Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighing Scheme for Efficient Indexation

Bouachir, Wassim; Kardouchi, Mustapha; Belacel, Nabil

Publisher's version / Version de l'éditeur:

Proceedings of the 5th IEEE International Conference on Signal Image Technology and Internet Based Systems(SITIS 2009), 2009-12-04

Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation

Wassim Bouachir[1], Mustapha Kardouchi[1], Nabil Belacel[2]
ewb8080@umoncton.ca, Mustapha.kardouchi@umoncton.ca, nabil.belacel@nrc-cnrc.gc.ca

[1]*Computer Science Department, University of Moncton, Moncton, NB, E1A3E9, Canada*
[2]*NRC, Institute for Information Technology, 100 des Aboiteaux, Suite 1100, Moncton, NB, E1A7R1, Canada*

Abstract

Recent works on Content Based Image Retrieval rely on bag of visual words to index visual content. Analogically to the bag of words approach in text retrieval, this model of description represents an image as a vector of weights, where each weight corresponds to the importance of a visual word in the image, and is computed according to the chosen weighting scheme. Instead of using the known weighting schemes directly migrated from text retrieval domain, we propose a new approach specifically for images. The proposed weighting scheme is based on a fuzzy model to take into account the fundamental difference that exists between textual words and visual words. For experiments, two datasets with very different properties are used. The tests clearly demonstrate that retrieval based on the proposed technique produces better results than standard term weighting schemes.

1. Introduction

Content Based Image Retrieval (CBIR) is a technology that aims to organize images in response to a query, based on visual content. This technology differs from traditional retrieval systems, based on keywords to describe images.

One of the basic problems in CBIR is how to transform visual contents into distinctive descriptors for dissimilar images, and into similar descriptors for images that look alike. In other words, the main problem is to translate the semantic similarity to visual similarity when indexing images.

A number of indexing methods for CBIR have been proposed. Most of works use global statistics of images [1]. Swain and Ballard [2] were the first to use color histograms and their intersections to compute a distance between images. Since then, many other features were applied for image indexation, for example the colorimetric moments [3] and the color sets [4].

Recently, the notion of keypoints was introduced. These local descriptors are used to describe interest points that form an object, and SIFT (Scale Invariant features Transform) [5, 6] is proven to be one of the best local descriptors [7], being reasonably invariant to changes in illumination, noise, rotation, scaling and viewpoint. Therefore, SIFT descriptors have been widely used as an effective image representation for several computer vision tasks like in object detection, image stitching and 3D scene modeling. Generally, this kind of applications exploits the SIFT descriptors in a local context, comparing and matching similar keypoints.

Local descriptors have also been adapted for image retrieval, to represent an image in a global context using a single vector, and Bag-of-Visual-Words is one of these techniques. Based on local descriptors, this approach is analogous to the bag-of-words representation of text document in automatic text retrieval in terms of form and semantics. The construction of a Bag-of-Visual-Words vector requires three main steps:

local description of the image, visual vocabulary construction and image indexation. Therefore, each image is represented by a Bag-of-Visual-Words signature which is traditionally a histogram of its patches, i.e. a bin of the histogram represents a visual word, and contains the associated information which depends on the chosen weighting scheme. We have seen the use of presence or absence information, its count in the image (the number of keypoints in the corresponding visual word), or the weighted count [8]. In effect, these are the most used term weighting techniques in text retrieval [9]. In Bag-of-Visual-Words approach, an image is described by its visual words just like a document is described by the terms in automatic text retrieval. However the visual words aren't quite as meaningful. For example, when describing a text document by a bag-of-words signature, each word is counted in the corresponding entry of the vocabulary, e.g. "walks", "walking" and "walked" would be counted in the entry of "walk". But when mapping an image's keypoints to visual words, each word is counted in the nearest entry of the visual vocabulary, based on a distance measure. This may introduce a loss of fidelity to image signature: two keypoints associated with the same vocabulary entry contribute in the same way to the construction of the signature, whether they are identical or appreciably different. Furthermore, two similar keypoints may be considered in two different entrees. Certainly, increasing the vocabulary size attenuates this disadvantage, but involves a longer processing time when responding to a query. The aim of this work is to propose a method keeping simplicity of the Bag-of-Visual-Words approach while minimizing the effects due to the vocabulary size choice and similarity between visual words. The proposed weighting scheme is based on a fuzzy modeling of the distribution of the keypoints. This paper is organized as follows: section 2 describes the development of the indexation system based on Bag-of-Visual-Words approach and reviews the existent weighting schemes. In Section 3, the proposed approach for visual-words weighting is presented. Section 4 provides detailed experimental results. Finally, section 5 concludes the paper.

