Information Sciences 366 (2016) 99–120

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Topic modeling and improvement of image representation for large-scale image retrieval

Nguyen Anh Tu, Dong-Luong Dinh, Mostofa Kamal Rasel, Young-Koo Lee*

*Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 24 September 2015
Revised 29 April 2016
Accepted 22 May 2016
Available online 26 May 2016

*Keywords:*
Topic modeling
Probabilistic graphical model
Image retrieval
Image representation
Image coding
Bag-of-visual words

ABSTRACT

In this paper, we present a new visual search system for finding similar images in a large database. However, there are a number of challenges regarding the robustness of the image representations and the efficiency of the retrieval framework. To tackle these challenges, we first propose an encoding technique based on soft-assignment of local features to convert an entire image into a single vector, which is a compact and discriminative representation. This encoded vector is suitable for most types of efficient indexing methods to produce an initial result. To compensate for the lack of incorporating geometric and object-related information during the encoding scheme, we then propose a probabilistic topic model to formalize the spatial structure among the local features. Moreover, the topic model allows us to effectively extract the object and background regions from the image. This is performed by a Markov Chain Monte Carlo algorithm for approximate inference. Finally, benefiting from the extracted objects in each image, we present a re-ranking scheme to automatically refine the initial search results. Our proposed retrieval framework has two major advantages: i) an aggregation strategy through soft-assignment improves the discriminative power of the representation, which has a determinative effect on the retrieval precision; and ii) the probabilistic latent topic model enables us to not only gain insight into the spatial structure of the image, but also handle a large variation in the object appearance. The experimental results from four benchmark datasets show that our approach provides competitive accuracy, and runs about ten times faster. Our studies also verify that proposed approach works effectively on large-scale databases of millions of images.



## 1. Introduction

In recent years, multimedia and networking technologies have significantly impacted our daily activities. In particular, the development of smart phones and other mobile devices have increased the demand for searching for information of the Internet, books, magazines, and reference materials. Many applications have been developed for automatic recognition of different objects of interest, such as product catalogs, landmarks, and art galleries. Therefore, the areas of visual recognition and image retrieval provide fascinating research opportunities. Typically, the aim of a retrieval task is to select from a collection of objects images that are similar to a query image. However, large-scale image retrieval poses a number of challenges regarding the desired quality of the image representations and the efficiency of the retrieval framework. Hence,

* Corresponding author. Tel.: +82 312013732.
*E-mail addresses:* tunguyen@khu.ac.kr (N.A. Tu), luongdd@ntu.edu.vn (D.-L. Dinh), rasel@khu.ac.kr (M.K. Rasel), yklee@khu.ac.kr (Y.-K. Lee).

http://dx.doi.org/10.1016/j.ins.2016.05.029
0020-0255/

we seek to obtain representations of images and image regions that are discriminative and robust to the various types of data content. We also seek to design a retrieval framework that can efficiently and effectively handle large image databases with high search accuracy.

Currently, Bag-of-visual words (BoV) [48] is a seminal framework for image retrieval. In this framework, local features (e.g., SIFTs [31]), which typically achieves invariance of orientation and scale in modern visual recognition, are quantized to form a vocabulary of visual words. An image is encoded as a sparse frequency histogram over the visual vocabulary. Inheriting the characteristics of BoV, an advanced encoding scheme called Vector of Locally Aggregated Descriptors (VLAD) [20,21] has been proposed to produce a higher-order representation by including the statistics of local features. The VLAD model also encodes an image as a single vector like the BoV model by aggregating its local features. The aggregated vector can be compressed with the dimension reduction method to obtain a compact representation. Using this encoding scheme with an indexing technique like product quantization [19], an image can be represented by a small number of bytes to provide competitive search accuracy [21]. Although popular retrieval models (e.g., BoV and VLAD) work generally well, they suffer from three main issues: (1) The discriminative power of the image representation is decreased due to the hard-assignment of feature quantization used in these models, where a local feature might be assigned to a wrong visual word. This results in reducing the similarity between two images containing similar objects. (2) When encoding an image, the lack of geometric (e.g., location, scale, and orientation) and object-related information makes these approaches very sensitive to large variations in objects, such as occlusion, deformation, and viewpoint change. (3) For similarity measurement, these models focus only on estimating the global change between two images. They fail to exploit the human cognition related to object appearance, background, and their relationship. Hence, they cannot capture the local change of each object appearing in the image. This leads to decreasing the accuracy and efficiency of a retrieval task when coping with complex data content.

In this paper, we address these issues by proposing a novel retrieval framework that enables a robust representation and an efficient re-ranking method for measuring the similarity between the query image and candidate database image. Each image is modeled as a set of local features with a two-phase procedure: encoding and topic modeling.

*Image encoding*. We enhance VLAD by developing an encoding scheme called soft-VLAD. In this scheme, we use two aggregation approaches based on soft-assignment to map each local descriptor to multiple visual words and aggregate them into a single vector. The first approach is distance-based assignment, where each assignment is assigned to a weight proportional to the distance from the visual word to the local descriptor. The second approach is sparse-coding-based assignment, which uses sparse coding [38] to project local descriptors onto the learned codebook, and then computes the weight of each assignment. This approach is motivated by successful applications of sparse coding in image classification [55,57]. By using these approaches, we overcome the limitation of hard-assignment and encode each image into a highly discriminative vector for indexing and retrieving an initial list of candidates.

*Topic modeling*. By exploiting human knowledge about the object appearance and the background, we propose a generative latent topic model called the *spatial latent topic model with background distribution* (SLTMB) to extract the background regions and topic regions from the image. In this probabilistic model, each topic corresponds to an object or a part of an object occurring frequently in the image corpus. Specifically, an image contains the object instances with a certain spatial arrangement, while each object instance can also be represented by appearance of the relevant set of visual words. The SLTMB model is therefore intended to exploit the probability distributions over visual words for different topics. Thus, visual words co-occurring often in the same image with a particularly spatial distribution tend to belong to the same object or topic region. Naturally, the similar object instances will have similar probability distributions, and so the SLTMB enables us to also infer what unknown objects are present in those images and where they are. The spatial location of visual words integrated into our topic model is effectively used to compensate for the limitations of encoding the image, and so strengthens our image representation.

*Re-ranking image*. The topic regions, meanwhile, have already been extracted in the candidate images, and we can take advantage of such information to refine the search results. We observe that a database image is similar to a query image if they contain similar topic regions. Consequently, we propose a re-ranking method with a fast and efficient geometric scoring scheme for large-scale image matching. We first establish matching feature pairs between the common topics of the query and candidate images. Then, similarity scores between the common topic regions are generated based on the geometric information of the matched features. Afterward, a new score for each image pair is computed as the sum of the topic similarity scores, and re-ranking is performed using the new scores. This strategy allows us to significantly speed up re-ranking as well as perform on medium-sized datasets (e.g., a thousand images). Moreover, using extracted topic regions, our method can handle local variations of object appearance in each image.

Our main contributions are three-fold: (1) We propose an encoding scheme called soft-VLAD to produce vector representation for the whole image, which extends VLAD by using the soft-assignment approaches for aggregation. (2) Our topic model built on human knowledge about image structure formulates the appearances and locations of the different topics and background regions. This allows us to effectively extract the objects from an image. (3) We present an efficient measuring scheme in conjunction with extracted topic regions to compute the similarity between two images. Using this scheme, the proposed re-ranking method shows very promising accuracy and fast processing on publically available datasets.

The remainder of this paper is organized as follows. Section 2 provides a discussion of related works. Section 3 presents an overview of proposed framework and then describes the details of our methods, including image preprocessing (Section 3.1), an encoding scheme called soft-VLAD (Section 3.2), the SLTMB model with learning and inference procedures

(Section 3.3), and a measuring scheme for the re-ranking stage (Section 3.4). Experimental results on standard datasets are conducted and discussed in Section 4. Conclusions and discussion are presented in Section 5.

## 2. Related works

In this section, we briefly describe the related methods, which can be divided into three areas, corresponding to the three stages of our proposed retrieval framework: 1) image encoding, 2) the topic model for visual recognition, and 3) the incorporation of geometric information and post-processing.

### 2.1. Image encoding

For large-scale visual searches, typical techniques usually rely on local features. Among these techniques, BoV encoding is the most popular approach adopted from text retrieval and widely applied to image searches [18,42,43]. However, conventional BoV approach [42] requires very high vocabulary size (e.g., 1M visual words) to guarantee reliable searches, and its sparse representation cannot be applied directly to efficient indexing techniques. Consequently, this method does not scale well to databases of more than 1 million images [21], because an inverted list for very high-dimensional representation can have the issue of storage. The other disadvantage of BoV encoding is quantization loss, as we introduced in Section 1. One way to decrease the quantization error is to use soft-assignment [18,43]. For instance, researchers in [18] proposed using better representation of the individual local descriptors with Hamming embedding (HE). Recently, alternative encoding approaches such as Fisher [41,45] and VLAD [20] have been introduced to extend BoV by utilizing higher-order statistics of the local feature distribution. Compared to BoV, both Fisher and VLAD produce dense representations with the much smaller vocabulary (e.g., hundreds of visual words). On the other hand, VLAD [20] also suffers from the issue of quantization loss like BoV due to the use of hard-assignment. Fisher [41,45] can avoid this problem because local descriptors are soft-assigned to multiple visual words by using GMM models. It has also been experimentally demonstrated [21] that Fisher outperforms BoV and VLAD in most cases. However, Fisher has the disadvantage of computational complexity, it requires more processing time to train visual words by GMM models, and in some cases, the convergence of GMM is very slow.

Besides the local feature-based approaches, several recent works [14,24,39] explored deep convolutional neural networks (CNNs) to compute image representation. These methods utilize the outputs of last network layers which are fully connected to encode an input of raw image as a high-dimensional vector. Although deep learning-based approaches have shown remarkable results in visual recognition (e.g., image classification), their training process is very costly due to a huge amount of labeled data and parameters to be learned. Apart from the CNNs-based approaches, more recent works used subspace learning [28,29] and metric learning [58] to produce the compact and discriminative representation by solving optimization problems. In [28], image understanding and feature learning are combined into a joint learning framework to discover suitable subspace for data representation. In [58], to construct a new feature space, the authors proposed a method to formalize feature and semantic similarities with pairwise constraints, which can be employed for image clustering. To guarantee improving the data separation, these methods require incorporating class labels or high-level semantics (e.g., image tags) into their frameworks. However, these types of information are not available for the task of visual search.

The image representations are usually described by the high-dimensional vectors, which can be prohibitive for large-scale image retrieval due to speed and storage cost. Hence, many researches worked on image indexing to improve search efficiency. Hashing-based methods [22,32,37,41,50] have been commonly used to compress the image representation into binary codes that have a small memory footprint. In [41], Fisher Vector is binarized by employing sign binarization and Locality-Sensitive Hashing (LSH)[12] based on random projections. Later, to better maintain the retrieval performance, the advanced methods [32,50] have been presented to learn hashing functions from data. In [32], authors proposed the asymmetric cyclical hashing scheme using two hash codes with different lengths for queries and database images. A hashing method that learns a discriminant hashing function is presented in [50] by preserving the neighborhood discriminative information. Besides the hashing-based approaches, product quantization [19] and its variant [23] have been proposed to compress image representation. They partition an vectorized representation (e.g. VLAD [21]) into disjoint sub-vectors and quantize each one separately with a pre-trained codebook to generate short codes that are composed of quantizer indices. Although above methods are efficient in computation and memory usage, they only focus on the global representation rather than exploring the geometric information incorporated in local features. As a result, they are sensitive to image transformations (e.g. scale and pose changes) that often occur in practice.

In this study, we are interested in the problem of local feature-based encoding by improving the design of VLAD. Our encoding approach is motivated by Fisher [41] and the soft-assignment term employed in [43]. However, in [43], the authors consider soft-assignment to improve the accuracy of the inverted index structure and the ranking stage, rather than for the aggregation of local descriptors as in our work. Fisher [41], instead, uses an expensive soft-assignment strategy via the GMM model. In our work, we employ soft-assignment techniques that are computationally cheaper without sacrificing much accuracy. Similar to VLAD and Fisher, our encoded vector can also leverage the efficient indexing techniques to guarantee the scalable search. Besides, by exploiting the information related local features, we can further improve the robustness of image representation.

### 2.2. Topic model for visual recognition

Topic models originated from natural language processing [5,16] and have been widely adopted for solving image understanding problems [44,47,49,54]. Specifically, each image is considered to be a visual document modeled as a bag of visual words [26] that can directly be applied to topic models, such as pLSA [16] and LDA [5]. Most works have employed topic models for the tasks of learning classification or annotation [27,47,54] and segmentation [27,54], rather than for retrieval tasks, which we are interested in here. One issue of conventional topic models is that they entirely ignore spatial information, which can be used to significantly improve image recognition as shown in previous works [26,49].

