality of the search, at the cost of reduced efficiency and increased memory usage. For instance, a common and simple strategy consists in simply considering the concatenating of the different BOW vectors as the image representation, as done by Nister et al. [2], who consider a hierarchical quantization method where the intermediate nodes correspond to smaller vocabularies. A late fusion technique based on rank aggregation was also proposed [15], but several inverted files have to be stored and queried in parallel. In addition, those techniques do not take into account the relationship between the vocabularies: their output is processed independently without considering the dependencies between the quantizers.

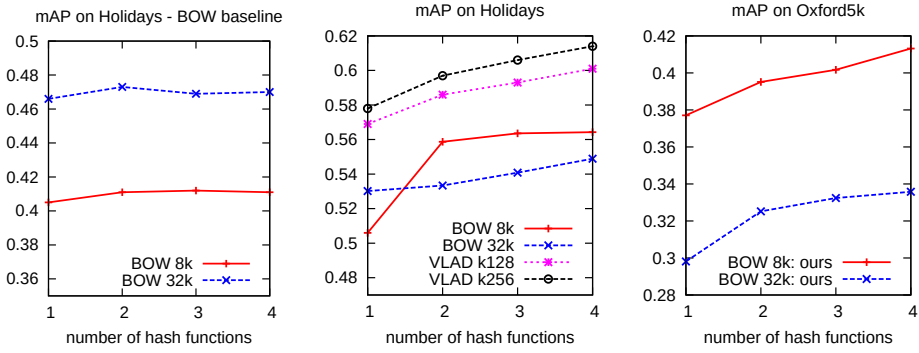

Fig. 4. The leftmost plot gives the BOW baseline when concatenating different numbers of BOW into a single vector (of increasing dimensionality), which provides only a small improvement, while linearly increasing the memory requirements. The middle and right plots show the search accuracy with the proposed joint reduction of the vocabularies to a *fixed* vector size of 128 components. Observe that the improvement brought by the use of several hash functions is comparatively better with our technique than for the concatenation of BOWs, thanks to the joint dimensionality reduction.

A more popular alternative consists of using multiple [15] or soft [22] assignment. This also increases the query time when the search is performed using an inverted file[1]. Since we are more particularly interested by image search in larger databases, taking care of the memory size of the representation is critical in order to keep the indexing structure in memory.

5.2 Joint reduction of multiple vocabularies

The key points of our multiple vocabulary method, which departs from those proposed in the literature, is that it performs *the joint dimensionality reduction* of the BOW vectors produced for the different vocabularies, and apply in turn the whitening technique proposed the previous section to correct the artifacts resulting from the use of multiple vocabularies. Indeed, the different vocabularies are redundant: two descriptors assigned to the same visual word for one vocabulary have higher probability to be assigned to the same visual word for another vocabulary, leading to co-occurrences as those mentioned in Subsection 4.

Another difference with [2] and [27] is that we consider overlapping quantizers. This, jointly with the dimensionality reduction, better addresses the problem of quantization artifacts than these approaches, or of multiple or soft quantization techniques [15, 22], as demonstrated later (Table 2) by comparing multiple vocabulary with VLAD (hard assignment) with Fisher Kernel (soft assignment based on a Gaussian mixture model) with a single vocabulary.

We therefore propose the following reduction for multiple vocabularies:

[1] The memory requirement is similarly increased if this assignment is performed in a symmetrical manner on query and database sides.

Vocabularies	weighting	mAP (%)	complexity
2 × 32k	N/A	53.3	65,536
4 × 16k	N/A	55.8	65,536
8 × 8k	N/A	56.7	65,536
32k+16k+...+128	1	58.5	65,408
32k+16k+...+128	k	54.3	65,408
32k+16k+...+128	log k	58.8	65,408
32k+16k+...+128	*idf*	57.8	65,408

Table 1. Performance of vocabularies of identical and multiple sizes (Holidays). Complexity: number of vectors comparison per local descriptor when constructing the aggregated representation formed by all BOW vectors. Weighting: see text for details.

1. The BOW or VLAD vectors are produced independently, using SSR component-wise normalization (see Section 2). The *idf* term is ignored, as it occurs that its influence is limited with multiple vocabularies. The SSR component-wise normalization is applied and the concatenated vector is normalized.
2. The different vectors are jointly reduced and whitened according to the guidelines of Subsection 4.

