Mean BoF per Quadrant
Simple and Effective Way to Embed Spatial Information in Bag of Features

Joan Sosa-García and Francesca Odone
Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi,
Università degli Studi di Genova, Genova, Italy

Keywords: Content-based Image Retrieval, Image Description, Embedding Space Information.

Abstract: This paper proposes a new approach for embedding spatial information into a Bag of Features image descriptor, primarily meant for image retrieval. The method is conceptually related to Spatial Pyramids but instead of requiring fixed and arbitrary sub-regions where to compute region-based BoF, it relies on an adaptive procedure based on multiple partitioning of the image in four quadrants (the NE, NW, SE, SW regions of the image). To obtain a compact and efficient description, all BoF related to the same quadrant are averaged, obtaining four descriptors which capture the dominant structures of the main areas of the image, and then concatenated. The computational cost of the method is the same as BoF and the size of the descriptor comparable to BoF, but the amount of spatial information retained is considerable, as shown in the experimental analysis carried out on benchmarks.

1 INTRODUCTION

In recent years, Content-Based Image Retrieval (CBIR) has been a very active research area (Liu et al., 2007; Rui et al., 1999). Besides its natural application to image datasets browsing, CBIR has been exploited in diverse domains, including location recognition (Crandall et al., 2009), image compression (Wu et al., 2014; Dai et al., 2012), Structure from Motion (Gherardi et al., 2011; Agarwal et al., 2009). Common to all these application domains is the need to represent, store, and access a huge number of images. Besides that, different applications pose different challenges and provide different insights.

Therefore, when designing image descriptors for CBIR engines, one must be aware about the peculiarities of the target application. For instance, in *partial-duplicate image search* the goal is to identify images containing the same scene captured from different point of views or variants of the query image altered in scale, contrast, containing occlusions or derived by cropping. In this case very accurate image descriptors are required, possibly robust to geometric transformations, noise, and appearance changes. Instances of this problem may be found in a variety of applications, ranging from copyright violation detection to place localization. Typically, this problem has been addressed by using local feature matching, which is not appropriate for large-scale datasets. Instead, pure *semantic-search* grounds on the idea that query and target images share the same concept more than content. Usually it addresses the problem of finding images containing objects of the same category to the query or somehow semantically related. In this setting the image descriptor should be able to capture the essence of the content, ideally discarding the influence of the specific instance.

Today most CBIR state-of-the-art methods rely on the Bag-of-Features (BoF) representations (Sivic and Zisserman, 2003; Nister and Stewenius, 2006; Csurka et al., 2004) or its derivatives (Perronnin et al., 2010; Lazebnik et al., 2006; Boureau et al., 2011). This approach has established a general framework of image retrieval. The n dataset images are scanned for representative elements and a descriptor is computed for each element (*feature extraction*). These descriptors are then clustered into a vocabulary of visual words (*visual dictionary*), and each descriptor is mapped to the closest visual word (*vector quantization*). An image is then represented as a bag of visual words (*image representation*), and these image descriptors will be used later for retrieval (*search*) through an appropriate similarity measure. The main idea of using the BoF model is to mimic text retrieval systems, and in particular to exploit the *inverted file* indexing structure (Zobel et al., 1998), which is efficient to compute Minkowski distance (Nister and Stewenius, 2006) between high dimensional sparse vectors.

The main drawback of the BoF model is that it disregards all information about the spatial distribu-

Sosa-Garcia J. and Odone F..
Mean BoF per Quadrant - Simple and Effective Way to Embed Spatial Information in Bag of Features.
DOI: 10.5220/0005281002970304
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 297-304
ISBN: 978-989-758-090-1