As generative models, topic models have good potential for handling large-scale databases with unsupervised learning, and they can effectively exploit informative priors. In this paper, we address the limitation of existing topic models through an explicit notion of spatial location for visual word and topic. Furthermore, we integrate background distribution into our generative latent model in a manner similar to [7].

### 2.3. Incorporation of geometric information and post-processing

In the retrieval system, re-ranking stage is essentially required as a post-process to filter out highly ranked false positives of the initial search results. By incorporating geometric information, re-ranking methods mainly depend on geometric consistency between the query and candidates. Precision is improved when images contain small numbers of consistent matches are re-ranked lower. Typically, the matching method like Geometric Verification (GV) [42] is widely used, where the consistency is estimated using inlier match counts. However, this method has a high computational cost and can only verify a limited number of top-ranked images. Consequently, various approaches have been proposed to incorporate relatively weak spatial constraints in the initial search step. Feature locations are the most frequently used geometric information, and they are usually integrated into inverted file [6,51,56,59]. The works [51,59] exploit a bag-of-phrases structure to assemble visual words containing high-order relationships into visual phrases. Alternatively, [6,56] use spatial co-occurrence within the feature space. Some methods include local affine frames for each feature [40], orientation and scale parameters [18], and feature spatial distances [9]. However, these spatial constraints are either too critical such that only translation can be managed [40,51,59] or too loose to capture enough information [6,18]. Another way to improve the accuracy of verification is to automatically expand the query [8,10]. Such methods refine the query model by adding words from spatially verified regions in result images. However, the performance of query expansion tends to be broken down by false positive search results. Consequently, it requires exact geometric verification with very high computational cost.

Unlike existing methods, we present a new re-ranking method based on the extracted topic and background regions. Then, our method can capture well the local changes of object appearance, whereas previous methods cannot, due to their dependence on global geometric transformation. Moreover, the extracted regions enable us to develop a simple measuring scheme that relies on weak geometric consistency of matching scale and orientation. As a result, while previous re-ranking methods are time-consuming and limit the list of candidates, we can significantly enlarge a number of candidates to be verified due to the fast processing.

## 3. Methodology

An overview of proposed framework is shown in Fig. 1. It also serves as a guide that describes how the proposed approaches presented in this paper come together during the retrieval process. The description of different parts of the framework is given in the following sections.

### 3.1. Image preprocessing and representation

In this section, we present the main steps to preprocess an image and construct the preliminary representation before applying it to each stage of the proposed framework. First, we extract the regions of interest from the image via an affine invariant detector. Then, the detected regions (or image patches) are described using the 128-D SIFT descriptor. Similar to [36], we build a vocabulary tree (VT) in an offline process based on the set of training SIFT descriptors. The VT is created for a hierarchical quantization by using hierarchical k-means clustering. In this work, we construct the VT that has 2 levels, excluding the root node, as illustrated in Fig. 2. Initially, large clusters or a coarse codebook containing $k_c$ centroids are generated from all the training descriptors by a k-means process. For each large cluster, k-means clustering is employed for the training descriptors assigned to that cluster, to generate $k$ smaller clusters. After the second level, from the VT, we obtain a refined codebook containing $k \times k_c$ centroids.

We use each type of codebook for the corresponding phases of the visual search system. The coarse codebook is applied to the encoding phase, while the refined codebook is utilized for the topic modeling phase. During the online process, each descriptor of the image is assigned to the nearest cluster of a coarse codebook of size $k_c$ and proceeds to two phases. For the topic modeling phase, the descriptors of each image are quantized further into visual words of the refined codebook, whose vocabulary size is $k \times k_c$. Our BoV representation in this phase includes words and their geometric information in each image $d$ as $\{w_{di}, l_{di}, \xi_{di}, o_{di}\}_{i=1}^{N_d}$, where $N_d$ is the number of visual words in image $d$; $w_{di}, l_{di}, \xi_{di}, o_{di}$ are the visual word identity, location, scale, and orientation, respectively.

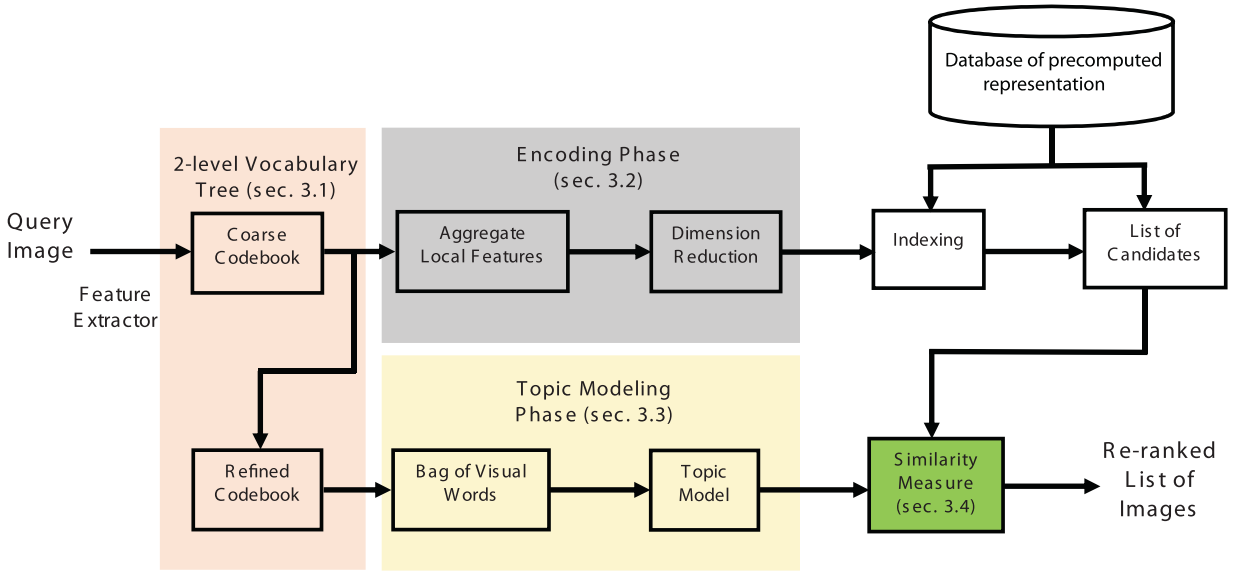


**Fig. 1.** Overview of our proposed framework. Background colors represent different stages of retrieval framework: image preprocessing (pink), image encoding (gray), topic modeling (gold), and similarity measure (green). For each colored box, we also denote a relevant section presented in the methodology. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

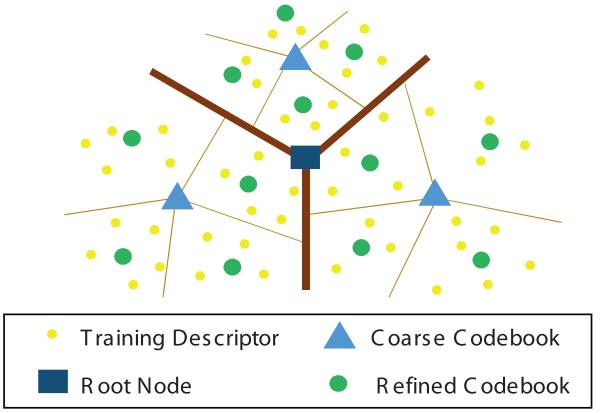


**Fig. 2.** The process of building the vocabulary tree by hierarchical k-means clustering of training descriptors.

*Discussion:* It should be noted that, for both tasks of image encoding and topic modeling, we can use most of local features such as: SURF [3], MSER [33], DAISY [52], or even deep learning-based local features [11,30], which have been developed recently using the activations from convolutional layers. Since these features have properties similar to SIFT, we can straightforwardly extract the image patches from an images, and such image patches are then used to compute local descriptors suitable for our tasks. For example, to extract visual topics, the region-based feature like MSER has been employed in several probabilistic topic models [46,47], because it captures well the homogeneous regions in an image [34]. In this work, we choose to use the keypoint-based SIFT due to two reasons: (1) SIFT is the most widely used feature for visual recognition; (2) SIFT is also suitable for our context of large-scale image retrieval. Compared to other types (e.g. region-based feature), the keypoint-based detector of SIFT outputs the higher number of interest regions, and hence performs better against complex databases such as images with occlusion and clutter [34].

### 3.2. Image encoding

In this section, we present the aggregation methods for image description, which play the key role in encoding phase to convert the whole image into a vector to be indexed. Such representations are widely employed in various computer vision tasks, such as object instance recognition, object category recognition, and image retrieval.

#### 3.2.1. Aggregation of local descriptors based on soft-assignment

Aggregating local descriptors essentially models an image as a set of local regions. This provides a certain level of robustness against changes in the object pose, as well as against local deformations. More specifically, it aggregates the descriptors

associated with a given codebook to produce a vector representation. Let $\mathbf{U} = \{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_N}\}$ be a set of local descriptors in a given image, and let $\mathbf{C} = \{\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_K}\}$ be pre-learned k-means codebook. Here, codebook $\mathbf{C}$ is obtained in a preprocessing step, where its clusters are coarse and K (or $k_c$) is typically 64 or 128.

VLAD is a commonly used method for image description [20,21]. Herein, each local descriptor $\mathbf{u}$ is first hard-assigned to its nearest centroid (or visual word) $\mathbf{c_i}$, and the residual vector $\mathbf{u} - \mathbf{c_i}$ is computed. More formally, the mapping of the 128-D descriptor $\mathbf{u}$ to the high dimensional vector $t(\mathbf{u})$ can be written as:

$$\begin{aligned} t(\mathbf{u}) &= [t_1(\mathbf{u}), \ldots, t_K(\mathbf{u})] \\ t_k(\mathbf{u}) &= \begin{cases} \mathbf{u} - \mathbf{c_k} & \text{if } k = \text{argmin}_j \|\mathbf{u} - \mathbf{c_j}\| \\ \vec{0} & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

Then, VLAD is aggregated over all mapped features as follows:

$$T(\mathbf{u}) = \sum_{i=1}^{N} t(\mathbf{u_i}) \tag{2}$$

For example, the codebook size is K = 64, and the resulting dimension of VLAD is 8192 for 128-D SIFT descriptors. However, as a consequence of hard-assigning each local descriptor into one visual word, error occurs called lost in quantization. This error can be caused by variability of local features, such as image noise, varying scene illumination, and instability in the feature extraction step. Because local features are unstable under changing environments, wrong assignments may occur. Furthermore, the number of descriptors in each image is much greater than the number of visual words ($N \gg K$), so there is high possibility of assignment error. This is the main factor that reduces the distinctness of VLAD because each wrong assignment (or mapping) of a local feature is considered as one noise addition at an aggregation step, where we sum all mapped vectors. Consequently, the search accuracy may be decreased significantly.

To overcome the limitation of hard-assignment, in this paper, we employ soft-assignment techniques with simple computation to produce the global vector representation called soft-VLAD. Different from the hard-assignment of VLAD, soft-VLAD assigns each descriptor to multiple visual words. In this way, we can avoid the quantization loss of local features under changing environment, as well as reducing the noise addition of wrong assignments. Our approach can be reformulated similar to VLAD, where each descriptor $\mathbf{u}$ is mapped to a high-dimensional vector, as follows:

$$\begin{aligned} t(\mathbf{u}) &= [\omega_1 t_1(\mathbf{u}), \ldots, \omega_K t_K(\mathbf{u})] \\ t_k(\mathbf{u}) &= \begin{cases} \mathbf{u} - \mathbf{c_k} & \text{if descriptor link to centroid } c_k \\ \vec{0} & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where $\omega$ is the weight of assignment between a descriptor and the assigned centroid. As shown in Eq. (3), vector $t(\mathbf{u})$ has K 128-D slots, and each slot corresponds to a visual word, like VLAD. However, a feature $\mathbf{u}$ is mapped to more than one slot, and the residual vector $\mathbf{u} - \mathbf{c_i}$ is computed with relevant weight $\omega_i$ in each slot. Then, to form the vector representation $\mathbf{y}$ for the whole image, all mapped vectors are aggregated together via two possible methods:

$$\text{Sum-aggregation:} \quad \mathbf{y} = T(\mathbf{u}) = \sum_{i=1}^{N} t(\mathbf{u_i}) \tag{4}$$

$$\begin{aligned} \text{Max-aggregation:} \quad \mathbf{y}(dim) &= \max\{\mathbf{g_i}(dim) | i = 1, \ldots, N\} \\ &\text{where} \quad \mathbf{g_i} = t(\mathbf{u_i}), \quad dim = 1, \ldots, 128 \times K \end{aligned} \tag{5}$$

We can see that the key enhancement of our approach is obtained by soft-assignment and weighted combinations of visual words. This leads to the problem of how to assign descriptors with appropriate visual words and guarantee avoiding the drawbacks of VLAD. Additionally, we must determine how to efficiently compute the weight for each assignment. In this work, we propose two approaches to solve these problems, as follows.

