Biomedical Image Retrieval in a Fuzzy Feature Space with Affine Region Detection and Vector Quantization of a Scale-Invariant Descriptor

Md Mahmudur Rahman, Sameer K. Antani, and George R. Thoma

U.S. National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA
{rahmanmm,santani,gthoma}@mail.nih.gov

Abstract. This paper presents a biomedical image retrieval approach by detecting affine covariant regions and representing them with an invariant fuzzy feature space. The covariant regions simply refers to a set of pixels or interest points which change covariantly with a class of transformations, such as affinity. A vector descriptor based on Scale-Invariant Feature Transform (SIFT) is then associated with each region, computed from the intensity pattern within the region. The SIFT features are then vector quantized to build a codebbok of keypoints. By mapping the interest points extracted from one image to the keypoints in the codebook, their occurrences are counted and the resulting histogram is called the "bag of keypoints" for that image. Images are finally represented in fuzzy feature space by spreading each region's membership values through a global fuzzy membership function to all the keypoints in the codebook. The proposed feature extraction and representation scheme is not only invariant to affine transformations but also robust against quantization errors. A systematic evaluation of image retrieval on a biomedical image collection demonstrates the advantages of the proposed image representation approach in terms of precision-recall.

1 Introduction

In recent years, rapid advances of software and hardware technology in medical domain facilitate the generation and storage of large collections of images by hospitals and clinics every day [1]. Such images of various modalities constitute an important source of anatomical and functional information for the diagnosis of diseases, medical research and education. In a heterogeneous medical collection with multiple modalities, such as ImageCLEFmed benchmarks [1] [2], images are often captured with different views, imaging and lighting conditions, similar to the real world photographic images. Distinct body parts that belong to the same modality frequently present great variations in their appearance due to changes in pose, scale, illumination conditions and imaging techniques applied. Ideally, the representation of such images must be flexible enough to cope with a large

[1] http://ir.ohsu.edu/image/

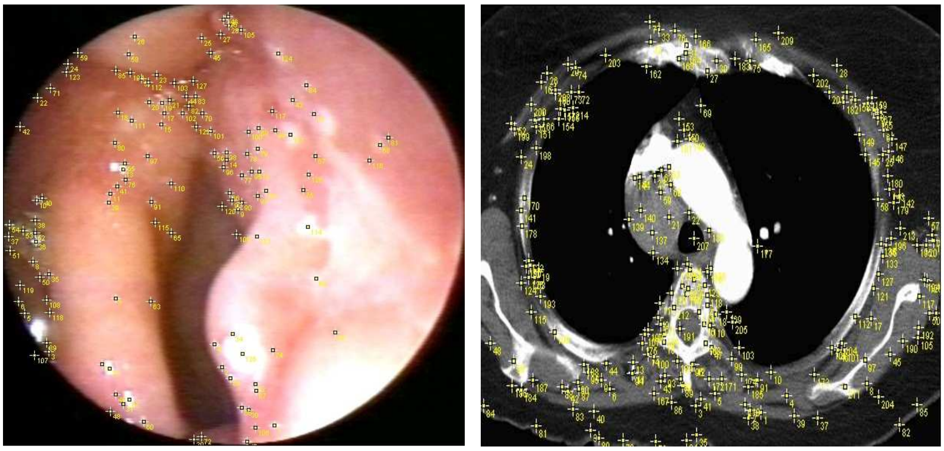
variety of visually different instances under the same category or modality, yet keeping the discriminative power between images of different modalities.

Recent advances in computer vision and pattern recognition techniques have given rise to extract such robust and invariant features from images, commonly termed as affine region detectors [3]. The regions simply refers to a set of pixels or interest points which are invariant to affine transformations, as well as occlusion, lighting and intra-class variations. This differs from classical segmentation since the region boundaries do not have to correspond to changes in image appearance such as color or texture. Often a large number, perhaps hundreds or thousands, of possibly overlapping regions are obtained. A vector descriptor, such as scale invariant feature transform (SIFT) [6] is then associated with each region, computed from the intensity pattern within the region. This descriptor is chosen to be invariant to viewpoint changes and, to some extent, illumination changes, and to discriminate between the regions. The calculated features are clustered or vector quantized (features of interest points are converted into visual words or keypoints) and images are represented by a bag of these quantized features (e.g., bag of keypoints) so that figures are searchable similarly with "bag of words" in text retrieval.

