Content-Based Image Retrieval Methodology for radio graphic images

G. Jalalu, K. Murali

1Department of EIE, V.R. Siddhartha Engineering College, Vijayawada.
2Department of ECE, Narayana Engineering College, Nellore

Abstract- We introduce a new method for efficient related to Content based medical image retrieval, it’s quite similar to the PANDA (Patterns for Next Generation Data based Systems) framework. This involves a low-level feature extraction from the salient regions of the images followed by clustering of the feature spaces to form high-level patterns. These components related are each similar to patterns consisting of a cluster representation. This measures quantifies the quality of image content representation achieved by the pattern. The similarity between two patterns is estimated as a function of similarity between both the structure and the measure components of the patterns. Experiments are conducted on a reliable set of radio graphic images, using the typical wavelet domain image features. The results show that the proposed scheme can be more efficient then the common ground screens for medical image retrieval, this has no concern with exhaustive nearest neighbor searching over the whole set of the available feature vectors. Keeping the patterns in unified form facilities further processing and analysis by mining or visualization algorithms.

Keywords- Radio graphic images, PANDA, Standard wavelet domain image features.

Introduction

The basic tools most commonly used by physicians are the comparison of previous and existing medical images related with pathologic conditions. Due to a rapid growth of pictorial information stored in both local and public medical databases, there is a requirement for image indexing and retrieval. During 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. The benefits emanating from the application of content-based approaches to medical image retrieval range from clinical decision support to medical education and research.

Proven benefits motivated researchers either to apply general purpose CBIR systems to medical images or to develop dedicated 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, breast cancer biopsy slides, positron emission tomographic (PET) functional images, ultrasound images, pathology images, and radiographic images. In this method we are likely to result in the retrieval of images with important perceived differences from the query image, since low-level features usually lack semantic interpretation.

We propose an unsupervised approach for efficient content-based medical image retrieval that utilizes similarity measures, defined over higher-level patterns that are associated with clusters of low-level image feature spaces. The term pattern is considered in the context of a state-of-the-art framework called patterns for Next generation database systems (PANDA) developed for the representation and the management of data mining results.

The proposed approach combines the advantages of the clustering-based CBIR methodologies with a semantically rich representation of medical images. Moreover, unlike related CBIR approaches that exploit multidimensional indexing techniques, such as R-trees, iconic index trees, and meshes of trees, the efficiency of the proposed approach is hardly affected by increasing the dimensionality of the low-level feature representation. The major contributions of this project are the following.

1) A novel representation of medical images treated as rich-in-semantics complex patterns. Each complex pattern comprises a set of simple patterns representing clusters of image regions associated with anatomic specimens in an unsupervised way. This pattern consists of clusters involving a combination of structural descriptors and quality measures.

2) Introducing a assessment for determining the similarity between complex patterns (i.e., medical images) for meeting CBIR purposes.

3) An a set of comprehensive experiments are conducted over a publicly available set of radiographic images. This is done in order to thoroughly evaluate our approach and also to demonstrate its effectiveness and efficiency.

Proposed Methodology:

I. Medical image retrieval using patterns:

![Figure 1: Proposed Content-Based Image Retrieval Methodology](image-url)
The proposed content-based medical image retrieval scheme is outlined in Figure 1. It involves four steps:

- Low-level feature extraction from each of the registered and query images
- Clustering of the extracted feature vectors per image
- Pattern instantiation of the resulted clusters
- Computation of pattern similarities.

The registration of a new image into the database involves steps 1, 2&3 whereas step 4 is processed during the retrieval task.

A. Low-Level Image Feature Extraction:
   Each of the images registered in the database, as well as the query image are raster scanned with a sliding window of user defined size, and then sampled image blocks at a given sampling step. The sampling step may allow consecutive blocks to overlap. For each block, a set of N features \( f_i \), \( i = 1 \ldots N \) is calculated to form a single feature vector \( F \). Typically, the sampling parameters and the features characterizing the low level image content are selected based on the details associated with the image collection and the retrieval task. Color, texture, and shape are the three major classes of image features commonly used in CBIR.

