Multiple Measurements and Joint Dimensionality Reduction for Large Scale Image Search with Short Vectors

Extended Version

Filip Radenović[1] Hervé Jégou[2] Ondřej Chum[1]
[1]CMP, Faculty of Electrical Engineering, Czech Technical University in Prague
[2]INRIA, Rennes
filip.radenovic@cmp.felk.cvut.cz, herve.jegou@inria.fr, chum@cmp.felk.cvut.cz

Abstract

This paper addresses the construction of a short-vector (128D) image representation for large-scale image and particular object retrieval. In particular, the method of joint dimensionality reduction of multiple vocabularies is considered. We study a variety of vocabulary generation techniques: different k-means initializations, different descriptor transformations, different measurement regions for descriptor extraction. Our extensive evaluation shows that different combinations of vocabularies, each partitioning the descriptor space in a different yet complementary manner, results in a significant performance improvement, which exceeds the state-of-the-art.

1. Introduction

LARGE-SCALE image retrieval techniques have been developing and improving greatly for more than a decade. Many of the current state-of-the-art approaches [21, 6, 11] are based on the bag-of-words (BOW) approach originally proposed by Sivic and Zisserman [27]. Another popular image representation arises from aggregating local descriptors like Fisher kernel [24] and Vector of Locally Aggregated Descriptors (VLAD) [13].

The BOW vectors are high dimensional (up to 64 million dimensions in [20]), so, due to the high memory and computational requirements, search is limited to a several million images on a single machine. There are more scalable approaches that tackle this problem by generating compact image representations [28, 24, 13], where the image is described by a short vector that can be additionally compressed into compact codes using binarization [28, 30], product quantization [12], or recently proposed additive quantization techniques [3]. In this paper we propose and experimentally evaluate simple techniques that additionally boost retrieval performance, but at the same time preserve low memory and computational costs.

Short vector image representations are often generated using the principal component analysis (PCA) [4] technique to perform the dimensionality reduction over high-dimensional vectors. Jegou and Chum [8] study the effects of PCA on BOW representations. They show that both steps of PCA procedure, i.e., centering and selection of decorrelated (orthogonal) basis minimizing the dimensionality reduction error, improve retieval performance. Centering (mean subtraction) of BOW vectors provides a boost in performance by adding a higher value to the negative evidence: given two BOW vectors, a visual word jointly missing in both vectors provides useful information for the similarity measure [8]. Additionnaly, they advocate the joint dimensionality reduction with multiple vocabularies to reduce the quantization artifacts underlying BOW and VLAD. These vocabularies are created by using different initializations for the k-means algorithm, which may produce relatively highly correlated vocabularies.

In this paper, we propose to reduce the redundancy of the joint vocabulary representation (before the joint dimensionality reduction) by varying parameters of the local feature descriptors prior to the k-means quantization. In particular, we propose: (i) different sizes of measurement regions for local description, (ii) different power-law normalizations of local feature descriptors, and (iii) different linear projections (PCA learned) to reduce the dimensionality of local descriptors. In this way, created vocabularies will be more complementary and joint dimensionality reduction of concatenated BOW vectors originating from several vocabularies will carry more information. Even though the proposed approaches are simple, we show that they provide significant boosts to retrieval performance with no memory or computational overhead at the query time.

Related work. This paper can be seen as an extension of [8], details of which are given later in Section 2.3. A number of papers report results with short descriptors obtained by PCA dimensionality reduction. In [14] and [24], aggregated descriptors (VLAD and Fisher vector respectively) are used followed by PCA to produce low dimensional image descriptors. In a paper about VLAD [2], authors propose a method for adaptation of the vocabulary built on an independent dataset (adapt) and intra-normalization (innorm) method that L_2 normalizes all VLAD components independently, which suppresses the burstiness effect [10]. In [15], a 'democratic' weighted aggregation method for burstiness supression is introduced. In this paper, we compare results of all the aforementioned methods using low dimensional descriptors $D' = 128$.

The rest of the paper is organized as follows: Section 2 gives a brief overview of several methods: bag-of-words (BOW), efficient PCA dimensionality reduction of high dimensional vectors, and baseline retrieval with multiple vocabularies. Used datasets and evaluation protocols are established in Section 3. Section 4 introduces novel methods for joint dimensionality reduction of multiple vocabularies and presents extensive experimental evaluations. Main conclusions are given in Section 5.

2. Background and baseline

This section gives a short overview of the background of bag-of-words based image retrieval and the method used in [8]. Key steps and ideas are discussed in higher detail to help understanding of the paper.

2.1. Bag-of-words (BOW) image representation

First efficient image retrieval based on BOW image representation was proposed by Sivic and Zisserman [27]. They use local descriptors extracted in an image in order to construct a high-dimensional global descriptor. This procedure follows four basic steps:

1. For each image in the dataset, regions of interest are detected [18, 17] and described by an invariant descriptor which is d-dimensional. In this work we use the multi-scale Hessian-Affine [23] and MSER [17] detectors, followed by SIFT [16] or RootSIFT [1] descriptors. The rotation of the descriptor is either determined by the detected dominant orientation [16], or by the gravity vector assumption [23]. The descriptors are extracted from different sizes of measurement regions [17], as described in detail in Section 4.