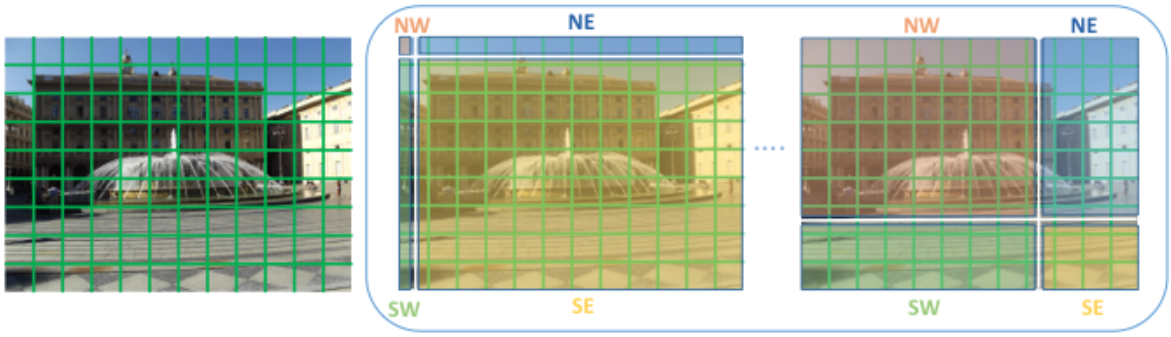

Figure 1: Image representation based on MBoFQ approach. A dense grid of local features is considered (left). Then different image partitioning in 4 quadrants are considered (right): each quadrant is associated with a different color and a label (NW - north west; NE - north east; SW - south west; SE - south east). A BoF descriptor is computed at each quadrant for all possible partitioning. An average of all BoF derived from each quadrant is obtained. Finally, a global vector concatenates the 4 quadrant descriptors.

tion of the visual words, which greatly reduces the descriptive power of the image representation and thus leads to inaccurate search results. Many approaches have been proposed to improve different stages of the classical pipeline based on BoF. To increase the quantization efficiency, hierarchical quantization (Nister and Stewenius, 2006), soft assignment (Philbin et al., 2008) and Hamming embedding (Jégou et al., 2010a) have been proposed. Alternative one may resort to quantization techniques producing very compact representations (e.g. 20 bytes), such as Fisher Kernel (Jaakkola and Haussler, 1999) or Vector of locally Aggregated Descriptors VLAD (Jégou et al., 2010b), followed by dimensionality reduction and appropriate indexing (Jegou et al., 2011). These recent methods provide excellent search accuracy with a reasonable vector dimensionality. However, these methods cannot work well in partial-duplicate image search where the object of interest only takes a small image region with cluttered background.

Some other schemes, particularly effective for partial-duplicate image search, improve the image search performance in the post-processing stage. RANSAC and neighboring-feature geometric consistency verification have been proposed to re-rank the results returned from BoF model and demonstrated that the spatial constraints consistently improve the search quality (Jegou et al., 2008; Philbin et al., 2007). This step is computationally expensive, since it is applied on a large number of local features, and is therefore non suitable for large scale image retrieval. Besides the above spatial verification techniques, query expansion is another important post-processing strategy. It reissues the initial highly ranked results to generate new queries so as to improve the recall performance (Chum et al., 2007; Kuo et al., 2009).

Incorporating spatial information a priori into the image descriptor is another relevant solution to improve the retrieval accuracy. There exist several papers in the literature for integrating spatial information into the image content descriptor (Jégou et al., 2010a; Zhou et al., 2010), which will be described in some details in the next section, where we also highlight the benefits of our contribution.

In this paper we propose a new approach to embed spatial information into the final image descriptors for image retrieval tasks. The method we propose, the Mean Bag-of-Features per Quadrant (henceforth MBoFQ) is inspired by the reasoning behind Spatial Pyramid Matching (SPM)(Lazebnik et al., 2006) but, instead of considering fixed hand crafted image partitioning, it considers multiple partitioning of the image in a four-cell grid and then it averages contributions obtained by the different partitioning. MBoFQ is more accurate than the BoF model, but still appropriate for semantic search. We adopt multiple partitioning of the image in four quadrants — north east (NE), north west (NW), south east (SE), south west (SW) — obtained by varying the origin of the considered reference system across the position of all possible image features (see Figure 1 for a visual impression of the concept). This multiple partitioning allows us to discover the different structures spread on the image, encode their relationships without the need to choose a fixed hand crafted partitioning before hand (as it is common practice in Spatial Pyramid models (Lazebnik et al., 2006)). All these partitioning produce a set of intermediate descriptors which are then averaged in a single *low dimensional* vector. The proposed approach can easily be used in conjunction with inverted file structure and its performances can be further boosted by adopting appropriate similarity measures (Jégou and Chum, 2012).

