A Survey on Automatic Classification and Retrieval Techniques for Medical Image Mining

Submitted by: Sugandha

NetID: sugandh2

Abstract: *The use of data mining techniques for medical images in health care is on a rise. Doctors are increasingly relying on mining existing images for computer aided detection (CAD) of new cases. These include X-ray, MRI, CT, mammograms and skin region images of existing cases. Data mining of such existing images can provide relevant knowledge for diagnosis. Some of the challenges being faced in using such techniques are a large data set of available images, low contrasts, inherent noise in medical images, and the problem of organization and retrieval from the image database. These problems can be solved by introduction of efficient medical image classification and retrieval methods. This survey presents a study of latest state of the art methods being used for automatic classification and retrieval of images. It is divided into two main sections: Image Classification and Image Retrieval.*

Introduction:

In cases such as cancer detection, any human error can prove very costly. Moreover, current methods of detection do not provide a 100% certain diagnosis. A false positive in can lead to an unnecessary biopsy, whereas a false negative can lead to a patient’s death. It has been shown that the use of image mining techniques in conjunction with a doctor’s diagnosis can significantly improve the accuracy of the diagnosis, thus reducing costs in terms of patients’ lives as well as health care spending.

This survey discusses latest techniques being used for image classification and retrieval. The general data mining procedures are applicable, but often need some variations to solve domain-specific challenges.

IMAGE CLASSIFICATION

Finding relevant images in a large database is not an easy task. This is further complicated for medical image retrieval by the fact that physicians might often be under time constraints. Moreover, medical image retrieval and classification requires high accuracy and low rate of false positives because of the critical nature of the task. Most hospitals use textual description for retrieving images from a database. This method is not efficient because it depends on individuals' perception and interpretation, and alphanumeric queries are often insufficient to describe an image. Text-based description tends to be incomplete, imprecise, and inconsistent in specifying visual information. This method is also costly and time consuming. This makes automatic categorization of images very important.

Medical image categorization mainly consists to two steps – image preprocessing to convert into a form suitable for classification, and the actual classification process. (Figure 1)

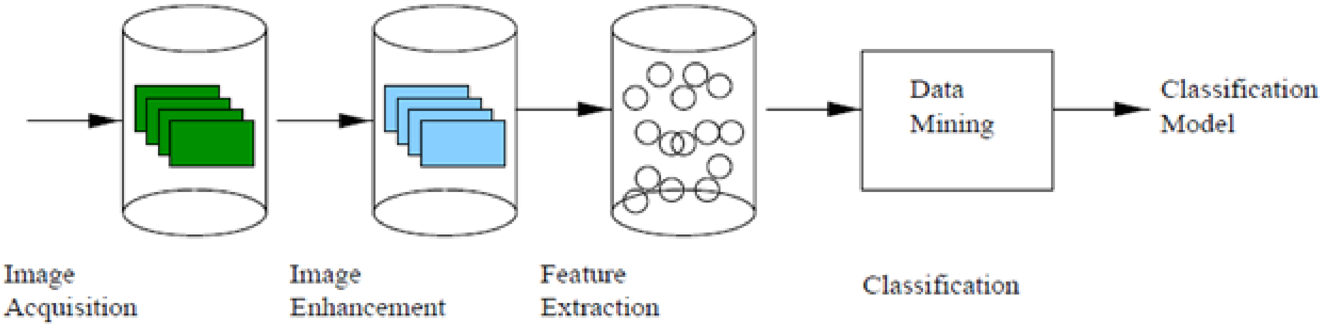


Figure 1: Image Classification Process [4]

Medical Image Pre-classification Steps

1. Image Acquisition

Access to actual medical records even for educational purposes is difficult because of patients' privacy reasons and bureaucratic hurdles. [4] relies on data collected from Mammographic Image Analysis Society for the study on breast cancer imaging. The training set consisted of images belonging to normal, benign and malign categories. They followed a hierarchical classification in the training images, where the abnormal images (with benign and malign tissues) were further classified into microcalcification, circumscribed masses, spiculated masses, ill-defined masses, architectural distortion and asymmetry. Factors such as the location, radius and type of abnormality were also recorded. In general, the quality of a dataset can be measured by the number and variety of images available. The images present must represent all possible categories or labels.