2. Descriptors extracted from the training (independent) dataset (see Section 3) are clustered into k clusters using the k-means algorithm, which creates a visual vocabulary.

3. For each image in the dataset, a histogram of occurrences of visual words is computed. Different weighting schemes can be used, the most popular is inverse document frequency (*idf*), which generates a D dimensional BOW vector ($D = k$).

4. All resulting vectors are L_2 normalized, as suggested in [27], producing final global image representations used for searching.

2.2. Efficient PCA of high dimensional vectors

In most of the cases BOW image representations have very high number of dimensions (D can take values up to 64 million [20]). In these cases the standard PCA method (reducing D to D') computing the full covariance matrix is not efficient. The dual gram method (see Paragraph 12.1.4 in [4]) can be used to learn the first D' eigenvectors and eigenvalues. Instead of computing the $D \times D$ covariance matrix $\boldsymbol{C}$, the dual gram method computes the $n \times n$ matrix $\boldsymbol{Y}^T\boldsymbol{Y}$, where $\boldsymbol{Y}$ is a set of vectors used for learning, and n is the number of vectors in the set $\boldsymbol{Y}$. Eigenvalue decomposition is performed using the Arnoldi algorithm, which iteratively computes the D' desired eigenvectors corresponding to the largest eigenvalues. This method is more efficient than the standard covariance matrix method if the number of vectors n of the training set is smaller than the number of vector dimensions D, which is usually the case in the BOW approach.

Jegou and Chum [8] analyze the effects of PCA dimensionality reduction on the BOW and VLAD vectors. They show that even though PCA successfully deals with the problem of negative evidence (higher importance of jointly missing visual words in compared BOW vectors), it ignores the problem of co-occurrences (co-occurences lead to over-count some visual patterns when comparing two image vector representation, see [5]). In order to tackle the aforementioned problem, they propose performing a whitening operation, similar to the one done in independent component analysis [7] (implicitly performed by the Mahalanobis distance), jointly with the PCA. In our experiments we will use dimensionality reduction from D to D' components, as done in [8]:

1. Every image vector $v = (v_1, \ldots, v_D)$ is post-processed using power-law normalization [24]: $v_i := |v_i|^\beta \times sign(v_i)$, with $0 \leq \beta < 1$ as a fixed constant. Vector v is L_2 normalized after processing. It has been shown [14] that this simple procedure reduces the impact of multiple matches and visual bursts [10]. In all our experiments $\beta = 0.5$, denoted as signed square rooting (SSR).

2. First D' eigenvectors of matrix $\boldsymbol{C}$ are learned using power-law normalized training vectors $\boldsymbol{Y} =$

$[Y_1|\ldots|Y_n]$, corresponding to the largest D' eigenvalues $\lambda_1, \ldots, \lambda_{D'}$.

3. Every power-law normalized image descriptor used for searching X is PCA-projected and truncated, and at the same time whitened and re-normalized to a new vector $\hat{X}$ that is the final short vector representation with dimensionality D':

$$\hat{X} = \frac{\operatorname{diag}(\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_{D'}^{-\frac{1}{2}}) \boldsymbol{P}^T X}{\left\| \operatorname{diag}(\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_{D'}^{-\frac{1}{2}}) \boldsymbol{P}^T X \right\|}, \tag{1}$$

where the $D \times D'$ matrix $\boldsymbol{P}$ is formed by the largest eigenvectors calculated in the previous step. Comparing two vectors after this dimensionality reduction with the Euclidian distance is now similar to using a Mahalanobis distance. It has been argued that the renormalization step is critical for a better comparison metric, see [8].

In order to compare results in a fair manner, we will use $D' = 128$ dimensions for all our experiments following the trend of previous research in short image representations.

2.3. The baseline method

This paper builds upon the work [8], which is briefly reviewed in this section. In [8], a joint dimensionality reduction of multiple vocabularies is proposed. Image representation vectors are separately SSR normalized for each vocabulary, concatenated and then jointly PCA-reduced and whitened as explained in the Section 2.2. The *idf* term is ignored, and it is noted that the influence is limited when used with multiple vocabularies. Results of this method are shown in Figure 1 (right plots). Comparing to the straightforward concatenation (Figure 1, left plots) where the results do not noticeably improve after adding multiple vocabularies, it can be noticed that an improvement in performance is achieved even when keeping low memory requirements by using PCA dimensionality reduction. However, for some vocabularies (i.e. $k = 2$k), performance is dropping after only few vocabularies used.

3. Datasets and evaluation

Results of our methods are evaluated on the datasets [25, 26, 9] that are widely used in the image retrieval area. Also, we compare our results with other approaches evaluated on the same datasets.

Oxford5k [25] and Paris6k [26]: Both datasets contain a set of images (5062 for Oxford and 6300 for Paris) having 11 different landmarks together with distractors, downloaded from Flickr by searching for tags of popular landmarks. For each of the 11 landmarks there are 5 different query regions defined by a bounding box, meaning that there are 55 different query regions per dataset. The performance is reported as mean average precision (mAP), see [25] for more details. In our experiments we use Paris6k as a training dataset in order to learn the visual vocabulary and projections of PCA dimensionality reduction. When evaluating our methods on Oxford5k, we always use the data learned on Paris6k.