2. Bag-of-Visual-Words approach

The visual words denote local features extracted from a large sampling of images and then quantized to form a visual vocabulary (codebook). Finally, an image is described by a histogram where each bin represents a visual word, and the associated weight represents its frequency in the image. Thus, constructing a BoVW signature requires three steps: extracting local descriptors, building a visual vocabulary and indexing images.

a. Local descriptors extraction

We use SIFT [5, 6] as the interest point detector and descriptor. The first step detects salient locations known as keypoints that are identifiable from different viewpoints and are usually around the corners and edges in images objects. Second, the keypoints are represented by a 128 elements vector summarising the edge information in the image patch centered at the keypoint. The output of this step is a set of SIFT local descriptors extracted from a large sampling of images to be used to build the visual vocabulary.

b. Building a visual vocabulary

The construction of the visual vocabulary is an important step of the BoVW model. In fact, each image in the dataset will be represented using the visual words of this vocabulary. To this end, the generated vocabulary must be the most representative possible. In practice, building the visual vocabulary means quantifying all the already extracted local descriptors. Since their space is not dense and uniform and certain vectors may not appear again whereas others appear frequently, a clustering algorithm is used to quantize SIFT vectors. The clustering is performed using the standard k-means algorithm, where the number of clusters is the vocabulary's size and the cluster's centers are the visual words.

c. Visual indexing

Once the visual vocabulary is built, we index the images in the collection constructing their BoVW signatures. An image's BoVW signature requires finding the weight of each visual word from the vocabulary. Thereby, each image is represented by a histogram where the bins are the vocabulary's entrees and the weights are the appearance frequencies in the image. Analogously to the term weighting techniques in text retrieval, a visual word's weight is formed by three factors. First the visual word is

frequently mentioned in an image which suggests a *term frequency (tf)* factor as a part of the weight. The second is the *inverse document frequency (idf)*.this is a collection-dependent factor used to favour visual words found in a few images of the collection. The intuition is that *tf* weights visual words occurring often in a particular image, whilst *idf* down-weights those that appear often in the collection. The third term-weighting factor is a *normalization* component introduced to treat all the images equally, even if their number of keypoints differs. Table 1 summarizes the popular term weighting schemes in text retrieval where they are named and described after the convention in [9].

name	value	description
		Term frequency factor
b	1,0	Binary i.e. 1 for visual words present, 0 if not
t	*tf*	Number of occurrence of the visual word.
		Collection frequency factor
x	1,0	No change in weight
f	$\log \frac{N}{n}$	Multiply by *idf* (N is the number of images in the collection, and n the number of images containing the visual word.
		Normalization factor
x	1,0	No normalization
c	$\frac{1}{\sum w_i}$	Each visual word weight w_i is divided by the sum of the of the image's weights.

Table 1: term weighting factors

For image search, we have seen the use of *term frequency-inverse document frequency* (*tfx)* in [10, 11] and the count of visual words *(txx)* in [12]. We have also seen the use the normalized term frequency *(txc)* [13] and binary weights *(bxx)*[14] for image classification. Instead of using a text retrieval weighting schema, we propose a more realistic approach to weight visual words using a fuzzy assignment.