The idea of "bag of keypoints"-based image representation has already been applied to the problem of texture classification and recently for generic visual categorization with promising results [8, 9]. For example, the work described in [9] presents a computationally efficient approach which has shown good results for objects and scenes categorization. Besides, being a very generic method, it is able to deal with a great variety of objects and scenes. However, the main limitation of keypoint-based approaches is that the quality of matching or correspondence (i.e., covariant region to keypoints) is not always exact. During the image encoding process, a region in general is classified or matched to a single keypoint only and the rest are simply overlooked or ignored. Hence, the correspondence of an image region to a keypoint is basically *"one-to-one"* due to the nature of hard classification. In reality, there are usually several keypoints with almost as closely match as the one detected for a particular image region. Although, two regions will be considered totally different if they match to different keypoints even though they might be very similar or correlated to each other.

To overcome the above limitation, this paper presents an image representation scheme with "bag of keypoints" based on a fuzzy soft annotation scheme. In this approach, the SIFT features are extracted at first from the covariant regions and then vector quantized to build a visual vocabulary of keypoints by utilizing the Self-Organizing Map (SOM)-based clustering. The images are presented in a fuzzy feature space by spreading each region's membership values through a global fuzzy membership function to all the keypoints in the codebook during the encoding and consequent feature extraction process. The organization of the paper is as follows: Section 2 describes the keypoint-based feature representation approach and an image representation scheme is a fuzzy feature space is presented in Section 3.Experiments and analysis of the results are presented in Sections 4 and 5. Finally, Section 6 provides our conclusions.

2 “Bag of Keypoints”-based feature representation



(a) Endoscopy Gastro Image

(b) Chest CT Image

Fig. 1. Images from the medical collection marked (white crosses) with interest points detected by the affine region detector.

A major component of this retrieval framework is the detection of interest points in scale-space, and then determine an elliptical region for each point. Interest points are those points in the image that possess a great amount of information in terms of local signal changes [3]. In this study, the Harris-affine detector is used as interest point detection methods [4]. In this case, scale-selection is based on the Laplacian, and the shape of the elliptical region is determined with the second moment matrix of the intensity gradient [5]. Fig. 1 shows the interest points (cross marks) detected in two images of different modalities from the medical collection.

A vector descriptor which is invariant to viewpoint changes and to some extent, illumination changes is then associated with each interest point, computed from the intensity pattern within the point. We use a local descriptor developed by Lowe [6] based on the Scale-Invariant Feature Transform (SIFT), which transforms the image information in a set of scale-invariant coordinates, related to the local features. SIFT descriptors are multi-image representations of an image neighborhood. They are Gaussian derivatives computed at 8 orientation planes over a 4×4 grid of spatial locations, giving a 128-dimension vector. Recently in a study [3] several affine region detectors have been compared for matching and it was found that the SIFT descriptors perform best. SIFT descriptor with affine covariant regions gives region description vectors, which are invariant to affine transformations of the image. A large number of possibly overlapping regions are obtained with the Harris detector. Hence, a subset of the representative region vectors is then selected as a codebook of keypoints by applying a SOM-based clustering algorithm [10].

For each SIFT vector of interest point in an image, the codebook is searched to find the best match keypoint based on a distance measure (generally Euclidean). Based on the encoding scheme, an image I_j can be represented as a vector of keypoints as

$$\mathbf{f}_j^{\mathrm{KV}} = [\hat{f}_{1j} \cdots \hat{f}_{ij} \cdots \hat{f}_{Nj}]^{\mathrm{T}} \tag{1}$$

where each element $\hat{f}_{ij}$ represents the normalized frequency of occurrences of the keypoints c_i appearing in I_j.

This feature representation captures only a coarse distribution of the keypoints that is analogous to the distribution of quantized color in a global color histogram. As we already mentioned, image representation based on the above hard encoding scheme (e.g., to find only the best keypoints for each region) is very sensitive to quantization error. Two regions in an encoded image will be considered totally different if their corresponding keypoints are different even though they might be very similar or correlated to each other. In the following section, we propose an effective feature representation scheme to overcome the above limitation.