Oxford105k [25]: This dataset is the combination of Oxford5k dataset and 99782 negative images crawled from Flickr using 145 most popular tags. This dataset is used to evaluate the search performance (reported as mAP) on a large scale. Paris6k is used as a training dataset for Oxford105k.

Holidays [9]: This dataset is a selection of personal holidays photos (1491 images) from INRIA, including a large variety of scene types (natural, man-made, water and fire effects, etc.). A sample of 500 images from the whole dataset is selected for query purposes [9]. The performance is reported as mAP, like for Oxford5k and Oxford105k, after excluding the query image from the results. As a training dataset for vocabulary construction and image representation level PCA learning we use Paris6k dataset in all experiments.

4. Sources of multiple codebooks

We propose combining multiple vocabularies that are differing not just in random initialization of clustering procedure, but also in the data used for clustering. The feature data are alternated in the process of local features description. This process is not trying to synthesize appearance deformations, but rather varying certain design choices in the pipeline of feature description, such as the relative size of the measurement region. Vocabularies created in this manner will contain less redundancy. This is combined with joint PCA dimensionality reduction (as described in Sections 2.2 and 2.3) in order to produce short-vector image representations that are used for searching the most similar images in the dataset.

Quantization complexity for all vocabularies used in experiments is given in Table 1. As stated in [8], time necessary to quantize 2000 local descriptors of a query image, for four $k = 8$k vocabularies, on 12 cores is 0.45s, using a multi-threaded exhaustive search implementation. Timings are proportional to the vocabulary size, i.e., to the number in the right column of Table 1.

Multiple measurement regions. An affine invariant descriptor of an affine covariant region can be extracted from any affine covariant constructed measurement region [17].

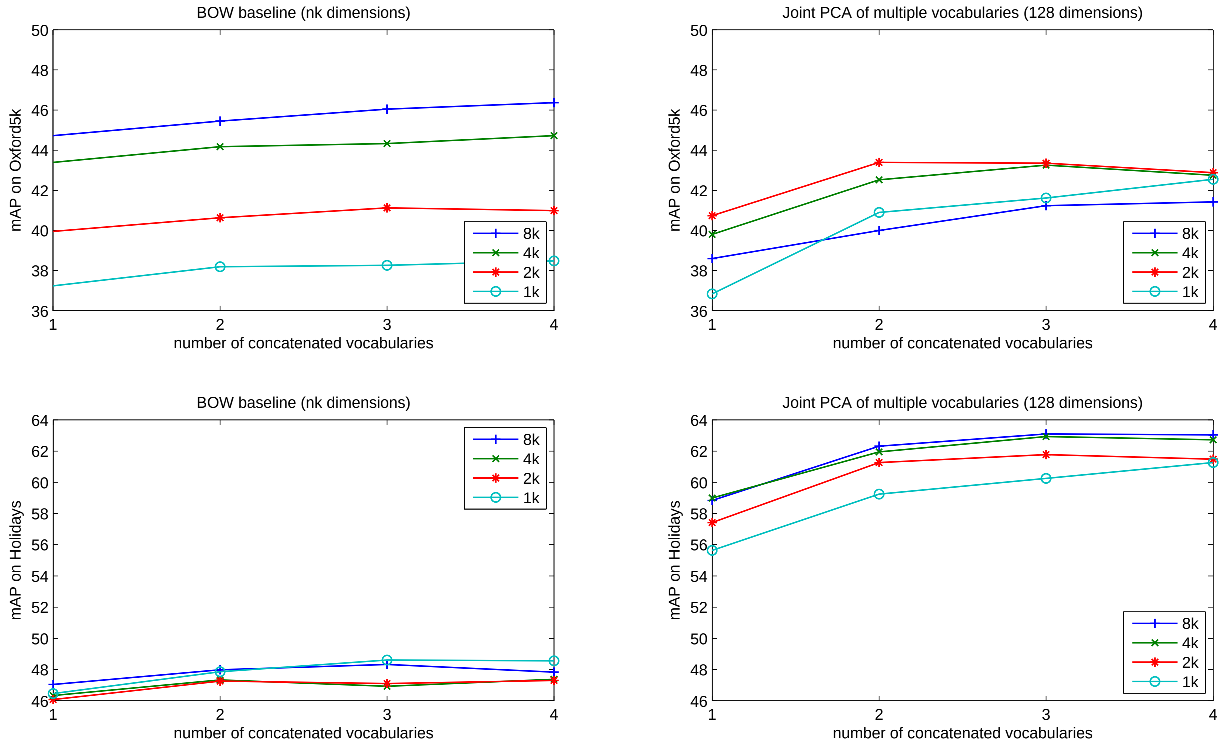

Figure 1. **Baseline methods:** Left plots show mAP performance on Oxford5k (upper plot) and Holidays (lower plot) after straightforward concatenation of BOW vectors (no PCA dimensionality reduction performed) generated using multiple vocabularies. Note that dimensionality of BOW grows linearly with every new concatenation. Right plots present mAP performance on Oxford5k and Holidays after joint PCA dimensionality reduction of concatenated BOW representations to a $D' = 128$ dimensional vector.

Table 1. **Complexity of vocabularies used throughout the experiments:** Complexity is given as a number of vector comparisons per local descriptor during the construction of the final BOW image representation.

Vocabulary	Complexity
8k	8192
4k	4096
2k	2048
1k	1024
4k+2k+...+128	8064
2k+1k+...+128	3968
1k+512+256+128	1920
512+256+128	896

As an example of a measurement region that is, in general, of a different shape than the detected region, is an ellipse fitted to the regions, as proposed by [29] and also used for MSERs [17]. An important parameter is the relative scale of the measurement region with respect to the scale of the detected region. Since the output of the detector is designed to be repeatable, it is usually not discriminative. To increase the disriminability of the descriptor, it is commonly extracted from area larger than the detected region. In case of [23], the relative change in the radius is $r = 3\sqrt{3}$. The larger the region, the higher discriminability of the descriptor, as long as the measurement region covers a close-to-planar surface. On the other hand, larger image patches have higher chance of hitting depth discontinuities and thus being corrupted. An example of multiple measurement regions is shown in Figure 2. To take the best of this trade off, we propose to construct multiple vocabularies over descriptors extracted at multiple relative scales of the measurement regions. Including lower scales leverages the disadvantages of large measurement regions, while joint dimensionality reduction eliminates the dependencies between the representations.

We consider using different sizes of measurement regions: $0.5 \times r$, $0.75 \times r$, $1 \times r$, $1.25 \times r$, $1.5 \times r$; creating slightly different SIFT descriptors used to learn every vocabulary. Implementation is very simple and during online stage the computation has to be done only for the features from query image region. Though simple, this method provides significant improvement even when concatenating vocabularies of small sizes (i.e. $k = 2\text{k}$ and $k = 1\text{k}$), see Figure 3 (left plot). We also explore the use of vocabularies with different sizes. All BOW vectors in this case are weighted proportionally to the logarithm of their vocabu-

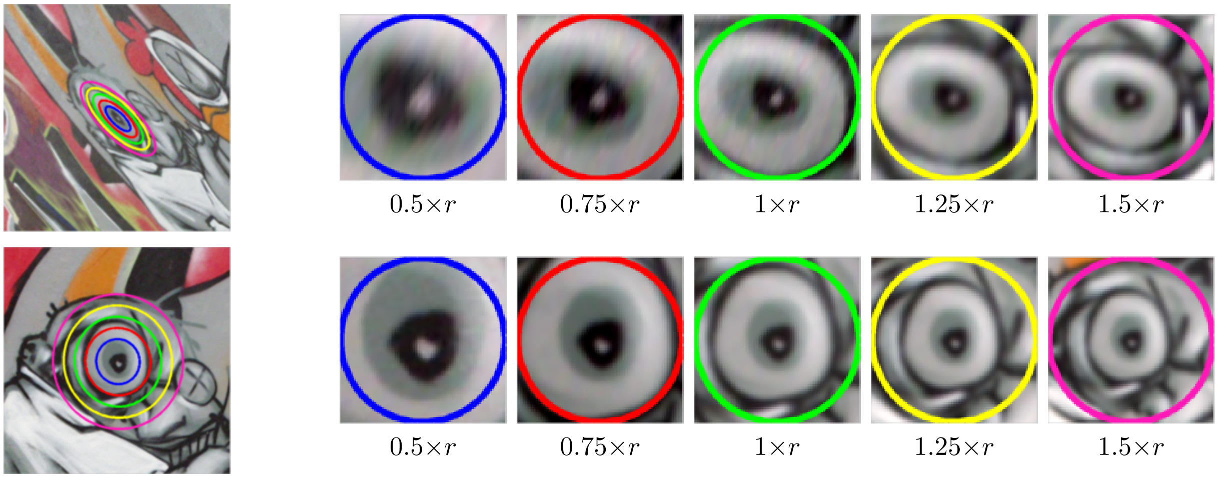

Figure 2. **Multiple measurement regions (mMeasReg):** A corresponding feature is detected in two images (left). Multiple measurement regions for a single detected feature are shown in each row. The normalized patches (right) show different image content described by the respective descriptor.

lary size [8]. In each step we concatenate a new bundle of vocabularies with multiple sizes, calculated with a different measurement region. We notice improvement when using multiple vocabulary sizes as well, see Figure 3 (right plot). For presentation of results on both plots in Figure 3, in every step we are adding a different vocabulary created on SIFT vectors with measurement regions in predefined order: $0.5 \times r$, $0.75 \times r$, $1 \times r$, $1.25 \times r$, $1.5 \times r$. This approach is denoted as mMeasReg.

Multiple power-law normalized SIFT descriptors. SIFT descriptors [16] were the popular choice in most of the image retrieval systems for a long time. Arandjelovic *et al.* [1] show that using a Hellinger kernel instead of standard Euclidian distance to measure the similarity between SIFT descriptors leads to a noticeable performance boost in retrieval system. The kernel is implemented by simply square rooting every component of SIFT descriptor. Using Euclidian distance on these new RootSIFT descriptors will give the same result as using Hellinger kernel on the original SIFT descriptors. In general, a power-law normalization [24] with any power $0 \leq \beta \leq 1$ can be applied to the descriptors ($\beta = 0.5$ resulting in RootSIFT [1]). Voronoi cells constructed in power-law normalized descriptor spaces can be seen as non-linear hyper-surfaces separating the features in the original (SIFT) descriptor space. Concatenation of such feature space partitionings reduces the redundant information.

There is no additional memory required and the change can be done on-the-fly with virtually no additional computational cost using simple power operation. We consider building four different vocabularies using: SIFT and SIFT with every component to the power of 0.4, 0.5, 0.6 (denoted as $\text{SIFT}^{0.4}$, $\text{SIFT}^{0.5}$, $\text{SIFT}^{0.6}$ respectively). Concatenation is done on single vocabularies (Figure 4, left plot) and on a bundle of vocabularies with different sizes (Figure 4, right plot). Adding all SIFT modifications to the process of vocabulary creation achieves noticeable improvement of retrieval performance in the case of all vocabulary sizes. We denote this method as mRootSIFT.

Combining vocabularies of different SIFT exponents improves over combining different vocabularies of a single SIFT exponent. For example, for 4×2k vocabularies, the mAP on Oxford5k is 46.5 for $4 \times \text{SIFT}^{0.5}$, and 47.7 (Figure 4 left) for exponent combination.

Multiple linear projections of SIFT descriptors. In locality sensitive hashing (random) linear projections are commonly used to reduce the dimensionality of the space while preserving locality. The idea pursued in this part of the paper is to use linear projections on the feature descriptors (SIFTs) before the vocabulary construction via k-means. However, random projections do not reflect the structure of the descriptors, resulting in noisy descriptor space partitionings. We propose to use PCA learned linear projections of SIFTs, learned on different training sets or subsets. The projections learned this way account for the statistics given by the training sets and hence produce meaningful distances, while inserting different biases into the vocabulary construction.

The improvement is twofold: (i) increased performance measured by mAP, and (ii) shorter quantization time during query due to shorter local descriptors after dimensionality reduction. On the other side there is a small amount of storage required to save learned projection matrices for every vocabulary, which we reuse at query. We consider and evaluate three different approaches for learning the eigenvectors used to project SIFT vectors from D to D' dimensions:

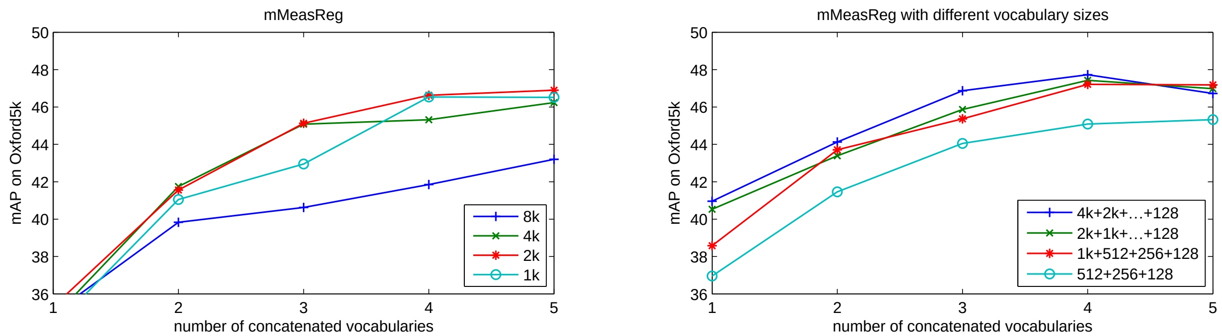

Figure 3. **Multiple measurement regions (mMeasReg):** mAP performance improvement on Oxford5k after PCA reduction to $D' = 128$ of concatenated BOW vectors produced on vocabularies created using SIFT descriptors with different measurement regions: $0.5{\times}r$, $0.75{\times}r$, $1{\times}r$, $1.25{\times}r$, $1.5{\times}r$.

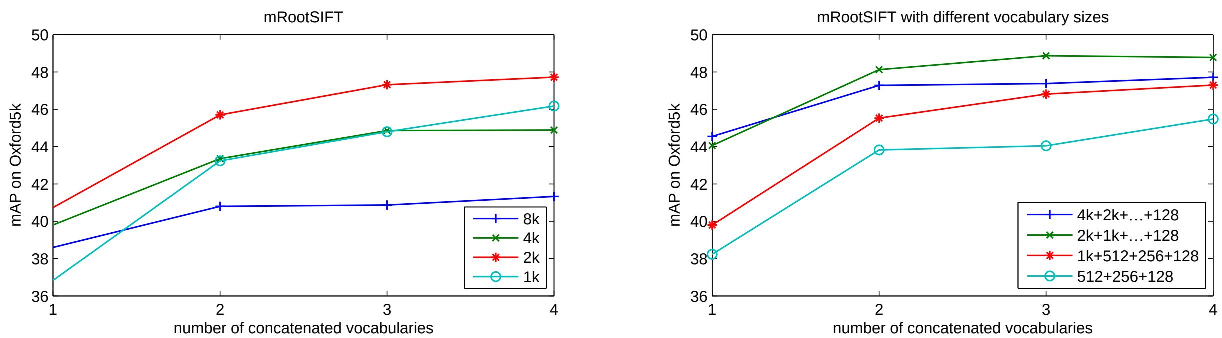

Figure 4. **Multiple power-law normalized SIFT descriptors (mRootSIFT):** mAP performance improvement on Oxford5k after PCA reduction to $D' = 128$ of concatenated BOW vectors produced on vocabularies created using multiple local feature descriptors: SIFT, SIFT$^{0.4}$, SIFT$^{0.5}$, SIFT$^{0.6}$.

1. We learn eigenvectors on Paris6k dataset and reduce the dimension of SIFT descriptors to $D' = 80, 64, 48, 32$ in the respective order for every newly created vocabulary (mPCA$_1$-SIFT). Results of this experiment are shown in Figure 5, 1$^{\text{st}}$ row.
2. We learn eigenvectors on different datasets: Paris6k, Holidays, University of Kentucky benchmark (UKB), PASCAL VOC'07 training in the respective order for every newly created vocabulary (mPCA$_2$-SIFT). Dimension of SIFT descriptors is reduced to $D' = 80$ in all cases. For the mAP performance on Oxford5k, see Figure 5, 2$^{\text{nd}}$ row.
3. We learn eigenvectors on different datasets: Paris5k, Holidays, UKB, PASCAL VOC'07 training and reduce the dimension of SIFT descriptors differently for each dataset ($D' = 80, 64, 48, 32$ respectively) creating different vocabularies (mPCA$_3$-SIFT). Performance is presented in Figure 5, 3$^{\text{rd}}$ row.

Note that first vocabulary in all three different approaches is produced using standard SIFT descriptors without PCA reduction. A new vocabulary is added in every step of the experiment having joint dimensionality reduction of 5 concatenated BOW vectors in the end.

Multiple feature detectors. In the Video Google approach [27] the authors combine vocabularies created from two different feature types. In this paper we attempt to combine Hessian-Affine [23] and MSER [17] detectors. Even though straightforward concatenation of BOW vectors created on $k = 8$k vocabularies (48.7 mAP on Oxford5k) gives improvement over using single BOW representations with Hessian-Affine (44.7) and MSER (40.1) features, after joint PCA reduction there is a decrease of performance when combining features (37.0 mAP on Oxford5k) compared to only doing PCA reduction on a single Hessian-Affine vocabulary (38.6), and an increase in performance when compared to PCA-reduced BOW vectors built on a single MSER vocabulary (24.4). Similar conclusions are made when combining smaller vocabulary sizes, i.e., there is always a drop in performance when comparing PCA reduction on a single vocabulary with Hessian-Affine features and PCA on combined vocabularies with Hessian-Affine and MSER features; mAP drop: from 39.8 to 39.1, from 40.7 to 38.7, from 36.8 to 35.1 for $k = 4$k, 2k, 1k respectively. We also experimented with combining Harris-Affine [19] with Hessian-Affine features in the same manner as with MSER, but the improvement is not significant. PCA reduction of a single $k = 8$k vocabulary on Hessian-