3. The proposed Visual-Words weighting approach

a. Drawbacks of existing approaches:

An empirical study of the impact of weighting scheme choice on classification performance [8] concluded that the best weighting scheme varies according to the vocabulary size and image properties. Since there is a fundamental difference between text words and images keypoints, we believe that using term weighting schemes directly migrated from automatic text retrieval domain is not an optimal choice. In fact, the text words vocabulary is generated from the text corpus according to a natural language. Hence, the document's term vector is constructed finding each word's vocabulary entry naturally according the language's grammar and semantic. By contrast, the visual words vocabulary is the output of vector quantization using the clustering algorithm. Thus an image's BoVW signature is generally obtained mapping keypoints to the most similar visual word based on a distance measure.

This may reduce the signature's discriminative power since two keypoints may be assigned to the same visual word (cluster) even if they are not equally similar to the visual word (i.e. they don't have the same distance to the cluster's center). Consequently, the two keypoints contribute in the same way in the signature's construction and the obtained value doesn't reflect the real weight of the visual word in the

image. Certainly, the more the vocabulary' size is increased, the more this effect is opposed. But in this case the vocabulary would be less generalizable, noise sensitive and incurs longer processing time for the retrieval.

b. The proposed fuzzy representation

Suppose that $V=\{v_1,v_2,..., v_i,...v_k\}$ is the visual vocabulary formed by the k centers of clusters (visual words) obtained after vector quantization with k-means algorithm. Let p_j, $j \in \{1, 2, \ldots, M\}$ be a SIFT local descriptor among M keypoints detected from an image. We associate to p_j a fuzzy description considering all the vocabulary's visual words. This represents the contribution of the keypoint in the weight of each visual word. To this end, a membership degree is defined using the fuzzy membership function of Fuzzy-C-Means algorithm [15].

$$U_{ij} = \frac{1}{\sum_{n=1}^{k}\left(\frac{||p_j - v_i||}{||p_j - v_n||}\right)^{\frac{2}{m-1}}}$$

U_{IJ}: The contribution of the keypoint whose the descriptor is p_j in the weight of the visual word whose the center is v_i.

m: the degree of fuzziness, $m \in]1,\infty[$.

Thus, a fuzzy histogram is obtained and each bin represents the fuzzy weight of the corresponding visual word. Such a representation takes into account the similarity between the keypoint and each visual word and resolves a major problem with existent weighting schemes.

To illustrate this effect, let's suppose that SIFT descriptors are 2 dimensions vectors.

