Classification Driven Semantic Based Medical Image Indexing and Retrieval

Yanxi Liu, Frank Dellaert and William E. Rothfus[a]

CMU-RI-TR-98-25

[a] Associate Professor of Radiologic Sciences, Diagnostic Radiology, Allegheny University of the Health Sciences (Allegheny General Hospital)

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This research is sponsored by the Allegheny-Singer Research Institute under prime contract through the Advanced Technology Program of the National Institute of Standards and Technology (NIST#70NANB5H1183). Views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of the National Institute of Standards and Technology or the United States Government.

ABSTRACT

The motivation for our work is to develop a truly semantic-based image retrieval system that can discriminate between images differing only through subtle, domain-specific cues, which is a characteristic feature of many medical images. We propose a novel image retrieval framework centered around classification driven search for a good similarity metric (image index features) based on the image semantics rather than on appearance.

Given a semantically well-defined image set, we argue that image classification and image retrieval share fundamentally the same goal. Thus, the distance metric defining a classifier which performs well on the data should be expected to behave well when used as the similarity metric for semantic image retrieval. In this paper we shall report our methodology and results on 3D grey-level medical image retrieval.

1 Introduction

1.1 Motivation

On-line image data is expanding rapidly in quantity, content and dimension. Many medical Picture Archiving and Communication Systems (PACS), for example, contain several terabytes of on-line medical image data [20]. However, the utilization of such data for research and education is limited by the lack of intelligent, effective retrieval capabilities [54]. The role of PACS has been noted as digital image archiving, since only simple retrieval using patient names or ID numbers is supported. Using text alone for medical database retrieval has been traditionally the dominating approach for database access. However, text-based methods are limited by their predefined vocabularies, which can be incomplete, coarse, subjective and/or ambiguous [44]. Since medical images form an essential and inseparable component of diagnosis, intervention and patient follow-ups, it is natural to use medical images as front-end index to retrieve relevant medical cases.

Existing "content-based" image retrieval systems, for example [13, 17, 36, 41, 35, 33, 38], depend on general visual properties such as color and texture to classify diverse, two-dimensional (2D) images. These general visual cues often fail to be effective discriminators for image sets taken within a single domain, where images have subtle, domain-specific differences. Furthermore, these global statistical color and texture measures do not necessarily reflect the meaning of an image, i.e. the image semantics. Experts in the content-based image retrieval community[1] have repeatedly pointed out the lack of objective evaluation standards for retrieval of general color images. Such standards are usually easy to establish in domain specific images.

Databases composed of (3D volumetric or 2D) images and their collateral information in a particular medical domain provide simple, semantically well-defined training sets, where the semantics of each image is the pathological classification indicated by that image (for example, normal, hemorrhage, stroke or tumor in neural radiology images). During system performance evaluation, the retrieval results are readily quantified by comparing the pathology (sub)classes in the query image and the retrieved images. As a test-bed, we have chosen a multimedia digital database in human neurology from the National Medical Practice Knowledge Bank project [24]. The medical knowledge bank is a database containing a large set of typical and rare medical cases in different sub-branches of medicine. Neurology is the current focus of the knowledge bank. Each case in the database is composed of at least one 3D image, either Computed Tomography (CT) or Magnetic Resonance images (MRI). In addition to images, other collateral information is also

[1] sources are from IEEE Content-based video and image retrieval workshop in conjunction with CVPR97, June 1997; IEEE Content-Based Access of Image and Video Libraries in conjunction with ICCV98, January 1998; and IEEE Content-Based Access of Image and Video Libraries in conjunction with CVPR98, June 1998

included in the case: the patient's age, sex, symptoms, test results, diagnosis, surgical operations, treatments, and outcomes in text, speech, or video formats.

As an illustration of the kind of semantic selectiveness, i.e. similarity based on medical/anatomical relevance, that our work is striving for, we show an example of a desired semantic image retrieval system. A query case presented as a 3D brain scan appears to the left of Figure 1. To the right appears the highest-ranking retrieved case. The same query case (left) and the second highest ranked case are shown in Figure 2. Note, first, that both retrieved cases have the same pathology as the query: acute bleed. Second, although acute bleeds in the query and the retrieved cases appear on opposite sides of the brain in Figure 1, they actually belong to the same brain anatomical structure, which has a high degrees of medical relevance, while the second ranked retrieved case (Figure 2) has an acute bleed on the same side of the brain as the query but in a different anatomical location, which is relatively less relevant medically than the top retrieved case.

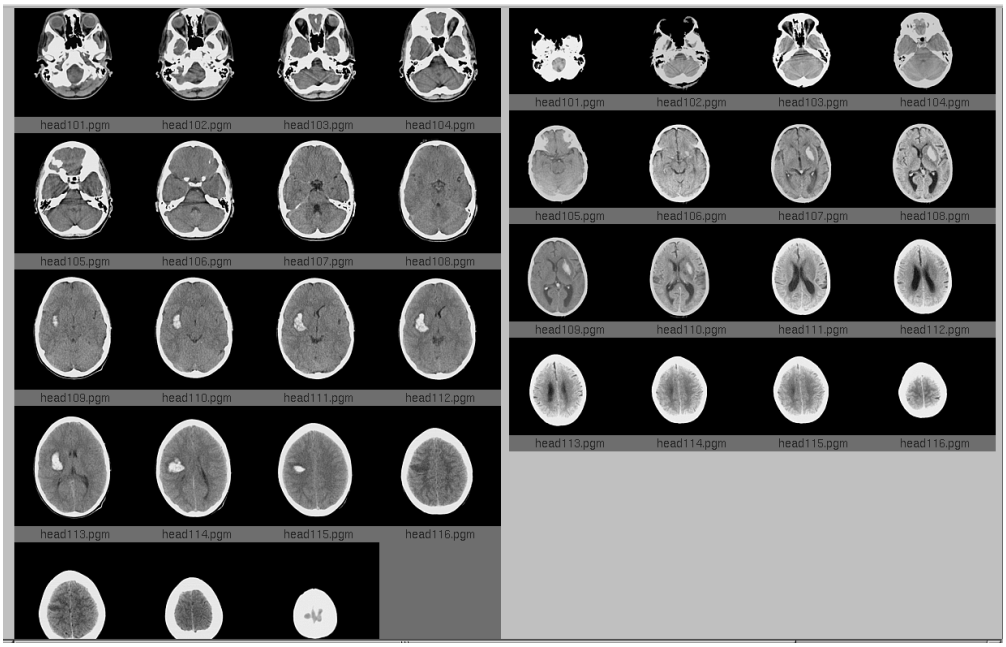

Figure 1: Left: query case, which is a right basal ganglion acute bleed. Right: top-ranked retrieved case, which is a left basal ganglion acute bleed. Although the bleeds appear on opposite sides of the brain, they belong to the same brain anatomical structure.

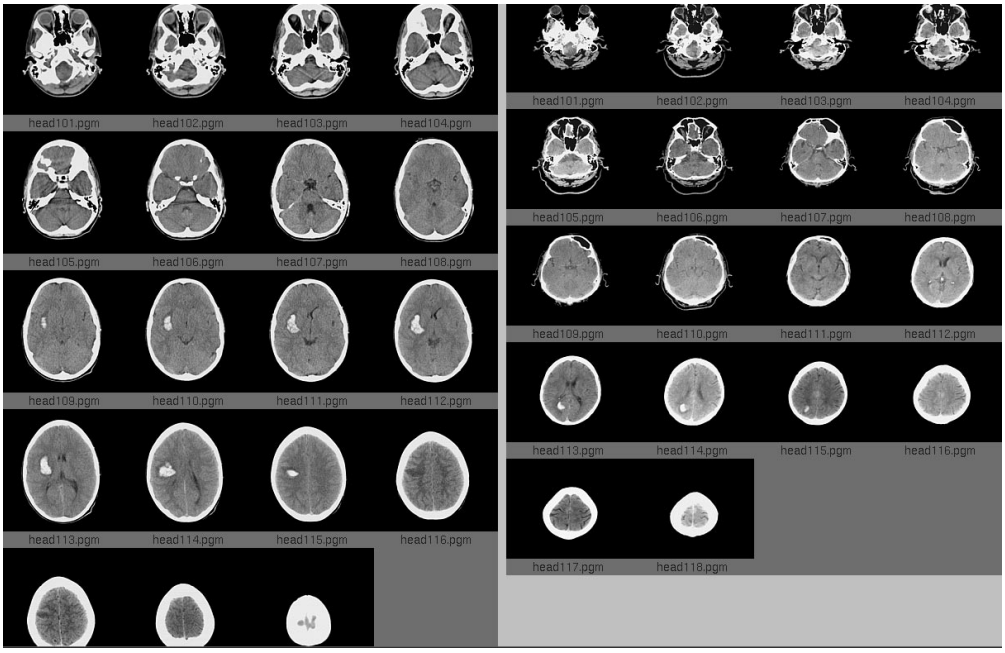

Figure 2: Left: same query as in the previous Figure. Right: second highest ranked retrieved case, indicating a right occipital/parietal acute bleed.

1.2 Semantic Image Retrieval

Semantic image retrieval is a hard problem. Common practice in the image retrieval and pattern recognition community is to map each image into a set of numerical or symbolic attributes called *features*, and then to find a mapping from feature space to image classes or labelings. To describe the semantic image retrieval problem more precisely, let us define f and p as mappings from images to image features (attributes) and from images to image meanings respectively (Figure 3). In order for a mapping from image features to image meanings to exist, it is mathematically **necessary** that (1). p be a function (if $p(x) = y1$ and $p(x) = y2$ then $y1 = y2$) and **sufficient** that (2). f be an invertible mapping (if $f(x) = y$ then $f^{-1}(y) = x$). Note, the second condition is not necessary and can rarely be satisfied in practice.

It is important to distinguish between a well-defined but hard problem (when p is a function) and an ill-defined problem (when p is not a function). For general images such as color scenery photos, the first requirement is difficult to satisfy. Often a general color photo image can have multiple or ambiguous subjective meanings (p is not a function). Therefore we propose the concept of a **semantically well-defined image set**:

- There is a finite number of possible distinct image classes (no arbitrariness) defined in the set.

- Any pair of image classes is mutually exclusive (no ambiguity).
- Each image corresponds to one class and one class only (p is a function).

Given such a semantically well-defined image set, semantic image retrieval becomes *finding a mapping f from images to image attributes such that the probability of mapping images to their respective classes via their extracted features can be maximized*. It is desirable that the mapping f can map images who belong to the same semantic class to the same or nearby feature points in a possibly multidimensional feature space (Figure 3). For example, if the image feature values for each image semantics class have a normal distribution in the feature space then the mean and the standard deviation of their values will become the discriminating characteristics for this image class, and can be mapped into corresponding image semantics straightforwardly. This problem set-up naturally leads to the use of Bayes law to compute the probability $P(c_j|[x_{i_1}, x_{i_2}, ...])$ of the most likely semantic class c_j given properly weighted image features $[x_{i_1}, x_{i_2}, ...]$.

