Mammographic Mass Characterization Using Qualitative Diagnostic Data and Image Texture Analysis

M. E. MAVROFORAKIS[1], H. V. GEORGIOU[1], D. CAVOURAS[2], N. DIMITROPOULOS[3], S. THEODORIDIS[1]

[1] Univ. of Athens, Informatics Dept., TYPA buildings, Univ. Campus, 15771, Athens, Greece
[2] Medical Imaging Technologies Dept., TEI-Athens, 12210, Athens, Greece
[3] Medical Imaging Dept., EUROMEDICA Medical Center, 2 Mesogeion ave, Athens, Greece

Contact: M. Mavroforakis (Mr), 43 Knossou str, 16561, Athens, Greece – mailto:memav@nbg.gr
H. Georgiou (Mr), 11 Vas.Dipla str, 11745, Athens, Greece – mailto:xgeorgio@middle-earth.gr

Abstract – A combined approach of image texture analysis and qualitative diagnostic data evaluation is presented in this study. The application of statistical approaches in detailed clinical findings and texture-related features have established the significance of image texture evaluation during the diagnostic assertion process. Multiple textural feature functions in various configurations were applied to a large database of digitized mammograms, in order to establish their discriminating value and statistical correlation with qualitative texture descriptions of breast mass tissue. A wide range of linear and non-linear classification models were applied, including linear discriminant analysis, least-squares minimum distance, K-nearest-neighbors, RBF and MLP neural networks. For texture-only classification, optimal accuracy rates reached 81.5%, while the introduction of patient's age increased the overall accuracy rates up to 85.4%.

I. INTRODUCTION

Mammographic image analysis and understanding is a complex cognitive task, which includes various aspects of medical expertise and clinical findings. The visual task of clinical evaluation and diagnosis, based on mammographic image screening, consists of a number of different factors in multiple scales and levels of decomposition. Fine-scale organization of the image is of most importance in the detection of malignancy, as it is expected to reflect the structural status of biological tissues [1].

In reference to clinical status estimation conducted by a physician, patient's age and history have been proven issues of outmost importance for the conduction of a successful clinical status evaluation [2]-[3]. The presence of suspicious areas in the form of tumors is often examined by using textural content of the mammographic image. A property of great importance is the presence and morphology of microcalcifications, as well as the morphology of the tumor itself [4].

This study was focused on three main areas of interest. First, a complete mammogram dataset, containing detailed human expert-provided qualitative information regarding verified clinical cases, is studied and the statistical significance regarding diagnostic discrimination information was estimated for each one of the components of the dataset, both separately and in combination. Second, same type of analysis was conducted over the whole range of different textural feature functions configurations, so that directly comparable results could be evaluated. Third, discrimination power and classification results from the qualitative dataset were compared with the textural features datasets accordingly. Finally, a hybrid configuration with combined data components from both the qualitative and textural features datasets was constructed, in order to evaluate the actual performance and practical usefulness of such mixed configurations, as well as the relative contribution of each data component grouping to the resulting efficiency of datasets and classifiers.

II. DIGITIZED MAMMOGRAPHIC DATA

For this study, a subset of 163 mammograms where selected for digitization by an expert. The selection was made on the basis of unbiased statistical distribution and the completeness of the dataset. All cases were positively verified clinically by biopsy and further diagnostic tests. For each mammogram, a complete list of qualitative information was provided by an expert physician, containing details about the age of the patient, the presence of tumor(s), microcalcifications, tumor density, percentage of fat, tumor boundary vagueness, tumor homogeneity, tumor morphological shape type and clinical diagnosis. The various qualitative details were included as explicit information related to various types of malignant mammogram abnormalities including architectural distortion, microcalcification clusters, nodular or stellate masses and lymphoadenomas.

The mammograms where digitized at a resolution of 63 μm (400 dpi) at 8-bit gray level and some post-processing was applied in order to further enhance the sharpness of the images. From the initial set of 163 mammograms, a total of 33 were not used in this study either because of the absence of distinct tumors, or due to ambiguity on the exact shape type or other qualitative property characterization. The final set of 130 mammograms was used in all cases with no reduction in spatial resolution or gray level depth.

Using the original mammographic database of the 130 images, mass boundaries were manually described by the expert and used later for the definition of mass inclusion masks and boundary zones for textural features extraction at these areas of interest.

III. FEATURES AND DATASETS

In order to study the relative effect in the quality of texture information extracted from the digitized images, a large set of textural feature functions were applied in multiple configurations.

A. Image Processing

For each sampled sub-region, co-occurrence and run-length matrices were computed for three separate neighboring pixel configurations according to a distance factor $d \in \{1,2,3\}$, creating three corresponding interleaving modes at pixel-level, essentially affecting the base on which the textural feature functions were applied. Multiple distance factors were used for the evaluation of the effect of sub-sampling at pixel level on the quality and consistency of the extracted textural features.