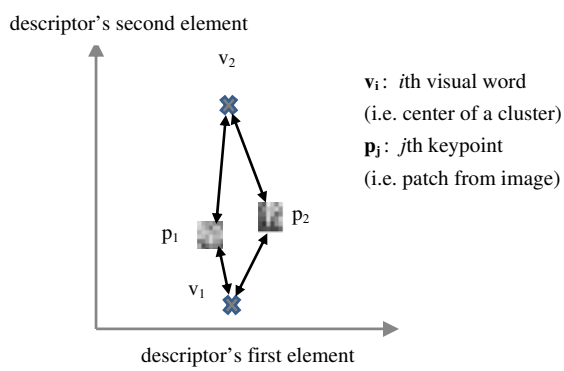

Figure 1: similarity measurment before assigning keypoints to visual words

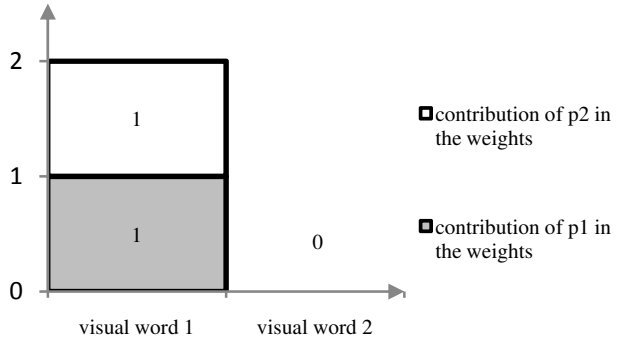

Figure 2- Crisp assignment

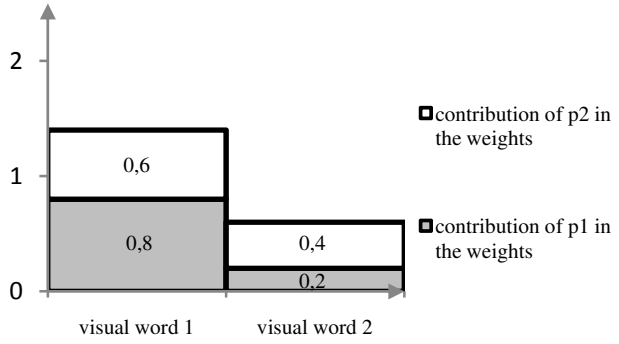

Figure 3- Fuzzy assignment

Figures 1,2 and 3 represent the contribution of the keypoints p_1 and p_2 in the weights of visual words, supposing that local descriptors are two dimensions vectors. In figure 2, the two keypoints p_1 and p_2 contribute in the same way to the weight of their nearest cluster's center even if they are not equally similar to this visual word (figure 1). Using the fuzzy assignment (figure 3), the two keypoints contribute in the weights of the visual words v_1 and v_2 and thus the distribution of weights is more equitable.

$m \in]1,\infty[$ controls the degree of fuzziness in the distribution of weights. Empirically, we found that m=1,1 is the best setting (ou bien; Empirically, we found that m=1,1 maximizes the retrieval performance

4. Experiments

a. Images collections

To evaluate the proposed approach and compare to the other weighting schemes, two databases with different properties are used: COREL-1000 and COIL-100.

COREL is a collection of about 60000 images created by the professor Wang's group at Penn State University. COREL-1000[1] is a well known sub-collection that contains 1000 natural images divided into ten categories with 100 images per category. Figure 4 shows ten scenes randomly selected for experiments (one image per category).

[1] available at http://wang.ist.psu.edu/docs/related.shtml

Figure 4- Sample images from COREL-1000 database

(The images have the same size: 384x256 or the inverse)

Figure 5- Sample images from COIL-1000 database

(The images have the same size 128x128)

The Columbia University COIL-100 database[2] contains 7200 images of 100 different objects, where 72 images where taken at 72 different viewpoints separated by 5°. Figure 5 represents the ten different objects selected also randomly for experiments.

The keypoints in two samples of databases are detected and described by SIFT. We used 300 and 3000 randomly sampled images from COREL-1000 and COIL-100 to extract and describe SIFT keypoints. For each sample, we use the k-means clustering algorithm to cluster keypoints descriptors into a visual word vocabulary of 100 entrees. To compare the proposed approach with existing weighting schemes, the images are finally indexed in 6 different ways using *bxx, txx, txc, tfx, tfc* and the proposed *Fuzzy weight* and several queries are performed using the Euclidian distance to compute the similarity between signatures.

[2] available at http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

b. Experimental results

In this section, we evaluate the proposed fuzzy weighting scheme and compare it with existing methods: *bxx, txx, txc, tfx, tfc*. To this end, we use statistics *recall* and *precision*. Where *precision* is defined as the number of correctly retrieved images by a search divided by the total number of images retrieved, and *recall* denotes the number of images retrieved by a search divided by the number of images of the class that the target image belongs to. The *precision/recall* curve is obtained making vary the number of images returned by a query. For each query, the recall and precision are computed for the 10, 20, 40, 60, 100 and 200 most similar images retrieved.

