Texture feature standardization in digital mammography for improving generalizability across devices

Yan Wang*, Brad M. Keller, Yuanjie Zheng, Raymond J. Acciavatti, James C. Gee, Andrew D.A. Maidment, Despina Kontos
Department of Radiology, Perelman School of Medicine at the University of Pennsylvania.
3600 Market St. Suite 370, Philadelphia, PA, USA, 19104

ABSTRACT

Growing evidence suggests a relationship between mammographic texture and breast cancer risk. For studies performing texture analysis on digital mammography (DM) images from various DM systems, it is important to evaluate if different systems could introduce inherent differences in the images analyzed and how to construct a methodological framework to identify and standardize such effects, if these differences exist. In this study, we compared two DM systems, the GE Senographe 2000D and DS using a validated physical breast phantom (Rachel, Gammex). The GE 2000D and DS systems use the same detector, but a different automated exposure control (AEC) system, resulting in differences in dose performance. On each system, images of the phantom are acquired five times in the Cranio-Caudal (CC) view with the same clinically optimized phototimer setting. Three classes of texture features, namely grey-level histogram, co-occurrence, and run-length texture features (a total of 26 features), are generated within the breast region from the raw DM images and compared between the two imaging systems. To alleviate system effects, a range of standardization steps are applied to the feature extraction process: z-score normalization is performed as the initial step to standardize image intensities, and the parameters in generating co-occurrence features are varied to decrease system differences introduced by detector blurring effects. To identify texture features robust to detectors (i.e. the ones minimally affected only by electronic noise), the distribution of each texture feature is compared between the two systems using the Kolmogorov-Smirnov (K-S) test at 0.05 significance, where features with $p>0.05$ are deemed robust to inherent system differences. Our approach could provide a basis for texture feature standardization across different DM imaging systems and provide a systematic methodology for selecting generalizable texture descriptors in breast cancer risk assessment.

Keywords: Digital mammography, x-ray detectors, MTF, phantom, parenchymal texture, feature standardization

1. INTRODUCTION

Breast cancer is the most common cancer in women worldwide, and about 1 in 8 U.S women will develop invasive breast cancer over the course of her lifetime. Early screening and proper treatment after diagnosis for individual women are both important aspects of current breast cancer research, and digital mammography (DM) is the main screening tool for cancer detection[1]. In western countries, 89% of women diagnosed with breast cancer are still alive 5 years after their diagnosis, which is due both to early detection and treatment.

Currently, the Gail Model[2] is one of the commonly used models to calculate a woman's risk of developing breast cancer within the next five years and within her lifetime. This model calculates the population risk for women with the similar risk factors, however with limited capacity at the individual level. Currently, there are increasing efforts to improve individualized breast cancer risk estimation. Mammographic density[3], estimated as the percent of dense tissue area within the breast, has been shown to be the strongest risk factor for breast cancer after age. Studies[4-7] also support a relationship between mammographic texture and breast cancer risk, as mammographic texture features can quantify the local distribution of the parenchymal pattern, providing complementary information for breast cancer risk assessment.

In studies of risk assessment based on mammographic texture features, it can be commonly the case that studies use DM images that are acquired from different DM systems. In such studies, it should be worthwhile to treat the imaging system as an additional parameter introducing potential bias in the analysis, as different imaging systems may possibly introduce

*wangyan1@sas.upenn.edu; phone 1 216 605-3705;

Medical Imaging 2013: Computer-Aided Diagnosis, edited by Carol L. Novak, Stephen Aylward, Proc. of SPIE
Vol. 8670, 867026 · © 2013 SPIE · CCC code: 1605-7422/13/$18 · doi: 10.1117/12.2008149

different inherent effects to the generated DM images. As a first step towards understanding these effects, in our study, the image intensity and extracted texture features from a breast phantom are compared between two DM systems for the raw (*i.e.,* "for processing") images. Using the modulation transfer function (MTF) property of the x-ray detectors, which is closely related to the spatial resolution of the acquired DM image, the effects of intensity standardization and related parameters in generating texture features, especially for co-occurrence features are considered and analyzed. A texture feature standardization scheme based on our previous preliminary studies[13] and statistical methodology is proposed to identify texture features that are robust to different DM systems. Our study addresses the importance of understanding the physics of imaging before extracting texture features. The observations from our experiments and the proposed general standardization scheme could be helpful for many studies or applications using texture analysis in digital mammography, including breast cancer risk assessment, breast tissue classification, and computed aided diagnosis.

