# Improving automatic image annotation: Approach by Bag-Of-Key Point

**Merad Boudia Mohammed Abdelaziz [1], Zaoui Lynda [2]**

**[1] University of Sciences and technology Mohamed Boudiaf, Oran , Algeria**

**[2] University of Sciences and technology Mohamed Boudiaf, Oran , Algeria**

**Abstract**

Automatic image annotation is to associate each image a set of keywords and describing the visual content of the image using an automatic system without any human intervention, many approaches have been proposed for the realization of such a system However, it is still inefficient in terms of semantic description of the image. Recent works show a frequent use of a special technique known as bag-of-key points that describes an image as a set of local descriptors using a histogram. Each bin of the histogram represents the importance of a visual pattern (called visual word) in the image. But crucial representation choices - such as the choice of local features, the steps of building the visual vocabulary - have not been thoroughly studied in existing works. In this paper, a novel approach based on Scale Invariant Features Transform (SIFT) features and treatment of the different steps of building de vocabulary are proposed. The proposed approach creates more robust signatures for images and better reflects the weight of visual words. The categorization of images has been the subject of the second phase of this approach. The purpose of this phase was to find a classification model that best suits the index method proposed, while avoiding problems due to large data and large dimension. Experiments with Corel-1000 dataset demonstrate that the proposed improvements outperform known techniques in scene categorization.

***Keywords***: interest region, bag-of-key points, visual vocabulary, image annotation**.**

## 1. Introduction and problems

The automatic image annotation is to transform the visual content of images in semantic information, for this the system must be able to analyze the visual characteristics of objects, for identify, describe and differentiate them. However, these objects do not always have the same visual characteristics; these characteristics vary depending on several factors: the considered instance, the conditions of shooting the image, and context of occurrence of the object in the image complicating the annotation that must cross the semantic gap problem that results.

To overcome these problems the researchers are trying to find regions in the images that contain visual information robust to visual variations, among the techniques used to detect these regions, there are extraction and description the interest regions. The bag of word approach is one of the most currently used technique, but crucial choices like the choice of detector and descriptor, the steps building vocabulary and the classification model of best suited to the method have not been thoroughly studied in the literature.

we have notes in the literature a multitude of implementation choices and several factors governing the effectiveness of each step, such as the choice of the local descriptors, vocabulary creation method, the size of the vocabulary, the calculation method of the visual words weights, the similarity measure between signatures of images (for the search for images), the classification model (for categorization)… etc. This research task explores all these factors by raising two principal problems: 1) the robustness of the visual vocabulary and 2)classification methods .

**Problem of vocabulary construction:**

The construction of the visual vocabulary is one of a sensitive step. Indeed, we must determine the nature of the weight that is associated with visual words, such as the presence or absence of words, their frequencies with the hypothesis that more the weight of visual word is higher, better describes it the image.

The problem is that each object of the base will be expressed in following by using the visual words of this

vocabulary. The result is that generate a most representative possible objects vocabulary allows to represent better these objects.

**Problem of classification**

Choosing the right classify significantly influences the learning time and the efficiency of recognition. The challenge is to find the best classification model adapts to the indexing approach by allowing establishing a good compromise between reduced learning time and high rate of correct classification.
Our objective is to study the approach "Bag of visual words" for indexing images and to apply it to research by the contents and classification. This study will make it possible to propose improvements compared to the known implementations, then to validate the implementation suggested. The objective is to optimize the performances of the search for images and classification, while maintaining the simplicity of the approach.

## 2. State of the art

Early work on automatic annotation of images appeared around 1999, and since this area has attracted much interest in the community of image processing. There are two families of approaches to address the problem of automatic annotation. The first approach considers the image as a whole, without trying to segment it, and tries to recognize the subject of the image by using the properties of color and texture of the complete image. The disadvantage of this approach is that it must overcome the problem of semantic gap between the low-level descriptors and semantics of the image, or the utility of using an intermediate semantic representation.

One of the first works is described by [10] who splits the image into a grid of rectangular regions and applies a model of co-occurrence between keywords and visual signatures. [9] For their part seek explicitly of objects in images to infer the semantic category of the scene. After an initial step of segmenting the image according to criteria of textures and colors, the local characteristics of texture, color and shape of the regions are extracted, and the global features such as the number of regions, the complexity of the image or its symmetry. These characteristics are used to recognize some regions such as skin, sky, water, mountains, artificial object. These objects are then used to assign a semantic category at the scene itself.

However these approaches depend strongly of segmentation techniques and the results of partitioning techniques used. This makes them sensitive to the problem of scalability. On the other hand, use a semantic level intermediate often involves a step of object recognition, and therefore uses analysis techniques local of image, particularly by regions or by interest points.
For this [5], calculates the bags of words from the local descriptor SIFT [8] and use them for learning of categories with PLSA algorithm probabilistic latent semantic analysis. The local descriptors are calculated from a regular grid. They classify 13 categories with an average rate of 73.4% of good recognitions. [12] Proposes a similar approach, involving only three categories, where SIFT descriptors were calculated around the points obtained from a detector of interest point. [6] Also uses bags of words with the SIFT descriptor and perform learning by proposing a variant of LDA (Latent Dirichlet Allocation) [1] to classify scenes into 13 categories.

## 3. Model of bag of visual words

The main idea is to represent images by collections of visual words called visual vocabulary, and to obtain a total signature from it cash the occurrences of these words. Then an algorithm of classification is applied to the signatures to build a classifier who will allow thereafter annotating the new images.

The principal stages of our method are:

- Detection and description of the interest points (IP) by detector SIFT.
- Vocabulary of visual words Construction.
- Construction of the signature associated with each image: Bag-Of-Features (BOF).
- To apply an algorithm of classification to the BOF, and thus to determine with which category to assign the image.

Our goal of is thus used a good vocabulary thus making it possible to build a good classifier while reducing the data-processing effort to the minimum, and to thereafter be able to annotate a new image while determining has which category/categories to assign the image.
We discuss now the choices made for each steps more in detail.

### 3.1 Descriptors Extraction

Among the detectors of interest points the most efficient we can find the Moravec detector, the Harris detector, and SIFT (Scale Invariant Feature Transform).For these detectors, the interest points are revealed more reliable than the outline approach because they provide more constraints on the intensity function, they are present in the vast majority of image and are robust to occlusions.

We chose the SIFT descriptor and this for the following reasons:

1 - It uses simple linear Gaussian derivatives. Therefore we expect it to be more stable to image disturbances such as noise.

2 - He obtained a greater vector descriptor (128 rather than 12 to 16). Therefore we have a representation richer and potentially more distinctive. Recently [2] compared several descriptors and found that the SIFT descriptor was the most efficient.

3 - This descriptor can find descriptions that are invariant with effect of the point of view, the scale, the rotation and lighting condition, allowing having a robust visual word to visual variations.

3.2 Vocabulary building

In practice, build the visual vocabulary (or visual dictionary) returns to quantify the space of all local features of images. Since this space is not dense and uniform seen that some patterns can never appear in images while others recur frequently, the use of effective methods of clustering is necessary. The number of clusters selected is actually the size of the dictionary. Clusters are the visual words vocabulary (Figure 1).

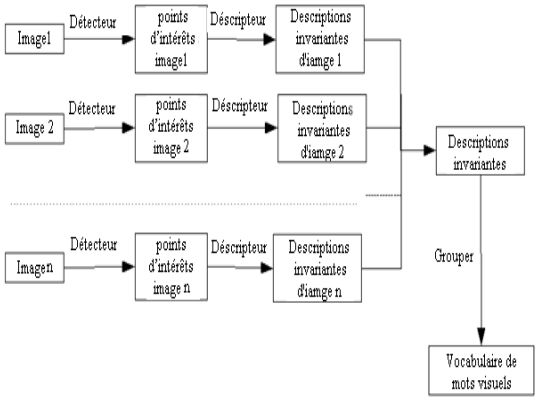


Fig. 1. Schema of creation of vocabulary.