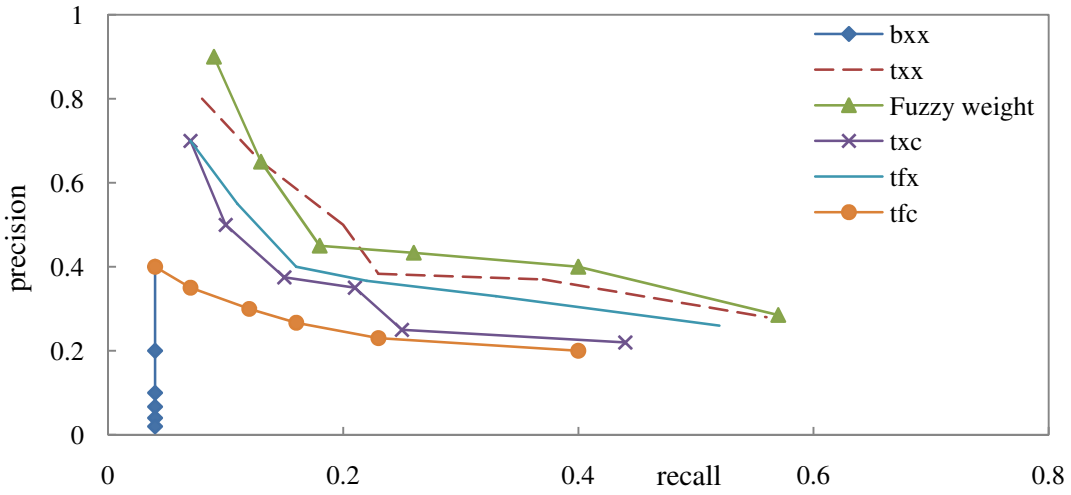

Figure 6 – Recall versus Precision: image COREL9

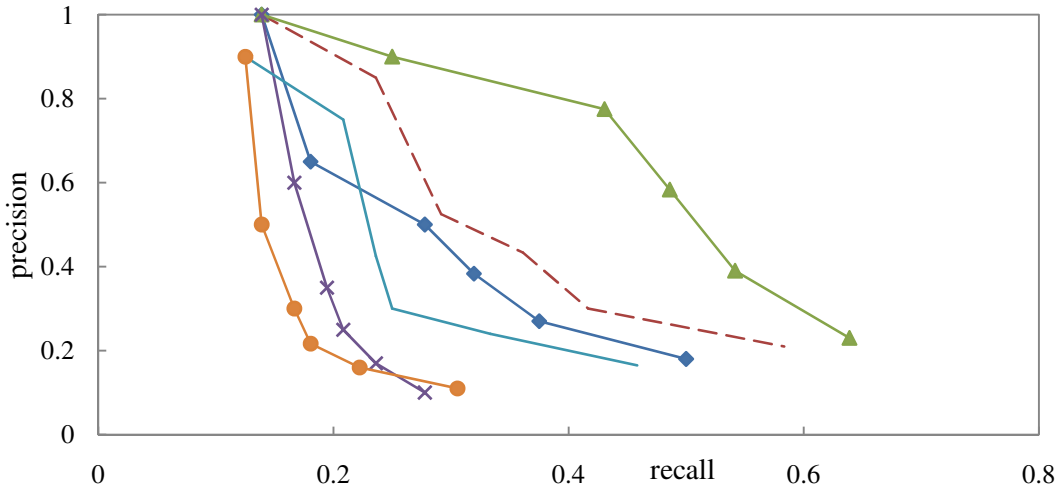

Figure 6 – Recall versus Precision: image COIL10

Figures 5 and 6 present the results of *precision* versus *recall* for two images COREL9 and COIL10 showing that the precision rate decreases as recall increases. In figure 5 *bxx* has the worst performance and the Fuzzy weight outperforms the others in most *recall/precision* points. For the image COIL10, it's clear that the proposed Fuzzy weight outperforms significantly the others.

