A Classification Based Similarity Metric for 3D Image Retrieval

Yanxi Liu and **Frank Dellaert**
Email: {yanxi,dellaert}@cs.cmu.edu
The Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
Keywords: bilateral symmetry, machine learning,image indexing, image semantics

ABSTRACT

We present a principled method of obtaining a weighted similarity metric for 3D image retrieval, firmly rooted in Bayes decision theory. The basic idea is to determine a set of most discriminative features by evaluating how well they perform on the task of classifying images according to predefined semantic categories. We propose this indirect method as a rigorous way to solve the difficult feature selection problem that comes up in most content based image retrieval tasks. The method is applied to normal and pathological neuroradiological CT images, where we take advantage of the fact that normal human brains present an approximate bilateral symmetry which is often absent in pathological brains. The quantitative evaluation of the retrieval system shows promising results.

1 Introduction

On-line image data is expanding rapidly in quantity, content and dimension. Existing *content-based* image retrieval systems [5, 7, 8, 15, 16] depend on general visual properties such as color and texture to classify diverse, two-dimensional (2D) images. However, subtle, domain-specific differences of image sets taken within a single domain, are difficult to capture using these global measures. Furthermore, little is known about three-dimensional (3D) volumetric image indexing, image semantics and systematic methods for feature selection. These are the motivations for our current research.

In this paper we present a principled method of obtaining a weighted similarity metric for 3D image retrieval, firmly rooted in Bayes decision theory. The basic idea is to determine a set of most discriminative features by evaluating how well they perform on the task of classifying images according to predefined semantic categories. Off-line, combinatorial search techniques are used to minimize a classification metric, *cross-entropy*, effectively generating a set of most useful features for image classification. These features are then used as the *index features* for image retrieval. We propose this indirect method as a more rigorous way to solve the difficult feature selection problem that comes up in most content based image retrieval tasks. It is our hypothesis that, since the user is mostly interested in the semantic content of the images, the metric that does well at *classifying* the images will also do well in finding similar images. Classification and retrieval share consistent evaluation criterions in terms of image semantic similarity.

3D grey-level volumetric images composed of normal and pathological neuroradiologic CT scans form the application domain for this research. These images are used as front-end indices to retrieve similar medical cases in a multimedia medical case database, like the National Medical Practice Knowledge Bank, currently under construction [10]. The purpose is to use the retrieval results to aid diagnosis, specialist consultation, patient surgical planning, and medical education. Image classification in this domain is to classify images based on their pathology (normal,blood,stroke), i.e. the semantics of an image. This domain also provides us with quantitative evaluation criterions to judge the performance of a retrieval system. Under this complete feature selection/retrieval evaluation loop the performance of a 3D image retrieval system can be improved quantitatively.

The remainder of this paper is structured as follows: in Section 2 we discuss 3D image preprocessing. Section 3 presents how the features are selected as potentially relevant to classifying images according to pathology type. Section 4 describes the machine learning framework and optimization techniques by which we search over a space of distance metrics for one that performs well for classification purposes. Section 5 defines a cost matrix which can be used to bias the classification result and provide a quantitative description of the performance of a classifier. In Section 6, we illustrate how this newly found metric can be used as a similarity metric for image retrieval, and we argue that quantitative evaluation criteria can be applied to the resulting system. We conclude with Section 7 where we also discuss future work.

2 3D Image Preprocessing

A 3D neuroradiology image can be expressed as a stack of 2D images (Figure 1). An *ideal head coordinate system* can be centered in the brain with positive X_0, Y_0 and Z_0 axes pointing in the right, anterior and superior directions respectively (Figure 2, 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,

