European Journal of Computer Science and Information Technology
Vol.1 No.1, pp.11-22, June 2013
Published by European Centre for Research Training and Development UK (www.ea-journals.org)

A SURF-COLOR MOMENTS FOR IMAGES RETRIEVAL BASED ON BAG-OF-FEATURES

Abdelkhalak Bahri and Hamid zouaki

Department of Mathematics and computer science, Faculty of Science, El Jadida, Morocco

Modélisation Mathématiques et Informatique Décisionnelle

Abstract: *An important research issue in multimedia databases is the retrieval of similar objects. Most of the Content-Based Image Retrieval (CBIR) system uses the low-level features such as color, texture and shape to extract the features from the images. In Recent years the Interest points are used to extract the most similar images with different view point and different transformations. SURF is fast and robust interest points detector/descriptor which is used in many computer vision applications. In the state-of-the-art the SURF is combined with Color Moments to improve the performance of the system. In this paper, we propose one presentation (LOWE 2004) to improving image search based on the color and shape descriptors. The representation is obtained by the quantification of the SURF (Herbert and all 2008) combined with the color moments (Stricker and all 1995), and so called Bag-of-Features and Colors (BOFC). Experiments show that our descriptor BOFC provides better results than a standard Bag of Visual Words approach based on SURF (BOF).*

Key words: bag of features, bag of words, local features, SURF, color moments, and CBIR.

INTRODUCTION

Feature-based approaches have recently become very popular in computer vision and image analysis applications, we can include the significant work of Lowe (LOWE 2004), Sivic and Zisserman (SIVIC 2003), and Mikolajczyk and Schmid (Mikolajczyk and all 2005a). In these approaches, an image is described as a collection of local features (”visual words”) from a given vocabulary, resulting in a representation referred to as a bag of features. The bag of features paradigm relies heavily on the choice of the local feature descriptor that is used to create the visual words. A common evaluation strategy of image feature detection and description algorithms is the stability of the detected features and their invariance to different transformations applied to an image. The main idea in the BOF is to quantize local invariant descriptors, for example obtained by an affine invariant interest point descriptor (LOWE 2004).

In recent years, the interest point detectors and descriptors (Mikolajczyk and all 2005a) are employed in many Content-based image retrieval (CBIR) systems. SURF (Speed Up Robust Feature) is one of the most and popular interest point detector and descriptor which has been published by Bay et al.(Herbert and all 2008). It is widely used in most of the computer vision applications. The SURF has been proven to achieve high repeatability and distinctiveness. It uses a Hessian matrix-based measure for the detection of interest points and a distribution of Haar wavelet responses within the interest point neighborhood as descriptor. An image is analyzed at several scales, so interest points can be extracted from both global and local image details. In addition to that, the dominant orientation of each of the interest points is determined to support rotation-invariant matching.

In this work, SURF algorithm is used to extract the features and the first order and second order color moments is calculated for the SURF key points to provide the maximum distinctiveness for the key points. In (Velmurugan and all 2011) the authors have combined the SURF with color Moments to improve the retrieval accuracy of the system. The most popular approach today initially proposed in (SIVIC 2003), relies on a bag-of-features (BOF) representation of the image.

The main advantages of the BOF representation are: (1) its compactness, i.e. reduced storage requirements, and (2) the rapidity of search due to an inverted file system. In detail, instead of storing a set of 64 dimensional SURF descriptors (Herbert and all 2008) for each image, we only have to store one entry for each existing visual word. In order to get better instances from an image, we introduce bag of features and colors (BOFC) based on SURF (Herbert and all 2008) combined with the color moments (Stricker and all 1995). However the BOFC comes from visual words more accurate. Since these visual words are built from the surfs combined with the first and the second color moments for each interest region of each image, which aims to have more distinctiveness of visual words, and consequently to have more discriminate description of each image.

The rest of this paper is organized as follows. The SURF descriptor described in section 2, In section 3, the bag of features image representation is discussed; the proposed descriptors are presented in section 4; the experiments are discussed in section 5; the paper is concluded in section 6.

SURF

Our method extracts salient features and descriptors from images using SURF. This extractor is preferred over SIFT (LOWE 2004) due to its concise descriptor length. Whereas the standard SIFT implementation uses a descriptor consisting of 128 floating point values, SURF condenses this descriptor length to 64 floating point values. As our aim is to obtain the image object segmentation in real-time, the SURF descriptor (Herbert and all 2008) becomes a natural choice, as its performance is similar to more complex descriptors like the IHOG (Zhu and all 2006), but with a lower computational cost. The SURF descriptor is a histogram of Haar wavelet responses accumulated at different spatial bins of the local region. Let dx and dy be the Haar wavelet response in the horizontal and vertical directions respectively. Then, a SURF descriptor with P×Q spatial bins is calculated by integrating the wavelets responses d_x and d_y over each bin sub-region. To take into account the polarity of the intensity images, the sum of absolute values of the responses, namely $\|d_x\|$and , is also extracted (see a SURF descriptor example in Figure 1). Hence, a four-dimensional descriptor vector of the form $\sum d_x, \sum\|d_x\|, \sum d_y, \sum\|d_y\|$ is generated for each spatial bin of the descriptor. Finally, the resulting P×Q×4 dimensional descriptor is normalized using L1-norm.

