Accurate Image Search using Local Descriptors into a Compact Image Representation

Soumia Benkrama[1], Lynda Zaoui[2] and Christophe Charrier[3]

**[1] University of Science and Techenology Mohamed Boudiaf, Department of Computer Science,
Laboratory Systems Signals Data
Oran, Algeria**

**[2] University of Science and Techenology Mohamed Boudiaf, Department of Computer Science,
Laboratory Systems Signals Data
Oran, Algeria**

**[3] University of Caen-Basse Normandie, GREYC, UMR CNRS 6072
Caen, France**

Abstract

Progress in image retrieval by using low-level features, such as colors, textures and shapes, the performance is still unsatisfied as there are existing gaps between low-level features and high-level semantic concepts.

In this work, we present an improved implementation for the bag of visual words approach. We propose a image retrieval system based on bag-of-features (BoF) model by using scale invariant feature transform (SIFT) and speeded up robust features (SURF). In literature SIFT and SURF give of good results. Based on this observation, we decide to use a bag-of-features approach over quaternion zernike moments (QZM). We compare the results of SIFT and SURF with those of QZM.

We propose an indexing method for content based search task that aims to retrieve collection of images and returns a ranked list of objects in response to a query image. Experimental results with the Coil-100 and corel-1000 image database, demonstrate that QZM produces a better performance than known representations (SIFT and SURF).

Keywords: *Content-Based Image Retrieval Systems, feature detection, Bag of visual words.*

1. Introduction

Content-based image retrieval (CBIR) is a long standing challenging problem in computer vision and multimedia, seek to represent the content of images automatically, using visual descriptors representing the multimedia data.

CBIR system views the query image and the images in the database as a collection of features, and ranks the relevance between the query and any matching image in proportion to a similarity measure calculated from the features. These features, or signatures of images, characterize the content of images; the key idea in the image search is an approximate search by using concept of proximity, similarity, and distance between objects [1]. The similarity measure is often based on the calculation of a distance in the feature space, one then seek the nearest neighbors of the query [1].

Recent works on Content Based Image Retrieval rely on Bag of Visual Words (BoVW) to index images. In BoVW, local features are extracted from the whole image dataset and quantized (termed as visual words). For compact representation, a visual vocabulary is usually constructed to describe BoF through the clustering of keypoint features. Each keypoint cluster is treated as a “visual word” in the visual vocabulary. This approach employs histogram based features for image representation.

In this paper, we address the problem of searching the most similar images in image database. We put an emphasis on the joint optimization of three constraints: storage, computational cost, and recognition performance.

2. State of the art

Many different approaches for CBIR have been proposed in the literature. Swain and Ballard [2] were the first to use color histograms features to describe images. Since, many other introduced other features like texture or colorimetric moments. These descriptors allow a quite efficient retrieval in many cases, but fail in precision, because global features lose most of local information expressed in the image. Recent approaches propose to use local features to describe interest regions in the image.

The idea is to detect interesting local patches, represent the patches as numerical vectors and consider image which allows comparing images by measuring the similarity between signatures.

Many approaches have been proposed to extract local features from images. In [4] and [5] the authors extract local patches using a regular grid. Other authors use also random sampling [6], [7] and segmentation methods. A more interesting approach is to extract keypoints.

The most popular approach today, initially proposed in [3], relies on a bag-of-features (BOF) representation of the image. The idea is to quantize local invariant descriptors.

Bag of features

This model allows describing an image as bag of elementary local features called visual words. As a result, an image is represented by a vector of weights, where weight corresponds to the importance of visual word in the image. The choice of local features and the weighting schema are very important to perform image retrieval.
Recent approaches that address the problem of indexing techniques and image search, trying to find areas in images that contain visual information for robust visual variations. In particular, the extraction and description of the regions of interest are successfully applied to detect these areas.