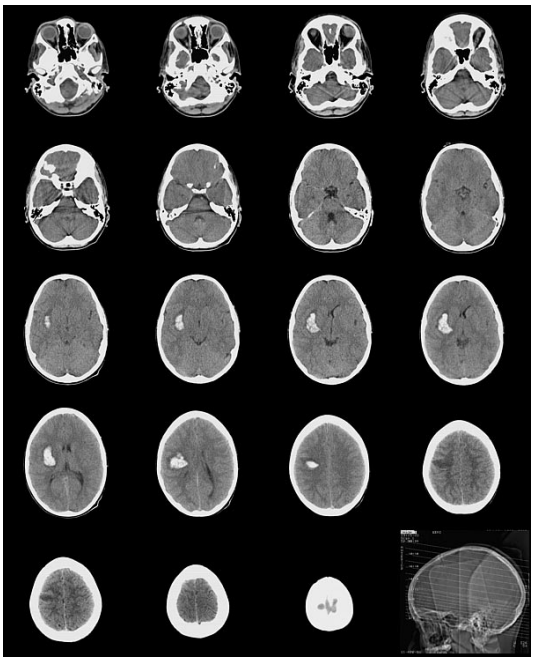

Figure 1: A 3D image which is composed of a set of clinical CT scans (axial), only a portion of a patient's head is captured as shown in a side view on the lower right corner. This is a case of acute right basal ganglion bleed.

but rather a *working 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 2, black coordinate axes). The orientation of the working 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.

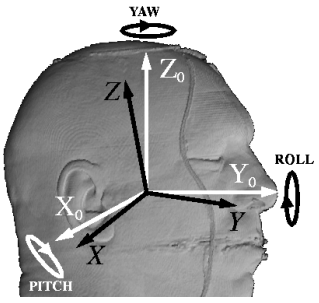

Figure 2: Ideal head coordinate system $X_0Y_0Z_0$ vs. the working coordinate system XYZ. Rendered head courtesy of the Visible Human Project.

Table 1 gives a sample of the 3D dataset we work with. They are all clinical CT images except the bottom MR image which is placed here for comparison. Notice, some of the voxel sizes of the CT images are far from cubical: the voxel height to its bottom-face-edge ratio can be as large as $10/0.5 = 20/1$. Compared with MR images, it is common that CT scans are usually sparsely and unevenly spaced (large, varied spacings between 2D slices) for minimizing radiation to the patients. If the images are collected from different scanners in different hospitals, each set of images may start and end at different portions of the brain and may be scanned in different angles or along different axes.

Table 1: A Sample of Input 3D Image Data

Size	Voxel (mm)	Pathology
512x512x23	0.5x0.5x5 (1-10) 0.5x0.5x8 (11-23)	Epidural Acute Bleed
512x512x21	0.5x0.5x5 (1-10) 0.5x0.5x8 (11-21)	Stroke
256x256x20	0.5x0.5x5 (1-11) 0.5x0.5x8 (12-20)	Stroke
514x518x19	0.5x0.5x5 (1-11) 0.5x0.5x10 (12-19)	Basal ganglion acute bleed
176x236x187	0.98x0.98x1.2 (1-187)	Normal

Until two 3D images are properly registered and segmented, existing techniques for content-based retrieval using color, texture, shape and position on 2D images cannot be applied directly for meaningful results. The current techniques for deformable registration of **dense** *normal* neurological images [18, 2, 4, 17, 11] are not directly applicable to pathological, sparsely- and partially-sampled brain images. For this type of data, no robust 3D registration algorithm yet exists that the authors know of. One basic observation of neuroradiologic images has helped us to get around the more difficult 3D image registration and segmentation problem stated above, that is *normal human brains present an approximate bilateral symmetry which is often absent in pathological brains*. Using this simple guidance, we developed a robust algorithm to automatically extract the ideal midsagittal plane (the plane $X_0 = 0$ in Figure 2) of a given 3D brain image [9]. The yaw and roll angle errors in the input images can be corrected by re-sampling the original 3D image. The region of interest (ROI) in a brain image can present itself as asymmetrical regions from simple comparisons of the two halves of the re-sampled brain image.

Strictly speaking, the pitch angle in each 3D image is not corrected after the midsagittal plane is identified. However, the images used in our experiments 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[1] while scanning. From now on we assume the pitch, yaw and roll angles are

[1] 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.

corrected in each 3D image, and the symmetry axis (plane) is centered in the middle of the image.

3 Feature Extraction

Each 3D image is composed of a set of 2D slices. Each 2D *half slice* is used as the basic descriptive unit for testing our approach. 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. This is also useful in identifying normal slices in an otherwise pathological 3D brain image, and which half of a pathological brain contains the lesion(s). Also, 2D images are easier to display than 3D images.

The first step of our approach is to extract a set F of easily computed features from each half slice. Three types of features are extracted: (1) Global statistical properties of the half slices are computed including features like the mean and standard deviation over the cropped slices. (2) Asymmetry features are obtained by comparing the half slices pixel-wise with their left-right counterpart. Two techniques are used to obtain those features. One is simply subtracting out a mirrored image of a 2D slice from the slice itself to obtain a difference image D, such that asymmetries show up as large positive or negative density values. Several features are counts of how many remaining pixels under different thresholds on D (significant difference pixels). The other technique: for each pixel, a Gaussian model of intensity is built, with its smoothing parameter set at 5, 9, or 15 respectively. Then the difference between the pixel and its counterpart in the symmetrical half slice is recorded. If that counterpart pixel fell significantly outside the estimated gray-value distribution, i.e. the difference with the estimated mean was greater than 3, 4, or 5 standard deviations, it was flagged as being significantly different. (3) By masking the original slice with the threshold images obtained in step (2), features were obtained that pertained only to these areas where asymmetries were present, i.e. values and differences within the asymmetrical regions.

There are total of 48 3D images which is 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. A total of 50 features are extracted from each 2D half slice of a 3D image.

4 Feature Selection by Classification

A set of discriminative features for image retrieval is found by evaluating how well the features perform on the task of classifying the images according to semantic categories. Three image categories are considered in this work. They are normal, blood (with light colored lesions) or stroke (i.e. infarct, with darker colored lesions). It is our hypothesis that the metric that does well at *classifying* the images will also do well in finding similar images, since similarity is exactly defined in semantic terms.

We use *Kernel Regression* (KR), a memory based learning (MBL) technique[1], to classify the images in different categories. MBL methods keep all the training data explicitly around, and calculate the posterior probability of a given a feature vector $\mathbf{x}$ by referring to all previously labeled instances. This can be implemented efficiently using augmented KD-trees[13].

We can compute the *posterior probability* $P(c|\mathbf{x})$ of a class c given $\mathbf{x}$ via Bayes law:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x}|c)P(c) + P(\mathbf{x}|\bar{c})P(\bar{c})} \qquad (1)$$

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:

$$P(c) \approx \hat{P}(c) = N_c/N \qquad (2)$$

The *conditional densities* $P(\mathbf{x}|c)$ in (1) can be approximated using Parzen window density estimation [6, 3]. Intuitively, this is done by placing an identical Gaussian kernel on each of the instances $\mathbf{x}_i$ in a given class, and approximating the density by a (appropriately normalized) sum of the identical Gaussian kernels:

$$P(\mathbf{x}|c) \approx \hat{P}(\mathbf{x}|c) = \frac{1}{N_c} \sum_{j \in c} G(\mathbf{x}, \mathbf{x}_j, \sigma) \qquad (3)$$

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

If we plug in (2) and (3) into Bayes law (1), it can be shown 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:

$$P(c|\mathbf{x}) \approx \hat{P}(c|\mathbf{x}) = \frac{\sum_i G(\mathbf{x}, \mathbf{x}_i, \sigma) f(\mathbf{x}_i)}{\sum_i G(\mathbf{x}, \mathbf{x}_i, \sigma)} \qquad (4)$$

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 $\mathbf{x}$.