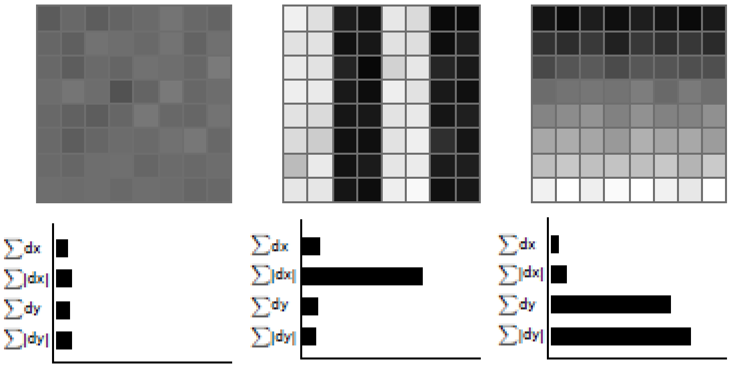


Figure 1: Example of SURF descriptors obtained from different image patches (from left to right): in a homogeneous region, in the presence of frequencies in the x direction, and when the intensity is gradually increasing along the y axis.

In SURF, a descriptor vector of length 64 is constructed using a histogram of gradient orientations in the local neighborhood around each key point.

Bag of features image representation

The model of Bag of Visual Words is used to express the visual content of images by compacting visual description of all regions of interest in a histogram structure. This model has become popular in the recent years due to the effectiveness and its results quality (Philbin and all 2007, Velmurugan and all 2011, Philbin and all 2008, Jegou and all 2010). It was presented first by (Stricker and all 1995) in the case of video retrieval and (Nister and all 2006) in the image categorization domain. To apply the Bag of Visual Words model, a clustering algorithm (e.g. K-Means) is applied on the visual descriptions of regions of interest, and the each resulted centroid represent a visual word. The set of visual words are called visual vocabulary. Then, images are viewed as bags of visual words represented as histograms. However, we note that according to (Wengert and all 2011), the visual words are much more ambiguous than text words: ”It is impossible to create words that are always observed on the same part of an object and anywhere else.”. In the Bag of Visual Word representation, we lose all information related to the topological organization of the regions of interest in the image, this information can be important to describe and differentiate objects. There have been several techniques put forth in the literature for improving the performance of a Bag of visual words image retrieval process by post query processing of the result set. As mentioned previously, Sivic et al. employ a spatial consistency re-ranking process that alters the initial query results based on how well an estimated affine homography maps the feature points between the query and gallery image (SIVIC and all 2006, Philbin and all 2008, Jegou and all 2010). While this method does improve the results, it adds significant additional computation and may not be feasible for massive data sets. Another way to improve query results is through rank aggregation (Wengert and all 2011). In rank aggregation, the query is performed multiple times using a separate vocabulary (index) for each query. The results are aggregated by, for example, taking the mean rank of the results. This increases the query time and index size. The separate queries can be processed in parallel to mitigate the time penalty, but there is still the cost of aggregation. From a storage perspective, multiple indexes increase the space requirements. At a high level, the procedure for generating a Bag of Features image representation is shown in Figure 2.

