Fast Visual Vocabulary Construction for Image Retrieval using Skewed-Split k-d trees

Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute, CERTH, Thessaloniki, Greece
{heliasgj,stefanos,ikom}@iti.gr

Abstract. Most of the image retrieval approaches nowadays are based on the Bag-of-Words (BoW) model, which allows for representing an image efficiently and quickly. The efficiency of the BoW model is related to the efficiency of the visual vocabulary. In general, visual vocabularies are created by clustering all available visual features, formulating specific patterns. Clustering techniques are k-means oriented and they are replaced by approximate k-means methods for very large datasets. In this work, we propose a faster construction of visual vocabularies compared to the existing method in the case of SIFT descriptors, based on our observation that the values of the 128-dimensional SIFT descriptors follow the exponential distribution. The application of our method to image retrieval in specific image datasets showed that the mean Average Precision is not reduced by our approximation, despite that the visual vocabulary has been constructed significantly faster compared to the state of the art methods.

1 Introduction

Image retrieval has become very challenging over the last years, due to the large amount of images, which are produced on a daily basis. Nowadays there are many applications of image retrieval based on the query by visual example paradigm in order to support personal photo organization, shopping assistance etc. However, one of the main challenges today is the scalability and the performance in terms of time of the image indexing and retrieval methods given the fact that they have to cope with large amounts of images in small amounts of time. 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, are computationally expensive and have been outperformed by Bag-of-Words (BoW) models.

The image representation using the BoW model is based on the construction of a visual vocabulary of all visual descriptors in a dataset, in analogy to the representation of text documents, using a vocabulary of visual words [12]. Efficient construction of visual vocabularies is done by clustering the set of all available descriptors in a dataset. In the case of large datasets, the k-means clustering techniques are replaced by approximate k-means methods [10], in order to reduce the computational cost of the visual vocabulary construction, in terms

This is a draft version. The final published version is available at: http://link.springer.com/chapter/10.1007/978-3-319-27671-7_39
Springer International Publishing Switzerland 2016
Q. Tian et al. (Eds.): MMM 2016, Part I, LNCS 9516, pp. 1–12, 2016.
DOI: 10.1007/978-3-319-27671-7 39

of time. However, even this approximate k-means algorithm [10] can be further elaborated in terms of its speed. For example, in [9] it is stated that for 17M descriptors and 1M clusters, a single iteration takes around 5 hours to complete (on a single CPU). Assuming that after 20 iterations, the approximate k-means algorithm is close to a solution, the visual vocabulary construction requires more than 4 days processing time.

Nowadays, the most popular way to represent an image in a set of vectors is using salient points such as SIFT descriptors [4].However, other image representations have been proposed for image retrieval and object detection, such as VLAD [3] and GIST [7]. In this work we focus our study on SIFT descriptors [4], which is one of the most popular descriptors used nowadays for image retrieval. After observing the values of SIFT in several datasets, we found strong evidence that the values of the SIFT descriptors are exponentially distributed, sharing a similar parameter λ, which can be quickly estimated as a function of the dataset.

The main research contributions of this work are:

- Show that the values of SIFT descriptors are exponentially distributed
- Construct faster k-d trees for SIFT descriptors by introducing a novel method called "skewed-splits"
- Build visual vocabularies quickly using "skewed-splits"

Finally we evaluate our approach by performing several image retrieval tasks with well-known image datasets.

In Section 2 we present existing approaches in the construction of visual vocabularies. In Section 3 we show that, for several datasets, the values of SIFT descriptors are exponentially distributed. The non-symmetric distribution of SIFT descriptors is utilized for a modified construction of k-d trees and visual vocabularies (Section 4), which are applied to two collections of images, in Section 5, for image retrieval.

2 Related Work

The image retrieval task was tackled as a Bag-of-Words (BoW) model initially in [12], where k-means clustering was employed for the construction of a visual vocabulary, in analogy to the text retrieval techniques. The most frequent visual words (they occur in almost all images) and very sparse terms are removed, and the final results are filtered in terms of spatial consistency. The query and each image are represented as a sparse vector of term (visual word) occurrences, which are weighted using tf-idf scores. The similarity between the query and each image is calculated, using the Mahalanobis distance.

Hierarchical k-means (HKM) was the first approximate method for fast and scalable construction of a visual vocabulary [6]. 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.

