Saliency map driven image retrieval combining the bag-of-words model and PLSA

Emmanouil Giouvanakis, Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
Email: {egiouvan, costas}@aiia.csd.auth.gr

***Abstract*—A new image retrieval system is proposed that combines the bag-of-words (BoW) model and Probabilistic Latent Semantic Analysis (PLSA). First, interest points on images are detected using the Hessian-Affine keypoint detector and Scale Invariant Feature Transform (SIFT) descriptors are computed. Graph-based visual saliency maps are then employed in order to detect and discard outliers in image descriptors. By doing so, SIFT features lying in non-salient regions can be deleted. All the remaining reliable feature descriptors are divided into a number of subsets and partial vocabularies are extracted for each of them. The final vocabulary used in the BoW model is obtained by the concatenating the partial vocabularies. The resulting BoW representations are weighted using the TF-IDF scheme. Finally, the PLSA is employed to perform a probabilistic mixture decomposition of the weighted BoW representations. Query expansion is demonstrated to improve the retrieval quality. Overall a 0.79 mean average precision is reported when the saliency filtering was applied on SIFTs and the BoW plus PLSA method was used.**

***Index Terms*—image retrieval, object retrieval, graph-based visual saliency, bag of words, probabilistic latent semantic analysis, query expansion**

I. Introduction

Image retrieval techniques can be classified into two classes, namely the text-retrieval techniques, resorting to key-words, and the content-based techniques (CBIR), using features extracted from images, which describe the image content. Despite the huge volume of related research, many problems still remain open. For example, in CBIR, one could mention the huge dimensionality of feature vectors or the big amount of data. Both problems lead to a compromise between the retrieval quality and the computational demands (e.g., time, memory). Another problem is related to image features chosen to describe the content. Global features, like histograms, describe an image in a holistic manner. Thus, they discard any local information at the expense of their discriminating power. On the other hand, local descriptors extracted from salient regions perform well even in the presence of illumination changes or occlusion.

One of the most popular image retrieval techniques is the so-called bag-of-words (BoW) model [1]. It relies on the idea of quantizing image features into visual words. Next, the frequency of visual words is estimated and it is exploited in any information retrieval method, such as the term-frequency - inverse document frequency (TF-IDF).

Many features have been proposed for image or object retrieval. The GIST descriptors [2], offer a low dimensional representation of a scene whose dominant spatial structure is represented by a set of perceptual dimensions (i.e. naturalness, openness, roughness, expansion, ruggedness). The GIST can perform well, when similar images are well-aligned. However, they lack efficiency if there are significant variations due to rotation, scaling, or viewpoint [3]. On the other hand, the popular scale-invariant feature transform (SIFT) [4] extracts a collection of local feature vectors at interest points determined by a keypoint detector that possess invariance to rotation and scale at some extent.

A new approach for image retrieval is proposed, which is scalable to large image datasets, being computationally efficient. First, a BoW model of SIFT descriptors is computed at keypoints derived by the Hessian-Affine detector [5]. Due to the fact that the keypoints are detected on the whole image, the keypoints may lie in non salient regions. Such keypoints are treated as outliers and their effect can be eliminated thanks to saliency maps. The aforementioned process reduces significantly the amount of descriptors per image leading to a computational efficient BoW representation. Furthermore, instead of using a subset of SIFT visual descriptors for vocabulary computation, the full set of SIFT descriptors is divided to a predefined number of disjoint subsets. From each subset, a partial vocabulary is derived and the full vocabulary is obtained by concatenating all the partial vocabularies. Second, the Probabilistic Latent Semantic Analysis (PLSA) [6], is performed on the BoW representations of the images in order to discover descriptive visual topics. Third, a query expansion technique, named Average Query Expansion (AQE) is used to further improve retrieval results.

To sum up, the novelty of this paper is in the combination of SIFT filtering with saliency maps and PLSA applied to BoW representations in order to extract low-dimensional feature vectors for image retrieval. An overview of the proposed system can be seen in Figure 1. It is proved that by using saliency maps the number of the total SIFT descriptors is reduced by approximately 50%, while the use of PLSA results in a just 10-dimensional image representation, which achieves a much greater performance than the standard BoW representation.