A good set of discriminative features is then found by searching in the space of distance metrics for a metric that optimizes classification performance. Kernel regression, as most instance based methods, suffers from the curse of dimensionality: in high-dimensional spaces, it is hard to get a good coverage of the space such that the assumptions of the underlying density estimation techniques are satisfied. The set of potential features obtained in previous section is rather large, and many are presumed to be either not so relevant or to provide redundant information.

To look for a distance metric in this feature space, a combinatorial search over the space of features is

performed. Each distance metric is tested using leave-one-out cross-validation on a training set, containing roughly two-thirds of the available data. The rest of the data was set aside for evaluation purposes. The error metric we seek to minimize is *cross entropy* [3]:

$$E = -\sum_i \sum_c \delta_{ic} ln \hat{P}(c|\mathbf{x}_i) \quad (5)$$

where δ_{ic} represents the 1-of-m multiple class membership encoding. In classification problems, minimizing the cross-entropy is equivalent to maximizing the likelihood of the training data[3]. Thus, a distance metric that achieves this is presumed to yield good classification performance. The actual search was done using the general memory-based learning system (GMBL) developed by Andrew Moore et al. [12], adapted by us to calculate cross-entropy as an error criterion rather than classification error. This system applies a battery of combinatorial search strategies to the given problem, ranging from exhaustive search over small sets of features to hill climbing.

From about 50 candidate features extracted from each 2D half slice, typically 5 or 6 features are identified to be the most discriminative feature subset for best classification results. Often, we have found that there are several feature subsets that have quantitatively equivalent discriminating power.

5 Evaluation of the Distance Metric

When a given distance metric has been found from the last section, its performance is evaluated on an independent test-set, consisting of one third of the 2D slices in our dataset. It is convenient to view this evaluation in a decision-theoretic framework. A given distance metric defines a specific instance of kernel regression that outputs an estimated posterior probability for each of the classes. This needs to be translated into a classification decision, by minimizing the expected cost associated with choosing class α: $\alpha = \arg\min_i \sum_{j=1}^{s} \lambda(\alpha_i|c_j) P(c_j|\mathbf{x})$ where $\lambda(\alpha_i|c_j)$ is the cost incurred when choosing class α_i given that the true class is c_j[6].

In the domain of medical image retrieval, it is imperative to minimize the *false negative rate* **FNR**, i.e. minimize the occurrence of pathological cases (blood,stroke) being classified as normal. This motivates a *cost matrix* structure $C_{ij} = \lambda(\alpha_i|c_j)$, where a *false negative penalty* P is incurred for the two cases where either a blood or stroke image is classified as normal, whereas other cases simply incur a unit cost or zero, if 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:

$$C = \begin{bmatrix} 0 & P & P \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \quad (6)$$

In Figure 3 we show, for one particular distance metric, how false negative rate **FNR** varies with increasing P, along with these other performance metrics: the *overall classification rate* **CR**, *false positive rate* **FPR**, *true positive rate* **TPR**, and *confusion rate* **CFR**, as defined in Table 2. As can be seen in the figure, the classification performance **CR** drops as we attempt to decrease **FNR**, since increasing P will also lead to increasingly more false positives.

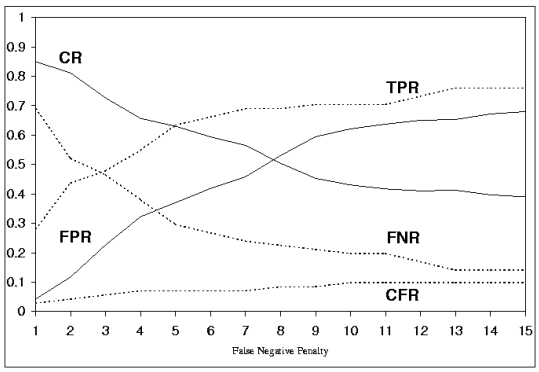

Figure 3: Classification rate (**PR**), True positive rate (**TPR**), False positive (**FPR**), False negative (**FNR**) and Confusion rates (**CFR**) vary while the false negative penalty P increases.