The remainder of this paper is organized as follows: Section 2 reviews state-of-the-art on encoding spatial information into image descriptors. Section 3 describes the proposed method. Section 4 reports an exhaustive experimental analysis on benchmark image retrieval datasets, while Section 5 provides a final discussion.

2 RELATED WORKS ON SPATIAL INFORMATION EMBEDDING

Integrating information about the spatial distribution of visual words into the image descriptor is a challenging task because of the combinatorial number of local features involved. However, several methods have been proposed in the last few years: the authors of (Wang et al., 2008) first cluster the salient regions into groups of neighbours, providing a set of *visual constellations* and second by representing each constellation with a BoF model. In (Sivic et al., 2005), the authors extend the BoF vocabulary to include *doublets*, i.e. pairs of visual words which co-occur within a local spatial neighborhood. Similarly, correlograms (Savarese et al., 2006) describe pairwise features in increasing neighborhoods or by appending the feature coordinates to their descriptors before building the dictionary (Mbanya et al., 2011). In (Yang et al., 2007), a descriptor is proposed to model the spatial relationship of visual words, by computing the average of the spatial distribution of a cluster center (called *keyton*) relative to all the key points of another cluster center. In (Yuan et al., 2007) the authors propose a higher-level lexicon, i.e.visual phrase lexicon, where a visual phrase is a set of spatially co-occurring visual words that form a pattern. This higher-level lexicon is less ambiguous than the lower-level one. Instead, Δ-TSR (Hoàng et al., 2010), describes triangular spatial relationships among visual entities with the aim of being invariant to image translation, rotation, scale and flipping. Many of these methods produce high-dimensional descriptors. Also, some approaches are very specific and appropriate for partial-duplicate image search (Jégou et al., 2010a; Zhou et al., 2010), therefore their use for semantic-search application is not straightforward.

Other important approaches describe the spatial layout into a hierarchy of local features. Bouchard et al. (Bouchard and Triggs, 2005) propose a generative model that encodes the geometry and appearance of generic visual object categories as a hierarchy of parts (the lowest level are local features), with probabilistic spatial relations linking parts to subparts. The Spatial Pyramid (SP) (Lazebnik et al., 2006) partitions the image into increasingly finer spatial subregions and computes a BoF vector from each sub-region. Although different image sub-divisions have been considered, typically, $2^l \times 2^l$ subregions, $l = 0,1,2$ are used. All the BoF vectors are weighted according to their level on the pyramid and concatenated to build the final image descriptor. The SPM model is a computationally efficient extension of the orderless BoF model, and has shown very promising performance on many image classification and retrieval tasks. The main drawback of SP is related to the dimension of the image descriptor ($21K$ for a 2 level pyramid, being K the vocabulary size), and for this reason SP is usually not applied to retrieval problems. Also, a hand crafted image partitioning is not always appropriate unless we know in advance the data we are considering suffer from some spatial bias. It is worth mentioning the fact that SP has instead proved very effective for image classification. In this domain various extensions to the scheme have been proposed (Boureau et al., 2011; Yang et al., 2009; Feng et al., 2011; Fanello et al., 2014).

The method we propose is related to the original Spatial Pyramid, as we use the same manner of partitioning the image into quadrants and obtain a descriptor for each quadrant. In our method, we only divide the image in four cells (equivalent to the first level of the pyramid), but this four-cell grid is moved among all local descriptors of the dense grid providing multiple 4-cell partitioning which capture the dominant structure of the image content and are not influenced by small changes in the scene. The size of the final descriptor is much smaller than a SP ($4K$ instead than $21K$), but the amount of spatial information retained is very meaningful.

3 THE PROPOSED IMAGE DESCRIPTOR

In this section, we present the main principles of our approach for incorporating spatial information into a BoF image descriptor. We first start by summarizing the general pipeline we refer to, then we describe our image descriptor, also discussing implementation and computational complexity issues.

3.1 The BoF Pipeline

We first review the stages of the standard BoF pipeline for what concerns data representation.