Figure 4 shows that multiple vocabularies, used jointly with our dimensionality reduction technique, provides an significant improvement for both BOW and VLAD representations, and this for *a fixed output vector size D'*. Larger vocabularies for BOW are not necessarily better than smaller ones: although k=32k provides better results on Holidays with a single vocabulary, we observe the opposite outcome for multiple vocabularies.

5.3 Merging vocabularies of different sizes

The goal of this subsection is to address the trade-off between absolute search quality (for a given vector size), and the quantization cost, and to provide a comparison with similar methods of the literature. The following analysis is mainly intended for BOW, since VLAD typically uses vocabularies of much smaller sizes (*e.g.*, k=256). Although the quantization cost does not depend on the dataset size, it delays the query (jointly with the extraction of the descriptor from the image), which might be critical for some applications. For reference, quantizing 2000 local descriptors of a query image, for 4 vocabularies comprising $k = 8,192$ centroids each, takes 0.45s on 12 cores, using an efficient multi-threaded implementation of exhaustive search (exact). The timings are in this case proportional to k.

To reduce the quantization cost, we consider vocabularies of different sizes, in the spirit of the hierarchical k-means method [2] and of the pyramid match kernel [27]. Vocabularies of different sizes have a different importance, and therefore their respective contribution should be adapted. We compare four different approaches to adjust the contribution of the vocabularies:

1. The same unit weight is applied for all vocabularies.

Method	Vocabulary size(s)	Holidays	Oxford5k	UKB
GIST [13]	N/A	36.5	-	1.64
BOW [1]	k=20k	45.2	19.4	2.95
Improved Fisher [7]	k=64	56.5	30.1	3.33
VLAD [8]	k=64	51.0	-	3.15
VLAD+SSR [19]	k=64	55.7	28.7	3.35
Ours/BOW	4×(8k)	56.7	**41.3**	3.19
Ours/BOW	2×(128+256+...32k)	**60.0**	-	3.28
Ours/VLAD	4×(256)	**61.4**	-	**3.36**

Table 2. Comparison against the state of the art on *short* vector image representations. We consider 128-D vectors for all methods (reduced by PCA, including for BOW). Most reference results are extracted from a paper [19] on compact representations.

2. Similar to [27], the weight of the vocabulary is proportional to its size (to the number of bins).
3. We consider weights proportional to the logarithm of the vocabulary size.
4. Similar to [2], the weights are determined by *idf* after all vocabularies are concatenated.

In the first three approaches (referred to as"1", "k" and "log k"), each descriptor for each vocabulary is first transformed by SSR and L2-normalized, then multiplied by the vocabulary weight. The descriptors of different vocabularies are concatenated and finally, the concatenated vectors are L2-normalized. In the fourth weighing scheme, the *idf* weighting is applied to the vectors after the concatenation of the vocabularies, as proposed in [2].

Table 1 shows the results with when considering multiple vocabularies of fixed and different sizes, and compares the different weighting techniques. For a fixed quantization cost, in this experiment the best choice is to use vocabularies of different sizes and our log weighting technique. Note however that the improvement of this latter is only 1% compared with equal weights for all sizes.

5.4 Comparison with the state-of-the-art

Table 2 compares our method to the state-of-the-art on short vector representations. The results obtained with the proposed short vector construction are consistently better than reference results. The improvement brought our method is higher when applied to BOW than to VLAD. Compared with BOW of 20k centroids reduced to 128D [19], our method, when applied to BOW with 4 vocabularies of size 8k, increases the mAP of +14.8% on Holidays, +21.9% on Oxford5k. The UKB score is 3.19/4 (BOW reference: 2.95). As a result, the BOW-based representation is competitive, when not better, than the best results reported with PCA-reduced VLAD and Fisher representations, where in [19] these representations are shown to significantly outperform BOW. By applying our method on the VLAD representation, we still obtain an improvement of +5.7% over the state of the art. The improvement is not significant on UKB.

Method	mAP on Holidays
VLAD [10]	46.0
Fisher vector [19]	50.6
Ours/BOW - 4×(k=8k)	49.8
Ours/VLAD - 4×(k=256)	**53.1**

Table 3. Comparison against the state of the art on image representations with short codes on Holidays. The code size is fixed to 16 bytes for all methods.

5.5 Encoding our short vectors with compact codes

The better results obtained with shorter vectors lead to better results when further coding the vectors using a compressed-domain approximate nearest neighbors search technique [10], as shown by Table 3. With codes of 16 bytes, we increase the mAP of the BOW baseline by +3.6% of mAP on Holidays. with VLAD (4×k=256) on Holidays, we outperform the state-of-the-art coded Fisher Vector [19] by +2.5% of mAP.