To further compare the performance of various weighting schemes, we performed on each database 10 queries returning the 20 most similar images, using the images in figure 4 and 5 as targets. Table 2 presents the average precision of retrieval results by using the different weighting schemes and shows that the *Fuzzy weight* has the best average precision for the two databases. We also completed the measurements for the 10, 40, 60, 100 and 200 most similar images retrieved to plot the average precision versus average recall in figure 7 and 8. The first figure shows that when indexing COREL-1000 images using fuzzy weights, retrieval results are slightly better than those obtained with *txx, txc* and *tfx* while *bxx* and *tfc* had the worst performance. For the COIL-100 database (figure 8), it's clear that indexation based on the proposed fuzzy model gives considerably better retrieval results than all other methods.

database	***bxx***	***txx***	***txc***	***tfx***	***tfc***	***Fuzzy weight***
COREL-1000	0,145	0,635	0,61	0,6	0,51	**0,655**
COIL-100	0,62	0,715	0,665	0,615	0,565	**0,8**

Table 2 – average precision of 10 queries for each database for different weighting schemes

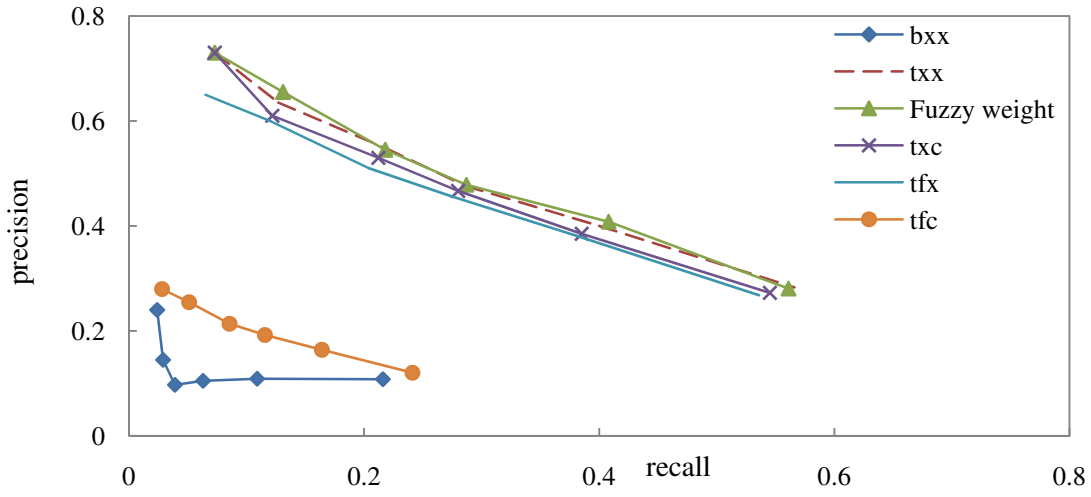

Figure 7 – average recall versus average precision for ten queries on COREL-1000

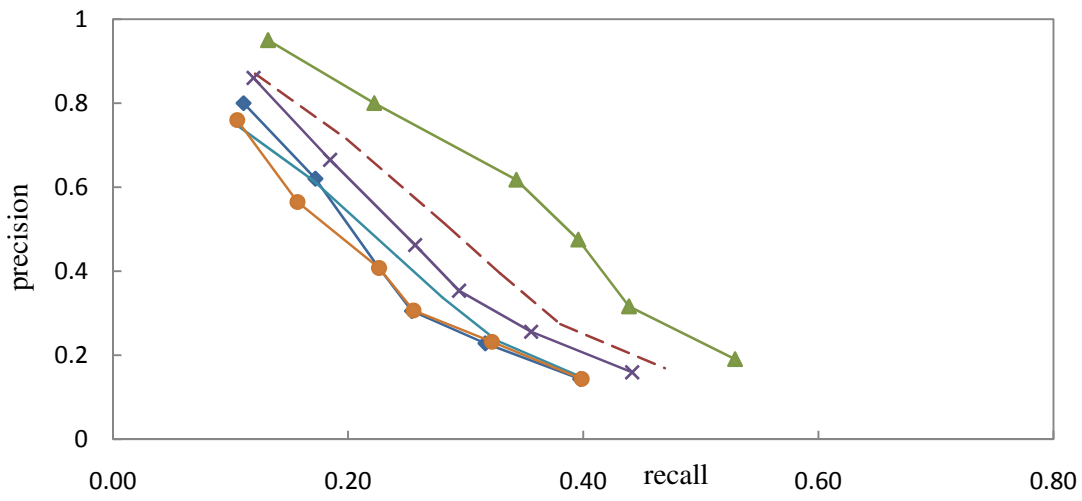

Figure 8 – average recall versus average precision for ten queries on COIL-100

5. Conclusion

This study confirms that Bag-of-visual-words is a reliable indexation method to represent visual content for image retrieval. Although BoVW is known for its simplicity and effectiveness, we have shown that using representation techniques directly migrated from automatic text retrieval domain is not an optimal choice. To remedy this drawback, we defined a fuzzy model, specifically for visual words instead of using known term weighting schemes. The proposed approach takes into account the fundamental difference between text and images and the conducted experiments proved its superiority.

BoVW approach can be improved by several other ways, such as using a more effective algorithm to create the visual vocabulary, taking into account their large number and noisy keypoints that may be considered . We believe also that the color provides valuable information in keypoints description. Since SIFT descriptors use only gray scale information and neglect color, a very important source of distinction may be lost. Consequently, a further improvement of BoVW would be by introducing the color information to describe keypoints.

