Incremental estimation of visual vocabulary size for image retrieval

Ilias Gialampoukidis, Stefanos Vrochidis and Ioannis Kompatsiaris

Abstract The increasing amount of image databases over the last years has highlighted our need to represent an image collection efficiently and quickly. The majority of image retrieval and image clustering approaches has been based on the construction of a visual vocabulary in the so called Bag-of-Visual-words (BoV) model, analogous to the Bag-of-Words (BoW) model in the representation of a collection of text documents. A visual vocabulary (codebook) is constructed by clustering all available visual features in an image collection, using k-means or approximate k-means, requiring as input the number of visual words, i.e. the size of the visual vocabulary, which is hard to be tuned or directly estimated by the total amount of visual descriptors. In order to avoid tuning or guessing the number of visual words, we propose an incremental estimation of the optimal visual vocabulary size, based on the DBSCAN-Martingale, which has been introduced in the context of text clustering and is able to estimate the number of clusters efficiently, even for very noisy datasets. For a sample of images, our method estimates the potential number of very dense SIFT patterns for each image in the collection. The proposed approach is evaluated in an image retrieval and in an image clustering task, by means of Mean Average Precision and Normalized Mutual Information.

Ilias Gialampoukidis
Centre for Research and Technology Hellas - Information Technologies Institute, 6th km Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece, e-mail: heliasgj@iti.gr

Stefanos Vrochidis
Centre for Research and Technology Hellas - Information Technologies Institute, 6th km Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece, e-mail: stefanos@iti.gr

Ioannis Kompatsiaris
Centre for Research and Technology Hellas - Information Technologies Institute, 6th km Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece, e-mail: ikom@iti.gr

1 Introduction

Image retrieval and image clustering are related tasks because of their need to efficiently and quickly search for nearest neighbors in an image collection. Taking into account that image collections are dramatically increasing (eg. Facebook, Flickr, etc.), both tasks, retrieval and clustering, become very challenging and traditional techniques show reduced functionality. Nowadays, there are many applications of image retrieval and image clustering which support image search, personal photo organization, etc.

Searching in an image collection for similar images is strongly affected by the representation of all images. Spatial verification techniques for image representation, like RANSAC and pixel-to-pixel comparisons, are computationally expensive and have been outperformed by the Bag-of-Visual-words (BoV) model, which is based on the construction of a visual vocabulary, known also as visual codebook, using a vocabulary of visual words [16], by clustering all visual features. The visual vocabulary construction is motivated by the Bag-of-Words (BoW) model in a collection of text documents. The set of all visual descriptors in an image collection is clustered using k-means clustering techniques, which are replaced by approximate k-means methods [14], in order to reduce the computational cost of visual vocabulary construction. However, both k-means techniques require as input the number of visual words k, which we shall estimate incrementally.

The visual vocabulary is usually constructed, using an empirical number of visual words k, such as $k = 4000$ in [17]. The optimal number k is hard to be tuned in very large databases, and impossible when ground truth does not exist. An empirical guess of k may lead to the construction of visual codebooks, which are not optimal when involved in an image retrieval of image clustering task. To that end, we propose a scalable estimation of the optimal number of visual words k in an incremental way, using a recent modification of DBSCAN [3], which also has a scalable and parallel implementation [6]. In the proposed framework, the final number of visual words is incrementally estimated on a sample of images and therefore, it can easily scale up to very large image collections, in the context of Big Data.

The most prominent visual features are SIFT descriptors [8], but several other methods have been proposed to represent an image, such as VLAD [7], GIST [11], Fisher vectors [12] or DCNN features [9]. In this work, we will restrict our study on the estimation of the visual vocabulary size, based on SIFT descriptors, and comparison in terms of the optimal visual features is beyond the scope of this work.

The main research contributions of this work are:

- Estimate the optimal size of a visual vocabulary
- Build the size estimation incrementally

Therefore, we are able to build efficient visual vocabularies without tuning the size or guessing a value for k. Our proposed method is a hybrid framework, which combines the recent DBSCAN-Martingale [5] and k-means clustering. The proposed hybrid framework is evaluated in the image retrieval and image clustering problems, where we initially provide an estimation for the number of visual words

k, using the DBSCAN-Martingale, and then cluster all visual descriptors by k, as traditionally done by k-means clustering.

