"Bag of Keypoints"-Based Biomedical Image Search with Affine Covariant Region Detection and Correlation-Enhanced Similarity Matching

Md Mahmudur Rahman, Sameer K. Antani, George R. Thoma
U.S. National Library of Medicine,
National Institutes of Health,
Bethesda, MD, USA

Abstract

This paper presents a "bag of keypoints" based biomedical image retrieval approach by detecting affine covariant regions. The covariant 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. To describe the intensity pattern within the interest points the Scale-Invariant Feature Transform (SIFT) is used. The SIFT features are then vector quantized to build a visual vocabulary of keypoints by utilizing the Self-Organizing Map (SOM)-based clustering. By mapping the interest points extracted from one image to the words in the visual vocabulary, their occurrences are counted and the resulting histogram is called the "bag of keypoints" for that image similar to the "bag of words" based representation of documents in text retrieval. To exploit the correlations between the keypoints in the collection, a global similarity matrix is constructed to be utilized in a distance measure function to compare the query and database images. A systematic evaluation of image retrieval on a biomedical image collection demonstrates the advantages of the proposed image representation and similarity matching approaches in terms of precision-recall.

1 Introduction

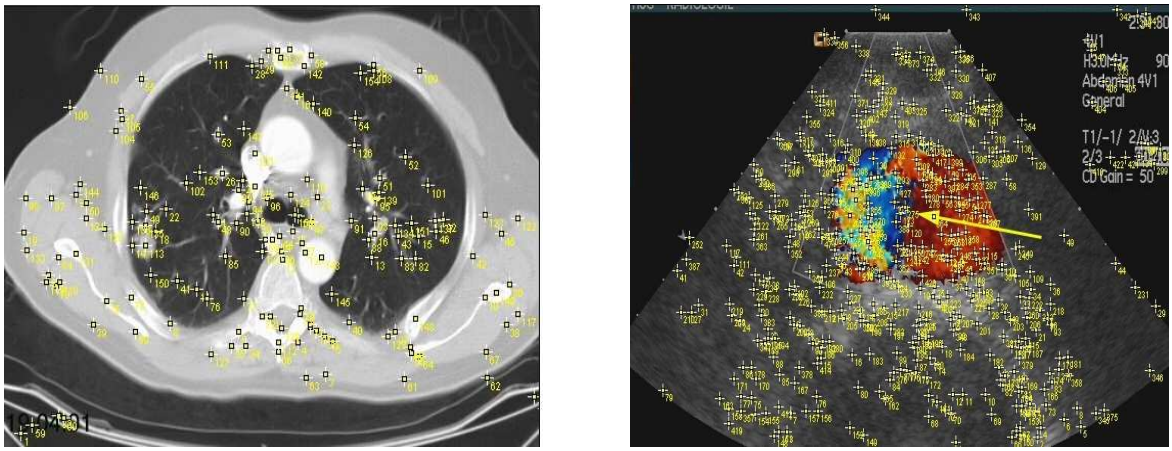
In recent years, the digital imaging revolution in the medical domain facilitate the generation and storage of large collections of images by hospitals and clinics every day. These images of various modalities constitute an important source of anatomical and functional information [1]. This exponential growth of the biomedical image data has created a compelling need for innovative tools for managing, retrieving, and visualizing images from large collections to support the clinical decision making, research, training and education.

In a heterogeneous medical collection with multiple modalities, such as ImageCLEFmed benchmarks [1], 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 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 [2]. 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) [3] 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 images are searchable in a similar manner with "bag of words" in text retrieval [4].

The idea of clustering invariant descriptors of image patches and represent images with "bag of keypoints" has already been applied to the problem of texture classification

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



(a) Chest CT Image (b) Doppler Ultra Sound Image

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

and recently for generic visual categorization with promising results [6, 7]. For example, the work described in [7] 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, similar to the text retrieval, in the *"bag of keypoints"* model each keypoints is considered independent of all the other keypoints besides the loss of all ordering structure. This independent assumption might not hold in many cases as in general there exists correlated keypoints in individual images as well as in a collection as a whole. For example, there is a higher probability of co-occurrence between the white teeth and red color tissue of the mouth in a dental photographic image whereas most of the time the teeth are surrounded by jaw bones and black background in dental X-ray images. In these examples, individual objects, such as teeth, mouth tissue, jaw bones and black background can be considered as the local region or interest points with their distinct color and texture properties. So, there is indeed a need to exploit the correlation or similarity patterns among the keypoints to improve the retrieval effectiveness.