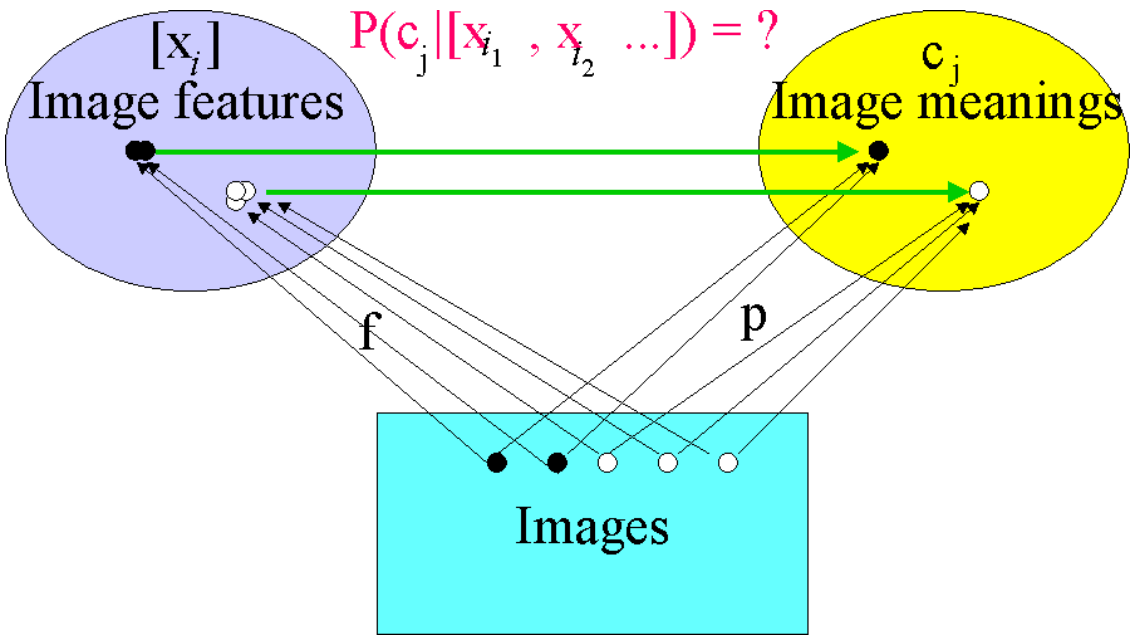

Figure 3: Given a semantically well-defined image set, semantic image retrieval becomes finding a mapping f from images to image attributes such that the probability of mapping images to their respective image classes via their extracted features can be maximized

1.3 Our Approach

The goal of this work is to "warp" the original multidimensional feature space, by putting different weights w_i on different feature dimensions. Such that the points corresponding to the same semantic

class are well-clustered in the warped feature space. Note, the feature weight w_i is a value between zero and one *inclusive,* i.e. $w_i \in [0, 1]$. By **feature selection** we mean the process of assigning 0-to-1-inclusive weights to a set of features. We define *similarity metric* to be a Euclidean distance metric defined in a non-uniformly scaled image feature space. The non-uniform scale on each feature dimension is equivalent to the 0-to1-inclusive weight for that feature, which is found while evaluating how well the features perform in image classification. In this paper we use the terms similarity metric, distance metric, and feature selection interchangeably.

The basic idea behind our approach is that, for a semantically well-defined image set, image classification and image retrieval share fundamentally the same goal, and thus they should also share the same similarity measurement and performance evaluation standards. Our hypothesis is that *a classifier that gives good classification results should also be expected to provide good image retrieval results when its distance metric is used as the image similarity metric for retrieval*. Therefore, in our work image classification is used as a similarity metric selection tool for semantic image retrieval.

Our ideas take concrete form in a medical image retrieval system for pathological neurological images. Instead of precisely locating lesions interactively or automatically by sophisticated segmentation or registration algorithms, we take a completely different approach: based on the assumption of bilateral symmetry existing in normal brains (but absent in pathological brains), a pool of candidate image features is extracted. An optimally weighted subset of these potentially discriminating features is then selected by evaluating how well they perform on the task of classifying medical images according to predefined pathological classes (semantics). Finally, the subset of the initial features that has the best performance in classification is used as a similarity metric for image retrieval.

1.4 Overview

In this paper we report on a semantic image retrieval system framework based on our ideas and the initial results of an implemented system. Section 2 reviews related work in image retrieval and medical image analysis domains. Section 3 describes in detail our approach of classification-driven similarity selection for semantic image retrieval. Section 4 reports experimental results. Section 5 discusses several important issues raised by the experiments. Section 6 concludes the paper with an analysis of the overall result and a discussion of future research.

2 Related Work

2.1 Content-Based Image Retrieval

Most content-based image retrieval work uses general image features such as color, texture, or line-segment histograms to classify diverse, two-dimensional (2D) images. Much recent work has focused on fine tuning of a user's query based on user input or feedback [38, 21, 10]. The two main methods in the literature are (1) query-point movement and (b) axis re-weighting. To compare our work with the Rosetta system of [10], for example, one can see that in [10] the semantic category is defined on the fly by a *set* of query images, and the statistics of the query set are used to construct a metric where the weights are determined by the inverse of the standard deviation of the feature values. In contrast, our approach relies on the *a priori* analysis of the database in terms of pathology semantic categories, and the database can be queried using a single 3D or 2D image. In addition, our search method can eliminate non-discriminative features that are common to all categories in the image database, while Rosetta does not make such a distinction. Therefore, the feature value that is common to the images in the query set (with small standard deviation value) can also be common to the rest of the images in the database thus containing little discriminating power. Our approach is better suited for application in specific databases where the differences between semantic categories are subtle, as in the medical imaging domain.

Image understanding based on image semantics has been an elusive goal of computer vision for many years. Content-based image retrieval systems based on visual feature syntax, for example [37, 33], give users the illusion that the system can understand the meaning of a query image: when the query image contains a lion, some lion pictures are returned. These systems are actually manipulating the syntax of an image exclusively. It is rather an accident than by design when images of the same semantics as the query image are retrieved. Several recent attempts have been made on classification of images based on image semantics of general color photos [43, 50], or videos [51]. It is, however, hard to evaluate given the problem being solved is not well-defined. As pointed out by the author of [50] "On Image Classification: City vs. Landscape" paper that "*The labels for the training set are assigned by a human subject. There are many images that can be placed into both (of) these categories ...*". On the contrary, Moghaddam et al's work on probabilistic matching for face recognition [29] has explicitly defined *two distinct and mutually exclusive classes*, thus their algorithm works on a semantically well-defined image set as we defined above, and the results can be easily evaluated.

Forsyth and Fleck have proposed "body plans" [15] as a semantic representation of specific objects. A hierarchical kinematic model of human bodies or horses from various viewing perspectives is constructed. First a skin (hide) filter is used to segment the image and then the segmented

image is matched against the predefined body plans to find possible corresponding articulated assemblies represented by the body plans. The hope is to use the body plans leading eventually to the recognition of human bodies or horses in a 2D colored image. Two distinct classifiers are built for finding human or horses. The best reported results so far are a recall rate of 42% (the proportion of relevant material actually retrieved in answer to a search request), 182 false positives and a response ratio 10 (a point on its ROC curve: true positive rate over false positive rate) for human on an image set of 565 targets and 4302 controls, and 15% recall with a response rate of 23 for horses on an image set of 100 targets and 1086 controls. It is not clear how many of these body plans are need to be hand built for better classification performance. Besides, this method cannot be applied if one does not know what possible shape variations the object to be recognized is going to take, such as the case in varied lesion shapes, densities and locations in medical images.

There is a large body of work on using eigenspace representations for object recognition in computer vision, especially for face recognition [49]. The basic idea is to project a high dimensional image feature space into a smaller subspace. Principal Component Analysis (PCA) is a mathematical method to find typical components and their multipliers from the original data. Normally in PCA, a covariance matrix C of the original data is created and eigenvectors and eigenvalues of the matrix are computed. The transformation is done by choosing several of those eigenvectors that ensure preserving as much information as possible (This could be determined by the size of corresponding eigenvalues). The number of the chosen eigenvectors p is clearly smaller than the original number of eigenvectors n. PCA can thus be considered as a way to compress images into a subspace. This transformation is also known as the Karhunen-Loève Transform (KLT). The distance in this subspace is the so-called *Mahalanobis distance*, which is equivalent to approximating the sum of squared errors between the two images. Thus PCA and KLT are commonly used for dimensionality reduction and object recognition.

Swets and Weng [42] have used eigenfeatures for image retrieval of well-framed face images. They pointed out that "more features do not necessarily imply a better classification success rate. ... Although the most expressive feature (MEF) projection is well-suited to object representation, the features produced are not necessarily good for discriminating among classes defined by the set of samples." In their work the most discriminating features (MDF) are sought after by a two-step projection from the n-dimensional image space to the m-dimensional MEF space, followed by a constrained discriminant projection. The goal is to find a linear space where the sample point are well separated by their classes, i.e. to maximize the between-class scatter while minimizing the with-class scatter as what is done in [14].

Moghaddam and Pentland's work reported in [28] is on finding "the saliency maps which have a probabilistic interpretation as object-class membership functions or likelihoods". In terms of

image semantics and probability based approach their work is directly relevant to our current effort.

Our work differs from previous work in this area in that (1). image classification or object recognition is not the goal but rather a feature evaluation tool used as an intermediate step to find a good similarity metric. Any good classifiers found in this step are simply a by-product of our semantic image retrieval process; (2). the feature generation is non-traditional in the sense that it is **not** simply the visual appearance we are looking for but from a collection of inexpensive image features to capture the implicit probabilistic model of human brain symmetry; (3). there is no limit on what kind of features to be used in our framework, including and not limited to eigenfeatures.

2.2 Medical Image Analysis

Work on neural image analysis has been concentrated more on morphological variations of normal brains[8, 18] or brains with mental disorders such as Alzheimer's disease[47], Schizophrenia [48] using high resolution, high density MRI data scanned especially for their research purposes. Also, the emphasis in these work has been on the visualization of variations in normal and psychiatric brains.

The image data used in this work, on the contrary, is primarily from clinical practice. For example, CT images are taken in the Emergency Room (ER) (See Data sets 1 through 5 in Table 1) to rule out certain possibilities such as hemorrhage. The sparseness and incompleteness of the image data poses more challenges in image understanding and indexing than does the complete, dense data used in other reported work. However, the images we use more realistically reflect the type of image database and query images that an end user will consult with in practice.

Martin et al [26] has studied shape descriptions of pathological brains caused by neurological disorders. They assume segmentation of the brain anatomic structure is given. Two anatomical brain structures are used as the shapes for comparison: putman and ventricle. The key issue here is to separate the shape changes in normal brain variations and those shape differences caused by schizophrenia, Alzheimer or normal-pressure hydrocephalus. Finite element method is used to characterize the global deformation where the intracranial cavity is modeled as a piece of linear elastic material. PCA is then applied and low-frequency high amplitude modes are chosen, the justification being: "... their amplitudes are larger and, therefore, for object discrimination, they are typically more powerful." For Schizophrenia case, 72% correct classification rate is reported on 13 patients and 12 healthy control subjects; and 88% correct rate is obtained for ventricular disorders on a dataset of 9 Alzhemiers, 7 normal-pressure hydrocephalus and 9 healthy control subjects. These rates are what we call leave-one-out training set rates. This work proposes one way to separate normal brain shape variations from pathology caused shape variations. However,

due to the simplicity of the brain stiffness model how much normal shape variations are actually modeled is unclear.

Both 3D image intra-subject multi-modality (same person's head, different modality) rigid registration [52, 53, 16],and inter-subject (across patients) normal brain deformable registration [2, 7, 32, 46, 27] have achieved excellent results. Using a brain atlas to deformably register with a normal subject's image has provided promising results in automatic anatomical segmentation of normal brain images [9]. However, **no** existing technology is yet ready for registration of inter-subject, pathological clinical brain images. Matching two 3D pathological brains remains a difficult problem. The topology of normal brains is the same while the geometric shapes vary from person to person, pathological brains violate this basic principle. New method to compare neurophysiological pathology brains are needed to get around this problem. In this work we take a different direction to understand brain pathology such that no deformable registration and segmentation are required.

2.3 Medical Image Database Retrieval with Text or with Imagery

With similar motivation but a completely different approach, Chu et al [5, 6] developed a text-based natural language interface for retrieval of semantically relevant neural images. Their work heavily depends on human experts to construct a layered query structure, and to provide extracted text information from 2D images. The manipulation of images in their work is minimal. No registration between images is done before comparison, and only 2D slices are used. Furthermore, the tumor location is defined as the X,Y image location regardless of the size and shape differences of human brains. In contrast, we focus primarily on automated image-based retrieval using a combination of image understanding technology, machine learning theory/algorithms, combined with a understanding of human anatomy. We believe that a true semantic retrieval system is not a simple addition of database and image processing techniques to domain knowledge but an intertwined process as demonstrated in this paper.

In the medical imaging domain, the only work on automatic 3D image retrieval has been dealing with dense (typically with $256 \times 256 \times 125$ voxels each with size $1 \times 1 \times 1.5mm^3$) MR images on normal subjects. The retrieval process reduces to a deformable registration process that is feature-based [11] or density-based [19]. In [19] a database of 10 MR images of healthy subjects are used to retrieve corresponding anatomical structure defined as a "volume of interest" (VOI) given by a user in a reference image. Cross correlation and stochastic sign-change are used as measures for morphological differences (similarity). The CANDID project [22], or more recently the TeleMedicine project, carried out at Los Alamos National Lab, deals with retrieving similar 2D

CT slices of lung diseases using texture analysis. In [40], 2D CT slices of lung diseases are also used to extract local features for content-based image retrieval, starting with physician identified region of interest.

There are several existing hip or joint medical databases. However, there are two main problems with them: one is the lack of digital image data in the database, and the other is that intelligent retrieval is not provided. In [3], for example, a surgeon's query has to be done by filling out a form first and then a skilled programmer is required to do the search through the database.