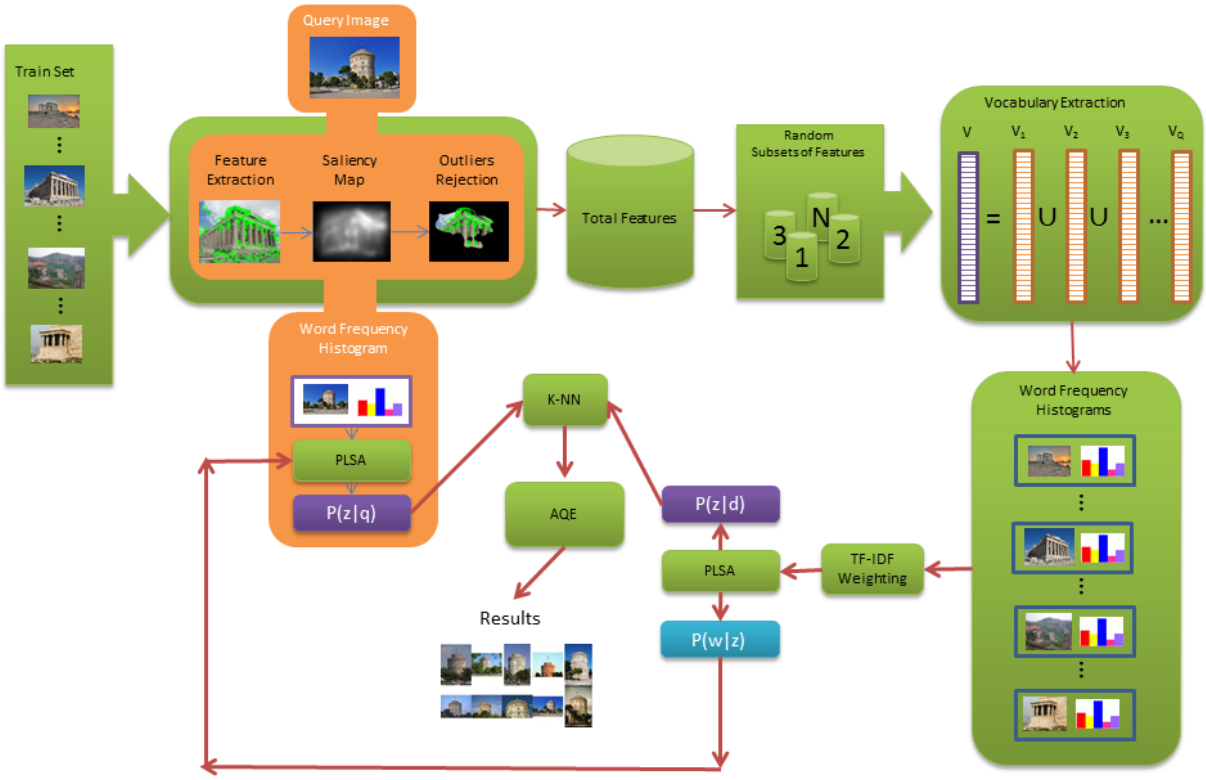

Fig. 1. Overview of the proposed system.

The rest of the paper is organized as follows. In Section II, related works are surveyed. The dataset used is presented in Section IV. Technical details for the implementation of the proposed system, including feature extraction, the BoW representation, the application of the PLSA to the BoW vectors, the query expansion, and the use of saliency maps, can be found in Section III. Evaluation results are demonstrated in Section V and conclusions are drawn in Section VI.

II. Related Work

Many approaches for image retrieval have been proposed so far in the literature [7]–[11]. In fact, most of them are based on the BoW model [1]. Their differences are mainly in the kind of image features extracted, the voting scheme applied during the computation of the word frequency vectors, and the method used for vocabulary creation.

In [8], the user selects a region of a query image containing an object, and the system returns a ranked list of images containing the same object. The system extracts the SIFT image features and uses the BoW representation to create a visual vocabulary. The authors state that flat k-means can be scaled to large collections of image descriptors by using approximate nearest neighbours techniques. Next, the TF-IDF is used to weigh the visual word frequency vectors. Similarity search is performed by calculating the ℓ_2 distance between a query vector and all the weighted image vectors. Finally, a re-ranking is performed on the top-ranked results using spatial constraints.