Local Features. In general image retrieval methods start by considering a feature detection step which selects meaningful local elements. As an alternative one

could consider a dense grid over the image, where each cell may be seen as a local feature. Regardless its origin, each local feature is normally associated with a local feature descriptor such as SIFT. Dense sampling usually inserts more information and more noise within the descriptor, thus is usually adopted primarily in image categorization (in this case noise may be filtered out by learning a discriminating function from many examples). Its main benefit is that it does not require a feature detection step and, also, it is equally applicable to different types of images, including the ones depicting poorly textured objects. In what follows we simply consider we have a set of N local features each one described by a local feature vector $\mathbf{x}_i \in R^d$, $i = 1,\ldots,N$. In this work we extract 128-dimensional SIFT descriptors densely over the image (Lazebnik et al., 2006).

Quantization. In this phase we assume we have a dictionary of visual words, which is a matrix D of size $K \times d$, where K is the size of the dictionary (number of atoms or visual words) and d is the dimensionality of the local descriptor. This dictionary is in general precomputed on an appropriate training set of images, for instance by clustering local features computed over the training data. The dictionary visual words are in this case the clusters centroids. At run time, each local feature is assigned to one visual word, via approximate nearest neighbor search (Csurka et al., 2004; Sivic and Zisserman, 2003). In this work we consider a hard assignment, where one local feature is associated with one visual word only. This choice produces a considerable sparseness, and introduces some level of arbitrarity.

3.2 Mean BoFs per Quadrant

In this section we describe the MBoFQ method for representing the image content developed with the purpose of retaining information on the spatial distribution of local features, and at the same time producing a relatively compact feature vector.

Figure 1 provides a pictorial description of the procedure, for each local feature $\mathbf{x}_i$ extracted in the image, we set its position p_i on the image plane as the origin of a $2D$ reference system. For each p_i we partition the image according to such a reference system obtaining a partitioning in 4 quadrants $\mathcal{P}_i = \{NE_i, NW_i, SE_i, SW_i\}$. We then compute a BoF representation for each quadrant, producing a BoF vector which embeds information on the different image structures belonging to the quadrant:

$$\mathbf{b}_i^q \in R^K \quad q \in \{NE, NW, SE, SW\}, i = 1\ldots,N.$$

To obtain a compact descriptor, we average all BoF vectors related to each quadrant, as follows

$$\text{avg}_q = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_i^q. \tag{1}$$

The final image *global feature vector* $MBoFQ \in R^{4K}$ is defined as follows

$$MBoFQ = [\text{avg}_{NE}, \text{avg}_{NW}, \text{avg}_{SW}, \text{avg}_{SE}].$$

The proposed description captures adaptively the dominant structure of image content on the four main regions of the image. This represents an improvement with respect to SPM because we are not considering a single fixed partition of an individual image on a fixed point, instead, we consider multiple possible partitioning. The immediate benefit for this is the reduced risk of arbitrarily dividing elements belonging to the same object; also our description is more robust to small view point changes, while it gives more weight to persistent structures.

Descriptor normalization. In each step of the algorithm we treat each quadrant of the image partition as a *subimage*. In order to avoid the negative effect of combining vectors coming from *subimages* with different sizes, after computation each BoF vector $\mathbf{b}_i^q$ is normalized with respect to the area of the current quadrant or *subimage*. The area of a quadrant is the number of local features inside a quadrant. The obtained normalized vector $\hat{\mathbf{b}}_i^q$ is used to compute the average description of quadrant q (as in Eq. (2)).

Implementation details. Instead of computing the mean BoF vectors after all partitions have been produced, we simply update cumulative averages as a new BoF vector becomes available. Let us assume we visit the local features row-wise, At iteration $i+1$ of the algorithm, while we are considering the position of the $i+1$-th feature as a reference system center, the accumulated BoF vector of the quadrant q (avg_q^{i+1}) it is updated with the new BoF vector of the quadrant ($\mathbf{b}_{i+1}^q$), as follows:

$$\text{avg}_q^{i+1} = \frac{\hat{\mathbf{b}}_{i+1}^q + i \cdot \text{avg}_q^i}{i+1}, \tag{2}$$

where the final descriptor $\text{avg}_q = \text{avg}_q^N$. For efficiency, at iteration $i+1$ the new vectors $\mathbf{b}_{i+1}^q$ are computed from the previous ones ($\mathbf{b}_i^q$) by adjusting the contributions of the current column. $\mathbf{b}_{i+1}^{NW}$ and $\mathbf{b}_{i+1}^{SW}$ vectors are updated by adding the appropriate local features, $\mathbf{b}_{i+1}^{NE}$ and $\mathbf{b}_{i+1}^{SE}$ are updated by subtracting the same features.