3 Our Approach

Given a semantically well-defined image set, image classification and image retrieval share fundamentally the same goal. Dividing images based on their semantic classes and finding semantically similar images should also share the same similarity measurement and performance evaluation standards. Our hypothesis is that *a distance metric* used by a classifier that gives good classification results should also be expected to provide good image retrieval result when used as the image similarity metric for retrieval. Therefore, in our work image classification is used as an image index feature selection tool to find a good similarity metric for semantic based image retrieval. The net effect is equivalent to warp the initial multidimensional feature space such that the image points corresponding to the same semantic class are well-clustered in the warped feature space. It should then be obvious that image retrieval using nearest neighbors in this feature space can achieve good results. Here two images are *semantically similar* means they belong to the same semantic class. In this paper we refrain ourselves from discussion of cross semantic class similarities.

Instead of being the goal, image classification is a tool for selecting a non-uniformly 0-to-1-inclusive scaled subset of image features which collectively form an optimal classifier on the given training set. This classifier is then used as the similarity metric for image retrieval.

We propose an image retrieval framework consisting of three stages as shown in Figure 4: 1) feature extraction, 2) feature selection via image classification, and 3) image retrieval using k nearest neighbors in the selected feature space. The function of feature selection stage is to determine, quantitatively via classification, the most discriminating subset of features across predefined semantic classes. This distinguishes our work from most existing systems, where no direct evaluation of the chosen image features is performed automatically and objectively. A typical content-based image retrieval system simply extracts certain predetermined image features from both the query image and the images in the database, finds the nearest N neighbors of the query image in the feature space and returns those images as the retrieved result. The difficult feature weighting problem is usually left for the human user to decide. Besides this approach provides no

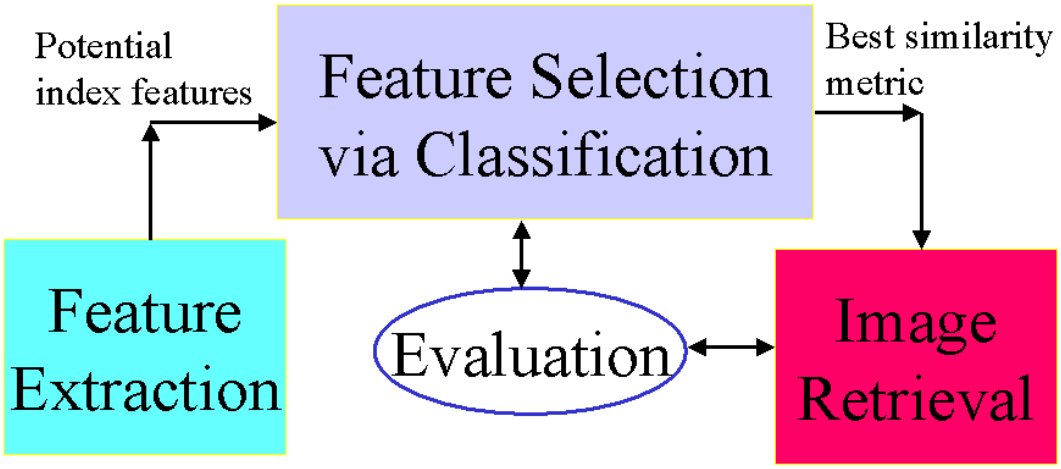

Figure 4: Three steps in classification driven image retrieval

insights to the distributions of the database image feature space, feature weighting is not always a manageable task, especially when the user is not experienced or when more than 5 indexing features are used. What is missing in the common image retrieval practice is an evaluation of the extracted image features **before** they are used for image retrieval. Critical questions like "*Are these features sufficient to capture the content of the image? Is there any redundancy in these features for discriminating different images? Which subset of these features has the most discriminating power across image semantic classes? and what are their relative weights?*" are left unanswered. In contrast, the method we propose directly answers these critical questions. Our goal is using memory based learning with database image features as the training set to discover and manipulate its feature space, and to find the most direct, most effective and most economical mapping from a proper subset of non-uniformly rescaled image features to their corresponding pathology classes.

3.1 Image Collection

The 3D images we use in this research are collected from Allegheny General Hospital (AGH) in Pittsburgh. Each of these images is a clinical image taken for diagnosis or treatment purposes. Most of them are taken in the emergency room while a patient first came in to the hospital. Follow the guidelines of the medical knowledge bank, medical cases that are typical or cases that are rare are chosen to be included in the database. Care is taken by a group of experienced neuroradiologists to include cases where images are not ambiguous. For each chosen case, a search for all related information is done and the patient's personal information, such as name and identification number, is deleted from the information of all forms to protect patient's confidentiality. The images for

each medical case are deposited on a workstation in the hospital using DICOM3 (Digital Imaging and Communications in Medicine, version 3) format, a commonly accepted medical image and communication format. A pseudo name is given for each 3D brain scan and each 2D slice is labeled orderly. The medical images used in this paper are CT images (axial slices) taken using Siemens CT scanner. The images are transmitted through Ethernet from the hospital computer workstation directly to an SGI workstation located in CMU. DICOM3 servers on both workstations in AGH and CMU transfer/receive images through a predefined communication protocol. Each 3D image is transmitted as a set of 2D slices, each 2D slice is an independent file composed of a 0.5MB-sized image file and a 0.001MB-sized header file.

3.2 Image Preprocessing

A 3D medical image is usually taken as a stack of parallel 2D images. As an example 3D neuroradiological image, see Figure 5 for a set of axial CT images placed on a plane, and Figure 6 for the same set of CT images interpolated and thresholded in 3D.

3D volumetric images pose some special challenges and possibilities for image computation compared with 2D images. An *ideal head coordinate system* can be centered in the brain [45] with positive X_0, Y_0 and Z_0 axes pointing in the right, anterior and superior directions respectively (Figure 7, white coordinate axes). Ideally, a set of *axial (coronal,sagittal)* slices is cut perpendicular to the $Z_0(Y_0, X_0)$ axis. In clinical practice, due to various positioning errors, we are presented not with the ideal coordinate system, but rather a *imaging coordinate system* XYZ in which X and Y are oriented along the rows and columns of each image slice, and Z is the actual axis of the scan (Figure 7, black coordinate axes). The orientation of the imaging coordinate system differs from the ideal coordinate system by three rotation angles, *pitch*, *roll* and *yaw*, about the X_0, Y_0 and Z_0 axes, respectively.

Since the images are collected from different scanners in different hospitals, they vary in modality, resolution, volume, scan angle and scanning axes. Until two 3D images are properly aligned and normalized, existing techniques for content-based retrieval using color, texture, shape and position can not be applied directly for meaningful results. This fact reflects one of the fundamental differences between 2D and 3D volumetric image retrieval: alignment before comparison is necessary and possible for 3D volumetric images.

It is important to identify the midsagittal plane of a given 3D neural image. This plane is helpful for the registration across different 3D images. Moreover, this plane can be used as the reflection plane to compute left and right brain differences. For dense image data where voxels are nearly cubical, a maximization of mutual information affine registration algorithm [25] is applied to

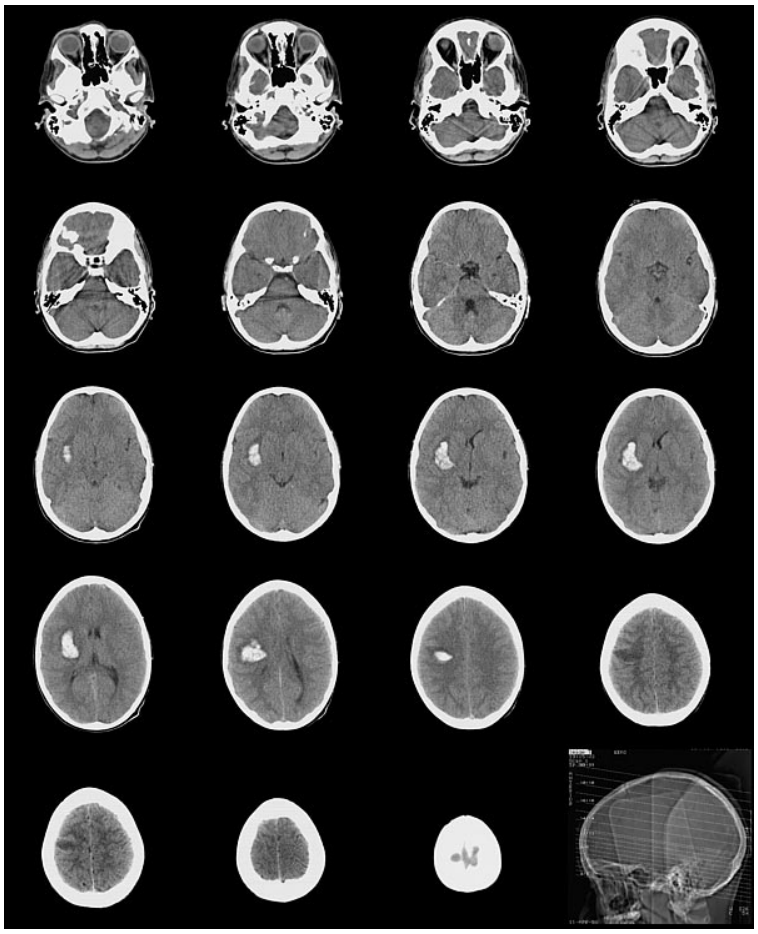

Figure 5: A 3D image which is composed of a set of clinical CT scans (axial), only a portion of a patient's head is captured as shown in a side view on the lower right corner. This is a case of acute right basal ganglion bleed.

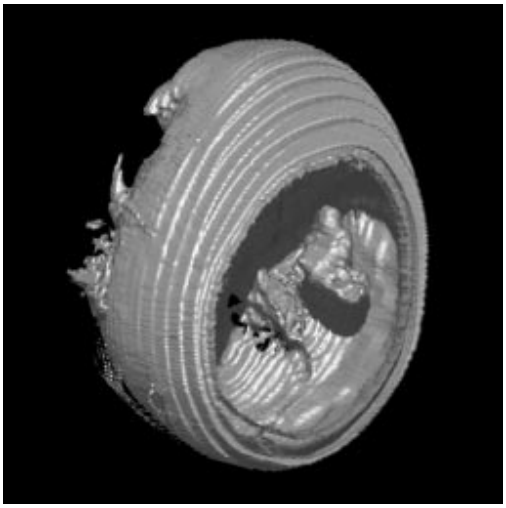

Figure 6: A 3D visualization of CT scan shows the appearance of the lesion with respect to the skull.

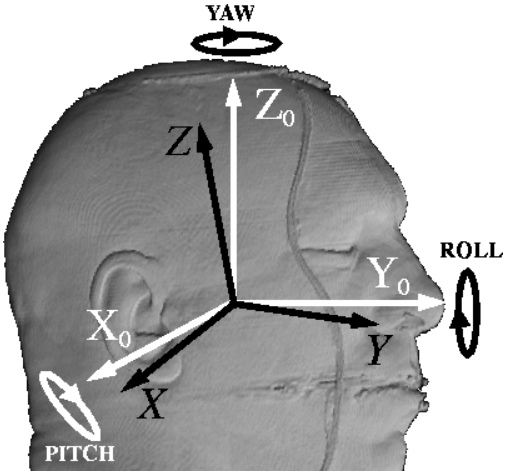

Figure 7: To compare different 3D images it is first necessary to register the 3D imaging coordinate coordinate system X, Y, Z with the ideal head coordinate system X_0, Y_0, Z_0. Rendered head courtesy of the Visible Human Project (supported by the National Library of Medicine).

register the original image with its own 3D image bilateral reflection about the Y-Z plane (Figure 7). After registration the image can be aligned with the ideal coordinate system X_0, Y_0, Z_0 by rotating half of the yaw and roll angles, and the full range of the pitch angle.

Table 1: A Sample of Input 3D Image Data

Image	Modality	Form	Size	Voxel (mm^3) (#slice)	Pathology
1	CT	axial	256x256x20	0.5x0.5x5 (1-11) 0.5x0.5x8 (12-20)	Stroke
2	CT	axial	514x518x19	0.5x0.5x5 (1-11) 0.5x0.5x10 (12-19)	Basal ganglion acute bleed
3	CT	axial	512x512x23	0.5x0.5x5 (1-10) 0.5x0.5x8 (11-23)	Epidural Acute Bleed
4	CT	axial	512x512x21	0.5x0.5x5 (1-10) 0.5x0.5x8 (11-21)	Stroke
5	CT	axial	512x512x20	0.5x0.5x5 (1-9) 0.5x0.5x8 (10-20)	Normal
6	MR (T2)	axial	176x236x187	0.98x0.98x1.2 (1-187)	Normal
7	MR (T1)	coronal	256x256x123	0.9375x0.9375x1.5 (1-123)	Atlas, Normal

For sparse 3D data where voxel has a voxel bottom edge to height ratio of 1/20 (Image 2 in Table 1), the above registration algorithm cannot be applied directly. To avoid poor re-sampling quality of the original image a method using 2D slice cross-correlation to align the midsagittal plane is developed [23]. The basic idea is to find symmetry axis on each 2D slice, then optimally fit a 3D plane through these parallel lines in space. This midsagittal plane extraction algorithm has been tested on both CT and MR normal and pathological images and exhibits about one degree of rotation error when compared against neuroradiologists' hand alignment.