*The distance-based assignment approach.* called *soft distance-based assignment* VLAD (or sd-VLAD) is motivated by Philbin et al. [43], where each local descriptor in a given image is assigned to its $m$ nearest visual words. In other words, the assignment is established if a visual word is among the $m$ nearest centroids of a descriptor. Then, the weight for each assignment is computed based on the distance between the descriptor and the visual word, as below:

$$\omega_i = e^{-\frac{\|\mathbf{u} - \mathbf{c_i}\|^2}{2\delta^2}} \tag{6}$$

Here, $\delta$ is the spatial scale to guarantee that significant weights are only assigned to a small number of visual words. The property of distance-based assignment is that the closer a visual word is, the bigger its weight is. Then, it can reduce significantly the possibility of wrong assignments caused by quantization error. The parameters for this approach are the number of nearest visual words $m$.

*The sparse-coding-based assignment approach.* called *soft sparse-coding-based assignment* VLAD (or ss-VLAD), has the objective of assignment to minimize the distance from each local descriptor to its weighted combination of visual words by using the sparse coding technique. We can see that VLAD hard-assignment is a special case of this approach. Subsequently,

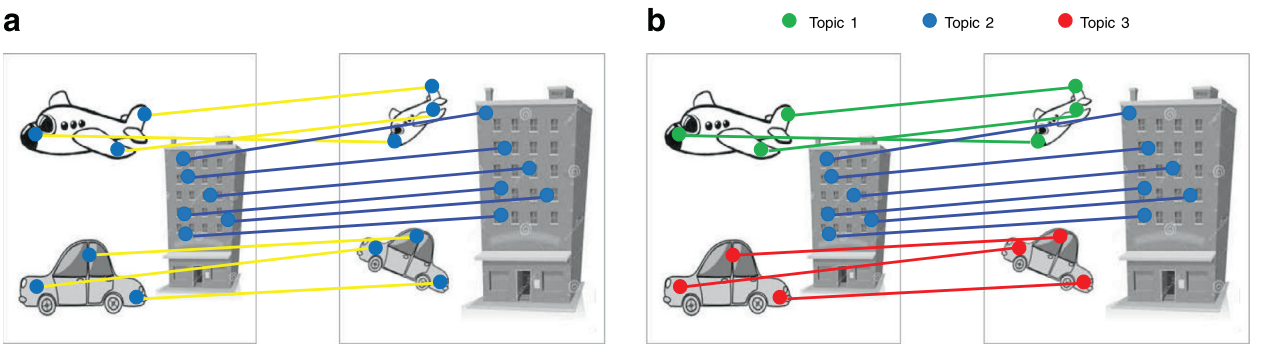


**Fig. 3.** Toy example to compare the GV and proposed method. (a) The global transform of GV reflects the variation of "the building" because the feature matches of the building dominate the feature matches of the other objects. Then, these matches are considered as true matches (shown in blue), and the remaining ones are treated as false matches (shown in yellow), even though they contain the correct matches of other objects (e.g., plane, car). (b) For local geometric consistency, the topic label is first assigned to the features associated with the same object. Subsequently, we proceed to estimate the geometric consistency in each local topic region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with a given codebook $\mathbf{C}$, the weight of the assignments between descriptor $\mathbf{u}$ and its visual words is computed by solving the following optimization problem:

$$\begin{aligned} &\min_{\omega} \|\mathbf{u} - \boldsymbol{\omega}\mathbf{C}\|^2 + \lambda\|\boldsymbol{\omega}\|_1 \\ &\quad \text{subject to} \quad \boldsymbol{\omega} \succeq 0 \end{aligned} \tag{7}$$

Here, vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ is a cluster membership indicator, and each element $\omega_i$ is the weight of the $i$th assignment. The second term is L1 norm regularization to guarantee sparsity of the vector $\boldsymbol{\omega}$ such that it only has a small number of nonzero elements. Formulation of Eq. (7), known as the Lasso problem, can be solved very effectively by using sparse coding algorithms. In our work, we use an accelerated proximal method, FISTA [4], to handle such objective functions.

*Discussion:* One can observe that both approaches assign local descriptors to small numbers of clusters. However, whereas the distance-based approach assigns each local feature to the nearest similar visual words, the sparse-coding-based approach selects fairly different visual words for the best description of the image patch. The former has simpler computation because it only performs a nearest neighbor search for each feature, whereas the latter is more accurate because it represents a local feature as a combination of visual words. Thus, using soft-assignment may not only avoid quantization error but also preserve the important properties of image patches. Furthermore, the dimension of the aggregated vector $\mathbf{y}$ is high (e.g. 8192) for the indexes memory usage. Hence, we use PCA to reduce the dimension of $\mathbf{y}$ and produce a low-dimensional and compact vector, which can be applied directly in many efficient indexing techniques (e.g., LSH [12,25], kd-tree [35], or product quantization [19]). In this work, we employ product quantization for indexing and creating an initial list of candidates, because this method has been shown superior performance in both memory and accuracy [21].

### 3.3. A probabilistic topic model for extracting object regions and re-ranking images.

In this section, we discuss existing limitations of popular re-ranking methods for refining retrieval results in practice. These limitations motivate us to develop a new probabilistic topic model to extract object regions from an image. Taking advantage of topic modeling allows us to address all the problems in similarity measurement.

To estimate the geometric transform between the query and target images, GV methods are widely adopted by employing robust regression techniques, such as RANSAC [13] or Hough [31]. The target images are then re-ranked based on the number of inliers by using estimated transforms. In other words, the goal of GV is to find a global transformation for a change in viewpoint between two images. Due to the iterative process of transform estimation, GV algorithms tends to be computationally expensive and so limit the list of candidate images to small numbers. For a large database of images with complex contents, we usually need to perform on a candidate list that is large enough for reliable re-ranking and to not miss true positives. But in this case, GV is impractical because it becomes very time-consuming. The other disadvantage of GV is that its accuracy strongly depends on the global transform between the query and target images. More specifically, two images should contain one common object, or the viewpoint of the objects in a target image is changed consistently with the viewpoint of the objects in a query image. It therefore fails when the viewpoint of two images containing more than one common object vary inconsistently. Hence, GV is infeasible for estimating local consistency, as shown in the toy example of Fig. 3(a).

To cope with local consistency, in each image, we first need to group features into clusters or segments, where each corresponds to an object, part of an object, or background region of the image. We then establish pairwise matching of local descriptors where they belong to the same object, as shown in Fig. 3(b). Subsequently, we estimate the geometric consistency locally upon pairwise matching of common regions. Consequently, this leads to the problem of how to effectively

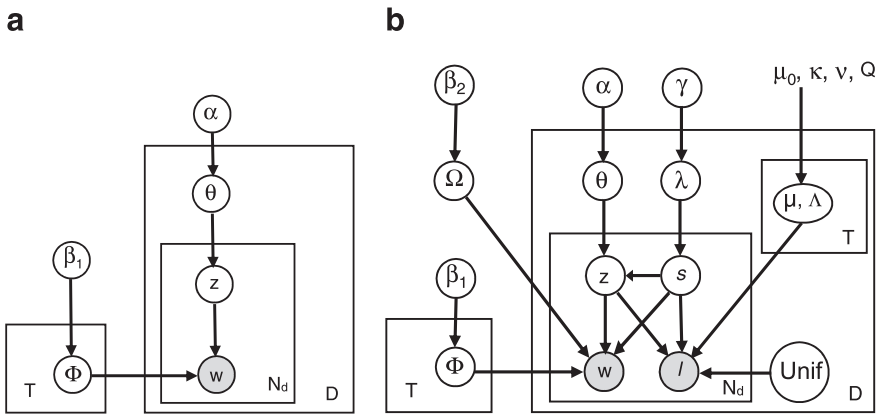


**Fig. 4.** Probabilistic topic model: (a) Latent Dirichlet allocation, (b) Spatial Latent Topic Model with Background Distribution. Gray circles represent observable variables, and plates represent replications.

**Table 1**
The notation of variables and parameters in SLTMB.

| Variable/parameters | Type | Description |
|---|---|---|
| $\alpha$, $\beta_1$, $\beta_2$, $\gamma$ | $\mathbb{R}^+$ | Dirichlet hyper parameters |
| $\mu_0$, $\kappa$, $\nu$, $Q$ | $\kappa, \nu \in \mathbb{R}^+$ $\mu_0 \in \mathbb{R}^2$, $Q \in \mathbb{R}^{2\times 2}$ | The parameters of Normal-Wishart distribution |
| $w_{di}$ | $\{1, \ldots, W\}$ | Identity of $i$th visual word $(i = 1, \ldots, N_d)$ in image $d (d = 1, \ldots, D)$ |
| $z_{di}$ | $\{1, \ldots, T\}$ | Identity of topic of $i$th visual word $(i = 1, \ldots, N_d)$ in image $d$ |
| $l_{di}$ | $\mathbb{R}^2$ | Coordinates of location of visual word $w_{di}$ |
| $\{\mu_{td}, \Lambda_{td}\}$ | $\mu_{td} \in \mathbb{R}^2$, $\Lambda_{td} \in \mathbb{R}^{2\times 2}$ | The parameters of Gaussian distribution for topic $t (t = 1, \ldots, T)$ in image $d$ |
| $s_{di}$ | $\{1, 2\}$ | Switch variable |
| $\lambda_d$ | $\mathbb{R}^2$ | Parameter of Bernoulli distribution of word type (i.e., topic and background) in image $d$ |
| $\Omega$ | $\mathbb{R}^W$ | Parameter of Mult distribution of words in background |
| $\theta_d$ | $\mathbb{R}^T$ | Parameter of Mult distribution of topics in image $d$ |
| $\phi_t$ | $\mathbb{R}^W$ | Parameter of Mult distribution of words in topic $t$ |

group features of a specific object into a cluster. To tackle this problem, we propose a probabilistic topic models in the topic modeling phase to extract the topics from the image, where each topic corresponds to an object instance occurring frequently in the image collection. In other words, local descriptors belonging to different objects will be indicated by different topic values. By employing the topic model, we can preserve object-related information that is lost during the encoding phase. Furthermore, the topic modeling phase in conjunction with the encoding phase enables us to develop an efficient re-ranking scheme to cope with large variations of object appearance. Also, using the topic model, we are able to gain insight into the spatial arrangement and relationships between objects and background. We describe our proposed topic model in more detail in the next subsections.

### *3.3.1. Spatial Latent Topic Model with Background Distribution (SLTMB)*

In topic modeling, we can treat an image as a document and directly apply topic models such as LDA [5] and pLSA [16] to learn or infer common topics. Especially, LDA (as shown in Fig. 4(a)) and its variants have been widely applied to solve visual recognition problems. However, the limitation of naively applying topic models is that important spatial structures among visual words are ignored. Unlike conventional topic models, we incorporate explicit notions of topic and visual word locations into our model. Furthermore, inspired by Chemudugunta and Steyvers [7], our topic model also considers the background distribution of visual words because we need to remove visual words belonging to background regions. In a retrieval task, this type of visual word, which appear in almost all images, is meaningless for similarity measurement. Otherwise, the vocabulary used for our topic model is obtained from a refined codebook, as described in Section 3.1.

Our proposed topic model is illustrated by the graphical model shown in Fig. 4(b). We can see that our model is an extension of LDA by incorporating spatial relation and background information. The SLTMB model is learned in an unsupervised manner. We summarize a formal description of SLTMB in Table 1.

The SLTMB model represents a collection of D images, and each image $I_d$ consists of $N_d$ image patches (i.e. local features). Each patch is described by its appearance $w_{di}$ and location $l_{di}$. Then, patches are formalized as groups of regions that are spatially clustered and have predictable appearances. We use latent variables (i.e., $z_{di}$) to characterize the topics, where topic $z$ for each patch indicates which objects or parts of object it comes from. More specifically, we factor the image into a combination of T topics. Each topic is modeled as a distribution over the appearance vocabulary of size W and over the spatial location of these words within each image. According to the graphical model, $w_{di}$ and $l_{di}$ are the only observable

variables, and the other variables are latent variables. Formally, the generative process of our SLTMB model for image corpus is as follows:

1. For each topic $t$: Draw an appearance distribution $\phi_t \sim Dir(\beta_1)$
2. Draw background distribution $\Omega \sim \text{Dir}(\beta_2)$
3. For each image $I_d (d = 1, \ldots, D)$:
   (a) Draw word type distribution $\lambda_d \sim \text{Beta}(\gamma)$
   (b) Draw topic proportion $\theta_d \sim \text{Dir}(\alpha)$
   (c) For each topic $t (t = 1, \ldots, T)$: draw a location distribution: $\{\mu_{td}, \Lambda_{td}\} \sim NW(\mu_0, \kappa, \nu, Q)$
4. For each word position $di$ where $i \in 1, 2, \ldots, N_d$:
   (a) Draw switch sample $s_{di} \sim \text{Bernoulli}(\lambda_d)$
   (b) if $s_{di} = 1$
      i. Draw topic $z_{di} \sim \text{Multi}(\theta_d)$
      ii. Draw visual word $w_{di} \sim \text{Multi}(\phi_{z_{di}})$
      iii. Draw word location $l_{di} \sim N(\mu_{dz_{di}}, \Lambda^{-1}_{dz_{di}})$
   (c) if $s_{di} = 2$
      i. Draw visual word $w_{di} \sim \text{Multi}(\Omega)$
      ii. Draw word location $l_{di} \sim \text{Uniform}$

Here, Dir, Multi, N, and NW respectively denote Dirichlet, Multinomial, Normal, and Normal-Wishart distributions. The priors including Multi and NW are chosen to conjugate to Dir and N for the word and location distributions, and hence, they simplify computation and guarantee efficient inference.

