**IOSR Journal of Engineering (IOSRJEN)**
**e**-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 3 (Mar. 2013), ||V1|| PP 18-31

# A pattern scheme for PANDA Medical Image Retrieval

[1] S.Yamuna, [2] Dr.S.Varadarajan, [3] K.Srinivasa Reddy

***Abstract:*** In this paper, a novel scheme for efficient content-based medical image retrieval, formalized according to the Patterns for Next generation Database systems (PANDA) framework for pattern representation and management. This scheme involves block-based low-level feature extraction from images followed by the clustering of the feature space to form higher-level, semantically meaningful patterns. The clustering of the feature space is realized by an expectation–maximization algorithm that uses an iterative approach to automatically determine the number of clusters. Experiments were performed on a large set of reference radiographic images, using different kinds of features to encode the low-level image content. Through this experimentation, it is shown that the proposed scheme can be efficiently and effectively applied for medical image retrieval from large databases, providing unsupervised semantic interpretation of the results, which can be further extended by knowledge representation methodologies.

***Keywords:*** *Patterns, CBIR, PANDA, Content based images, color HSV*

## I. INTRODUCTION

The use of images in human communication is hardly new. The use of maps and buildings plans to convey information almost certainly dates back to pre-roman times. But, the twentieth century has witnessed unparalleled growth in the number, availability and importance of images in all walks of life. Images now play a crucial role in the fields as diverse as medicine, journalism, advertising, design, education and entertainment. Technology, in the form of inventions such as photography and television, has played a major role in facilitating the capture and communication of image data. But, the real engine of imaging revolution has been the computer, bringing with it a range of techniques for digital image capture, processing, storage and transmission. Once computerized, imaging became affordable and it soon penetrated into areas that were traditionally depending heavily on images for communication, such as engineering, architecture and medicine. Photograph libraries, art galleries and museums, too, began to see the advantages of making their collections available in electronic form. The creation of the World-Wide Web in the early 1990's, enabling users to access data in a very variety of media from anywhere on the planet, has provided a further massive stimulus to the exploitation of digital images. The number of images on the Web was recently estimated to be between 10 and 30 million. The process of digitization does not in itself make image collections easier to image. Some form of cataloguing and indexing is still necessary to manipulate relevant images.

Content-based image retrieval (CBIR) has become an important practicable technique to support effective searching and browsing of larger and larger collections of unstructured images and videos. Content-based image retrieval - CBIR uses visual content (low-level features) of images such as color, texture, shape, etc. to represent and to index images. These features are described by multi-dimensional vectors called feature vectors that are used in the process of retrieve similar images. Extensive experiments on CBIR show that low-level features not represent exactly the high-level semantic concepts and can fail when used to retrieve similar images. In order to overpass this problem, different approaches aim to propose new methods that use different techniques combined with low level descriptors.

## II. CONTENT BASED IMAGE RETRIEVAL (CBIR)

Content based image retrieval is a technique for retrieving images from a database on the basis of automatically derived features such as color, texture and shape etc. The features used for retrieval can be either primitive or semantic, but the extraction process must be predominantly automatic. Retrieval of images by manually assigned keywords describes image content. CBIR defers from the classic information retrieval in that the image data bases are essentially unstructured, since digitized images consists purely of arrays of pixel intensities, with no inherent meaning. One of the key issues with any kind of image processing is the need to extract useful information from the raw data( such as recognizing the particular shapes or textures) before any kind of reasoning about the image contents is possible. Image data bases thus defer fundamentally for text data base where the raw materials (words stored as ASCII character strings) had already been logically structured.

CBIR draws many of its methods from the fields of image processing and computer vision. It defers from these fields principally through its emphasis on the retrieval of images with desired characteristics from a collection of significant size. Image processing covers a much wider field, including image enhancement,

compression, transmission and interpretation. While there are gray areas (such as object reorganization by feature analysis) the distinction between main stream image analysis and CBIR is usually fairly clear cut. An example may make this clear.

In this case the two images are matched; a process few observe would call CBIR. Second, the entire database may be search to find the most closely matching images. This is a genuine of example of CBIR. This work focuses on CBIR systems representing information content of image by visual features such as color, texture, and shape and retrieve images based on the similarity of features.

## III. SYSTEM ARCHITECTURE OF CBIR:

The general scheme of image retrieval from a database. The basic idea behind content based image retrieval is the extraction of feature vectors (the features can be color shape, texture, region, or spatial features, or features in some compressed domain, etc.) These vectors are then stored in a database for future use. When given a query image, its feature vectors are similarly extracted and matched with those in the database. If the distance between the query image features and feature vector available in the database is small enough, the corresponding image in the database is considered a match to the query. The search is usually based on similarity rather than on exact match. The retrieval results are then ranked according to a similarity index and a group of similar target images is usually presented to the users.

The main Block diagram includes digitizer, feature extractor, image data base, feature data base, matching and multi dimensional indexing.

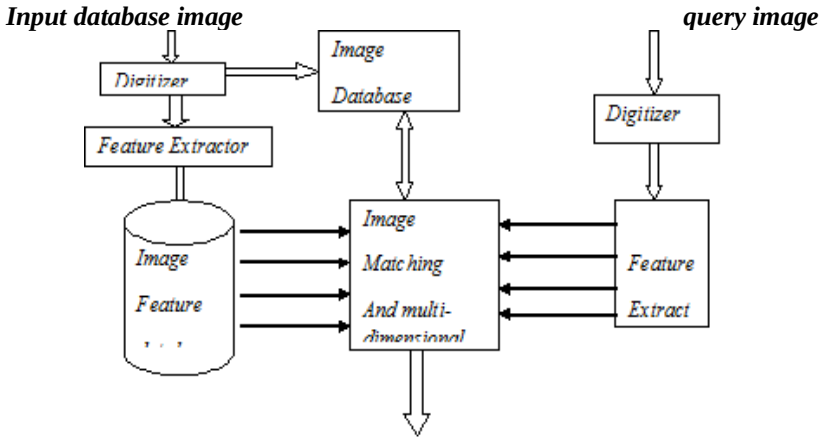


Fig: General scheme of Content Based Image Retrieval

**Digitizer:**

To add new images to data base, quire images acquired from CCD camera, X-ray imaging system, micro densitometer, image de-sectors, Videocon cameras etc, or needed to digitized, so that computer can process those images.