3 Image representation in a fuzzy feature space

There are usually several keypoints in the codebook with almost as good match as the best matching one for a particular covariant region. This scheme considers this fact by spreading each region's membership values through a global fuzzy membership function to all the keypoints in the codebook during the encoding and consequent feature extraction process. The vector $\mathbf{f}^{\mathrm{Keypoint}}$ is viewed as a keypoint distribution from the probability viewpoint. Given a codebook of size N, each element f_{i_j} of the vector $\mathbf{f}_j^{\mathrm{Keypoint}}$ of image I_j is calculated as $f_{i_j} = l_i/l$. It is the probability of a region in the image encoded with label i of keypoint $c_i \in C$, and l_i is the number of regions that map to c_i and l is the total number of regions detected in I_j.

According to the total probability theory [11], f_{i_j} can be defined as follows

$$f_{i_j} = \sum_{k_j=1}^{l} P_{i|k_j} P_k \ = \frac{1}{l} \sum_{k_j=1}^{l} P_{i|k_j} \tag{2}$$

where P_k is the probability of a region selected from image I_j being the k_jth region, which is $1/l$, and $P_{i|k_j}$ is the conditional probability of the selected k_jth region in I_j maps to the keypoint c_i. In the context of the keypoint-based vector $\mathbf{f}^{\mathrm{keypoint}}$, the value of $P_{i|k_j}$ is 1 if the k_jth region is mapped to c_i or 0 otherwise. Due to the crisp membership value, this feature representation is sensitive to quantization errors.

In such a case, fuzzy set-theoretic techniques can be very useful to solve uncertainty problem in classification tasks [12, 15, 16]. This technique assigns an observation (input vector) to more than one class with different degrees instead of a definite class by crisp classification. In traditional two-state classifiers, an input vector $\mathbf{x}$ either belongs or does not belong to a given class A; thus, the characteristic function is expressed as [12]

$$\mu_A(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{otherwise.} \end{cases}$$

In a fuzzy context, the input vector $\mathbf{x}$, belonging to the universe X, may be assigned a characteristic function value or grade of membership value $\mu_A(\mathbf{x})$ $(0 \leq \mu_A(\mathbf{x}) \leq 1)$ which represents its degree of membership in the fuzzy set A.

Many methods could be adapted to generate membership from input observations. These include the histogram method, transformation of probability distributions to possibility distributions, and methods based on clustering [12, 15]. For example, fuzzy-c-means (FCM) [15] is a popular clustering method, which embeds the generation of fuzzy membership function while clustering. Few schemes have been proposed to generate fuzzy membership functions using SOM [13, 14], where the main idea is to augment the input feature vector with the class labeling information. However, without any class label information (as in our case), it might be difficult to generate such fuzzy membership functions. Due to this, we perform a two-step procedure, where in the first step we generate the proper clusters (e.g., keypoints in the codebook) based on the SOM clustering and next the fuzzy membership values are generated according to the generated clusters in the first step as follows [16]:

The membership degree μ_{ik_j} of a region vector $\mathbf{x}_{k_j} \in \Re^d, k = 1, 2, \cdots, l$, of the k_jth region in I_j to keypoint vectors $\mathbf{c}_i, i = 1, 2, \cdots, N$ is:

$$\mu_{ik_j} = \frac{\frac{1}{\left\|\mathbf{x}_{k_j} - \mathbf{c}_i\right\|^2}^{\frac{2}{m-1}}}{\sum_{n=1}^{N} \frac{1}{\left\|\mathbf{x}_{k_j} - \mathbf{c}_n\right\|^2}^{\frac{2}{m-1}}} \tag{3}$$

The higher the distance of an input SIFT vector from a keypoint vector, the lower is its membership value to that keypoint based on (3). It is to be noted that when the distance is zero, the membership value is one (maximum) and when the distance is infinite, the membership value is zero (minimum). The values of μ_{ik_j} lies in the interval $[0, 1]$. The fuzziness exponent $\frac{2}{m-1}$ controls the extent or spread of membership shared among the keypoints.

