Improving Visual Vocabularies: a More Discriminative, Representative and Compact Bag of Visual Words

Leonardo Chang, Airel Pérez-Suárez and José Hernández-Palancar
Advanced Technologies Application Center, 7A #21406, Siboney, Playa, Havana, Cuba, C.P. 12220
E-mail: {lchang,asuarez,jpalancar}@cenatav.co.cu

Miguel Arias-Estrada and L. Enrique Sucar
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, México, C.P. 72840
E-mail: {ariasmo,esucar}@inaoep.mx

Keywords: bag of visual words improvement, bag of features, visual vocabulary, visual codebook, object categorization, object class recognition, object classification

Received: July 21, 2016

In this paper, we introduce three properties and their corresponding quantitative evaluation measures to assess the ability of a visual word to represent and discriminate an object class, in the context of the BoW approach. Also, based on these properties, we propose a methodology for reducing the size of the visual vocabulary, retaining those visual words that best describe an object class. Reducing the vocabulary will provide a more reliable and compact image representation. Our proposal does not depend on the quantization method used for building the set of visual words, the feature descriptor or the weighting scheme used, which makes our approach suitable to any visual vocabulary. Throughout the experiments we show that using only the most discriminative and representative visual words obtained by our proposed methodology improves the classification performance; the best results obtained with our proposed method are statistically superior to those obtained with the entire vocabularies. In the Caltech-101 dataset, average best results outperformed the baseline by a 4.6% and 4.8% in mean classification accuracy using SVM and KNN, respectively. In the Pascal VOC 2006 dataset there was a 1.6% and 4.7% improvement for SVM and KNN, respectively. Furthermore, these accuracy improvements were always obtained with more compact representations. Vocabularies 10 times smaller always obtained better accuracy results than the baseline vocabularies in the Caltech-101 dataset, and in the 93.75% of the experiments on the Pascal VOC dataset.

Povzetek: S pomočjo rudarjenja podatkov se prispevek ukvarja z iskanjem besed za razločevanje razredov objektov.

1 Introduction

One of the most widely used approaches for representing images for object categorization is the Bag of Words (BoW) approach [5]. BoW-based methods have obtained remarkable results in recent years and they even obtained the best results for several classes in the recent PASCAL Visual Object Classes Challenge on object classification [8]. The key idea of BoW approaches is to discretize the entire space of local features (e.g., SIFT [22]) extracted from a training set at interest points or densely sampled in the image. With this aim, clustering is performed over the set of features extracted from a training set in order to identify features that are visually equivalent. Each cluster is interpreted as a visual word, and all clusters form a so-called visual vocabulary. Later, in order to represent an unseen image, each feature extracted from the image is assigned to a visual word of the visual vocabulary; from which a histogram of occurrences of each visual word in the image is obtained, as illustrated in Figure 1.

One of the main limitations of the BoW approach is that the visual vocabulary is built using features that belong to both the object and the background. This implies that the noise extracted from the image background is also considered as part of the object class description. Also, in the BoW representation, every visual word is used, regardless of its low representativeness or discriminative power. These elements may limit the quality of further classification processes. In addition, there is no consensus about which is the optimal way for building the visual vocabulary, i.e., the clustering algorithm used, the number of clusters (visual words) that best describe the object classes, etc. When dealing with relatively small vocabularies, clustering can be executed several times and the best performing vocabulary can be selected through a validation phase. However, this becomes intractable for large image collections.

In this paper, we propose three properties to assess the ability of a visual word to represent and discriminate an

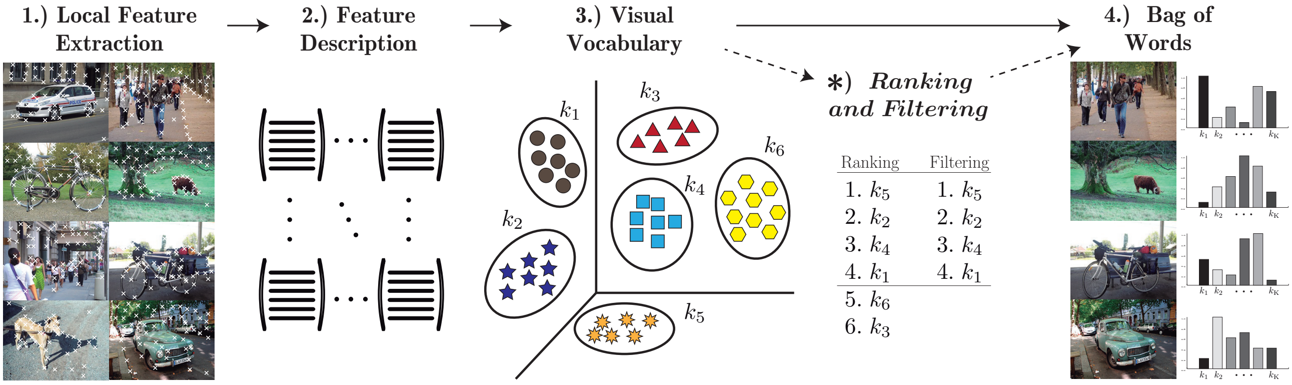


Figure 1: Classical BoW approach overview (steps 1 to 4). First, regions/points of interest are automatically detected and local descriptors over those regions/points are computed (step 1 and 2). Later in step 3, the descriptors are quantized into visual words to form the visual vocabulary. Finally, in step 4, the occurrences in the image of each specific word in the vocabulary for constructing the BoW feature are found. In this work, we propose to introduce step (∗) in order to use only the most discriminative and representative visual words from the visual vocabulary in the BoW representation.

object class in the context of the BoW approach. We define three measures in order to quantitatively evaluate each of these properties. The visual words that best represent a class, best generalize over intra-class variability and best differentiate between object classes will obtain the highest scores for these measures. A methodology for reducing the size of the visual vocabulary based on these properties is also proposed. Our proposal does not depend on the clustering method used to create the visual vocabulary, the descriptor used (e.g., SIFT, SURF, etc.) or the weighting scheme used (e.g., *tf*, *tf-idf*, etc.) Therefore, it can be applied to any visual vocabulary to improve its representativeness, since it does not build a new visual vocabulary, it rather finds the best visual words of a given visual vocabulary.

Experiments conducted on the Caltech-101 [10] and Pascal VOC 2006 [9] datasets, in a classification task, demonstrate the improvement introduced by the proposed method. Tested with different vocabulary sizes, different interest points extraction and description methods, and different weighting schemas, the classification accuracies achieved using the entire vocabulary were always statistically inferior to those achieved by several of the vocabularies obtained by filtering the baseline vocabulary, using our proposed vocabulary size reducing methodology. Moreover, the best results were obtained with as few as the 13.4% and 17.2%, in average, of the baseline visual words for the Caltech-101 and Pascal VOC 2006 datasets, respectively. Compared with a state-of-the-art mutual information based method for feature selection our proposal obtains superior classification accuracy results for the highest compression rates and comparable results for the other filtering sizes.

The paper is organized as follows: Section 2 gives an overview on related works for building more discriminative and representative visual vocabularies. Section 3 introduces the proposed properties and measures for the evaluation of the representativeness and distinctiveness of visual words. The performance of our proposed method on two data sets and a discussion of the obtained results are presented in Section 4. Finally, Section 5 concludes the paper with a summary of our findings and a discussion of future work.

2 Related work

Several methods have been proposed in the literature to overcome the limitations of the BoW approach [25]. These include part generative models and frameworks that use geometric correspondence [30, 23], works that deal with the quantization artifacts introduced while assigning features to visual words [15, 11], techniques that explore different features and descriptors [24, 12], among many others. In this section, we briefly review some recent methods aimed to build more discriminative and representative visual vocabularies, which are more related to our work.

Kersorn and Poslad [17] presented a framework to improve the quality of visual words by constructing visual words from representative keypoints. Also, domain specific non-informative visual words are detected using two main characteristics for non-informative visual words: high document frequency and a small statistical association with all the concepts in the collection. In addition, the vector space model of visual words is restructured with respect to a structural ontology model in order to solve visual synonym and polysemy problems.