Computational complexity. The time complexity of the proposed approach is $O(N)$, recalling N is the number local features. The traditional BoF model has also a time complexity $O(N)$. The total number of

densely located patches N, is defined by two parameters: distance between each patch center and the size of the patches (see section 4 for details). The order of the algorithm ($O(N)$) is due to the fact that in the first step of the algorithm all local features need to be used to obtain a BoF vector for each quadrant. Subsequently, when a new origin is selected, only one column at a time is considered.

4 EXPERIMENTAL ANALYSIS

In this section, we evaluate the performance of MBoFQ with respect to the image descriptors previously introduced in the image retrieval literature. As we focus on the intrinsic quality of our proposed approach, we do not apply the post-processing stage which is usually performed on a shortlist to filter out geometrically inconsistent results. We start by reviewing the properties of the benchmark datasets we consider.

4.1 Datasets

INRIA Holidays (Jegou et al., 2008) is a dataset mainly containing high resolution holidays photos. The images were taken on purpose to test the robustness to various attacks: image rotation, viewpoint and illumination changes, blurring, etc. The dataset includes a very large variety of scene types: natural, man-made, water and fire effects, etc. This dataset contains 1491 holiday images of 500 objects and scenes manually annotated to provide a ground truth. In the experimental protocol suggested in (Jegou et al., 2008) one image per object/scene is used as a query to search within the remaining 1490 images. The retrieval performance is measured in terms of mean average precision (mAP) over the 500 queries. This dataset is targeted at large scale content-based image retrieval rather than object retrieval, due to limited changes in viewpoint and scale of each object/scene. Queries are defined only in terms of complete images and not specific image regions (objects). Note that the query image is ignored in retrieval results, unlike for Oxford 5k and Paris 6k datasets where it is counted as a positive.

Oxford 5k (Philbin et al., 2007) consists of 5062 high-resolution images collected from Flickr using queries such as "Oxford Christ Church", "Oxford Radcliffe Camera" and "Oxford". The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. Each of the 55 queries is defined by a rectangular region delimiting a building on an image. The relevant results for a query are images of this building. The accuracy is measured by mAP. This dataset was originally built for object retrieval and it is quite challenging due to substantial variations in scale, viewpoint, occlusions, distorsion and lighting conditions for a single object.

Datasets for the Learning Stages. Following a common practice in the literature, we use an independent dataset for building the vocabulary and for the other learning stages when evaluating on Holidays dataset. This independent dataset consists of 12502 images (Flickr 12k) selected from Flickr100K (Philbin et al., 2007). Instead we use Paris 6k to learn the meta-data associated with the evaluation on Oxford 5k. Analogously to Oxford 5k, the Paris 6k dataset (Philbin et al., 2008) consists of 6412 images collected from Flickr by searching for particular Paris landmarks. As it contains images of Paris it is considered to be an independent dataset from Oxford 5k and thus commonly used to test effects of computing a visual vocabulary from it while evaluating performance on Oxford 5k.

4.2 Experimental Protocol

Features. We extract 128-dimensional SIFT descriptors densely over the images similarly to (Lazebnik et al., 2006). Each image is first resized proportionally, to a maximum value of width and height of 600 pixels. The SIFT features are extracted from densely located patches centered at every 8 pixels on the image and the size of the patches is fixed as 16×16 pixels. The number of samples used to build the dictionary is 1M, selected randomly from all local features of the current independent dataset (Flickr 12k or Paris 6k). Each SIFT descriptor is encoded into a K-dimensional code vector, based on the learnt dictionary, by hard vector quantization.

Improving Image Retrieval Quality. Simple techniques may help improving the quality of BoF and VLAD representations (Jégou and Chum, 2012). These heuristics include *(i)* a transformation of the original vector representation $\mathbf{v} = \mathbf{v} - \alpha \cdot \bar{\mathbf{v}}$ which allows a similarity measure such as the cosine transform to appreciate the co-occurrence of missing words in two different feature vectors; *(ii)* a whitening the vector space jointly with the dimensionality reduction (PCA). The benefit of these heuristics will be assessed in the reminder of the section.