We can see that switch variable $s$ is used to control the generation of the visual word. An image contains two types of visual words, where one is generated from topic distribution Multi($\Phi_z$) and the other is generated from background distribution Multi($\Omega$).

An important characteristic of our probabilistic graphical model is that human cognition can be formalized. As described in generative process, each topic $t$ in image $d$ has a probability distribution $p(\mu_{td}, \Lambda_{td})$ of locations and a probability distribution $\Phi_t$ of visual words. The distribution of locations $p(\mu_{td}, \Lambda_{td})$ is not shared among images, whereas the distribution of visual words $\Phi_t$ are shared across images. This is because we assume that the appearance of an object captured by the $\Phi_t$ distribution is similar in all images. In contrast, the location of an object in a specific image can be assumed to be independent of the location in other images. For instance, a building can appear in various image locations. However, its appearance, which is described by visual words, is the same in all images. Similar to topic distribution, background distribution is also shared across images, but its parameter $\Omega$ is identical in the entire corpus. Moreover, objects are typically compact, and solid while backgrounds tend to spread across the image. This knowledge is formalized via the Gaussian spatial distributions for topics and the uniform distribution for background. Thus, objects emerge from the background and, therefore, correspond to the notion of saliency across all images in the corpus. With parameter estimation and inference algorithms, which we will describe in the next subsections, we can determine where the patches are spatially arranged for a particular topic.

### 3.3.2. Parameter estimation in SLTMB

In this subsection, we describe a method for parameter estimation in the SLTMB model where we will use an independent training set of D images. Let $\Pi = \{\alpha, \beta_1, \beta_2, \gamma, \mu_0, \kappa, \nu, Q\}$ be the set of model parameters. Given a corpus of image data $\{\mathbf{w_d}, \mathbf{l_d}\}_{d=1}^{D}$, the parameters $\Phi$ and $\Omega$ of topic and background distributions respectively, can be found by maximization of the following log likelihood function.

$$L(\Phi, \Omega) = \sum_{d=1}^{D} \log\left(p(\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_d}, \mathbf{s_d} | \Phi, \Omega, \Pi)\right) \tag{8}$$

The distribution in Eq. (8) is intractable to direct estimation, so one effective approach is to estimate using Monte Carlo EM algorithms [1], as summarized in Algorithm 1.

**Algorithm 1** Parameter estimation of SLTMB.

**Input:** Corpus of image data formed as a bag of visual words and locations $\{\mathbf{w_d}, \mathbf{l_d}\}_{d=1}^{D}$
**Output:** The estimated parameters $\Phi$ and $\Omega$

1. **Initialization**. Initialize set of parameters $\left\{\Phi^{(0)}, \Omega^{(0)}\right\}$
2. **For each** $k = 1, \ldots, K$ **do**:
   (a) Given $\left\{\Phi^{(k-1)}, \Omega^{(k-1)}\right\}$, sample latent variables with N Gibbs steps for each image $I_d$ from the posterior distribution $p(\mathbf{z_d}, \mathbf{s_d} | \mathbf{w_d}, \mathbf{l_d}, \Phi^{(k-1)}, \Omega^{(k-1)}, \Pi)$ using Eqs. (9) and (10).
   (b) Compute $\left\{\Phi^{(k)}, \Omega^{(k)}\right\}$ using as Eqs. (15) and (16).
3. **End**

As shown in Algorithm 1, rather than directly computing the posterior of latent variables, which is intractable, we draw samples from it. Then, the parameters are estimated by examining this posterior distribution. Here, we use the collapse Gibbs sampling algorithm [15] for joint sampling of latent variables $z$ and $s$ for each visual word $w$, as in the following equations:

$$p(z_{di}=t, s_{di}=1|\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{d1,-di}+\gamma}{N_{d,-di}+2\gamma} \times \frac{n^{WT}_{wt,-di}+\beta_1}{\sum_{w'} n^{WT}_{w't,-di}+W\beta_1} \times \frac{n^{TD}_{td,-di}+\alpha}{\sum_{t'} n^{TD}_{t'd,-di}+T\alpha} \times t_{\nu^{TD}_{td,-di}-q+1}\left(\mu^{TD}_{0,td,-di}, \frac{Q^{TD}_{td,-di}\left(\kappa^{TD}_{td,-di}+1\right)}{\kappa^{TD}_{td,-di}\left(\nu^{TD}_{td,-di}-q+1\right)}\right) \tag{9}$$

$$p(s_{di}=2|\mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{d2,-di}+\gamma}{N_{d,-di}+2\gamma} \times \frac{n^{W}_{w,-di}+\beta_2}{\sum_{w'} n^{W}_{w',-di}+W\beta_2} \tag{10}$$

where the subscript *-di* indicates whole variables excluding the $i$th variable in image $d$. $N_{d1}$ and $N_{d2}$ are the numbers of visual words in image $d$ assigned to the related topic and background words, respectively; $n^{TD}_{td}$ is the number of words assigned to topic $t$ in image $d$; $n^{WT}_{wt}$ is the number of times word $w$ is assigned to topic $t$; $n^{W}_{w}$ is the number of times word $w$ is assigned to the background words distribution in the image corpus. The last term of Eq. (9) denotes the t-student distribution with the corresponding parameters computed as follows:

$$\mu^{TD}_{0,td,-di} = \frac{\kappa\mu_0 + n^{TD}_{td,-di}\bar{l}_{td,-di}}{\kappa + n^{TD}_{td,-di}} \tag{11}$$

$$A = \sum_{i=1}^{n^{TD}_{td,-di}} \left(l_{di}-\bar{l}_{td,-di}\right)\left(l_{di}-\bar{l}_{td,-di}\right)^T$$
$$\nu^{TD}_{td,-di} = \nu + n^{TD}_{td,-di}; \kappa^{TD}_{td,-di} = \kappa + n^{TD}_{td,-di} \tag{12}$$

$$Q^{TD}_{td,-di} = Q + A + \frac{\kappa n^{TD}_{td,-di}}{\kappa + n^{TD}_{td,-di}}\left(\mu_0-\bar{l}_{td,-di}\right)\left(\mu_0-\bar{l}_{td,-di}\right)^T \tag{13}$$

Here, we skip certain details of the derivation due to space. The reader may refer to the Appendix for details. Since all latent variables are computed from sampling equations, parameters $\Phi$, $\Omega$ are then estimated by examining posterior distributions. Following some iterative steps, the parameters will converge to $\Phi^*$, $\Omega^*$. The posterior of the topic-word multinomial is computed as below:

$$p(\Phi_t|\mathbf{w}, \mathbf{z}, \mathbf{s}) = \mathrm{Dir}\left\{\beta_1 + n^{WT}_{wt}\right\} \tag{14}$$

where $\mathbf{w} = \{\mathbf{w_d}\}_{d=1}^{D}$, $\mathbf{z} = \{\mathbf{z_d}\}_{d=1}^{D}$, $\mathbf{s} = \{\mathbf{s_d}\}_{d=1}^{D}$. Thus, $\Phi$ can be estimated as the posterior mean of $p(\Phi_t|\mathbf{w}, \mathbf{z}, \mathbf{s})$, which is simply the normalized Dirichlet parameters, as follows:

$$\Phi_t = \frac{n^{WT}_{wt}+\beta_1}{\sum_{w'} n^{WT}_{w't}+W\beta_1} \tag{15}$$

Similarly, we can estimate $\Omega$ of the background distribution as the posterior mean of $p(\Omega|\mathbf{w}, \mathbf{z}, \mathbf{s})$, as follows:

$$\Omega = \frac{n^{W}_{w}+\beta_2}{\sum_{w'} n^{W}_{w'}+W\beta_2} \tag{16}$$

*Observations*: In Eqs. (9) and (10), the first term indicates the ratio of words assigned to the topic distribution and the background distribution. We also observe that the second and third terms of Eq. (9), which represent the probability of word $w_{di}$ under topic $t$ and the probability of topic $t$ in the image, cluster visual words often co-occurring in the same image into one object. Furthermore, the last term of Eq. (9) shows that a word tends to have the same topic model as other words in the image if it is closer to a particular topic location. Consequently, it forces visual words with a consistent distribution to be grouped together. The multinomials $\Phi_t$ forces objects in different images to have the same visual words statistics. This agrees with our prior knowledge that visual words from the same object are consistent with the distributions of their appearance and location. This also enables us to combine object appearance statistics across multiple images based on the distribution of the visual words they contain.

*3.3.3. Inference of unseen image*

An unseen image (e.g. a query, database image) can be applied to our model. With known parameters (i.e. $\Phi$, $\Omega$) obtained from the training process, we infer the latent variables of the unseen image, such as the topic label $z_{di}$ and parameters $\mu_{td}$, $\Lambda_{td}$ of the topic location. By using SLTMB, the inference algorithm is similar to the estimation. However, the second terms in Eqs. (9) and (10) will be fixed and replaced by $\Phi_t$ and $\Omega_t$, respectively. The update equations of the Gibbs sampling algorithm are modified as Eqs. (17) and (18). This reflects the fact that all the learned object and background could be present without any prior knowledge.

$$p(z_{di}=t, s_{di}=1|\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{d1,-di}+\gamma}{N_{d,-di}+2\gamma} \times \Phi_t \times \frac{n^{TD}_{td,-di}+\alpha}{\sum_{t'} n^{TD}_{t'd,-di}+T\alpha} \times t_{\nu^{TD}_{td,-di}-q+1}\left(\mu^{TD}_{0,td,-di}, \frac{Q^{TD}_{td,-di}\left(\kappa^{TD}_{td,-di}\right)}{\kappa^{TD}_{td,-di}\left(\nu^{TD}_{td,-di}-q+1\right)}\right) \tag{17}$$

$$p(s_{di}=2|\mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{d2,-di}+\gamma}{N_{d,-di}+2\gamma} \times \Omega \tag{18}$$

Note that the inference of different images are independent of each other, and the update equation of the Gibbs sampling for inference can be factored into the terms that only depend on variables related to a single image. Therefore, we can distribute images to multiple machines and process them in parallel. Hence, there is no issue with the scalability of our SLTMB model.

*3.4. Local geometric consistency for similarity measure*

As mentioned in Section 3.3, the re-ranking stage based on GV algorithms tends to be computationally expensive and fails to deal with local variations of object appearance. Consequently, we proceed to estimate the local geometric consistency according to the matching of common topic regions between two images. Due to the high computation cost of the original geometric transformation [42], we use weak geometric verification motivated by Jégou et al. [18] to estimate local consistency for each topic matching. Different from [18], where they estimate the global consistency and integrate it into an inverted file, we utilize this method to speed up the re-ranking stage by employing a simple similarity scheme.

To evaluate the consistency between two images, we use orientation and scale information obtained from the SIFT detector [31]. If we assume that the change in viewpoint of topic regions between the query and database images are consistent, then matching feature pairs should have consistent differences of orientation and scale. For that reason, our scheme for the similarity measure is to validate the consistency of these differences between matching pairs. To represent the local geometric consistency between two images, we build a 2D histogram for each matching topic region, where the $x-$ and $y-$axes of the histogram correspond to the orientation and log scale differences. Since we set T as the number of topics for the image corpus, each image is then divided into T topic regions along with the background region. Therefore, we need to estimate T histograms for each image in the candidate list when matching them with a given query image.