The construction of the visual vocabulary is one of the sensitive steps in our model. Indeed, every object in the training set or test database will be expressed in the following by means of visual words of this vocabulary. It follows that generate a vocabulary as representative as possible of the objects allows to best represent these objects.

## 3.3 Quality of vocabulary

How to judge the quality of a visual vocabulary? Performances for automatic annotation are obviously the final criteria. However, they are factors that may affect the relevance of vocabulary among them:

**The partitioning algorithm**, the most used algorithm that is simple and rather efficient is k-means algorithm; however, it is not guaranteed that the solution provided is optimal. It depends strongly on the initialization phase. To overcome this problem [7] proposed another algorithm called Quality Threshold (QT) which is initially used for the analysis of data on gene expression. The main idea is to fill the space with partitions having a fixed radius RQT, it starts with the partition with the largest number of signatures. All signatures belonging to this partition are removed and iterates until the base is empty. The number of words is determined by the algorithm and depends on the radius chosen. The main disadvantage is its computational cost quadratic. So to avoid the defects of k-means and QT we adapted the algorithm proposed QT (QT-Plus), but instead of imposing a fixed radius for partitions, we impose a fixed number of signatures by partition , which we denote λ. At each iteration, the partition that is retained is the one with the smallest radius.

The creation of a vocabulary is how to account for the density of patches in the visual space. So we introduced the QT-Plus to avoid overrepresentation of visual space dense areas.

To compare performance, a comparison between the random, K-means and QT-Plus is shown in Figure 2, the average precision or MAP (Mean Average Precision) is calculated on a vocabulary of up to 1,000 visual words , showing the advantage provided by the QT-Plus.

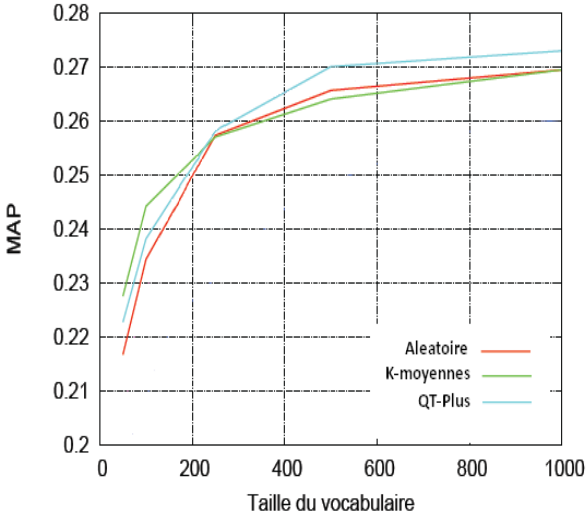


Fig. 2. Performance according to the partitioning algorithms (corel-1000).

**Strategy to create the vocabulary**, indeed, the combination of several detectors or more vocabularies can influence the outcome of vocabulary.

**Influence of the number of patches by image**, [11] shows that one of the more important criteria is the number of patches extracted from each image.

### 3.4 Representation by F-list

**Motivation**

It was observed that the detector points are usually attracted in some areas of the image. So we think that creation algorithm must take into account vocabulary the best possible data distribution. It is hollowing that the zones of visual space that are important should be represented by a high frequency vocabulary in order to maintain the potential of description they provide.

The proposed model created for each category a subset that representing the vocabulary for this category.

**Description**

Suppose that the vocabulary $V = \{v_1, v_2, \ldots, v_i, \ldots, v_k\}$ is formed by the centers of clusters obtained with the QT-Plus, and $U_{ij}$ $j \in \{1,2,\ldots\ldots,k\}$ the frequency of occurrence of each word in the vocabulary Vi following the local descriptors calculated by the SIFT descriptor for an image, and F-list is the sum of $U_{ij}$ for all images after each category.

Once the F-lists of each category calculated and ordered by uncrossing order, we obtain a reduced table will be called F-Tab-list or each line in this table will help to represent a class of image-based, thus passing from a global SIFT to a reduced table Tab-f- list.