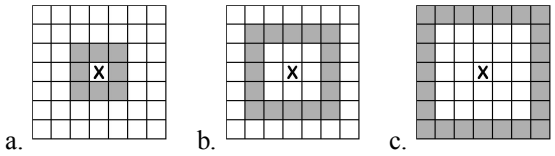

Figure-1. Sub-region sampling schemes over various pixel neighboring distances: (a) d=1, (b) d=2, (c) d=3.

For each different pixel interleaving mode, all available angular directions were used for the calculation of co-occurrence and run-length matrices. The average and range values of each feature over all primary directions were stored as possible discriminating texture information.

B. Textural Features Extraction

From each sampled sub-region, texture was examined by extracting first order statistics, second order statistics and graylevel runs. All subsequent analysis preserved the graylevel depth to the original 8-bit depth and all runs were examined up to their full length.

First order statistics of the grayscale distribution for each image sub-region matrix were examined through six commonly used metrics proposed by Haralick et al [5]-[6]. Namely, min value, max value, average value, standard deviation, skewness and kurtosis were used as descriptive measurements of the overall grayscale histogram of the mass.

Second order statistics of the grayscale distribution, derived from spatial distribution grayscale matrices (SDGM), were examined through fourteen commonly used metrics also proposed by Haralick et al [5]-[6]. Namely, angular second moment, contrast, correlation, variance, inverse difference moment, sum of averages, sum of variances, sum of entropy, entropy, difference of variance and difference of entropy were used as standard co-occurrence matrix descriptors of texture, while information measure of correlation 1, information measure of correlation 2 and maximal correlation coefficient were used as co-occurrence matrix measures of textural information content.

Graylevel runs, derived from run-length matrices (RLM), were examined through five commonly used run-length metrics proposed by Galloway [7]. Namely, short runs emphasis, long runs emphasis, run-length non-uniformity and run percentage were used as descriptive measurements of each of the run-length matrix calculated over the sampled image sub-regions.

C. Qualitative Dataset Construction

The qualitative dataset consisted of detailed data estimated and quantified by an experienced physician. After extensive research and suggestions made by experienced physicians, a subset of nine most dominant and suitable properties were selected for the construction of the final dataset. Furthermore, experts also defined the appropriate method, scale and resolution for translating each feature into quantitative data. Resulting forms with filled data were verified and any case containing ambiguities was removed from the database prior to any further processing, both for textural features and qualitative datasets construction.

IV. CLASSIFICATION AND TESTING

Features extracted from each separate dataset were rated by applying multivariate statistical significance analysis (MANOVA) and feature subsets combinations were optimized by exhaustive search during the training process.

Several classifier architectures were applied during the classification phase. A LDA model was used in the form of linear classifier. A least-squares minimum distance classifier (LSMD) was employed, using Mahalanobis distance measure and least-squares dataset pre-processing on the input. A K-nearest neighbor model was also used, including estimation of an optimal value K for the size of the neighborhood set.

Two different types of neural network architectures were employed: a RBF network with Gaussian activation functions and linear output functions, and a MLP network with hyperbolic tangent internal activations and softmax output functions, both implemented with topology adapted in each configuration and dataset. All topologies included one hidden layer with optimized size.

All configurations used leave-one-out method for dataset manipulation during training and testing phases, combined with optimal feature subset selection for the linear classifiers, or the selected (optimal) feature sets for the neural networks.

V. RESULTS AND DISCUSSION

Classification tests were conducted upon four primary factors. First, to evaluate the discriminative content of the qualitative data. Second, to study the effect of different textural feature functions configurations in relation to sampling image sub-regions box sizes, pixel neighboring distance values and sampling region spatial positioning (using complete mass region or mass borderline only). Third, to evaluate the discriminative content captured by the various textural feature functions. Finally, a hybrid

configuration of mixed qualitative and textural features dataset was constructed by using previously optimized options, in order to estimate the peak performance of the system and evaluate its usefulness in practical diagnostic applications as a computer-aided mammographic double-screening process [8].

A. Qualitative Dataset Classifications

Preliminary studies on the original mammographic database have confirmed the strong significance of morphological shape characterization with the clinical diagnostic result. Specifically, it was statistically verified that round and lobulated cases exhibit only 5% to 16% malignancy, while nodular and stellate cases exhibit 90% to 97% malignancy. This result shows that utilizing the shape characterization alone as discrimination measure between benignancy and malignancy can establish a success rate over 90%.

	Round	Lobulated	Nodular	Stellate
Benign	25	18	1	1
	83%	95%	2%	3%
Malignant	5	1	42	37
	17%	5%	98%	97%

Table-1. Training patterns true distribution against morphological shape types and clinical diagnosis.