To overcome the above limitation, this paper presents a correlation-enhanced "bag of keypoints" based biomedical image retrieval approach. In this approach, the SIFT features are extracted at first from the interest points and then vector quantized by the Self-Organizing Map (SOM)-based clustering to build a visual vocabulary of keypoints. By mapping the interest points extracted from one image to the words in the visual vocabulary, their occurrences are counted and the resulting histogram is called the "bag-of-keypoints" for that image. The similarities/correlations between the keypoints are analyzed in the collection as a whole to construct a global similarity thesaurus that is finally utilized in a distance measure function to compare query and target images in a database. The organization of the paper is as follows: In Section 2, the "Bag of Keypoints"-based image representation approach is discussed. Section 3 presents the correlation-enhanced similarity matching approach based on the generation of a global matrix. Exhaustive experiments and analysis of the results are presented in Sections 4 and 5. Finally, Section 6 provides our conclusions.

2 "Bag of Keypoints" based Image Representation

A major component of this retrieval framework is the image representation as the "bag of Keypoints" similar to the "bag of words" based representation of documents in text retrieval. The main steps for the feature representation scheme are:

- Detection and description of covariant regions or interest points.
- Generation of a codebook (a vocabulary) by applying vector quantization or clustering based on the region descriptors in the training set.
- Assigning the region descriptors to a set of predetermined clusters or codewords of the codebook.
- Constructing a bag of keypoints of individual images, which counts the number regions assigned to each

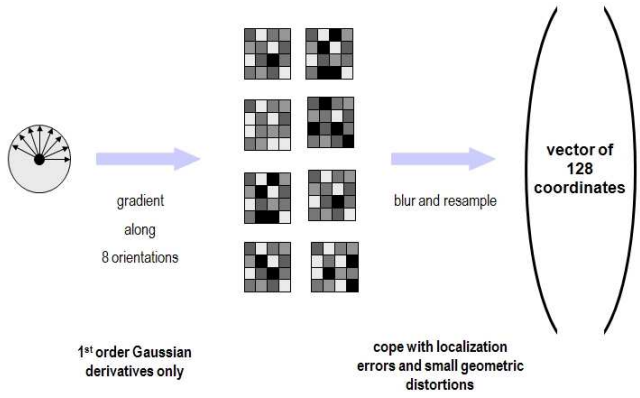


Figure 2. SIFT descriptor generation process

cluster and treating the bag of keypoints as the feature vector.

Each steps are described briefly in following:

2.1 Detection and Description of Interest Points

First, a set of covariant regions or interest points is detected in an image. Often a large number, perhaps hundreds or thousands, of possibly overlapping regions are obtained. Interest points are those points in the image that possess a great amount of information in terms of local signal changes [2]. In this study, the Harris-affine detector is used as interest point detection methods [5]. Harris affine points are detected by an iterative process. Firstly, positions and scales of interest points are determined as local maxima (in position) of a scale adapted Harris function, and as local extrema in scale of the Laplacian operator. Then an elliptical (i.e. affine) neighborhood is determined. This has a size given by the selected scale and a shape given by the eigenvalues of the images second moment matrix. The selection of position/scale and the elliptical neighborhood estimation are then iterated and the point is kept only if the process converges within a fixed number of iterations. The affine region is then mapped to a circular region, so normalizing it for affine transformations. 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 [3] 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 [2] several affine region detectors have been compared for matching and it was found that the SIFT descriptors perform best. Fig. 2 shows an example of the maps of gradient magnitude corresponding to the 8 orientations and vector generation process.