5.6 Large scale experiments

We have evaluated our approach on one million images, by merging the Holidays dataset with the Flickr1M distractor set, as done in [4, 8]. Our method is compared (from curves in [19]) with the BOW representation (k=200k).

The results are presented in Figure 5. Our approach significantly outperforms the baseline, and this by using an image representation which is a 128 dimensional vector only, i.e., using significantly less memory than sparse BOW, which typically requires 4 bytes per encoded local descriptors. The efficiency is also much better than BOW. With an efficient implementation of exhaustive search, querying the whole 500 query images from Holidays takes 3.08 seconds using 12 cores of a 3 Ghz machine, which corresponds to 6 ms per query. This is about two orders of magnitude faster than the timings reported for BOW [4].

5.7 Improving BOW for classification

Although the primary goal of this paper is to consider image retrieval on a large scale with short vectors, we report some preliminary results showing the interest of our method in a context of classification with very short vectors and efficient linear classifiers. For this purpose, we improve the BOW baseline with our approach (SSR, whitening joint de-correlation of 4 vocabularies) and compare to methods combined with a linear classifier. On Pascal VOC'07 [28], by considering the same protocol as the one proposed in [29], our technique is significantly better than the corresponding BOW: we obtain mAP=46.9% with D'=256 dimensions, instead of 41.4% for BOW with 4k dimensions. The result is approximately the same of the one of spatial pyramid matching (SPM, [30]) with a linear classifier, but with a vector which is 100 times shorter.

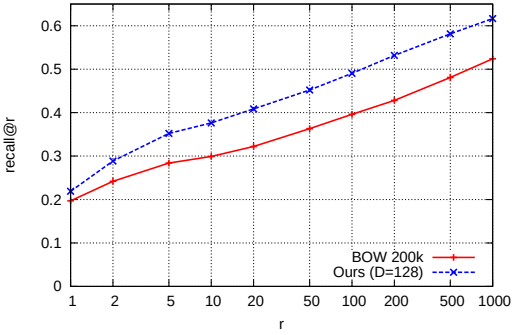

Fig. 5. Holidays+1M distractors: proportion of true positives returned in the first r ranks (recall@r), for BOW with 200k (from [19]) and our method with a 128-D vector.

6 Conclusion

Different techniques to improve dimensionality reduction by PCA for large scale image retrieval were proposed. First, a solution is proposed for giving more importance to jointly non-occurring visual words, leading to improved image search quality with bag-of-features at a negligible cost in memory and computational complexity. This approach can be also integrated into an inverted file. Then, we considered the problem of co-occurring and correlated visual words, jointly with the dimensionality reduction and the use of multiple vocabularies. This method produces short vectors (128-dimensional, *i.e.*, the size of a *single* SIFT local descriptor) yielding a high retrieval accuracy, as demonstrated by our results on popular image search benchmarks. Finally, it was shown on image classification that the methods are generally applicable.

Acknowledgments. Ondřej Chum was supported by the GACR P103/12/2310 project. Hervé Jégou was supported by the Quaero project, funded by OSEO.

References

1. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)
2. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006) 2161–2168
3. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV. (2007)
4. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. IJCV **87** (2010) 316–336
5. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop Statistical Learning in Computer Vision. (2004)

6. Torralba, A., Fergus, R., Weiss, Y.: Small codes and large databases for recognition. In: CVPR. (2008)
7. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: CVPR. (2010)
8. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010)
9. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NIPS. (2008)
10. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. Trans. PAMI **33** (2011) 117–128
11. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
12. Perronnin, F., J.Sánchez, Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010)
13. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
14. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2007)
15. Jégou, H., Schmid, C., Harzallah, H., Verbeek, J.: Accurate image search using the contextual dissimilarity measure. Trans. PAMI **32** (2010) 2–11
16. Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC. (2002) 384–393
17. Mikolajczyk, K., schmid, C.: Scale and affine invariant interest point detectors. IJCV **60** (2004) 63–86
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
19. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local descriptors into compact codes. In: Trans. PAMI. (2012)
20. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR. (2009)
21. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)
22. 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)
23. Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: CVPR. (2009)
24. Chum, O., Matas, J.: Unsupervised discovery of co-occurrence in sparse high dimensional data. In: CVPR. (2010)
25. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65** (2005) 43–72
26. Comon, P.: Independent component analysis, a new concept? Signal Processing **36** (1994)
27. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV. (2005)
28. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88** (2010) 303–338
29. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. (2011)
30. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)