Zhang *et al.* [29] proposed to obtain a visual vocabulary comprised of descriptive visual words and descriptive visual phrases as the visual correspondences to text words and phrases. Authors state that a descriptive visual element can be composed by the visual words and their combinations and that these combinations are effective in representing certain visual objects or scenes. Therefore, they define visual phrases as frequently co-occurring visual word pairs.

Lopez-Sastre *et al.* [21] presented a method for building a more discriminative visual vocabulary by taking into account the class labels of images. The authors proposed a cluster precision criterion based on class labels in order to obtain class representative visual words through a Reciprocal Nearest Neighbors clustering algorithm. Also, they introduced an adaptive threshold refinement scheme aimed to increase vocabulary compactness.

Liu [19] builds a visual vocabulary based on a Gaussian Mixed Model (GMM). After K-Means clusters are obtained, GMM is then used to model the distribution of each cluster. Each GMM will be used as a visual word of the visual vocabulary. Also, a soft assignment schema for the bag of words is proposed based on the soft assignment of image features to each GMM visual word.

Liu and Shah [20] exploit mutual information maximization techniques to learn a compact set of visual words and to determine the size of the codebook. In their proposal two codebook entries are merged if they have comparable distributions. In addition, spatio-temporal pyramid matching is used to exploit temporal information in videos.

Most popular visual descriptors are histograms of image measurements. It has been shown that with histogram features, the Histogram Intersection Kernel (HIK) is more effective than the Euclidean distance in supervised learning tasks. Based on this assumption, Wu *et al.* [28] proposed a histogram kernel k-means algorithm which use HIK in an unsupervised manner to improve the generation of visual codebooks.

In [4], in order to use low level features extracted from images to create higher level features, Chandra *et al.* proposed a hierarchical feature learning framework that uses a Naive Bayes clustering algorithm. First, SIFT features over a dense grid are quantized using K-Means to obtain the first level symbol image. Later, features from the current level are clustered using a Naive Bayes-based clustering and quantized to get the symbol image at the next level. Bag of words representations can be computed using the symbol image at any level of the hierarchy.

Jiu *et al.* [16], motivated by obtaining a visual vocabulary highly correlated to the recognition problem, proposed a supervised method for joint visual vocabulary creation and class learning, which uses the class labels of the training set to learn the visual words. In order to achieve that, they proposed two different learning algorithms, one based on error backpropagation and the other one based on cluster label reassignment.

In [27], the authors propose a hierarchical visual word mergence framework based on graph-embedding. Given a predefined large set of visual words, their goal is to hierarchically merge them into a small number of visual words, such that the lower dimensional image representation obtained based on these new words can maximally maintain classification performance.

Zhang *et al.* [31] proposed a supervised Mutual Information (MI) based feature selection method. This algorithm uses MI between each dimension of the image descriptor and the image class label to compute the dimension importance. Finally, using the highest importance values, they reduce the image representation size. This method achieve higher accuracy and less computational cost than feature compression methods such as product quantization [14] and BPBC [13].

In our work, similarly to [17, 21, 16], we also use the class labels of images. However, we do not use the class labels to create a new visual vocabulary but for scoring the set of visual words, according to their distinctiveness and representativeness for each class. It is important to emphasize that our proposal does not depend on the algorithm used for building the set of visual words, the descriptor used nor the weighting scheme used. The previously mentioned characteristics make our approach suitable to any visual vocabulary since it does not build a new visual vocabulary, it rather finds the best visual words of a given visual vocabulary. In fact, our proposal could directly complement all the above discussed methods, by ranking their resulting vocabularies according to the distinctiveness and representativeness of the obtained visual words, although is out of the scope of this paper to explore it.

3 Proposed method

Visual vocabularies are commonly comprised by a lot of noisy visual words due to intra-class variability and the inclusion of features from the background during the vocabulary building process, among others. Later, for image representation every visual word is used, which may lead to an error-prone image representation.

In order to improve image representations, we introduce three properties and their corresponding quantitative evaluations to assess the ability of a visual word to represent and discriminate an object class in the context of the BoW approach. We also propose a methodology, based on these properties, for reducing the size of the visual vocabulary, discarding those visual words that worst describe an object class (i.e., noisy visual words). Reducing the vocabulary in such a manner will allow to have a more reliable and compact image representation.

We would like to emphasize that all the measures proposed in this section are used during the training phase; therefore, we can use all the knowledge about the data that is available during this phase.

3.1 Inter-class representativeness measure

A visual word could be comprised of features from different object classes, representing visual concepts or parts of objects common to those different classes. These common parts or concepts do not have necessarily to be equally represented inside the visual word because, even when similar, object classes should also have attributes that differentiate them. Therefore, we can say that, in order to represent an object class the best, a property that a visual word must satisfy is to have a high representativeness of this class. In order to measure the representativeness of a class c_j in visual word k, the measure $\mathcal{M}_1$ is proposed:

$$\mathcal{M}_1(k, c_j) = \frac{f_{k,c_j}}{n_k}, \tag{1}$$

where f_{k,c_j} represents the number of features of class c_j in visual word k and n_k is the total number of features in visual word k.

Figure 2 shows $\mathcal{M}_1$ values for two example visual words. In Figure 2 a) the 'blue' class has a very high value of $\mathcal{M}_1$ because most of the features in the visual word belong to the ○ class, being the opposite for the classes □ and △ that are poorly represented in the visual word. Figure 2 b) shows an example visual word where every class is nearly equally represented, therefore every class have similar $\mathcal{M}_1$ values.

3.2 Intra-class representativeness measure

A visual word could be comprised of features from different objects, many of them probably belonging to the same object class. Even when different, object instances from the same class should share several visual concepts. Taking this into account, we can state that a visual word best describes a specific object class while more balanced are the features from that object class comprising the visual word, with respect to the number of different training objects belonging to that class. Therefore, we could say that, in order to represent an object class the best, a property that a visual word must satisfy is to have a high generalization or intra-class representativeness over this class.

To measure the intra-class representativeness of a visual word k for a given object category c_j, the measure μ is proposed:

$$\mu(k, c_j) = \frac{1}{O_{c_j}} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{f_{k,c_j}} - \frac{1}{O_{c_j}} \right|, \tag{2}$$

where O_{c_j} is the number of objects (images) of class c_j in the training set. o_{m,k,c_j} is the number of features extracted from object m of class c_j in visual word k, and f_{k,c_j} is the number of features of class c_j that belong to visual word k. The term $1/O_{c_j}$ represents the ideal ratio of features of class c_j that guarantees the best balance, i.e., the case where each object of class c_j is equally represented in visual word k.

The measure μ evaluates how much a given class deviates from its ideal value of intra-class variability balance. In order to make this value comparable with other classes and visual words, μ could be normalized using its maximum possible value, which is $\frac{2 \cdot O_{c_j} - 2}{O_{c_j}^2}$.

Taking into account that μ takes its maximum value in the worst case of intra-class representativeness, the measure $\mathcal{M}_2$ is defined to take its maximum value in the case of ideal intra-class variability balance and to be normalized by $\max(\mu(k, c_j))$:

$$\mathcal{M}_2(k, c_j) = 1 - \frac{O_{c_j}}{2 \cdot (O_{c_j} - 1)} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{f_{k,c_j}} - \frac{1}{O_{c_j}} \right|. \tag{3}$$

Figure 3 shows the values of $\mathcal{M}_2$ on two example visual words. In Figure 3 a), the number of features from the different images of the ○ class in the visual word is well balanced, i.e., the visual word generalizes well over intra-class variability for the ○ class, hence this class presents a high $\mathcal{M}_2$ value. In contrast, in Figure 3 b) only one image from the ○ class is well represented by the visual word. As the visual word represents a visual characteristic only present in one image, it is not able to well represent intra-class variability, therefore, the ○ class will have a low value of $\mathcal{M}_2$ in this visual word.