2.2 Codebook Generation by SOM

A codebook $C = \{c_1, \cdots, c_j, \cdots, c_N\}$ is a set of codewords (visual words) where each c_j is associated a vector $\mathbf{c}_j = [c_{j_1} \ \cdots c_{j_2} \cdots c_{j_d}]^{\mathrm{T}}$ of dimension d in an Euclidean space. The codebook or visual vocabulary is generated by clustering the interest points detected in a subset of images, and each discovered cluster represents a codeword of the codebook. In this work, the clustering is performed by applying the SOM [8], a competitive learning-based algorithm that maps the high dimensional input vectors to a low-dimensional regular lattice or grid of map units. The basic structure of the SOM consists of two layers: an input layer and a competitive output layer as a map. The input layer consists of a set of input node vectors and the output map consists of a set of N units organized into either a one-or two-dimensional lattice structure where each unit m_j is associated with a weight vector $\mathbf{w}_j \in \Re^d$.

The first step of the learning process is to initialize the weight vectors of the output map. Then, for each input SIFT vector $\mathbf{x}_i \in \Re^d$, the distances (e.g., Euclidean) between the $\mathbf{x}_i$ and weight vectors of all map units are calculated. The unit that has the smallest distance is called the *best-matching unit* (BMU) or the *winning node*. The next step is to update the weight vectors associated with the BMU, m_c as

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)\theta_{cj}(t)(\mathbf{x}_i(t) - \mathbf{w}_j(t)) \quad (1)$$

Here, t is the current iteration, $\mathbf{w}_j(t)$ and $\mathbf{x}_i(t)$ are the weight vector and the target input vector respectively at the iteration t, and $\theta(t)$ and $\alpha(t)$ are the smooth neighborhood function and the time-dependent learning rate. Due to the process of self-organization, the initially chosen $\mathbf{w}_j$ gradually attains new values such that the output space acquires appropriate topological ordering. After the learning phase, the map can be used as a codebook where a weight vector $\mathbf{w}_j$ of unit m_j resembles a codeword vector vector $\mathbf{c}_j$ of the codebook C. We refer to the codewords (e.g., map units) as "keypoints" by analogy with "keywords" in text retrieval.

2.3 Image encoding and representation

The codebook described above is effectively utilized as a representation scheme. In this approach, an image is characterized by a histogram of keypoints, i.e., by counting the number of interest points that belong to every codewords

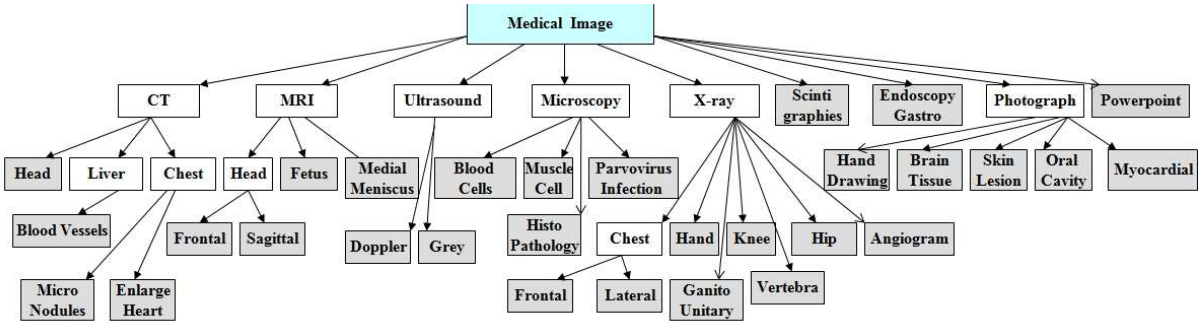


Figure 3. Classification structure of the medical image data set.

formed in the previous step. For each SIFT vector of interest point in an image, the codebook is searched to find the best match codeword (e.g., BMU in the map) based on a distance measure. Based on the encoding scheme, an image I_j can be represented as a vector of codewords (keypoints) as

$$\mathbf{f}_j = [\hat{f}_{1j} \cdots \hat{f}_{ij} \cdots \hat{f}_{Nj}]^{\mathrm{T}} \tag{2}$$

where each element $\hat{f}_{ij}$ represents the normalized frequency of occurrences of the keypoints c_i appearing in I_j. The resulting vector is the "bag of keypoints" representation for the image, which is going to be used for retrieval purposes.