**Image data base:**

The comparison between query image and images from the data base can be done pixel to pixel which will give precise match. But on the other hand recognizing objects entirely at query time will limit the retrieval speed of the system. Due to high expenses of such computing, generally this crude method of comparison not used, but image data base contains raw images, is required for visual display purpose.

**Feature Extraction:**

To avoid the above problem of Pixel-by-pixel comparison, a better abstraction level for representing images is the feature level. Every image characterized by a set of these features such as Texture, color and shape etc., and the extraction of these features are summarized in a reduced set of k indices and stored in the feature database, The query image is then processed in the same way as the images in the database. Thus, matching and on the feature database.

**Application of CBIR.**

There are many applications where content-based image retrieval is important. Some of them are:

- In architecture, real estate, and interior Design: Allows users to find similar building and decoration of rooms that correspond to more appealing structures from database.

- In education: In history, for example, it's always helpful to have immediate access to images and short video sequences of relevant and people. in such cases CBIR would be extremely useful.
- In geographical information systems.
- In medicine, the medical literature contains volumes of photograph of normal versus pathological condition in every part of the day.
- Diagnosis may require recalling the current condition and checking its resemblance with the conditions from the literature.
- In remote sensing. In finding relevant data from satellite images.
- In film and video archives, to find video shots quickly for particular characteristics such as color, texture and shape of even high level concepts such as particular persons, places or objects.

## IV. MOTIVATION

CBIR systems gained a lot of prominence in the last few years mainly due to the increase in multimedia database and information repositories.CBIR systems can focus on retrieving images for a given image based on color, shape or texture. Image texture is an important visual primitive to search and browse through large collections of similar looking patterns, hence this project focuses on CBIR for face images. The CBIR system is to create a database of features from a database of images.

**low level features for image retrieval**

Different techniques were proposed for extracting low-level features. Color feature is one of the most widely used features in image retrieval because it is efficient in describing colors although it is not directly related to high-level semantics. MPEG-7 is an ISO/IEC standard developed by Moving Pictures Expert Group for standardizing the description of multimedia content data. This standard defines seven color descriptors: Color space, color quantization, dominant colors, scalable color, color layout, color-structure, and GoF/GoP color. The scalable color descriptor is a color histogram in HSV color space, which is encoded by a Haar transform. Its binary representation is scalable in terms of bin numbers and bit representation accuracy over a broad range of data rates. Color layout descriptor represents the spatial distribution of color of visual signals in a very compact form. This compactness allows visual signal matching functionality with high retrieval efficiency at very small computational costs.

**COLOR**

Color does not only add beauty to objects but also give more information, which is used as powerful tool in content based image retrieval. In color indexing, given a query image, the goal is to retrieve all the images whose color composition is similar to the color composition of query image. In color image retrieval there are various methods, but here the discussion is on some prominent methods.

Typically the color composition is characterized by color histograms. In 1991 Swain and Ballard proposed the method, called color indexing, which identifies the objects using color histogram indexing. Color histograms are a way to represents the distribution of colors in images where each histogram bin represents a color in a suitable histogram can be used to define similarity match between the two distribution. The core idea is to compute

$$H[I, M] = \sum_{n} \min (I_j, M_j)$$

where H (I,M) is the match value, and Ij and Mj are jth image(query image) and model (an image in the data base) histogram respectively, each containing n bins. The min (Ij,Mj ) indicates the minimum value pixel among all the pixels. The match value is computed for very model histogram and the value is closer to unity if the model image is more similar, it is obvious that a match value of unity is obtained for an image compared with it. It is clear that for this approach, the feature 'f' used to characterized the color information of an image is the 3-D color histogram h(x, y, z) and the similarity measure between features is given by the match value as shown in equation (2.1).

As seen earlier, the histogram intersection technique takes in to account every color bin o f the 3-D color histogram of the two images compared and it does a detailed comparison. However, for many synthesized images like trade mark images, flags textile design patterns etc there large regions of uniform colors, therefore, the 3-D histogram will have a few dominate peaks and the rest of the bins do not capture much color information of the images. Hence detailed comparison for such images is not required. Also there is some noise

introduced during the process of scanning images. Hence fine Comparison is not necessary and may even produce in correct results.

To Overcome above problem in 1995 Mehtre proposed two new color-matching methods the "Distance Method "and "Reference color Table Method", for image retrieval. In the Distance method they proposed a coarse comparison of the color histogram of the query and model images. The feature they used for capturing the color information is the mean value, μ, of the 1-D histogram of each of the three color components of the image, these components could be, R, G, B for the RGB representation or the three opponent color axes Rg, By and $\omega_b$ (Swain and Ballard, 1991) given by

$$\mathbf{R_g = R-G,}$$

$$\mathbf{B_y = 2*B-R-G,}$$

$$\mathbf{\omega_b = R+G+B}$$

Therefore, the feature vector Fv for characterizing an RGB image will be

$$\mathbf{f_v = (\mu_R, \mu_G, \mu_B)}$$

The histogram can be normalized by considering the retrieval fraction of pixels (compared to the total number of pixel in the image) in each bin of the histogram. Then use a distance measure to compute similarity or match value for given pair of images. Depending on the type of distance measure used-Manhattan (city block) or Euclidean-the following measures are taken:

$$\mathbf{D^M_{ql} = | f_q - f_i | = \sum_{R,G,B} | \mu q - \mu I |}$$

$$\mathbf{D^E_{qi} = \sqrt{(fq - fi)^2} = \sqrt{\sum_{R,G,B} ( \mu q - \mu I )^2}}$$

