A Pattern Similarity Scheme for Medical Image Retrieval

Dimitris K. Iakovidis, Member, IEEE, Nikos Pelekis, Haralampos Karanikas, Evangelos Kotsifakos, Ioannis Kopanakis, and Yannis Theodoridis

Abstract— In this paper we propose a novel scheme for efficient content-based medical image retrieval, formalized according to the PANDA (PAttterns for Next generation Database systems) framework. The proposed scheme involves low-level feature extraction from image regions followed by clustering of the feature space to form higher-level patterns. The components of each pattern include a cluster representation and a measure that quantifies the quality of the image content representation achieved by the pattern. The similarity between two patterns is estimated as a function of the similarity between both the structure and the measure components of the patterns. Experiments were performed on a reference set of radiographic images, using standard wavelet domain image features. The results show that the proposed scheme can be more efficient than the common ground schemes for medical image retrieval, as it does not involve exhaustive, nearest neighbor searching over the whole set of the available feature vectors. Keeping the patterns in a unified form facilitates further processing and analysis by mining or visualization algorithms.

I. INTRODUCTION

One of the primary tools used by physicians is the comparison of previous and current medical images associated with pathologic conditions. As the amount of pictorial information stored in both local and public medical databases is steadily growing, efficient image indexing and retrieval is increasingly becoming a necessity.

In the last decade the advances in information technology allowed the development of Content-Based Image Retrieval (CBIR) systems, capable of retrieving images based on their similarity with one or more query images. Some of these systems are QBIC [1], VisualSEEK [2], Virage [3], Netra [4], PicSOM [5], SIMPLicity [6], CIRES [7], and FIRE [8]. More than fifty CBIR systems are surveyed in [9].

The benefits emanating from the application of content-based approaches to medical image retrieval range from clinical decision support to medical education and research [10]. These benefits motivated researchers either to apply general purpose CBIR systems to medical images [8] or to develop new ones explicitly oriented to specific medical domains. Specialized CBIR systems have been developed to support the retrieval of various kinds of medical images, including High Resolution Computed Tomographic (HRCT) images [11], breast cancer biopsy slides [12], Positron Emission Tomographic (PET) functional images [13], ultrasound images [14], endoscopic images [15], pathology images [16], spine radiographs [17], and mammographic images [18].

Common ground for most of the systems cited above is that image retrieval is based on similarity measures estimated directly from low-level image features, whereas it involves multidimensional, usually exhaustive, nearest neighbor searching over the whole set of the available feature vectors. However, such an approach can be time consuming with large image databases.

Research on improving the efficiency of the image retrieval process has mainly focused on image indexing techniques by utilizing data structures, such as R-trees [19], feature index trees [20], iconic index trees [21], and meshes of trees [22]. Other approaches to improving efficiency, include clustering of the image feature spaces [23][24], and utilization of alternative similarity measures, usually dependent on feature sets [25][26].

Motivated by these studies, we propose a novel scheme for efficient content-based medical image retrieval that utilizes similarity measures defined over higher-level patterns associated with clusters of low-level image feature spaces. The term pattern is defined in the context of a state of the art framework named PANDA (PAAttterns for Next generation Database systems) [27].

The rest of this paper comprises four sections. Section 2 outlines the PANDA framework. The proposed pattern similarity scheme is described in Section 3, and the results obtained from an indicative application for the retrieval of radiographs are apposed in Section 4. Finally, Section 5 summarizes the conclusions and the future perspectives of this study.

II. THE PANDA FRAMEWORK

The efficient management of patterns extracted from medical image databases is of vital importance due to the
extremely big storage requirements as well as the complexity of such kind of raw data. Taking advantage of the PANDA framework [27] we adopt the idea of a Pattern Based Management System (PBMS) as the infrastructure for managing patterns extracted from our CBIR scheme, in contrast to DataBase Management Systems (DBMS). The key concept of PANDA is that any type of pattern can be represented in a compact and unified way. This can be achieved through a Pattern-Base (PB) keeping information about extracted patterns. Such a PB introduced in [27] consists of three basic layers: the pattern, the pattern type and the class. A pattern type is a description of the pattern structure. A pattern is an instance of the corresponding pattern type and class is a collection of semantically related patterns of the same pattern type.

Formally, a pattern type \( pt \) is a quintuple \( pt = \langle n, ss, ds, ms, f, \rangle \), where \( n \) is the name of the pattern type. The structure schema \( ss \) defines the pattern space by describing the structure of the pattern type, while the source schema \( ds \) defines the data source space by describing the dataset wherefrom patterns of this pattern type are derived. The measure schema \( ms \), quantifies the quality of the source data representation achieved by patterns of this pattern type and the formula \( f \) describes the relationship between the source space and the pattern space. In this notation, if \( pt \) is a pattern type then \( p = \langle \text{pid}, s, d, m, e \rangle \) is an instance of \( pt \), where \( \text{pid} \) is a unique pattern identifier, \( s, d, m \) are the corresponding structure, source and measure of the pattern, while \( e \) is an expression indicating the section of the source space related to pattern \( p \).