Qualitative Feature	LSMD succ% - diag.
Mass Shape Type	93.1%
Boundary Vagueness	86.1%
Fat Percentage	74.6%
Mass Density	73.1%
Mass Homogeneity	73.1%
Patient's Age	68.5%
Microcalcifications?	60.8%

Table-2. Least-squares minimum distance discrimination efficiency of every individual qualitative feature.

Individual features classifications proved the strong statistical correlation between clinical diagnosis and most of the qualitative features. All features except patient's age and mass shape type refer directly to textural properties of the mass. With regard to clinical diagnosis, the morphological mass shape type was the most correlated feature in the set. Using all other features except mass shape type and patient's age, the optimal LSMD classifier selected fat percentage, boundary vagueness and mass homogeneity as the best feature set, yielding 86.9% accuracy. Including patient's age in that feature set selection was also optimal, yielding 89.2% accuracy. For datasets with no reference to mass shape type information, accuracy rates ranged from 87.7% up to 91.5%, while for datasets including mass shape type information, accuracy rates ranged from 91.5% up to 93.1%.

B. Textural Features Evaluation

The four texture datasets (box sizes of 20, 50, 40 at borderline only and complete mass) were used in a total of eight dataset configurations. Four initial configurations included primary tests for feature functions parameter optimization using the 20 and 50 box size cases, while the other two datasets exploited the acquired optimized parameterization for feature functions to derive four more (optimized) texture datasets used for the conclusive classification configurations. Multivariate statistical significance selection (MANOVA), along with subset combinations optimization, was employed for final feature set construction.

The first configuration included sampling box size of 50 and pixel neighboring of all three available distances (d=1,2,3). Evaluation was conducted by using LDA and least-squares minimum distance classifiers, with the best accuracy rate 72.6% was achieved by a LSMD classifier. The second configuration used the same box size and pixel neighboring distance equal to one (d=1) only. In this case, the best accuracy rate achieved was 67.5% by LDA. The third configuration included sampling box size of 20 and pixel neighboring of all three available distances (d=1,2,3). The best accuracy rate achieved was 58.3% by LSMD. Finally, the forth configuration used the same box size and pixel neighboring distance equal to one (d=1) only. The best accuracy achieved was 65.8% by LSMD.

Based on the results of these four initial configurations, it was determined that all the available pixel neighboring distances exhibited important discrimination value, while larger sampling boxes increased the quality of feature content. Thus, the complete mass region was used instead of the previous box size of 50 and the borderline sampling box size was increased to 40 instead of the previous box size of 20.

The two texture datasets were organized into three configurations, namely two for processing each one individually and one for processing the merging of the two datasets into one feature-based set union. Finally, a forth configuration was created by mixing the merged texture dataset with the single feature of patient age from the qualitative dataset. All configurations included training patterns grouped according to their source image. In this way, every training pattern constituted one complete texture descriptor for each identified mass.

Using the entire range of the available classifiers in all cases against clinical diagnosis, textural features based on complete mass statistics yielded the best texture-only results, with 81.5% accuracy rate achieved by LSMD classifier, along with 80.0% achieved by MLP classifier. Configurations based on borderline only texture evaluation resulted in lower accuracy rates between 69.2% and 77.1% over all classifiers. Combining the two texture datasets increased the overall performance ranging from 75.3% to 79.2%, but still lower in comparison with the complete-mass dataset configuration.

Finally, the hybrid texture dataset containing patient's age significantly increased the overall performance in all classifiers. Specifically, accuracy rates ranged from 76.9% up to 85.4% with the best two peak values achieved by LSMD (83.8%) and MLP (85.4%) classifiers.

Provided the statistical significant analyses' results regarding optimal textural features ordering in various configurations, along with the optimal feature set combinations estimated by LDA and LSMD classifiers, a list of most valuable textural features was constructed. Combining the results and optimizations of multiple configurations, Table-3 summarizes these features.

Feature	Description
1	Patient's age
105	SDGM difference of entropy, range, d=3
111	SDGM max. correlation coef., range, d=3
114	RLM long runs emphasis, mean, d=3
76	RLM long runs emphasis, mean, d=2
100	SDGM entropy, mean, d=3
99	SDGM sum of entropy, range, d=3
95	SDGM sum of average, range, d=3
118	RLM non-uniformity, mean, d=3
117	RLM graylevel non-unif., range, d=3

Table-3. Multivariate (MANOVA) textural features selection for combined texture datasets plus patient's age. Underlines indicate features selected in optimal sets.