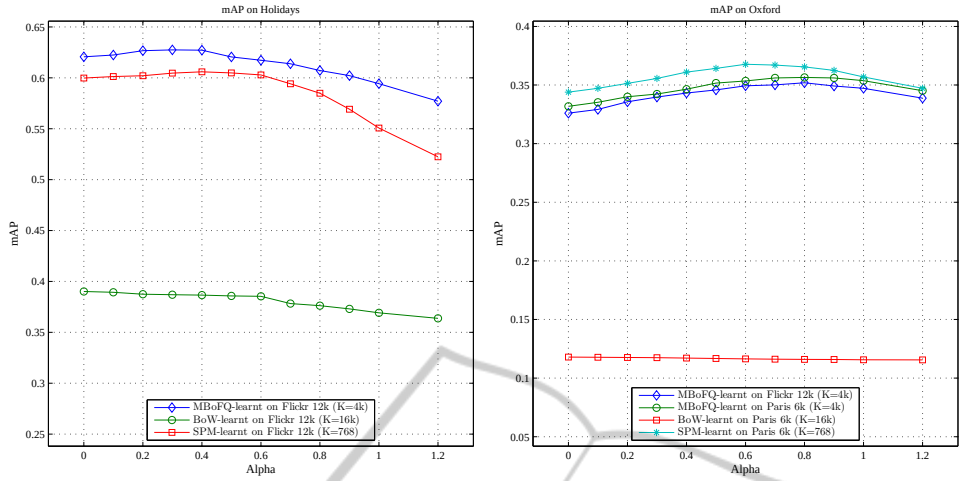

Figure 2: Different descriptors for Holidays and Oxford 5k datasets and the effect of co-missing words (see text).

4.3 Co-missing Visual Words Effect

As a first experiment we evaluate the appropriateness of our descriptor on the benchmark retrieval tasks previously described and, contextually, analyze the benefits of applying the vector normalization heuristic described above followed by cosine similarity (Jégou et al., 2012). To this purpose, we mimic the experimental setup of (Jégou and Chum, 2012), but using dense extraction instead of interest-point detector. We learn the vocabulary D and estimate the mean vector ($\bar{v}$) on Paris 6k and Flickr 12k for the Oxford 5k tests and on Flickr 12k for the Holidays tests. We also report the performance of SPM and BoW descriptors. The vector size is 16k for all descriptors compared in Figure 2. Notice the value $\alpha = 0$ corresponds to the case of non-transformation and thereby are using the original image descriptors. Figure 2 illustrates the impact of the novel cosine similarity as a function of α. It can be observed that the proposed update of the descriptors produces significant improvements: for Oxford 5k dataset there is an increase in performance of about 2.5% with respect to the original formulation (corresponding to $\alpha = 0$) for the best increase (MBoFQ-learnt on Paris 6k) with $\alpha = 0.8$, while for Holidays it is almost 1% (MBoFQ-learnt on Flickr 12k) with $\alpha = 0.3$. The results reported on the next sections are those corresponding to the best value of α for each configuration.

4.4 Comparative Analysis

We now perform a comparative analysis with other descriptors from the literature.

Full Size Feature Vectors. We first report, in Table 1, the performance of image representation based on our approach against the current state-of-the-art for descriptors of medium dimensionality (20k-D to 30k-D). It is worth emphasizing that our approach only uses dense features and the reported results employ sparse features. Therefore, for fair comparison we also include our implementation of VLAD, SPM and BoW using dense features and with vector sizes comparable to our descriptors. The retrieval accuracy of the full size vectors of our approach is evaluated at different vocabulary sizes. For the Holidays dataset, the proposed approach is in line with the best performing method (Improved Fisher), while it outperforms the rest of the state-of-the-art. It is worth noting that the vector size of our descriptors is much lower than the others. Instead, the results achieved by our method on the Oxford 5k dataset are not as encouraging. The reason is the fact our descriptor is primarily