Where $D^M_{ql}$ is the Manhattan and $D^E_{qi}$ is the Euclidean distance between the query image and the database image, fq is the color feature vector of the query image fi and is the color feature vector of the database image. It is obvious that the distance of an image from itself is zero. The results show that both the new methods perform better than the existing histogram intersection method. Currently both histogram intersection and their methods require a linear search, which can be very time consuming for large database. Most color histograms are very sparse and thus sensitive to noise. In 1995 Stricker and Orengo proposed cumulated color histogram. Observing the fact that the color histograms lack information about how color is spatially distributed, in 1997 Huang, introduced a new color feature for image retrieval called color correlogram. This feature characterizes how the spatial correlation of pairs of color changes with distance in an image. Usually, because the size of color correlogram is quite large, the color autocorrelation is often used instead. This feature only captures spatial correlation between identical colors.

In 1999, Greves and Smeulders analyzed and evaluated various color features for the purpose of image retrieval by color-metric histogram matching under varying illuminations environments. They introduced new color model and concluded that this color model is most appropriate color model to be used for image retrieval by color-metric histogram matching under the constraint of white illumination source.

In 2001 Kim et al. proposed an efficient indexing/matching algorithm that is independent of the changes in the illuminant color and the geometric conditions for 3-D objects with multiple color.

**HSV color space:**

HSV Color space: Basically there are three properties or three dimensions of color that being hue, saturation and value HSV means Hue, Saturation and Value. It is important to look at because it describes the color based on three properties. It can create the full spectrum of colors by editing the HSV values. The first dimension is the Hue. Hue is the other name for the color or the complicated variation in the color. The quality of color as determined by its dominant wavelength. This Hue is broadly classified into three categories. They are primary Hue, Secondary Hue and Teritiary Hue. The first and the foremost is the primary Hue it consists of three colors they are red, yellow and blue. The secondary Hue is formed by the combination of the equal amount of colors of the primary Hue and the colors of the secondary Hue which was formed by the primary Hue are Orange, Green and violet. The remaining one is the teritiary Hue is formed by the combination of the primary Hue and the secondary Hue. The limitless number of colors is produced by mixing the colors of the primary Hue in different amounts. Saturation is the degree or the purity of color

**Properties of the HSV color space:**

Sensing of light from an image in the layers of human retina is a complex process with rod cells contributing to scotopic or dim-light vision and cone cells to photopic or bright-light vision (Gonzalez and Woods, 2002). At low levels of illumination, only the rod cells are excited so that only gray shades are perceived. As the illumination level increases, more and more cone cells are excited, resulting in increased color perception. Various color spaces have been introduced to represent and specify colors in a way suitable for

storage, processing or transmission of color information in images. Out of these, HSV is one of the models that separate out the luminance component (Intensity) of a pixel color from its chrominance components (Hue and Saturation). Hue represents pure color, which is perceived when incident light is of sufficient illumination and contains a single wavelength. Saturation gives a measure of the degree by which a pure color is diluted by white light. For light with low illumination, corresponding intensity value in the HSV color space is also low.

The HSV color space can be represented as a Hexa cone, with the central vertical axis denoting the luminance component, I (often denoted by V for Intensity Value). Hue, is a chrominance component defined as an angle in the range [0,2p] relative to the red axis with red at angle 0, green at 2p/3, blue at 4p/3 and red again at 2p. Saturation, S, is the other chrominance component, measured as a radial distance from the central axis of the hexacone with value between 0 at the center to 1 at the outer surface. For zero saturation, as the intensity is increased, we move from black to white through various shades of gray. On the other hand, for a given intensity and hue, if the saturation is changed from 0 to 1, the perceived color changes from a shade of gray to the most pure form of the color represented by its hue. When saturation is near 0, all the pixels in an image look alike even though their hue values are different.

As we increase saturation towards 1, the colors get separated out and are visually perceived as the true colors represented by their hues. Low saturation implies presence of a large number of spectral components in the incident light, causing loss of color information even though the illumination level is sufficiently high. Thus, for low values of saturation or intensity, we can approximate a pixel color by a gray level while for higher saturation and intensity, the pixel color can be approximated by its hue. For low intensities, even for a high saturation, a pixel color is close to its gray value. Similarly, for low saturation even for a high value of intensity, a pixel is perceived as gray. We use these properties to estimate the degree by which a pixel contributes to color perception and gray level perception.

## V. FEATURE EXTRACTION OF THE HSV COLOR:

**Image retrieval:**

Image retrieval is nothing but a computer system used for browsing searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval use some method of adding metadata by captioning, Keywords or the descriptions to the images so that the retrieval can be performed. Manual image annotation is time consuming, expensive and laborious. For addressing this there has been a large amount of research done on automatic image annotation. It is crucial to understand the scope and nature of the image data in order to determine the complexity of the image search system design. The design is also largely dependent on the factors. And some of the factors include archives, Domain specific collection, Enterprise collection, Personal collection and web etc..,

Invention of the digital camera has given the common man the privilege to capture his world in pictures, and conveniently share them with others. one can today generate volumes of images with content as diverse as family get-togethers and national park visits. Low-cost storage and easy Web hosting has fueled the metamorphosis of common man from a passive consumer of photography in the past to a current-day active producer. Today, searchable image data exists with extremely diverse visual and semantic content, spanning geographically disparate locations, and is rapidly growing in size. All these factors have created innumerable possibilities and hence considerations for real-world image search system designers.