3 Correlation-Enhanced Similarity Matching

This section presents a similarity matching approach by considering the correlations/similarities between the keypoints in the collection. For the correlation analysis, we construct a similarity matrix where each element of the matrix defines the keypoint similarities. To measure the similarities, we rely on how the keypoints in the collection are indexed by images, i.e., for each keypoint there is an image vector space. This idea of measuring this similarity was originally proposed in [9] for the query expansion in text retrieval. In this approach, each keypoint $c_i, i \in \{1, \cdots, N\}$ is associated with a vector $\mathbf{c}_i =< w_{i1}, \cdots, w_{ij}, \cdots, w_{iM} >$ where M is the number of images in the collection. The element w_{ij} is the weight for the keypoint c_i in image I_j, which is computed in a rather distinct form as [9]:

$$w_{ij} = \frac{(\frac{f_{ij}}{max_j\ (f_{ij})})\ ikf_j}{\sqrt{\sum_{l=1}^{M}(\frac{f_{il}}{max_l\ (f_{il})})^2\ ikf_l^2}} \tag{3}$$

where f_{ij} be the frequency of occurrence of the concept c_i in the image I_j and $max_j\ (f_{ij})$ computes the maximum frequency of c_i under all images in the collection. Further, the inverse concept frequency ikf_j for I_j, (e.g., analogous to the inverse image (document) frequency), is computed as $ikf_j = log\frac{L}{k_j}$, where k_j be the number of distinct keypoints in the I_j.

After generating the vectors, a similarity matrix $\mathbf{S}_{N\times N} = [s_{u,v}]$ is built through the computation of each element $s_{u,v}$ as the normalized cosine relationship or dot product between two vectors $\mathbf{c}_u$ and $\mathbf{c}_v$ as

$$s_{u,v} = \mathbf{c}_u \cdot \mathbf{c}_v = \sum_{j=1}^{M} w_{uj} * w_{vj} \tag{4}$$

Although, the construction of the matrix $\mathbf{S}$ is prohibitively difficult for large collections. Many collections are available now-a-days, with several hundred thousand images. However, the matrix needs to be computed only once and can be computed off-line. The only component done on a per query basis is the utilizing the matrix elements in the distance matching function.

Finally, the global matrix is utilized in a quadratic form of distance measure to compare a query and database images as

$$Dis_{\mathrm{S}}(I_q, I_j) = \sqrt{(\mathbf{f}_q - \mathbf{f}_j)^T \mathbf{S} (\mathbf{f}_q - \mathbf{f}_j)} \tag{5}$$

Here, $\mathbf{f}_q$ and $\mathbf{f}_j$ are the feature vector for the query image I_q and a target image I_j respectively.

4 Experiments

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 [10]. In this collection, images are classified into three levels as shown in Figure 3. In the first level, images are categorized according to the imaging modalities (e.g., X-ray, CT, MRI, etc.). At the next level, images at each of the modalities is classified according to

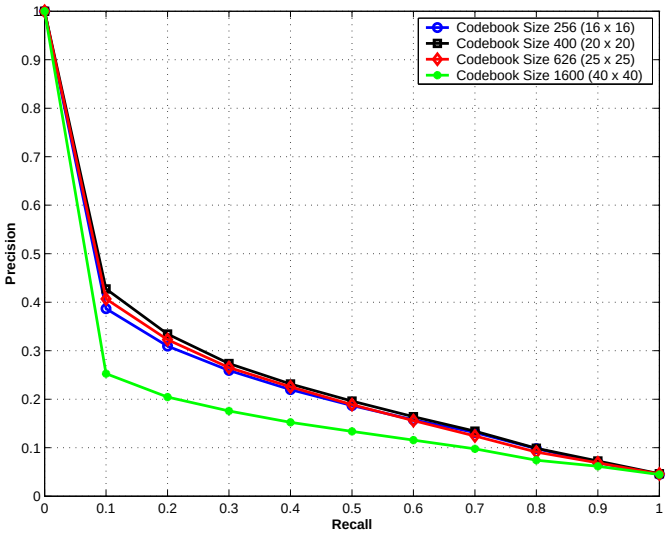


Figure 4. PR-graphs of different codebook sizes.