3.3 Inter-class distinctiveness measure

$\mathcal{M}_1$ and $\mathcal{M}_2$ provide, under different perspectives, a quantitative evaluation of the ability of a visual word to describe a given class. However, we should not build a vocabulary just by selecting those visual words that best represent each object class, because this fact does not directly imply that the more representative words will be able to differentiate well one class from another, as a visual vocabulary is expected to do. Therefore, we can state that, in order to be used as part of a visual vocabulary, a desired property of a visual word is that it should have high values of $\mathcal{M}_1(k, c_j)$ and $\mathcal{M}_2(k, c_j)$ (represents well the object class), while having low values of $\mathcal{M}_1(k, \{c_j\}^C)$ and $\mathcal{M}_2(k, \{c_j\}^C)$ (misrepresents the rest of the classes), i.e., it must have high discriminative power.

In order to quantify the distinctiveness of a visual word for a given class, the measure $\mathcal{M}_3$ is proposed. $\mathcal{M}_3$ expresses how much the object class that is best represented by visual word k is separated from the other classes in the $\mathcal{M}_1$ and $\mathcal{M}_2$ rankings.

Let $\Theta_{\mathcal{M}}(K, c_j)$ be the set of values of a given measure $\mathcal{M}$ for the set of visual words $K = \{k_1, k_2, ..., k_N\}$ and the object class c_j, sorted in descending order of the value of $\mathcal{M}$. Let $\Phi(k, c_j)$ be the position of visual word $k \in K$ in $\Theta_{\mathcal{M}}(K, c_j)$. Let $P_k = \min_{c_j \in C}(\Phi(k, c_j))$ be the best position of visual word k in the set of all object classes $C = \{c_1, c_2, ..., c_Q\}$. Let $c_k = \arg\min_{c_j \in C}(\Phi(k, c_j))$ be the object class where k has position P_k. Then, the inter-class distinctiveness (measure $\mathcal{M}_3$), of a given visual word k for a given measure $\mathcal{M}$, is defined as:

$$\mathcal{M}_3(k, \mathcal{M}) = \frac{1}{(|C| - 1)(|K| - 1)} \sum_{c_j \neq c_k} (\Phi(k, c_j) - P_k). \tag{4}$$

In Figure 4, the $\mathcal{M}_3$ measure is calculated for two visual words (i.e., k_2 and k_5) of a six visual words and three

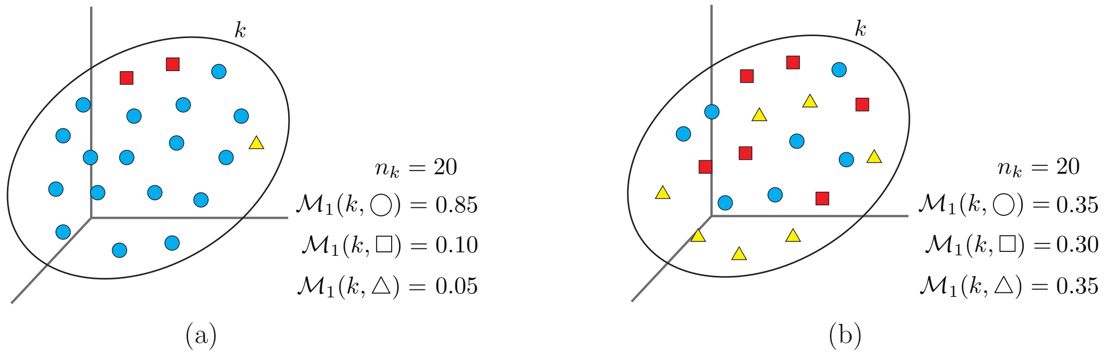

k
$n_k = 20$
$\mathcal{M}_1(k, \bigcirc) = 0.85$
$\mathcal{M}_1(k, \square) = 0.10$
$\mathcal{M}_1(k, \triangle) = 0.05$
(a)
k
$n_k = 20$
$\mathcal{M}_1(k, \bigcirc) = 0.35$
$\mathcal{M}_1(k, \square) = 0.30$
$\mathcal{M}_1(k, \triangle) = 0.35$
(b)

Figure 2: *(best seen in color.)* Examples of $\mathcal{M}_1$ measure values for a) a visual word with a well-defined representative class (○ class with high $\mathcal{M}_1$ value, □ and △ classes with low $\mathcal{M}_1$ values) and b) a visual word without any highly representative class (○, □ and △ classes have low and very similar $\mathcal{M}_1$ values).

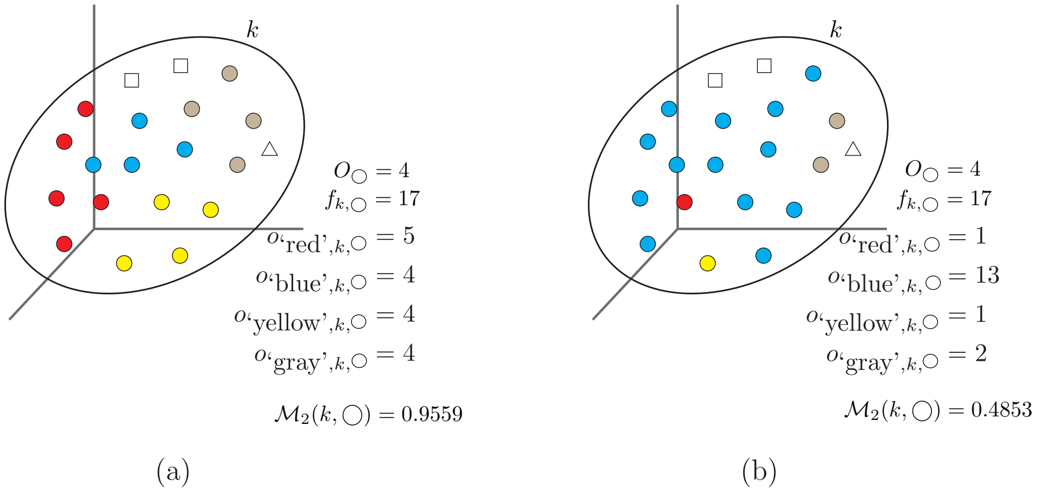

k
$O_{\bigcirc} = 4$
$f_{k,\bigcirc} = 17$
$o_{\text{'red'},k,\bigcirc} = 5$
$o_{\text{'blue'},k,\bigcirc} = 4$
$o_{\text{'yellow'},k,\bigcirc} = 4$
$o_{\text{'gray'},k,\bigcirc} = 4$
$\mathcal{M}_2(k, \bigcirc) = 0.9559$
(a)
k
$O_{\bigcirc} = 4$
$f_{k,\bigcirc} = 17$
$o_{\text{'red'},k,\bigcirc} = 1$
$o_{\text{'blue'},k,\bigcirc} = 13$
$o_{\text{'yellow'},k,\bigcirc} = 1$
$o_{\text{'gray'},k,\bigcirc} = 2$
$\mathcal{M}_2(k, \bigcirc) = 0.4853$
(b)

Figure 3: *(best seen in color.)* Examples of $\mathcal{M}_2$ measure values for the ○ class in a) a visual word where there is a good balance between the number of features of different images of the ○ class (high $\mathcal{M}_2$ value), and in b) the opposite case where only one image for the ○ class is predominantly represented in the visual word (low $\mathcal{M}_2$ value). In the figure, different fill colors of each feature in the visual word represent features extracted from different object images of the same class.

classes example. Visual word k_2 is among the top items of the representativeness ranking for every class in the example. Despite this, k_2 has low discriminative power because describing well several classes makes harder the process of differentiate one class from another. In contrast, visual word k_5 is highly discriminative because it describes well only one class.

3.4 On ranking and reducing the size of visual vocabularies

The proposed measures, provide a quantitative evaluation of the representativeness and distinctiveness of the visual words in a vocabulary for each class. The visual words that best represent a class, best generalize over intra-class variability and best differentiate between object classes will obtain the highest scores for these measures. In this section, we present a methodology for ranking and reducing the size of the visual vocabularies, towards more reliable and compact image representations.