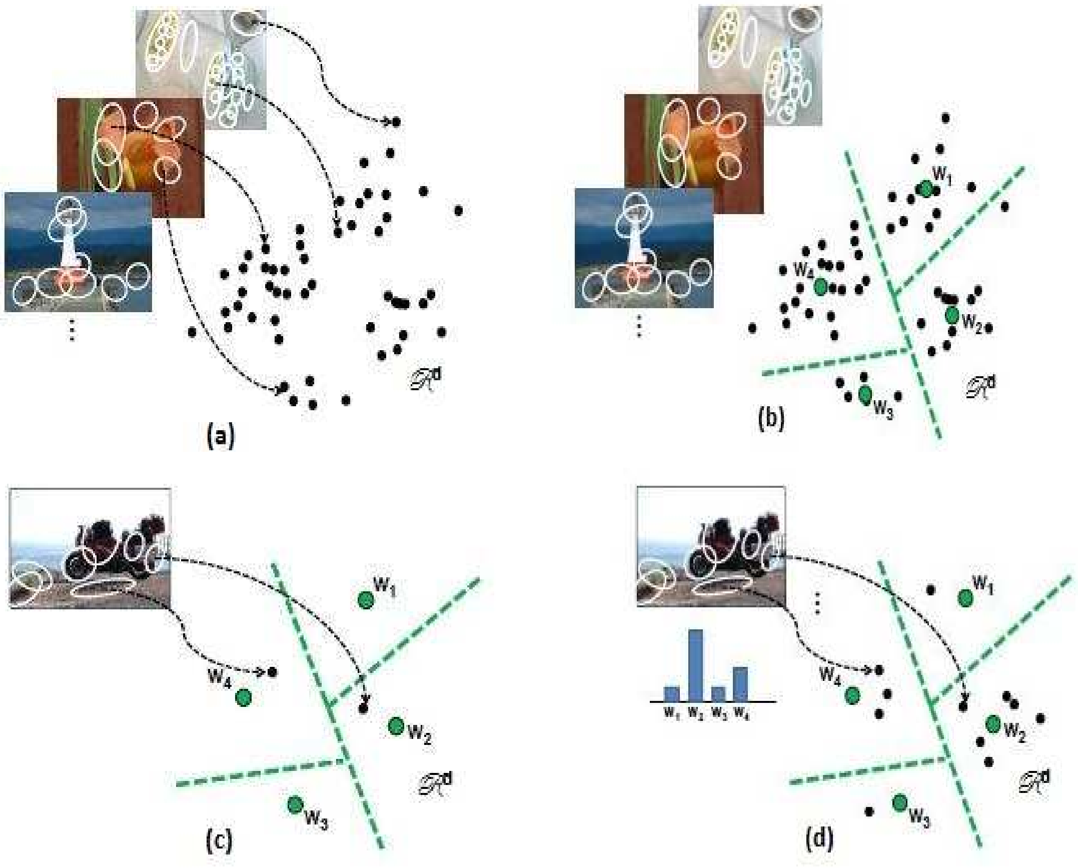


Figure 2: Process for Bag of Features Image Representation

Algorithm 1 Algorithm to building the BOF

(a) A large corpus of representative images are used to populate the feature space with descriptor instances. The white ellipses denote local feature regions, and the black dots denote points in some feature space, e.g., SURF.
(b) Next the sampled features are clustered in order to quantize the space into a discrete number of visual words. The visual words are the cluster centers, denoted with the large green circles. The dotted green lines signify the implied Voronoi cells based on the selected word centers.
(c) Now, given a new image, the nearest visual word is identified for each of its features. This maps the image from a set of high-dimensional descriptors to a list of word numbers.
(d) A bag-of-visual-words histogram can be used to summarize the entire image. It counts how many times each of the visual words occurs in the image.

There is a number of design choices involved at each step in the BoF representation. One key decision involves the choice of feature detection and representation. Many use an interest point operator, such as the Harris-Affine detector (Mikolajczyk 2005b) or the Maximally Stable External Regions (MSER) detector (Matas and all 2004). At every interest point, often a few thousand per image, a high-dimensional feature vector is used to describe the local image patch. Lowe’s 128-dimension SIFT descriptor is a popular choice (LOWE 2004). Another pair of design choices involve the method of vector quantization used to generate the vocabulary and the distance measure used to assign features to cluster centers. A distance measure is also required when comparing two term vectors for similarity (as is done with image retrieval), but this measure operates in the term vector space as opposed to the feature space. Vector Quantization (Clustering) is used to build the visual vocabulary in Bag of Features algorithms. Nearest-neighbor assignments are used not only in the clustering of features but also in the comparison of term vectors for similarity ranking or classification. Thus, it is important to understand how quantization issues, and the related issues involving measuring distances in feature and term vector space, affect Bag of Features based applications. There are a great many clustering/vector quantization algorithms, and this report does not attempt to enumerate them. Many BoF implementations are described as using Kmeans (SIVIC and all 2003, Lazebnik and all 2006, Jiang and all 2007), or an approximation thereof for large vocabularies Nister and all 2006, Philbin and all 2007). Given any clustering method, there will be points that are equally close to more than one centroid. These points lie near a Voronoi boundary between clusters and create ambiguity when assigning features to terms. With K-means and similar clustering methods, the choice of initial centroid positions affects the resultant vocabulary. When dealing with relatively small vocabularies, one can run K-means multiple times and select the best performing vocabulary during a validation step. This becomes impractical for very large data sets. When determining the distance between two features, as required by clustering and term assignment, common choices are the Manhattan (L1), Euclidean (L2), or Mahalanobis distances. A distance measure is also needed in term vector space for measuring the similarity between two images for classification or retrieval applications. Euclidean and Manhattan distances over sparse term vectors can be computed efficiently using inverted indexes, and are thus popular choices.