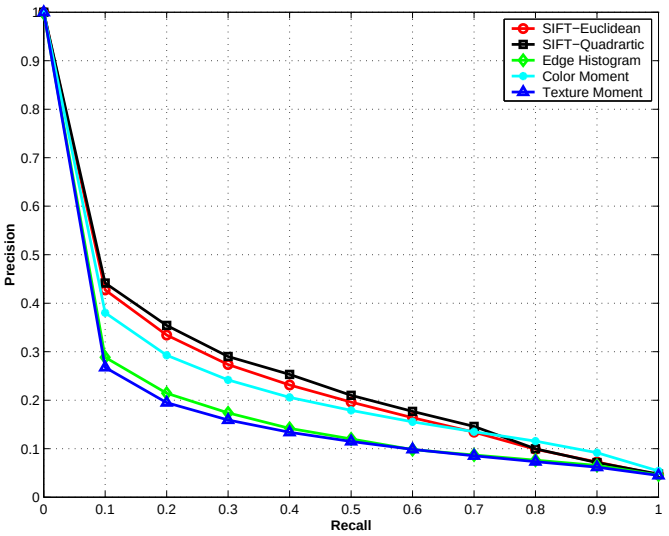


Figure 5. PR-graphs of different feature spaces.

the examined body parts (e.g., head, chest, etc.) and finally images are 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. 3. 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 [4]. The average precision and recall are calculated over all the queries to generate the precision-recall (PR) curves in different settings.

5 Results

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 (20×20), 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. 4 shows the PR-curves on four different codebook sizes. It is clear from Fig. 4 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. 5 shows the PR-curves of the keypoints-based image representation by performing the Euclidean (e.g., "SIFT-Euclidean") and proposed correlation-enhanced similarity matching (e.g., "SIFT-Quadratic"). 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 medical 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 [11]. 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 [12]. 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. 5, we can observe that the performance of the proposed keypoints-based feature representation is much 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 features are more localized in nature and invariant to viewpoint and illumination changes. From Fig. 5, we can also observe that, the correlation enhanced similarity matching approach performed slightly better when compared to the Euclidean similarity matching. Overall, the improved result indicate that the correlations among the keypoints are not negligible and can be exploited effectively in the similarity matching function.

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 as a vector of keypoints. The proposed technique also exploit the similarities/correlations between the keypoints based on the generation of a global matrix and utilized in a similarity matching function. Experimental results justified the initial assumption of representing medical images with affine region descriptors and validated the exploitation of correlations between keypoints to improve retrieval performance.

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 [10] 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 Future Directions," *International Journal of Medical Informatics*, vol. 73, pp. 1–23, 2004.

[2] 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.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60 (2), pp. 91-110, 2004.

[4] R. B. Yates, and B. R. Neto, *Modern Information Retrieval*, 1st ed., Addison Wesley, 1999.

[5] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector", *Proc. of European Conference on Computer Vision*, pp. 128–142, 2002.

[6] 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.

[7] 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.

[8] T. Kohonen, *Self-Organizating Maps*, New York, Springer-Verlag, 1997.

[9] Y. Qiu, and H. P. Frei, Concept Based Query Expansion. *Proc. 16th Int. ACM SIGIR Conf. on R&D in Info. Retrieval*, SIGIR Forum, ACM Press, June, 1993.

[10] 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)*, Proc. of LNCS, 5152, 2008.

[11] R. M. Haralick, Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybernetics*, vol. 3, pp. 610–621, 1973.

[12] J. Canny, "A computational approach to edge detection", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 679–698, 1986.