Bundled features, that is, image features bundled into local groups, are proposed in [9]. Using a group of bundled features, a higher discriminative power is achieved, and simple and robust geometric constraints were efficiently enforced within each group. The main idea is to detect maximally stable extremal regions and describe these regions with SIFT descriptors at points in these regions.

A recent approach is proposed in [10], where locality sensitive hashing is used to solve the large memory consumption problem during visual vocabulary creation. In particular, exact Euclidean locality sensitive hashing (E2LSH) is used to hash the SIFT features in order to form a group of random visual vocabularies. Next, visual vocabulary histograms are computed and the TF-IDF weighting scheme is applied to the visual word frequency vectors. Finally, a query expansion strategy is used to achieve better results.

III. Image Retrieval System Development

The implemented system can be seen in Figure 1. The training images are first passed into the feature extraction step, where the Hessian-Affine keypoint detector is used and SIFT features are computed. Saliency maps are employed to discard any outliers. Next, all the SIFT features are divided into Q disjoint subset. The final vocabulary is the concatenation of all the partial vocabularies derived from each subset. Then, word frequency histograms are computed for each image and weights for each word are assigned using the TF-IDF method. The weighted BoW representations are passed into the PLSA to acquire a low-dimensional vector for each image. Each of the previous steps is discussed in more depth next.

A. Feature Extraction

The code from [12] was used for feature extraction. This code implements a modified version of the Hessian-Affine keypoint detector [5], the SIFT [4] features are computed on affine normalized image patches and they are normalized to unit ℓ_2 norm vectors.

B. Outlier Rejection Using Saliency Maps

Taking into account that we are interested in image retrieval applied to images related to tourism, two images cannot be considered as similar unless they both depict the same landmark (e.g. archaeological monument, landscape). More specifically, the system must not be distracted by common features found in abundance in any image dataset, such as trees, roads.

Due to the fact that the Hessian-Affine keypoint detector examines the whole image, the sets of derived keypoints contain many outliers. To alleviate this drawback, saliency maps are used to detect the most salient regions and retain only the descriptors lying within these salient regions.

The saliency map is computed thanks to the bottom-up graph-based visual saliency (GBVS) model [13]. The latter model is preferred from classical algorithms (e.g., that in [14]), because the GBVS predicts human fixations more reliably. The saliency maps admit values in [0,1] with low values indicating less salient regions. A binary saliency mask is obtained by thresholding the saliency map at a desired level, which distinguishes the salient regions from the non salient ones. For example, by defining a threshold of 75%, the 75% most-salient regions are retained. By using this mask, keypoints whose coordinates lie within the non salient regions are discarded along with the corresponding SIFT descriptors. In Figure 2, saliency maps are demonstrated for several images of the ATLAS dataset which is descibed in Section IV. Approximately, half of the original features are thus retained for visual vocabulary computation.