Hierarchical k-means has been outperformed by approximate k-means [10] which allows for building scalable visual vocabularies. The exact k-means algorithm involves the computation of the distances between all points and cluster centers. In contrast, this computation is replaced by the computation of the distances between points and the approximately nearest cluster centers. The approximate nearest neighbor search is performed using 8 randomized k-d trees. The efficiency of k-means increases as the number of nearest centers increases in the distance computation, but the algorithm becomes slower.

Scalability issues are often tackled using distributed processing and more than 100 processing nodes [8]. The performance of offline indexing have been improved, on a semantic level, by adding semantic attributes on the set of visual descriptors [5].

Contrary to the aforementioned approaches we present a novel method for creating visual vocabularies, in which we exploit the exponential distribution of SIFT descriptors in order to provide a faster and more efficient visual vocabulary construction. After the extraction of SIFT descriptors, we fit all SIFT values to the exponential distribution. From the estimated parameter λ of the exponential distribution, we construct a k-d tree with split value the third quartile of the exponential distribution, namely "k-d tree with skewed split". Our method is comparable to the construction of visual vocabularies using approximate k-means [10], but in each k-d tree the split value is neither the median nor the mean.

3 The exponential distribution of SIFT descriptors

The SIFT descriptors has been proved very reliable for the representation of an image [4]. For SIFT extraction we used the LIP-VIREO toolkit[1], where keypoints are detected using the Fast Hessian detector. After the extraction of SIFT features, we examine each one of the 128 coordinates separately as a sample of SIFT values, in order to test their fit to the exponential distribution.

For our experiments we have used the following image collections:

1. The Pascal voc 2007 dataset[2], containing 9,962 images (Pascal 10K) and its test set (Pascal 5K), containing 4952 images.
2. The Flickr logos dataset[3], containing 8240 images.
3. The Oxford buildings dataset[4] (Oxford 5K) has 5062 images.
4. The Caltech 101 dataset[5] has pictures of objects belonging to 101 categories, from which we get the category "airplanes" for the "Caltech 0.8K" dataset (800 images) and the categories airplanes, barrel, binocular, bonsai, brain, buddha, butterfly, camera, car_side, cellphone, chair, crab, faces, kangaroo,

[1] http://pami.xmu.edu.cn/ wlzhao/lip-vireo.htm
[2] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html
[3] http://www.multimedia-computing.de/flickrlogos/
[4] http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/
[5] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

lamp, rhino, saxophone, scissors, snoopy, umbrella, water_lilly for the "Caltech 2.5K" dataset (2,516 images).

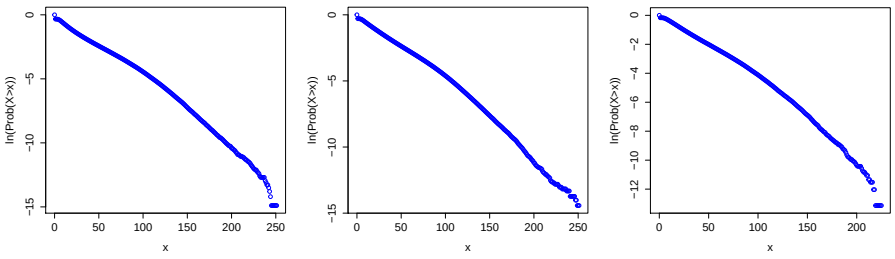
5. The WANG dataset[6] has 1K images, belonging to 10 categories.

We test whether the values of the aforementioned SIFT descriptors are exponentially distributed. Let $x \in [0, x_{max}]$ be the SIFT values for some coordinate $j = 1, 2, \ldots, 128$, and X be the random variable that generates the sample of SIFT values. We shall show that there is strong evidence that the cumulative density function fits well (for some parameter λ) to the form:

$$Prob(X \leq x) = 1 - e^{-\lambda x} \tag{1}$$

In order to test the validity of Eq. (1) we sort the SIFT values x so as to compute $Prob(X > x)$, i.e. the fraction of them that are greater than x, for all values of x. We also define $y = Prob(X > x)$ in order to test if the logarithm $\ln y$ fits to a straight line:

$$\ln y = \alpha + \beta x \tag{2}$$



(a) Pascal (b) Oxford (c) WANG

Fig. 1: The linear fit of the logarithm $\ln Prob(X > x)$

Table 1: Formulation of the hypothesis tests

Hypothesis	**for** α	**for** β
Null	$H_0 : \alpha = 0$	$H_0 : \beta = 0$
Alternative	$H_1 : \alpha \neq 0$	$H_1 : \beta \neq 0$