Vocabulary Size

While the size of text vocabulary is predefined by the corpus, the size of visual vocabulary is obtained by clustering methods. Therefore, the chosen size of the vocabulary becomes tricky and interesting. A small vocabulary may lack the discriminative power since two local descriptors may be assigned into the same cluster even if they are not similar to each other. A large vocabulary, on the other hand, is less generalizable, less forgiving to noises, and incurs extra processing overhead (Jiang and all 2007). Additionally, distinct datasets may also prefer different size of vocabulary since the vocabularies themselves are not similar due to the topics of datasets. Actually, the vocabulary size varies a lot as mentioned in different papers, from 1K to 1M. But it is suggested that the broader dataset covers, the larger the vocabulary should be. And constrained by the memory size and scale of computing power, large vocabulary could be impractical.

Proposed descriptor

The BOF has been inspired by the Bag-Of-Words (BOW) for text document representation. In the BOW, a text document is represented by the number of occurrences of the words in the document. Despite the simplicity of the model, which neither takes into account the order of the words, nor the relationships between them, this model is very efficient for document classification tasks (Hofmann 2000). Sivic and Zisserman (SIVIC 2003) proposed to compute a visual dictionary by clustering similar visual entities inspired by BOW. Hence, to build a visual dictionary, we must define two key concepts: what entity defines the spatial support for a visual word and which descriptor underpins the notion of similarity in the clustering process. In the BOF, local interest points are used as salient image patches and SIFT or related (Mikolajczyk and all 2005a) descriptors used to describe the patches. In this paper we use the SURF descriptor (Herbert and all 2008) as based features.

Bag of features and colors (BOFC)

Surf works only on gray scale images. The Color Moments are used to extract the color features from the region of 5x5 pixels around the SURF interest point. Since most of the information is concentrated on the low order moments, only the first moment (mean) and the second moments (variance) will be used as the color features.

The value of the ith color channel at the jth image pixel is P_{ij}. The index entries related to this color channel are calculated by:

$$E_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij}, \qquad and \qquad \sigma_i = \left[\frac{1}{N}\sum_{j=1}^{N}(P_{ij} - E_i)^2\right]^{1/2}$$

Where N is the number of pixels in the image patch. For each interest point, color moments are calculated for the region of 5×5 pixels around the SURF interest point for the RGB channel. The first order and second order color information are concatenated with the SURF descriptor to obtain the descriptor vector length 70 (64 bins for SURF, 3 bins for mean, and other 3 bins for variance).

After this step, the Algorithm 1 is applied to build the BOFC for each image. The goal of using the color moments and SURF descriptor is to provide the maximum distinctiveness for the key points.

Experiment and results

We performed our proposed BOCF, on COIL-100(Nene and all 2006) databases. COIL-100(Nene and all 2006) is a popular image database for benchmark which contains 72 views for 100 objects acquired by rotating the object under study about the vertical axis. Figure 3, shows one example image from each category in the database. In this database we choose 15 different categories consisting of 72 images in each category.

A schematic illustration of the experiment is shown in Figure. 2. In the training process, SURF, the first, and the second color moments are extracted from each image in the database, and then be integrated for constructing the SURF-Color moments descriptors. After the K-means clustering, the tow visual vocabularies are constructed by the SURF, and the integrated SURF-color moments descriptors. A descriptor is categorized into its cluster centroid using a Euclidean distance metric. Each image is mapped to the tow visual vocabularies in order to obtain its BOF, and BOFC histograms. The BOF, and the BOFC are building for all images in the database. An experimental training set of 75 query images is created by randomly choosing 5 images from each class.

Figure 3: Shows a sample database of 15 images by randomly selecting one image from each category of the COIL-100 database.

For a query image, each extracted descriptor is mapped into its nearest cluster centroid. A histogram of counts is constructed by incrementing a cluster centroid's number of occupants each time a descriptor is placed into it. The result is that each image is represented by a histogram (BOF, BOFC) vector of length N. It is necessary to normalize each histogram to make this procedure invariant to the number of descriptors used. First we study the influence of the size N of the vocabulary on retrieval performance of the system. We separately let N = 50, N = 100, N=150, N = 200, and N=250 in the comparison experiments.

In order to explain the well performance of BOFC, we give the result of the experiments by the same platform and the same image database.