Because of the success of BoF approaches for image retrieval, several authors propose extensions of BoVF representations. [16] propose the random locality sensitive vocabulary (RLSV) scheme towards visual vocabulary construction in such scenarios, this method is not advantageous in terms of query complexity. A several range of methods has been proposed to incorporate spatial information to improve the BoVW model [17, 18]. [19] [19] propose a image retrieval system based on bag-of-features (BoF) model by integrating scale invariant feature transform (SIFT) and local binary pattern (LBP). Using a weighted Kmeans clustering algorithm, the image-based SIFT-LBP integration achieves the superior performance on a given benchmark problem. Most of them aimed at human actions recognition.
Current techniques are based on the extraction of many local information like SIFT [8] or SURF [9] in that produce average over 1000 points per image. Each of these points is then characterized by a dictionary of visual (visual pattern characteristic of a portion of the image).
Each image is thus represented by a "bag of visual words". Indexing techniques involve comparing each visual word in an image with all the words that exist in database. However, the computational cost of such techniques makes them difficult to use in an environment where many images are available. The method bag of words is to quantify the local descriptors calculated on regions of interest, which are previously extracted by a detector affine invariant. The quantization indices of these descriptors are called visual words, by analogy with the search of textual records. The image is represented by a histogram of the frequency of occurrence of visual words, this representation allows efficient computation of similarity between images.
The steps for the construction of the bags of words are as follows: (i) the identification or extraction of information in the image, (ii) quantifying with matching characteristics with local visual words, (iii) creating a histogram of frequency used to evaluate a global representation of the image. So using the method of word bag, it is possible to reduce the size of the vectors descriptors of images, while providing a compact representation of the images.
The BOF representation groups local descriptors. It requires the definition of a codebook of k “visual words” usually obtained by *k*-means clustering [10]. Each local descriptor of dimension *d* from an image is assigned to the closest centroid. The BOF representation is obtained as the histogram of the assignment of all image descriptors to visual words. Therefore, it produces a *k* dimensional vector, which is subsequently normalized.
There are several variations on how to normalize the histogram. When seen as an empirical distribution, the BOF vector is normalized using the Manhattan distance. Another common choice consists in using Euclidean normalization.
Several variations have been proposed to improve the quality of this representation. One of the most popular [11, 12] consists in using soft quantization techniques instead of a k-means.
The main advantages of the BOF representation are 1) its compactness, i.e., reduced storage requirements and 2) the rapidity of search. [13] propose to use more visual words in the BoW algorithm, and showed that using multiple independent dictionaries built from different subsets of the features increases significantly the recognition performance of BoW systems. [14] propose the structural features for object recognition are nested multi-layered local graphs built upon sets of SURF feature points with Delaunay triangulation. This representation conserves the invariance to affine transformations of image plane which the initial SIFT/SURF features have. A Bag-of-Visual-Words framework is applied on these graphs, giving birth to a Bag-of-Graph-Words representation. For each layer of graphs its own visual dictionary is built.

Our contribution consists in proposing a representation that provides excellent search accuracy with a reasonable vector dimensionality. We propose three descriptors, derived from both BOF, to produce a compact representation.

3. Proposed approach

This article addresses precise image search based on local descriptors. A variety of feature detection algorithms have been proposed in the literature to compute reliable descriptors for image matching [8], [9]. SIFT and SURF descriptors are the most promising due to good performance and have now been used in many applications. [20] summarize the performance of two robust feature detection algorithms namely Scale Invariant Feature Transform (SIFT) and Speeded up Robust Features (SURF) on several classification datasets. Based on this observation, we decide to use a bag-of-features approach over quaternion zernike moments (QZM). We compare the results of SIFT and SURF with those of QZM.

SIFT, SURF and QZM features are used because reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint. We propose an indexing method for content based search task that aims to retrieve a large collection of images and returns a ranked list of objects in response to a query image. The visual words quantize the space of descriptors. Here, we use the k-means algorithm to obtain the visual vocabulary. For each query image, the signature is calculated from these points can be compared over the signature of the query image with all signatures from the base.