Let $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$ be the rankings of vocabulary K, using measures $\mathcal{M}_3(K, \mathcal{M}_1)$ and $\mathcal{M}_3(K, \mathcal{M}_2)$, respectively. $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$ provide a ranking of the vocabulary based on the distinctiveness of visual words according to inter-class and intra-class variability, respectively.

In order to find a consensus, $\Theta(K)$, between both rankings $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$ a consensus-based voting method can be used; in our case, we decided to use the Borda Count algorithm [7] although any other can be used as well. The Borda Count algorithm obtains a final ranking from multiple rankings over the same set. Given $|K|$ visual words, a visual word receive $|K|$ points for a first preference, $|K| - 1$ points for a second preference, $|K| - 2$ for a third, and so on for each ranking independently. Later, individual values for each visual word are added and a final ranking obtained.

From this final ranking a reduced vocabulary can be obtained by selecting the first N visual words. As pointed in [20], the size of the vocabulary affects the performance and there is a vocabulary size which can achieve maximal accuracy, which depends on the dataset, the number of classes and the data nature, among others. In our experiments, we explore different vocabulary sizes, over different datasets, different interest points extraction and description methods, different weighting schemas, and different classifiers.

4 Experimental evaluation

In this work we have presented a methodology for improving BoW-based image representation by using only the most representative and discriminative visual words in the vocabulary. As it was stated in previous sections, our proposal does not depend on the algorithm used for building the set of visual words, the descriptor used nor the weighting scheme used. Therefore, the proposed methodology could be applied for improving the accuracy of any of the methods reported in the literature, which are based on a BoW approach.

The main goal of the experiments we present in this section is to quantitatively evaluate the improvement introduced by our proposal to the BoW-based image representation, over two standard datasets commonly used in object categorization. The experiments were focused on: a) to assess the validity of our proposal in a classic BoW-based classification task, b) to evaluate the methodology directly with respect to other kinds of feature selection algorithms, and c) to measure the time our methodology spent in order to filter the visual vocabulary built for each dataset. All the experiments were done on a single thread of a 3.6 GHz Intel i7 processor and 64GB RAM PC.

The experiments conducted in order to evaluate our proposal were done in two well-known datasets: Caltech-101 [10] and Pascal VOC 2006 [9].

The Caltech-101 dataset [10] consists of 102 object categories. There are about 40 to 800 images per category and most categories have about 50 images. The Pascal VOC 2006 dataset [9] consists of 10 object categories. In total, there are 5304 images, split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets.

4.1 Assessing the validity in a BoW-based classification task

As it was mentioned before, the goal of the first experiment is to assess the validity of our proposal. With this aim, we evaluate the accuracy in a classic BoW-based classification task, with and without applying our vocabulary filtering methodology.

In the experiments presented here, we use for image representation the BoW schema presented in Figure 1 with the following specifications:

- Interest points are detected and described using two methods: SIFT [22] and SURF [2].
- K-means, with four different K values, is used to build the visual vocabularies; these vocabularies constitute the baseline. For both Caltech-101 and Pascal VOC 2006 datasets we used K=10000, 15000, 20000 and 25000.
- Each of the baseline vocabularies is ranked using our proposed visual words ranking methodology.
- Later, nine new vocabularies are obtained by filtering each baseline vocabulary, leaving the 10%, 20%, ..., 90%, respectively, of the most representative and discriminative visual words based on the obtained ranking.
- Two weighting schemas are used for image representation: *tf* and *tf-idf*.

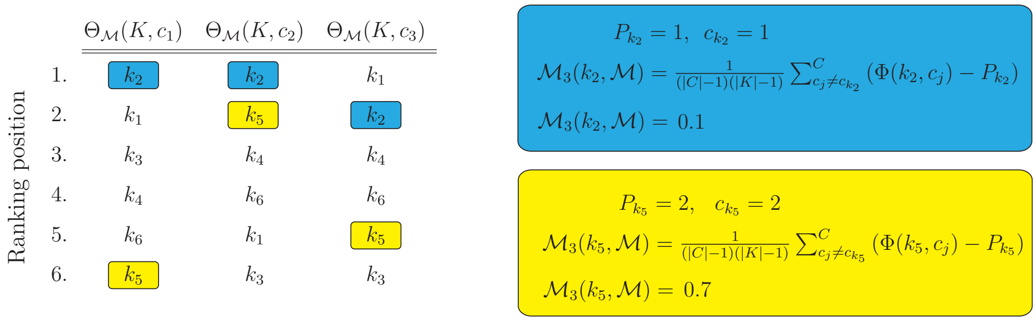

$\Theta_{\mathcal{M}}(K,c_1)$
$\Theta_{\mathcal{M}}(K,c_2)$
$\Theta_{\mathcal{M}}(K,c_3)$
Ranking position
1. k_2 k_2 k_1
2. k_1 k_5 k_2
3. k_3 k_4 k_4
4. k_4 k_6 k_6
5. k_6 k_1 k_5
6. k_5 k_3 k_3
$P_{k_2} = 1, \ c_{k_2} = 1$
$\mathcal{M}_3(k_2, \mathcal{M}) = \frac{1}{(|C|-1)(|K|-1)} \sum_{c_j \neq c_{k_2}}^{C} (\Phi(k_2, c_j) - P_{k_2})$
$\mathcal{M}_3(k_2, \mathcal{M}) = 0.1$
$P_{k_5} = 2, \ c_{k_5} = 2$
$\mathcal{M}_3(k_5, \mathcal{M}) = \frac{1}{(|C|-1)(|K|-1)} \sum_{c_j \neq c_{k_5}}^{C} (\Phi(k_5, c_j) - P_{k_5})$
$\mathcal{M}_3(k_5, \mathcal{M}) = 0.7$

Figure 4: *(best seen in color.)* Example of $\mathcal{M}_3$ measure for two visual words. k_2 has low discriminative power because it represents well several classes, while k_5 has high discriminative power because it describes well only one class.

– For both datasets we randomly selected 10 images from all the categories for building the visual vocabularies. The rest of the images were used as test images; but, as in [18], we limited to 50 the number of test images per category.

After that, we tested the obtained visual vocabularies in a classification task, using SVM (with a linear kernel) and KNN (where K is optimized with respect to the leave-one-out error) as classifiers. For each visual vocabulary, test images are represented using this vocabulary and, a 10-fold 10-times cross-validation process is conducted, where nine of the ten partitions are used for training and the other one for testing the trained classifier. The mean classification accuracy along the ten iterations is reported.

Figures 5 and 6 show the mean classification accuracy results over the cross-validation process using SVM and KNN, respectively, on the Caltech-101 dataset. Figures 7 and 8 show the same for the Pascal VOC 2006 dataset. In Figures 5 to 8, subfigures (a) and (b) show the results using SIFT descriptor; results for SURF are shown in subfigures (c) and (d). Results for the two different weighting schemas, i.e., *tf* and *tf-idf*, are shown in subfigures (a) (c), and (b) (d), respectively.

It can be seen that in both datasets, for every configuration, our proposed methodology allows to obtain reduced vocabularies that outperformed the classical BoW approach (baseline).

Table 2 summarizes the results presented in Figures 5 and 6 for the Caltech-101 dataset. The results in Figures 7 and 8 are summarized in Table 3. For every experiment configuration, Tables 2 and 3 show the baseline classification accuracy against the best result obtained by the proposed method with both SVM and KNN classifiers. The size of the filtered vocabulary in which the best result was obtained is also showed.

4.1.1 Discussion

The experimental results presented in this section validate the claimed contributions of our proposed method. As it can be seen in Tables 2 and 3, the best results obtained with our proposed method outperform those obtained with the whole vocabularies. For the experiments conducted in the Caltech-101 dataset, our average best results outperformed the baseline by a 4.6% and 4.8% in mean classification accuracy using SVM and KNN, respectively. In the Pascal VOC 2006 dataset there was a 1.6% and 4.7% improvement for SVM and KNN, respectively. As noticed on Figures 5 to 8, there is a trend of the performance with respect to the filter size in the two considered datasets, i.e., for smaller filter sizes higher accuracy.