To check the performance of proposed technique the precision and recall is used. The standard definitions of these two measures are given by following equations.

$$\Pr ecision = \frac{Number\ \ of\ \ relevant\ \ images\ \ retrieved}{Number\ \ of\ \ images\ \ retrieved} \tag{1}$$

$$\operatorname{Re} call = \frac{Number\ \ of\ \ relevant\ \ images\ \ retrieved}{Total\ \ number\ \ of\ \ relevant\ \ images\ \ in\ \ the\ \ database} \tag{2}$$

Table 1 lists the average precision for each image class using BOF. Table 2 lists the average precision for each image class using BOFC. we recall that the image numerated form the top left to the bottom right. The figure 4 shows the average precision according to recall for each average vocabulary size. And the figure 5 shows also the average precision for each average vocabulary size.

Table 1: Retrieved images using BOF descriptor

Image categories	Vocabulary size of BOF descriptor					AVERAGE PRECISION
	K=50	K=100	K=150	K=200	K=250	
1	69.44	80.56	94.44	86.11	83.33	82.78
2	51.39	87.50	68.06	75.00	55.56	67.50
3	19.44	18.06	22.22	18.06	20.83	19.72
4	86.11	81.94	83.33	86.11	81.94	83.89
5	47.22	50.00	31.94	43.06	29.17	40.28
6	93.06	91.67	86.11	91.67	66.67	85.83
7	62.50	66.67	62.50	68.06	47.22	61.39
8	63.89	79.17	77.78	66.67	81.94	73.89
9	83.33	80.56	87.50	86.11	72.22	81.94
10	25.00	20.83	25.00	25.00	19.44	23.06
11	77.78	95.83	93.06	100.00	98.61	93.06
12	72.22	81.94	76.39	76.39	31.94	67.78
13	95.83	88.89	95.83	91.67	95.83	93.61
14	68.06	63.89	68.06	63.89	61.11	65.00
15	30.56	34.72	16.67	27.78	37.50	29.44
Average precision	63.06	68.15	65.93	67.04	58.89	64.61

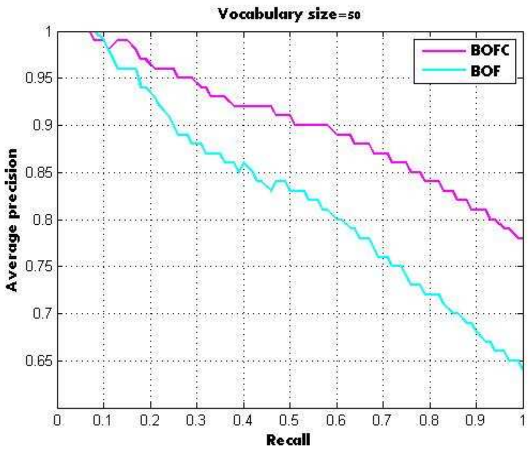
Table-1 gives number of total relevant images in the set of first 72 retrieved images for all 15 categories. The percentage of Precision/Recall for all categories of the resultant images are shown in Table-1.The Average Precision for all 15 categories is 64.61%.

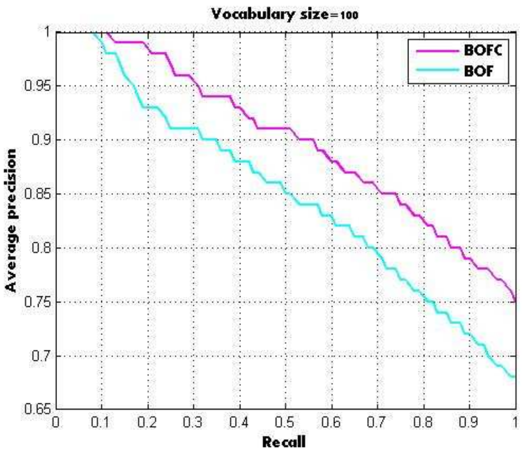
Table 2: Retrieved images using BOFC descriptor

Image categories	vocabulary size of BOFC descriptor					Average Precision
	K=50	K=100	K=150	K=200	K=250	
1	87.50	97.22	100.00	97.22	91.67	94.72
2	73.61	69.44	72.22	59.72	79.17	70.83
3	20.83	25.00	22.22	22.22	20.83	22.22
4	81.94	91.67	91.67	93.06	93.06	90.28
5	44.44	45.83	34.72	40.28	31.94	39.44
6	76.39	91.67	70.83	97.22	84.72	84.17
7	97.22	95.83	72.22	61.11	65.28	78.33
8	94.44	83.33	91.67	94.44	95.83	91.94
9	94.44	97.22	95.83	95.83	87.50	94.17
10	97.22	52.78	95.83	88.89	45.83	76.11
11	95.83	100.00	100.00	100.00	100.00	99.17
12	77.78	84.72	93.06	83.33	91.67	86.11
13	100.00	100.00	98.61	97.22	93.06	97.78
14	93.06	68.06	69.44	66.67	70.83	73.61
15	29.17	29.17	55.56	38.89	38.89	38.33
Average precision	77.59	75.46	77.59	75.74	72.69	75.81

Table-2 gives number of total relevant images in the set of first 72 retrieved images for all 15 categories. The percentage of Precision/Recall for all categories of the resultant images are shown in Table-2.The Average Precision for all 15 categories is 75.81 %.