The use of the cost-matrix and such summary graphs provide a good quantitative overview of the strengths and weaknesses of a particular distance metric, and can also be used to select an appropriate value for P. This is important in scenarios where the classifier associated with the distance metric is used to aid the retrieval process, as discussed in the next section.

6 Application to Image Retrieval

A similarity metric has been found to perform well for classification, i.e machine learning is used to find the best semantically meaningful metric. The same distance metric 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.

Figure 4 shows a randomly chosen query image from the hold-out test set on the left, and its five nearest neighbors retrieved from the database on its right. The column B of Table 2 shows the evaluation on the retrieved result.

Several advantages to combine an image retrieval system with a classifier are

Smaller feature space: Current indexing feature space is reduced from 50 dimensions to 5 or 6 dimensions, retrieval speed is increased.

Unlabeled image retrievable: Either a labeled or

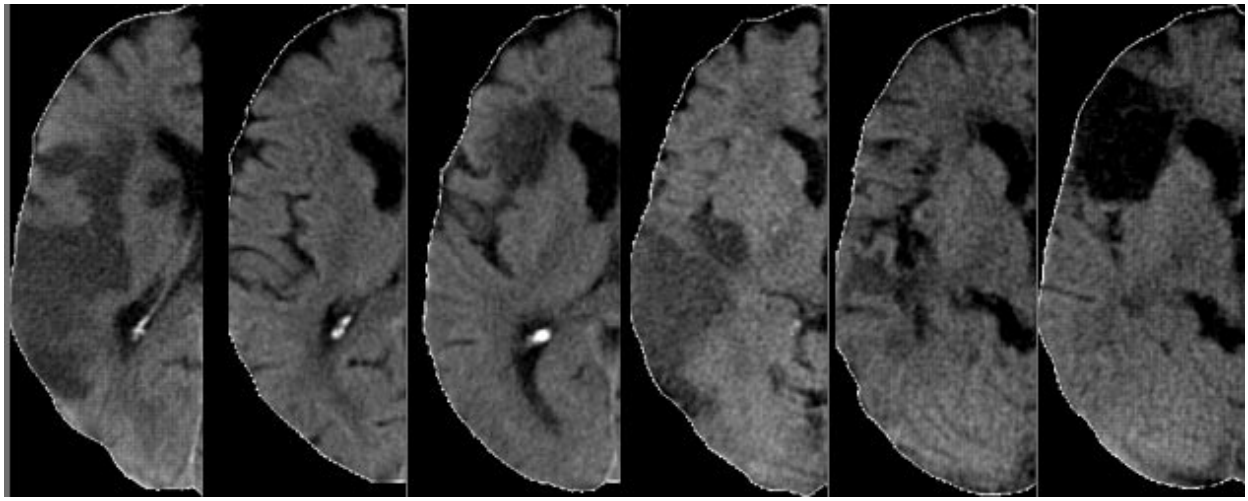

Figure 4: The first five retrieved half slices. The left most is the target image containing a stroke. Descending order of similarities is from left to right. The pathologies in the retrieved images are:normal (with partial volume effect),stroke, stroke,stroke,stroke.

unlabeled image database can be used for retrieval. In the case images are not labeled (in medical image databases, usually each 3D image is labeled as a whole, but its 2D brain slices are not labeled), the found classifier can be used in a decision theoretic framework to classify the images in accordance with a cost matrix.

Tunable performance: Besides more discriminative features are used for indexing, the performance of a chosen classifier can be tuned towards the performance a user desires by varying its cost matrix.

Consistent quantitative evaluations: Once the retrieved images are labeled, a set of quantitative evaluations can be given to the retrieved results as defined in Table 2. The same evaluations can be applied to classification results or retrieval results.

Multidimensional and multimedia data retrieval: Given a 2D slice, a set of similar slices are retrieved. If the user desires, the 3D images containing these slices and their collateral information can also be retrieved.