In this approach, during the image encoding process, the fuzzy membership values of each region to all keypoints are computed for an image I_j based on (3), instead of finding the best matching keypoint only. Based on the fuzzy membership values of each region in I_j, the *fuzzy keypoint vector* (FKV) is represented as $\mathbf{f}_j^{\mathrm{FKV}} = [\hat{f}_{1_j}, \cdots, \hat{f}_{i_j}, \cdots \hat{f}_{N_j}]^{\mathrm{T}}$, where

$$\hat{f}_{i_j} = \sum_{k=1}^{l} \mu_{ik_j} \ P_k = \frac{1}{l} \sum_{k=1}^{l} \mu_{ik_j}; \quad \text{for } i = 1, 2, \cdots, N \tag{4}$$

The proposed vector essentially modifies probability as follows. Instead of using the probability $P_{i|k_j}$, we consider each of the regions in an image being related to all the keypoints in the codebook based on the fuzzy-set membership function such that the degree of association of the k_j-th region in I_j to the keypoint c_i is determined by distributing the membership degree of the μ_{ik_j} to the corresponding index of the vector. In contrast to the keypoint-based vector (e.g., $\mathbf{f}^{\mathrm{Keypoint}}$), the proposed vector representation (e.g., $\mathbf{f}^{\mathrm{FKV}}$)considers not only the similarity of different region vectors from different keypoints but also the dissimilarity of those region vectors mapped to the same keypoint in the codebook.

4 Experiments

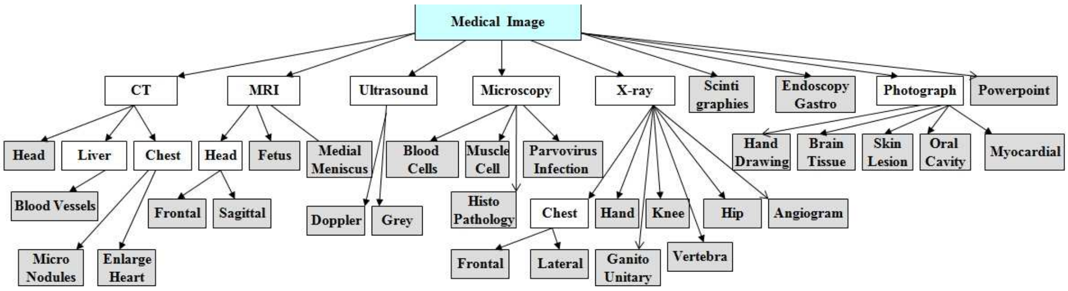


Fig. 2. Classification structure of the medical image data set.

The image collection for experiment comprises of 5000 bio-medical images of 32 manually assigned disjoint global categories, which is a subset of a larger collection of six different data sets used for medical image retrieval task in ImageCLEFmed 2007 [2]. In this collection, images are classified into three levels as shown in Fig. 2. In the first level, images are categorized according to the imaging modalities (e.g., X-ray, CT, MRI, etc.). At the next level, each of the modalities is further classified according to the examined body parts (e.g., head, chest, etc.) and finally it is further classified by orientation (e.g., frontal, sagittal, etc.) or distinct visual observation (e.g. CT liver images with large blood vessels). The disjoint categories are selected only from the leaf nodes (grey in color) to create the ground-truth data set.

To build the codebook based on the SOM clustering, a training set of images is selected beforehand for the learning process. The training set used for this purpose consists of 10% images of the entire data set (5000 images) resulting in a total of 500 images. For a quantitative evaluation of the retrieval results, we selected all the images in the collection as query images and used *query-by-example (QBE)* as the search method. A retrieved image is considered a match if it belongs to the same category as the query image out of the 32 disjoint categories at the global level as shown in Fig. 2. Precision (percentage of retrieved images that are also relevant) and recall (percentage of relevant images that are retrieved) are used as the basic evaluation measure of retrieval performances [7]. The average precision and recall are calculated over all the queries to generate the precision-recall (PR) curves in different settings.