An image retrieval system designed to serve a personal collection should focus on features such as personalization, flexibility of browsing, and display methodology. For example, Google's Picasa system [Picasa 2004] provides a chronological display of images taking a user on a journey down memory lane. Domain specific collections may impose specific standards for presentation of results. Searching an archive for content discovery could involve long user search sessions. Good visualization and a rich query support system should be the design goals. A system designed for the Web should be able to support massive user traffic. One way to supplement software approaches for this purpose is to provide hardware support to the system architecture. Unfortunately, very little has been explored in this direction, partly due to the lack of agreed-upon indexing and retrieval methods.

**Characteristics of the color cumulative histogram:**

Research on image retrieval technology based on color feature, for the color histogram with a rotation, translation invariance of the advantages and disadvantages of lack of space, a color histogram and color moment combination Image Retrieval. The theory is a separate color images and color histogram moment of extraction, and then two methods of extracting color feature vector weighted to achieve similar distance, similar to the last distance based on the size of the return search results, based on the realization of the characteristics of the color image Retrieval system. The results show that the method is rotation, translation invariance, a single method of extracting color features, enhanced image search and improve the accuracy of the sort.

Color retrieval is one of the most significant features of the image retrieval. There are many color models to express color such as the RGB color model YUV color model and the HSV color model. HSV color model is thereof most consistent with the induction of the human visual model. *H* represents color hue, and it is the wavelength of the light reflected from an object or throughout the object; *S* represents color saturation, means how much white is added to the color; *V* represents brightness (vale), is the degree of color shading. But, the computer can only identify the RGB color component of an image, in which *R* represents the red component, *G* represents the green component, *B* represents the blue component. Therefore, we need the following formula for the image conversion from RGB color space to HSV color space

**Color histogram:**

Color have a certain stability, and it is not sensitive to size and direction, so the image retrieval technology using color characteristics has been pay great attention to. The retrieval method of using color characteristic was originally proposed by Swain and Ballard, they put forward t the color histogram method of which the core idea is to use a certain color space quantization method for color quantization, and then do statistics for the proportion of each quantitative channel in the whole image color. Abscissa represents the normalized color value, ordinate represents the sum of image pixels which corresponding to each color range. Image statistical histogram is a one-dimensional discrete function:

$$h_k = \frac{n_k}{n}, \quad k=0,1,\cdots,L-1$$

The letter k presents eigen values of color, letter l presents the number of features of value . So we get the color histogram of the image P as follows:

$$H_{(p)}=[h_1,\ h_2,\ \ldots h_{L-1}]$$

There are many color histogram methods such as the global color histogram, cumulative histogram and sub-block histogram. However, color histogram has its own drawbacks, such as the color histograms of different images may be the same:

There are two unrelated images in Figure 1 , but they are the same as the color histogram in Figure 2

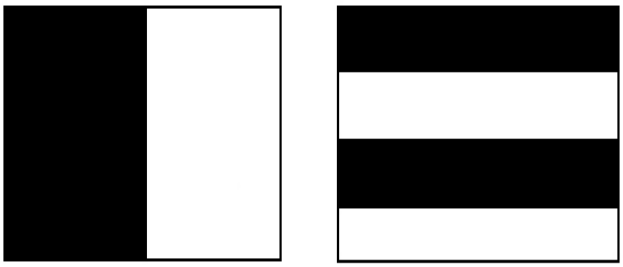

Figure : Two unrelated images

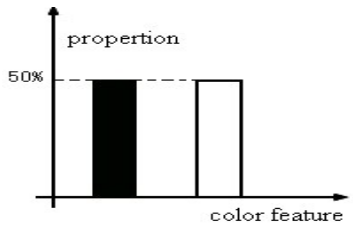


Figure: Color histogram

**Color moment:**

Stricker and Orengo who propose the method of color moment consider that the color information focus on the low-level color moment of the image, and they mainly do statistics for the first order, second-order and third-order moment of each color component. For image retrieval, the color moment is a simple and effective representative method of color features. Such color moment as first-order (mean) and second (variance) and third-order (gradient), is proved to be very effective in presenting color distribution of images. The three colors moments are defined with figures

as follows:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij}$$

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^2}$$

$$s_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^3}$$

In which $f_{ij}$ represents the probability of the color components of the pixels j is I, N represents the number of pixels in the image. Each pixel has three color components and each component has three lower order moments so there are nine color characteristic components. They are represented as follows:-

$$\mu_1, \sigma_1, s_1, \mu_2, \sigma_2, s_2, \mu_3, \sigma_3, s_3.$$

However the color moment is only the initial color characteristic extraction of the image and the effect of the extraction is very rough. So is regularly integrated used for other extraction methods. In practice there are other color feature extraction methods such as color sets color Correlogram color polymerization vector and consistent color vector.

**Low level Features for Image Retrieval**

Different techniques were proposed for extracting low-level features. Color feature is one of the most widely used features in image retrieval because it is efficient in describing colors although it is not directly related to high-level semantics. MPEG-7 is an ISO/IEC standard developed by Moving Pictures Expert Group for standardizing the description of multimedia content data. This standard defines seven color descriptors: Color space, color quantization, dominant colors, scalable color, color layout, color-structure, and GoF/GoP color. The scalable color descriptor is a color histogram in HSV color space, which is encoded by a Haar transform.

**color histogram features**

In image processing and photography, a color histogram is a representation of the distribution of colors in an image. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space, the set of all possible colors.

The color histogram can be built for any kind of color space, although the term is more often used for three-dimensional spaces like RGB or HSV. For monochromatic images, the term intensity histogram may be used instead. For multi-spectral images, where each pixel is represented by an arbitrary number of measurements (for example, beyond the three measurements in RGB), the color histogram is *N*-dimensional, with N being the number of measurements taken. Each measurement has its own wavelength range of the light spectrum, some of which may be outside the visible spectrum.