2. METHODS

2.1 Material

The Gammex 169 "Rachel" anthropomorphic breast phantom was used in our experiments[10]. Image acquisition was performed on both GE Senographe 2000D and DS Full Field Digital Mammography (FFDM) systems, with 0.1mm/pixel resolution, 14 bit gray-levels. The GE 2000D/DS systems were introduced in years 2000/2004 respectively, with the same size flat panel detector, while the DS system has a smaller tube with stereotactic capabilities. The two systems both have indirect-conversion detectors using cesium iodide with TFT, which compared to direct conversion process, blurring effects can be introduced by the phosphor and can cause loss of spatial resolution. A study comparing the MTF of the two detector systems has shown that the spatial resolution for 50% MTF for the 2000D and DS systems is 3.19/mm and 3.29/mm respectively, and the corresponding MTF at 5/mm is 0.27 and 0.28. The GE 2000D and DS also have a slightly different automated exposure control (AEC) system, resulting in differences in dose performance. Using the fixed Rachel phantom, the clinically optimized phototimer setting of (kVp, mAs) was chosen at 29 kVp, 71 mAs for 2000D and 29 kVp, 90 mAs for the DS system. The phantom image acquisition process was repeated 5 times for both machines. The mean of these 5 images was used for subsequent analysis to decrease the effect of imaging noise, and these 5 images were also used to assess the effect of imaging noise on the texture features within each DM system.

2.2 Image preprocessing

The breast area is segmented by 1) manually removing the box boundary, and 2) a synchronic thresholding scheme is used to generate the breast region mask. The details of these preprocessing steps have been previously explained[14].

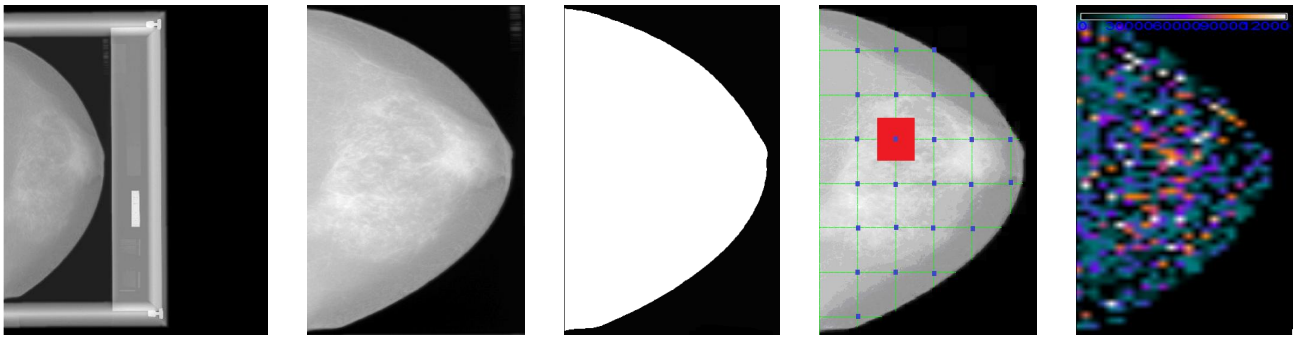

Figure 1. For visual convenience, processed DM image acquired on the 2000D system is shown here; From left to right: 1) original phantom image; 2) removal of bounding case; 3) mask for breast region; 4) the lattice for texture image generation; 5) feature image (co-occurrence feature 'cluster shade').

2.3 Feature extraction

Three classes of texture features[8,9] are extracted (a total of 26 features) using an automated breast image analysis software pipeline[11], including 1) grey-level histogram features, 2) co-occurrence features, and 3) run-length features.

Table 1. List of texture features used in this study.

