Content-based Image Retrieval for Alzheimer's Disease Detection

Mayank Agarwal
Lexical Informatics
New Delhi 110063, INDIA
Email: mayank@lexicalinfo.com

Javed Mostafa
Biomedical Research and Imaging Center &
NC Translational Clinical Sciences Institute
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3306, USA
Email: jm@unc.edu

Abstract

This paper describes ViewFinder Medicine (vfM) as an application of content-based image retrieval to the domain of Alzheimer's disease and medical imaging in general. The system follows a multi-tier architecture which provides the flexibility in experimenting with different representation, classification, ranking and feedback techniques. Classification is central to the system because besides providing an estimate of what stage of the disease the input query may belong to, it also helps adapt and rank the search results. It was found that using our multi-level approach, the classification performance matched the best result reported in the medical imaging literature. Up to 87% of patients were correctly classified in their respective classes, leading to an average precision of about 0.8 without any relevance feedback from the user. To encourage engagement and leverage physicians' knowledge, a relevance feedback function was subsequently added and as result precision improved to 0.89.

1 Intoduction

The rate at which medical images are produced everyday is increasing exponentially. Such images are a rich source of information about shape, color and texture, which can be exploited to improve the diagnosis and ultimately the treatment of complex diseases. Alzheimer's Disease (AD) has been successfully associated with structural changes in the brain. However, with the volume of Magnetic Resonance Imaging (MRI) scans growing at a rapid rate, it is becoming increasingly difficult to perform a search over these scans. Research has shown that the performance of clinicians has improved through the use of content-based image retrieval systems [2].

Typically, a radiologist or a physician diagnosing a particular patient is also interested in finding similar cases pertinent to the case in hand. Such information in fact is mutually beneficial to both, the physician and the system. While the physician can use this information to arrive at an informed decision about the patient's health condition, the system uses this information for better organization.

Combining texual and visual elements in a single query model and user interface is a challenge. The problem is compouned due to the lack of of a translational model between the user's information need, which is typically at a higher semantic level, and the low level image features. A fusion model based system such as ViewFinder Medicine (vfM) that leverages the advantages of both, text and visual retrieval, and brings the physicians in the loop, can be useful in such an environment.

In this paper, a fusion model and framework, vfM, for content-based retrieval and access of images, is proposed. vfM combines the textual information, such as various biomarkers and cognitive scores, with the low level features of the MRI scans to classify and eventually find similar scans. The clinicians play a central role in the retrieval process indicating how well the system is performing in terms of relevance of the search results. In the following sections, we describe the vfM architecture. The system performance is demonstrated and discussed through experiments and results, and the conclusions are drawn in the final sections.

2 Methods

2.1 vfM System Architecture

vfM builds upon the multi-level information retrieval architecture. Critical components of the system include representing the content in the right format, a classifier and a ranking engine that includes feedback mechanism. Both local and global texture are used to index and search images along with the cognitive score markers. A multinomial logistic regression model is trained to identify the statistically significant features. Singular Value Decomposition (SVD) is conducted to obtain an independent feature sub-space. A Support Vector Machine (SVM) is used to train a radial-basis function kernel classifier to sort the images into multiple classes. Classes are combined while performing Euclidean distance based matching, based on the probability of the image falling in a particular class. Subsequently, an inter-session and intra-session feedback is generated. In general, the relevance feedback mechanism involves presenting the user with an initial set of results and requiring him/her to generate an estimate of how well the system performed its task. The system then utilizes the feedback to improve the retrieval in subsequent sessions in an iterative manner. Ideally we hope that our system, improved with relevance feedback mechanism, will satisfy the users information need in as few steps as possible, optimally requiring only one or two iterations. In the following sections we discuss each component in detail.

2.2 Feature Representation and Indexing

To describe the visual content of interest it is important to understand the structural changes in the brain that are responsible for AD, which has been associated with volumetric reductions in medial temporal lobes and hippocampal regions of the brain [13].

Inter-subject and intra-subject registration is conducted on the MRI scans to align them to the same space. The brain volume is extracted from the registered scans. The medial temporal lobes and hippocampus regions are then identified using Harvard-Oxford cortical and subcortical structural atlases. The cortical regions thus extracted are then segmented into the three tissue types: gray matter, white matter and cerebrospinal fluid. The segmentation method is elaborated in [1].