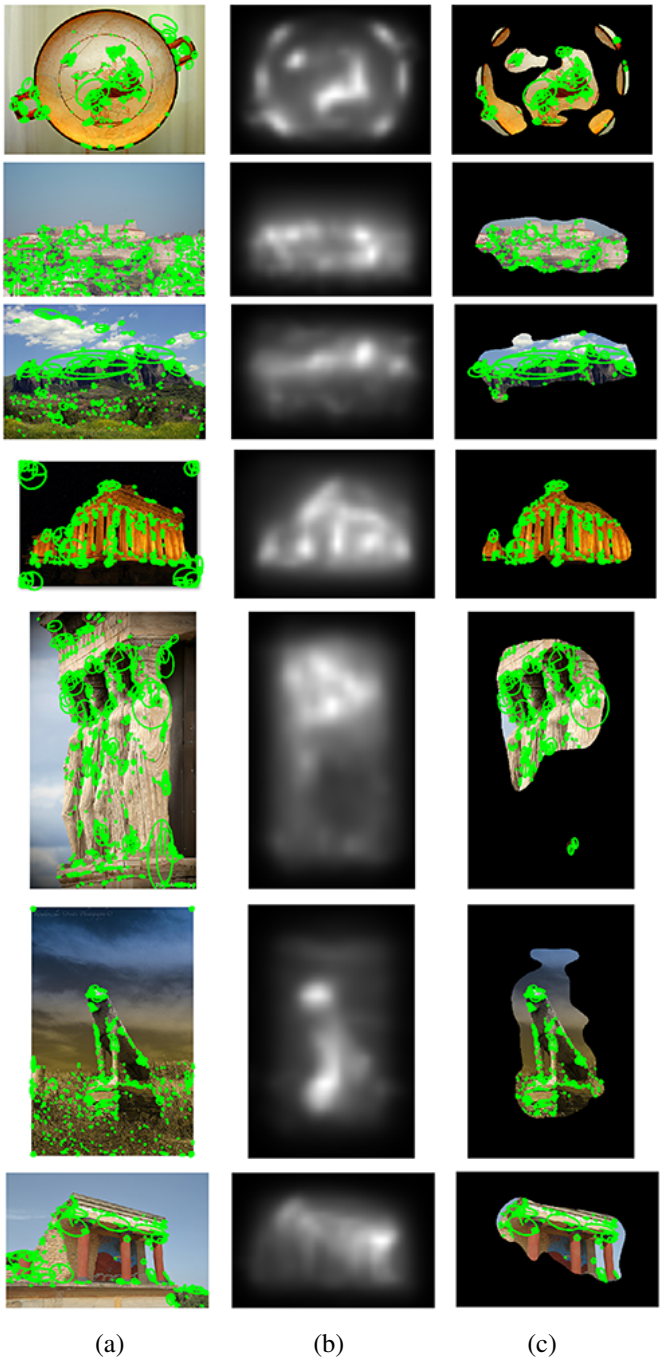

Fig. 2. a) Images along with their detected keypoints. b) The graph-based saliency map of the images. c) By thresholding the saliency map, the images are segmented and only the keypoints lying in salient regions are retained.

C. BoW Representation

The BoW representation is a state-of-the-art method in image retrieval. It is based on the idea of quantizing image features to clusters called visual words and applying then standard information retrieval techniques originally proposed for text retrieval [15]. To do so, a vocabulary of visual words is generated and each descriptor is assigned to its nearest visual word.

Accordingly, the generation of a visual vocabulary is a vector quantization problem and k-means or one of its variants is employed. Early systems (e.g., in [1]) used a flat k-means algorithm, which is not scalable to large datasets.

Here, the k-means algorithm employs an approximate nearest neighbor (ANN) algorithm to accelerate the sample-to-center comparisons. In fact, the ANN uses a best-bin-first randomized kd-tree algorithm to approximately and quickly quantize each descriptor into its nearest visual word. This results in speeding-up the execution and allows for handling large datasets.

No matter how fast and efficient a clustering algorithm is, there will always be a computational issue regarding the number of descriptors passed to it. To deal with the data explosion problem, besides speeding-up the k-means algorithm execution, the full set of descriptors is divided randomly into Q disjoint subsets. For example, if there are S descriptors divided into Q subsets, a vocabulary W of size $|W|$ is generated by concatenating the partial vocabularies W_i, $i = 1, 2, ..., Q$ of size approximately $|W|/Q$ words, i.e., $W = \bigcup_{i=1}^{Q} W_i$.

After vocabulary extraction, SIFT descriptors have to be quantized to their nearest visual word. A kd-tree [16] is built to efficiently find the nearest visual word to each 128-dimensional SIFT descriptor. Then, each image is described by a sparse vector of visual word frequencies called the BoW vector. Many proposals have been made for the voting scheme to be used to calculate the frequency vectors. In most cases, each SIFT is assigned exclusively to the word. Alternatively, the top-n nearest visual words can be used [17]. Here, the exclusive assignment of SIFT descriptors to one nearest visual word is used, because no significant performance improvement was noticed, when the top-n nearest visual words were employed.

Finally, the TF-IDF weighting scheme [15] is applied to visual word frequency vectors, which punishes the common visual words bearing no discriminating power, while rewarding the visual words appearing in a small portion of images. The latter are considered as discriminant descriptors for the images.

When a query image is submitted, the SIFT descriptors are computed and then quantized into visual words using the vocabulary created during training, resulting in its BoW vector. Then it is weighted using the weights given by applying the TF-IDF technique on the BoW vectors of the training images. Finally, the resulting vector is normalized so that it admits a unit ℓ_2 norm.