In all datasets examined (Table 2) we formulated two hypothesis tests, as shown in Table 1. The hypothesis $H_0 : \alpha = 0$ cannot be rejected for all levels

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

of significance because the average p-value is greater than 1%. However, the hypothesis $H_0 : \beta = 0$, under the alternative $H_1 : \beta \neq 0$, is rejected for all levels of significance because the average p-value is less than 10^{-100}. We conclude that $\alpha = 0$ and $\beta \neq 0$ and Eq. (2) is written:

$$y = e^{-\lambda x}, \quad \lambda = -\beta \tag{3}$$

Eq. (1) follows from the fact that $Prob(X > x) + Prob(X \leq x) = 1$ and $y = Prob(X > x)$. The linear fit model is also evaluated by the R-squared statistic, known also as coefficient of determination [11]. We average over the 128 coordinates of the SIFT descriptors in order to have the average R-squared statistic, which is reported in Table 2 for all datasets examined. For illustrative purposes we demonstrate, in Figure 1, the linear fit of the logarithm $\ln Prob(X > x)$ of selected coordinates for three datasets.

Table 2: The examined datasets and their corresponding fit to the exponential distribution. The average R-squared statistic is very close to 1 in all cases examined. Even for datasets which are very small, such as 800 airplane images from the Caltech dataset, the SIFT descriptors fit to the exponential distribution very well.

Dataset	**SIFT values per coordinate**	**estimated parameter λ**	**R-squared $\pm$ std**
Pascal 10K	5,884,677	0.0545	0.9914 $\pm$ 0.0066
Flickr logos 8.2K	5,803,263	0.0492	0.9927 $\pm$ 0.0052
Oxford 5K	3,678,453	0.0592	0.9883 $\pm$ 0.0055
Pascal 5K	2,940,834	0.0536	0.9893 $\pm$ 0.0094
Caltech 2.5K	769,546	0.0492	0.9877 $\pm$ 0.0101
WANG 1K	505,834	0.0632	0.9789 $\pm$ 0.0096
Caltech 0.8K	174,091	0.0567	0.9461 $\pm$ 0.0402

Using the exponential distribution of SIFT descriptors we shall provide, a faster and more efficient visual vocabulary construction, in the special case of SIFT descriptors.

4 Visual vocabulary construction using k-d trees with skewed split

In this chapter we introduce a novel methodology for creating visual vocabularies by exploiting the exponential distribution of SIFT descriptors. In order to construct the vocabulary we apply the well established framework for extracting SIFT descriptors [4]. Initially, the keypoints are detected and SIFT descriptors are extracted, which are clustered in order to provide a set of visual words (the visual vocabulary). The clustering technique is usually approximate k-means due

to the fact that exact k-means is not applicable for large datasets with billions of descriptors. A general framework is presented in Figure 2, where k-means clustering of the SIFT descriptors is replaced by approximate k-means methods. The construction of the visual vocabulary results to the image representation using tf-idf scores, i.e. weighted term frequencies. The similarity between the query and each image is calculated, using the Euclidian distance.

The main novelty of our approach is on the approximate k-means part of 2. We cluster the extracted SIFT descriptors using the conjunction of 8 k-d trees with skewed split, in order to improve the construction of the visual vocabulary.

In the following, we first discuss the construction of one single k-d tree with skewed split and, secondly, we describe the approximate clustering method using the conjunction of 8 k-d trees with skewed split.

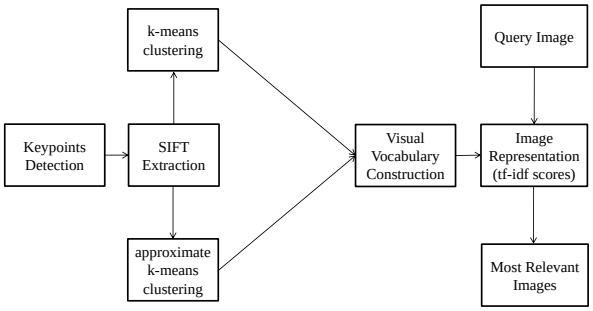


Fig. 2: Image retrieval using SIFT descriptors. After the SIFT extraction, approximate k-means methods may be used alternatively to k-means clustering for the construction of the visual vocabulary.