2.2.1 Visual Features

Both global and local texture features have been used to characterize AD. Global features include image textures computed using Discrete Cosine Transform (DCT) and Daubechie's Wavelet Transform (DWT).

DCT is an integer transform widely used to characterize texture of an image, e.g., in the JPEG compression standard. Due to the excellent energy-compaction properties of DCT, much of the information is concentrated in few low-frequency components [14].

DWT has been used for decades in computer vision and robotics to characterize the image texture. Recently it has been used in content-based image retrieval systems to aid texture based search and retrieval [8, 17]. A 2D-DWT divides the input image into 4 sub-bands at the first level of decomposition. Each approximation sub-band is further divided recursively into 4 sub-bands till the desired level is achieved. DWT works on the principle that the most prominent information in the signal appears in high amplitudes and the less prominent information appears in very low amplitudes and thus is able to compress the information contained for analysis.

Local Binary Pattern (LBP) is a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood [11, 12]. For each pixel of the image a binary code is produced by thresholding the pixel value by center value of its neighborhood. LBP is designed such that it is invariant to changes in illumination and intensity, and can be easily adapted to make it robust against rotation as well. The application of LBP to medical images and specifically MRI images has been explored in [15].

Based on discussions with radiologists a subset of slices relevant to detect Alzheimer's disease has been identified. On a slice by slice basis, 2D-DCT is computed on segmented gray matter, white matter, and cerebrospinal fluid to get an arrangement of frequency components. Since most of the frequency components are concentrated at the beginning of the 2D matrix, a lower dimensional matrix is selected to optimally capture most common frequency components. Normalized average of the three sub-matrices thus generated is taken which is then stored linearly to make up the DCT feature.

For DWT, Daubechies's D4, 2-level, 2D wavelets are used. The wavelets are computed on a per-slice basis for both, the skull-removed brain and the segmented cortical regions. The 2D matrix forms the DWT component of the feature vector.

LBP is computed for each pixel in a slice by considering the neighboring pixels. The radius of the neighborhood circle is varied from 3 to 5 to generate various LBP features.

2.2.2 Textual Features

The imaging data is accompanied by textual information including the detailed demographic information of the patients, and the scores for various cognitive tests conducted to identify any possible dementia. Cognitive scores have been established as reliable indicators of the patient's condition [4]. Mini-Mental State Examination (MMSE) has been proved to accurately predict cognitive decline [6]. Clinical Dementia Rating (CDR) uses six domains of cognitive and functional performance: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care; to identify the possible cognitive decline [3]. Multinomial logistic regression was used to identify the variables that contribute to the prediction of Alzheimer's disease. It was found that Age, MMSE, and CDR contribute significantly to the predicting the right class.

2.2.3 Fusion Model

The visual and textual features once generated are combined together to form a feature space for the further processing. The fusion model is based on multinomial regression training on the generated visual and textual characteristics. Dimensionality of the resulting feature vector was reduced by applying SVD to identify the independent dimensions of the feature space. The sub-space thus found is used to conduct the classification and retrieval. The feature is represented as a 2D matrix and will be refrerred to as a feature matrix.

2.3 Classification

When a user searches for a particular scan, the system first automatically classifies the image into one of the three possible classes of AD, MCI, and normal. SVMs are arguably one of the best classification schemes for generic data, and their application has been explored in [9, 16, 5]. The classifier plays a dual role in the context of vfM: at a basic level the classifier is important to give radiologist or physician an estimate of the class to which any new scan may potentially belong to; for a physician interested in finding similar cases, this is important as it gives the systems insight about what may be the scenario, and may well provide the first step towards an accurate diagnosis. During the actual retrieval operation, the estimate of class can further be utilized to enhance the retrieval performance of vfM

2.4 Retrieval

Once the input scan is classified the distance between the input scan and the rest of the images belonging to that class is calculated. The results are ranked on the basis of Euclidean distance.

$$R = p \times \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - y_{ij})^2} \tag{1}$$

The decision estimates provided by the classifier is combined with the distance calculated to rank the search results. The rank R for a search result is estimated by Equation 1 where p indicates the estimate that the input query belong to a particular class.