In the case of radiographic medical image retrieval, local grey level intensity and texture features have proved to discriminate best the depicted specimens. Standard multiscale statistical approach is used for the representation of the radiographic image regions that preserves local features, and does not depend on spatial coordinates. It is based on the 2-D discrete wavelet transform (2D-DWT).

B. Clustering:
   The low-level feature vectors are clustered using mixture models that model the data by a number of Gaussian distributions. A cluster corresponds to a set of distributions, one for each dimension of the dataset. Each distribution is described in terms of mean and standard deviation. A probabilistic approach to assigning feature vectors to clusters is used. For 1-D datasets, a mixture is a set of \( c \) Gaussian probability distributions, representing \( c \) clusters. The parameters of a mixture model are determined by the expectation maximization (EM). With \( c \) Gaussians, the probability density function of a variable \( X \) is

\[
f(X \mid \theta) = \sum_{i=1}^{c} PP_i \frac{1}{\sqrt{(2\pi)^d \left| \Sigma_i \right|}} e^{-\frac{1}{2} (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)}
\]

---- eq. 1

The algorithm results in a set of distributions, a vector of pairs of means \( \mu \) and standard deviations \( \sigma \), each of which corresponds to a feature, and outputs the size of the cluster. The vector of means \( \mu \) of the distributions for every feature represents the centroid of the cluster. Combining EM with the \( v \)-fold cross-validation algorithm, the number of clusters in the output of the algorithm can automatically be determined. The \( v \)-fold cross-validation technique works by partitioning the data into \( v \) equally sized segments. Most importantly, the distributions representing the clusters at the output of the EM algorithm can be easily utilized for pattern instantiation by the PANDA framework.

C. Pattern Instantiation:
   The clusters resulting from the EM algorithm are considered as patterns extracted from the image database, and are represented and handled according to the PANDA formalization. Hence, given a clustered image comprising of \( M \) simple patterns \( P_i \), \( i = 1 \ldots M \). A Specimen \( i \) is instantiated for each pattern \( P_i \) representing a physical anatomic specimen

\[
\text{Specimen}_i = \begin{pmatrix} SS : (D \mid \mu_{[\text{Real]}}, \sigma_{[\text{Real}]}) \\ MS : (pp_{[\text{Real]}}, SV_{[\text{Real}]}) \end{pmatrix}.
\]

---- eq. 2

The structure schema SS of a specimen is represented by the pair \( (\mu, \sigma) \) of the distribution \( D \) for each of the \( N \) features \( j = 1 \ldots N \) in pattern \( P_i \), respectively. The measure schema MS of a specimen is represented by two values:

- The prior probability (pp) and
- The scatter value (SV) of Pi.

pp is defined as the fraction of the feature vectors of the image that belong to pattern \( P_i \). SV is a measure of the cohesiveness of the data items in a cluster with respect to the centroid of the cluster, and it is a commonly used intrinsic measure of the quality of a cluster. The scatter value SV of a specimen is defined as

\[
SV = \sum_{k \in P_i} (x_k - c_{P_i})^2
\]

---- eq. 3

Where \( x_k \) are the feature vectors that belong to pattern \( P_i \) and \( c_{P_i} \) is the corresponding centroid, which is also a vector having the same dimensionality as \( x_k \).

In this context, a medical image MI is considered as a complex pattern and it is defined as

\[
MI = \left( SS \mid \{\text{Specimen}\} \right)
\]

---- eq. 4

D. Computation of pattern similarities:
   CBIR system is the estimation and ranking of the similarity between query and registered images. The similarity between two medical images is to define the distance over the structures and the measures of two simple patterns \( P1 \) and \( P2 \). Since complex patterns are decomposed into a number of simple patterns, in comparing two medical images, \( MI1 \) and \( MI2 \), it is a way to associate component patterns of \( MI1 \) to component patterns of \( MI2 \). The coupling type constrains is the way for associating the component patterns. Next, an effective way is proposed to measure the distance between two simple patterns, and then coupling process is presented. The distance between the measures of two patterns is defined as the absolute difference of the scatter values, each one weighted by the corresponding prior probability of the patterns, normalized by the sum of the two scatter values.

\[
d_{\text{mean}}(P1, P2) = \frac{|P1_{pp} \cdot P1_{SV} - P2_{pp} \cdot P2_{SV}|}{P1_{SV} + P2_{SV}}.
\]