Table 1: **Full size image descriptors.** Comparison of image descriptors of medium-dimensionality (20k-D to 32k-D). Reference results are obtained from Jégou et al. (Jégou et al., 2012). For fair comparison, we also include our implementation of VLAD, SPM and BoW using dense features (denoted by: *Dense: Method).*

Method	*size*	Holidays	Oxford
BoW 200k-D	200k	0.540	0.364
BoW 20k-D	20k	0.452	0.354
Improved Fisher	$20-32$k	0.626	**0.418**
VLAD	$20-32$k	0.526	-
VLAD+SSR	$20-32$k	0.598	0.378
Dense : *VLAD*	16k	0.547	0.266
Dense : *SPM*	16k	0.605	0.367
Dense : *BoW*	16k	0.390	0.117
$MBoFQ_{K=2048}$	8192	0.583	0.286
$MBoFQ_{K=4096}$	16384	**0.627**	0.357

Table 2: **Low dimensional image descriptors.** Comparison of image descriptors of low dimensionality (128-D). Most reference results are obtained from the paper of Jégou et al. (Jégou et al., 2012). Multiple vocabulary (Multivoc) methods are from (Jégou and Chum, 2012). For fair comparison, we also include our implementation of VLAD, SPM and BoW using dense features (denoted by: *Dense: Method).*

Method	Holidays	Oxford 5k
GIST	0.365	-
BoW	0.452	0.194
Improved Fisher	0.565	0.301
VLAD	0.510	-
VLAD+SSR	0.557	0.287
Multivoc-BoW	0.567	**0.413**
Multivoc-VLAD	0.614	-
Dense : *VLAD*	0.553	0.266
Dense : *SPM*	0.620	0.322
Dense : *BoW*	0.388	0.122
$MBoFQ_{K=2048}$	0.641	0.296
$MBoFQ_{K=4096}$	**0.665**	0.325

meant for image retrieval and not for object retrieval, and indeed it tends to favor the overall image structure including the background information, which is not beneficial for object retrieval.

Low Dimensional Feature Vectors. Today, in image retrieval, is common practice to include a dimensionality reduction step over the final feature vector. This process helps reducing the size of the descriptor, improving retrieval performances, but as an additional benefit controls data redundancy. Table 2 compares our descriptor with others in the literature, after a PCA and whitening procedure (see (Jégou and Chum, 2012)). The image vectors are produced independently, using the method described in Section 3.2 and then l_2 normalized. The different descriptors are reduced into vectors of 128 components by using PCA and whitening. We mimic the experimental setup of (Jégou and Chum, 2012) (but using dense features), and learn the vocabulary and PCA on Paris 6k for the Oxford 5k tests. For the Holidays tests it is used Flickr 12k for learning the PCA and vocabulary. Table 2 also reports the results of our implementation of VLAD, SPM and BoW with dense features. Here, for the Holidays dataset we outperform the best method proposed so far (Multivoc-VLAD) by 5%. It should also be noticed that Multivoc-VLAD uses multiple vocabularies to obtain multiple VLAD descriptions of one image; instead we use only one vocabulary prior dimensionality reduction with a benefit on a reduced computation to obtained the descriptor. Also, it can be observed how, in this case, dimensionality reduction greatly improves the accuracy we obtained with the original descriptor. All these elements strongly speak in favor of the appropriateness of our descriptor for an image retrieval problem. Instead, here again, on the Oxford 5k our performances are lower than the best performing method (Multivoc-BoF which uses again multiple vocabularies). It should be noticed, though, how the performances of our descriptor are stable to the reduction of dimensionality, while most of the other methods experience a remarkable decrease of performances.

5 DISCUSSION

In this paper we presented a new approach for incorporating spatial information into BoF image descriptors. The image was partitioned adaptively by using different four-quadrant partitioning and BoFs were computed within each quadrant. Then all BoF relative to a given quadrant were averaged to obtain a robust overall description of an image region. The main advantage of the proposed approach is that it relies on simplicity to embed spatial information within the widely spread BoF.

Experimental analysis on two different benchmarks highlighted how the proposed method is very appropriate for image retrieval and quasi-duplicate search. This opens the possibility to apply the method to view-based localization and way-finding, which are our reference applications. As expected, the method is not as effective when object retrieval is needed, as it provides a structured global picture of the image content. Further evaluations need to be performed in the case of large-scale image retrieval (up to 10 million images), to asses our representation in this scenario. An analysis of the possible benefits of detecting sparse features is necessary and will be carried out in future works. Also, the proposed approach only takes into consideration the quantized local features (hard assignment) within a quadrant in the image partition and build a BoF vector from this information. The representation can be easily extended for the case of soft assignment. Also, more effective aggregation procedures, e.g. pooling operations, may also be applied.

REFERENCES

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building rome in a day. In *ICCV*, pages 72–79. IEEE.

Bouchard, G. and Triggs, B. (2005). Hierarchical part-based visual object categorization. In *CVPR*, volume 1, pages 710–715. IEEE.

Boureau, Y., Le Roux, N., Bach, F., Ponce, J., and LeCun, Y. (2011). Ask the locals: multi-way local pooling for image recognition. In *ICCV*, pages 2651–2658. IEEE.

Chum, O., Philbin, J., Sivic, J., Isard, M., and Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, pages 1–8.

Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In *Proc. WWW*, pages 761–770. ACM.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *(SLCV, ECCV 2004*, volume 1, page 22.

Dai, L., Yue, H., Sun, X., and Wu, F. (2012). Imshare: instantly sharing your mobile landmark images by search-based reconstruction. In *Proc. MM*.

Fanello, S., Noceti, N., Ciliberto, C., Metta, G., and Odone, F. (2014). Ask the image: supervised pooling to preserve feature locality. In *CVPR*.

Feng, J., Ni, B., Tian, Q., and Yan, S. (2011). Geometric ℓ_p-norm feature pooling for image classification. In *CVPR*, pages 2609–2704. IEEE.

Gherardi, R., Toldo, R., Garro, V., and Fusiello, A. (2011). Automatic camera orientation and structure recovery with samantha. *ISPRS*, pages 38–5.

Hoàng, N. V., Gouet-Brunet, V., Rukoz, M., and Manouvrier, M. (2010). Embedding spatial information into image content description for scene retrieval. *Pattern Recognition*, 43(9):3013–3024.

Jaakkola, T. and Haussler, D. (1999). Exploiting generative models in discriminative classifiers. *NIPS*, pages 487–493.

Jégou, H. and Chum, O. (2012). Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *ECCV*, pages 774–787.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317. Springer.

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

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

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010b). Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311. IEEE.

Jégou, H., Perronnin, F., Douze, M., Schmid, C., et al. (2012). Aggregating local image descriptors into compact codes. *PAMI, IEEE Tran. on*, 34(9):1704–1716.

Kuo, Y.-H., Chen, K.-T., Chiang, C.-H., and Hsu, W. H. (2009). Query expansion for hash-based image object retrieval. In *Proc. MM*, pages 65–74. ACM.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE.

Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282.

Mbanya, E., Gerke, S., and Ndjiki-Nya, P. (2011). Spatial codebooks for image categorization. In *ICMR*, page 50. ACM.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *CVPR*, volume 2, pages 2161–2168. IEEE.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156. Springer.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8. IEEE.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8. IEEE.

Rui, Y., Huang, T. S., and Chang, S.-F. (1999). Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62.

Savarese, S., Winn, J., and Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *CVPR*, volume 2, pages 2033–2040.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering objects and their location in images. In *ICCV*, pages 370–377.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477. IEEE.

Wang, W., Luo, Y., and Tang, G. (2008). Object retrieval using configurations of salient regions. In *CIVR*, pages 67–74. ACM.

Wu, X., Hu, S., Li, Z., Tang, Z., Li, J., and Zhao, J. (2014). Comparisons of threshold ezw and spiht wavelets based image compression methods. *TELKOMNIKA*.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801. IEEE.

Yang, L., Meer, P., and Foran, D. J. (2007). Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, pages 1–8. IEEE.

Yuan, J., Wu, Y., and Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, pages 1–8. IEEE.

Zhou, W., Lu, Y., Li, H., Song, Y., and Tian, Q. (2010). Spatial coding for large scale partial-duplicate web image search. In *Proc. ICMM*, pages 511–520. ACM.

Zobel, J., Moffat, A., and Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACMTDS*, 23(4):453–490.