Algorithm

1. Extract local features $\{d_n\}$ from every image i.
2. Constructing codebooks: randomly selecting centres from among the sampled training patches, and online k-means initialized using this.
3. The dictionaries are built using Approximate K-Means. Histograms are normalized to have unit l_1 norm, then the l_1 distance is used to measure similarity between histograms.

In the conventional bag-of-visual-words model, at first each image I is represented in terms of image descriptors: $I = \{d_1, d_2, \ldots.., d_n\}$, where d_i is the description of an image patch and n is the total number of patches in the image. By this way, we get numerous descriptors from all the local patches of all the images for a given dataset. Typically, K-means unsupervised clustering is applied on these descriptors to find clusters $W = \{w_1, w_2, \ldots.., w_N\}$, that constitutes the visual vocabulary, where N is the predefined number of clusters.
Figure 1 illustrates how each of these operations impacts our representation.

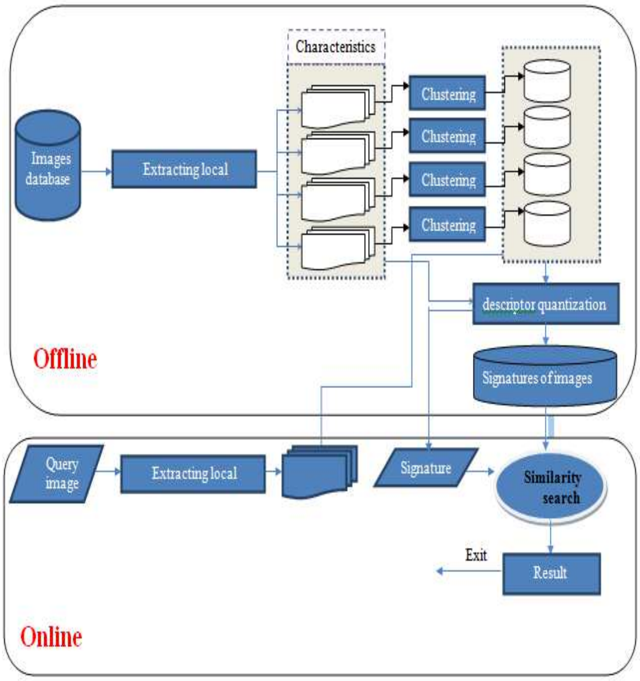

Fig. 1 The architecture of the approach suggested

3.1 Experiments

In this section, we first evaluate our descriptors and the joint dimensionality reduction.

3.1.1 Datasets and Evaluation

Two datasets have been used to perform the k-means clustering: the COIL-100 set, as well as the Corel dataset.
Columbia Object Image Library (COIL-100) is a database of color images of 100 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects were taken at pose intervals of 5 degrees. Figure 2 shows The COIL-100.

Fig. 2 Samples of Coil-100 dataset. The dataset includes 128x128.

COREL database composed of 1000 images distributed in 10 classes. The tags of the classes are: Africans, Beach, Architecture, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and Food. Figure 3 shows one sample per each class.

(a) Africans (b) Beach (c) Architecture (d) Buses (e) Dinosaurs

(f) Elephants (g) Flowers (h) Horses (i) Mountains (j) Food

Fig. 3 Samples of Corel 1000 dataset. The dataset includes 256x384 or 384x256 images.

The objective of this work is focused on finding images, and not detection. SIFT and SURF are two recent and competitive alternatives to image local featuring that we compare through QZM.
To do this, a set of points of interest is extracted from each descriptor, SIFT (128 dimensions), SURF (64 dimensions), and QZM (34 dimensions), it is mandatory for the construction of the visual dictionary, using the *K*-means algorithm applied to the associated descriptors. Each corresponds to a visual word class center.
For each query image, the signature is calculated from these points than we can compare the signature of the query image with all signatures from the database, then the user can send the images most similar to his query. In a retrieval system images, the user is interested in answers relevant system. So the image search systems require the evaluation of the accuracy of the response. This type of evaluation is considered performance evaluation research. We describe the two most common measures: recall and precision. This relationship is often described by a curve of recall and precision.