As described in Section 3.1, the query image and the candidate image $k$ are formed as $\mathbf{I_q} = \left\{w_{qi}, l_{qi}, \xi_{qi}, o_{qi}\right\}_{i=1}^{N_q}$ and $\mathbf{I_k} = \left\{w_{kj}, l_{kj}, \xi_{kj}, o_{kj}\right\}_{j=1}^{N_k}$, respectively. Note that word location has been integrated into the topic model to extract the object regions. Then, in this step, we use the remaining geometric information for the similarity measure. Let $M_t$ be the matching feature pairs within topic $t$. Subsequently, the histogram of the $t$th topic region is formed as follows:

$$M_t = \left\{ \begin{cases} \Delta^{\xi}_{t,ij} = \log\left(\frac{\xi_{qi}}{\xi_{kj}}\right), \\ \Delta^{o}_{t,ij} = (o_{qi} - o_{kj}) \end{cases} \middle| \begin{array}{c} z_{qi} = z_{kj} = t, w_{qi} = w_{kj} \\ i \in \{1, \ldots, N_q\}, j \in \{1, \ldots, N_k\} \end{array} \right\} \tag{19}$$

$$h^k_t(\rho_x, \rho_y) = \sum_{(\Delta^{\xi}_{t,ij}, \Delta^{o}_{t,ij}) \in M_t} \mathbb{I}\left(\lfloor \varepsilon_x \Delta^{\xi}_{t,ij} \rfloor = \rho_x, \lfloor \varepsilon_y \Delta^{\varnothing}_{t,ij} \rfloor = \rho_y\right) \tag{20}$$

where $\mathbb{I}(.,.)$ is the indicator function. The 2D histogram bin index is represented as $(\rho_x, \rho_y)$. $\varepsilon_x$, $\varepsilon_y$ are the factors to adjust the bin width of the $x$- and $y$-axes, respectively. The geometric score $S_k$ between the image $k$ of the candidate list and the query image is then given by:

$$\left(\rho^t_x, \rho^t_y\right) = \underset{\rho_x, \rho_y}{\operatorname{argmax}}\, h^t_k(\rho_x, \rho_y) \tag{21}$$

$$S_{k,t} = \sum_{\substack{(i,j) \\ (\Delta^{\xi}_{t,ij}, \Delta^{o}_{t,ij}) \in M_t \\ \lfloor \varepsilon_x \Delta^{\xi}_{t,ij} \rfloor = \rho^t_x, \lfloor \varepsilon_y \Delta^{\varnothing}_{t,ij} \rfloor = \rho^t_y}} \frac{1}{\sqrt{n^{WT}_{w^*_{qi}t}}} \frac{1}{\sqrt{n^{WT}_{w^*_{kj}t}}} \tag{22}$$

$$S_k = \sum_t S_{k,t} \tag{23}$$

Here, $S_{k,t}$ is the matching score of the $t$th topic region; $n^{WT}_{w^*_{qi},t}, n^{WT}_{w^*_{kj},t}$ correspond to the number of times that visual word $w^*_{qi} = w^*_{kj}$ is assigned to topic $t$ in the query and candidate images, respectively. In Eq. (22), the two denominators are the term frequencies of visual words $w^*_{qi}, w^*_{kj}$ assigned to topic $t$ of the query and candidate images. This term is used to eliminate the visual words repeatedly occurring in the same topic of the image and so avoid the "burstiness phenomenon" that usually corrupts the similarity measure, as described in [17].

Using this scheme enables us to perform similarity measurement very quickly and much faster than the original GV, due to its simplicity. Furthermore, it is reasonable to expect that SLTMB and the local geometric consistency contain complementary information. While our topic model uses the locations of local features to indicate the relationship and arrangement among objects, our measuring scheme employs orientation and scale information of the local feature to evaluate the local consistency between two images via their common topic regions. Thus, our re-ranking stage completely exploits the geometric information of the SIFT feature in the visual search system and overcomes the limitation of recent works, which fail to capture enough information.

## 4. Experimental results

We evaluated the performance of the proposed approaches by comparing them with state-of-the-art methods on public benchmark datasets for image retrieval. To study the scalability of proposed framework, we further conducted experiments on large datasets of millions of images. Our results were reported for the following two tasks: (i) Evaluate the quality of the initial ranking by using soft-VLAD; and (ii) Examine the qualitative and quantitative results of SLTMB for evaluation of the re-ranking performance.

**Datasets**: All results were reported on four benchmarks and extended datasets that were constructed by combining benchmarks with a distractor dataset. The description of each dataset is as below:

- Oxford building [42]: This dataset comprises 5,062 images collected from Flickr by searching 11 particular Oxford landmarks. Each landmark contains 5 different queries, each of which is within an associated bounding box. The dataset also provides ground truth for these 11 different landmarks.
- Paris [43]: This dataset contains 6412 images collected from Flickr by searching particular landmarks in Paris. Similar to the Oxford 5K dataset, most of images in this dataset are buildings.
- Inria Holidays [20]: This is a collection of 1,491 images, each of which represents a scene or object. In each group, the first image can be used as a query, and the other images are the correct retrieval results. The content of the dataset is related to various types, such as natural, man-made, water, and fire effects.
- University of Kentucky Recognition Benchmark (UKB) [36]: This dataset contains 10,200 images of 2,550 objects, each of which is represented by 4 images and taken from different viewpoints.
- Flickr60k [18]: This is a set of 60k images collected from Flickr and made available by the authors of [18]. This dataset is used to learn the PCA and construct vocabularies.
- Flickr5M: To test our scalability, we randomly downloaded 5M images from Flickr. This dataset is considered as a distractor for large-scale experiments, where the benchmark datasets are then merged with this collection to create a ground truth for extended datasets.

**Experimental setup**: Our visual search system contained the following components: local descriptor extraction, vocabularies, encoding, and topic modeling. To evaluate the performance of the proposed method, our experimental settings for each component are described below:

- *Local descriptors*. We used Difference-of-Gaussian (DoG) [31] to find the salient interest points in the image. The local descriptor was then computed on a patch around the interest point by using the SIFT descriptor presented in [42]. Most of our experiments used the default parameters of the tool.
- *Vocabularies*. For all methods based on vocabularies (or codebooks), we considered only distinct datasets for learning. More precisely, following the common practice, the k-means were learned on Flickr60k. As described in Section 3.1, we employed hierarchical k-means to generate two types of codebooks: a coarse codebook of $k_c$ visual words and a refined codebook of $k_c \times k$ visual words. The sizes of coarse and refined codebooks were set to 64 and 65k, respectively. We relied on the Yael library [18] for codebook construction.
- *Encoding*. This stage had three main parts: aggregation, dimension reduction, and indexing with PQ. For aggregation, we trained and tested with different methods as presented in Section 3.2. In addition, to reduce the influence of peaky components [41], we used intra-normalization [2] for the aggregated vector. For dimension reduction, we also used Flickr60k for learning the transformation matrix by using PCA, and so generate various PCA-reduced dimensions. Finally, we applied PQ with the same setting as in [19] for indexing.
- *Topic modeling*. For our model, we evenly split the data into two sets. The model was then trained on the first set, which was constructed by combining subsets of benchmark datasets, such as Holidays, UKB, Oxford5k, and Paris. We used the second set, containing the remaining subsets of the benchmark datasets, for testing. We ran our model based

**Table 2**
Proposed encoding methods based on soft aggregation in comparison with state-of-the-art methods without dimension reduction.

| Methods | Oxford | Paris | UKB | Holidays |
|---|---|---|---|---|
| BoV | 0.410 | 0.395 | 0.607 | 0.583 |
| HE | 0.527 | 0.518 | 0.714 | 0.691 |
| Fisher | 0.593 | 0.491 | 0.681 | 0.570 |
| VLAD,sum | 0.517 | 0.457 | 0.683 | 0.563 |
| VLAD,max | 0.482 | 0.422 | 0.624 | 0.533 |
| sd-VLAD,sum | 0.533 | 0.482 | 0.688 | 0.571 |
| sd-VLAD,max | 0.495 | 0.432 | 0.667 | 0.541 |
| ss-VLAD,sum | 0.557 | 0.479 | 0.689 | 0.587 |
| ss-VLAD,max | 0.599 | 0.535 | 0.730 | 0.625 |

on the Gibbs sampler for around 100 iterations to guarantee convergence. The parameter settings of STLMB are further discussed in Section 4.2.

**Evaluation criterion**: We used mean average precision (mAP) to quantitatively evaluate the retrieval performance of the competing methods. The retrieval performance of a single query was measured by the average precision (AP), which is the area under the precision recall curve. Subsequently, the mean value over multiple queries was the final measurement of the retrieval performance.

### *4.1. Image encoding with soft-VLAD*

In this section, we investigate the improvement of performance using our proposed encoding methods based on soft aggregation. By using different datasets, we compared our methods with a number of well-known methods including BoV [42], HE [18], VLAD [20,21], and Fisher [41,45].

The first experiment shows the benefit of our soft aggregation when compared to other aggregation-based methods (e.g., VLAD and Fisher) and standard BoV-based methods. As shown in Table 2, the proposed method outperforms the existing aggregation-based methods for all datasets, with ss-VLAD achieving approximately 8% and 4% higher mAP than VLAD and Fisher, respectively. Furthermore, sd-VLAD always has a higher mAP than VLAD, and it is competitive with Fisher. Particularly, a parameter associated with sd-VLAD is the number of nearest neighbor $m$. We set this number to 7 for all experiments as its optimal value. Interestingly, our methods are computationally much cheaper than the Fisher method, which is known as a powerful approach that also employs soft-assignment via GMM with higher complexity. Moreover, we also compared our proposed methods and VLAD using different aggregation strategies including sum- and max-aggregation. As shown in Table 2, VLAD and sd-VLAD both perform best using sum-aggregation, while ss-VLAD performs significantly better with max-aggregation rather than with sum-aggregation. This is because, when employing sparse coding, max-aggregation tends to pick up the distinctive features that are more likely to be repeated, which is an important factor in image retrieval. Therefore, we applied sum-aggregation to VLAD and sd-VLAD, and max-aggregation to ss-VLAD, when we conducted the remaining experiments.

In addition, we compared the performance of our method with that of the standard BoV-based methods including BoV and HE. Note that HE is a complicated version of BoV that integrates geometric information and Hamming embedding into an inverted file to improve retrieval accuracy. For these BoV-based methods, the learned vocabulary had a size of 65k visual words, which is similar to the size of the vocabulary used in our topic modeling phase. As shown in Table 2, we obtain approximately 10% higher mAP than that of the standard BoV, although the codebook size is much smaller than the BoVs vocabulary size. Soft-aggregation also provides performance on the four datasets that is competitive with HE, which requires higher complexity. Notice that all results so far were obtained by using full vector representation.

Because the memory usage is directly proportional to the number of dimensions of the image representation, it is necessary to reduce the dimensions of the representation vector before applying it to indexing stage. Fig. 5 shows the respective performances of the aggregation methods, measured for different dimensions reduced by PCA. We see that dimension reduction reduces the accuracy in most cases. For the Holidays dataset, the accuracy is even improved with some reduced dimensions. We also note that, for 128 dimensions, VLAD, sd-VLAD, and Fisher have small reductions of mAP compared to those of the original dimensions. With the Paris, UKB, and Holidays datasets, ss-VLAD suffers from dimension reduction more than the others. Therefore, the optimal value of the dimensions is 128, which not only provides competitive results but also effectively saves memory usage. Again, ss-VLAD outperforms the other aggregation methods, whereas sd-VLAD performs better than VLAD and is comparable with Fisher for low dimensions.

We further studied the scalability of our approach when conducting experiments on large-scale datasets, as shown in Fig. 6. Note that the representation vectors of the aggregation methods are reduced to 128 dimensions by PCA. Subsequently, reduced-dimensional vectors are indexed by PQ, which has been shown in [21] to achieve good retrieval accuracy with less than a hundred bytes per image. Therefore, the scalability of the aggregation methods is significantly better than that of the BoV-based methods (e.g., BoV and HE): they scale well to 5 million images, whereas BoV-based methods do not scale to more than 1M images due to the high memory usage of the inverted list.