The result of midsagittal extraction is not to find where the midsagittal plane is, but where it is supposed to be. This is especially true for pathology brains since the midsagittal plane often is distorted (shifted and bended) due to large lesion. See Figures 8 and 9 for a sample of the results obtained on 2D slices and 3D images. One thing to keep in mind is that although the result is shown on 2D slices the method is using 3D computation of the whole brain. So local asymmetry on 2D slices has little effect on the 3D orientation and location of the computed midsagittal plane.

For sparse 3D images, the pitch angle in each 3D image is not corrected after the midsagittal plane is identified. A 2D maximization of mutual information registration algorithm is used to register each image to a common midsagittal plane to correct the pitch angle. However, little pitch

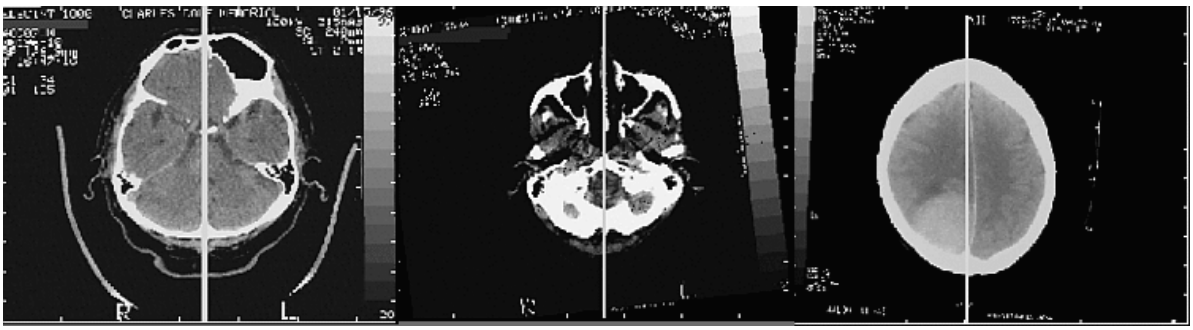

Figure 8: The ideal symmetry axis is extracted as the intersection of a 3D brain image and its midsagittal plane. The method works even in the presence of obvious asymmetries, for example the sinus near the frontal lobe in the left image.

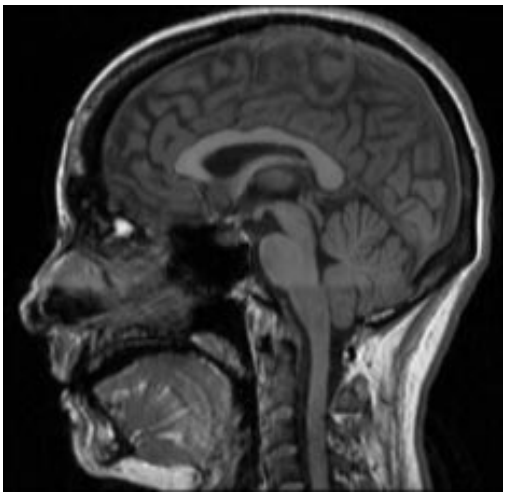

Figure 9: Automatically extracted midsagittal plane from a 3D MR image

angle difference has been observed in the CT images used in our experiments. This is because the images are taken in the same hospital and from the same CT scanner, and the technicians who took the images follow the same orbital meatal line[2] while scanning. To maintain the integrity of each 2D slice, no further re-sampling is done. From now on we assume the pitch, yaw and roll angles are corrected in each 3D image, and symmetry axis (plane) of the brain is centered in the middle of the image.

3.3 Feature Extraction

A set of most significant visual features for detecting pathologies has been identified by neuroradiologists, this set includes:

1. mass effect (departure of anatomical structure from normal brain bilateral symmetry),
2. anatomical location of the lesion,
3. density of the lesion,
4. contrast enhancement on the lesion,
5. lesion boundary,
6. lesion shape,
7. edema around the lesion,
8. lesion size,
9. lesion texture,
10. lesion station,
11. patient age (brains shrunk while aging which can be observed on an image).

Notice, 9 out of the 11 features require the explicit detection of the lesion. As it is well known that medical image segmentation is a hard problem. Especially, each 3D image can contain multiple lesions without knowing beforehand how many there are. These lesions take no specific forms, density values, shapes or geometric and anatomic locations (see Figure 10). In attempting an automatic segmentation algorithm, it is hard to avoid some sort of subjective thresholding.

[2]This is an approximate line from the angle of the orbit to the external auditory meatus and then go up about 12 degrees or so.

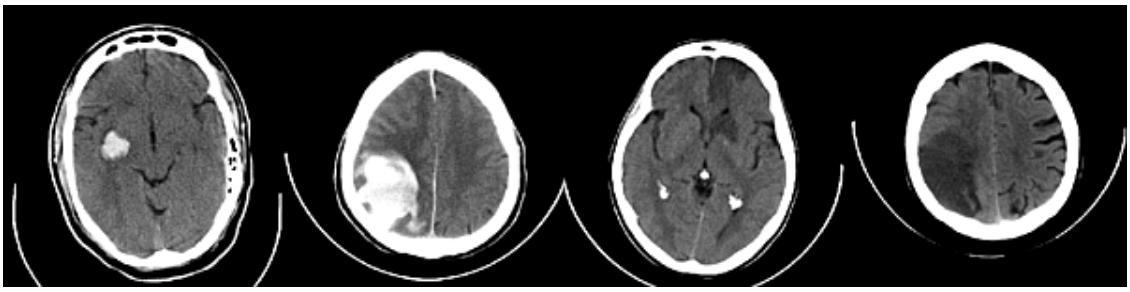

Figure 10: Variations of lesions.From left to right the pathologies are: acute blood, acute blood, infarct, infarct

Though many sophisticated segmentation algorithms are developed (Section 2), few are universally applicable. In medical image retrieval, many systems depend on human experts to hand pick the lesion as the initial input.

We take a completely different approach: instead of trying to precisely locating the lesion interactively or automatically, we collect a set of relatively simple and computationally inexpensive image features based on the assumption of normal brain symmetry. One must realize that though the intention is to capture those features human experts suggest to be useful, there are NO necessary one-to-one correspondences between the features human experts identify and the statistical features a computer algorithm generates from an image. The hypothesis is that some (appropriately weighted) subset of these simple features, collectively, may capture the statistical 3D brain model of the pathology accurately. The feature selection process will be carried out in the next stage using Bayesian classifiers. If none of these selected features, or any subsets of them, are effective, the result of the classification will reflect that fact. Only then other, perhaps more expensive means, will be utilized to extract more relevant features.

After image preprocessing, the ideal midsagittal plane is identified and aligned with plane $X_0 = 0$ (Figure 7). To extract features that will quantitatively describe the amount and type of asymmetry in the image, we use three types of symmetry-descriptive features: 1) global statistical properties, 2) measures of asymmetry of half brains, and 3) local asymmetrical region-based properties. These features are extracted from the original image I with its midsagittal plane aligned with the $X_0 = 0$ plane, the difference image D of the original image and its mirror reflection with respect to $X_0 = 0$, the thresholded difference image, and the original image **masked** by the thresholded binary image (Figure 11).

Global statistical properties are computed over the entire half brain, and include features like the mean and standard deviation of grey-level intensity. *Asymmetry features* are obtained by voxel-wise comparison of corresponding left and right half brains. Two techniques are used to obtain

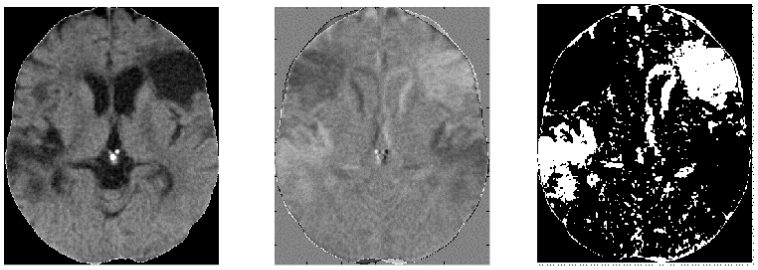

Figure 11: Left: original image I with midsagittal plane aligned. Middle: difference image $D = I - reflect_v(I)$. Right: thresholded D

these features. (1) The first is to simply subtract out a vertically reflected image of the left brain from its corresponding right half, to obtain a difference image D (Figure 11). Asymmetries show up as large positive or negative density values in the difference image. Numeric features are then computed as counts of how many voxels remain after thresholding D with different values. (2) In the second technique, each image is smoothed with a Gaussian filter having a standard deviation of 5, 9, or 15 pixels respectively. This equivalent to examine images at different resolutions. Then the difference between each voxel and its counterpart in the opposite symmetrical half is recorded. If that counterpart voxel falls significantly outside the estimated image gray-value distribution, i.e. the difference is greater than 3, 4, or 5 standard deviations, it is flagged as being significantly different, and again the number of voxels that pass this threshold test are counted. Finally, a set of *local region statistics* are generated by masking the original image with the threshold images obtained in the previous step. Intensity statistics such as mean and variance are then computed over these asymmetrical areas, and thus pertain only to the local areas where asymmetries are present.

3.4 Feature Selection via Classification

We hypothesize that *the similarity metric that does well at classifying the images will also do well in finding similar images.* We propose a principled method, firmly rooted in Bayes decision theory, to determine the set of most discriminative features by evaluating how well they perform on the task of classifying images according to predefined semantic categories.

Kernel regression (KR), a memory based learning (MBL) [1, 31] technique, is applied to classify images in the database. It works by approximating the posterior densities involved in the classification process using a technique called Parzen window density estimation [34, 12, 4]. To understand this process, recall that the a posterior probability $P(c|\mathbf{x})$ of an image being in class c

when feature vector $\mathbf{x}$ is observed can be computed via Bayes law [12]:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x}|c)P(c) + P(\mathbf{x}|\overline{c})P(\overline{c})} \tag{1}$$

The *prior probability* $P(c)$ of a class c can be easily estimated from labeled training data by dividing N_c, the number of instances in class c, by the total number of instances N in the training set: $P(c) \approx N_c/N$. The conditional density $P(\mathbf{x}|c)$ for each class is estimated using the aforementioned Parzen window density estimation technique. Intuitively, this is done by placing an identical Gaussian kernel on each of the training instances $\mathbf{x}_i$ in a given class, and approximating the density by a (appropriately normalized) sum of the identical Gaussian kernels:

$$P(\mathbf{x}|c) \approx \hat{P}(\mathbf{x}|c) = \frac{1}{N_C} \sum_{j \in C} G(\mathbf{x}, \mathbf{x}_j, \sigma) \tag{2}$$

where $G(u, \mu, \sigma)$ is a multivariate Gaussian kernel, and σ acts as a smoothing parameter.

If we plug in $P(c) \approx \hat{P}(c) = N_C/N$ and (2) into Bayes law (1), it is easily verified that the posterior probability $P(c|\mathbf{x})$ can be approximated by a weighted sum over the labeled instances, where each instance $\mathbf{x}_i$ is assigned a value $f(\mathbf{x}_i)$ of 1 if it belongs to the class, and 0 otherwise:

$$P(c|\mathbf{x}) \approx \hat{P}(c|\mathbf{x}) = \frac{\sum_i G(\mathbf{x}, \mathbf{x}_i, \sigma) f(\mathbf{x}_i)}{\sum_i G(\mathbf{x}, \mathbf{x}_i, \sigma)} \tag{3}$$

KR uses the above formula to approximate the posterior $P(c|\mathbf{x})$, and thus simply calculates a weighted average of the classification function f, averaged over the entire training set of training instances $\mathbf{x}_i$ and weighted by a Gaussian kernel centered around $\mathbf{x}$.

To evaluate individual KR classifiers, we compute the cross-entropy of the training data given a classifier. There is a whole space of classifiers, as the KR density estimation will yield a different result for each value of the smoothing parameter σ, or when the distance metric (feature weighting) is changed. *Cross-entropy* is defined as the negative log-likelihood of the training data given a specific classifier, and is therefore a measure for how well a classifier performs. Formally, the cross entropy is defined as $E = -\sum_i \sum_c \delta_{ic} ln \hat{P}(c|\mathbf{x}_i)$ where δ_{ic} represents the 1-of-m multiple class membership encoding, and $\hat{P}(c|\mathbf{x}_i)$ is the approximation of the a posterior probability $P(c|\mathbf{x})$ of a class c given a feature vector $\mathbf{x}$ via Bayes law. Minimizing this function will yield a metric and a kernel width σ for which kernel regression best approximates the a posterior probabilities $P(c|\mathbf{x})$, and is thus optimally suited for classification [4].

The core of our approach consists of the fact that we perform an off-line, combinatorial search in the space of classifiers to find one that minimizes a classification performance measure. We will then use the associated similarity metric for image retrieval. Since kernel regression is almost

uniquely defined by a distance metric (apart from σ), it is ideally suited to our purpose. However, different classification methods could conceivably be used (as long as they include the notion of a distance metric) making our method very general. To prevent overfitting to the training data, each classifier is evaluated using leave-one-out cross-validation[30] on a training set, containing roughly two-thirds of the available data in the database. The rest of the data is set aside for evaluation purposes, and is on this set that our results are reported. The specific optimization technique used and the methodology of our approach will be explained in more detail in the experiments section below.

The main result of the optimization process is a similarity metric, i.e. a weighted set of the most discriminating features for use in a classifier. The classification process screens out the majority of the 50 or so features that were extracted for each image, and ends up with a proper subset of the initial feature attributes. Often, we have found that there are several feature subsets that have quantitatively equivalent discriminating power. We believe this is explained by the large redundancy that exists between the features themselves.

3.5 Classification Evaluation

Table 2: Evaluation Measurements for Classification and Retrieval

Measurement Definitions
T = number of total image instances
P = number of pathological instances in T
N = number of normal instances in T
TP = total number of correctly classified pathological instances
P_as_N = number of pathological instances which are classified as normal
N_as_P = number of normal instances which are classified as pathological
B_as_S = number of instances which are blood but classified as stroke
S_as_B = number of instances which are stroke but classified as blood
True positive rate **TPR** = TP/P
False positive rate **FPR** = N_asP / N
False negative rate **FNR** = P_as_N / P
Confusion rate **CFR** = (B_as_S + S_as_B) / P
Classification Rate **CR** = 1 - (P_as_N + N_as_P + B_as_S + S_as_B) / T = 1- ((1-**TPR**)*P/T + **FPR***N/T)

The quantitative evaluation of potential classifiers (similarity metrics) are carried out using

measurements defined in Table 2. These definitions are established for a three-class classification problem, one normal and two pathology classes.

In general, the classification will never be perfect, as the densities of the respective classes overlap in feature space, and it is expected that errors will be made. This motivates the use of decision theory to minimize the cost of deciding on a particular class for a given sample. Specifically, in the domain of medical image classification/retrieval, it is imperative to minimize the *false negative rate* **FNR**, i.e. minimize the occurrence of pathological cases being classified as normal. This motivates a *cost matrix* structure $C_{ij} = \lambda(\alpha_i|c_j)$, where a *false negative penalty* w is incurred whenever a pathological image is classified as normal, whereas a normal image is classified as pathological simply incur a unit cost and zero cost presents when a class chosen is the correct class. Thus, given three classes: 1. normal, 2. blood and 3. stroke, we typically use a cost matrix of the form shown in Figure 12. A classification decision is made by minimizing the expected risk associated with choosing class (Figure 13). For example, $R1$ in Figure 13 is the risk to take when choosing class 1, i.e. normal, for the given image. From the expression one can see that the higher the value of w, the higher the risk. Thus by increasing w false negative decisions are effectively punished.

Cost Matrix

	True Class N	True Class S	True Class B
Classify As N	0	w	w
Classify As S	1	0	1
Classify As B	1	1	0

Figure 12: One example cost matrix for three semantic classes: class 1. normal (N); class 2. stroke (S) and class 3. blood (B).

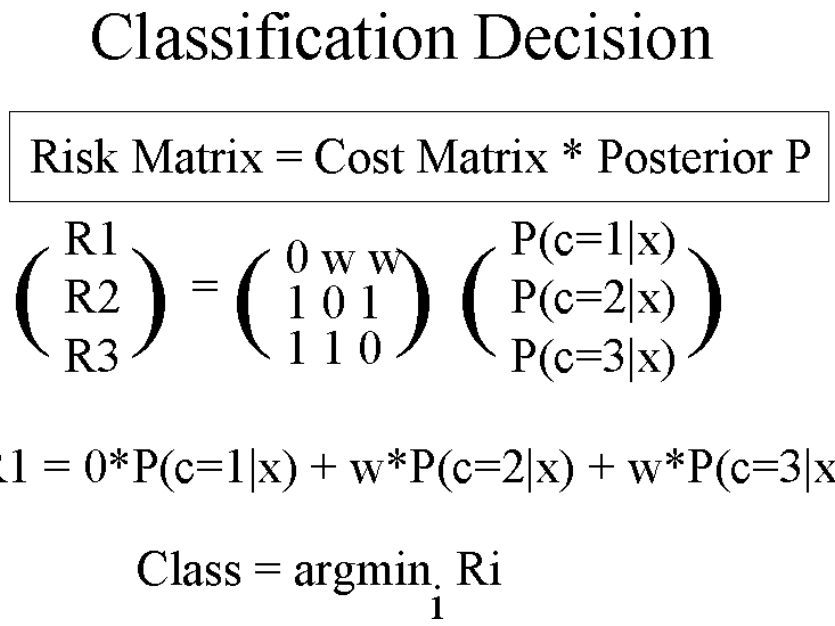

Figure 13: Risk matrix. R1 is the risk to take when choosing class 1, i.e. normal. The higher the value of w, the higher the risk. Thus by increasing w false negative decisions are effectively punished.

3.6 Semantic Based Image Retrieval and Evaluation

Once a (locally) optimal classifier and associated distance metric have been found, we apply this same metric to the image retrieval problem. Assume the extracted image features form an N dimensional vector space, the overall distance function D of two images a, b is defined as

$$D(a,b) = \sqrt{(\vec{A}-\vec{B})^T \mathbf{\Sigma}^{-1} (\vec{A}-\vec{B})} \tag{4}$$

where $\vec{A} = f(a), \vec{B} = f(b)$ are the N dimensional image feature vectors of images a and b respectively, $\mathbf{\Sigma}$ is an $N \times N$ covariance matrix. When $\mathbf{\Sigma}$ is a diagonal matrix, D is the weighted Euclidean distance. When D is the Mahalanobis distance, $\mathbf{\Sigma}$ is a symmetric but not necessarily diagonal matrix. In this paper we use non-uniformly scaled Euclidean distance. The weights in $\mathbf{\Sigma}$ are discovered during the classifier selection process described above.

The image retrieval is done using K nearest neighbor (KNN) as commonly done in image content-based retrieval. The evaluation of retrieved images can be simply the true positive rate or *precision* in the top K retrieved images. Let C_p be the number of correct pathology class in the top k retrieved images, then the top K precision rate is $\mathbf{TPR_K} = C_p/K$.

4 Experiments

To test our ideas, we have built a small, end-to-end database retrieval system, starting with data collection and preprocessing, feature definition and extraction, classification-based search for a similarity metric, and finally quantitative evaluation of system retrieval performance. The following subsections will discuss each of these aspects in turn. Because our approach builds a similarity metric based on a classification problem which is query-dependent, the retrieval metric obtained is tuned to the specific query. This point is illustrated in the last subsection.

4.1 Data Collection/Feature Extraction

This study was performed using a data set of 48 volumetric CT brain scans containing normal, stroke and blood cases. Though our approach can be applied directly on 3D volumetric images, in these experiments we use 2D slices. 2D image are easier to display, and physicians primarily use 2D images as queries to access digital databases. It is also useful to identify normal slices in an otherwise pathological 3D brain image. Finally, it is hard to obtain a sufficient number of 3D images to make accurate statistical inferences from them, whereas a small number of 3D images yields a large set of 2D slices. In fact, we go one step further and ultimately use *2D half slices* as the basic descriptive unit for testing our approach, doubling the amount of data at our disposal once more. This is justified under the assumption that the normal human brains are approximately symmetrical, thus each half of a brain slice is potentially equivalent to the other half. The violation of this symmetry is used as an implication of pathology.

After the data collection is done, each of the half slices is labeled according to its pathology with the help of an expert, after which part of the data is set aside for evaluation purposes. In this specific data-set, there were three distinct classes at the level of the half-slice: normal, stroke and blood. The vast majority of the 2D half slices are labeled as normal, with only 127 stroke and 83 blood on a total of 1250 (See bottom half of Table 3). The 3D image set is then divided into a training set containing 31 3D images and a hold-out test set containing 17 3D images, amounting to a total of 1250 half slices (Table 3). We took care to separate the test set in such a way that there is no mixing of 3D images between training and test set at the level of 2D slices. Instead, the test set is a completely separate set of 3D images, and is only used for evaluation purposes.

The feature extraction for each of the half-slices is performed exactly as explained in the approach section, yielding a total of 1250 data points in a 48-dimensional feature space. Next we will look for a metric in this space that yields good classification.

Table 3: Priori of 3D Image Data for Training and Testing

	Normal	Stroke	Blood	Total (rate)
3D Image	26	14	8	48
Set	54%	29%	17%	100%
Training	16	10	5	31
Set	51.6%	32.3%	16.1%	64.6%
Testing	10	4	3	17
Set	58.8%	23.5%	17.7%	35.4%
2D Half	1040	127	83	1250
Slices	83.2%	10.2%	6.59%	100%
Training	665	86	53	804
Set	82.7%	10.7%	6.6%	64.3%
Testing	375	41	30	446
Set	84.1%	9.2%	6.7%	35.7%

4.2 Searching the Space of Classifiers

To look for a distance metric in feature space with locally optimal classification performance we used a proprietary combinatorial search engine called 'Vizier', implemented here at CMU by Moore et al. [39]. A classifier is defined by a distance metric and a smoothing parameter σ. We are looking for a kernel regression classifier that minimizes the leave-one-out cross-entropy of the training data.

The 'Vizier' search engine works by searching through a large set of possible classifiers within the user's specifications, and the search stops when either it exhausts all the possible choices or a time limit given by the user is reached. Vizier applies a number of combinatorial search algorithms ranging from hill-climbing in metric space to standard feature selection searches common in the machine learning literature. Typically, we ran the search for a full hour on standard high-performance workstations. The output from Vizier is a specification of the best similarity metric found so far: the weights on each input attribute (a warped/scaled Euclidean space which is a proper subspace of the original feature space in terms of dimensions), and the smoothing parameter σ.

4.3 Evaluation of Classification Results

Since the search engine contains some non-determinism and yields an 'anytime' best classifier, the exact classifier found can vary from run to run. Different combinations of features can yield the same discriminative power because of the redundancy between features. One such classifier found used the features specified below. To explain these features, we introduce some notation: given a 2D slice I, we will name its its left or right half IH, the difference image $D = I - fliplr(I)$, and the right or left half of the difference image DH. The classifier in this example used 9 features out of the original 48, all equally weighted in the distance metric:

-
- the sequence labeling of 2D slices
- the mean absolute value of the X direction gradient of IH
- the ratio of non-zero pixels remaining in D when thresholded at $20/255$
- the ratio of non-zero pixels remaining in D when thresholded at $60/255$
- the ratio of non-zero pixels remaining in DH when thresholded at $60/255$
- after Gausian smoothing of image I with $\sigma = 5$ and computing a new difference image D of I, the mean of non-zero pixels remaining in DH when thresholded at 5σ
- after Gausian smoothing of image I with $\sigma = 15$ and computing a new difference image D, the mean of non-zero pixels remaining in D when thresholded at 3σ
- after canny-edge the image I with threshold 15 and $\sigma = 4$, the total number of white pixels in the top-left quarter
- let A be the sum of a pixel-wise multiplication of the absolute value of DH and the absolute value of vertical gradient of IH, and B be the sum of the absolute values of DH, then A/B.