Noted that each visual word according to each line of Tab-F-list that will attributed an important factor, which will be used later in the classification.

### 3.5The classification

Once the visual vocabulary built, and the signature of each image according to the vocabulary calculated, we reduce the problem of visual categorization than the supervised classification with many classes as defined visual categories. That is to learn from indexed images with a known class, a model that predicts the membership of a new image to one of the classes known a priori. In the learning phase, two major problems are encountered: the complex classification models generally guarantee good recognition, but often require significant learning time with large databases. Moreover, the approach "bag of visual words" generates high-dimensional data, which further increases the learning time and complicates the search for correlation between data. The Choice of the good classifier influences greatly on the learning time and efficiency of recognition. The challenge is to find the best classification model suited to the indexing approach, allowing establishing a good compromise between reduced learning time and high rate of correct classification.

**Classification of F-list:**

Our goal in using the Tab-F-list is to classify images according to classes; we seek to bring out some visual words that best represent these classes.

The method we propose here is a very simple method. To classify an image just visual vocabulary are calculated and ordered in descending order who be called **t-F-list**.

Once the **t-F-list** calculated, a set of visual words will be selected (exp: the first 3 words) Fig3. These words are compared to each line of **Tab-F-list**, thereby calculating the factor of importance of each word following each line. Thereafter the image will be assigned a category in the important factor is the greatest.

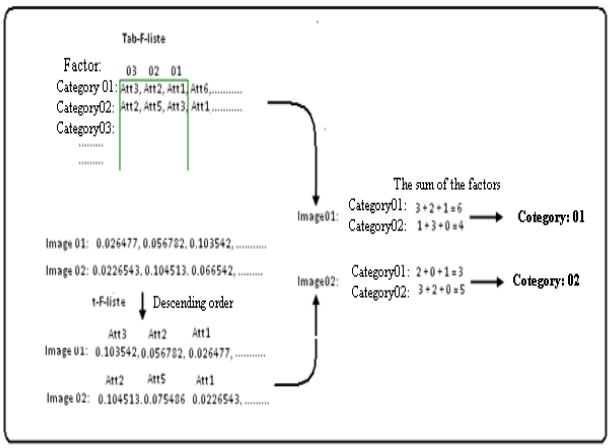


Fig. 3. Classification by F-list.

## 4. Experiments and result

We explored the performance of the proposed method based on image Corel-1000. Corel is a collection of about 60,000 images created by the group of Professor Wang at the University of Pennsylvania. Corel-1000 is a sub-collection that contains 1000 natural images divided into ten categories with 100 images per category.

To test our approach we divided the image into two basic groups:

- annotated sets of images
- unannotated sets of images, on what approaches can be tested.

The evaluation measure is average precision of the annotation object class, and the average over all classes "MAP - Mean Average Precision".

We randomly selected 700 images for extracting SIFT key points. Then, we used clustering algorithm proposed QT-Plus for grouping local features extracted from a visual

vocabulary. For our experiments, we fixed the vocabulary size to 100 visual words.

Corel-1000 is known to be a collection of images and very difficult to classify it because the high number of classes and the high variability of background images even for images belonging to the same class. However, experiments have shown that when the proposed method is applied one manages to have satisfactory classification rate.

We present in Figure 4 the results obtained by the F-list approach described in 3.5. The average accuracy over all classes of objects is 0.608. We note that the results are superior to standard approach, and this is due to the inclusion of the order uncrossed visual information to create a model for each category in which just the most representative visual words it's used.

We note that the accuracy of annotation depends on the class of objects. Two classes are better detected than others: "dinosaur" 91% and "flowers"93%, while the lowest rate is 42% and 41% respectively obtained for Mountain and Building. These rates can be explained by the fact that these two categories may be included in other categories. For example, for example, 17% of *building* scenes were confused with the category *Bus* because many images from the latter contain also buildings.