If the set of possible color values is sufficiently small, each of those colors may be placed on a range by itself; then the histogram is merely the count of pixels that have each possible color. Most often, the space is divided into an appropriate number of ranges, often arranged as a regular grid, each containing many similar color values. The color histogram may also be represented and displayed as a smooth function defined over the color space that approximates the pixel counts. Like other kinds of histograms, the color histogram is a statistic that can be viewed as an approximation of an underlying continuous distribution of colors values.

Color histograms are flexible constructs that can be built from images in various color spaces, whether RGB, rg chromaticity or any other color space of any dimension. A histogram of an image is produced first by discretization of the colors in the image into a number of bins, and counting the number of image pixels in each bin. For example, a Red–Blue chromaticity histogram can be formed by first normalizing color pixel values by dividing RGB values by R+G+B, then quantizing the normalized R and B coordinates into N bins each.

In this module first the RGB image is changed to grayscale image, also known as the intensity image, which is a single 2-D matrix containing values from 0 to 255. After the conversion from RGB to grayscale image, we perform quantization to reduce the number of levels in the image. We reduce the 256 levels to 16 levels in the quantized image by using uniform quantization. The segmentation is done by using color histograms.

## VI. CLUSTER ANALYSIS

Cluster analysis or clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research. By clustering, one can identify dense and sparse regions and therefore, discover overall distribution patterns and interesting correlations among data attributes.

As a branch of statistics, cluster analysis has been studied extensively for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS. In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. In conceptual clustering, a group of objects forms a class only if it is describable by a concept. This differs from conventional clustering, which measures similarity based on geometric distance. Conceptual clustering consists of two components: (1) it discovers the appropriate classes, and (2) it forms descriptions for each class, as in classification. The guideline of striving for high intra class similarity and low interclass similarity still applies.

An important step in most clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point ($x = 1$, $y = 0$) and the origin ($x = 0$, $y = 0$) is always 1 according to the usual norms, but the distance between the point ($x = 1$, $y = 1$) and the origin can be 2, or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

## VII. COMMON DISTANCE FUNCTIONS:

- The Euclidean distance (also called distance as the crow flies or 2-norm distance). A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.
- The Manhattan distance
- The maximum norm
- The Mahalanobis distance corrects data for different scales and correlations in the variables
- The angle between two vectors can be used as a distance measure when clustering high dimensional data.
- The Hamming distance measures the minimum number of substitutions required to change one member into another. Another important distinction is whether the clustering uses symmetric or asymmetric distances. Many of the distance functions listed above have the property that distances are symmetric.

**Future Enhancement**

- . This classification of feature set can be enhanced to heterogeneous (shape, texture) so that we can get more accurate result.
- It can also enhanced to merging of heterogeneous features and neural network.
- The schemes proposed in this work can be further improved by introducing fuzzy logic concepts into the clustering process.

## VIII. K-MENS CLUSTERING ALGORITHM

Algorithm: k-means. The k-means algorithm for partitioning based on the mean value of the objects in the cluster. Input: The number of clusters k and a database containing n objects.
Output: A set of k clusters that minimizes the squared-error criterion. Method:

- arbitrarily choose k objects as the initial cluster centers:
- repeat
- (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
- Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
- Until no change.

The benefits emanating from the application of content-based approaches to medical image retrieval range from clinical decision support to medical education and research . These benefits have 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. Common ground for most of the systems cited earlier is that image retrieval is based on similarity measures estimated directly from low-level image features. This approach is likely to result in the retrieval of images with significant perceived differences from the query image, since low-level features usually lack semantic interpretation. This has motivated researchers to focus on the utilization of higher-level semantic representations of image contents for content-based medical image retrieval. Recent approaches include semantic mapping via hybrid Bayesian networks, semantic error-correcting output codes (SECC) based on individual classifiers combination, and a framework that uses machine learning and statistical similarity matching techniques with relevance feedback . However, these approaches involve supervised methodologies that require prior knowledge about the dataset and introduce constraints to the semantics required for the image retrieval task. 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 paper are the following.

1) We define 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. The pattern representation of clusters involves both structural descriptors and quality measures.
2) We propose a novel scheme for the assessment of the similarity between complex patterns (i.e., medical images) for CBIR purposes.
3) We conduct a comprehensive set of experiments over a publicly available set of radiographic images, in order to thoroughly evaluate our approach and demonstrate its effectiveness and efficiency in comparison to state-of-the-art techniques. The rest of this paper is structured as follows. Section II outlines the PANDA framework, which provides necessary background information to the reader. The proposed pattern similarity scheme for medical image retrieval is presented in Section III. The results obtained from the experimental evaluation of the proposed scheme are apposed in Section IV. The conclusions along with the future perspectives are summarized in Section V.

## IX. PANDA FRAMEWORK

The efficient management of patterns extracted from medical image databases is of vital importance due to the extremely large storage requirements as well as the complexity of such kind of raw data. Taking advantage of the PANDA framework , we adopt the idea of a pattern-base (PB) keeping information about extracted patterns in a compact and unified way. A PB consists of three basic layers: the *pattern type*, the *pattern*, 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 defined as a pair PT=_SS,MS_, where SS defines the pattern space by describing the structure schema of the pattern type, while the measure schema MS quantifies the quality of the source data representation achieved by patterns of this pattern type. As an example, consider a pattern type representing Euclidean-distance, spherical-like clusters in a *D*-dimensional space. The structure of such a pattern type may be modeled by specifying the cluster center (a *D*-dimensional vector) and a radius (a real value). The measure for a cluster might be, for instance, its support, that is, the fraction of the data points represented by the cluster. As such

$$\text{EuclideanCluster}=\begin{pmatrix}\text{SS}:(\text{center}:[\text{Real}]_1^D,\text{radius}:\text{Real})\\ \text{MS}:(\sup p:\text{Real})\end{pmatrix}.$$

A pattern-type PT is called *complex* if its structure schema SS includes another pattern type, otherwise PT is called *simple*. Thus, a Euclidean Cluster is a simple pattern type, whereas a clustering extracted, e.g., by a partitioning clustering algorithm is considered a complex pattern type since it can be modeled as a set of clusters with no measure component

## X. MEDICAL IMAGE RETRIEVAL USING PATTERNS

The proposed content-based medical image retrieval scheme is outlined in Fig. 1. It involves four steps: 1) low-level feature extraction from each of the registered and query images; 2) clustering of the extracted feature vectors per image; 3) pattern instantiation of the resulted clusters; and 4) computation of pattern similarities. The registration of a new image into the database involves steps 1)–3), whereas step 4) is processed during the retrieval task.