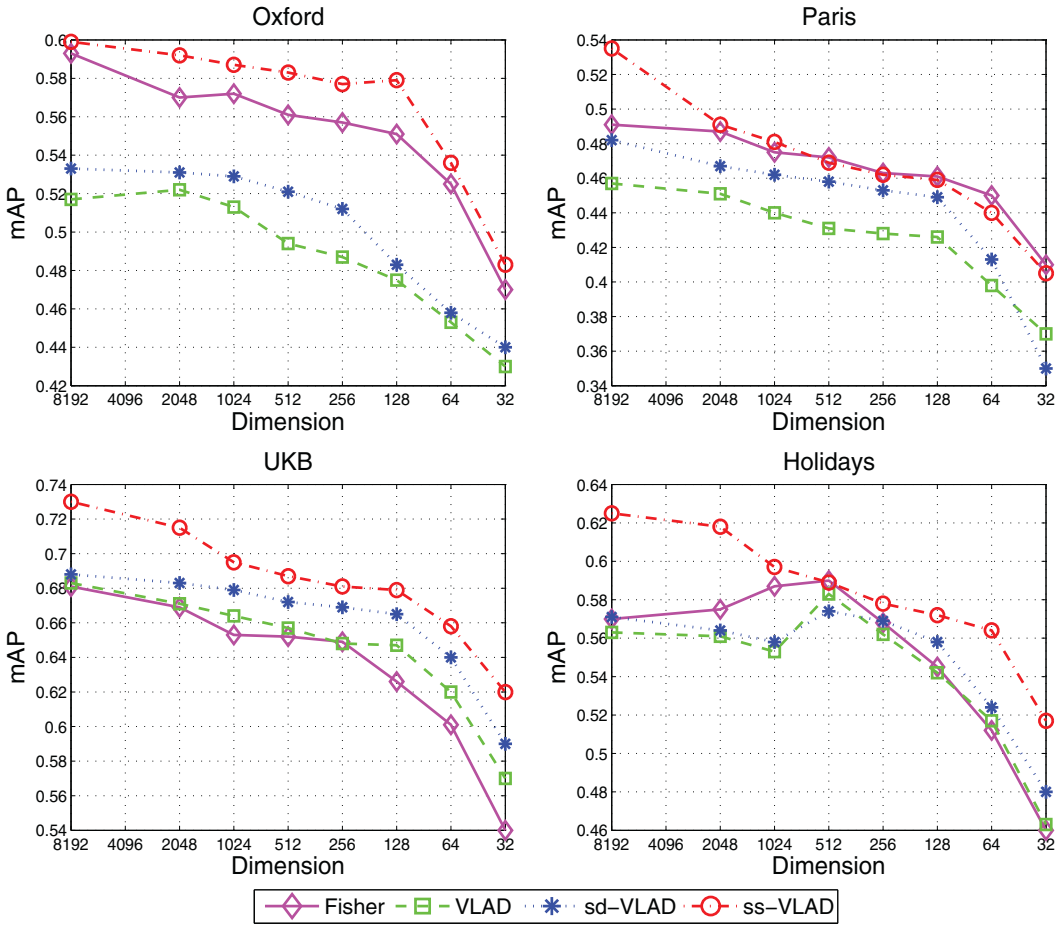


**Fig. 5.** Comparison results of different aggregation-based methods with respect to PCA dimensions.

Fig. 6 also shows the impact of database size on performance. The mAP of ss-VLAD is better than the other aggregation methods for all the datasets with their different sizes. Moreover, soft-aggregation obtains significant gains in performance over standard BoV and also requires less memory. Especially, for the Oxford dataset, ss-VLAD even outperforms the highly complex HE. For the remaining three datasets, HE achieves higher accuracy than the other methods. This is because the integration of geometric information allows HE to deal with a high degree of change in viewpoints. Additionally, the accuracy of the aggregation-based methods is affected by dimension reduction, when it is necessary for their scalability. Particularly, HE outperforms the others by a large margin on Holidays, because this dataset has high variability of the objects and scenes. On Paris and UKB, due to less variability of data contents, mAP differences between HE and the others are reduced significantly. Interestingly, as the database size increases, mAP of ss-VLAD is closer to mAP of HE. Hence, our proposed methods show promising behavior for searching on large-scale datasets. Note that the mAP was measured in this experiment without using geometric verification or a re-ranking scheme. The performance could be greatly improved by incorporating geometric information.

### *4.2. SLTMB model for extracting topic regions*

In our SLTMB model, the free parameter (T) is the number of topics. To select this number empirically, we examined the effect of this parameter on the likelihood of the SLTMB model. In this study, log-likelihood is used as a standard criterion [15] for evaluation of topic learning because it provides a quantitative measurement to reflect the fitting of the topic model with given training data. The higher score of log-likelihood is, the better generative model fits. As shown in Fig. 7, we achieve the maximum likelihood with 180 topics, and after that, the likelihood slightly decreases with increasing T. Note that the training data for this experiment was collected from subsets of each benchmark, and for each number of topics we then ran 100 iterations of the Gibbs sampler to compute the likelihood of the probabilistic model. Generally speaking, we can obtain good performance with a high rather than a low number of topics because the performance depends on the variety of content in the images in the dataset. In our work, we chose $T = 180$ as the optimal number of topics for our remaining experiments.

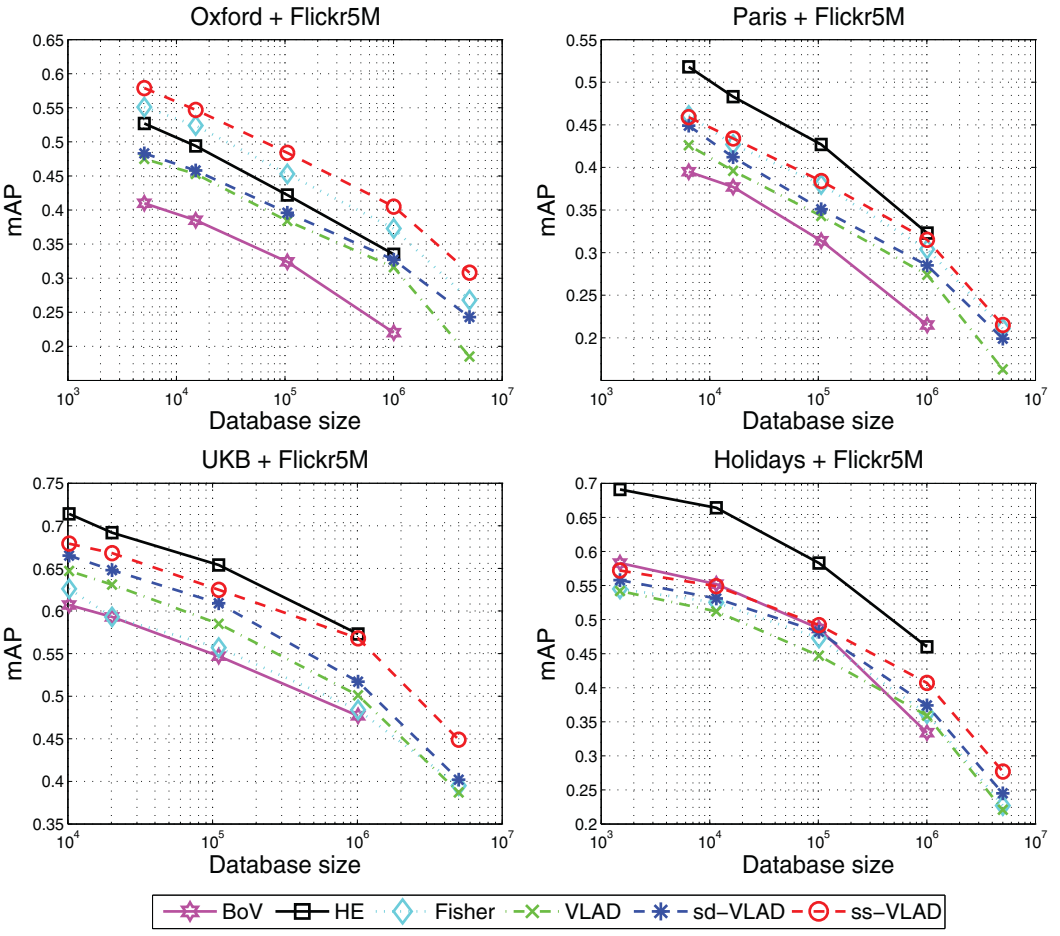


**Fig. 6.** Comparison results of proposed methods with state-of-the-art methods with respect to data size. (Note: aggregation-based methods use PCA 128-D).

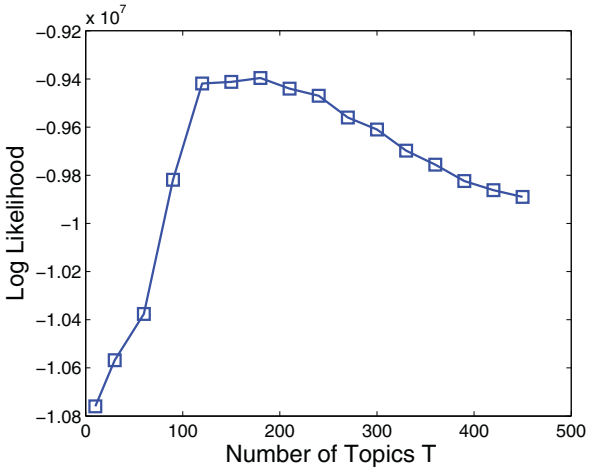


**Fig. 7.** Impact of the number of topics.

The qualitative results of the topic models are shown in Fig. 8, where we compared our proposed SLTMB model with the standard LDA. In the example, the four images of the object are taken from different viewpoints. Each image contains a set of patches extracted by the affine invariant detector to represent the local features. Besides the keypoint-based detector of the SIFT feature, we further conducted experiments with the region-based detector of the MSER feature. This type of feature has been successfully applied to several visual topic models [46,47]. Subsequently, the patches are represented as the circles for SIFT features at different scales and oriented along the dominant gradient (Fig. 8(a) and (b)). Otherwise, to guarantee the

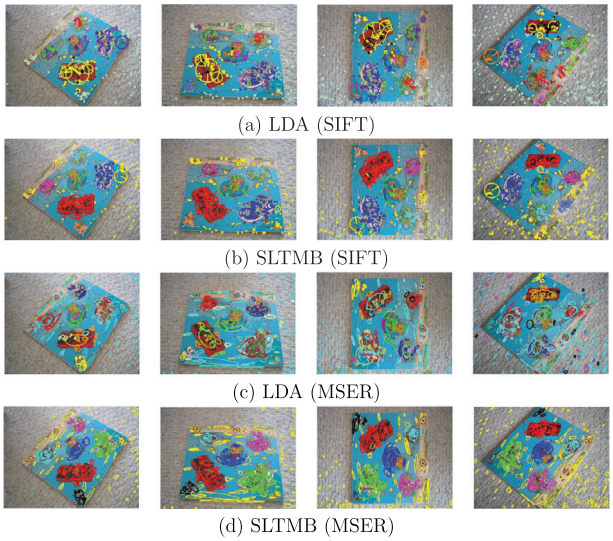


**Fig. 8.** Comparing the qualitative results of LDA and SLTMB by using SIFT (Fig. 8(a) and (b)) and MSER (Fig. 8(c) and (d)) features.

good description for MSER features, the extracted patches are fitted into elliptical regions as shown in Fig. 8(c) and (d). Each patch then denotes a visual word produced by a quantizer, as presented in Section 3.1. In these qualitative examples, each topic corresponds to a part of the object, or we consider it as a sub-object. Therefore, the visual words associated with a sub-object should be assigned to the same topic or labeled in the same color. In particular, the background words extracted by the SLTMB model are labeled in yellow, as shown in Fig. 8(b) and (d). The background words usually appear in a large number of images in the corpus. For example, in this example, the background is related to the "carpet" object of the UKB dataset.

As shown in Fig. 8(a), the results of LDA with SIFT features are quite noisy in most cases, where visual words associated with a specific sub-object are labeled to more than one topic. Moreover, the visual words related to the carpet object dominate and interfere with other topic regions in the image. We also obtain the similar results by using LDA with MSER features as shown in Fig. 8(c), where they are noisy and the sub-objects are interfered by other topic regions. In contrast, by incorporating spatial information and learning with the background distribution, SLTMB obtains much better results than the standard LDA with both SIFT and MSER features. As shown in Fig. 8(b) and (d), visual words having consistent spatial distributions that are close together are usually assigned to the same topic. There is very limited noise in each topic region, and visual words associated with a sub-object are quite accurately labeled in the same color across the four images. Furthermore, the background words are well extracted and tend to spread across the image; these words are meaningless to the similarity measure and will be removed for remaining steps of re-ranking. Thus, SLTMB not only effectively extracts the latent topics from the images, but also localizes the topic (e.g., an object or a part of an object) in each image. Furthermore, different types of features can be applied to our topic model and gain the better topic extraction than the LDA model.

Another benefit of our approach is that the convergence of SLTMB in estimating model parameters is much faster than for LDA because we integrated the spatial constraints of the visual words and the background word distribution into the Gibbs sampler, as described in Section 3.3. Usually, LDA needs to run more than 100 iterations for convergence, whereas SLTMB runs fewer than 60 to 70 iterations. In other words, the Gibbs sampler of SLTMB converges approximately two time faster than does LDA.

### *4.3. Performance comparison after re-ranking images*

In this section, we investigate the improvements in retrieval accuracy with different re-ranking methods. We evaluate two types of re-ranking methods: geometric verification and local geometric consistency. These methods limit the list of candidates to a few hundred images because large candidate lists tend to be computationally expensive when employed on a large number of images.