---- eq. 5

The above equation quantifies the inter pattern divergence between the cohesiveness of two clusters which defines that overrides the inefficiency of the relativeness of
the scatter value with respect to the number of items in the cluster, as each scatter value is weighted by the fraction of the feature vectors of the image that belong to pattern \( P_i \). The standardized difference \( d \) between two distributions \( P_1 \) and \( P_2 \) is defined by Cohen. Cohen's \( d \) is defined as the absolute difference between the means of the distributions, divided by the root-mean square of the two standard deviations

\[
\begin{align*}
    d[D_1, D_2] = \begin{cases} 
    \frac{|D_1 \mu - D_2 \mu|}{\sqrt{\frac{D_1 \sigma^2 + D_2 \sigma^2}{2}}}, & \text{if } D_1 \sigma \neq 0 \text{ or } D_2 \sigma \neq 0, \\
    |D_1 \mu - D_2 \mu|, & \text{otherwise}
    \end{cases}
\end{align*}
\]

Cohen's distance \( d \) is a nonnegative real number interpreting the overlap between two distributions. If \( d \) is zero, the distributions are identical. Low \( d \) indicates quite similar distributions whereas high \( d \) indicates quite dissimilar distributions. If both standard deviations are zero, the absolute difference between the means is used as the distance between the distributions.

II. Pattern Instantiation via Clustering:

The feature vectors extracted from each image were clustered using an implementation of the EM algorithm using the 10-fold cross-validation algorithm to determine the number of clusters. Each cluster was represented by a pattern Specimen, \( i = 1, \ldots, M \) and each image was represented by a complex pattern \( M_l \). Figure 2(a) illustrates three radiographic images from breast, abdomen, and hand categories (from left to right). The respective clustering’s obtained are illustrated in Figure 2(b). The different grey levels in Figure 2(b) indicate the different specimen patterns found in the images. Figure 2(c) illustrates projections of the 2-D feature vectors to a 3-D space constructed according to the centroid preserving projection technique [3].

Effectiveness

The effectiveness of the proposed pattern is to evaluate the retrieve images by adopting the popular recall and precision measures, where recall is defined as the ratio of the relevant images retrieved over the total relevant images in the database, and precision is defined as the ratio of the relevant images retrieved over the total number of images retrieved, relevant or not.

Efficiency:

The efficiency of the proposed medical image similarity scheme involves the pattern comparisons. A vector comparison in the conventional approach is considered equivalent to a pattern comparison in the proposed scheme. The sequential and exhaustive scan is chosen as the yardstick for our method, as other common methods such as R-trees are sensitive to the high dimensionality of the feature vectors, which is usual in CBIR applications. The performance of these approaches degrades rapidly as dimensionality increases. For instance, it has been shown that even for a dimensionality of as low as 5, the R*-tree behavior in similarity search is problematic. The main reason is that, with the growth of the dimensionality, the overlap in the internal nodes of the tree increases and, as such, its discrimination ability decreases. It can be observed that the advantage of the proposed approach increases with the number of blocks per image (e.g., by increasing the sampling step), and for a few hundreds of blocks per image, it requires almost three orders of magnitude fewer comparisons than the conventional approach. The average processing time (CPU plus I/O time) for the comparison of a pair of images is estimated. For the aforesaid experimental setting, the proposed pattern similarity scheme requires always less than 0.1 ms. The average time required for the mixture model parameters to converge to a constant or almost-constant estimate is 0.22 ± 0.04 s.

RESULTS

Case 1: When the input image is existing in the database

Case 2: When the input image is not matched with the database images

Figure 3: Result when input image is recognized

Figure 4: Result when input image is not recognized
CONCLUSION

In this paper, a scheme for efficient content-based medical image retrieval is presented. This scheme utilizes rich-semantics pattern representations of medical images, defined in the context of PANDA, a framework for representing and handling data mining results. The results advocate both its efficiency and effectiveness in comparison with state of the art.

Future perspectives of this paper include: 1) systematic evaluation of the proposed scheme for the retrieval of various kinds of medical images, such as endoscopic and ultrasound images according to their pathology; 2) the enhancement of the retrieval performance by using image indexing techniques based on specialized data structures.

REFERENCES:


