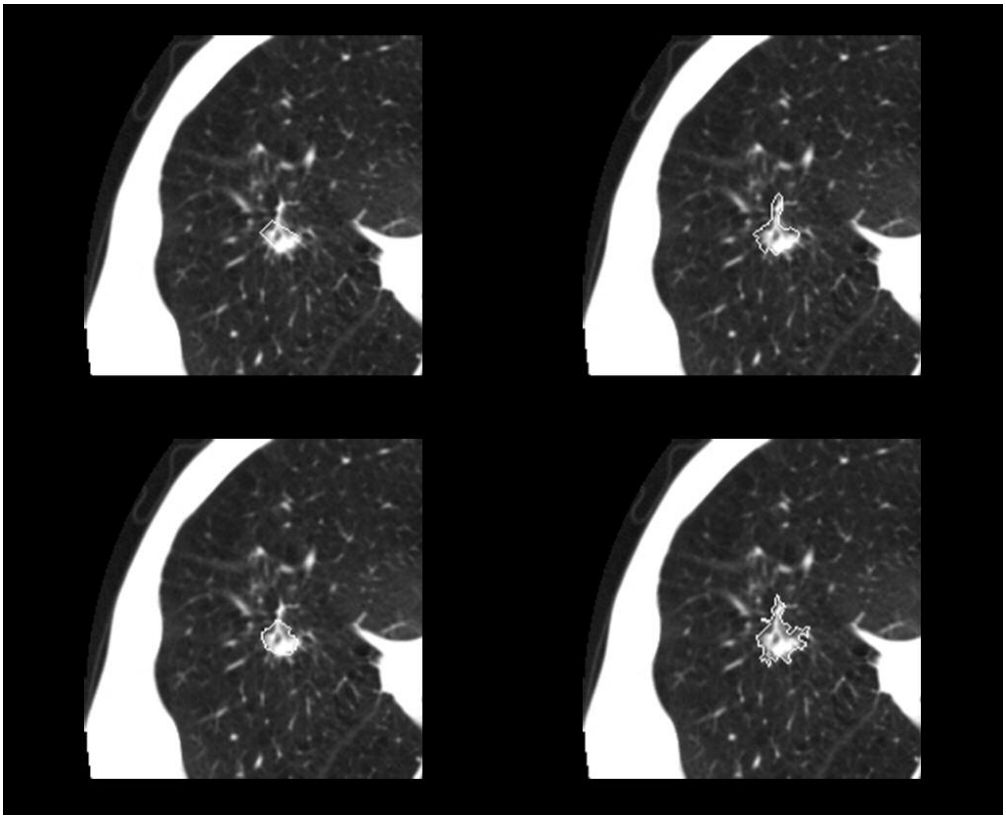
Springer

Texture Model Comparison for Lung Nodules Interpretation and Retrieval

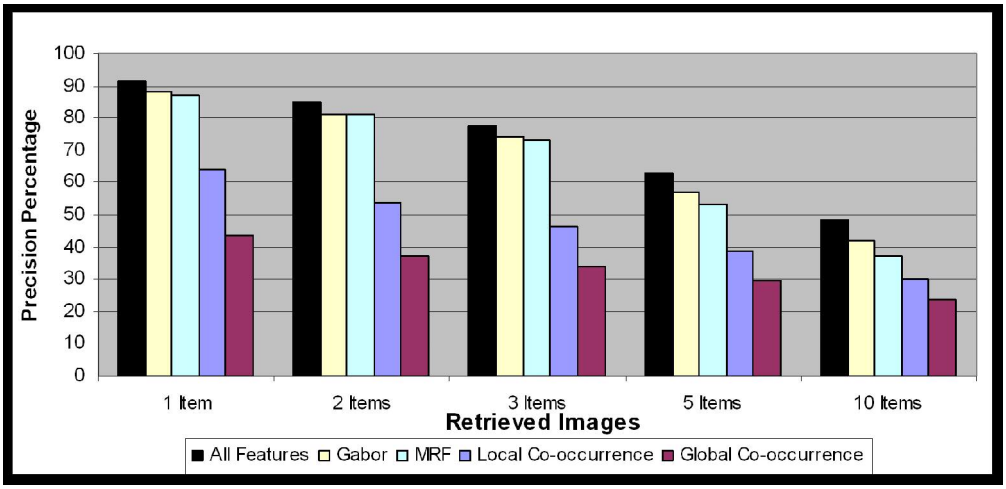
Journal:	*Journal of Digital Imaging*
Manuscript ID:	draft
Manuscript Type:	Hypothesis-Driven Research
Keywords:	Computed Tomography, Content-Based Image Retrieval, Image Processing, Image Retrieval, Lung
Additional Keywords:	Texture Extraction, co-occurrence, BRISC