Once a locally optimal classifier is found, we evaluate its classification performance on the test set by plotting standard ROC curves in function of the cost matrix parameter w (Figure 12) that punishes false negatives. Figure 14 displays the ROC curves (true positive rate / false positive rate) of the chosen classifier on classifying 2D half-slices, both for the training data and the separate test set. As can be seen from the figure, the performance on the training set is much better than the performance on the unseen test-set, which is to be expected, as we searched to minimize the former. Thus, reporting performance figures on the training data will give optimistic if not false information

on the performance of the resulting classifiers. The use of the cost-matrix and such summary graphs provide a good quantitative overview of the strengths and weaknesses of a particular set of selected features. It can also be used to select an appropriate w value to use when classifying new, unlabeled images.

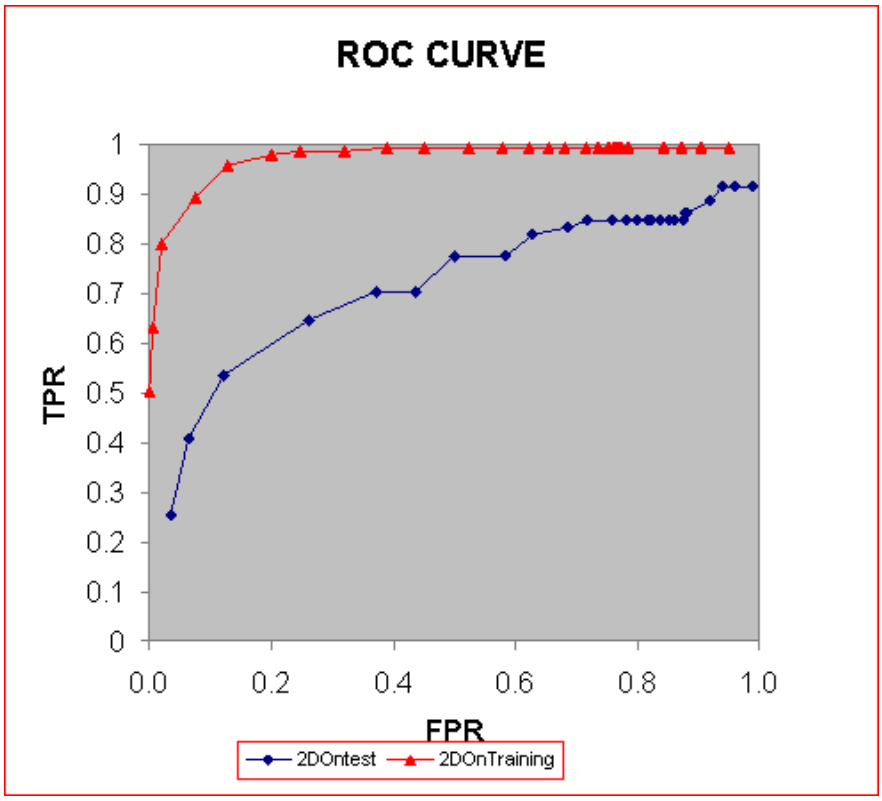

Figure 14: ROC curve for 2D classification on training images using LOO, and on a hold-out testing image set.

4.4 From 2D Image To 3D Image Classification

From the result of the 2D half slice classifiers, can we build a classifier to predict what pathological class the corresponding 3D image belongs to? This exercise is to demonstrate that following the same approach it is possible to construct 3D image classifiers on top of its 2D classifiers. The output of the 2D classifier is the input data for the 3D classifier. For a given 3D image with n 2D half slices, n_N, n_S, n_B are the numbers of predicted normal, stroke and blood classes, where $n = n_N + n_S + n_B$. We use the ratios of $r_S = n_S/n$ and $r_B = n_B/n$ as the two input features for the intended 3D classifier. So the problem becomes to find the posterior probability $P(c|[r_S, r_B])$

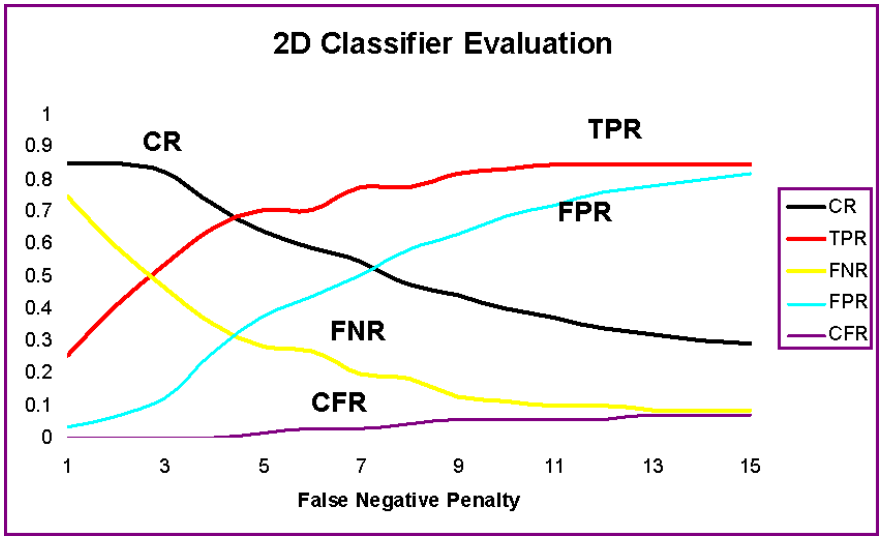

Figure 15: 2D classification rates for increasing value of w.

of a 3D image being in class c when feature vector $[r_S, r_B]$ is observed. Using Bayes law, this probability can be expressed as

$$P(c|[r_S, r_B]) = \frac{P([r_S, r_B]|c)P(c)}{P([r_S, r_B]|c)P(c) + P([r_S, r_B]|\overline{c})P(\overline{c})} \tag{5}$$

Using the same optimization approach we have found several 3D image classifiers. The performance of a 3D classifier strongly depends on the setting of the cost matrix parameter w: for a low value of w, the 2D classifier will classify most of the pathologies as normal, because the densities overlap and normal slices are far more common. For a high value of w, the classification rate is sacrificed to the true positive rate, and the ratios r_S and r_B become unreliable. This is illustrated graphically in Figure 16, where the ratios r_S and r_B are plotted in successive 2D plots, in function of w.

Using a simple 3D classifier with 50% strength on feature r_S and full strength on feature r_B, we evaluate its performance on the testing data for all different settings of w. The result of that experiment is shown in Figure 17. As with 2D classification, classification rate is traded in for true positive rate, although the effect is more indirect through the double classifier structure (and so an ROC curve does not make sense here).

4.5 Two-Class Classification and Retrieval

Our approach is ideally suited to tuning a similarity metric to the specific type of query that will be submitted to the system. As an example, one might only be interested in finding whether an image

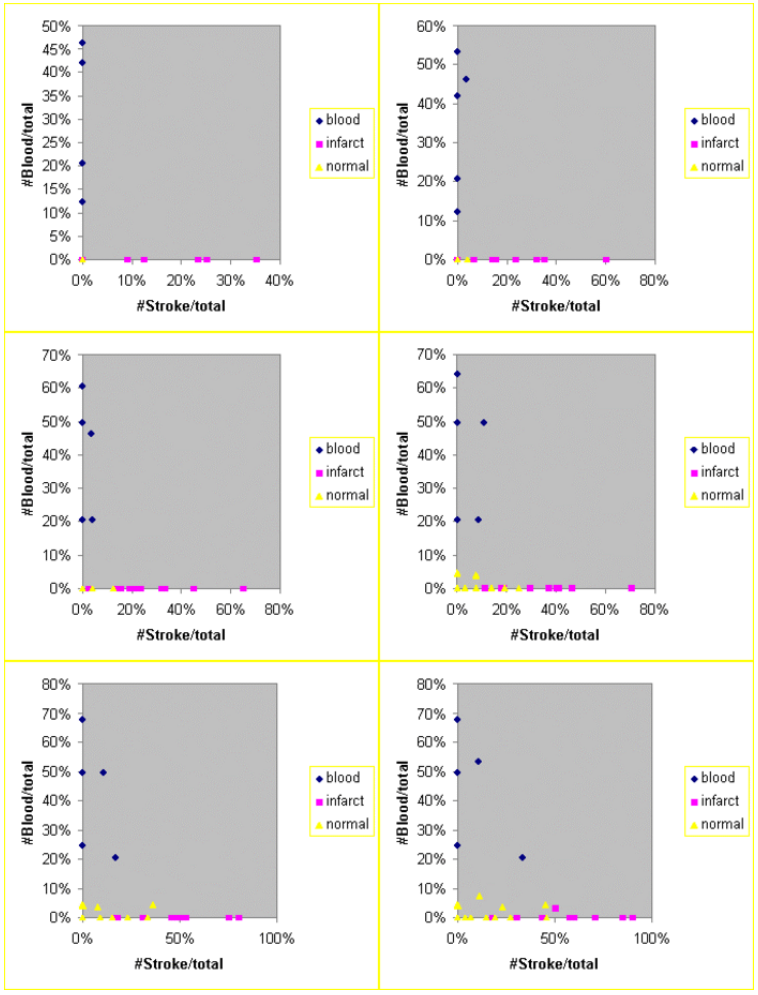

Figure 16: 3D classifier training data is migrating while the false negative penalty of the 2D classifier increases. Read from left to right, top down.

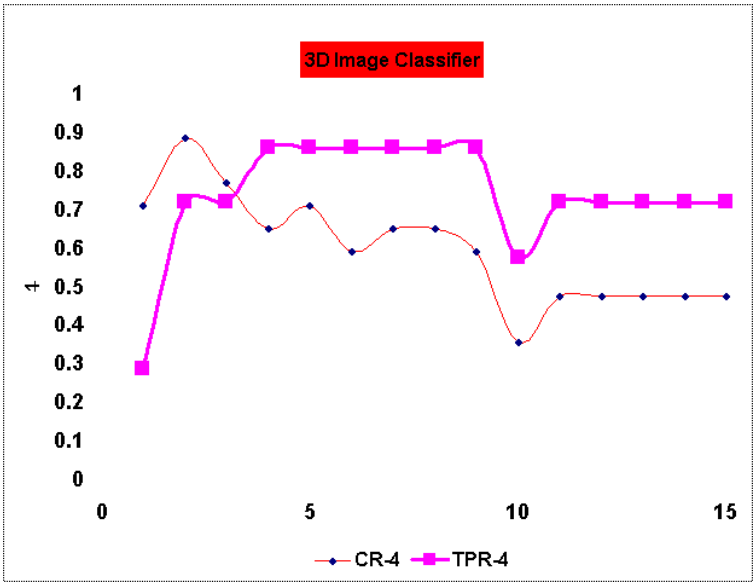

Figure 17: Test-set performance of the best 3D classifier. CR-4: classification rate. TPR: true positive rate.

shows a pathology or not, without regard for the type of pathology. In this case, the metric that classifies the images into normal/pathological might be different from one that needs to make the additional distinction between stroke and blood.

When we tried this, the locally optimal classifier found by the search engine was indeed different from the ternary classification case, although the features used overlapped substantially. The resulting classifier was again evaluated using ROC curves (Figure 18) and the various classification rates on the hold-out test set (Figure 19), as before.

4.6 Image Retrieval

The same similarity metric found above can now serve as an image index vector for retrieving images by finding the nearest neighbors in the feature space, as is conventionally done in content-based image retrieval. However, the dimension of the index feature vector is now much reduced with respect to the initial feature space.