4.1 Construction of k-d trees with skewed split

A k-d tree selects the coordinate with maximum variance and splits the data at the median or the mean value. A randomized k-d tree picks the coordinate to split, at random, from a set of coordinates with the highest variance and the split value s_{median} is chosen to be a value close to the median. The conjunction of 8 randomized k-d trees has been proved very efficient for approximate nearest neighbor search and approximate k-means clustering, for the construction of visual vocabularies [10].

Motivated by the lack of symmetry of the exponential distribution of SIFT values, we propose an alternative split value for the construction of a k-d tree

i.e. $s_{skewed} = \ln(4)/\lambda$, which is the 3$^{\text{rd}}$ quartile of the exponential distribution, Eq. (1), with parameter λ. The mean value of the exponential distribution is $s_{mean} = 1/\lambda$ and the median is $\ln(2)/\lambda$.

In order to make the split more efficient, we need to take into account the mutual distances. To that end, we perform one simple k-means by $k = 2$ clusters, which results to two centers c_1, c_2. The split value, as obtained by k-means, is the border of the two clusters, i.e. $s_{kmeans} = (c_1 + c_2)/2$. Given a sample of 1K random numbers $u = \{u_1, u_2, \ldots, u_{1000}\}$ from the uniform distribution we generate a sample of 1K exponentially distributed points, using the transformation $-\ln(u)/\lambda$ [2]. After several simulations of exponentially generated datasets, we conclude that s_{kmeans} is much closer to s_{skewed} rather than to s_{mean} or s_{median}. Our statement is verified in Figure 3, where the candidate split values $s_{mean}, s_{median}, s_{skewed}, s_{kmeans}$ are compared.

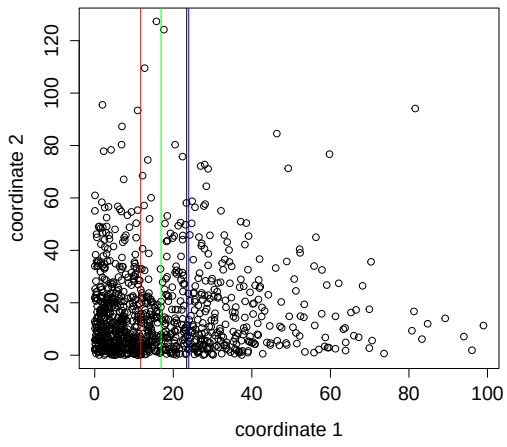


Fig. 3: The projection of 1K exponentially generated points onto the 2 dimensional plane. Choosing the coordinate 1 to split, we demonstrate the median (red line), the mean (green line), the 3$^{\text{rd}}$ quartile (black line) and the value s_{kmeans} (blue line) as provided by a simple k-means algorithm on the coordinate 1, by $k = 2$ clusters. We observe that it is hard to distinguish the 3$^{\text{rd}}$ quartile (black line) from the value s_{kmeans} (blue line), which we have chosen to be the split value.

We note that we tested the correlation of all pairs of coordinates and we did not observe any high correlation. The largest observed square of the Pearson correlation coefficient is 0.75, which is far from the values 0.90-0.99.

Algorithm 1: k-d tree with skewed split

Input: Number of generations $n < 128$, $s_{skewed} = \ln(4)/\lambda$
Output: 2^n leaves
1: Sort the 128 coordinates in decreasing variance in an index set S
2: **for** $i = 1$ **to** n **do**
3: Choose the coordinate with the i-th highest variance $S[i]$ and split at s
4: **end for**

Algorithm 1 shows the construction of a k-d tree in four steps. The parameter λ is computed by Eq. (3), as $\lambda = -\beta$, where the maximum likelihood estimate for β, for any 2-dimensional set of points $(x_i, y_i), i = 1, 2, \ldots, N$, is [1, 11]:

$$\beta = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}(\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2} \tag{4}$$

There are two main reasons that our tree is constructed faster than a k-d tree. Firstly, the variance of each coordinate is computed only once, because we sequentially split the dataset without re-calculating the variance. The variance of each coordinate is initially computed and all coordinates are sorted in decreasing variance. Starting from the coordinate with the largest variance, we split the dataset at the split value, into two leaves. For each newly generated leaf, we pick the coordinate with the second largest variance to split. The split process is repeated for all newly generated leaves, until the desired number of leaves is generated. The overall process involves the computation of all variances once for each coordinate and uses them at each step, in order to choose the coordinate to split. Secondly, the split value is fixed for all new splits and is not computed as a function of the dataset. In case a new split results to an empty new node, we set the split value to be the median, which occurs rarely. For k-d trees with 2^k leaves, it never occurred for $k < 12$. As the number of SIFT descriptors increase, it is more unlikely to get an empty new node. In all case examined, even for k-d trees with 2^{16} leaves, the split value is set to be the median for the 3% of the newly generated leaves.

4.2 Clustering using a forest of 8 k-d trees with skewed split

In the approximate nearest neighbors search, points close to the boarder of a leaf are likely to be assigned to incorrect clusters. Philbin et al. [10] used a conjunction of 8 randomized k-d trees to overcome this issue and expand the search area. In contrast, we propose the conjunction of 8 k-d trees with skewed split, tuning the split value s, as shown in Table 3, which results to an overlapping partition of the dataset. The overlapping regions of this partition define the search area of each point, in order to be assigned to its closest center. An illustrative example of overlapping regions is shown in Figure 4.

The overlapping regions defined by the conjunction of 8 k-d trees with skewed split are used for clustering. For each point, we use only one search for the closest center, within each overlapping region. The number of overlapping regions

Table 3: Tuning the split value s. The 3$^{\text{rd}}$ tree has split value the 3$^{\text{rd}}$ quartile and the 8$^{\text{th}}$ tree has split value the median.

Tree ID	**split value** s	**Tree ID**	**split value** s
1	$-\ln(0.15)/\lambda$	5	$-\ln(0.35)/\lambda$
2	$-\ln(0.20)/\lambda$	6	$-\ln(0.40)/\lambda$
3	$-\ln(0.25)/\lambda$	7	$-\ln(0.45)/\lambda$
4	$-\ln(0.30)/\lambda$	8	$-\ln(0.50)/\lambda$

determines the number of clusters because after n generations, 2^n leaves are created. Each leaf determines one region which is expanded by a collection of 8 trees with (skewed) split values close to the third quartile. Two points in each expanded region of 8 leaves are considered to be approximately close to each other and, in the case of k-means clustering, the search for the closest center is restricted to each region. Therefore, our overall clustering method coincides with one iteration of the approximate k-means algorithm [10], in terms of time, and the computational cost of our approach is $\mathcal{O}(N \log K)$, since we search over the (approximate) closest centers only once, in order to assign points to clusters.

In the following section, we test whether our significantly faster method provides also visual vocabularies of high quality.

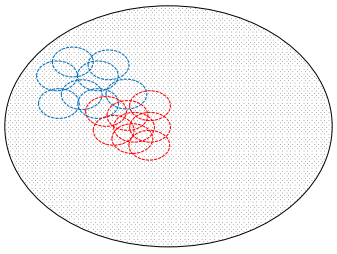


Fig. 4: Two cells of the overlapping partition. Each tree leaf is illustrated by a circle. The union of 8 blue leaves is not disjoint with the union of the 8 red leaves.

5 Application to Image Retrieval

In order to evaluate our method we perform image retrieval experiments with the datasets Caltech 2.5K and the WANG 1K (Table 2). Since the best performing methods for fast construction of visual vocabularies are based on approximate k-means clustering and k-d trees, we create a visual vocabulary based on approximate clustering and k-d trees with skewed split.

We build one visual vocabulary using the conjunction of 8 randomized k-d trees as in [10] and another one visual vocabulary using the conjunction of 8 k-d trees with skewed-split. For the approximate k-means clustering of [10], we allow 20 iterations and the same number of nearest neighbors to search.

After the construction of each visual vocabulary, the tf-idf scores are computed [12]:

$$tdidf_{ij} = \frac{n_{id}}{n_d} \log \frac{D}{n_i}$$

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

We evaluate our method on 2 datasets of Table 2, namely the Caltech 2.5K and the WANG 1K. The 21 query images of the Caltech dataset are demonstrated in Figure 5 and the 10 query images of the WANG dataset are demonstrated in Figure 6. The experiments were performed on an Intel Core i7-4790K CPU at 4.00GHz with 16GB RAM memory, using a single thread. For the statistical analysis of SIFT descriptors and the construction of k-d trees (with and without skewed-splits) we used the R programming language[7].

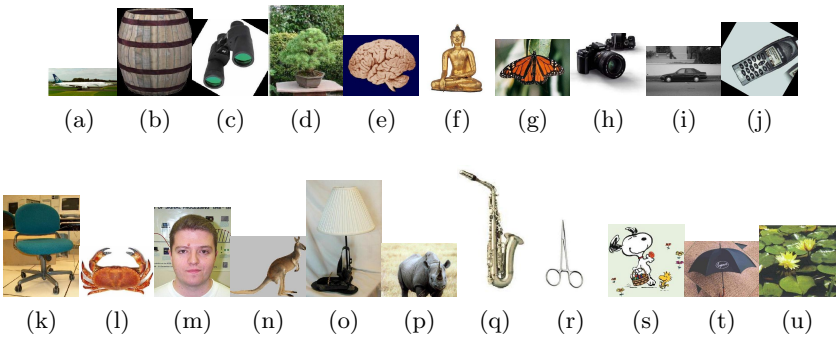


Fig. 5: The query images used for evaluation from the Caltech dataset

For each query image, we compute the average precision, defined as the area under the precision-recall curve. Averaging for all queries, we obtain the mean Average Precision (mAP) for each visual vocabulary. In Table 4 we present the results in two selected datasets of Table 2, with different numbers of clusters, tuning the number of visual words in $2^n, n \in \{10, 13, 14, 15, 16\}$.

The reported mAP of the WANG dataset for 1024 clusters is higher in the baseline method. In all other cases we outperform the baseline method, not only in terms of speed, but also in terms of the mean Average Precision. As the number of SIFT descriptors and clusters increase, we observe that our method

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

(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 6: The query images used for evaluation from the WANG dataset

Table 4: The mean Average Precision for two datasets of Table 1 for several number of leaves in the conjunction of 8 trees.

mAP	Caltech		WANG	
clusters	without skewed split	with skewed split	without skewed split	with skewed split
1024	0.1100	**0.1167**	**0.2061**	0.1562
8192	0.0812	**0.1168**	0.1405	**0.1457**
16384	0.0769	**0.1009**	0.1382	**0.1457**
32768	0.0733	**0.0932**	0.1317	**0.1389**
65536	0.0731	**0.0935**	0.1314	**0.1378**

performs better than the baseline approach, due to the statistical approach we have adopted. For relatively small datasets, such as the WANG dataset, and low levels of the k-d trees, we expect that the statistical laws become weak.

6 Conclusion

In this paper we present strong evidence that SIFT descriptors are exponentially distributed. Using the highly skewed distribution of SIFT values, we proposed an alternative split value for the construction of k-d trees. Using the BoW model for image representation, we introduced a novel method for the construction of visual vocabularies. Our tree construction of k-d trees with skewed split and the proposed clustering method are significantly faster than the corresponding baseline method. However, there are some limitations, which need to be considered, for example the fact that the number of visual words cannot be greater than 2^{128}. This is a very large number (greater than 10^{38}) and cannot easily be reached. The application of our model to the image retrieval task has shown that we obtain slightly better mAP than the baseline method, in most cases, but at a small percentage of its computational cost. Even for datasets which are very small, the SIFT descriptors fit to the exponential distribution very well and the mAP is not reduced, when compared to the baseline method.

In the future, we plan to test the statistical properties of other visual features beyond SIFT descriptors. The distribution of the visual features is crucial for fast construction of visual vocabularies.

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

References

1. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
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. 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.
4. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
5. Luo, Q., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2014, April). Superimage: Packing semantic-relevant images for indexing and retrieval. In *Proceedings of International Conference on Multimedia Retrieval* (p. 41). ACM.
6. 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.
7. 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.
8. Moise, D., Shestakov, D., Gudmundsson, G., & Amsaleg, L. (2013, April). Indexing and searching 100M images with Map-Reduce. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval* (pp. 17-24). ACM.
9. Philbin, J. (2010). *Scalable object retrieval in very large image collections* (Doctoral dissertation, Oxford University).
10. 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.
11. Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer Science & Business Media.
12. 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.