In the retrieval process, we input a query image, by comparing its BOF, and BOFC histograms and other BOF, and BOFC histograms in the tow databases; we can obtain a ranked set of most similar images based on the Euclidian distance. Figure 4 and 5 shows the average precision by using BOF and BOFC descriptors for nearest neighbor search. It shows that our proposed descriptor BOFC is more outperforms than the standard BOF in term of accuracy.





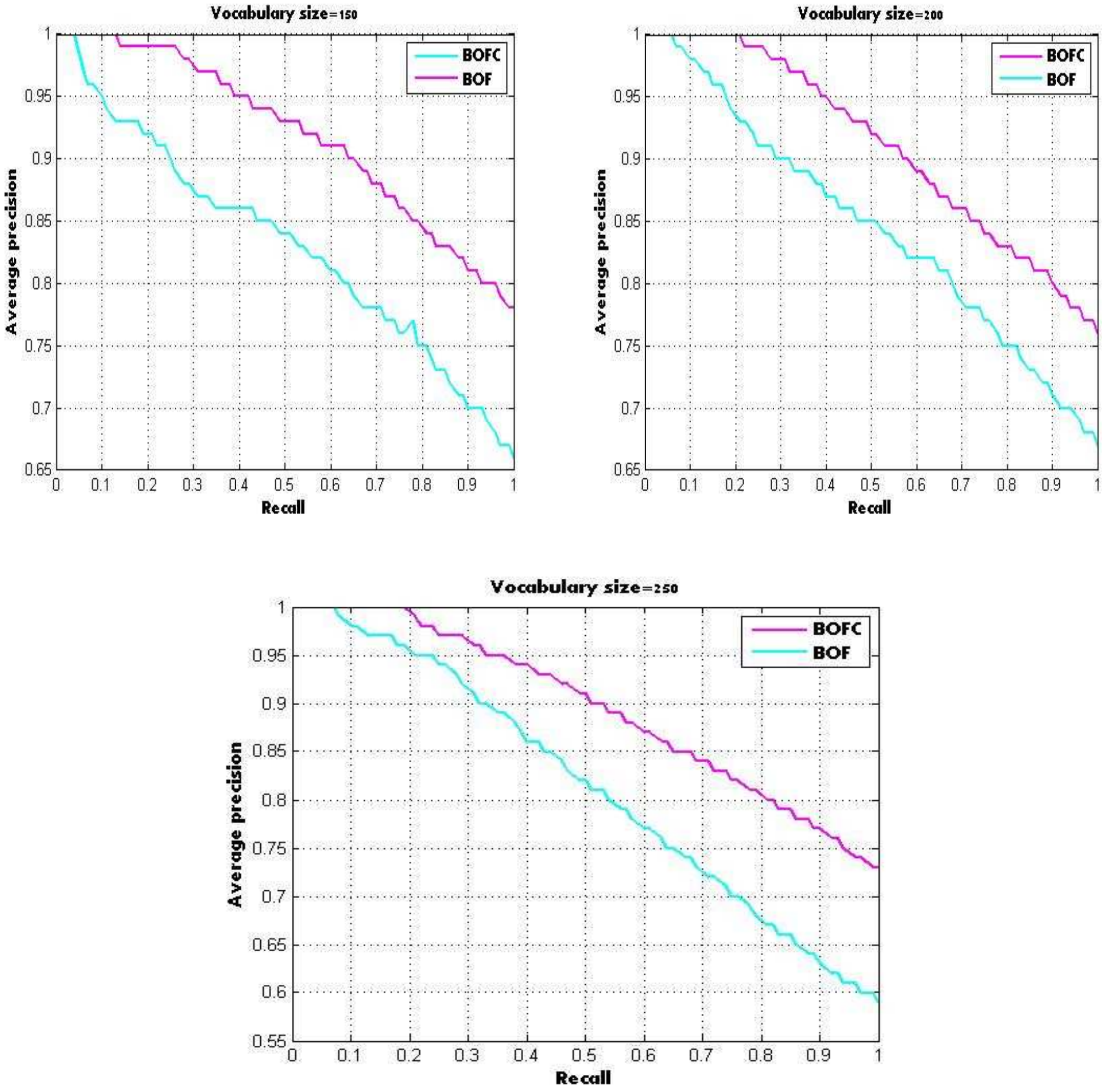


Figure 4: The average precision according to recall for each vocabulary size

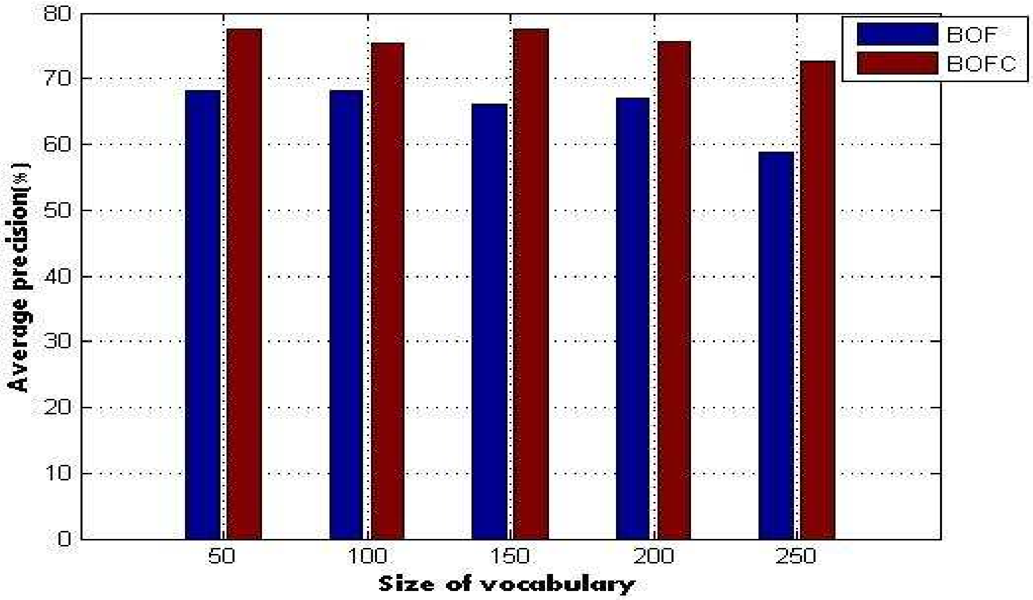
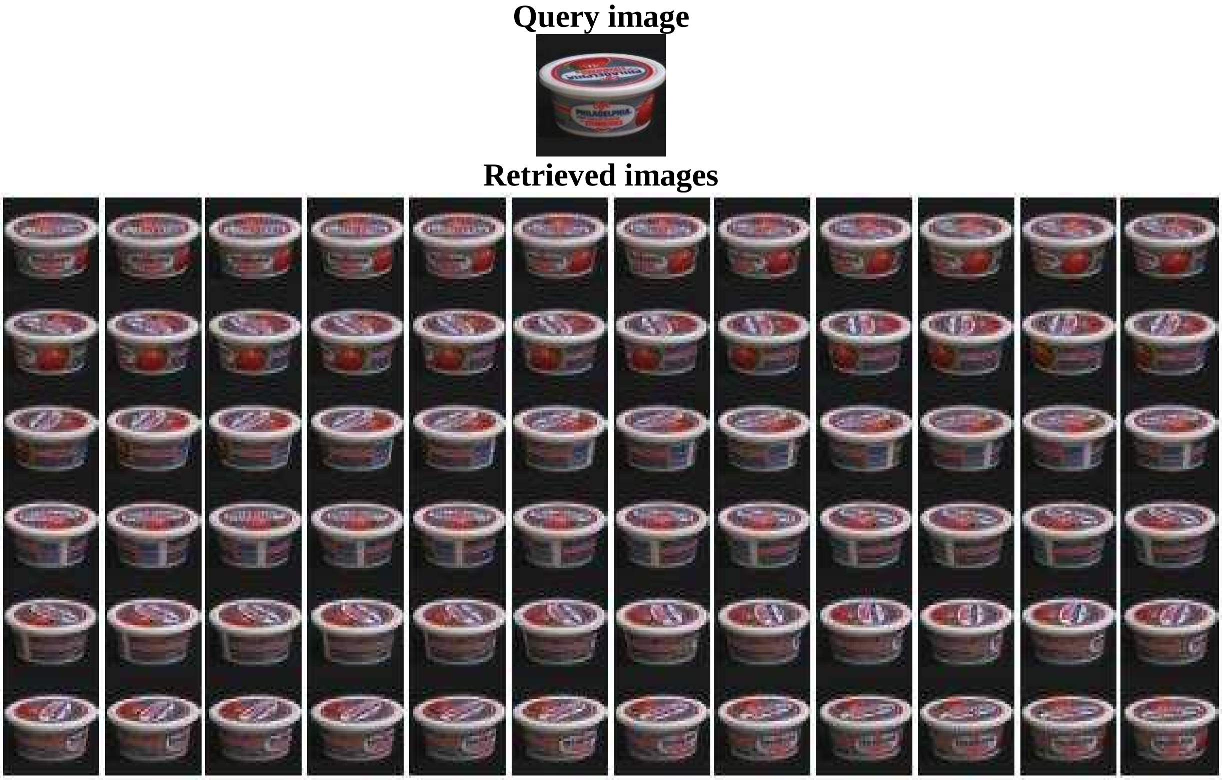