Figure 20 shows the mean retrieval precision rate for all the hold out test images, one for the three-class case and one for the two-class case described in the previous section. The two different optimal metrics found during classification are used here as the similarity metric for retrieval. One can observe a slightly better performance for the 2-class than for the 3-class image set. This is to be expected since 2-class classification leaves less space for errors to be made than the 3-class

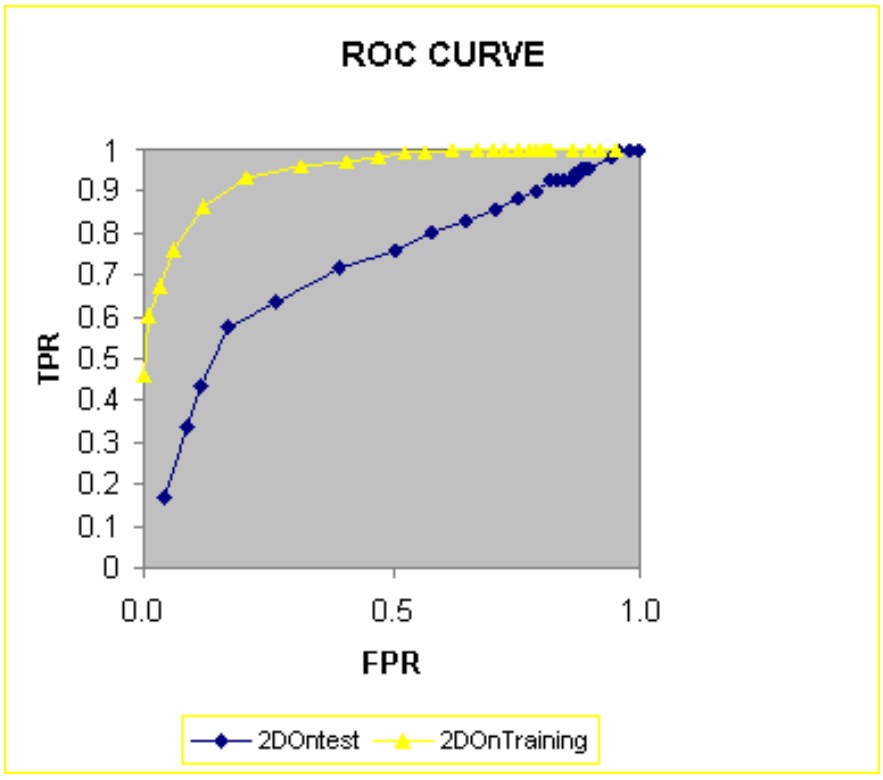

Figure 18: ROC curve for binary 2D classification on training images using LOO, and on a hold-out testing image set.

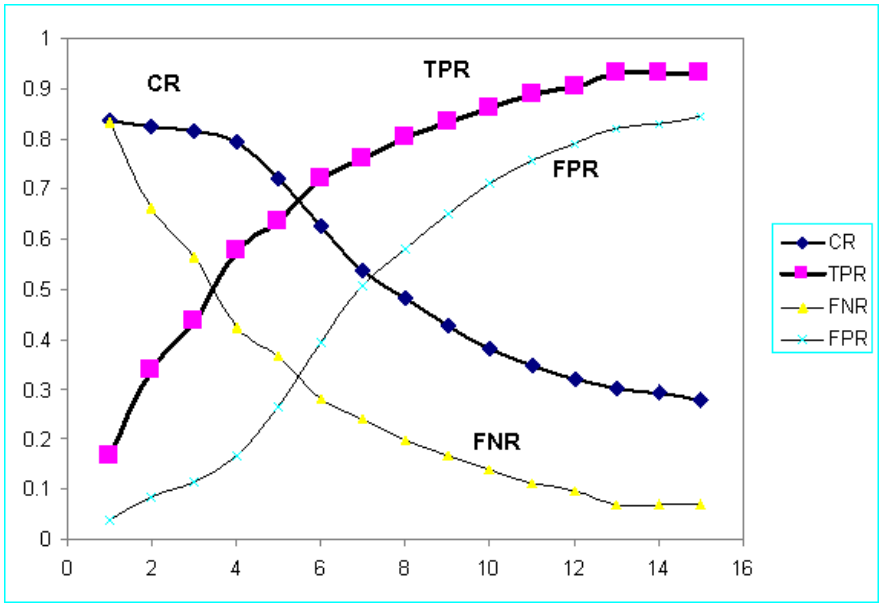

Figure 19: Binary 2D classification rates on a hold-out testing image set.

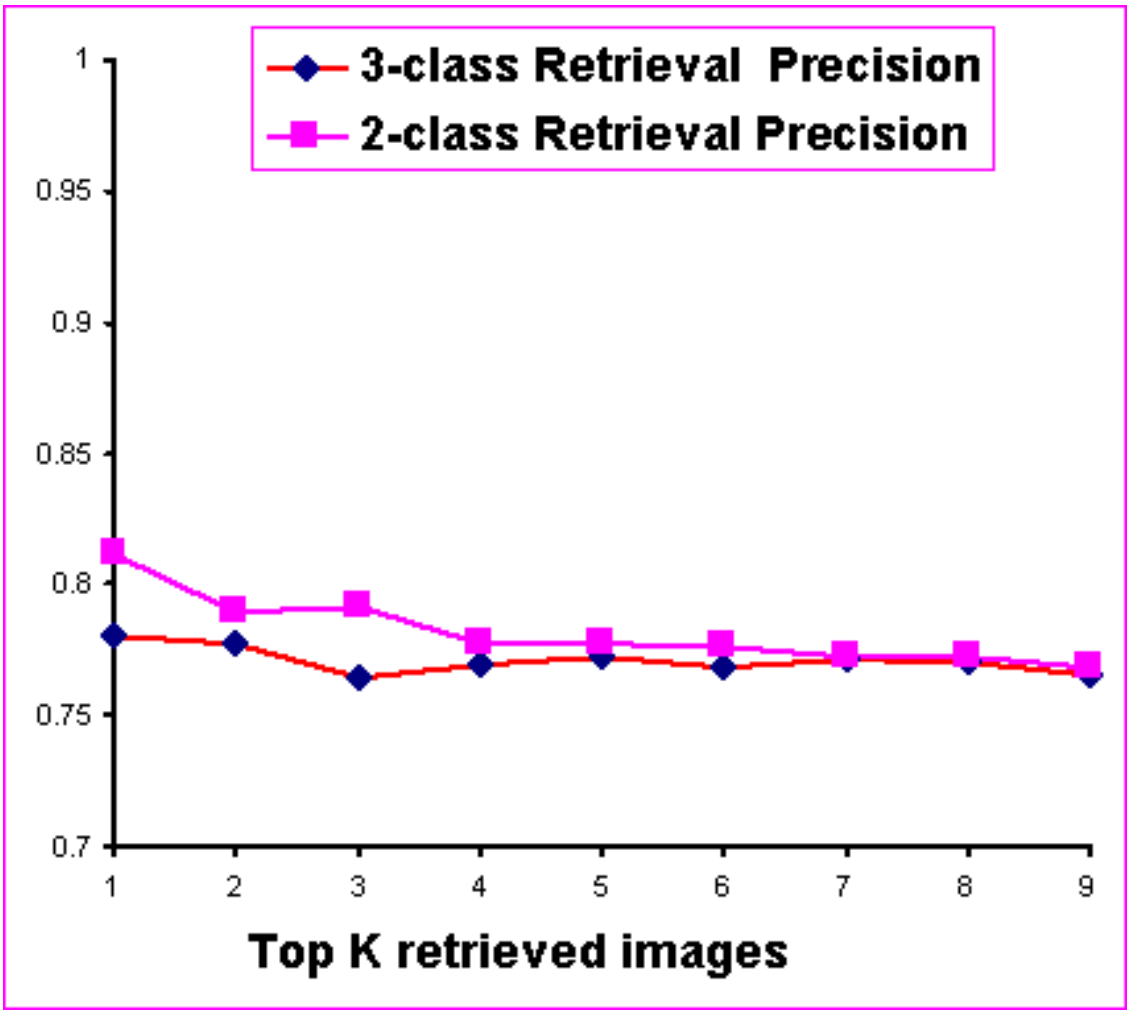

Figure 20: The mean value of retrieval precision as a function of the top K retrieved images.

classification problem. However, in this problem the difficult class separation is not between blood and the normal and stroke, but between normals and strokes. This observation is shown clearly in Figure 16.

Given near 80% precision rate in average, this result implies that in average 8 out of 10 top ranked retrieved images have the same pathology as the query image. Figures 21 and 22 show two of the best retrieved results for blood and for stroke. One can observe the visual difficult in stroke images.

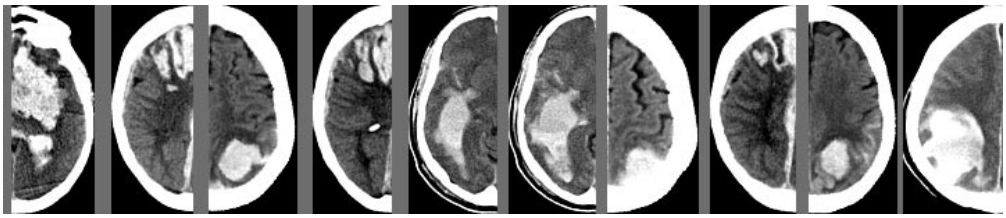

Figure 21: Left: query image with acute blood. The first nine retrieved half slices follow,from left to right in descending order are in descending order of similarity. The pathologies in the retrieved images are all acute blood.

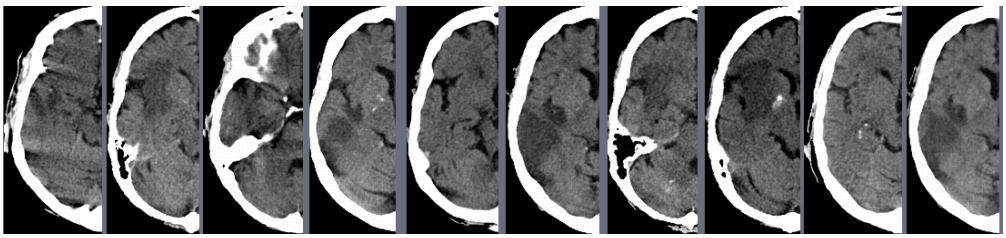

Figure 22: Left: query image with stroke (infarct). The first nine retrieved half slices follow,from left to right in descending order are in descending order of similarity. The pathologies in the retrieved images are: infarct, infarct infarct, normal,infarct, infarct infarct, normal,infarct.

5 Discussion

It should be emphasized that since the initial feature set contains many different attributes trying to depict brain image symmetry from different perspectives, it is less important how many and which features are chosen as a similarity metric. This is because the initial feature set contains a high degree of redundancy. Several feature subsets have quantitatively equivalent discriminating power.

It is somewhat surprising that the classification result and retrieval result show about 80% accuracy given the following facts:

- no detailed segmentation or deformable registration are done on the training, testing or query images, i.e. the pathology is determined without knowing the exact shape, size and the location of the lesion (if any);
- the assumption of human brain symmetry does not always hold, i.e. normal human brains can be grossly asymmetrical and pathology can appear symmetrically in a brain. The probabilistic model in Bayes law is capable of capturing this kind of variable distribution.

During classification-driven semantic image retrieval process, the key step is to learn the conditional density function $P(x|c)$ — the statistical model of brain is of class c given feature subset x. This is where the variations of degrees of symmetry are modeled.

Also worth noting are the following points:

- the retrieved 2D images are registered, thus returned images are located in similar anatomical locations;
- no one-to-one correspondence between expert identified features and algorithm generated features
- classification process is used as a feature evaluation process, the idea is to use simple, inexpensive features first, when the classification rate is low, go back to find more expensive features

From our preliminary study, two stages with five essential components are identified for constructing an intelligent semantic-based image retrieval system (Figure 23). The two stages are (A) off-line similarity metric learning (feature selection) and (B) on-line image retrieval. The five essential components are (1) Image preprocessing, (2) Image feature extraction, (3) Feature selection via image classification, (4) Image retrieval, and (5) Quantitative evaluation. Though the two stages share some common components the goals and constraints differ. In stage (1) the goal is to find the best and smallest subset of image features that capture image semantics. It requires an explicitly labeled image database, sufficient computer memory space to store large feature attribute matrix and to support extensive search on this matrix. High computational speed is a plus but not necessary. Stage (2), on the other hand, demands fast retrieval speed and presentation of retrieved images in addition to a given similarity metric and a query image.