In order to validate the improvement obtained by the proposed method, the statistical significance of the obtained results was verified. For testing the statistical significance we used the Mann-Whitney test, with a 95% of confidence. A detailed explanation about Mann-Whitney test, as well as an implementation, can be found in [1]. As a result of this test, it has been verified that the results obtained in both datasets, by the proposed method, are statistically superior to those obtained by the baseline.

In addition, the best results using the filtered vocabularies were obtained with vocabularies several times smaller than the baseline vocabularies, i.e., 6 and 10 times smaller in average using SVM and KNN, respectively, for the Caltech-101 datasets, and 8 and 5 times smaller in average for the Pascal VOC 2006 dataset with SVM and KNN, respectively. Furthermore, vocabularies 10 times smaller always obtained better accuracy results than the baseline vocabularies in the Caltech-101 dataset, and in the 93.75% of the experiments on the Pascal VOC 2006 dataset. Obtaining smaller vocabularies implies more compact image representations, that will have a direct impact on the efficiency of further processing based on these image representations, and less memory usage.

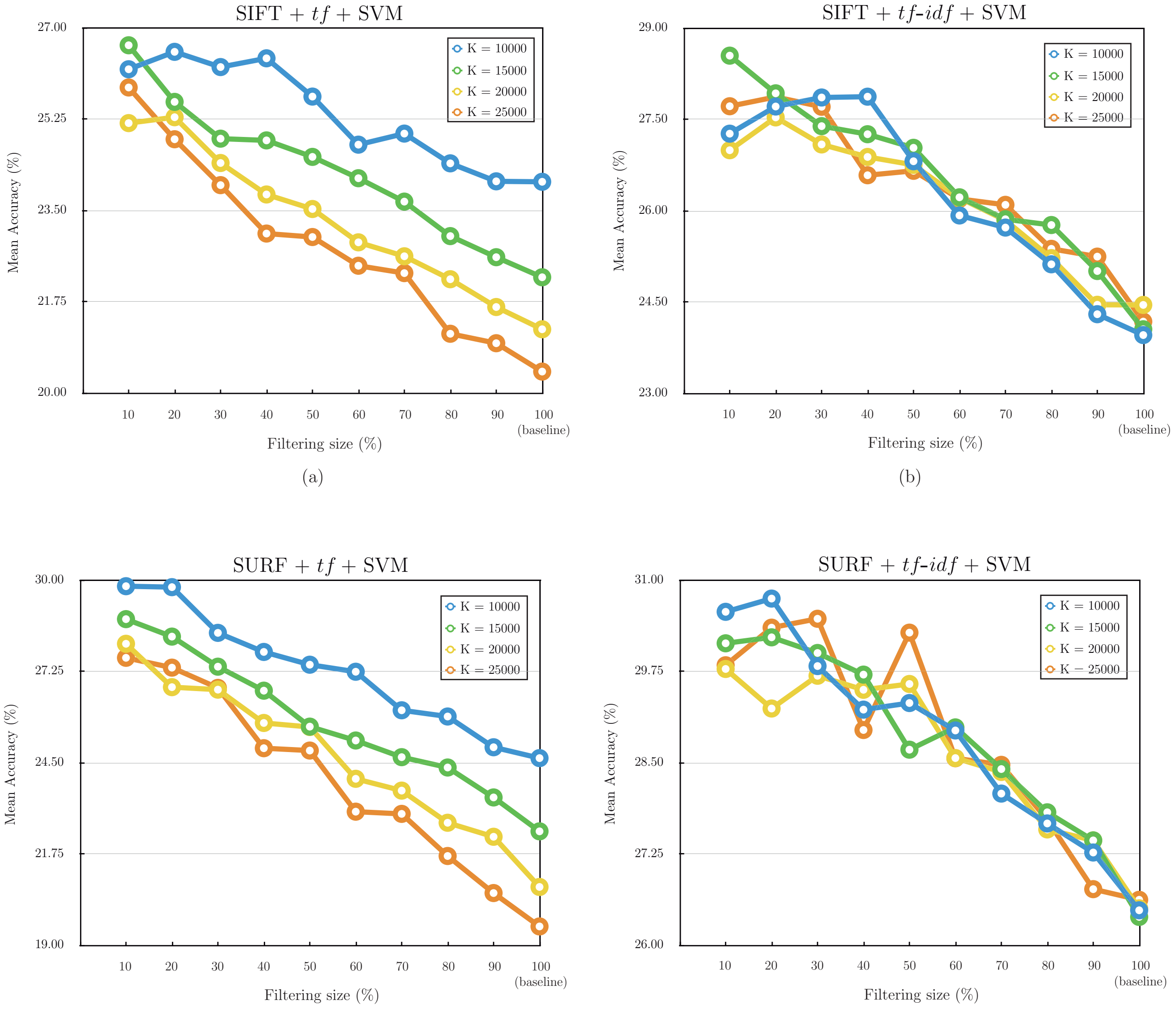
Also, the conducted experiments provide evidence that a large number of visual words in a vocabulary are noisy or little discriminative. Discarding these visual words allows for a better and more compact image representations.

4.2 Comparison with other kinds of feature selection algorithms

The aim of the second experiment is to compare our proposal with respect to other kind of feature selection algorithm. With this purpose, we compare the accuracy of our vocabulary filtering methodology with respect to the accu-

Table 1: Computation time (in seconds) of visual vocabulary ranking compared to vocabulary building.

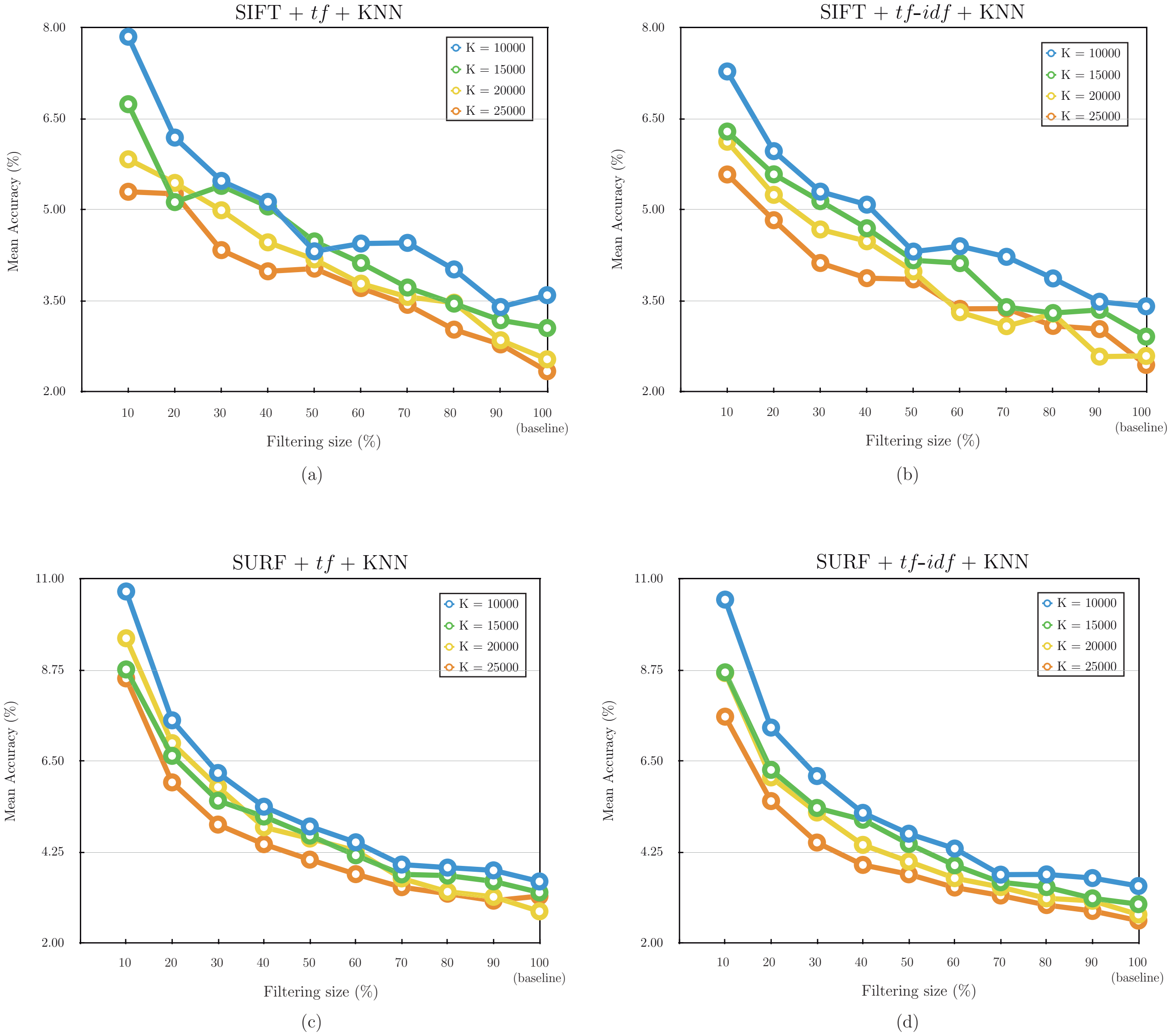
Dataset	K	Vocabulary building (K-means) computation time (s)	Vocabulary ranking (proposed method) computation time (s)	Vocabulary ranking (MI-based method [31]) computation time (s)
Caltech-101 (188 248 training features)	10000	4723.452	**8.111**	5.754
	15000	6711.089	**18.622**	16.825
	20000	7237.885	**33.890**	27.478
	25000	9024.024	**54.338**	49.963
Pascal VOC 2006 (114 697 training features)	10000	4126.487	**7.985**	5.285
	15000	5922.134	**18.320**	15.879
	20000	6441.980	**30.259**	28.458
	25000	8563.692	**51.743**	49.132