2. Image Preprocessing

Preprocessing of images before classification is necessary for improving the image quality. The goal of medical image pre-processing is to transform the image into a format that can be more easily and effectively classified. It is also needed for 'data cleansing' because the images available might be noisy, incomplete or inconsistent. For digital images, digitization process also introduces some noise into the image. Different illumination condition while clicking the image might also lead to differences between two images of the same object, taken from the same angle. Efficient pre-processing increases the accuracy of statistical inferences used by classifiers.

It is a frequent requirement to convert a 3-D image into 2-D for display and query purposes. Furthermore, images need to be rotated before storing to align their midsagittal planes along a common direction for matching. This rotation about the central axis is measured in terms of 3 angles – *pitch*, *roll* and *yaw*. The pitch angle must be zero for images to be aligned. Maximization of mutual information affine registration algorithm and 2D-to-3D cross correlation registration algorithms have been used in research to minimize the rotational errors among 3D images. Midsagittal extraction can tell us where the midsagittal plane is supposed to be. For example, a brain lesion in a CT scan can cause the midsagittal plane to be distorted (shifted or bended).

Some steps common to pre-processing across all types of images are:

- **Cropping** removes non-relevant area, as well as the noise in background. Cropping may be done before image enhancement as in [4], so that the background noise does not get enhanced.

- **Feature subset Selection** may be used as a way to remove irrelevant information and reduce data dimensionality. The most discriminating set of features must be selected, weighted by the importance of features.
- **Segmentation** may be used to divide the image into smaller areas for processing. It can be very hard to segment abnormalities during pre-processing as they do not have a fixed shape, size or location. Many systems depend on human experts for this initial labeling.

Recent Advances in Segmentation Techniques

- Active Shape Modeling

A technique known as *Active Shape Modeling* (ASM) has been shown to efficiently segment irregular, noisy images. ASM works with two models of the image to be segmented – a shape model and a grayscale model. The shape model describes the target image shape and can be used to derive constraints on the range of shapes to which the deformed template may converge. The grayscale model is used to obtain this deformation from the initial shape and alignment of database images, based on the constraints imposed by shape model. Both models are based on what the target image is expected to look like, and are derived from the input image. These models are statistical in nature. Initial research into this method was carried out by researchers at Texas Tech University using 80-point vertebral models built from each of 40 C-spine images. For each of the 40 images, shape and grayscale models were constructed based on the remaining 39 images and leave-one-out testing was carried out for performance evaluation. Mean Squared Error (MSE) between the converged ASM boundary and a manually crated real boundary was used as the metric for performance. This testing was further carried out for two different poses (position, orientation, shape) – once with the pose calculated as an average of the remaining 39 images, and the second time, line integral processing method was used to estimate spine region location and orientation. In 35 out of 40 cases, the second run gave lower MSE. Conclusions of this work were that (1) significant improvement in ASM performance on these images is feasible with an automated template placement method and (2) some anomalous cases exist where even a good initial pose will not assure good ASM performance.

Latest improvements in this algorithm involve improvements in both grayscale and shape models. The grayscale image is preprocessed with edge enhancement technology to improve the informational content. In a test of 100 NHANES II x-ray images, it has been shown that the ASM convergence accuracy obtained with this new grayscale model is clearly improved over the old grayscale model, and in fact is comparable to the accuracy obtained when ASM is applied to a computer-generated vertebra [13].

Landmark Tool in MATLAB can be used to collect boundary landmarks and generate the corresponding grayscale and gradient values.

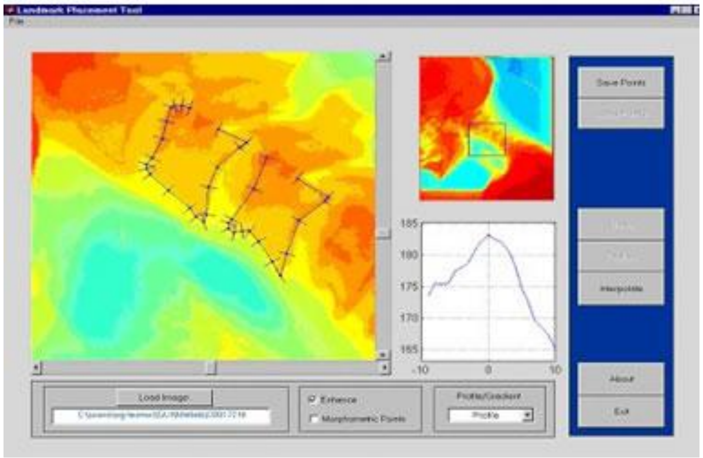


Figure 2: MATLAB Landmark tool for segmentation (Image depicts its use in segmentation of spine vertebrae)

- <u>Active contours</u>

This is a framework for delineating image outlines. It works by creating a model of the object to be segmented from a shape template and generating a search grid in the neighborhood of this template boundary, and mathematically minimizing an energy function over all feasible curves within this grid. The incorporation of dynamic programming can make this optimization step very fast. Disadvantages of this segmentation method are that this method is highly local in nature and thus strongly dependent on good initial placement of the template and certain boundaries (e.g. spinal vertebra boundaries) are sometimes missed because they are close to stronger edges formed by tissue/background interfaces.

- <u>Unified segmentation</u>

This technique uses Generalized Hough Transform (GHT) for template pose initialization. Hough Transform is used to detect straight line in images. GHT is a variation that allows the detection of arbitrary shapes. It scans the entire image, while varying the scale, position and orientation of the input image, to find the best match using the Hough 'bin' counting method. Authors of [13] have used this method to match a vertebral template to the "truth" vertebral pose, up to possible shifts of the template along the spine by integral numbers of vertebrae. The results obtained were robust, but suffered from slight inaccuracies, such as a template created from vertebrae C2-C5 might be matched by the GHT to C3-C6 instead. Current research is investigating methods to increase the accuracy of the matching and to reduce the computation time.

- **Data registration** is to bring images into a common coordinate system**.** Data acquired by recording the same subject at different times and from different perspectives may in different coordinate systems. Registration is a data calibration process that transforms these different sets of data both spatially and temporally, in order to be able to compare or model different images.
- **Image enhancement** accentuates the areas of interest. Reduces the effect of varying illumination conditions by improving image contrast. This is done using a technique called histogram equalization. Histogram equalization makes dark regions darker and light regions lighter. This helps to standardize the images to an extent, for further processing.
- **Feature extraction** extracts relevant features and stores them, to be used as an input to classification algorithms. In general, four parameters can be computed for the evaluation of extracted features:

i) mean: the average of available data points

ii) variance: the spread of data points

iii) skewness: the measure of assymmetry

iv) kurtosis: the peakedness of an object

Medical Image Classification Algorithms

This section discusses some commonly used classification algorithms for medical image classification.

1) **Neural networks with back propagation**

Neural networks consist of nodes connected in several layers (input, output, hidden), with associated weights that are tuned in the training phase to achieve a high performance. Back propagation algorithm approximates steepest descent algorithm and works by minimizing least mean squares error. Authors in [4] conducted an experiment to classify mammograms as 'normal' or 'abnormal' using a neural network consisting of 69 input nodes, one hidden layer with 10 nodes and one output node.

The classifier was found to perform well on an average, when compared to other studies, but the classification success ratio ranged from 65.6% to 93.7% depending on database split. This inconsistency makes the method nonviable in real life applications. Therefore, this method can be helpful for initial categorization, but cannot be taken as a reliable classifier.

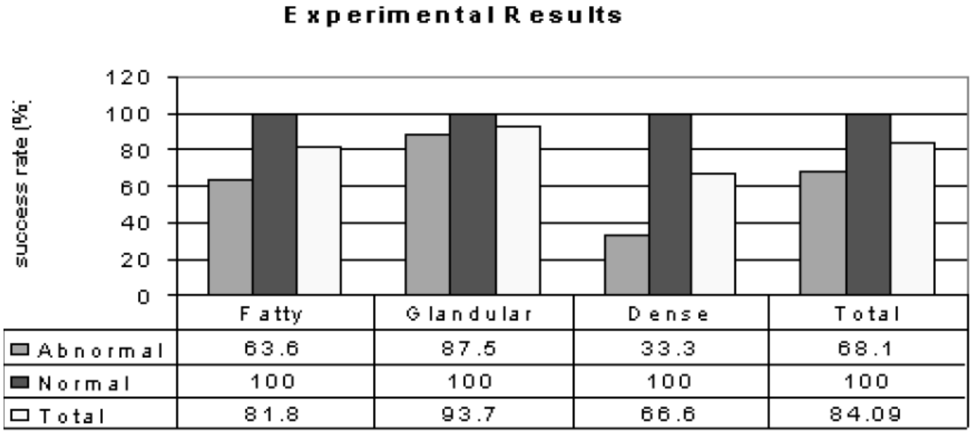
2) **Association Rule mining**

Authors in [4] have used association rule mining using apriori algorithm to classify mammograms in suspected breast cancer cases. The first step in association rule discovery is frequent itemset discovery. An association rule is a rule of the form A=>B where A and B are transactions in a database D, and A and B have no common items. Such a rule is derived if the conditional probability of B given A is higher than a predefined threshold (called confidence). The association rules discovered were of the form:

(Set of features of mammogram) => Cancer Category

For a given category (normal or abnormal), the authors worked at discovering all possible combinations of features from the mammogram features database that may imply this category, and the association rules to validate these. These association rules constitute the classifier model.