6 Conclusion and Future Work

The main novelty of our approach is to construct a similarity metric suited to semantic image retrieval by finding a metric that does well at classification. Future work includes the study of

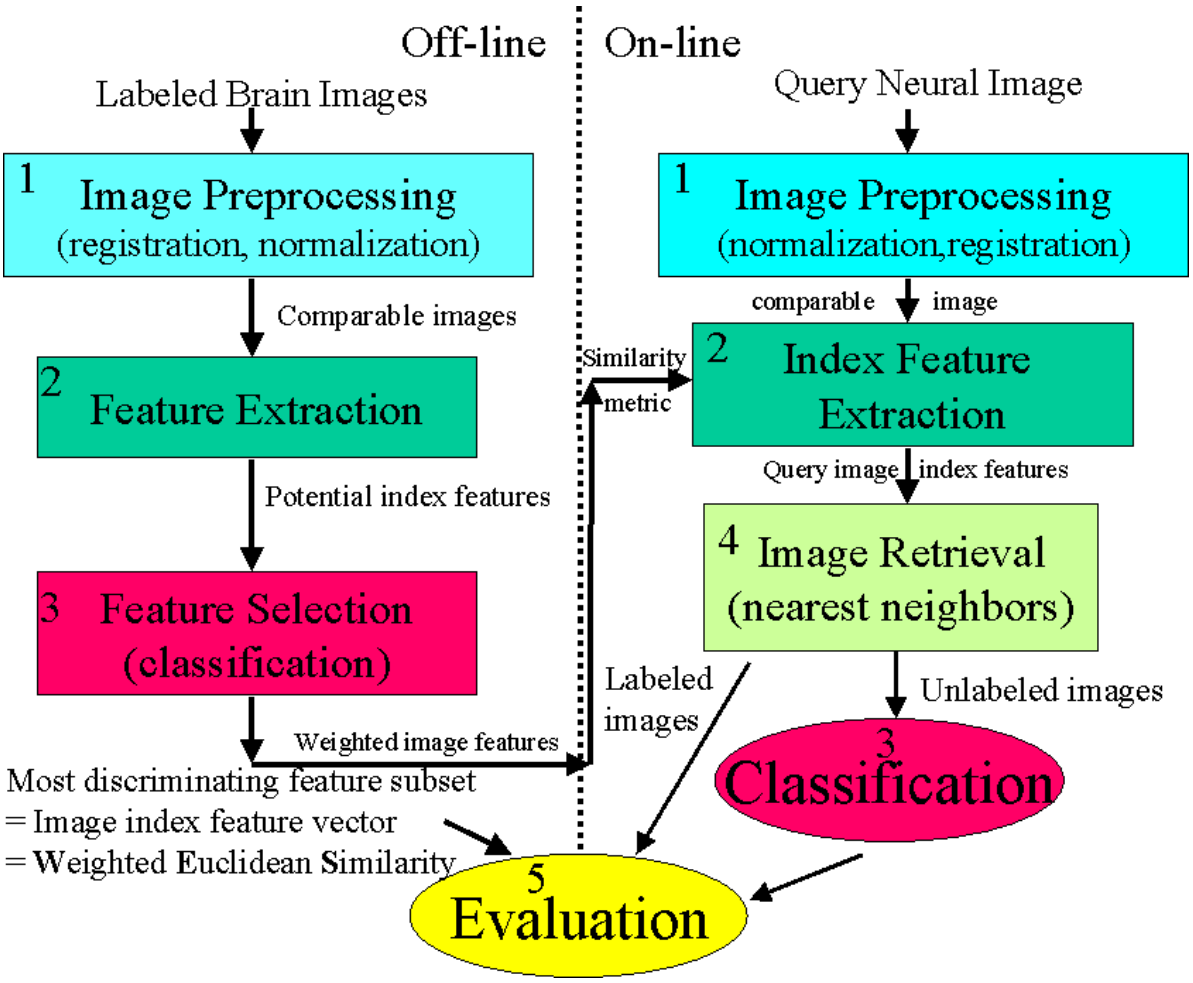

Figure 23: A basic framework and individual components for classification driven semantic image retrieval

different sets of most discriminative features for different purposes or on different subsets of the image database, and dynamically switching from one similarity metric to another during retrieval. For example, within the same pathology such as bleed the anatomical location of the lesion serves as the dominating cue for further detailed classifications, but for tumors, lesion texture is also crucial in their classifications. We expect feature weights to vary accordingly from the initial classification. We also would like to combine visual features with collateral information such as age, sex, and symptoms of the patient to obtain a better retrieval rate and faster retrieval speed. Other kind of features such as eigenfeatures can also be added. The basic framework presented here has provided us with such an *information fusion* capability.

We are planning to apply this framework of classification driven semantic image retrieval to different medical image databases. A Total Joint Registry containing 3D volumetric images (CT, MR) and 2D X-rays. We envision that these two different applications only differ in the image preprocessing and image feature extraction components, the rest of the methodology is general. In both of these domains 3D volumetric images are used. In particular, 2D X-ray images are collected in the Total Joint Registry for each patient over a long period of time. This sequence of 2D X-rays forms a 3D volumetric image along the time axis. We intent to use such 3D image as query image to find similar cases in the Registry. The scalability of our approach is yet to be tested on a much larger sized image database. Larger database provides better, more convincing training and testing image sets for memory-based learning approach.

In any application domain, coming up with the potential feature set is still an important and not easily automated step. Candidate features need to be selected using considerable domain knowledge. One advantage of classification driven approach is that the quantitative classification result is telling you whether the initial features are sufficiently reflecting the image semantics. If it flags not sufficient, some more work on understanding the domain knowledge is called for.

Though this work is carried out in 3D medical image domain, our emphasis on statistical feature extraction, index feature selection via classification and retrieval evaluation, in particular, should provide valid methods regardless of image domain or dimensions.

References

[1] C. Atkeson, S. Schaal, and Andrew Moore. Locally weighted learning. *AI Review*, 11:11–73, 1997.

[2] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Procession*, 46:1–21, 1989.

[3] D.J. Berry, M. Kessler, and B.F. Morrey. Maintaining a hip registry for 25 years. *Clinical Orthorpaedics and Related Research*, (344):61–68, 1997.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995. ISBN:0198538499.

[5] W.W. Chu, F.C. Alfonso, and K.T. Ricky. A knowledge-based multimedia medical distributed database system. *Information Systems*, 19(4):33–55, 1994.

[6] W.W. Chu, F.C. Ieong, and K.T. Ricky. Semantic modeling approach for image retrieval by content. *VLDB Journal*, 3:445–477, 1994.

[7] C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comp. Vis. and Image Understanding, Spefial Issue on Medical Imaging*, May 1997.

[8] C. Davatzikos, M. Vaillant, S. Resnick, J.L. Prince, S. Letovsky, and R.N. Bryan. A computerized approach for morphological analysis of the corpus callosum. *Comp. Ass. Tomography.*, 20:88–97, Jan./Feb. 1996.

[9] B.M. Dawant, J. P. Thirion, F. Maes, D. Vandermeulen, and P. Demaerel. Automatic 3d segmentation of internal structures of the head in mr images using a combination of similarity and free form transformation. In *Proc. of SPIE, Medical Imaging 1998: Image Processing*, pages 545–554. Vol. 3338, 1998.

[10] J.S. De Bonet and P. Viola. Rosetta: An image database retrieval system. In *Proceedings 1977 DARPA Image Understanding Workshop*, 1997.

[11] J. Declerck, G. Subsol, J-P Thirion, and N. Ayache. Automatic retrieval of anatomical structures in 3d medical images. Technical Report 2485, INRIA, Sophia-Antipolis, France, 1995.

[12] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[13] D. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 1994.

[14] R.A. Fisher. The statistical unilization of multiple measurements. *Annals of Eugenics*, 8:376,386, 1938.

[15] D. Forsyth and M. Fleck. Body plans. In *CVPR1997*, June 1997.

[16] E. et al Grimson. An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced realigy visualization. In *The Proceedings of CVPR*. IEEE Computer Society, 1994.

[17] V.N. Gudivada and V.V. Raghavan. Content-based image retrieval systems. *Computer*, pages 18–22, September 1995.

[18] R. Guillemaud, M. Sakuma, P. Marais, J. Feldmar, R. Crow, L. DeLisi, A. Zisserman, and M. Brady. Cerebral symmetry analysis from mri scans. *Submitted to Psychiatry Research Neuroimaging*.

[19] A. Guimond and G. Subsol. Automatic mri database exploration and applications. *Pattern Recognition and Artificial Intelligence*, 11(8), December 1997.

[20] H.K. Huang and R.K. Taira. Infrastructure design of a picture archiving and communication system. *American Journal of Roentgenology*, 158:743–749, 1992.

[21] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *VLDB 1998*, pages 433–438. also available as CMU Technical Report: CMU-CS-98-119, 1998.

[22] P.M. Kelly and T.M. Cannon. CANDID: Comparison algorithm for navigating digital image databases. In *Proceedings of Seventh International Working Conference on Scientific and Statistical Database Management*, pages 252–258, September 1994.

[23] Y. Liu, R.T. Collins, and W.E. Rothfus. Automatic Bilateral Symmetry (Midsagittal) Plane Extraction from Pathological 3D Neuroradiological Images. *SPIE's International Symposium on Medical Imaging 1998*, 3338(161), February 1998.

[24] Y. Liu, W.E. Rothfus, and T. Kanade. Content-based 3d neuroradiologic image retrieval: Preliminary results. *IEEE workshop on Content-based accessof Image and Video Databases in conjunction with ICCV'98*, January 1998.

[25] F. Maes, A. Collignon, D. Vandermeulun, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187,198, 1997.

[26] J. Martin, A. Pentland, A. Sclaroff, and R. Kikinis. Characterization of neuropathological shape deformations. *IEEE Transaction son Pattern analysis and machine intelligence*, 20(2):97–112, February 1998.

[27] T. McInerney and D. Terzopoulos. Deformable models in medical images analysis: a survey. *Medical Image Analysis*, 1(2):91–108, 1996.

[28] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern analysis and machine intelligence*, 19(7):696–710, July 1997.

[29] B. Moghaddam, W. Wahid, and A. Pentland. Beyong eigenfaces: Probabilistic matching for face recognition. In *The 3rd IEEE Int'l Conference on Automatic Face and Gesture Recognition*, April 1998.

[30] A. W. Moore, D. J. Hill, and M. P. Johnson. An Empirical Investigation of Brute Force to choose Features, Smoothers and Function Approximators. In S. Hanson, S. Judd, and T. Petsche, editors, *Computational Learning Theory and Natural Learning Systems, Volume 3*. MIT Press, 1994.

[31] A. W. Moore, J. Schneider, and K. Deng. Efficient locally weighted polynomial regression predictions. In *Proceedings of the 1997 International Machine Learning Conference*. Morgan Kaufmann, 1997.

[32] M Moshfeghi. Elastic matching of multimodality medical images. *CVGIP:Graphical Models and Image Processing*, 53:271–282, 1991.

[33] Virage Inc. Home page. http://www.virage.com.

[34] E. Parzen. On estimation of a probability density function and mode. *Amn. Math. Stat.*, 33:1065–1076, September 1962.

[35] A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *IJCV*, 18(3):233–254, June 1996.

[36] R.W. Picard. A societry of models for video and image libraries. *IBM Systems Journal*, 1997.

[37] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *Proceeding of Sixth International Conference on Computer Vision*, pages 59,66. Narosa, 1998.

[38] Y. Rui, T.S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases VI*. volume 3312, January 1998.

[39] J. Schneider and A.W. Moore. A locally weighted learning tutorial using vizier 1.0. 1997.

[40] C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A. Aisen, and L. Broderick. Local versus global features for content-based image retrieval. In *IEEE workshop on content-based access of image and video libraries*, pages 30–34, June 1998.

[41] M.J. Swain. Interactive indexing into image databases. *SPIE*, 1908, 1993.

[42] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831,836, August 1996.

[43] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE International workshop on content-based access of image and video databases*, pages 42–51, January 1998.

[44] H.D. Tagare, C.C. Jaffe, and J. Duncan. Medical image databases: a content-based retrieval approach. *J Am Med Inform Assoc*, 4(3):184–198, May 1997.

[45] J. Talairach and P. Tournoux. *Co-Planar Steriotaxic Atlas of the Human Brain*. Thieme Medical Publishers, 1988.

[46] J.P. Thirion. Fast intensity-based non-rigid matching. In *Proc. of 2nd Intl. Symp. on Med. Robotics and Comp. Ass. Surgery*, pages 47–54, 1995.

[47] P.M. Thompson, C. Schwartz, R.T. Lin, A.A. Khan, and A.W. Toga. Three-dimansional statistical analysis of sulcal variability in the human brain. *Journal of Neuroscience*, 16(13):4261–74, 1996.

[48] P.M. Thompson, C. Schwartz, and A.W. Toga. High-resolution random mesh algorithms for creating a probabilistic 3d surface atlas of the human brain. *Neuroimage*, 3(1):19–34, Feb 1996.

[49] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Congnitive Neuroscience*, 3(1):71–86, 1991.

[50] A. Vailaya, A. Jain, and H.J. Zhang. On image classification: City vs. landscape. In *IEEE workshop on content-based access of image and video libraries*, pages 1–8, June 1998.

[51] N. Vasconcelos and A. Lippman. A Bayesian framework for semantic content characterization. In *The proceeding of CVPR*, pages 566–571, June 1998.

[52] P. Viola. Entropy, information, computer vision and image processing. In *Ph.D., MIT,*, 1995.

[53] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, March 1996.

[54] S.T. Wong and H.K. Huang. Design methods and architectural issues of integrated medical image data base systems. *Comput Med Imaging Graph*, 20(4):285–299, Jult 1996.