5 Results

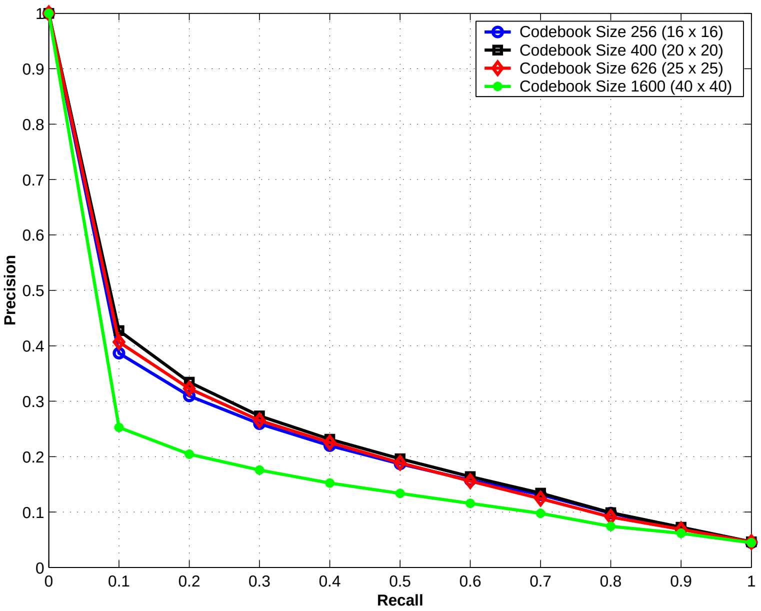


Fig. 3. PR-graphs of different codebook sizes.

To find an optimal codebook that can provide the best retrieval accuracy in this particular image collection, the SOM is trained at first to generate two-dimensional codebook of four different sizes as 256 (16×16), 400 $(20 \times 20\ )$, 625 (25×25), and 1600 (40×40) units. After the codebook construction process, all the images in the collection are encoded and represented as "bag of keypoints" as described in Section 2. For training of the SOM, we set the initial learning rate as $\alpha = 0.07$ due to its better performance.

Fig. 3 shows the PR-curves on four different codebook sizes. It is clear from Fig. 3 that the best precision at each recall level is achieved when the codebook size is 400 (20×20). The performances are degraded when the sizes are further increased, as a codebook size of 1600 (40×40) showed the lowest accuracies among the four different sizes. Hence, we choose a codebook of size 400 for the generation of the proposed keypoints-based feature representation and consequent retrieval evaluation.

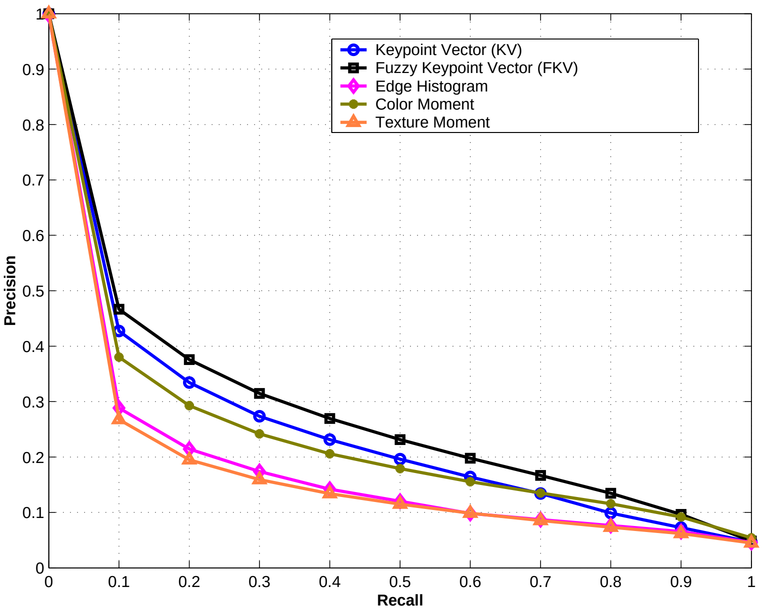


Fig. 4. PR-graphs of different feature spaces.

Fig. 4 shows the PR-curves of the keypoints-based image representation by performing the Euclidean distance measure in the “bag of keypoints”-based feature space (e.g., “KV”) and the proposed fuzzy keypoints-based feature space (e.g., “FKV’). The performances were also compared to three low-level color, texture, and edge related features to judge the actual improvement in performances of the proposed methods. The reason of choosing these three low-level feature descriptors is that they present different aspects of images. For color feature, the first (mean), second (standard deviation) and third (skewness) central moments of each color channel in the RGB color space are calculated to represent images as a 9-dimensional feature vector. The texture feature is extracted from the gray level co-occurrence matrix (GLCM). A GLCM is defined as a sample of the joint probability density of the gray levels of two pixels separated by a given displacement and angle [18]. We obtained four GLCM for four different orientations (horizontal 0°,vertical 90 °, and two diagonals 45 ° and 135 °). Higher order features, such as energy, maximum probability, entropy, contrast and inverse difference moment are measured based on each GLCM to form a 5-dimensional feature vector and finally obtained a 20-dimensional feature vector by concatenating the feature vector for each GLCM. Finally, to represent the shape feature, a histogram of edge direction is constructed. The edge information contained in the images is processed and generated by using the Canny edge detection (with $\sigma = 1$, Gaussian masks of size = 9, low threshold = 1, and high threshold = 255) algorithm [19]. The corresponding edge directions are quantized into 72 bins of 5° each. Scale invariance is achieved by normalizing this histograms with respect to the number of edge points in the image.