(a) (b)

(c) (d)

Figure 5: Mean classification accuracy results for SVM cross-validation on the Caltech-101 dataset. As can be seen, using the reduced vocabularies always resulted in better classification accuracies.



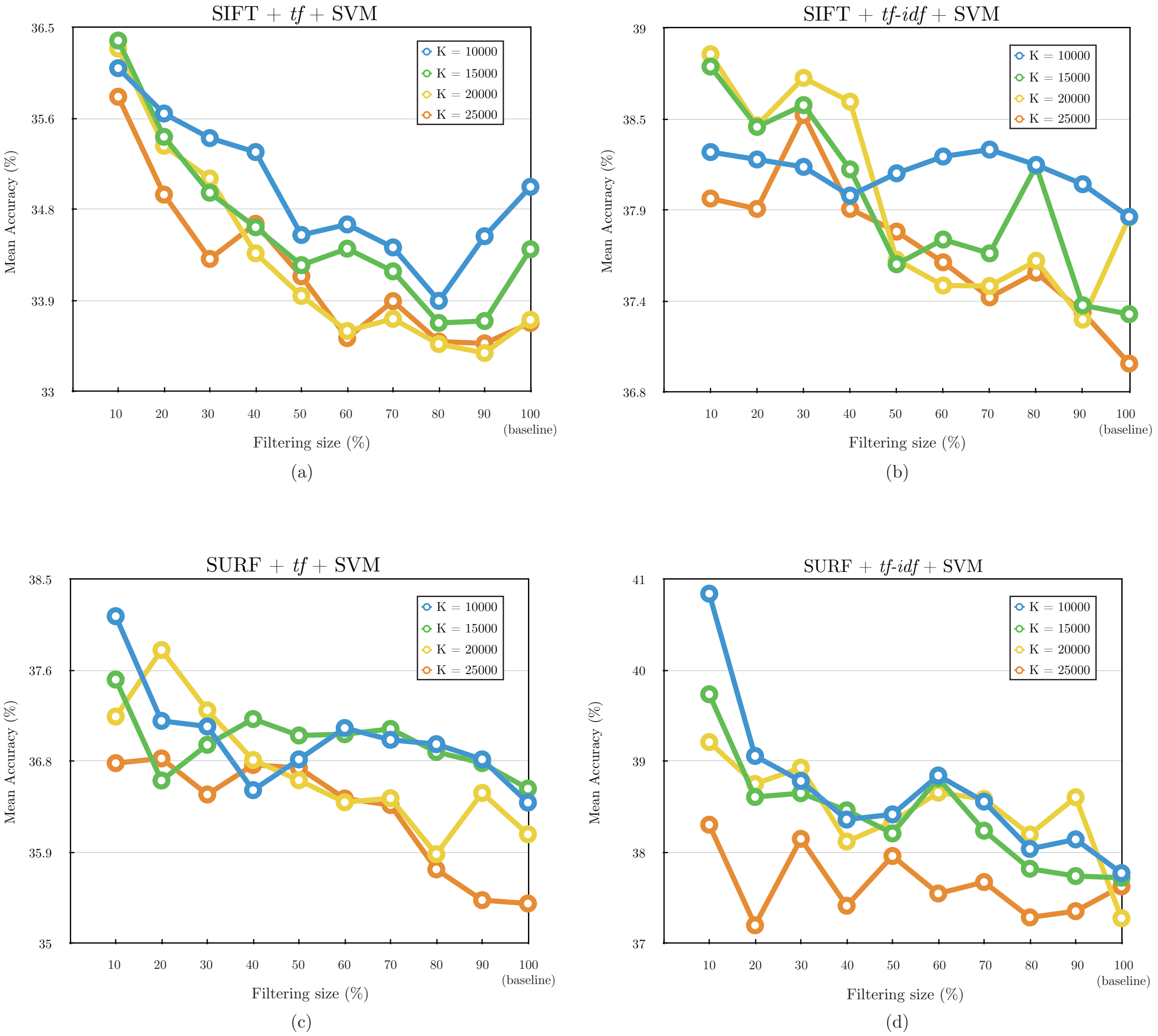
(a)

(b)

(c)

(d)

Figure 6: Mean classification accuracy results for KNN cross-validation on the Caltech-101 dataset. As can be seen, using the reduced vocabularies always resulted in better classification accuracies.