D. PLSA on the BoW representation

PLSA [6] is used to perform a probabilistic mixture decomposition of the weighted BoW representations. Although the PLSA was originally used to discover the topics where words and documents could be attributed to, it has also been used to discover the object categories in image classification [18]. By applying PLSA to the weighted term-document matrix formed by the BoW representations of the images, the relations between the words and the images are captured by the probability distribution between the images and the generated topics as well as between the topics and the the words. Let $D = \{d_1, \ldots, d_N\}$ be a set of images and $W = \{w_1, \ldots, w_M\}$ be the vocabulary of visual words, where $M = |W|$. The joint probability model is defined by the mixture:

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \quad (1)$$

where $z \in Z = \{z_1, z_2, \ldots, z_L\}$ is a set of latent topics, $P(w|z)$ is the word per topic distribution and $P(z|d)$ is the topic per image distribution.

By applying the PLSA to the weighted BoW representations, the goal is to derive the distributions $P(w|z)$ and $P(z|d)$. During training, the latent visual topics z are learnt and the images can now be represented by the L-dimensional vector having elements $P(z_i|d), i = 1, 2, \ldots, L$.

When a query image q is submitted, it is represented by the distribution $P(z|q)$ which is calculated by running the M step of the Expectation Maximization (EM) algorithm for $P(z|q)$ until convergence, keeping $P(w|z)$ fixed to that learnt during the training (see Figure 1).

E. Query Expansion

Given a query, its nearest images are returned by calculating the Euclidean distances between the query and the BoW vectors of the training images. In order to improve performance, query expansion can be performed. This is a technique borrowed from text retrieval where a number of highly ranked documents are reissued as a new query. Many query expansion techniques are met in the literature [19]–[21].

In this paper, AQE is used in which a query is applied and the top-k retrieved images are tested for spatial consistency with the query image. For spatial verification, firstly, visual matches between the images are found using the SIFT descriptors and then these matches are passed to the RANSAC algorithm [22] in order to find inliers. Only images sharing a number of inliers above a predefined threshold are consider as spatially verified. Then, the BoW vectors of the spatially verified images along with the query vector are averaged and reissued as a new query. By doing so, the performance is significantly improved due to the fact that the query vector is further enriched with information of similar images, i.e., words appearing in multiple views of the same landmark.

Inspired by this, the AQE is tested on the topic per document distribution $P(z|d)$ as well. This means that after calculating $P(z|q)$, the Euclidean distances between the $P(z|q)$ and the $P(z|d)$ are measured and only the top-k images are tested for spatial consistency. The spatial verification step is performed in the same way as described previously. Finally, the distribution $P(z|d')$ for d' that were spatially verified, along with the distribution $P(z|q)$ are averaged and the resulting distribution is reissued as a new query.

IV. Image Dataset Description

The ATLAS dataset contains images of tourist interest which were crawled from *Flickr* using keywords/tags about Greece. The total size of the collected dataset is about 650,000 images including archaeological monuments, landscapes around Greece. The dataset is stored in an SQL database locally.

In this paper, we use a portion of the ATLAS dataset, containing 1,320 images gathered by querying the SQL database for images bearing specific tags. The subset of images are of various size and are compressed in JPG format. The 1,320 images are grouped into 6 classes, each one depicting a different landmark or monument. Ground truth for this subset of images is manually defined in order to evaluate the system.

V. Evaluation results

In order to evaluate the proposed system and assess the contribution of saliency filtering and PLSA, results are disclosed for two approaches. The first one is the standard BoW model, and the second one is the BoW + PLSA, in which the PLSA is performed on the weighted BoW vectors. In both cases, the AQE is applied.

During training, a subset of 1,320 images was used for feature extraction and two sets of features were stored. The first one, namely the raw set, contains the whole set of descriptors, while the second one is the filtered set containing only the descriptors lying in salient regions of the images. Both the BoW and the BoW + PLSA approaches were tested on the raw and the filtered set.

For evaluation purposes, vocabularies of different sizes were computed on the raw and the filtered set. The vocabulary sizes vary between 100 and 10K words, while the number of topics chosen for the PLSA varies between 10 and 100 topics.

120 query images were used, i.e., 20 images for each of the 6 classes. For spatial verification, the top-200 ranked images was tested for spatial consistency, while a threshold of 20 inliers was chosen in order to consider two images as geometrically consistent.

Best performance was obtained when saliency filtering was applied along with a 6K vocabulary and 10 topics for the PLSA. The mean Average Precision (mAP) is chosen as the evaluation measure. The mAP measured in the conducted experiments is shown in Table I. It is seen that the saliency filtering improves the retrieval performance considerably. Moreover, the PLSA yields a great performance gain.