3.1.2 Results and analysis

In this section, we present some results, giving details of the curves - reminder for each descriptor.
Indeed it has been demonstrated in the literature, these methods currently give the best results.
The approach is based on an unsupervised classification (K-means) that we use to build the dictionary. The comparison result by varying the number of classes (K) to obtain the visual vocabulary. We note that for large dictionaries, the approach provides better results.

Regarding the descriptors according to their performance varis bases and therefore the application.
The mosr common evaluation measures used in CBIR are precision and recall, usually presented as a precision vs. recall.

$$\text{Precision} = \frac{\text{Number.relevant images retrieved}}{\text{Total number.images retrieved}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number relevant images retrieved}}{\text{Total relevant images in the collection}} \quad (2)$$

of images of the class that the target image belongs to. In literature SIFT and SURF give of good results, we have to compare this with QZM.
From these results, we can calculate the recall/ precision curve:

Table: Impact of the dataset used for k-means clustering (uncorrelated Corel dataset or the Coil-100) and of the vocabulary size.

Descriptors	Image database	Vocabulary size	Recall	Precision
SIFT	Coil	10	30%	60%
	Coil	50	40%	80%
	Corel	30	25%	48%
	Corel	50	35%	70%
SURF	Coil	10	20%	40%
	Coil	50	30%	60%
	Corel	30	30%	45%
	Corel	50	40%	60%

QZM	Coil	10	50%	100%
	Coil	50	65%	100%
	Corel	30	35%	70%
	Corel	50	40%	80%

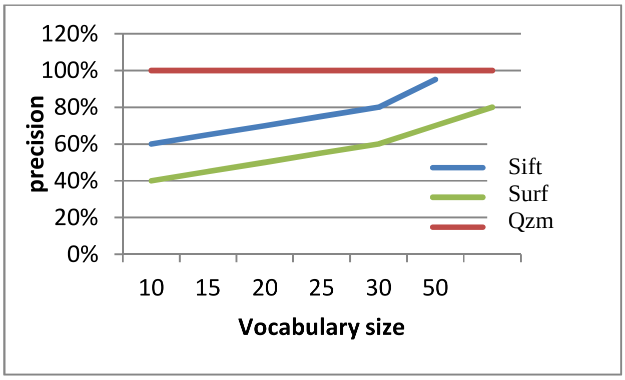

Fig. 4 Comparing results for SURF-based, SIFT-based and QZM- based approaches at various vocabulary sizes for coil-100

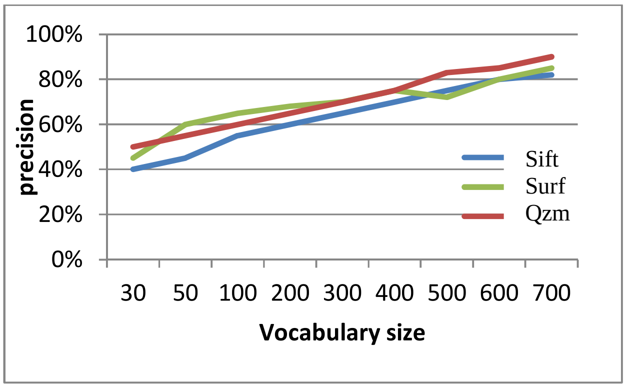

Fig. 5 Comparing results for SURF-based, SIFT-based and QZM- based approaches at various vocabulary sizes for corel-1000

3.1.3 Discussion

In this paper we have reported a analysis. We explored ways to boost the performance of BoW image search methods by using more visual words. We presented algorithm, QZM. Experimental results with Coil-100 and Corel image databases, demonstrate that the method QZM produces better performance.

4. Conclusion

The method allowed bags of words obtaining a compact representation of images, so that from a set of descriptors for each image, we obtain a single feature vector for each image, then we could therefore reduce the size of the search space using clustering, and the size of the descriptors (signatures) with an index structure that will limit the number of images to check to accelerate research (early treatment the query).
Furthermore, we develop descriptor QZM based on the bag-of-features approach and use the benchmark to demonstrate that they significantly outperform other descriptors in the literature such as SIFT and SURF.
The proposed approach is based on a reduction in the size of the feature vectors. The originality consists in an adaptation of a bag of words representation of the use of descriptors point of interest SIFT, SURF and QZM.