Example CT slice with nodule [calcification = 6, internal structure = 1, lobulation = 3, malignancy = 5, sphericity = 3, speculation = 3, subtlety = 5, texture = 5, and margin = 3]

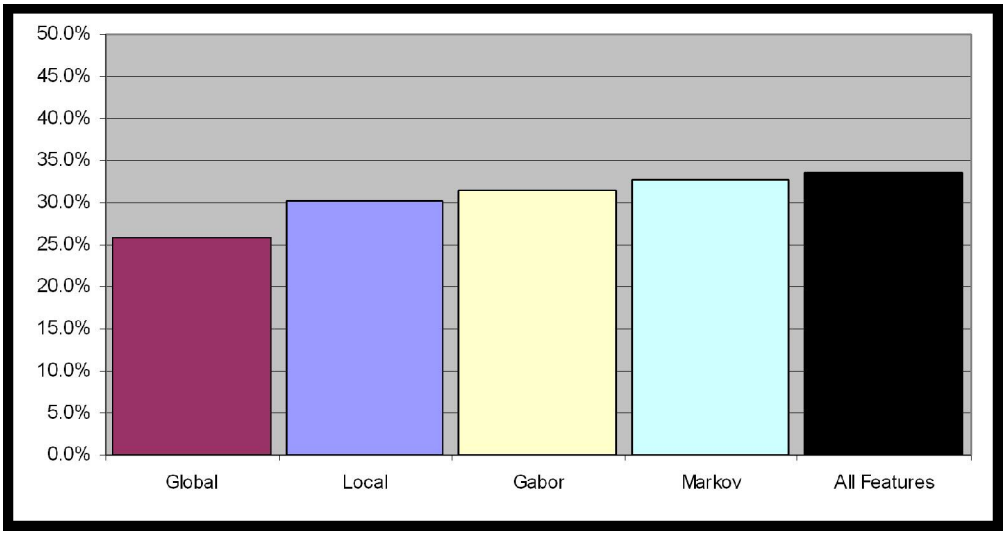
45x29mm (300 x 300 DPI)



Four distinct outlines for a nodule delineated by four radiologists
86x70mm (300 x 300 DPI)

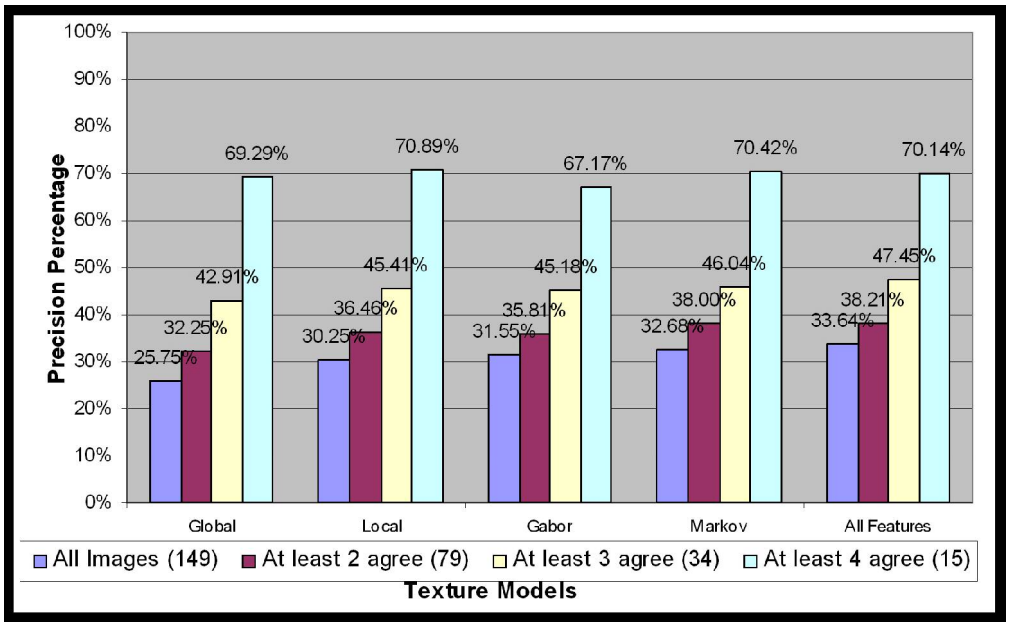


Images Retrieved Comparison
152x73mm (300 x 300 DPI)