Aside from the physical storage of the patterns we appropriately utilize the PANDA framework for the comparison of patterns [27] as the base of our image retrieval methodology in our CBIR scheme. In PANDA the similarity \( sim \), \( sim \in [0,1] \) between two patterns \( p_1, p_2 \) of the same type can be computed by combining, by means of an aggregation function \( f_{agg} \), the similarity between both the structure \( s \) and the measure \( m \) components:

\[
sim(p_1, p_2) = f_{agg}(\text{sim}_{struct}(p_1.s, p_2.s), \text{sim}_{meas}(p_1.m, p_2.m)) \tag{1}
\]

Efficient definition of the structure and measure of patterns extracted from medical images, as well as appropriate selection of an aggregation logic and distance functions to assess the respective similarities, is one of the challenges adopted in this paper.

III. MEDICAL IMAGE RETRIEVAL USING PATTERNS

The proposed content-based medical image retrieval scheme is outlined in Fig. 1. It involves four steps: a) low-level feature extraction from each of the stored and the query images, b) clustering of the extracted feature vectors per image, c) pattern instantiation of the clusters, and d) computation of pattern similarities. The registration of a new image in the database involves the first three of the four steps described for image retrieval (a, b, and c).

![Fig. 1. Outline of the content-based medical image retrieval methodology that embodies the proposed pattern similarity scheme.](image)

A. Low-Level Image Feature Extraction

Each of the images stored in the database, as well as the query image is raster scanned with a sliding window of user-defined size and sliding step. The sliding step may allow windows to overlap between each other. For each window \( N \) features \( f_i, i=1,2, N \) are calculated to form a single feature vector \( F \). The number of feature vectors produced for each image depends on the size, the dimensions and the step of the sliding window.

Typically, the features characterizing the image content should be selected based on the details associated with the image collection and the retrieval task [28]. Color, texture and shape are the three major classes of image features used in image retrieval [1][9][28]. However, this paper focuses on the utility of the proposed pattern similarity scheme rather than on the selection of an optimal feature set for a particular image retrieval task.

B. Clustering

For the clustering of the extracted image features we used the Expectation Minimization (EM) algorithm [37]. The EM algorithm is a widely-used statistical clustering method. It performs clustering by estimating the mean and standard deviation of each cluster, so as to maximize the likelihood of the observed data.

C. Pattern Instantiation

Given a clustering of an image comprising \( M \) clusters \( C_i \), \( i=1,2,...,M \), according to the PANDA formalization and with respect to the output of the EM algorithm a pattern \( \text{specimen} \) is instantiated for each cluster \( C_i \) representing a physical anatomic specimen depicted in a medical image:

\[
\text{specimen} = \langle \text{SS} : \langle \langle \text{Mean} : \text{Real}, \text{StdDev} : \text{Real} \rangle \rangle, \langle \text{MS} : \langle \langle \text{GG} : \text{Real} \rangle \rangle \rangle \tag{2}
\]

where \( \text{mean} \) and \( \text{stdDev} \) are the mean and the standard deviation of the distribution \( D_i \) for every one of the \( N \)
features \((j = 1, 2, \ldots, N)\) in cluster \(C_j\), respectively, and \(pp\) is the prior probability of \(C\). Here prior probability is defined as the fraction of the feature vectors of the image that belong to cluster \(C_j\). Intuitively, prior probability \(pp\) is equivalent with the support measure widely used in data mining models. In our case, in addition to the qualitative aspect of the prior probability, it also provides an indication of the size of the specimen.

In this connection, a medical image is considered as a complex pattern defined by (3), consisting of a set of simple clusters each one of them represented by the mean and standard deviation values of a distribution.

\[
\text{medical_image} = \left\{ SS : \{\text{Specimen}\} \right\}
\]

\[
D. \text{ Computation of Pattern Similarities}
\]

Aiming at the definition of the similarity of two medical images (i.e., complex patterns), we have first to define the measures’ and the structural similarity between two clusters \(C_1\) and \(C_2\) (i.e., simple patterns). The measures’ similarity between two clusters is computed using the Euclidean distance as in Eq.(4)

\[
sim_{\text{meas}}(C_1, C_2) = 1 - \left[ |C_1\cdot pp - C_2\cdot pp| \right]
\]