References

[1] I. Daoudi, K. Idrissi, et S. E. Ouatik, “A multi-class metric learning for contentbased image retrieval,” IEEE international Conference on Image processing, ICIP 2009.
[2] M. J. Swain and D. H. Ballard, “Color Indexing,” International Journal of Computer Vision, pp. 11-32, 1991.
[3] J. Sivic and A. Zisserman. “Video Google: A text retrieval approach to object matching in videos,” In ICCV, pages 1470– 1477, 2003.
[4] L. Fei-Fei and P.Perona. “A Bayesian hierarchical model for learning natural scene categories,” Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 524- 531, 2005.
[5] J. Vogel and B. Schiele, “On Performance Characterization and Optimization for Image Retrieval,” Proceedings of European Conference on Computer Vision, pp. 51-55, 2002.
[6] M. Vidal-Naquet and S.Ullman. “Object recognition with informative features and linear classification,” Proceedings of IEEE International Conference Computer Vision, pp. 281-288, 2003.
[7] R. Maree, P. Geurts, J. Piater, J. And L. Wehenkel, “Random subwindows for robust image classification,” Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 34-40, 2005.
[8] David G. Lowe. “Object recognition from local scale invariant features,” In ICCV, pages 1150–1157, 1999.
[9] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. “Speeded-up robust features (surf),”. Computer Vision and Image Understanding, 110(3) :346– 359, 2008.
[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” CVPR, 2007.
[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. “Lost in quantization: Improving particular object retrieval in large scale image databases,” In CVPR, June 2008.
[12] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. “Visual word ambiguity,”. PAMI. 2009
[13] Mohamed Aly 1 Mario Munich 2 Pietro Perona 1 “MULTIPLE DICTIONARIES FOR BAG OF WORDS LARGE SCALE IMAGE SEARCH,” IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, September 2011.
[14] Svebor Karaman, Jenny Benois-Pineau, Rémi Mégret and Aurélie Bugeau. “ Mots visuels issus de graphes locaux

multi-niveaux pour la reconnaissance d'objets,” RFIA (Reconnaissance des Formes et Intelligence Artificielle). 2012.

[15] H. Mahi, N. Benkablia, “ Moments de Zernike Quaternioniques pour la Classification des Bâtiments sur des données à Très Haute Résolution Spatiale,” Proceedings of International Conference on image processing, Avril 2012.

[16] Y.Mu, J. Sun, Tony X. Han, Loong-Fah Cheong, Shuicheng Yan, “Randomized Locality Sensitive Vocabularies for Bag-of-Features Model,” The 11th European Conference on Computer Vision (ECCV), 2010

[17] S.Kim, X. Jin, and J. Han. “Disiclass: discriminative frequent patternbased image classification,” In Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD’10, pages 7:1–7:10, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0220-3.

[18] D. Liu, G. Hua, Paul A. Viola, and Tsuhan Chen. “Integrated feature selection and higher-order spatial feature extraction for object categorization,”. In CVPR, 2008.

[19] X. Yuan, Jing Yu, Z. Qiny , and T.Wan “A SIFT-LBP IMAGE RETRIEVAL MODEL BASED ON BAG-OF-FEATURES,” 18th IEEE International Conference on Image Processing. 2011.

[20] N. Younus Khan, B. McCane *and* G. Wyvill. “SIFT and SURF Performance Evaluation Against Various Image Deformations on Benchmark Dataset,” International Conference on Digital Image Computing: Techniques and Applications. 2011.

Soumia BENKRAMA
PhD Student at the University of Sciences and Technology of Oran

Lynda ZAOUI
HdR at the University of Sciences and Technology of Oran

Christophe CHARRIER
HdR at the University of Caen Basse-Normadie