When a new mammogram is to be classified, the categorization system extracts the image features and identifies the association rules that apply to that image. Some systems assign images to a category for which the most number of association rules apply, but this will work only if the number of association rules discovered per class is balanced. A tuning of the classifiers has to be done for this purpose. This means finding some optimal confidence intervals that



	Fatty	Glandular	Dense	Total
▨ Abnormal	63.6	87.5	33.3	68.1
■ Normal	100	100	100	100
□ Total	81.8	93.7	66.6	84.09

Figure 3: Success Rates of Association Rule Mining Classifier [4]

maximize the recognition rate. Maximization of recognition of abnormal cases is taken as the criterion in medical research, because the rate of false negatives needs to be minimized. A system should rather give a false positive than a false negative (i.e. classify abnormal tissue as normal).

The classification results of using this algorithm on 44 test images – 22 normal and 22 abnormal are given in Figure 3. Abnormal cases can be further subdivided according to tissue type as: fatty (11 cases), fatty-glandular (8 cases) and dense (3 cases). The overall success rate for classification was found to be 78.69%.

3) **ID3 Decision Tree**

Authors in [10] have used this algorithm for classification of lung X-rays. It works by choosing the most informative attribute at each step, so that the expected number of tests needed for classification is minimized. The authors also considered entropy as one of the deciding factors at this stage, to minimize noise and chaos. Suppose S is a set of training examples, and C_j, j=1,2,.....,n is the set of decision classes. A decision tree is constructed by repeatedly calling the decision tree algorithm in each generated node of the tree. The most informative attribute Ai, is selected as the root of the subtree and the current training set S is split into subsets Si according to the values of the most informative attribute. Recursively, a subtree T_i is built for each subset S_j. Tree construction stops when all examples in a node are of the same class. Each leaf node generated is labeled by a value of exactly one class variable. However, leaves can also be empty, if there are no training examples having attribute values that would lead to a leaf or can be labeled by more than one class name if there are training examples with same attribute values and different class name. For the classification of lung x-rays, different feature subsets were created by selecting the most dissimilar features. Figure 4 shows the resulting decision tree obtained:

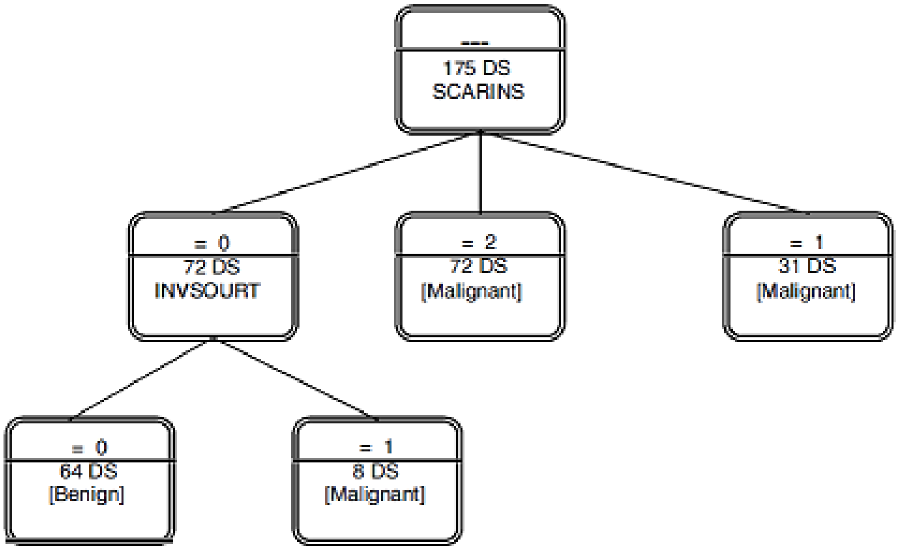


Figure 4: Decision Tree representing classification of lung x-rays [10]

The split attributes from the tree are: SCARINS – representing scar like changes inside the node, and INVSOURT – representing invasion into surrounding tissues.