Although the classifications are done on 2D half slices, one can attempt to classify 3D images as well. We have experimented with a varying cost matrix and a simple rule that says if the same pathology appears in more than 2 adjacent 2D slices, the 3D brain is considered pathological. The column A in Table 2 shows the classification result on 3D images when $P = 4$ in its cost matrix.

7 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. It is instructive to compare this with the Rosetta system of [5]. There, the semantic category is defined on the fly by querying with a *set* of query images, and the statistics of the query set are used to construct a metric. 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 image. In addition, our search method can eliminate non-discriminative features that are common to all categories. Thus, whereas Rosetta is more a general purpose tool, our approach might be better suited for application specific databases where the differences between semantic categories are subtle, as in the medical imaging domain.

3D image retrieval is a relatively new area, especially when one is seeking for the semantics of the images. Our initial attempt has shown that given the proper application domain and an effective feature selection scheme we can achieve good results for retrieving relevant cases using a 3D image database. We believe that feature selection by image classification is a powerful tool for any image retrieval practice regardless of image dimensionality. The experimental results support our hypothesis that the distance metric that does well at *classifying* the images will also perform well as a similarity metric for image retrieval.

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. In our case, the bilateral symmetry property in normal brains has been exploited to construct features predictive of pathologies due to asymmetry. One weakness in demonstrating the effectiveness of this approach is the small number of 3D images we have. We expect the database to grow much larger in the near future, so that a larger hold-out test set can be used.

Future work includes the study of different sets

Table 2: Evaluation Measurements for Classification and Retrieval: A - classification result on hold-out test 3D images, B - retrieval result (top 5 only) for an infarct query image (Figure 4)

Measurement Definitions	A	B
T = # of instances	17	5
P = # pathological instances in T	7	5
N = # of normal instances in T	10	0
TP= # of correctly classified pathological instances	7	4
P_as_N=# pathological instances which are classified as normal	0	1
N_as_P=# normal instances which are classified as pathological	6	0
B_as_S= # of instances which are blood but classified as stroke	0	0
S_as_B=# instances which are stroke but classified as blood	0	0
CR = 1 - (P_as_N + N_as_P + B_as_S + S_as_B) / T	0.65	0.8
TPR = TP/P	1	0.8
FNR = P_as_N / P	0	0.2
FPR = N_as_P /N	0.6	0
CFR = (B_as_S + S_as_B) / P	0	0

of most discriminative features for different purposes or on different subsets of the database, and dynamic switching from one set of features 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, the dominating cue for further classifications would be lesion texture. We expect feature weights to vary accordingly from the initial classification. An analogy can be made between *classification based image retrieval* and appearance-based learning [14], where an object recognition process is followed by a pose estimation process performed on different appearance manifolds. Finally, 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. The basic framework presented here has provided us with such an *information fusion* capability.

8 Acknowledgement

The authors would like to thank Dr. Rothfus of Radiologic Sciences, Allegheny General Hospital for his medical guidance, Prof. Andrew Moore and Dr. Jeff Schneider of CMU for helpful discussions. This research is partly supported by the Allegheny-Singer Research Institute under prime contract through the National Institute of Standards and Technology (NIST#70NANB5H1183).

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] Christopher M. Bishop. *Neural Networks and Pattern Recognition*. Oxford, 1995.

[4] 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.

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

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

[7] 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.

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

[9] 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.

[10] Y. Liu, W. Rothfus, and T. Kanade. Content-based 3d neuroradiologic image retrieval: Preliminary results. *IEEE CAIVD'98, in conjunction with ICCV*, pages 91,100, January 1998.

[11] 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.

[12] 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.

[13] Andrew W. Moore, Jeff Schneider, and Kan Deng. Efficient locally weighted polynomial regression predictions. In *Proceedings of the 1997 International Machine Learning Conference*. Morgan Kaufmann, 1997.

[14] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

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

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

[17] 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.

[18] 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.