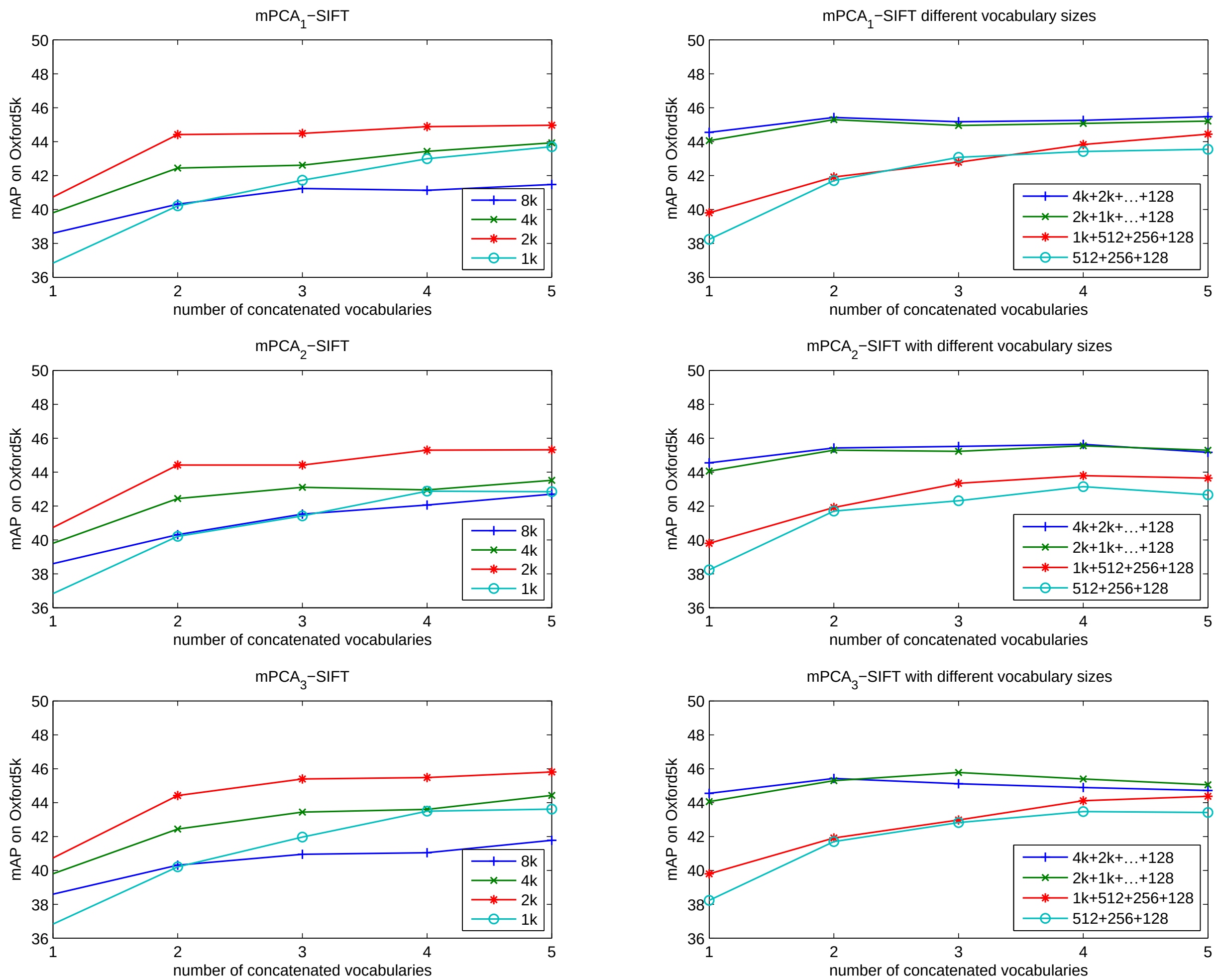

Figure 5. **Multiple linear projections of SIFT descriptors (mPCA-SIFT):** mAP performance improvement on Oxford5k after PCA reduction to $D' = 128$ of concatenated BOW vectors produced on vocabularies created using different PCA-reduced SIFT descriptors. For more details about all three presented methods see Section 4.

Affine yields 38.6 mAP on Oxford5k while joint PCA after adding a vocabulary of the same size built on Harris-Affine improves mAP to 39.0, which is smaller improvement than using two vocabularies built on Hessian-Affine features with different randomization (40.0 mAP).

Discussion. In order to better understand the impact of using multiple vocabularies we count the number of unique assignments in the product vocabulary. It corresponds to the number of non-empty cells of the descriptor space generated by all vocabularies simultaneously. The maximum possible number of unique assignments is equal to the product of number of clusters (cells) of all joint vocabularies. The number is related to the precision of reconstruction of each feature descriptor from its visual word assignments. For combination of vocabularies with different SIFT exponents (mRootSIFT) the number of unique assignments for Oxford5k dataset is shown in Figure 6. The plots are similar for all vocabulary combinations.

4.1. Comparison with the state-of-the-art

Comparison with the current methods dealing with short vector image representation is given in Table 2. Authors of the baseline approach on multiple vocabularies (mVocab) did not provide results for Oxford5k and Oxford105k datasets using all of their proposed methods, so we reimplemented and presented the corresponding results. Compared to their best method on Oxford5k that achieves 42.9 mAP, our best method (48.8 mAP) obtains significant relative improvement of 13.8%. In fact, all our methods outperform mVocab baseline methods on Oxford5k by a noticeable margin, with an improvement of 6.1% in the case of our worst performing method. When evaluating large-scale retrieval on Oxford105k dataset our methods again outperform the baseline method, relative improvement is

Table 2. **Comparison with the state-of-the-art on short vector image representation** ($D' = 128$)**:** Results in the first section of the table are mostly obtained from the paper [14], except for the recent method on triangulation embedding and democratic aggregation with rotation and normalization (ϕ_Δ+ψ_d+RN) proposed in [15]. In the second section we present results from methods that are using joint PCA and whitening of high dimensional vectors as we do. Results marked with * are obtained after our reimplementation of the methods using feature detector and descriptor as described in Section 2.1 and Paris6k as a learning dataset. In the last section of the table we present results of our methods. All methods are described in detail in Section 4.

Method	Vocabulary	Oxford5k	Oxford105k	Holidays
GIST [22]	N/A	–	–	36.5
BOW [27]	k=20k	19.4	–	45.2
Improved Fisher [24]	k=64	30.1	–	56.5
VLAD [13]	k=64	–	–	51.0
VLAD+SSR [14]	k=64	28.7	–	55.7
ϕ_Δ+ψ_d+RN [15]	k=16	43.3	35.3	61.7
mVocab/BOW [8]	k=4×8k	41.3/41.4*	–/33.2*	56.7/63.0*
mVocab/BOW [8]	k=2×(32k+...+128)	–/42.9*	–/35.1*	60.0/64.5*
mVocab/VLAD [8]	k=4×256	–	–	61.4
mVocab/VLAD+adapt+innorm [2]	k=4×256	44.8	37.4	62.5
mMeasReg/mVocab/BOW	k=5×2k	46.9	38.9	66.9
mMeasReg/mVocab/BOW	k=4×(4k+...+128)	47.7	39.2	67.3
mRootSIFT/mVocab/BOW	k=4×2k	47.7	39.8	64.3
mRootSIFT/mVocab/BOW	k=4×(2k+...+128)	48.8	41.4	65.6
mPCA$_3$-SIFT/mVocab/BOW	k=5×2k	45.8	38.1	63.2
mPCA$_1$-SIFT/mVocab/BOW	k=5×(4k+...+128)	45.5	37.8	64.6

17.9% for our best performing method, and 7.7% for the worst performing one. In order to make a fair comparison when evaluating on Holidays dataset we again reimplemented the baseline approach, using Paris6k for learning the vocabularies and PCA projections (as we did in all our methods). In this case, the relative improvement is 4.3% with our best method (from 64.5 mAP to 67.3 mAP). We also compare our methods to two recent state-of-the-art approaches on short representations [2, 15]. On Oxford5k and Oxford105k we improve as much as 8.9% and 10.7%, respectively, compared to VLAD based approach [2], and 12.7% and 17.3%, respectively, compared to T-embedding based approach [15]. On Holidays dataset relative improvement is 7.7% compared to the former and 9.1% compared to the latter. Note that the dataset used for learning of the meta-data for Holidays is different: we use Paris6k, while both [2] and [15] are using an independent dataset comprising of 60k images downloaded from Flickr.

5. Conclusions

Methods for multiple vocabulary construction were studied and evaluated in this paper. Following [8], the concatenated BOW image representations from multiple vocabularies were subject to joint dimensionality reduction to 128D descriptors. We have experimentally shown that generating diverse multiple vocabularies has crucial impact on search performance. Each of the multiple vocabularies was learned on local feature descriptors obtained with varying parameter settings. That includes feature descriptors extracted from measurement regions of different scales, different power-law normalizations of the SIFT descriptors, and applying different linear projections to feature descriptors prior to k-means quantization. The proposed vocabulary constructions improve performance over the baseline method [8], where only different initializations were used to produce multiple vocabularies. More importantly, the *all* proposed methods exceed the state-of-the-art results [2, 15] by a large margin. The choice of the optimal combination of vocabularies to combine still remains an open problem.

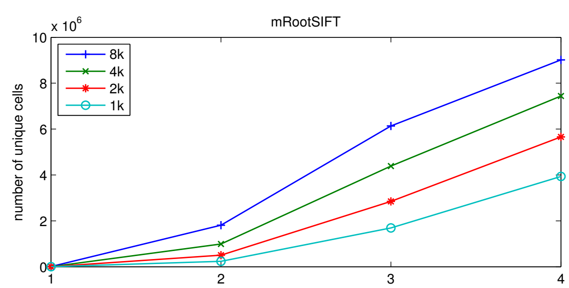

Figure 6. **Number of unique assignments (vocabulary cells) for Oxford5k dataset when combining vocabularies built on multiple power-law normalized SIFT descriptors (mRootSIFT):** SIFT, SIFT$^{0.4}$, SIFT$^{0.5}$, SIFT$^{0.6}$.

Acknowledgements. The authors were supported by MSMT LL1303 ERC-CZ and ERC VIAMASS no. 336054 grants.

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, pages 2911–2918, 2012. 2, 5

[2] R. Arandjelović and A. Zisserman. All about VLAD. In *Proc. CVPR*, 2013. 2, 8

[3] A. Babenko and V. Lempitsky. Additive quantization for extreme vector compression. In *Proc. CVPR*, pages 931–938. IEEE, 2014. 1

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 1, 2

[5] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. CVPR*, 2010. 2

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007. 1

[7] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 2

[8] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *Proc. ECCV*, Firenze, Italy, Oct. 2012. 1, 2, 3, 5, 8

[9] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008. 3

[10] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Proc. CVPR*, 2009. 2

[11] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, 2010. 1

[12] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE PAMI*, 33(1):117–128, 2011. 1

[13] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010. 1, 8

[14] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE PAMI*, 34(9):1704–1716, 2012. 2, 8

[15] H. Jégou, A. Zisserman, et al. Triangulation embedding and democratic aggregation for image search. In *Proc. CVPR*, 2014. 2, 8

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Proc. ICCV*, 60(2):91–110, 2004. 2, 5

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, volume 1, pages 384–393, 2002. 2, 3, 4, 6

[18] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004. 2

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005. 6

[20] A. Mikulik, M. Perd̆och, O. Chum, and J. Matas. Learning vocabularies over a fine quantization. *IJCV*, pages 1–13, 2012. 1, 2

[21] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006. 1

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 8

[23] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR*, 2009. 2, 4, 6

[24] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proc. CVPR*, 2010. 1, 2, 5, 8

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007. 3

[26] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in largescale image databases. In *Proc. CVPR*, 2008. 3

[27] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003. 1, 2, 6, 8

[28] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proc. CVPR*, pages 1–8. IEEE, 2008. 1

[29] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. BMVC*, 2000. 4

[30] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Proc. NIPS*, pages 1753–1760, 2009. 1