By analyzing the Fig. 4, we can observe that the performance of the keypoints-based feature representation (e.g., "KV") is better when compared to the global color, texture, and edge features in term of precision at each recall level. The better performances are expected as the keypoints-based feature representation is more localized in nature and invariant to viewpoint and illumination changes. In addition, we can observe that the fuzzy feature-based representation (e.g., "FKV') approach performed slightly better when compared to the similarity matching in the normalized keypoints-based feature space. Overall, the improved result justifies the soft annotation scheme by spreading each region's membership values to all the keypoints in the codebook. Hence, the proposed fuzzy feature representation scheme is not only invariant to affine transformations but also robust against the distribution of the quantized keypoints. For generation of the fuzzy feature, we consider the value of $m = 2$ of the fuzziness exponent due to its better performance in the ground truth dataset.

6 Conclusions

We have investigated the "bag of keypoints" based image retrieval approach in medical domain inspired by the ideas of the text retrieval. In this approach, interest points are detected and described by affine SIFT descriptor at first. Based on the construction of a SOM generated codebook, images are represented in a fuzzy feature space by spreading each region's membership values through a global fuzzy membership function to all the keypoints in the codebook. The proposed feature representation scheme is invariant to affine transformations, as well as occlusion, lighting and intra-class variations and robust against quantization errors. Experimental results in a medical image collection justified the validity of the proposed feature extraction and image representation approach.

Acknowledgment

This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). We would like to thank the CLEF [2] organizers for making the database available for the experiments.

References

1. H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A Review of Content-Based Image Retrieval Systems in Medical Applications Clinical Benefits and Fu-

ture Directions," *International Journal of Medical Informatics*, vol. 73, pp. 1–23, 2004.

2. H. Müller, T. Deselaers, E. Kim, C. Kalpathy, D. Jayashree, M. Thomas, P. Clough, and W. Hersh, "Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks", *8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007)*, Proceedings of LNCS, 5152, 2008.
3. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A Comparison of Affine Region Detectors", *International Journal of Computer Vision*, vol. 65, pp. 43–72, 2005.
4. K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector", *Proc. of European Conference on Computer Vision*, pp. 128–142, 2002.
5. A. Baumberg, "Reliable feature matching across widely separated views", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 774-781, 2000.
6. D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60 (2), pp. 91-110, 2004.
7. R. B. Yates, and B. R. Neto, *Modern Information Retrieval*, 1st ed., Addison Wesley, 1999.
8. S. Lazebnik, C. Schmid, and J. Ponce, "Sparse texture representation using affine-invariant neighborhoods", *Proc. International Conference on Computer Vision & Pattern Recognition*, pp. 319–324, 2003.
9. G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray, "Visual categorization with bags of keypoints," *Proc. Workshop on Statistical Learning in Computer Vision*, pp. 1-22, 2004.
10. T. Kohonen, *Self-Organizating Maps*, New York, Springer-Verlag, 1997.
11. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Academic Press, 1990.
12. J.C. Bezdek, and S.K. Pal, Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data, NY: IEEE Press, 1992.
13. S. Mitra, and S.K. Pal, Self-organizing neural network as a fuzzy classifier, IEEE Trans Syst Man Cybernet. 24 (3) (1994) 385-399.
14. C.C. Yang, N. K. Bose, Generating fuzzy membership function with self-organizing feature map, Pattern Recog Letters, 27 (5) (2006) 356–365.
15. J.C. Bezdek, M.R. Pal, J. Keller, and R. Krisnapuram, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers, Boston, 1999.
16. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.
17. J. Han, and K.K. Ma, Fuzzy Color Histogram and Its Use in Color Image Retrieval, IEEE Trans Image Process. 11 (8) (2002) 944–952.
18. R. M. Haralick, Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybernetics*, vol. 3, pp. 610-21, 1973.
19. J. Canny, "A computational approach to edge detection", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 679–698, 1986.