During feature extraction it was noticed that saliency filtering offers a reduction of the initial set of descriptors by over 50%, while the overall performance is increasing. Saliency filtering may sometimes decrease the performance and this is due to the nature of the images, i.e., the monument depicted, the point of view the image was taken etc. However, the decrease in performance in these classes is relatively small compared to the great increase for the others, not to mention the huge reduction in the amount of descriptors. Specifically, the number of descriptors extracted from the 1,320 images is 930,316 and it is reduced to 412,359, when saliency filtering is applied.

To test if the total mAP differences are statistically significant, we assume that the mAPs p_1 and p_2 delivered by two methods are binomially distributed random variables. If $\hat{p_1}$, $\hat{p_2}$ denote the empirical mAPs listed in the last row of Table I and $\bar{p} = \dfrac{\hat{p_1} + \hat{p_2}}{2}$, the hypothesis $H_0 : p_1 = p_2 = \bar{p}$ is tested at 95% level of significance. The variance of the mAP difference is given by $\beta = 2\,\dfrac{\bar{p}(1-\bar{p})}{T}$, where T is the number of test images (i.e., 120). For $\varphi = 1.65\sqrt{\beta}$, if $\hat{p_1} - \hat{p_2} \geq \varphi$, we reject H_0 with risk 5% of being wrong. The aforementioned analysis yields that the performance gain between the BoW + PLSA without saliency filtering and the BoW + PLSA with saliency filtering ($\varphi = 9.3\%$) is statistically significant, while between the BoW without saliency filtering and the BoW with saliency filtering ($\varphi = 10.2\%$) is not.

TABLE I
Mean average precision for the 6K vocabulary and 10 topics.

	No Saliency Filtering		*Saliency Filtering*	
Class	**BoW**	**BoW + PLSA**	**BoW**	**BoW + PLSA**
Erechtheum	0.47	**0.47**	0.60	**0.83**
Odeon	0.76	**0.79**	0.70	**0.74**
Parliament	0.85	**0.93**	0.80	**0.90**
Parthenon	**0.56**	0.54	0.63	**0.81**
Sounio	0.43	**0.52**	0.45	**0.54**
White Tower	0.58	**0.90**	0.70	**0.88**
Total	0.61	**0.69**	0.65	**0.79**

VI. Conclusions

An image retrieval system has been described. The BoW model was employed and subsets of the features were used for vocabulary creation. Graph-based visual saliency maps were exploited in order to detect non-salient regions on each images and discard local features that do not belong in them. The Probabilistic Latent Semantic Analysis was applied to the BoW representation of the images. It has been demonstrated that the just mentioned combination of the BoW model and the PLSA offers performance gains. In the future, more sophisticated query expansion techniques will be examined.

VII. Acknowledgments

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Competitiveness-Cooperation 2011" - Research Funding Program: 11SYN-10-1730-ATLAS.

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, pp. 1470–1477.

[2] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[3] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 19:1–19:8.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, 2004.

[5] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[6] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.

[7] A. Mikulik, O. Chum, and J. Matas, "Image retrieval for online browsing in large image collections," in *Proc. Similarity Search and Applications*. Springer, 2013, pp. 3–15.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching." in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2007.

[9] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009, pp. 25–32.
[10] Z. Yongwei, L. Bicheng, and G. Haolin, "Bag-of-visual-words based object retrieval with E2LSH and query expansion," in *Instrumentation, Measurement, Circuits and Systems*, 2012, vol. 127, pp. 713–725.
[11] L. Dai, X. Sun, F. Wu, and N. Yu, "Large scale image retrieval with visual groups," in *Proc. IEEE Int. Conf. Image Processing*, 2013, pp. 582–2586.
[12] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 9–16.
[13] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 545–552.
[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[15] R. A. B.-Yates and B. R.-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.
[16] J. L. Bentley, "Multidimensional divide-and-conquer," *Commun. ACM*, vol. 23, no. 4, pp. 214–229, 1980.
[17] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2007, pp. 494–501.
[18] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," in *Proc. European Conf. Computer Vision*, 2006, pp. 517–530.
[19] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2012.
[20] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–8.
[21] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 889–896.
[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.