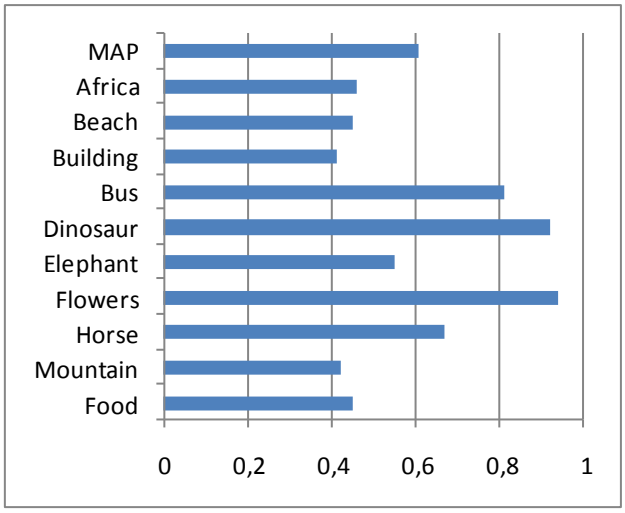


Fig. 4. Average Precision of annotation in corel 1000 (approach of F-list).

## 5. Conclusions and perspective

In this paper we have presented a state of the art on the main research works have addressed the problem of semantic gap, trying to overcome the visual variations greatly complicate the analysis of images visual features and, consequently, the annotation based on this analysis.

In this work, we tried to remedy at two critical problems: the visual variations and accuracy of annotation.

The first study we demonstrated that the constructed vocabulary does not guarantee effective representation of images. To overcome these limitations, we tried to overcome the various factors which diminish its relevance.

The categorization of images has been the subject of the second phase of this research. The purpose of this phase was to find the classification model most adapts to the proposed method, while avoiding the problems caused by voluminous data and large dimensions. Indeed, experimental studies have confirmed a priori hypotheses by demonstrating the efficiency and simplicity of the proposed method, as well as to search for image than for classification.

The extensions and improvements that can be added to the studied approach are many and involved in different steps of this approach. The first perspective is to merge with other descriptors SIFT to better characterize images. In fact, add more information such as color and texture is an interesting view of the lack of this information in the SIFT descriptor.

The second direction would be to add information on the spatial distribution of visual words in the image.

## References

[1] D. Blei, A. Ng, M. Jordan. "Latent dirichlet allocation", Journal of Machine Learning Research, 3, pp. 993–1022, (2003).

[2] K. Mikolajczyk and C. Schmid, A performance evaluation of local descriptors,CVPR(2003).

[3] Ni D., Qu Y., Yang X., Chui Y.-P.,Wong T.-T., Ho S. S. M., Heng P.-A., « Volumetric Ultrasound Panorama Based on 3D SIFT », *MICCAI (2)*, p. 52-60, (2008).

[4] Zhou H., Yuan Y., Shi C., « Object tracking using SIFT features and mean shift », *Computer Vision and Image Understanding*, vol. 113, n° 3, p. 345-352, March, (2009).

[5] A. Bosch, A. Zisserman, X. Mu˜noz. "Scene classification via plsa". In European Conference on Computer Vision, (2006).

[6] L.Fei-Fei, P. Perona. "A bayesian hierarchical model for learning natural scene categories".In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 524–531, Washington, DC, USA, (2005) .

[7] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In IEEE International Conference on Computer Vision, (2005).

[8] D. Lowe. "Distinctive image features from scale-invariant keypoints", (2003).

[9] A. Mojsilovic, J. Gomes, B. Rogowitz. "Isee : perceptual features for image library navigation". In Proceedings SPIE Human Vision and Electronic Imaging, Volume 4662, pages 266–277, San Jose, Californie, (2002).

[10] Y. Mori, H.Takahashi, and R.Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In International Workshop on Multimedia Intelligent Storage and Retrieval Management, (1999).

[11] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In European Conference on Computer Vision, (2006).

[12] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, L. V. Gool. "Modeling scenes with local descriptors and latent aspects". In International Conference on Computer Vision, pages 883–890, Pékin, Chine, (2005).