Figure 5: Comparison of retrieval results with different number of vocabulary size. Average retrieval precision (ARP) for different vocabulary size are presented using BOF and BOFC. Vocabulary size 50 and 150 gives the best performances for BOFC. And also the vocabulary size 50, 100, and 200 and gives the best performances for BOF.

The previous figures give an overview of the performance of our descriptor BOFC. Curves precision / recall are shown for each descriptor with different sizes. So we can see the goal of our descriptor. Results currency change for each visual vocabulary size, but the performance of our descriptor (BOFC) is always better than the standard descriptor (BOF). According to the previous schemas, the result of BOFC is really better especially for k = 50 and K = 150 (for example for k=50 and k=150 the average precision equal a 77.59% for BOFC, but the one in the same size is equal a 63.6% and 65.93 for BOF respectively). However, the result with k = 50 and k = 150 are the same. Another interesting point is that the result with k = 100, k = 200, k = 250 are (almost) the same. This means that it is enough with k = 50 visual words and if we increase the amount of words, the result does not change or changes slightly. And consequently there will be a significant gain in terms of accuracy, response time and memory occupation. Then it says that the combination of colors moments and SURF descriptor before the creation of visual words is a good idea. In the end, results of the proposed descriptor (BOFC) is more efficiency than the BOF for all vocabulary size, and this is explained by the contribution of using multiple descriptors at creating visuals words of BOFC.


Query image

Retrieved images

Figure 6: Shows Results of query image using BOFC. Note that the database contains total 72 images. For the query image as shown in this figure for 72 retrieved images, the total relevant images obtained are all 72.

1.0 Conclusion

In this paper, we have proposed a novel method for image retrieval based on the bag-of-features model. The SURF and color moments are integrated in to a single descriptor. Based on the experimental studies on a COIL-100 images retrieval problem, the BOFC that is based on the SURF and the color moments give the best performance comparing to the standard BOF based on the SURF only. In the future work we plan to improve the efficiency of the search in term of accuracy by combining them with the textual vocabulary.

References

Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, ”SURF: Speeded Up Robust Features”, Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

Hofmann T. Learning the similarity of documents : an information-geometric approach to document retrieval and categorization. In Advances in Neural Infor- mation Processing Systems, 2000.

Jegou H, Schmid C, Harzallah H, Verbeek J, Accurate image search using the contextual dissimilarity measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(1):2-11,2010

Jiang YG, Ngo CW, Yang J (2007) towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proc. CIVR

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR, vol 2

LOWE D.: Distinctive image features from scale invariant key points. IJCV 60, 2 (2004), 91-110

Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10):761-767

Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Gool LV (2005b) A comparison of affine region detectors. International Journal of Computer Vision 65(1):43-72

Mikolajczyk K. and C. Schmid, A Performance Evaluation of Local Descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (10), pp. 1615-1630, 2005a.

Nene S.A., S.K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, CUCS, 1996.

Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proc. CVPR

Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Proc. CVPR

Philbin J, Chum O, Isard M, Sivic J, Zisserman A , Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proc. CVPR 2008

SIVIC J., ZISSERMAN A.: Video Google: A text retrieval approach to object matching in videos. In Proc. ICCV (2003), vol. 2, pp. 1470-1477.

Stricker M.A. and M. Orengo. Similarity of color images. In SPIE, Storage and Retrieval for image Video Databases, pages 381-392, 1995.

Velmurugan K., Lt.Dr.S. Santhosh Baboo Content-Based Image Retrieval using SURF and Color Moments, Global Journal of Computer Science and Technology Volume 11 Issue 10 Version 1.0 May 2011.

Wengert C., M Douze, H Jgou Bag-of-colors for improved image search, ACM Multimedia, 2011.

Zhu Q., M.C. Yeh, K.T. Cheng and S. Avidan, Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 14911498, 2006.

Authors email addresses: Bahri_abdelkhalak@yahoo.fr, hamid_zouaki@yahoo.fr