4) Bayesian Classifiers

Authors of [11] used a Memory Based Learning (MBL) technique called Kernel Regression (KR). Posterior Probability of an image belonging to a class C, when feature x is observed can be computed by Bayes' Law as:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x}|c)P(c) + P(\mathbf{x}|\bar{c})P(\bar{c})}$$

The prior probability P(c) of a class c can be estimated from labeled training data by dividing Nc, the number of instances in class c, by the total number of instances N in the training set: P(c) ~ Nc/N. Changing the kernel width in KR density estimation or the distance metric (feature weighting) gives the whole space of classifiers. The approach consists of finding classifiers that minimize cross-entropy, since cross entropy is the negative log-likelihood of the training data given a specific classifier, and is therefore a measure for how well a classifier performs. Minimizing this cross-entropy will yield a metric and a kernel width for which kernel regression best approximates the posterior probabilities and is thus optimally suited for classification. An experiment with 48 CT brain scan images classified images based on features such as density, shape and location of lesions.

Image retrieval by matching depends directly on classification, and the classification technique used must aid retrieval as well. The same distance metric that is used for classification can also be used for efficient retrieval.

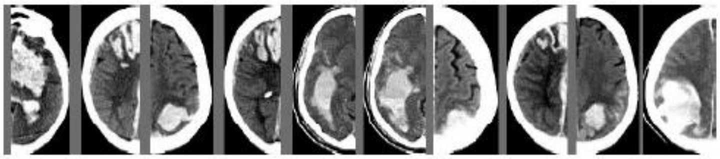


Figure 5: Retrieval results using Bayesian classifier to identify similarity. Leftmost image is the query image and the remaining 9 images are retrieval results based on decreasing order of similarity.

IMAGE RETRIEVAL

Medical image retrieval conventionally depends on manual indexing and text queries that match these keywords or tags.

Latest advances have made it possible to retrieve medical images by giving the input query as an image and finding a match from the database through the computation of a similarity measure, such as k nearest neighbors, a technique used in [11]. Such a methodology to retrieve images is known as *Content Based Image Retrieval*.

Content Based Image Retrieval

With the advances in database and multimedia technologies, it has become possible to store a large number of images. For medical images, a doctor would like to retrieve images with certain types of symptoms as his present case, e.g. a lesion in a particular location. For this we need pattern recognition and image matching techniques as well as a way for the doctor to describe the image. Content Based Image Retrieval (CBIR) is a popular method that solves these challenges. Based on the semantics of input picture, it generates a description that can be used as an index to retrieve relevant images from the database. CBIR is also known as Query by Image Content (QBIC) and Content Based Visual Information retrieval (CBVIR). CBIR relies on the actual 'content' of image, as opposed to metadata (such as keywords or tags assigned by humans) about it, for retrieval. It utilizes context free grammars and graph grammars to generate a description based on image contents. 'Intelligent algorithms' that are beyond conventional human indexing capabilities are used to generate these descriptions and indexes. They include capabilities such a non-textual input queries and return their output based on similarity, or fuzzy membership in a category. The features used by CBIR systems to identify images can be classified as primitive (e.g. color and texture), logical and abstract (e.g. significance of features).

CBIR is particularly useful in the area of radiology, because radiologists often refer to existing HRCT (High Resolution Computed Tomography) scans for diagnosis. The aim of CBIR in medicine is to enable the doctors to retrieve an image with known pathologies that are similar to the patient's scan. Apart from medical images, CBIR is used in other fields such as retrieval of images of art work, postage stamps etc.