In Section 2 we present the related work in visual vocabulary construction and in Section 3 we briefly present the DBSCAN-Martingale estimator of the number of clusters. In Section 4, our proposed hybrid method for the construction of visual vocabularies is described in detail, and finally, in Section 5, it is evaluated under the image retrieval and image clustering tasks.

2 Related Work

The Bag-of-Visual-Words (BoV) model initially appeared in [16], in which k-means clustering is applied for the construction of a visual vocabulary. The constructed visual vocabulary is then used for image retrieval purposes and is similar to the Bag-of-Words model, where a vocabulary of words is constructed, mainly for text retrieval, clustering and classification. In the BoV model, the image query and each image of the collection are represented as a sparse vector of term (visual word) occurrences, weighted by tf-idf scores. The similarity between the query and each image is calculated, using the Mahalanobis distance or simply the Euclidian distance. However, there is no obvious value for the number of clusters k in the k-means clustering algorithm.

Other approaches for the construction of visual vocabularies include Approximate k-means (AKM) clustering, which offers scalable construction of visual vocabularies. Hierarchical k-means (HKM) was the first approximate method for fast and scalable construction of a visual vocabulary [10], where data points are clustered by $k = 2$ or $k = 10$ using k-means clustering and then k-means is applied to each one of the newly generated clusters, using the same number of clusters k. After n steps (levels), the result is k^n clusters. HKM has been outperformed by AKM [14], where a forest of 8 randomized k-d trees provides approximate nearest neighbor search between points and the approximately nearest cluster centers. The use of 8 randomized k-d trees with skewed splits have recently been proposed, in the special case of SIFT descriptors [4]. However, all AKM clustering methods require as input the number of clusters k, so an efficient estimation of k is necessary.

The need to estimate the number of visual words emerges from the computational cost of k-means algorithm, either in exact or approximate k-means clustering [18]. Apart from being a time consuming process, tuning the number of clusters k may affect significantly the performance of the image retrieval task [14]. Some studies assume a fixed value of k, such as $k = 4000$ in [17], but in general the choice of k varies from 10^3 up to 10^7, as stated in [11]. In another approach, 10 clusters are extracted using k-means for each one of the considered classes (categories), which are then concatenated in order to form a global visual vocabulary [19]. In contrast, we shall estimate the number of clusters using the DBSCAN-Martingale [5], which automatically estimates the number of clusters, based on an extension of DBSCAN [3], without a priori knowledge of the density parameter *minPts* of DB-

SCAN. DBSCAN-Martingale generates a probability distribution over the number of clusters and has been applied to news clustering, in combination with LDA [5].

3 The DBSCAN-Martingale estimation of the number of clusters

In this section, we briefly describe the DBSCAN-Martingale, which has been introduced for the estimation of the number of clusters in a collection of text documents. DBSCAN [3] uses two parameters ε and $minPts$ to cluster the points of a dataset without knowing the number of clusters. DBSCAN-Martingale overcomes the tuning of the parameter ε and shows robustness to the variation of the parameter $minPts$ [5].

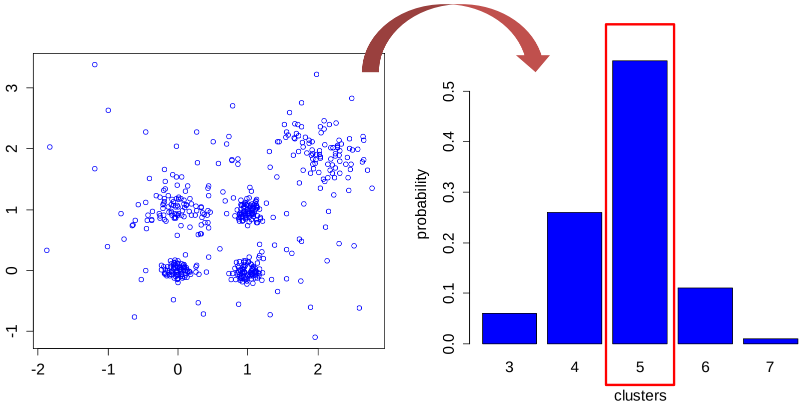


Fig. 1: The number of clusters as estimated by the DBSCAN-Martingale on an illustrative dataset. The generated probability distribution states that it is more likely to have 5 clusters, although they appear in different density levels and there are points which do not belong to any of the clusters.