Comparison of the texture models using the Relevance Index defined in equation (1). The number of retrieved images under consideration is n=10
150x79mm (300 x 300 DPI)

	Global	Local	Gabor	Markov
Euclidean	25.75%			
Manhattan	25.61%			
Chebychev	25.43%			
Jeffrey Div.		30.25%	30.81%	24.40%
Chi-Square		30.25%	31.55%	32.68%



Comparison of Radiologist Agreement Evaluated Using the Second Method (10 Items Returned)

150x93mm (300 x 300 DPI)

	Texture Models	Evaluation Method 1	Evaluation Method 2	Radiologist Annotation (Eval. #2) 4 Agree
1	All Features	91.14	33.64	70.14
2	MRF	87	32.68	70.42
3	Gabor	88	31.55	67.17
4	local co-occurrence	64.21	30.25	70.89

1. Introduction

In medicine to date, virtually all Picture Archiving and Communications Systems (PACS) retrieve images simply by textual indices based on patient name, technique, or some observer-coded text of diagnostic findings [23]. Fields of text tags, such as patient demographics, diagnostic codes (e.g. ICD-9, American College of Radiology diagnostic codes), image view-plane (e.g. saggital, coronal, etc) and so on usually are the first handles on this process. This textual approach, however, may suffer considerably from observer variability, high cost of manual classification and manipulation of images by medical experts, and failure to fully account for quantitative relationships of medically relevant structures within an image that are visible to a trained observer but not codable in conventional database terms. The *objective* of this paper is to extend our previous work [1] on content-based image retrieval for lung nodules in Computed Tomography studies. This work is significant given the fact that there were an estimated 160,390 deaths in the United States due to lung cancer in 2007 [2] and lung cancer accounts for around 29% of all cancer deaths [3].

The hypothesis is that the uncertainty of the radiologist in identifying suspicious lesions can be reduced by providing a visual comparison of a given lesion to a collection of similar lesions of known pathology. To test this hypothesis, we propose to develop a CBIR system whose similarity results match the human perception. The human observer (radiologist) manually (or semi-automatically or automatically) segments a lesion from a clinical case. The system computes a set of quantitative descriptors for that lesion and compares those descriptors to the descriptors of known lesions. The underlying assertion is that if a known malignant lesion has certain computable features then unknown lesions with similar computable features would be malignant. Simply put, the *expected outcome* is a system that provides a way of "looking-up" an image in a collection of images such that similar images are retrieved.

BRISC was previously implemented by Lam et al [1] and the code is also available as open source [4]. Lam at al. showed that the global co-occurrence texture model performs worse (retrieval precision of 29%) than the Gabor filters and MRF texture models (retrieval precision of 88%). In this paper, we implement the co-occurrence texture model at the local level (within a small neighborhood for each pixel of a lung nodule image instead of the entire nodule image); furthermore, we investigate the effect of using all three texture models with respect to their similarity retrieval power. Given that each one of the texture models captures different properties of the texture, in this paper we show that the combination of the three texture models produces better results than the individual texture models.

There are several CBIR projects in the medical field currently underway. Kinoshita et al. [5] discussed a CBIR system for mammograms. They utilized a large amount of visual features including shape, texture, and granulometric. Kinoshita et al. combined many different features using principal component analysis to improve the system. They report a precision of around 85%. However, based upon the ground truth that they used (BI-RADS categorization) a random selection of images would return a precision of 62.5%. Wei. et al. [6] discussed image analysis and image retrieval using gray level co-occurrence matrices for the mammography domain. They used a distance of 5 on the co-occurrence matrices which resulted in a maximum precision of 51% and recall of 19% (average = 49% and 18% respectively). They grouped each region of interest into 6 categories and if the returned image was placed in the right category it was relevant.

Furthermore, Muramatsu et al. [7] presented strong evidence that CBIR systems could help to improve the accuracy of identifying benign or malignant clustered micro-calcifications on mammograms and that breast radiologists are able to provide a reasonable ground truth. Muramatsu [8] discusses the development of an Artificial Neural Network that shows promising ability to retrieve images similar to those of an unknown lesion. Tourassi et al. [9] evaluated similarity measures utilized in a scheme for content-based retrieval and detection of masses in screening mammograms. They found that the measures interestingly fell into two categories: "one category is better suited to the retrieval of semantically similar cases while the second is more effective with knowledge-based decisions regarding the presence of a true mass in the query location". Zheng et al. [10] has also done interesting work developing an automated

interactive computer-aided diagnosis scheme that performs just as well as the subjective rating method.

Computed Tomography (CT) scanning has been found to increase the detection rate of pulmonary nodules [11]. Much work has been done to develop computer assisted diagnosis and detection (CAD) systems for pulmonary nodules in CT. For a detailed description of CAD systems, we suggest the review by Muller et al [12].

One of the largest CBIR projects currently underway using lung CT images is the ASSERT project [13], which is being developed at Purdue University and was first published in 1999. It proposed a "physician-in-the-loop" system where radiologists highlight a region and the system would return similar images. The system used a variety of image features including co-occurrence statistics, shape descriptors, Fourier transforms and global gray level statistics. The system also utilized physician-provided ratings of features such as homogeneity, calcification and artery size. The best precision reported by the system was 76.3%.

There are many difficulties involved with content-based retrieval of medical images, including the difficulty of automatic segmentation, the large variability of feature selection, and the lack of standardized toolkits and evaluation methods [14][15][16]. There have been efforts recently to solve these problems including the Lung Image Database Consortium (LIDC) collection which was specifically developed to support evaluation and comparison of chest CAD systems [17]. It can be used similarly to develop, evaluate, and compare CBIR systems.

2. Methods

2.1. LIDC Data

The data in our study was obtained from the Lung Image Database Consortium (LIDC) database [17]. The database contains 149 unique pulmonary nodules that have been segmented and annotated by up to four different radiologists amounting to a total of 2020 images. These images were taken from a total of 90 Computer Tomography studies of the chest, each containing between 100 and 400 Digital Imaging and Communication (DICOM) images. Four radiologists marked the contour of the nodules and assigned nine semantic terms/characteristics to each nodule: calcification, internal structure, lobulation, malignancy (as interpreted by the radiologists based on imaging findings), sphericity, spiculation, subtlety, texture, and margin. Calcification and internal structural are nominal while the other seven annotations are ordinal. Calcification contains six different categories, internal structure contains four different categories, and the other seven annotations each are rated on a scale from one to five.

Figure 1

2.2. Texture Models

We extract low-level image features that encode the *texture* of the lung nodules while satisfying the main requirements for feature extraction: a) *completeness/expressiveness* (features should be a rich enough representation of the image contents to reproduce the essential information); b) *compactness* (the storage of the features should be compact to allow efficient access) and c) *tractability* (the distance between features should be efficient to compute).

The three texture models satisfying these properties and used for this research are: local and global co-occurrence matrices [18], Gabor filters [21], and MRF [18][22]. The co-occurrence texture models generated 11 texture descriptors which represented the statistical properties of the nodules' texture. Separate co-occurrence matrices were calculated for each direction (0, 45, 90, and 135 degrees) and displacement (1, 2, 3, and 4 pixels). In global co-occurrence the texture descriptors were extracted per nodule image while in local co-occurrence the texture descriptors were extracted for each relevant pixel in the nodule image. The intensities of the nodule image were binned for global and local co-occurrence to allow statistical relevance to appear in the co-occurrence matrices; otherwise the information gained is usually noise. For pixel level feature extraction local co-occurrence extracts a set of pixels that surrounds each pixel

and performs co-occurrence on that subset of the original image. The size of that subset is determined by the variable 'window size'. Also, to compute similarity, local co-occurrence is placed into a histogram. The variables for window size, number of bins used for the histogram, and the number of bins for the intensities were varied in an attempt to find the parameters that achieved the best results in local co-occurrence.

In contrast to the statistical based co-occurrence methods, Gabor filtering is a transform-based method of extracting texture information in the form of a response image. A Gabor filter is a sinusoid function modulated by a Gaussian and produces 12 filter images tuned to four orientations ($0, \pi/2, \pi/4, 3\pi/4$) and three frequencies (.3, .4, and .5) encoding the texture properties in the frequency space [1]. Markov Random Fields capture the local contextual information of an image. The value utilized for each pixel in MRF is dependant on its neighbors. The MRF model produced five images corresponding to four orientations (0, 45, 90, and 135 degrees) between pairs of neighboring pixels plus variance [1].

2.3. Similarity Measures

There are many similarity measures proposed for general CBIR systems and the choice of a similarity measure is dependent on both the feature space representation and its ability to capture the visual human perception of similarity. We investigate similarity measures from three categories of similarity measures: 1) *Heuristic distance metrics (*Minkowski distance), 2) *Non-parametric test statistics (*Chi-square statistics), and 3) *Information Theory Divergences (*Jeffrey-Divergence).

Global co-occurrence results in a one dimensional feature vector for each image, therefore Euclidean, Manhattan, and Chebyshev were used to measure the texture-based similarity of the nodules. Local co-occurrence, MRF, and Gabor features are local, so they result in a two dimensional feature response for each image. Thus the Chi-Square and Jeffrey-Divergence measures were used to measure the texture-based similarity between these models. More information about these similarity measures can be found in the papers by Lam et al [1] and Puzicha et al. [19], [20].

2.4. Retrieval Performance Evaluation

We evaluate the retrieval system using precision as the performance metric and the expert relevance feedback. The precision is calculated for all images from the database; the overall precision of the system is then calculated as the average of all precision values obtained when each image becomes the query image. The general formula for calculating precision is:

$$\Pr ecision = \frac{\#\_of\_relevant\_images}{\#\_of\_retrieved\_images} \quad (1)$$

We calculate the precision considering a "relevant image" in response two a query in two different ways: 1) objective evaluation: a "relevant image" is an image belonging to the same nodule but appearing in another slice or even in the same slice but outlined by another radiologist (see figure 2), and 2) subjective evaluation: a "relevant image" is an image that appears in the list of the most similar images with the query image based on the radiologists' annotations. Furthermore, we say that the computer retrieval results match perfectly the human perception when the order in the list L_c of the most similar images based on low-level features corresponds exactly to the order in the list L_a of the most similar images based on the radiologists' annotations and the Jaccard coefficient distance. More formally, if we assign scores $S_{L_{c,i}}$ and $S_{L_{a,i}}$ to the retrieved images from n to 1 (n to the first item in the list, n - 1 to the second, etc . . .) for each one of the two lists L_c and L_a respectively, a perfect correspondence between the texture-based results and the human annotations will happen when the *Relevance Index* (RI) defined by the left term of equation 2 is equal to the term from the right hand side.

$$\sum_{i=1}^{n} S_{L_{c,i}} * S_{L_{a,i}} = \sum_{i=0}^{n-1} (n-i)^2 \qquad (2)$$

where n is the number of retrieved images. For example, for n=10 items retrieved, a perfect match will result in a value equal to 385 in the above equation. In the worst case scenario, the value will be equal to 0 since no items from one list will be found in the other list.

Figure 2

3. Experimental Results

3.1: Texture Retrieval Results

We implemented all possible combinations of the proposed texture features and similarity measures and evaluated the precision results using both the objective (Figure 3) and subjective (Figure 4) criteria. For each criterion, we found that a combination of local co-occurrence, Gabor, and MRF performs the best, followed by either the Gabor filtering (objective criterion) or MRF (subjective criterion), and the worst being the global co-occurrence texture model. Therefore, we have reaffirmed the conclusions that we had made based on the first evaluation method (objective) in our previous work [1].

Figure 3

Figure 4

3.2: Similarity Measure Comparison, Second Evaluation Method

Table 1 shows the relationship between the similarity measures utilized when calculating Global Co-occurrence, Local Co-occurrence, Gabor filters, and MRF evaluated using the second method. The results utilize n = 10. When radiologists' agreement was taken into account the results stayed constant, as they did for different values of n. For MRF, however, Chi-Square performs much better while Jeffrey Divergence performed very poorly. Therefore, when combined with MRF, Chi-Square is much more related to radiologists' perception than when using Jeffrey-Divergence.

Table1

3.3: Second Evaluation Method Using Radiologist Agreement

Figure 5 describes the comparison between the texture models using the second evaluation method while varying the number of radiologists that agree on the texture of a lung nodule. From this we can see that when radiologist agreement isn't taken into account, the differences between the similarity measures is very similar to the results we received using the first evaluation method. When four radiologists agree Local Co-occurrence performs the best with 70.89% precision. Gabor, which performed the best using evaluation method 1, performed the worst with 67.17% precision. Therefore Local Co-occurrence is more related to the radiologists' perceptions while Gabor is less so. Results may not be statistically significant because of the small number of nodules that are included in the set when 4 radiologists agree.

Figure 5

4. Conclusion

Our second evaluation method shows similar results with our first evaluation method, reinforcing the results we received from our first evaluation method. In terms of similarity, there is little difference between the Euclidean, Manhattan, and Chebychev although Euclidean repeatedly performs slightly better followed closely by Manhattan then by Chebychev. Jeffrey-Divergence and Chi-Square perform exactly the same for Local Co-occurrence yet Chi-Square performs slightly better for Gabor consistently. For MRF, however, Chi-Square is a much better similarity measure, especially when considering radiologists' perceptions. Gabor performs poorly when radiologists' agree on the annotation of texture while MRF and Local Co-occurrence perform the best. Overall, considering both evaluation methods, MRF was the best single texture retrieval method while the combination of the texture models performed the best (see Table 2).

Table 2

References

[1] M. Lam, T. Disney, M. Pham, D. Raicu, J. Furst, R. Susomboon, "Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images", SPIE Medical Imaging Conference, Vol. 6516, 65160N, March, 2007.

[2] *Cancer Facts and Figures*, American Cancer Society, 2006.

[3] "What are the Key Statistics about Lung Cancer?" American Cancer Society, 2007. http://www.cancer.org/docroot/CRI/content/CRI_2_4_1x_What_Are_the_Key_Statistics_About_Lung_Cancer_15.asp?sitearea=.

[4] Lam M, Disney *T*, Raicu DS, Furst JD, Channin DS. "BRISC - An Open Source Pulmonary Nodule Image Retrieval Framework". Journal of Digital Imaging, 2007, doi 10.1007/s10278-007-9059-y.

[5] S. K. Kinoshita, P.M. de Azevedo-Marques, R.R. Pereira Júnior, J.A.H. Rodrigues, and R.M. Rangayyan, "Content-based retrieval of mammograms using visual features related to breast density patterns". Vol 20; number 2, pages 172-190, February 2007.

[6] C. Wei, C. LI, R. Wilson. "A General Framework for Content-Based Medical Image Retrieval with its application to Mammograms", Proc. SPIE Medical Imaging 2005, April 2005; 134-143.

[7] Muramatsu C. Li Q. Schmidt R. Suzuki K. Shiraishi J. Newstead G. Doi K. "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results". *Medical Physics. 33(9):3460-8, 2006 Sep.*

[8] Muramatsu C. Li Q. Suzuki K. Schmidt RA. Shiraishi J. Newstead GM. Doi K. "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results". Medical Physics. 32(7):2295-304, 2005 Jul.

[9] Tourassi GD. Harrawood B. Singh S. Lo JY. Floyd CE. "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms". Medical Physics. 34(1):140-50, 2007 Jan.

[10] Zheng B. Mello-Thoms C. Wang XH. Abrams GS. Sumkin JH. Chough DM. Ganott MA. Lu A. Gur D. "Interactive computer-aided diagnosis of breast masses: computerized selection of visually similar image sets from a reference library". Academic Radiology, Volume 14, Issue 8, Pages 917-927, 2007 Aug.

[11] Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP: "Early lung cancer action project: overall design and findings from baseline screening". The Lancet 354:99-105, 1999, July.

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

[13] C.-R. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," Computer Vision and Image Understanding 75, pp. 111=132, July/August 1999.

[14] Smeulders AW, Worring M, Santini S, Gupta A, Jain R: "Content-based image retrieval at the end of the early years". IEEE Trans Pattern Anal Mach Intelligence 22(12):1349-1380, 2000, (December).

[15] Antani S, Long LR, Thoma GR: "Content-based image retrieval for large biomedical image archives". In Proceedings of 11th World Congress on Medical Informatics (MEDINFO) 2004 (September).

[16] Muller H, Michoux N, Bandon D, Geissbuhler A: "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions". International Journal of Medical Informatics 73(1):1-23, 2004 (February).

[17] LIDC Lung Nodule Image Database. National Cancer Imaging Archive (https://imaging.nci.nih.gov/ncia/). Accessed July 2007.

[18] A. Materka and M.Strzelecki, "Texture analysis methods – a review," tech. rep., Technical University of Lodz, Institute of Electronics, 1998. COST B11 report.

[19] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), 1997.

[20] J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in ICCV (2), pp. 1165-1172, 1999.

[21] T. Andrysiak and M. Choras, "Image retrieval based on hierarchical gabor filters," International Journal Applied Computer Science 15 (4), pp. 471-480, 2005.

[22] C. Chen, L. Pau, and P. W. (eds.), The Handbook of Pattern Recognition and Computer Vision (2nd Edition), World Scientific Publishing Company, 1998.

[23] Tagare HD, Jaffe CC, Duncan J. Medical image databases: a content-based retrieval approach. JAMIA, 1997(4), pp. 184-198.