It is evident that these optimally selected features are highly specialized. Run-length features, especially those related to non-uniformity, short runs and long runs emphasis, seem to have a significant role in the correct characterization of fine-scaled structures against benignancy or malignancy of breast tumor tissue. From the co-occurrence matrices statistical features, variance and its statistical properties seem to be equally important, along with statistics on average and information content measures like entropy. Notably, at least half of the textural feature measures refer to the range of variance, rather than just their mean value averaged over all available angular directions. Patient's age is, again, proved to be of most importance in relation to clinical diagnosis.

Finally, with the exception of two individual feature choices, all other features refer to texture extraction at pixel neighboring distance of 3. Features of great statistical and discriminating significance produced by similar configurations imply that texture extraction at larger scales may be more appropriate for breast mass tissue characterization via image texture. Consequently, further studies have to be conducted towards the formulation and application of scalable image texture descriptors like 2-D wavelet decomposition structures [9].

C. Classifier Architectures Evaluation

Conclusive classification results proved the value of non-linear architectures versus most of the linear models. Between linear classifiers, least-squares pre-processing of input data (LSMD) resulted in significant overall success rate and it was marginally the best classifier choice over texture-only datasets, with peak accuracy at 81.5%, over 80.0% accuracy achieved by MLP classifiers for the same complete-mass texture dataset. However, in the case of combined texture datasets with the patient's age, MLP classifiers achieved accuracy rates up to 85.4% against LSMD classifiers' accuracy rates of 83.8%.

For neural networks, comparison between RBF and MLP architectures proved that RBF networks resulted in somewhat lower overall accuracy. Differences in success rates ranged between 0% and +5%, marginally favoring the choice of MLP architectures over all optimized topology sizes.

VI. CONCLUSION

Texture analysis is one of the most valuable and promising areas in breast tissue analysis and characterization. Textural features have been widely used for a variety of computer vision applications, including diagnostic systems for ultrasonic images and the segmentation of mammographic images. Extensive study of these features in various scales and configurations is necessary in order to select a viable set of texture descriptors, specialized for the specific task of breast mass tissue characterization. The application of robust classifier models proved to be of outmost importance as well. Multi-layered perceptrons outperformed all other linear and neural architectures, while least-squares minimum distance classifiers performed equally well. In both cases, the overall performance and accuracy rates were over 81% for texture-only data, or higher than 85% when combining them with patient's age. As the texture addresses only one property of mammographic images, the prospect of using texture in conjunction with other methodologies, like structural or morphological mass analysis, into a combined diagnostic tool are very promising.

REFERENCES

[1] L.Bocchi, G.Coppini, et al., Tissue characterization from X-ray images, Med.Eng.Phys., Vol.19, No.4, 1997, 336-342.

[2] C.D.Orsi, D.J.Getty, J.A.Swets, et al., Reading and decision aids for improved accuracy and standardization of mammographic diagnosis, Radiology, Vol.184, 1992, 619-622.

[3] Y.Wu, M.L.Giger, K.Doi, et al., Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer, Radiology, Vol.187, 1993, 81-87.

[4] I.Christoyianni, E.Dermatas, G.Kokkinakis, Fast detection of masses in computer-aided mammography, IEEE Sig.Proc.Mag., Jan.2000, 54-64.

[5] R.M.Haralick, K.Shanmugam, I.Dinstein, Textural features for image classification, IEEE Trans.Sys.Man.Cyb., Vol.SMC-3, No.3, Nov.1973, 610-621.

[6] R.M.Haralick, Statistical and structural approaches to texture, Proc.IEEE, Vol.67, No5, May 1979, 786-804.

[7] M.Galloway, Texture analysis using gray level run lengths, Comp.Graph.Im.Proc., 4, 1975, 172-179.

[8] E.Thurfjell, K.A.Lernevall, A.Taube, Benefit of independent double reading in a population-based mammography screening program, Radiology, Vol.191, 1994, 241-244.

[9] M.Vetterli, C.Herley, Wavelets and filter banks: theory and design, IEEE Trans.Sig.Proc., Vol.40, No.9, 1992, 2207-2232.

[10] Huai Li, Yue Wang, K.J.Ray Liu, et al., Computerized radiographic mass detection – Part I: Lesion site selection by morphological enhancement and contextual segmentation, IEEE Trans.Med.Im., Vol.20, No.4, Apr.2001, 289-301.

[11] L.V.Ackerman, A.N.Mucciardi, et al., Classification of benign and malignant breast tumors on the basis of 36 radiographic properties, Cancer, Vol.31, 1973, 342-352.

[12] R.Reed, Prunning algorithms – A survey, IEEE Trans. Neural Networks, Vol.4, No.5, Sept.1998, 740-747.

[13] L.V.Ackerman, A.N.Mucciardi, et al., Classification of benign and malignant breast tumors on the basis of 36 radiographic properties, Cancer, Vol.31, 1973, 342-352.

[14] S.Theodoridis, K.Koutroumbas, Pattern recognition (San Diego, CA: Academic Press, 1999).