The estimation of the number of clusters is a probabilistic method and assigns a probability distribution over the number of clusters, so as to extract all clusters for all density levels. For each randomly generated density level ε, density-based clusters are extracted using the DBSCAN algorithm. The density levels $\varepsilon_t, t = 1, 2, \ldots, T$ are generated from the uniform distribution in the interval $[0, \varepsilon_{max}]$ and sorted in increasing order.

Each density level ε_t provides one partitioning of the dataset, which then formulates a $N \times 1$ clustering vector, namely $C_{DBSCAN(\varepsilon_t)}$ for all stages $t = 1, 2, \ldots, T$, where N is the number of points to cluster. The clustering vector takes as value the cluster ID κ of the j-th point, i.e. $C_{DBSCAN(\varepsilon_t)}[j] = \kappa$.

In the beginning of the algorithm, there are no clusters detected. In the first stage $(t = 1)$, all clusters detected by $C_{DBSCAN(\varepsilon_1)}$ are kept, corresponding to the lowest density level ε_1. In the second stage $(t = 2)$, some of the detected clusters by $C_{DBSCAN(\varepsilon_2)}$ are new and some of them have also been detected at previous stage $(t = 1)$. DBSCAN-Martingale keeps only the newly detected clusters of the second stage $(t = 2)$, by grouping the numbers of the same cluster ID with size greater than *minPts*. After T stages, we have progressively gained knowledge about the final number of clusters $\hat{k}$, since all clusters have been extracted with high probability.

The estimation of number of clusters $\hat{k}$ is a random variable, because of the randomness of the generated density levels $\varepsilon_t, t = 1, 2, \ldots, T$. For each realization of the DBSCAN-Martingale one estimation $\hat{k}$ is generated, and the final estimation of the number of clusters have been proposed [5] as the majority vote over 10 realizations of the DBSCAN-Martingale. The percentage of realizations where the DBSCAN-Martingale outputs exactly $\hat{k}$ clusters is a probability distribution, as the one shown in Fig. 1.

4 Estimation of the visual vocabulary size using the DBSCAN-Martingale

Motivated by the DBSCAN-Martingale, which has been applied in several collections of text documents in the context of news clustering [4], we propose an estimation of the total number of visual words in an image collection, as shown in Fig. 2. The proposed method is incremental, since the estimation of the final number of visual words is progressively estimated and updated when a new image is added to the collection.

Starting from the first image, keypoints are detected and SIFT descriptors [8] are extracted. Each visual feature is represented as a 128-dimensional vector, hence the whole image i is a matrix M_i with 128 columns, but the number of rows is subject to the number of detected keypoints. On each matrix M_i, the 128-dimensional vectors are clustered using the DBSCAN-Martingale, which outputs the number of dense patterns in the set of visual features, as provided by several density levels. Assuming that the application of 100 realizations of the DBSCAN-Martingale has output k_1 for the first image, k_2 for the second image and k_l for the l-th image (Fig. 2), the proposed optimal size of the visual vocabulary is:

$$k = \sum_{i=1}^{l} k_i \tag{1}$$

DBSCAN-Martingale extracts clusters sequentially, combines them into one single clustering vector and outputs the most updated estimation of the number of clusters, in each realization. The DBSCAN-Martingale requires T iterations of the DBSCAN algorithm, which runs in $\mathscr{O}(n \log n)$, when kd-tree data structures are employed for fast nearest neighbor search and in $\mathscr{O}(n^2)$ without tree-based spatial in-

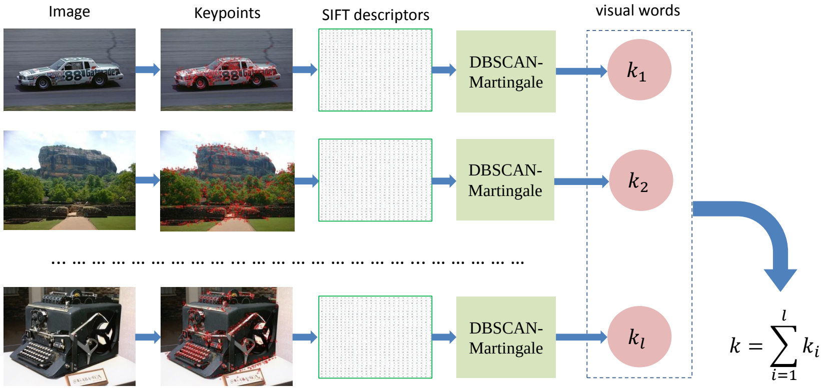


Fig. 2: The estimation of the number of visual words in an image collection. Each image i contributes with k_i visual words to the overall estimation of the visual vocabulary size.

dexing [1]. We adopt the implementation of DBSCAN-Martingale in R[1], which is available on Github[2], because the R-script utilizes the dbscan[3] package, which runs DBSCAN in $\mathscr{O}(n \log n)$. Thus, the overall complexity of the DBSCAN-Martingale is $\mathscr{O}(Tn \log n)$, where n is the number of visual descriptors per image. Assuming r iterations of the DBSCAN-Martingale per image and given an image collection of l images, the overall estimation of the size of a visual vocabulary is $\mathscr{O}(lrTn \log n)$.

In order to reduce the complexity, we sample l' out of l images to get an average number of visual words per image. The final number of visual words is estimated from a sample of images $S = \{i_1, i_2, \dots, i_{l'}\}$ of size l', so the overall complexity becomes $\mathscr{O}(l'rTn \log n)$. The final estimation for the number of visual words $\hat{k}$ of Eq. (1) becomes:

$$\hat{k} = \frac{l}{l'} \sum_{i \in S} k_i \tag{2}$$

We utilize the estimation $\hat{k}$, provided by Eq. (2), in order to cluster all visual features by $\hat{k}$ using k-means clustering. Therefore, a visual vocabulary of size $\hat{k}$ is constructed. After the construction of a visual vocabulary, as shown in Fig. 3, images are represented using term-frequency scores with inverse document frequency weighting (tf-idf) [16]:

$$\mathrm{tdidf}_{ij} = \frac{n_{id}}{n_d} \log \frac{D}{n_i} \tag{3}$$

[1] https://www.r-project.org/

[2] https://github.com/MKLab-ITI/topic-detection/blob/master/DBSCAN_Martingale.r

[3] https://cran.r-project.org/web/packages/dbscan/index.html

where n_{id} is the number of occurrences of visual word i in image d, n_d is the number of visual words in image d, n_i is the number of occurrences of visual word i in the whole image collection and D is the total number of images in the database.

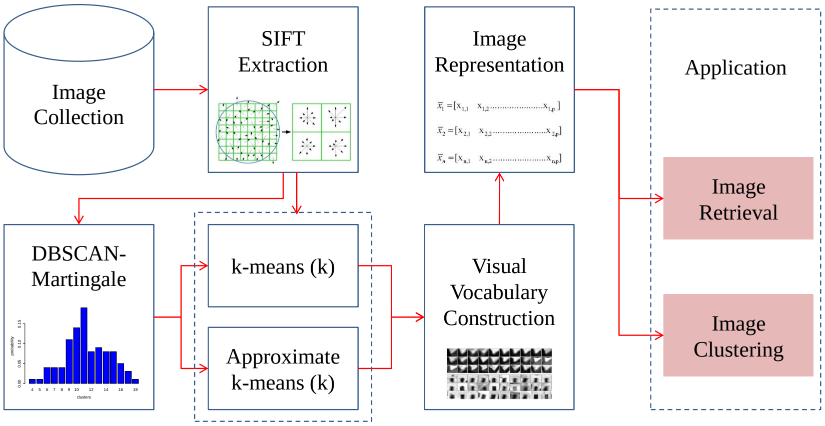


Fig. 3: The hybrid visual vocabulary construction framework using the DBSCAN-Martingale for the estimation of k and either exact or approximate k-means clustering by k. After the visual vocabulary is constructed, the collection of images is efficiently represented for any application, such as image retrieval or clustering.

In the following section, we test our hybrid visual vocabulary construction in the image retrieval and image clustering problems.

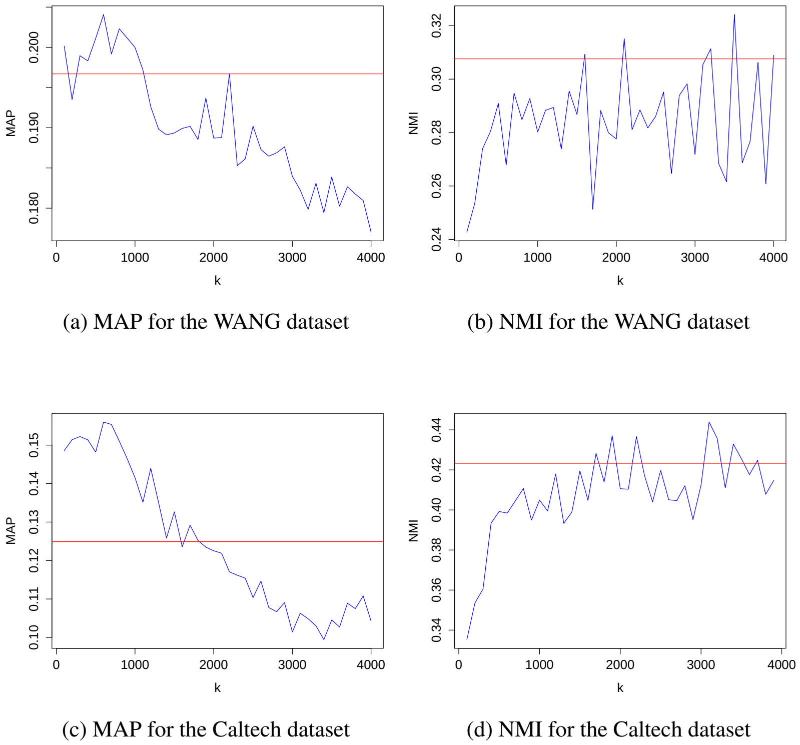
5 Experiments

We evaluate our method in the image retrieval and image clustering tasks, in which nearest neighbor search is performed in an unsupervised way. The datasets we have selected are the WANG[4] 1K and Caltech[5] 2.5K with 2,516 images, with queries as described in [4] for the image retrieval task. The number of extracted visual descriptors (SIFT) is 505,834 and 769,546 128-dimensional vectors in the WANG 1K and Caltech 2.5K datasets, respectively. The number of topics is 10 for the WANG dataset and 21 for the Caltech, allowing also image clustering experiments with the considered datasets. We selected these datasets because they are appropriate for performing both image retrieval and image clustering experiments and tuning the number of visual words k may be done in reasonable processing time, so as to eval-

[4] http://wang.ist.psu.edu/docs/related/

[5] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

uate the visual vocabulary construction in terms of Mean Average Precision (MAP) and Normalized Mutual Information (NMI).



(a) MAP for the WANG dataset

(b) NMI for the WANG dataset

(c) MAP for the Caltech dataset

(d) NMI for the Caltech dataset

Fig. 4: Evaluation using MAP and NMI in image retrieval and image clustering tasks for the WANG and Caltech datasets. The MAP and NMI scores which are obtained by our k estimation is the straight red line.

Keypoints are detected and SIFT descriptors are extracted using the LIP-VIREO toolkit[6]. For the implementation of DBSCAN-Martingale we used the R-script, which is available on Github[7] for $\varepsilon_{max} = 200$ and 100 realizations. We build one visual vocabulary for several number of visual words k, which is tuned in $k \in \{100, 200, 300, \ldots, 4000\}$. The parameter $minPts$ is tuned from 5 to 30 and the final number of clusters per image is the number which is more robust to the variations of $minPts$. In k-means clustering, we allow a maximum of 20 iterations with 5 random initial starts.

[6] http://pami.xmu.edu.cn/ wlzhao/lip-vireo.htm

[7] https://github.com/MKLab-ITI/topic-detection/blob/master/DBSCAN_Martingale.r

Our estimations for the number of visual words is $\hat{k} = 2180$ and $\hat{k} = 1840$ for the WANG and Caltech datasets, respectively, given a sample of 200 images. The corresponding MAP and NMI are compared to the best MAP and NMI scores in $k \in \{100, 200, 300, \dots, 4000\}$. The results are reported in Table 1, where apart from the best MAP and NMI scores, we also present the ratio of MAP (NMI) provided by our $\hat{k}$-estimation to the maximum observed MAP (NMI) score, denoted by r_{MAP} (r_{NMI}). In particular, in the WANG dataset, MAP is 96.42% of the best MAP observed and NMI is 94.91% of the best NMI. Similar behavior is observed in the Caltech dataset, where NMI is approached at 95.36% and MAP at 80.06%, respectively. In Fig. 4 we observe that our incremental estimation method, when combined with k-means, approaches the highest observed MAP and NMI scores in all cases examined.

Table 1: Evaluation in image retrieval and image clustering tasks.

Dataset	k visual words	MAP	r_{MAP}	NMI	r_{NMI}
WANG	best k	0.2040		0.3241	
	DBSCAN-Martingale $\hat{k}$	0.1967	0.9642	0.3076	0.9491
Caltech	best k	0.1560		0.4439	
	DBSCAN-Martingale $\hat{k}$	0.1249	0.8006	0.4233	0.9536

6 Conclusion

We presented an incremental estimation of the optimal size of the visual vocabulary, which efficiently estimates the number of visual words, and evaluated the performance of the constructed visual vocabulary in an image retrieval and an image clustering task. The proposed hybrid framework utilizes the output of DBSCAN-Martingale on a sample of SIFT descriptors, in order to be used as input in any k-means or approximate k-means clustering for the construction of a visual vocabulary. The estimation is incremental, i.e. the final number of visual words is updated when a new sample image is used. A potential limitation of our approach could appear in the case where an image exists more than once in the image collection and therefore needlessly contributes with extra visual words the final estimation. However, if the sample of images on which the DBSCAN-Martingale is applied does not have duplicate images, the overall estimation will not be affected. In the future, we plan to test our method using other visual features and in the context of multimedia retrieval, where multiple modalities are employed.

Acknowledgements This work was supported by the projects MULTISENSOR (FP7-610411) and KRISTINA (H2020-645012), funded by the European Commission.

References

1. Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod Record* (Vol. 28, No. 2, pp. 49-60).
2. Devroye, L. (1986, December). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation* (pp. 260-265). ACM.
3. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).
4. Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2016, January). Fast visual vocabulary construction for image retrieval using skewed-split kd trees. In *MultiMedia Modeling* (pp. 466-477). Springer International Publishing.
5. Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2016, July). A hybrid framework for news clustering based on the DBSCAN-Martingale and LDA., In *Machine Learning and Data Mining*, New York, USA, accepted for publication.
6. He, Y., Tan, H., Luo, W., Feng, S., & Fan, J. (2014). MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. *Frontiers of Computer Science*, 8(1), 83-99.
7. Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010, June). Aggregating local descriptors into a compact image representation. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3304-3311). IEEE.
8. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
9. Markatopoulou, F., Mezaris, V., & Patras, I. (2015, September). Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection. In Image Processing (ICIP), 2015 IEEE International Conference on (pp. 1786-1790). IEEE.
10. Mikolajczyk, K., Leibe, B., & Schiele, B. (2006, June). Multiple object class detection with a generative model. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 1, pp. 26-36). IEEE.
11. Mikulik, A., Chum, O., & Matas, J. (2013). Image retrieval for online browsing in large image collections. In *Similarity Search and Applications* (pp. 3-15). Springer Berlin Heidelberg.
12. Perronnin, F., Sanchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Computer Vision-ECCV 2010* (pp. 143-156). Springer Heidelberg.
13. Philbin, J. (2010). *Scalable object retrieval in very large image collections* (Doctoral dissertation, Oxford University).
14. Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007, June). Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (pp. 1-8). IEEE.
15. Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer Science & Business Media.
16. Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 1470-1477). IEEE.
17. Van De Sande, K. E., Gevers, T., & Snoek, C. G. (2010). Evaluating color descriptors for object and scene recognition. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9), 1582-1596.
18. Wang, J., Wang, J., Ke, Q., Zeng, G., & Li, S. (2015). Fast approximate k-means via cluster closures. In *Multimedia Data Mining and Analytics* (pp. 373-395). Springer International Publishing.
19. Zhang, J., Marszalek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2), 213-238.