**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, sampling image blocks at a given sampling step. The sampling step may allow consecutive blocks to overlap. For each block, a set of *N* features *fi* , *i* = 1, . . ., *N*, is calculated to form a single feature vector *F*. The number of feature vectors produced for each image depends on the size, the dimensions of the sliding window, and the sampling step. 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. Considering an image as a set of block samples, the features used with the proposed pattern similarity scheme should describe properly the local content of the image. it should be noted that 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.

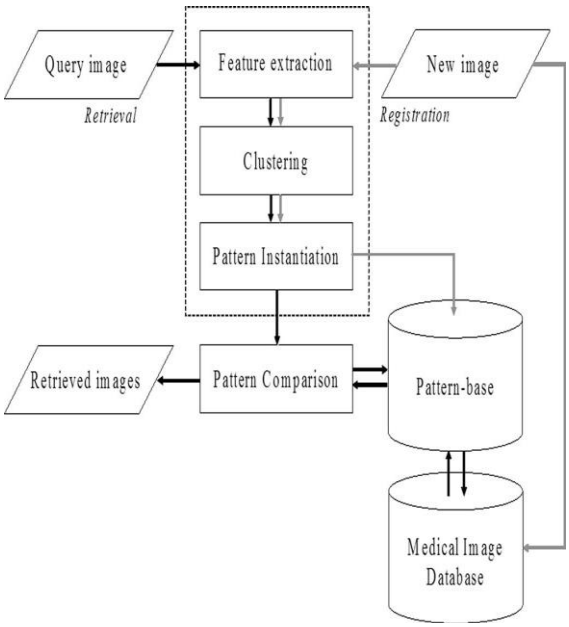


Fig. 1. Outline of the proposed content-based image retrieval methodology. The black arrows indicate the data flow for image retrieval, whereas the grey arrows indicate the data flow for the registration of a new image.

**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) algorithm . With *c* Gaussians, the probability density function of a variable *X*

Where $pp_i > 0, \sum_{i=1}^{c} pp_i = 1,$ and d is the dimension of the feature vector. The set of model parameters $\theta\{pp_i, \mu_i, \sum_i\}, i = 1, \ldots, c,$ consists of the prior probabilities *pp*i of the Gaussian *i*, the mean vector *μi* , and the covariance matrix $\sum_i$ for the Gaussian *i*, respectively. The EM algorithm is used to estimate the maximum likelihood *L* of $\theta$ given a set of features $\{x_1, \ldots, x_N\}$

The model parameters are initialized with random values. The algorithm starts by calculating the probabilities that a vector should belong to each distribution. These probabilities are used to compute a new estimate for the parameters. The whole process is repeated until the parameters converge to a constant or almost

constant estimate. 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 number of vectors that belong to the cluster). The vector of means $\mu$ of the distributions for every feature represents the centroid of the cluster. The EM algorithm exhibits many advantages over other clustering algorithms that make it appealing for use with the CBIR methodology described in this paper. 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. Starting with one cluster, EM is performed *v* times holding out one segment at a time for test purposes and the likelihood is averaged over all the results. Next, EM is performed over two clusters, and if the likelihood increases, the number of clusters is set to two and the process is repeated until the estimated likelihood begins to decrease. Furthermore, the EM algorithm is more general than, e.g., *K*-means, as it can find clusters of different sizes and ellipsoidal shapes. 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, which is discussed next.

**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 presented. Hence, given a clustered image comprising of *M* simple patterns $P_i, i = 1, \ldots, M,$ and with respect to the output of the EM algorithm, a Specimen *i* is instantiated for each pattern $P_i$ representing a physical anatomic specimen in a medical image.

$$\text{Specimen}_i = \begin{pmatrix} \text{SS} : (D : [\mu : [\text{Real}], \sigma : [\text{Real}]]_1^N) \\ \text{MS} : (pp : [\text{Real}], \text{SV} : [\text{Real}]) \end{pmatrix}.$$

More specifically, the structure schema SS of a specimen is represented by the pair ($\mu$, $\sigma$) of the distribution *Dj* for each of the *N* features $(j = 1, \ldots, N)$ in pattern *Pi* , respectively. Correspondingly, the measure schema MS of a specimen is represented by two values, the prior probability (pp) and the scatter value (SV) of *Pi*. Formally, the prior probability *pp* is defined as the fraction of the feature vectors of the image that belong to pattern *Pi* . Intuitively, *pp* is equivalent with the support measure widely used in data mining models. In this case, it provides an indication of the size of the specimen. On the other hand, 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. Formally, the scatter value SV of a specimen is defined as

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

where *xk* are the feature vectors that belong to pattern *Pi* and *cP i* is the corresponding centroid, which is also a vector having the same dimensionality as *xk* , and its value in each dimension is computed as the average from the corresponding features' values belonging to pattern *Pi* . A low scatter value indicates good scatter quality, but it should be noted that this is a relative measure of quality, since it depends on the number of items in the cluster. In this context, a medical image MI is considered as a complex pattern.