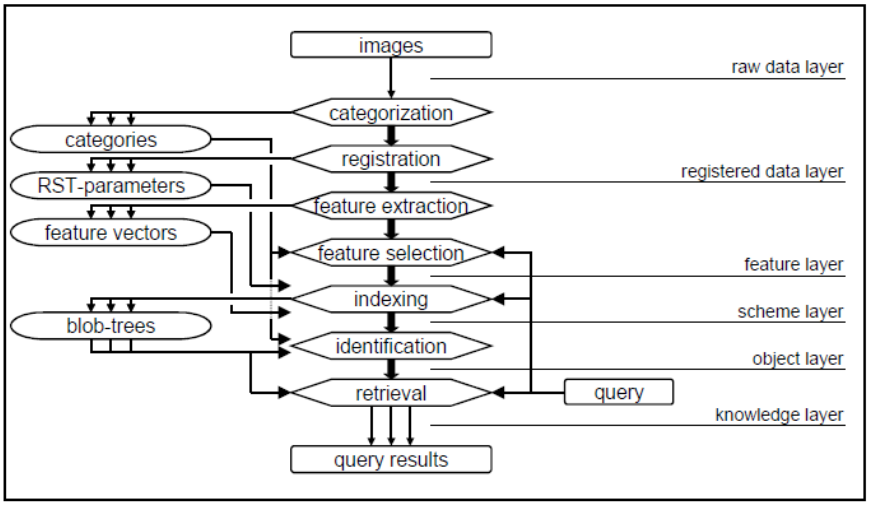


Figure 6: CBIR Processing Steps and Semantic Layers [9]

A CBIR system is able to answer queries such as,

Query 1: *Find patients with MRI scans similar to patient ID '1123'*

Query 2: *Retrieve the image frames in which a micro-lesion is nearby the lateral ventricle and approximately 9 mm in diameter. The micro-lesion evolves into a macro-lesion with diameter equal or larger than 25 mm and invades the lateral ventricle in approximate one year.* [14]

1. Feature Vector Approach to CBIR

CBIR uses a feature vector to represent an image in the database. For retrieval against an input image, a similar feature vector is computed for the input image and compared against the vectors in database. A distance measure such as Euclidean distance gives the similarity measure for retrieval. In medical terms, this similarity means a similarity in disease location, type and severity, and therefore the treatment. This is called feature vector approach to CBIR. The use of high dimensional feature spaces can lead to overwhelming storage and processing capabilities (curse of dimensionality) and the choice of distance measurement must be made cautiously.

2. Customized Queries Approach to CBIR

Depending on the disease class, CBIR only utilizes the features that best discriminate images within that class. It works as a two tier approach, consisting of two steps – identification of class and subclass. The first step is query classification, according to the class labels of images, and using the features that best discriminate the classes. The second step then retrieves the most similar images within the predicted class using the features customized to distinguish "subclasses" within the selected class. For example, for an HRTC of a lung, the first step will identify the features that most accurately describe the disease class and give the highest classification accuracy, and the second step will look for images with similar disease location, type etc. [2] found that most accurate classifiers for first level classification is the C5.0 boosting algorithm that utilizes decision trees. Second level classification is the problem of feature selection using unsupervised (unlabeled) data, because there is no prior listing of subclass labels within a given class. *FSEM* (Feature Subset Selection using Expectation Maximization) clustering algorithm is used for this purpose. This algorithm searches through the feature space to locate every feature subset F_t by EM clustering and evaluating the clusters formed based on a feature selection criterion. The search converges in a feature subset that maximizes a predefined criterion.

Search Method – For d features, a total of 2^d feature subsets are possible. Exhaustive search through all subsets is intractable. A greedy search algorithm called Sequential Forward Selection (SFS) can be used. It starts with a single feature and sequentially adds the feature that best optimizes the search, when considered along with the existing selected features. It does not lead to a globally optimal solution, but reduces the complexity to $O(d^2)$.

Clustering Algorithm – EM algorithm is used as the clustering method in CQA approach to estimate the maximum likelihood model parameters and return the clusters, along with their associated probabilities (soft clusters). Choosing a suitable initial value is important to prevent the algorithm from being stuck at a local maximum.

Feature Evaluation Criterion – The feature sets obtained are evaluated in terms of the compactness and separatability of clusters. This means that intra cluster distances should be small and inter cluster distances should be large.

Picture Archiving and Communication system (PACS) Integrated CBIR

PACS is a medical imaging system that provides good quality image storage and retrieval capability in a fast and efficient way, at the same time, eliminating the need to manually tag and index images. PACS images are stored in DICOM (Digital Imaging and Communications in Medicine) format. PACS by itself uses only alphanumerical descriptions of study, patient, and technical parameters for image retrieval, but its integration with CBIR systems provides combined benefits of both systems.

PACS consists of imaging modalities (like CT, X-Ray, MRI), a communication network, workstations and storage and retrieval capabilities.

Integration of CBIR with PACS can be done on different levels, and for standardization, both components must use protocols that are independent of underlying CBIR techniques being used, and the system hardware. The integration scheme simply considers CBIR as one of the components connected by the PACS network, along with the other three components mentioned in the last paragraph. Figure 7 represents the integrated architecture. The storage location remains unchanged and PACS functions as a service provider component that answers queries made by IRMA (Image Retrieval in Medical Applications) core. PACS core provides and API and a unique DICOM identifier (ID) of the image currently being displayed. IRMA functionality (like feature extraction in the image) are managed through this ID. This API can be invoked through an IRMA option in the PACS menu. Selecting this option passes the image ID to IRMA core. If any human labeling and categorization is required by the CBIR system being used, the IRMA handle initiates the IRMA server and GUI window for data entry. Thus, the CBIR system is loosely coupled with PACS viewing and reporting system and the integration is not affected by the CBIR algorithms being used.

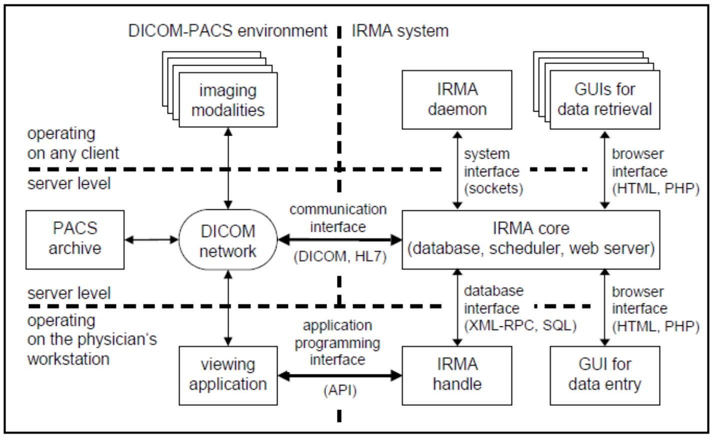


Figure 7: IRMA PACS Integration [9]

Knowledge Based Image Retrieval

[14] presents a framework where a physician can input a visual query (e.g. CT scan to be matched), along with additional constraints to be considered while retrieval. This model is called *Knowledge Based Semantic Temporal Image Model*.

An image model is proposed which consists of a Raw Data Layer (RDL), Feature and Content Layer (FCL), a Schema Layer (SL), and a Knowledge Layer (KL).

RDL contains abstracts image storage details such as format and compression from rest of the upper layers.

FCL stores image features such as contours, spatial relationship characteristics, and temporal sequences. Spatial feature computation provides for shape and spatial relationship modeling. Shape modeling is a decomposition approach to describe an object's shape and is used to match contoured objects. Spatial relationship information is based on relationship features described by domain experts. For example, the spatial relationship for a lesion that is near another object can be captured using the distance of the centroids of the two contours on the x-axis and y-axis, the angle of coverage (the angle for viewing a contour from the centroid of another contour), and the ratio of area to classify the spatial relationship [14].

SL can contain visual as well as stream entities (whose sequence over a period of time is stored- e.g. progressive stages of lung cancer).

KL contains a hierarchical of images and their relationships as a structure called Type Abstraction Hierarchy (TAH). Higher nodes in the TAH represent more generalized concepts (i.e., wider range of feature values) than that of the lower nodes (i.e. narrower range of feature values).

[14] also introduces a Visual Query Language that can be used to make spatial and temporal queries in this framework.

Figure 8 shows an example of a visual query expression.

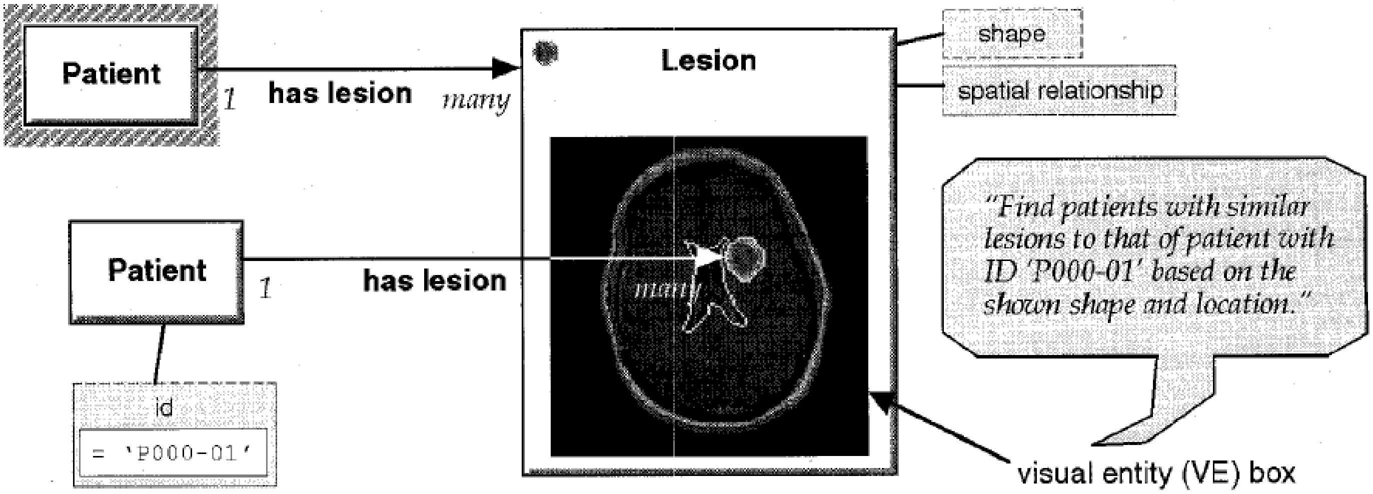

Patient
1
has lesion
many
Lesion
shape
spatial relationship
"Find patients with similar lesions to that of patient with ID 'P000-01' based on the shown shape and location."
Patient
1
has lesion
many
id
= 'P000-01'
visual entity (VE) box

Figure 8: A visual query expression to retrieve CT scans with a similar lesion. [14]

Conclusion:

Research in the past decade has led to significant advances in the field of medical image classification and retrieval. Image classification and retrieval must be preceded by an image pre-processing step which includes image segmentation, image enhancement and feature extraction. Techniques such as neural networks, associative rule mining and Bayes classification can be used to build classifiers. Retrieval systems employ similar classification techniques to match images in a medical database with the input image. Such systems range from text-query description based retrieval to more sophisticated content based retrieval. A mixed model that combines both modes of retrieval and supports temporal queries has also been proposed [14]. Due to a very high level of accuracy required, medical image classification and retrieval continues to be an area of active research.

References:

[1] W. Moudani, A.R. Sayed, "Efficient Image Classification using Data Mining," IJCOPI Vol. 2, No. 1, Jan-April 2011

[2] J.G. Dy, C.E. Brodley, Lynn S. Broderick, A.M. Aisen, "Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 3

[3] D. Keysers, J. Dahmen, H. Ney, B.B. Wein, T.M. Lehmann, "Statisical Framework for Model-Based Image Retrieval in Medical Applications," Journal of Electronic Imaging 12(1), 59–68 (January 2003).

[4] M.L. Antonie, O.R. Zaiane and A. Coman, "Application of Data Mining Techniques for Medical Image Classification," In. Proc Second International Workshop on Multimedia Data Mining

[5] Z.S. Zubi, R. A. Saad, "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer"

[6] P. Rajendran, M. Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm," In Journal of Computing, Volume 2, Issue 1, January 2010

[7] H. Muller, N. Michoux, David Bandon and Antoine Geissbuhler, "A Review of Content Based Image Retrieval Systems in Medical Applications- Clinical Benefits and Future Directions," IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004.

[8] N. M. Sirakov and P. A. Misna, "Search Space Partitioning using Convex Hull and Concavity Features for Fast Medical Image Retrieval,"

[9] T.M. Lehmann, M.O. Guld, C. Thies, B. Fischer, D. Keysers, M. Kohnen, H. Schubert, B.B. Wein , "Content Based Image Retrieval in Medical Applications for Picture Archiving and Communication Systems," Proceedings Vol. 5033, Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation

[10] S.M. Khan, M.R. Islam, M.U. Chowdhury, "Medical Image Classification Using an Efficient Data Mining Technique," in Proceedings of International Conference on Machine Learning and Applications (ICMLA'04), Louisville, KY, USA, December 2004.

[11] Y. Liu and F. Dellaert "Classification Driven Medical Image Retrieval"

[12] M. R. Ogiela and R. Tadeusiewicz ,“Semantic Oriented Syntactic Algorithms for Content Recognition and Understanding of Images in Medical Databases” in IEEE International Conference on Multimedia and Expo, 2001

[13] L. R. Long, S. Antani, D. J. Lee, D. M. Krainak, G. R. Thoma. “Biomedical Information from a National Collection of Spine X-rays: Film to Content Based Retrieval”

[14] W. Chu, C. Hsu, A. Cardenas, R. Taira. “Knowledge Based Image Retrieval with Spatial and Temporal Constructs,”