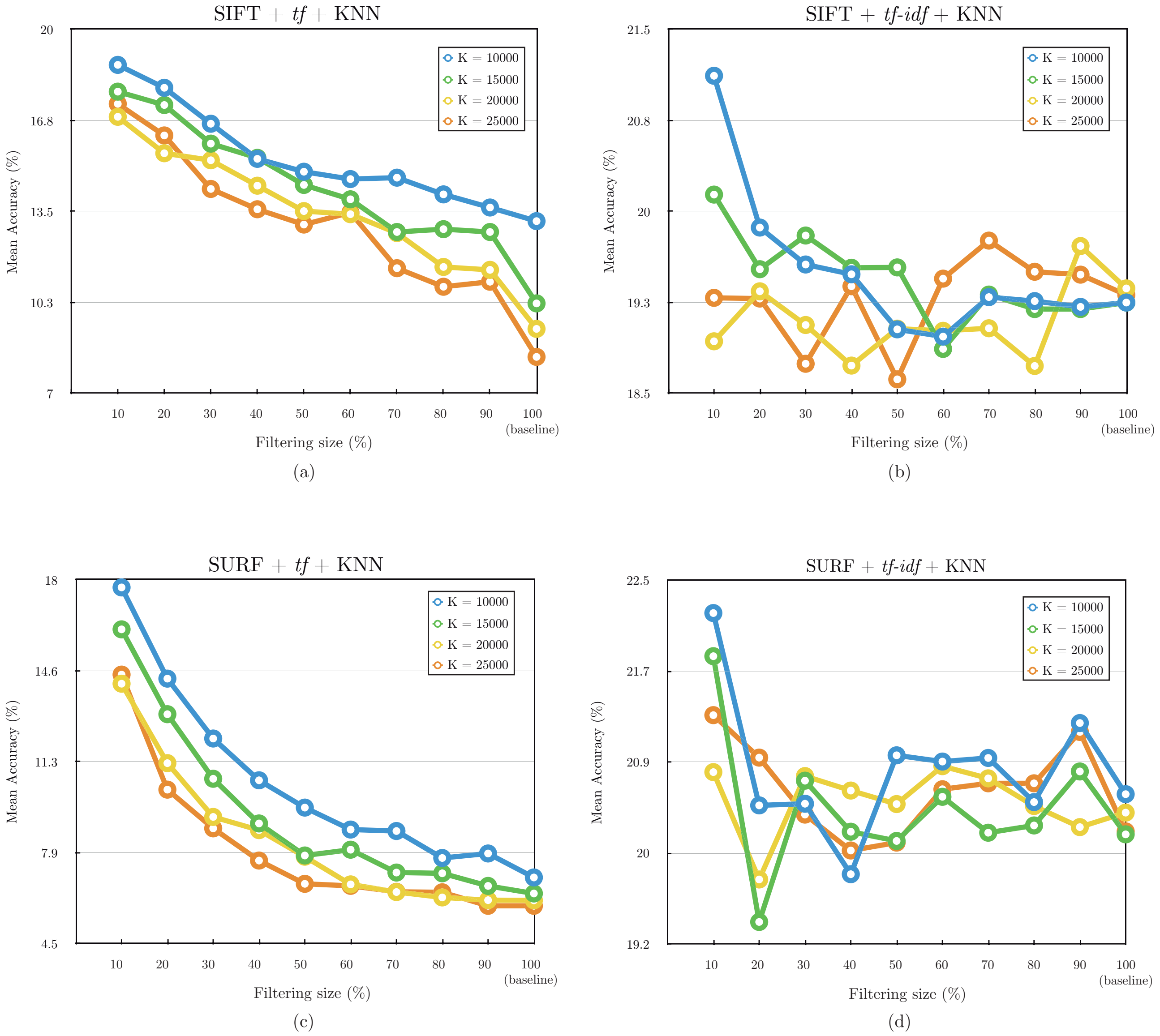
(a)

(b)

(c)

(d)

Figure 7: Mean classification accuracy results for SVM cross-validation on the Pascal VOC 2006 dataset. As can be seen, using the reduced vocabularies always resulted in better classification accuracies.



(a)

(b)

(c)

(d)

Figure 8: Mean classification accuracy results for KNN cross-validation on the Pascal VOC 2006 dataset. As can be seen, using the reduced vocabularies always resulted in better classification accuracies.

Table 2: Summarized results for the Caltech-101 dataset.

Descriptor	**Weighting schema**	**K**	**SVM**			**KNN**		
			Baseline	**Best result**	**Best filter size (%)**	**Baseline**	**Best result**	**Best filter size (%)**
SIFT	*tf*	10000	24.05	**26.54**	20	3.60	**7.86**	10
		15000	22.22	**26.66**	10	3.06	**6.74**	10
		20000	21.22	**25.28**	20	2.54	**5.84**	10
		25000	20.41	**25.85**	10	2.34	**5.30**	10
	tf-idf	10000	23.96	**27.87**	40	3.41	**7.28**	10
		15000	24.05	**28.55**	10	2.91	**6.29**	10
		20000	24.45	**27.53**	20	2.60	**6.13**	10
		25000	24.18	**27.87**	20	2.45	**5.59**	10
SURF	*tf*	10000	24.63	**29.81**	10	3.53	**10.70**	10
		15000	22.43	**28.82**	10	3.26	**8.77**	10
		20000	20.75	**28.08**	10	2.80	**9.54**	10
		25000	19.56	**27.66**	10	3.17	**8.55**	10
	tf-idf	10000	26.48	**30.74**	20	3.42	**10.50**	10
		15000	26.39	**30.21**	20	2.97	**8.70**	10
		20000	26.50	**29.78**	10	2.72	**8.69**	10
		25000	26.62	**30.47**	30	2.57	**7.61**	10
Average			23.62	**28.23**	16.8	2.96	**7.76**	10

Table 3: Summarized results for the Pascal VOC 2006 dataset.

Descriptor	**Weighting schema**	**K**	**SVM**			**KNN**		
			Baseline	**Best result**	**Best filter size (%)**	**Baseline**	**Best result**	**Best filter size (%)**
SIFT	*tf*	10000	34.97	**36.11**	10	13.16	**18.74**	10
		15000	34.37	**36.38**	10	10.22	**17.79**	10
		20000	33.69	**36.30**	10	9.31	**16.89**	10
		25000	33.66	**35.84**	10	8.31	**17.35**	10
	tf-idf	10000	37.86	**38.27**	10	19.25	**21.12**	10
		15000	37.27	**38.77**	10	19.25	**20.14**	10
		20000	37.86	**38.84**	10	19.37	**19.72**	90
		25000	36.97	**38.48**	30	19.31	**19.76**	70
SURF	*tf*	10000	36.36	**38.15**	10	6.96	**17.72**	10
		15000	36.49	**37.54**	10	6.37	**16.17**	10
		20000	36.05	**37.82**	20	6.11	**14.15**	10
		25000	35.39	**36.78**	20	5.91	**14.49**	10
	tf-idf	10000	37.77	**40.85**	10	20.56	**22.20**	10
		15000	37.72	**39.74**	10	20.19	**21.81**	10
		20000	37.28	**39.21**	10	20.39	**20.81**	60
		25000	37.63	**38.31**	10	20.22	**21.28**	10
Average			36.33	**37.96**	12.5	14.06	**18.76**	21.88

racy of the MI-based method proposed in [31], in a classification task; the experiment was done over the Caltech-101 dataset. As it was mentioned in Section 2, the MI-based method proposed in [31] obtains the best results among the feature selection and compression methods of image representation for object categorization.

In the experiments presented here, we use for image representation a BoW-based schema with the following specifications:

- PHOW features (dense multi-scale SIFT descriptors) [3].
- Spatial histograms as image descriptors.
- Elkan K-means [6], with five different K values (K= 256, 512, 1024, 2048 and 4096), is used to build the visual vocabularies; these vocabularies constitute the baseline.
- Each of the baseline vocabularies is ranked using the MI-based method proposed in [31] and our proposed visual vocabulary ranking methodology.
- Later, nine new vocabularies are obtained by filtering each baseline vocabulary, leaving the 10%, 20%, ..., 90%, respectively.
- We randomly selected 15 images from each of the 102 categories of Caltech-101 dataset, in order to build the visual vocabularies. For each category, 15 images were randomly selected as test images.

We tested the obtained visual vocabularies in a classification task, using a homogeneous kernel map to transform a χ^2 Support Vector Machine (SVM) into a linear one [26]. The classification accuracy is reported in Figure 9.

As it can be seen in Figure 9, for each value of K used in the experiment, our proposal obtains the best classification accuracy results for the highest compression rates. Besides, for the other filtering sizes our proposal and the MI-based method attains comparable results.

4.3 Computation time of the visual vocabulary ranking

The computation time of the visual vocabulary ranking methodology has also been evaluated. Table 1 shows the time in seconds taken for the ranking method in different size vocabularies, for the Caltech-101 and the Pascal VOC 2006 dataset. In Table 1, the ranking time is compared with the time needed to build the visual vocabulary.

As can be seen in Table 1, the proposed methodology can be used to improve visual vocabularies without requiring much extra computation time.

5 Conclusion and future work

In this work we devised a methodology for reducing the size of visual vocabularies that allows to obtain more discriminative and representative visual vocabularies for BoW image representation. The vocabulary reduction is based on three properties and their corresponding quantitative measures that express the inter-class representativeness, the intra-class representativeness and inter-class distinctiveness of visual words. The experimental results presented in this paper showed that, in average, with only 25% of the ranked vocabulary, statistically superior classification results can be obtained, compared to the classical BoW representation using the entire vocabularies. Therefore, the proposed method, in addition to providing accuracy improvements, provides a substantial efficiency improvement. Also, compared with a mutual information based method our proposal obtained superior results for the highest compression rates and comparable results for the other filtering sizes.