$$\text{MI} = \begin{pmatrix} \text{SS} : \{\text{Specimen}\} \\ \text{MS} : \perp \end{pmatrix}$$

Consisting of a set of simple patterns (i.e., specimens), which follow the definition

## XI. RESULTS

A number of experiments were performed with radiographic images from the image retrieval in medical applications (IRMA) dataset ,which is often used as a reference for medical image retrieval tasks. It currently contains 10 000 arbitrarily selected anonymous radiographic images taken randomly from patients of different ages, genders, and pathologies during medical routine. The images are categorized into 116 classes according to the IRMA code. This code comprises four fields: 1) the imaging modality; 2) direction of the imaging device and the patient; 3) the anatomic body part that is examined; and 4) the system under investigation. The particular dataset comprises only plain X-ray images of various directions (such as anteroposterior and mediolateral), anatomic body parts (such as cranium, spine, arm, elbow, and chest) and systems under investigation (such as musculoskeletal, gastrointestinal, and uropoietic). iorthogonal spline wavelet decomposition of each sampled block and the estimation of the first two wavelet moments from each band. This process results in a 20-D feature vector per block. The determination of the sampling parameters was

based on preliminary experiments seeking the maximum average distance between complex patterns MI of the different categories comprising the registered dataset. The sampling parameters tested before each CBIR experiment include sliding windows of 32 × 32, 64 × 64 and 128 × 128 pixels. In all cases, the maximum average distance was obtained with windows of 64 × 64 pixels. Variation of the overlap (0%, 25%, 50%, and 75%) between the sampled blocks did not affect this result. Increasing the overlap provides better localization of the patterns but produces many more sampled blocks, affecting the efficiency of both the feature extraction and the pattern instantiation tasks. Thus, a 50% overlap, i.e. a 32 pixel step, was used as a compromise between localization and efficiency.

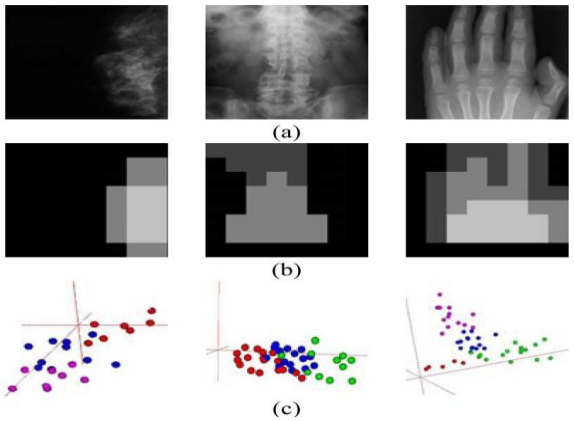


Fig. 2. (a) Original radiographic images, (b) clustering output, and (c) 3-D visual representation of the feature spaces.

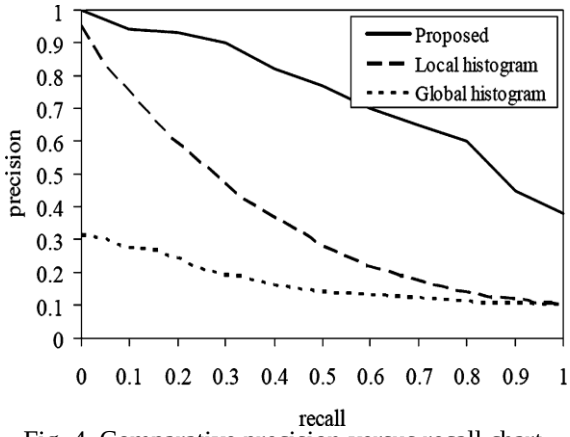


Fig. 4. Comparative precision versus recall chart.

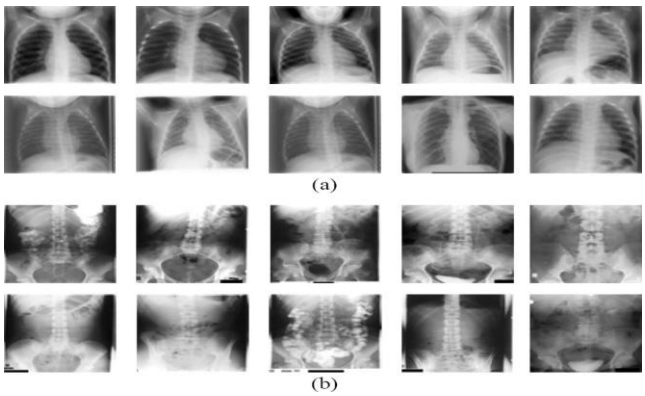


Fig. 5. (a) Query requesting nine chest images similar to the upper-left image (1,1):

All retrieved images belong to the same category; (b) A query requesting nine abdomen-gastrointestinal system images similar to the upper-left image (1,1): all retrieved images belong to the same category, except (1,4) and (2,5), which belong to the abdomen-uropoietic system. (Notation (*i*, *j*) indicates the positioning of an image at the *i*th row, *j*th column in the figure.).

## XII. CONCLUSION

We presented a novel scheme for efficient content-based medical image retrieval. This scheme utilizes rich-in-semantics *pattern* representations of medical images, defined in the context of PANDA, a framework for representing and handling data mining results. The theoretical contributions of this paper are validated by comprehensive experimentation on the IRMA reference collection of radiographic images. 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; and 3) 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. By storing the semantically rich patterns along with low-level features in a unified way, according to the PANDA framework, will enable the extension of the CBIR methodologies with knowledge representation techniques for semantic processing and analysis.

## REFERENCES


[1] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *J. Intell. Inf. Syst.*, vol. 3, pp. 231–262, 1994.

[2] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947–963, Sep. 2001.

[3] T. Deselaers, D. Keysers, and H. Ney, "FIRE—flexible image retrieval engine: ImageCLEF 2004 evaluation," in *Lecture Notes in Computer Science*, vol. 3491, 2004, pp. 688–698.

[4] R. Veltcamp and M. Tanase, "Content-based image retrieval systems: A Survey," Dept. Comput. Sci.Utrecht Univ., Tech. Rep. UU-CS-2000-34, 2000.

[5] H. M¨uller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval Systems in medicine—clinical benefits and future directions," *Int. J. Med. Inf.*, vol. 73, pp. 1–23, 2004.

[6] C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "ASSERT: A physician-in-the-loop content-based image retrieval system for HRCT image databases," *Comput. Vis. Imag. Understand.*, vol. 75, pp. 111–132, 1999.

[7] F. Schnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," *Technol. Health Care*, vol. 8, pp. 291–297, 2000.

[8] W. Cai, D. D. Feng, and R. Fulton, "Content based retrieval of dynamic PET functional images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 4, no. 2, pp. 152–158, Jun. 2000.

[9] D.-M. Kwak, B.-S. Kim, O.-K. Yoon, C.-H. Park, J.-U. Won, and K.-H. Park, "Content-based ultrasound image retrieval using a coarse to fine approach," *Ann. NY Acad. Sci.*, vol. 980, pp. 212–224, 2002.

[10] L. Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, pp. 249–255, Dec. 2003.

[11] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N.Wernick, "Asimilarity learning approach to content-based image retrieval:Application to digital mammography," *IEEE Trans. Med. Imag.*, vol. 23,no. 10, pp. 1233–1244, Oct. 2004.

[12] C.-Y. Lin, J.-X.Yin, X. Gao, J.-Y.Chen, and P.Qin, "Asemantic modelling approach for medical image semantic retrieval using hybrid Bayesiannetworks," in *Proc. 6th Int. Conf. Intell. Syst. Des. Appl. (ISDA)*, 2006,pp. 482–487.

[13] J. Yaoa, I, S. Antani, R. Longb, G. Thoma, and Z. Zhanga, "Automaticmedical image annotation and retrieval using SECC," in *Proc. 19th Int.Symp. Comput.-Based Med. Syst. (CBMS)*, Salt Lake City, UT, Jun. 2006.

[14] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," *IEEE Trans. Inf. Technol.Biomed.*, vol. 11, no. 1, pp. 58–67, Jan. 2007.

[15] H. Greenspan and A. T. Pinhas, "Medical image categorization and retrievalfor PACSusing theGMM-KLframework," *IEEE Trans. Inf. Techol.Biomed.*, vol. 11, no. 2, pp. 190–202, Mar. 2007.

[16] I. Bartolini, P. Ciaccia, I. Ntoutsi, M. Patella, and Y. Theodoridis, "Aunified and flexible framework for comparing simple and complex patterns,"in *Proc. 8th Eur. Conf. Principles Pract. Knowl. Discov. Database (PKDD)*, 2004, pp. 496–499.

[17] R. O. Stehling, A. X. Falcao, and M. A. Nascimento, "An adaptive andefficient clustering-based approach for content-based image retrieval inimage databases," in *Proc. Int. Symp. Database Eng. Appl.*, pp. 356– 365.

[18] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002.
[19] C. Yixin, J. Z. Wang, and R. Krovetz, "CLUE: Cluster-based retrieval ofimages by unsupervised learning," *IEEE Trans. Image Process.*, vol. 14,no. 8, pp. 1187–1201, Aug. 2005.
[20] E. G. M. Petrakis and C. Faloutsos, "Similarity searching inmedical imagedatabases," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 3, pp. 435–447,May/Jun. 1997.
[21] J. K. Wu and A. D. Narasimhalu, "Identifying faces using multiple retrievals,"*IEEE Multimedia*, vol. 1, no. 2, pp. 20–38, Aug. 1994.
[22] W.-M. Jeng and J.-H. Hsiao, "An efficient content based image retrievalsystem using the mesh-of-trees architecture," *J. Inf. Sci. Eng.*, vol. 21,pp. 797–808, 2005.
[23] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval—aquantitative comparison," in *Proc. 26th DAGM Symp., Lect. Notes Comput.Sci.*, 2004, pp. 228–236.
[24] L. Setia, A. Teynor, A. Halawani, and H. Burkhardt, "Image classificationusing cluster-cooccurrence matrices of local relational features," in *Proc.8th ACM Int. Workshop Mult. Inf. Retrieval*, Santa Barbara, CA, 2006,pp. 173–182.
[25] F. S. Bongard and D. Y. Sue, *Current Critical Care: Diagnosis and Treatment*, 2nd ed. New York: McGraw-Hill/Appleton and Lange, 2002.

AUTHORS:

Yamuna.S she is pursuing M.Tech in Instrumentation & Control Systems at SVU College Of Engineering, Tirupati. She completed B.Tech in Electronics & Instrumentation Engineering from Sreenivasa Institute Of Technology And Management Studies, Chittoor.

**Dr.S.Varadarajan**
he did his M.Techfrom NIT,Warangal, and Ph.Dfrom SVU,Tirupathi. His specializations include Signal Processing and digital Communications. He is working as Professor in the department of Electronics and Communication Engineering, Sri VenkateswaraUniversity. He is a fellow of Institution of Electronics and Telecommunication Engineers,India and member of IEEE.

**K. Srinivasa Reddy** is Associate Professor and Head of the Electronics and Communication Engineering, **Nagole Institute of Technology** Hyderabad .He received his B.Tech degree in Electronics and Communication Engineering from JNT University, Hyderabad, and M.Tech degree in Embedded Systems from JNT University, Hyderabad. His areas of interest are Communication Systems, Cellular and Mobile Communication, He is a member of The International Association of Engineers (IAENG).