Table 3 summarizes the performance of the re-ranking methods for different datasets compared to the performance of the initial search without re-ranking. It also shows the results of using different types of features (i.e. SIFT, MSER, and their combination) in different rows of the table. First, we examined the performance of proposed method using SIFT features. One can see that GV provides the best performance improvement for all the datasets by obtaining a performance gain of about 20% on average. The reason is that, with the given datasets, the variability of objects or scenes in many similar images is globally consistent. This is suitable for estimating the global transformation of the GV algorithm. However, the performance in terms of local geometric consistency with SLTMB is comparable with that of GV by obtaining a performance gain of about 16% on average. Not surprisingly, the improvement of re-ranking with LDA is significantly lower than with SLTMB due to the

**Table 3**
Performance of ss-VLAD (PCA,128-D) with and without re-ranking.

| Methods | Oxford | Paris | UKB | Holidays |
|---|---|---|---|---|
| **SIFT** | | | | |
| W/o re-ranking | 0.579 | 0.459 | 0.679 | 0.572 |
| GV | 0.762 | 0.67 | 0.854 | 0.811 |
| LDA | 0.624 | 0.548 | 0.739 | 0.667 |
| SLTMB | 0.719 | 0.626 | 0.828 | 0.754 |
| **MSER** | | | | |
| W/o re-ranking | 0.384 | 0.309 | 0.553 | 0.442 |
| GV | 0.543 | 0.517 | 0.738 | 0.641 |
| LDA | 0.417 | 0.366 | 0.599 | 0.519 |
| SLTMB | 0.534 | 0.498 | 0.696 | 0.615 |
| **SIFT + MSER** | | | | |
| W/o re-ranking | 0.579 | 0.459 | 0.679 | 0.572 |
| GV | 0.705 | 0.632 | 0.845 | 0.783 |
| LDA | 0.612 | 0.523 | 0.734 | 0.628 |
| SLTMB | 0.681 | 0.605 | 0.833 | 0.714 |

lack of spatial constraints and wrong topic assignment, as described for the previous experiment. By re-ranking with LDA, the gain in mAP is only about 7%.

Apart from the SIFT feature, we further examined the retrieval performance when using the MSER feature. Particularly, the MSER feature is applied to two phases consisting of encoding to retrieve an initial list and topic modeling to perform re-ranking. Note that, for the re-ranking stage, we cannot apply the efficient matching scheme based on the 2D histogram (presented in Section 3.4) with MSER features. The reason is that the MSER features are described by a set of elliptical regions, which are not specified by orientation and scale corresponding to circle regions as SIFT features. Therefore, instead of using the proposed histogram-based matching scheme, we utilize an alternative scheme that employs the parameters of co-variant matrix attached to each elliptical regions. The matches reflecting the transformation between two topic regions are determined by using the RANSAC-like algorithm. Then, the similarity of two images based on the feature matches is measured by Eqs. (22) and (23). Although this scheme is more computationally expensive than our proposed scheme, it guarantees the matching accuracy to make the fair comparison between SIFT and MSER. As shown in the second row of Table 3, with MSER features, we can draw nearly the similar conclusions of performance improvement as those on SIFT features, where SLTMB performs much better than LDA and slightly worse than GV. However, due to the significantly lower mAP of initial search, the re-ranking performance with MSER is much lower than the one with SIFT in all cases. This implies that encoding scheme captures the property of keypoint-based feature (i.e. SIFT) better than region-based feature (i.e. MSER).

To further compare the effectiveness of using different features in topic modeling, we conducted the other experiments that extract two types of features for each image, where SIFT is applied to the encoding phase while MSER is applied to the topic modeling phase. Compared to the experiment using SIFT for both phases, this allows us to avoid the effects of initial search result since we use the same encoded vectors to retrieve initial lists. As reported in the first and third rows of Table 3, we can see that SIFT generally outperforms MSER, which well demonstrates the effectiveness of using SIFT in our topic model. The reason is that the SIFT detector outputs the larger number of features than the MSER detector as illustrated in Fig. 8. Therefore, SIFT deals with occlusion and clutter more effectively. We note that these phenomena often occur in the building images (e.g. Oxford and Paris), or the outdoor images with high variety of scene types (e.g. Holidays). On UKB, the difference of performance between two types of features is small. Interestingly, with our proposed SLTMB, MSER even performs slightly better. This is because most images of the UKB dataset are the indoor objects with insignificant occlusions. Hence, UKB images take advantage of the property of MSER that captures well the local homogeneous parts in objects. In general, experimental results show that the keypoint-based feature like SIFT is more suitable for various types of datasets in large-scale image retrieval.

In addition, we evaluated the benefits of the re-ranking methods and their trade-offs between processing time and accuracy when increasing the size of the candidate list. Fig. 9 shows the results of mAP and time when short lists were used for re-ranking with GV and SLTMB. The experiment was conducted using Oxford +Flickr1M. In this case, the distractor images were collected from Flickr 1M, which is a subset of Filckr5M. Fig. 9(a) shows that increasing the size of the candidate list yields superior performance with both methods. However, there is a trade-off between processing time and mAP, where the processing time increases linearly with the short-list size.

Otherwise, as shown in Fig. 9(a), the mAP of our approach is slightly lower than that of GV as the short-list size increases, which is consistent with the results shown in Table 3. This is because we use a parametric representation of the object for SLTMB. In more detail, we assume that the distribution of the visual words within a topic region (or object) is Gaussian, which may not fit the shape of a complex object, such as a building or landmark. This may negatively affect the similarity measure at the re-ranking stage. But, as shown in the presented results, this effect is not significant.

Fig. 9(b) further shows the efficiency of our approach compared to GV. With the same short-list, re-ranking with SLTMB processes much faster than with GV. For a short-list size of 100 images, GV takes around 7.4 s, whereas our approach

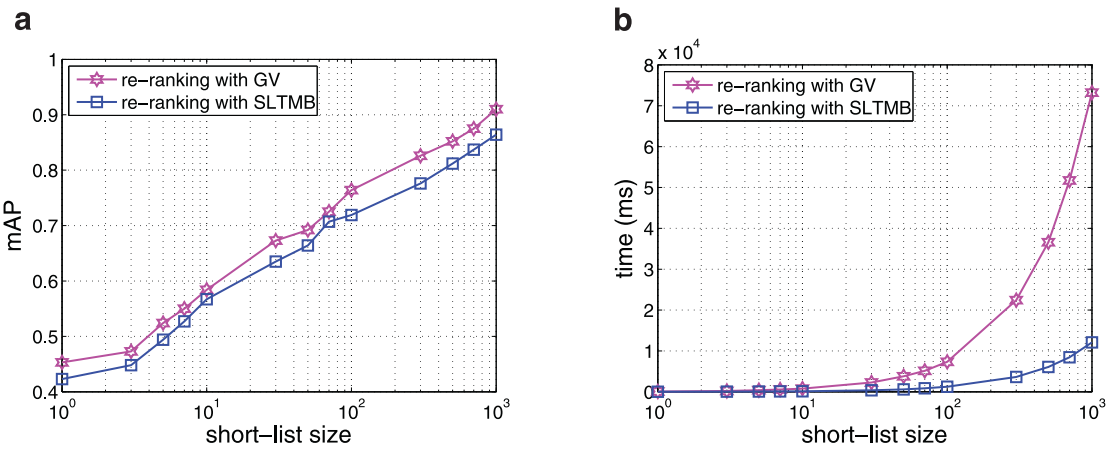


**Fig. 9.** GV vs. SLTMB based on: (a) mAP and (b) processing time.

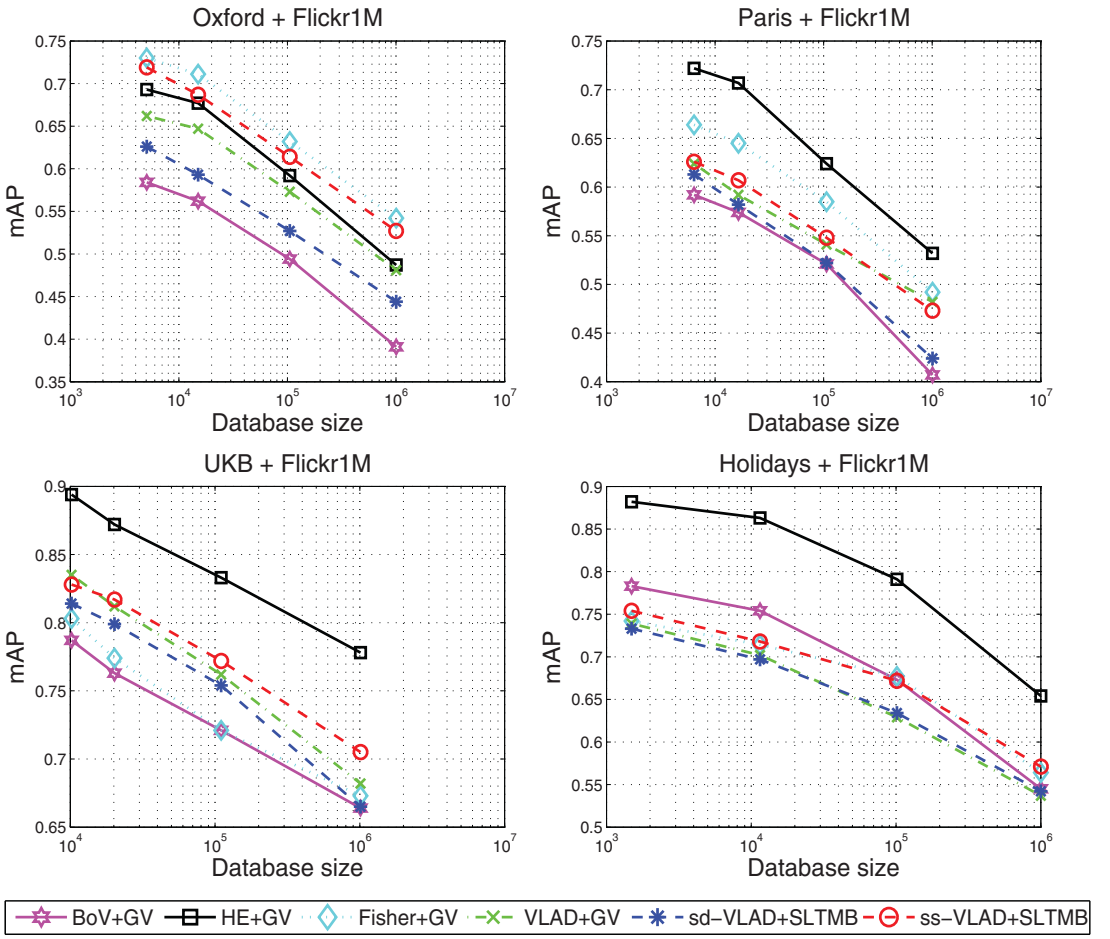


**Fig. 10.** Comparison results of proposed system with state-of-the-art methods using re-ranking with respect to data size.

only takes 1.1 s. On average, our approach runs about ten times faster than GV. This enables us to handle a larger list of candidates and increase the probability of ranking relevant images. Interestingly, as shown in Fig. 9, SLTMB obtains higher mAP with a short-list of 500 images that does GV with a short-list of 100 images, but the processing time is about the same. For that reason, our re-ranking method achieves a better trade-off between accuracy and efficiency.

Fig. 10 shows the behavior of our approach in combination with a re-ranking stage on the large datasets. We constructed these datasets by merging benchmarks with different numbers of Flickr 1M. Note that the methods used in this experiment

are limited to databases of 1M images. This is because additional memory is required for geometric information and identity of visual words in the inverted file to estimate the geometric consistency. Therefore, the scalability is limited when we consider the re-ranking stage in retrieval system. But this issue can be effectively processed in parallel when the database is larger than 1M images. Otherwise, the size of the candidate lists used in this experiment was 100 for all methods.

One can see in Fig. 10 that HE + GV outperforms the other methods in most cases due to its complexity of both the encoding and re-ranking stages. For the Oxford + Flickr1M database, the Fisher + GV and ss-VLAD + SLTMB show the best results. Because our approach is much faster than Fisher+GV, the efficiency of searching is much higher. With the remaining datasets, our proposed approach also achieves competitive results compared with the GV results presented in previous works. Recall that the performance of our approach can be significantly improved with larger short-lists. In contrast, this is a limitation in previous works, where large candidate lists made their searching times become very slow and impractical.

## 5. Conclusions

In this paper, we have presented a novel retrieval framework for finding similar images in a large-scale database. We first employed an encoding scheme to aggregate local descriptors into a single vector using two soft-assignment approaches based on distance and sparse coding. This encoded vector was then compressed and indexed with a small number of bytes so that it could scale well to very large database for producing initial search. Our experimental results show that this scheme outperformed well-known aggregation-based methods (e.g. VLAD, Fisher), and achieved the performance comparable to complicated method like HE. Subsequently, we introduced a probabilistic topic model to extract the latent topic and background regions from an image by using Gibbs sampling algorithm for approximate inference. Unlike conventional topic modeling approaches, our topic model explicitly exploits the spatial notion of visual words and the correlation between objects and background. This enables us to gain insight into the structure of the image, and handle large variation of object appearance. When compared to conventional LDA model, our SLTMB model provided much better results for extracting topics. Finally, we performed re-ranking in conjunction with the extracted topic regions on the list of candidates to refine the initial search result. Using an efficient scheme to calculate the similarity score between two images based on the common topic regions, our approach sped up the search and verification by ten times and could achieve competitive performance compared to popular GV method.

It should be noted that though proposed framework works effectively on large-scale datasets, there are existing limitations related to our SLTMB model. In this work, we assumed that topic regions are distributed according to a specific form like Gaussian, which may not fit the shape of the object. This may reduce the accuracy of topic prediction and the matching during measuring scheme, as well as negatively affect the performance of overall framework. Furthermore, it is time-consuming to manually select the optimal number of topics T during the training of SLTMB model. In the future, we plan to consider nonparametric Bayesian models [49] to automatically determine the optimal number of topic. Otherwise, different types of local features or their combination can be incorporated into topic model to examine the improvement of topic extraction. Another possible direction is that a combination of topic model and CNNs in the similar manner of [53] can open many opportunities to develop a powerful learning model. Hence, further study is needed to address these issues and provide a highly efficient search system.

## Acknowledgment

We would like to thank the anonymous reviewers for their valuable comments that helped us very much to enrich the quality of the paper.This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-(H8501-16-1015) supervised by the IITP (Institute for Information & communications Technology Promotion).

## Appendix

For our proposed model, computing the posterior distribution of latent variables given the observed data (visual word appearance and its location) is intractable. Therefore, we developed an approximate inference algorithm based on Gibbs sampling. Then, the goal was to approximate the distribution $p(\mathbf{z}, \mathbf{s}|\mathbf{w}, \mathbf{l}, \Pi)$ by deriving the following conditional probabilities:

$$p(z_{di} = t, s_{di} = 1|\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) = \frac{p(z_{di} = t, s_{di} = 1, \mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}|\Pi)}{p(\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}|\Pi)} \tag{24}$$

$$p(s_{di} = 2|\mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}, \Pi) = \frac{p(s_{di} = 2, \mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}|\Pi)}{p(\mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}|\Pi)} \tag{25}$$

Based on Bayes rule and d-separation property of our probabilistic graphical model, Eq. (24) yields:

$$\begin{aligned}
&p(z_{di}=t, s_{di}=1|\mathbf{w_d}, \mathbf{l_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&\propto p(z_{di}=t, s_{di}=1, w_{di}, l_{di}, \mathbf{w_{-di}}, \mathbf{l_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}|\Pi)\\
&\propto p(w_{di}|z_{di}=t, s_{di}=1, \mathbf{w_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&\times p(l_{di}|z_{di}=t, s_{di}=1, \mathbf{l_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&\times p(z_{di}=t|s_{di}=1, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) \times p(s_{di}=1|\mathbf{s_{-di}}, \Pi)
\end{aligned} \tag{26}$$

The first term of Eq. (26) is the posterior predictive of unobserved visual word $w_{di}$. Given the conjugate prior over $\Phi_t$ and marginalizing this parameter, we obtain:

$$\begin{aligned}
&p(w_{di}|z_{di}=t, s_{di}=1, \mathbf{w_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&= \int_{\Phi_t} p(w_{di}|z_{di}=t, s_{di}=1, \Phi_t, \Pi) \times p(\Phi_t|\mathbf{w_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) d\Phi_t\\
&= \frac{n^{WT}_{wt,-di} + \beta_1}{\sum'_w n^{WT}_{w't,-di} + W\beta_1}
\end{aligned} \tag{27}$$

The second term of Eq. (26) is the posterior predictive of unobserved location $l_{di}$. We can marginalize over the parameter of normal distribution $(\mu_{td}, \Lambda_{td})$ in image $d$ and obtain the closed form expression:

$$\begin{aligned}
&p(l_{di}|z_{di}=t, s_{di}=1, \mathbf{l_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&= \iint_{(\mu_{td}, \Lambda_{td})} p(l_{di}|z_{di}=t, s_{di}=1, \mu_{td}, \Lambda_{td}, \Pi) \times p(\mu_{td}, \Lambda_{td}|\mathbf{l_{-di}}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) d\mu_{td} d\Lambda_{td}\\
&= t_{\nu^{TD}_{td,-di}-q+1}\left(\mu^{TD}_{0,td,-di}, \frac{Q^{TD}_{td,-di}\left(\kappa^{TD}_{td,-di}\right)}{\kappa^{TD}_{td,-di}\left(\nu^{TD}_{td,-di} - q + 1\right)}\right)
\end{aligned} \tag{28}$$

Similarly, the third and fourth terms are respectively the posterior predictive of unobserved topic assignment $z_{di}$ and switch variable $s_{di}=1$. They are computed by marginalizing over parameters $\theta_d$ and $\Omega$, respectively.

$$\begin{aligned}
&p(z_{di}=t|s_{di}=1, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&= \int_{\theta_d} p(z_{di}=t|s_{di}=1, \theta_d, \Pi) p(\theta_d|\mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) d\theta_d\\
&= \frac{n^{TD}_{td,-di} + \alpha}{\sum'_t n^{TD}_{t'd,-di} + T\alpha}
\end{aligned} \tag{29}$$

$$p(s_{di}=1|\mathbf{s_{-di}}, \Pi) = \int_{\lambda} p(s_{di}=1|\lambda, \Pi) p(\lambda|\mathbf{s_{-di}}, \Pi) = \frac{N_{d1,-di} + \gamma}{N_{d,-di} + 2\gamma} \tag{30}$$

Combining Eqs. (27)–(30), we obtain the Gibbs sampling in Eq. (9).
In the same way, Eq. (25) can be factorized as:

$$\begin{aligned}
&p(s_{di}=2|\mathbf{w_d}, \mathbf{l_d}, \mathbf{s_{-di}}, \Pi)\\
&\propto p(s_{di}=2, w_{di}, l_{di}, \mathbf{w_{-di}}, \mathbf{l_{-di}}, \mathbf{s_{-di}}|\Pi)\\
&\propto p(w_{di}|s_{di}=2, \mathbf{w_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&\times p(l_{di}|s_{di}=2, \mathbf{l_{-di}}, \mathbf{s_{-di}}, \Pi) \times p(s_{di}=2|\mathbf{s_{-di}}, \Pi) d\lambda
\end{aligned} \tag{31}$$

Similar to the above derivation, each term of Eq. (31) can be computed by the posterior predictive and expressed as follows:

$$\begin{aligned}
&p(w_{di}|s_{di}=2, \mathbf{w_{-di}}, \mathbf{s_{-di}}, \Pi)\\
&= \int_{\Omega} p(w_{di}|s_{di}=2, \Omega, \Pi) p(\Omega|\mathbf{w_{-di}}, \mathbf{s_{-di}}, \Pi) d\Omega\\
&= \frac{n^{W}_{w,-di} + \beta_2}{\sum'_w n^{W}_{w',-di} + W\beta_2}
\end{aligned} \tag{32}$$

$$p(l_{di}|s_{di}=2,\mathbf{l}_{-\mathbf{di}},\mathbf{s}_{-\mathbf{di}},\Pi)=\text{uniform}\propto 1 \tag{33}$$

$$p(s_{di}=2|\mathbf{s}_{-\mathbf{di}},\Pi)=\int_{\lambda}p(s_{di}=2|\lambda,\Pi)p(\lambda|\mathbf{s}_{-\mathbf{di}},\Pi)d\lambda=\frac{N_{d2,-di}+\gamma}{N_{d,-di}+2\gamma} \tag{34}$$

Combining Eqs. (32)–(34), we obtain Eq. (10).

Given the samples from the posterior distributions by Gibbs sampling, it is possible to assign each visual word to a topic $t$ or background using the MAP estimator.

## References


[1] C. Andrieu, N. De Freitas, A. Doucet, M.I. Jordan, An introduction to mcmc for machine learning, Mach. Learn. 50 (1–2) (2003) 5–43.
[2] R. Arandjelovic, A. Zisserman, All about Vlad, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 1578–1585.
[3] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.
[4] A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal recovery, Convex Optim. Signal Process. Commun. (2009).
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[6] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, Spatial-bag-of-features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, IEEE, 2010, pp. 3352–3359.
[7] C. Chemudugunta, P.S.M. Steyvers, Modeling general and specific aspects of documents with a probabilistic topic model, in: Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems 19, vol.19, MIT Press, 2007, p. 241.
[8] O. Chum, A. Mikulik, M. Perdoch, J. Matas, Total recall ii: query expansion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 889–896.
[9] O. Chum, M. Perdoch, J. Matas, Geometric min-hashing: finding a (thick) needle in a haystack, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR., IEEE, 2009, pp. 17–24.
[10] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: Automatic query expansion with a generative feature model for object retrieval, in: Proceedings of the IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[11] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3828–3836.
[12] M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Proceedings of the Twentieth Annual Symposium on Computational Geometry, ACM, 2004, pp. 253–262.
[13] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[14] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 392–407.
[15] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences, 101, 2004, pp. 5228–5235.suppl 1
[16] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.
[17] H. Jégou, M. Douze, C. Schmid, On the burstiness of visual elements, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 1169–1176.
[18] H. Jégou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, Int. J. Comput. Vis. 87 (3) (2010) 316–336.
[19] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, Pattern Anal. Mach. Intell. IEEE Trans. 33 (1) (2011) 117–128.
[20] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 3304–3311.
[21] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, Pattern Anal. Mach. Intell. IEEE Trans. 34 (9) (2012) 1704–1716.
[22] K. Jiang, Q. Que, B. Kulis, Revisiting kernelized locality-sensitive hashing for improved large-scale image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4933–4941.
[23] Y. Kalantidis, Y. Avrithis, Locally optimized product quantization for approximate nearest neighbor search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2321–2328.
[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[25] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: Proceedings of the IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2130–2137.
[26] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 2169–2178.
[27] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 2036–2043.
[28] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2085–2098.
[29] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, Knowl. Data Eng. IEEE Trans. 26 (9) (2014) 2138–2150.
[30] J.L. Long, N. Zhang, T. Darrell, Do convnets learn correspondence? in: Advances in Neural Information Processing Systems, 2014, pp. 1601–1609.
[31] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[32] Y. Lv, W.W. Ng, Z. Zeng, D.S. Yeung, P.P. Chan, Asymmetric cyclical hashing for large scale image retrieval, Multimed. IEEE Trans. 17 (8) (2015) 1225–1235.
[33] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image Vis. Comput. 22 (10) (2004) 761–767.
[34] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1–2) (2005) 43–72.
[35] M. Muja, D.G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration., in: VISAPP (1), 2, 2009.
[36] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 2161–2168.
[37] M. Norouzi, A. Punjani, D.J. Fleet, Fast exact search in hamming space with multi-index hashing, Pattern Anal. Mach. Intell. IEEE Trans. 36 (6) (2014) 1107–1119.
[38] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1? Vis. Res. 37 (23) (1997) 3311–3325.
[39] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1717–1724.

[40] M. Perd'och, O. Chum, J. Matas, Efficient representation of local geometry for large scale object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 9–16.
[41] F. Perronnin, Y. Liu, J. Sánchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2010, pp. 3384–3391.
[42] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2007, pp. 1–8.
[43] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2008, pp. 1–8.
[44] J. Philbin, J. Sivic, A. Zisserman, Geometric latent dirichlet allocation on a matching graph for large-scale image datasets, Int. J. Comput. Vis. 95 (2) (2011) 138–153.
[45] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.
[46] J. Sang, C. Xu, Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012, pp. 19–28.
[47] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, W.T. Freeman, Discovering object categories in image collections, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV, IEEE, 2005, pp. 370–377.
[48] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV, IEEE, 2003, pp. 1470–1477.
[49] E.B. Sudderth, A. Torralba, W.T. Freeman, A.S. Willsky, Describing visual scenes using transformed objects and parts, Int. J. Comput. Vis. 77 (1–3) (2008) 291–330.
[50] J. Tang, Z. Li, M. Wang, R. Zhao, Neighborhood discriminant hashing for large-scale image retrieval, Image Process. IEEE Trans. 24 (9) (2015) 2827–2840.
[51] W. Tang, R. Cai, Z. Li, L. Zhang, Contextual synonym dictionary for visual object retrieval, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 503–512.
[52] E. Tola, V. Lepetit, P. Fua, Daisy: An efficient dense descriptor applied to wide-baseline stereo, Pattern Anal. Mach. Intell. IEEE Trans. 32 (5) (2010) 815–830.
[53] L. Wan, L. Zhu, R. Fergus, A hybrid neural network-latent topic model, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2012, pp. 1287–1294.
[54] C. Wang, D. Blei, F.-F. Li, Simultaneous image classification and annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 1903–1910.
[55] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2010, pp. 3360–3367.
[56] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, T.X. Han, Contextual weighting for vocabulary tree based image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 209–216.
[57] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 1794–1801.
[58] J. Yu, D. Tao, J. Li, J. Cheng, Semantic preserving distance metric learning and applications, Inf. Sci. 281 (2014) 674–686.
[59] Y. Zhang, Z. Jia, T. Chen, Image retrieval with geometry-preserving visual phrases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2011, pp. 809–816.