As future work, we aim to propose a weighting schema that takes advantage of the proposed measures, in order to improve image representation. Also we would like to explore the use of hierarchical classifiers for dealing with inter-class variability. Finally, we also aim to define a measure that help us to automatically choose the filter size.

References

[1] Concepts and applications of inferential statistics, 2013.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

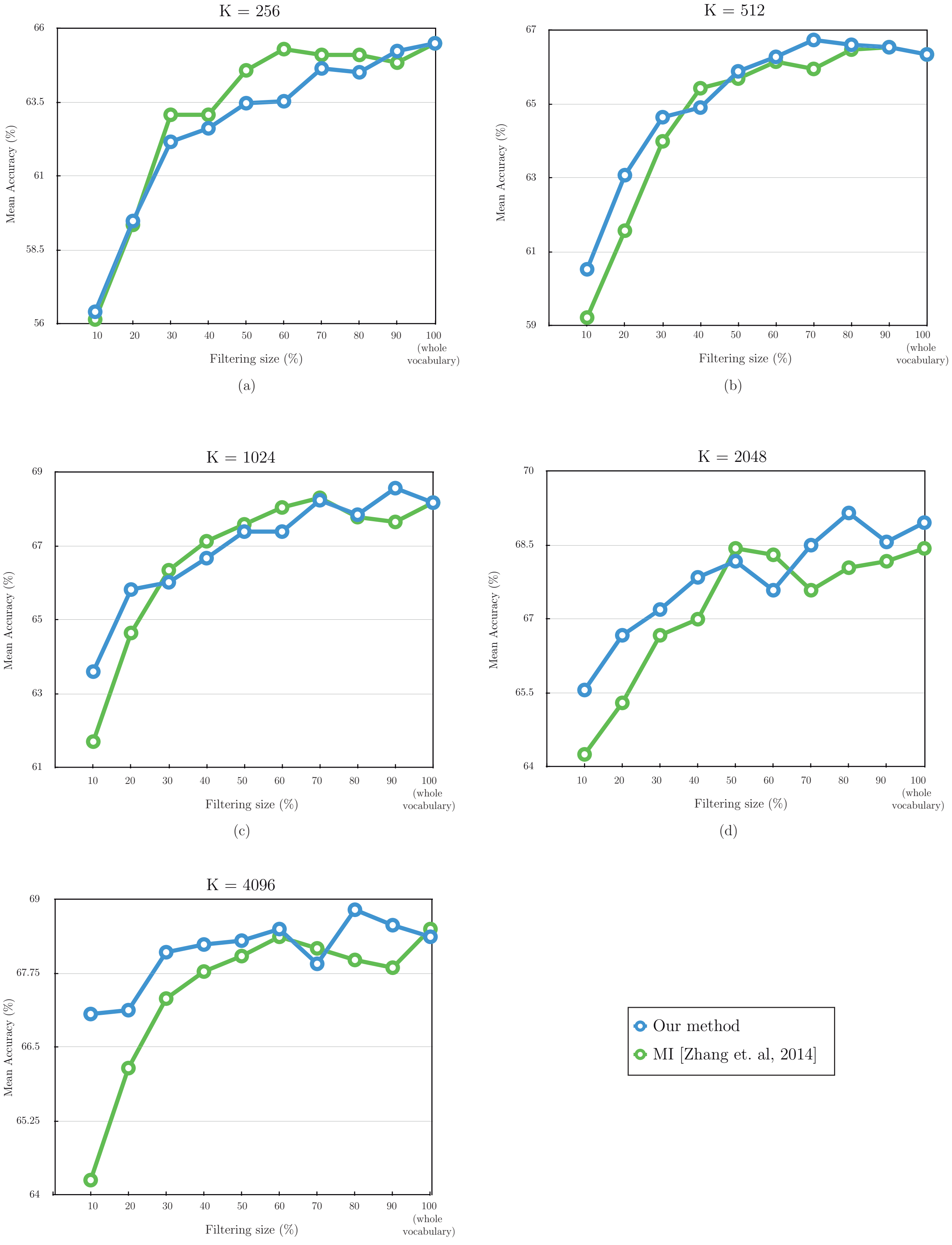
[3] A Bosch, Andrew Zisserman, and X Munoz. Image classification using random forests and ferns. *IEEE 11th International Conference on Computer Vision (2007)*, 23(1):1–8, 2007.

[4] Siddhartha Chandra, Shailesh Kumar, and C. V. Jawahar. Learning hierarchical bag of words using naive bayes clustering. In *Asian Conference on Computer Vision*, pages 382–395, 2012.

[5] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[6] Charles Elkan. Using the triangle inequality to accelerate k-means. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 147–153. AAAI Press, 2003.

[7] Peter Emerson. The original borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358, 2013.



(a)

(b)

(c)

(d)

(e)

Figure 9: Comparison of mean classification accuracy results, on the Caltech-101 dataset, between the proposed methodology and the MI-based method proposed in [31].

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/ workshop/index.html.

[9] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/ results.pdf.

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, April 2007.

[11] Basura Fernando, Élisa Fromont, Damien Muselet, and Marc Sebban. Supervised learning of gaussian mixture models for visual vocabulary generation. *Pattern Recognition*, 45(2):897–907, 2012.

[12] Peter V. Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *ICCV*, pages 221–228. IEEE, 2009.

[13] Y. Gong, S. Kumar, H. A. Rowley, and S. Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *CVPR 2013*, 2013.

[14] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intellingence*, 33(1):117?128, 2011.

[15] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, 2011.

[16] Mingyuan Jiu, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Supervised learning and codebook optimization for bag of words models. *Cognitive Computation*, 4:409–419, December 2012.

[17] Kraisak Kesorn and Stefan Poslad. An enhanced bag-of-visual word vector space model to represent visual content in athletics images. *IEEE Transactions on Multimedia*, 14(1):211–222, 2012.

[18] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Volume 2 CVPR06*, 2(2169-2178):2169–2178, 2006.

[19] Gang Liu. Improved bags-of-words algorithm for scene recognition. *Journal of Computational Information Systems*, 6(14):4933 – 4940, 2010.

[20] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[21] R.J. Lopez-Sastre, T. Tuytelaars, F.J. Acevedo-Rodriguez, and S. Maldonado-Bascon. Towards a more discriminative and semantic visual vocabulary. *Computer Vision and Image Understanding*, 115(3):415 – 425, 2011. Special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[23] Zhiwu Lu and Horace Ho-Shing Ip. Image categorization with spatial mismatch kernels. In *CVPR*, pages 397–404. IEEE, 2009.

[24] Jianzhao Qin and Nelson Hon Ching Yung. Feature fusion within local region using localized maximum-margin learning for scene categorization. *Pattern Recognition*, 45(4):1671–1683, 2012.

[25] Chih-Fong Tsai. Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, 2012, 2012.

[26] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intellingence*, 34(3), 2011.

[27] Lei Wang, Lingqiao Liu, and Luping Zhou. A graph-embedding approach to hierarchical visual word mergence. *IEEE Trans. Neural Netw. Learning Syst.*, 28(2):308–320, 2017.

[28] Jianxin Wu, Wei-Chian Tan, and James M. Rehg. Efficient and effective visual codebook generation using additive kernels. *Journal of Machine Learning Research*, 12:3097–3118, 2011.

[29] Shiliang Zhang, Qi Tian, Gang Hua, Qingming Huang, and Wen Gao. Generating descriptive visual words and visual phrases for large-scale image applications. *IEEE Transactions on Image Processing*, 20(9):2664–2677, 2011.

[30] Shiliang Zhang, Qi Tian, Gang Hua, Wengang Zhou, Qingming Huang, Houqiang Li, and Wen Gao. Modeling spatial and semantic cues for large-scale near-duplicated image retrieval. *Computer Vision and Image Understanding*, 115(3):403–414, 2011.

[31] Y. Zhang, J. Wu, and J. Cai. Compact representation for image classification: To choose or to compress? In *CVPR 2014*, 2014.