2.5 Relevance feedback

Due to the presence of initial gaps in the query representation in capturing users information need, the retrieval results in the initial presentation may not be satisfactory. The problem is more pronounced in image retrieval due to the ambiguities and uncertainty that arise while interpreting an image. The uncertainty makes it necessary to seek additional information from the query initiator as feedback to improve performance. vfM assumes binary relevance (yes or no). The system uses the user-provided feedback to perform inter-session and intra-session re-ranking.

A session is defined in terms of steps involved in answering a single query. Initiation of a new query indicates the start of a new session. Intra-session feedback refers to using the user-feedback to improve the search results for the query currently under execution. Inter-session feedback leverages the user-feedback collected in previous query or queries to improve the search results for the current query. Figure 1 shows how intra-session and inter-session feedback are used in vfM. With this framework, a feedback loop is defined as a cycle during which the user generates relevance value for the search results and system uses those values to improve the quality of search results. The re-ranking procedure is explained in detail in the following sections.

2.5.1 Intra-session feedback Re-ranking

During the course of a single query session upon presentation of the first set of results if the user wishes to look at results beyond the first set, the system tries to re-rank the rest of the results for the query based on the feedback provided by the user. The re-ranking in this case can be thought of as an optimization problem. Depending on a pre-selected threshold, for every query there are two sets of images; relevant and not relevant. The rank of any image in the retrieved set can be expressed in terms of the query image , the set of relevant images and the set of non-relevant images . The rank for image can be computed by modifying Equation 1 as:

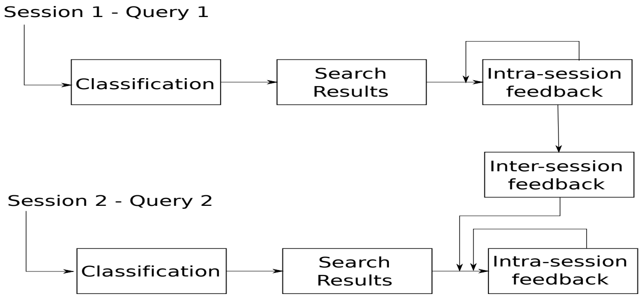

Figure 1. Intra-session and inter-session feedback

$$R = p \times D \times \frac{min\{d_{ik}\}}{max\{d_{jk}\}} \tag{2}$$

2.5.2 Inter-session feedback Re-ranking

vfMs architecture can be leveraged to constantly improve its performance. By the time, the user gives feedback to the system, the query image has already been classified into one of the three possible classes. For each session the system logs the images that have been marked relevant for each class and creates a profile for each class. This profile is called a class-profile. When a user initiates a new query, the image is classified into one of the three classes. Using the predicted class the rankings for each image in the retrieved set is calculated based on the class-profile according to the Equation 2. The class-proflie gets richer with each feedback loop.

3 Experiments

To evaluate the system performance experiments were conducted using the MRI data. Alzheimers Disease Neuroimaging Initiative, ADNI [1] is a five-year research grant to study the rate of decline or atrophy associated with MCI and AD and to provide a large collection of image dataset that can be used for clinical trials and improved diagnosis for AD. A subset of the ADNI collection containing 354 cases with a total of 1020 scans, with an average of 2.88 scans per subject has been used for this study. All the MR scans were T1 weighted contrast enhanced MP-RAGE images.

3.1 Performance Measures

Effectiveness and efficiency are two important performance dimensions for a retrieval system. Effectiveness defines how well the underlying algorithms are performing in

[1] http://www.loni.ucla.edu/ADNI/

Table 1. Average Precision at cut-off without classification, with classification and with feedback and classification

Avg Precision	Without Classification	With Classification	With Feedback
at 10	0.420	0.816	0.895
at 20	0.380	0.783	0.889
at 30	0.370	0.776	0.8831
at 40	0.334	0.750	0.88
at 50	0.332	0.730	0.875

terms of providing accurate results. Efficiency is concerned with how fast the underlying algorithms and the whole system are performing. In [10] it was acknowledged that only a limited number of studies have considered time as a performance measure.

In this study, 10 fold cross-validation was used to establish the classifier accuracy. The classification accuracy indicates the percentage of images or scans that were classified correctly in their respective classes. The retrieval performance is measured in terms of precision at cut-off. In vfM, the user is presented with a “matrix” of eight images with the top left image representing the closest search result, and other results arranged in a descending order of their ranks. The average performance measure is calculated over 25 randomly selected queries. The queries were distributed with 9 queries in AD, 8 in MCI and 8 belonging to normal class.

3.2 Effectiveness Analysis

The first set of experiments were conducted to estimate the efficiency of using a classification module in the system. With this goal the retrieval performance without classification was compared with retrieval performance with classification. In table 1 the retrieval performance for whole brain images, with and without DCT-based classfication, were compared. The high frequency DCT components were computed with a window size of 5×5. It was noted that the system performance improved with a classifier in between. These first set of experiments led us to further explore the role of the classifier in detail.

The second set of experiements was conducted to measure the accuracy of classification and its effect on retrieval performance. We compare the classification accuracy using DCT, DWT, LBP and their possible combinations. The classification accuracy with the above models was contrasted with the accuracy with the fusion model. It should be mentioned that all the scans were segmented into corresponding gray matter, white matter, and cerebrospinal fluid. It was noted that the classification accuracy with skull-information removed is siginificantly higher ($p = 0.022$) than classification accuracy with whole-brain images. No statistically significant difference was found between the classification accuracy using whole brain and only the hippocampus and skull removed brain and the hippocampus.

It was also found that the performance with LBP is significantly higher than with DCT or DWT ($p = 0.015$ and $p = 0.005$ respectively). The fusion model leads to a maximum classification accuracy of 86.7%.

The fusion model was based on an experiment showing that, using only the limited information provided by the available metadata, it is possible to train a high accuracy classifier. An SVM based classifier was trained and tested using the cognitive scores and age as feature to achieve a classification accuracy of 99%. The high accuracy of a text based classifier allowed us to hypothesize that combining the available metadata with the visual features should improve the classifier performance.

Exploring the classification results further led to examining the classification accuracy for each class. Based on experimental observation it was noted that the maximum classification accuracy was achieved for MCI class. The results exceed the previously published results for classifying the images using skull removed brain into MCI and Control [7]. The classification accuracy using the hippocampus region is 86.69% which also exceeds the state-of-the-art classification accuracy of 85.6% reported in [7].

Literature suggests that increasing the amount of available information should improve the system performance. Experiments were conducted to gauge the effect of centrality and window size on classification and retrieval as a measure of varying available information. Centrality defines the number of slices we chose to represent the scan. With a centrality of 11, 21 and 151 the classfication accuracy achieved was 86.7%, 87.2% and 85.8% respectively. The system uses two window sizes. The first window size refers to the number of DCT coefficients used as the feature. The second window size refers to the number of pixels considered in the neighborhood of any pixel while computing LBP.

Table 2. Effect of DCT window size on classification accuracy

Window Size	Classification Accuracy
3×3	69.7279
5×5	71.7687
7×7	70.068
9×9	70.7483

Table 3. Effect of LBP window size on classification accuracy

Window Size	Classification Accuracy
3×3	85.76
4×4	85.91
5×5	85.76

From Table 2 and Table 3 it can be seen that increasing the amount of information available to the system does not necessarily improve the classification or retrieval performance.

From these (apparently counter-intuitive) results, it can be inferred that, since most of the components are clustered at the beginning of the feature matrix, adding more information does not necessarily improve the quality of the available information and does not contribute any new knowledge.

The amount of information available to the system can also be varied by processing the input image from which the features are generated. It was interesting to note the effect of varying segmentation levels on retrieval performance. Using the fusion model, the effect of segmentation on retrieval performance is summarized in Table 4. Though there is an increase in the average precision for skull removed brain, it is not statistically significant. This is in contrast with the significant increase in the classification accuracy with segmentation.

Table 4. Effect of varying segmentation level on retrieval performance

Avg Precision	Whole Brain	Without Skull Brain	Hippocampus
at 10	0.8422	0.8567	0.8464
at 20	0.8486	0.8412	0.841
at 30	0.8306	0.84	0.8296
at 40	0.827	0.8274	0.8264
at 50	0.808	0.8122	0.81

Following the analysis of classification, we wanted to understand its effect on retrieval performance. Table 5 highlights the effect of varying LBP window size on the average preicision of the system using LBP based classifier. It can be noted that highest retrieval performance was obtained for a window size of 3×3.

Table 5. Effect of varying LBP window size on retrieval performance

Window Size	3×3	4×4	5×5
Precision at 10	0.8342	0.8138	0.8103
Precision at 20	0.791	0.7792	0.7585
Precision at 30	0.77	0.7665	0.756
Precision at 40	0.7685	0.7603	0.7617
Precision at 50	0.75	0.741	0.732

To evaluate the system performance with relevance feedback it is important to model the user selection. The user in vfM is modeled according to the following rules:

1. Any image (i.e. scan) that does not belong to the estimated class is considered non-relevant.

2. Any other image is assigned relevant or irrelevant based on a normal probability on the distance between the search result and the query.

The set of distances between the the query image X and the search result is denoted by $D = \{d_{1X}, d_{2X} \ldots\}$, and the probability of an image being relevant to the query image is defined according to the following equation:

$$p_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(d_{iX}-\mu)^2}{2\sigma^2}} \ \forall d_{iX} \in D \qquad (3)$$

where σ and μ represents the standard deviation and the mean of D. Depending on the threshold value p_t, an image is marked relevant only if $p_i \geq p_t$ where p_t can take any value between 0 and 1.

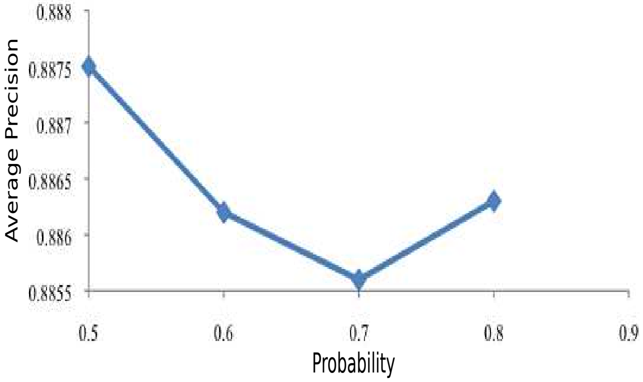

Figure 2. Effect of varying probability threshold on average precision

From Table 1 it can be seen that the performance of the system with relevance feedback is considerably higher than without feedback. The results observed are due to the promotion of relevant images to the top of the list. When a user initiates a new query, the inter-session re-ranking of the images considers the feedback provided by the other users at the class level besides taking into account the confidence in the class prediction. As those images that are marked as relevant influence the learning process, we also wanted to study the effect of varying the threshold for determining relevance on the retrieval performance.

3.3 Efficiency Analysis

vfM was conceptualized with an aim towards providing almost real-time access to images relevant to the scan in question. Not many researchers have attempted to define efficiency in this way. With a growing collection of MRI scans and relatively short time available to clinicians for diagnosis, it becomes imperative to bring this aspect of system performance into medical image retrieval.

Experiments were conducted to characterize the times required to build a classification model and to classify a scan. The different times were contrasted for the different types of features (DCT, DWT, LBP, fusion model) and their combinations. The time efficiency was also noted by varying the window sizes for DCT and LBP.

Figure 3 highlights the time needed to train the classifier along with the time needed to classify any scan in one of three classes. As can be seen, as the number of features increases, the times needed to train the model and for classification increase accordingly. The important observation here is that,

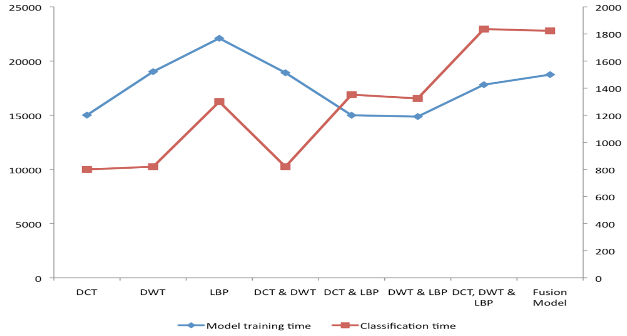

Figure 3. Model training and classification time for different feature types

although the training time is large, model generation is only a one time process. Once a model has been generated, when a new scan comes in, it only needs to be classified into one the classes using the model that was generated offline. The classification time is not large and works well for a real-time system.

4 Discussion and Conclusion

Traditional retrieval models, including those designed to handle images, have become inadequate as image collections have grown in size and complexity. Relevance feedback is a means to leverage users impact (and expertise) to improve retrieval performance. Most of the current relevance feedback schemes for information retrieval are designed to treat feedback as input to learning or a classification procedure. This study validates a realistic content-based image feedback retrieval framework that physicians can use for retrieving cohorts to assist them in diagnosing a condition or analyzing a case. Using vfM we demonstrate a means to employ texture as the fundamental description of MRI scans, combine it with associated textual information, and learn from previous queries to produce optimal search results. The results show that vfM can successfully leverage user feedback and associated information to achieve retrieval performance that exceeds the best results reported so far.

5 Acknowledgement

Partial funding for this paper came from UNC BRIC and the CTSA grant.

References

[1] M. Agarwal and J. Mostafa. Image retrieval for alzheimer's disease detection. In *Proceedings of Medical Content-based Retrieval for Clinical Decision Support (MCBR-CDS) Workshop in conjunction with MICCAI*, 2009.

[2] A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori. Automated Storage and Retrieval of Thin-section CT Images to Assist Diagnosis: System Description and Preliminary Assessment. *Radiology*, 228(1):265–270, 2003.

[3] L. Berg. Clinical dementia rating (cdr). *Psychopharmacol Bull*, 24(4):637–9, 1988.

[4] J. Brown, G. Pengas, K. Dawson, L. A. Brown, and P. Clatworthy. Self administered cognitive screening test (TYM) for detection of Alzheimer's disease: cross sectional study. *BMJ*, 338:b2030, 2009.

[5] X. Du, Y. Li, and D. Yao. A Support Vector Machine Based Algorithm for Magnetic Resonance Image Segmentation. In *ICNC '08: Proceedings of the 2008 Fourth International Conference on Natural Computation*, pages 49–53, 2008.

[6] M. F. Folstein, S. E. Folstein, and P. R. McHugh. "mini-mental state" : A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189 – 198, 1975.

[7] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr., J. Ashburner, and R. S. J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[8] M. Lee and C. Pun. Rotation and scale invariant wavelet feature for content-based texture image retrieval. *Journal of the American Society for Information Science and Technology*, 54(1):68–80, 2003.

[9] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005. Special Section: Social Cognitive Neuroscience.

[10] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23, 2004.

[11] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *Computer Vision - ECCV 2000*, pages 404–420, 2000.

[12] M. Pietikäinen. *Image Analysis*, chapter Image Analysis with Local Binary Pattern. Lecture Notes in Computer Science. Springer Berlin, 2005.

[13] B. H. Ridha, J. Barnes, L. A. van de Pol, J. M. Schott, R. G. Boyes, M. M. Siddique, M. N. Rossor, P. Scheltens, and N. C. Fox. Application of automated medial temporal lobe atrophy scale to alzheimer disease. *Archives of Neurology*, 64(6):849–854, June 2007.

[14] P. K. Singh. Unsupervised segmentation of medical images using dct coefficients. In *VIP '05: Proceedings of the Pan-Sydney area workshop on Visual information processing*, pages 75–81, 2004.

[15] D. Unay, A. Ekin, M. Cetin, R. Jasinschi, and A. Ercil. Robustness of local binary patterns in brain mr image analysis. In *EMBS 2007. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007.*, pages 2098–2101, August 2007.

[16] C.-M. Wang, X.-X. Mai, G.-C. Lin, and C.-T. Kuo. Classification for Breast MRI Using Support Vector Machine. In *CIT-WORKSHOPS '08: Proceedings of the 2008 IEEE 8th International Conference on Computer and Information Technology Workshops*, pages 362–367, 2008.

[17] J. Z. Wang. Pathfinder: multiresolution region-based searching of pathology images using IRM. In *Proceedings of AMIA Annual Symposium*, pages 883–887, 2000.