One other interesting direction for future work is to decompose the image's signature into sub-histograms each corresponding to a part of the described image. As a result, the BoVW signature is enriched by the information on the spatial relation between visual words.

Acknowledgement

References

[1] Y. Rui, T. S. Huang and S. Fu. Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Visual Communication and Image Representation 10*, 1999.

[2] M. J. Swain and D. H. Ballard. Color Indexing. *International Journal of Computer Vision*, 1991.

[3] M. Stricker and M. Orengo. Similarity of color images. *Proceedings of SPIE Vol. 2, Storage and Retrieval for Image and Video Databases*, 1995.

[4] J. R. Smith and S. Chang. Single color extraction and image query. *Proceedings of the 1995 International Conference on Image Processing*.

[5] David G. Lowe. Object Recognition from Local Scale-Invariant Features. *Proceedings of the International Conference on Computer Vision*, Corfu, September 1999.

[6] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *The International Journal of Computer Vision*, 2004.

[7] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 27, Number 10*, 2005.

[8] J. Yang, Y. Jiang, A. G. Hauptmann, C. Ngo. Evaluating bag-of-visual-words representations in scene classification. *Proceedings of the international workshop on multimedia information retrieval*, 2007.

[9] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an Int'l Journal*, 1988.

[10] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2003.

[11] W. Zhao, Y. Jiang, C. Ngo. Keyframe retrieval by keypoints : Can point-to-point matching help? *Proceedings of the 5th international Conference on Image and Video Retrieval*, 2006.

[12] S. Newsam and Y. Yang. Comparing Global and Interest Point Descriptors for Similarity Retrieval in Remote Sensed Imagery. *Proceedings of the 15th International Symposium on Advances in Geographic Information Systems*, 2007.

[13] Y. Yang and S. Newsam. Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery. *Proceedings of the 15th IEEE International Conference on Image Processing*, 2008.

[14] E. Nowak, F. Jurie and B. Triggs, Sampling strategies for bag-of-features image classification. *Proceedings of the European Conference on Computer Vision*, 2006.

[15] J. Bezdeck, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press Ed., New-York, 1981.