Rephrasing the problem of defining the structural similarity between \(C_1\) and \(C_2\) we need to find a measure for evaluating the closeness of two sets of distributions, as \(C_1\) and \(C_2\) are. Further decomposing the problem, we should first define a method of computing the similarity between just two distributions \(D_1\) and \(D_2\). To achieve this, we extend the methodology proposed in [38] where two distributions can be compared by discovering the expected percentage the first distribution overlaps the other. This is achieved by consulting a nomogram and more specifically by interpolating between contours of the nomogram to estimate the expected percentage. The previously mentioned nomogram is constructed by taking as \(x\) axis a Standardized Absolute Distance between the means of the distributions (SADM, Eq. 5) whereas the \(y\) axis is computed as the Ratio of the Standard Deviations of the distributions (RSD, Eq. 6).

\[
\text{SADM} = \frac{|D_{\text{mean}} - D_{\text{mean}}|}{D_{\text{stdDev}}}
\]

\[
\text{RSD} = \frac{D_1\cdot \text{stdDev}}{D_2\cdot \text{stdDev}} \text{ where } D_1\cdot \text{stdDev} < D_2\cdot \text{stdDev}
\]

Given the above, we define the similarity between distributions \(D_1\) and \(D_2\), by Eq.(7).

\[
sim(D_1, D_2) = \begin{cases} 
\frac{|D_{\text{mean}} - D_{\text{mean}}|}{D_{\text{stdDev}}^2}, & \text{if } |D_{\text{mean}} - D_{\text{mean}}| < D_2\cdot \text{stdDev}^2 \\
0, & \text{otherwise}
\end{cases}
\]

This is a means to automate and materialize the intuitive overlap between two distributions. In other words, the similarity is defined by finding the proximity of the mean values, normalized by the standard deviation of the first and by weighting with a factor indicating how bigger the standard deviation of the second in contrast to the first is. Having this, we let the structural similarity between two sets of distributions (i.e. two clusters \(C_1\) and \(C_2\)) be the average among the similarities computed for each pair of the \(N\) features:

\[
sim_{\text{struct}}(C_1, C_2) = \frac{1}{N} \sum_{j=1}^{N} \sim_{\text{meas}}(D_j^1, D_j^2)
\]

We aggregate the similarities between the qualitative \(sim_{\text{meas}}\) and the structural \(sim_{\text{struct}}\) similarities between the clusters by using the following aggregation function \(\text{f}_{\text{agg}}\), which gives the same weight to either of the above similarities, while further weights the overall similarity by the prior probabilities of the clusters, as a bias towards similar and concurrently big clusters.

\[
sim(C_1, C_2) = \frac{sim_{\text{meas}}(C_1, C_2) + sim_{\text{meas}}(C_1, C_2)}{2}
\]

Having defined the similarity between clusters (i.e., simple patterns), to compare two medical images \(M_1\) and \(M_2\) (i.e., complex patterns) we need to determine the coupling methodology between the different clusters of each image. Though various coupling types can be applied in the context of the PANDA framework [27], we adopt the matching of Eq.(10), allowing each cluster of the first image to match more than one cluster of the second, and vice versa.

\[
sim(M_1, M_2) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \sim_{\text{meas}}(C_i^{\text{pp}}, C_j^{\text{pp}})
\]

IV. RESULTS

Experiments were performed with radiographic images from the IRMA (Image Retrieval in Medical Applications) dataset [29]. This is a growing collection of radiographic images acquired in RWTH Aachen University of Technology Hospital, Germany. It is used as reference for medical image retrieval tasks. It currently contains 10,000 arbitrarily selected anonymous radiographic images for which the ground truth information is provided. The radiographs span 116 categories and depict various anatomic specimens, including cranium, spine, arm, chest, abdomen, leg, pelvis, breast, and hands of patients of various ages, genders, and pathologies. All radiographic images are in 8-bit greyscale format and have been downscaled to fit into a 512x512 bounding box maintaining the original aspect ratio.

The aim of the experiments was to demonstrate the efficiency of the proposed pattern similarity scheme for CBIR over the conventional scheme used by CBIR systems. To evaluate its performance a subset of 90% of the images was registered in the database, whereas a subset of 10% of the images was used for querying the pattern-base. Each image was sampled using overlapping windows of 32x32-pixel dimensions with an 8-pixel sliding step, resulting in a total of 625 samples.

From each sample we considered the extraction of a standard, yet effective set of features comprising 3-level
Discrete Wavelet Transform (DWT) energy features [10][31]. The dimension of each feature vector is $N=10$.

The feature vectors extracted from each image were clustered by means of the EM algorithm implemented in the WEKA data mining tool [36].

In order to determine the optimal number of distinctively discriminated clusters per image, we followed a visualization approach to representation and understanding of the spatial relationships between the categories of the image content. This was achieved by means of geometric projection techniques [31] and the three-dimensional class-preserving projection algorithm [32]. During the projection procedures, class-preserving projection techniques preserve the properties of the clustered data in the $R^3$ space also to the projection plane in order to construct corresponding representations from which accurate inferences could be extracted. The results from the application of this visualization method on a representative image from each anatomic specimen showed that the optimal number of distinctive clusters per image is $M=4$. An example three-dimensional projection of the clustered feature space of an image is illustrated in Fig. 2.

For each cluster a pattern $\text{specimen}_i, i=1,2,3,4$ (Eq.2) was assigned, whereas each image was represented by a complex pattern $\text{medical_image}$ (Eq.3). The collection of patterns originating from the images registered in the database was used to build the pattern-base. Subsequent queries were executed to evaluate the performance of the proposed scheme.

An indicative example retrieval of five radiographic images, based on the content of a query image is illustrated in Fig. 3. It can be noticed that the retrieved images contain semantically relevant regions (spine) although they belong to different categories. This could be justified if one considers that the proposed CBIR scheme involves features extracted locally from the images. Similar observations are valid for queries performed using radiographic images from other categories, and indicate that the patterns used for image representation carry substantial semantic information.

![Figure 2](image2.png)

**Fig. 2.** Three-dimensional projection of the feature space of a radiographic image clustered into four categories by the EM algorithm.

For each cluster a pattern $\text{specimen}_i, i=1,2,3,4$ (Eq.2) was assigned, whereas each image was represented by a complex pattern $\text{medical_image}$ (Eq.3). The collection of patterns originating from the images registered in the database was used to build the pattern-base. Subsequent queries were executed to evaluate the performance of the proposed scheme.

An indicative example retrieval of five radiographic images, based on the content of a query image is illustrated in Fig. 3. It can be noticed that the retrieved images contain semantically relevant regions (spine) although they belong to different categories. This could be justified if one considers that the proposed CBIR scheme involves features extracted locally from the images. Similar observations are valid for queries performed using radiographic images from other categories, and indicate that the patterns used for image representation carry substantial semantic information.

![Figure 3](image3.png)

**Fig. 3.** Indicative example retrieval of radiographic images: (a) Query image (category: abdomen, uroperitoneal system), (b)-(d) Correct retrievals, (e)-(f) False retrievals (category: abdomen, gastrointestinal system).

Figure 4 illustrates the performance of the proposed pattern similarity scheme that involves pattern comparisons, as compared with the performance of the conventional scheme that involves exhaustive comparisons of the feature vectors. The performance is measured in terms of the logarithm of the number of comparisons between the query and the registered data. The results show that the proposed scheme achieves approx. 2,400 less comparisons.

![Figure 4](image4.png)

**Fig. 4.** Number of comparisons between the query and the registered data for the conventional and the proposed scheme.

**V. CONCLUSIONS**

We presented a novel scheme for efficient content-based medical image retrieval. This scheme utilizes patterns, defined in the context of the state of the art framework PANDA. Image retrieval is based on similarities defined over complex patterns. Searching for similar images involves comparison of patterns instead of exhaustive, nearest neighbor searching over the whole set of the feature vectors stored in the image database. The results advocate to the efficiency of the proposed scheme for image retrieval from large databases over the commonly employed methods.
The retrieval performance of a content-based medical image retrieval system implementing the proposed scheme is highly associated with the feature set used for data representation. Selection of representative features, can lead to separable clusters, and provided that the clustering method chosen is compatible with the geometry of the feature space used, it is possible to minimize the performance divergence of the proposed and the conventional schemes.

By storing clustering patterns along with the low-level features set in a unified format we facilitate further processing and analysis. To date, most data mining algorithms have concentrated on the extraction of interesting rules directly from low-level data [33]. Our approach provides the means for deriving rules from the results of other data mining algorithms that is, mining from rules set. In the current work our initial low-level feature set is further processed and represented via clustering by higher-level patterns which are in a machine processible format. A significant advantage of this approach is that the nature of the rules to be extracted by this process contains different higher order semantics.

Future perspectives of this work include: a) the systematic evaluation of the proposed scheme for the retrieval of various medical images, such as endoscopic [34] and ultrasound images [35] according to their pathology, b) the enhancement of the retrieval performance by using image indexing techniques based on specialized data structures [19]-[22], c) the integration of the proposed scheme with ontology-based information extraction and data mining techniques for the retrieval of medical images using heterogeneous data sources.

ACKNOWLEDGMENT

The collection of images used in this study is courtesy of T.M. Lehmann, Image Retrieval in Medical Application (IRMA) group, Dept. of Medical Informatics, RWTH Aachen, Germany, http://irma-project.org.

REFERENCES

Patterns”, in Proc. 8th European Conference on Principles and Practice of Knowledge Discovery in Database, pp. 496-499, 2004.