Feature Class	Feature Name(:Notation)
Grey-Level Histogram	5th (:5TH) 5thmean(:5THM) 95th (:95TH) 95thmean(:95THM) max(:MAX) min(:MIN) sum(:SUM) mean(:MEAN) entropy(:ETP) kurtosis(:KTS) sigma(:STD) skewness(:SKEW)
Co-occurrence	cluster shade (:CSD) energy(:ENG) entropy(:CETP) inertia(:INT) Correlation(:COR) haralick correlation(:HCOR) inverse difference moment (:IDM)
Run-Length	grey level nonuniformity(:GLN) run length nonuniformity(:RLN) run percentage(:RP) high grey level run emphasis(:HGRE) long run emphasis(:LRE) low grey level run emphasis(:LGRE) short run emphasis(:SRE)

These features have been shown to have values in breast cancer risk estimation[4-7]. For each DM image, and each texture feature in Table 1, the texture image is generated by calculating the feature within a series of adjacent square regions (e.g., lattice) covering the original breast region, as shown in the fourth figure in Figure 1. There are several parameters involved in generating the texture feature images. Co-occurrence features are a class of features describing the spatial relationship between the neighborhood pixels and features in Table 1 are calculated based on a pre-constructed Grey Level Co-occurrence Matrix[8] (GLCM). This matrix depends on parameters including the number of grey-levels, the length and angle of offset, where offset defines the size and direction of the neighborhood region for each pixel. In our study, the texture feature images with offset directions 0°,45°,90°, and 135° are averaged such that features are orientation independent. The proper choice of offset length can be potentially dependent on the spatial resolution of images, which is a property related to the MTF of x-ray detectors. It is well known that in digital mammogram systems, the loss of spatial resolution is caused by the blurring introduced by the phosphor. If the offset length is not chosen properly, the blurring effect can be enhanced by considering a relatively small neighborhood, and the detector effects on texture can be significant. With this consideration, we analyzed the relationship between offset length and effects of system differences on the co-occurrence features. In this study, the offset length was increased from 1 to 10 pixels consecutively.

2.4 Statistical analysis

We used the two-sample two-sided Kolmogorov-Smirnov (K-S) test[12,13] to compare the texture feature distributions between the two DM systems, where only the distributions within the breast region are considered. The two sample K-S test is known as one of the most useful non-parametric and distribution free statistical test. As the distribution of texture features within the breast region is not normally distributed in general, it makes the K-S test a proper choice for the distribution comparisons. The statistic used in the K-S test is called the K-S distance (denoted as D in (1)), defined as the maximum of the absolute vertical difference between two cumulative distribution function (CDF) curves from two distribution samples. The CDF curve describes the overall distribution of the texture feature. To be more specific, the K-S distance is calculated as:

$$D(T) = \max_x | F(S_1, T) - F(S_2, T) | \quad (1)$$

Here, $F(S_1,T)$, $F(S_2,T)$ is the cumulative distribution function of the texture image for feature T generated for the images acquired from S_1 and S_2 respectively. S_1 and S_2 stand for GE 2000D, DS in our study, but could also be generalized to other systems. This indirect comparison of such a distribution property has the advantage of avoiding biases and artifacts for example from pixel shifting, when performing pixel-wise comparisons, and other such slight noise effects. The significance level of the K-S test is chosen to be at 0.05. Features with p-value<0.05 in the test will be labeled as significantly affected by the different image systems. In addition, the K-S distance is used here as a metric to describe the effects of system differences as well as the effects of imaging noise.

2.5 Robust feature identification and standardization

Our proposed texture feature standardization process can be divided into four main steps: First, we compare the CDFs of the grey-level intensities and the texture feature distributions generated from the original raw images of the two imaging

systems. Second, as the two CDFs of the raw image intensities differ by a scale factor, the z-score normalization within the breast region is performed to alleviate this scaling effect. Third, for features that the system difference can not simply be alleviated by z-scoring, a study of feature extraction parameters is further investigated, including the offset length for extracting the co-occurrence features. Finally, the K-S test with significance level 0.05 is applied as a selection scheme for texture features considered robust to system differences and potentially best suitable for texture feature analysis.

3. RESULTS

3.1 Image intensity comparison

As the first step study, the distribution of image intensity within the breast region is compared between the two systems for both original images and the z-score normalized images.

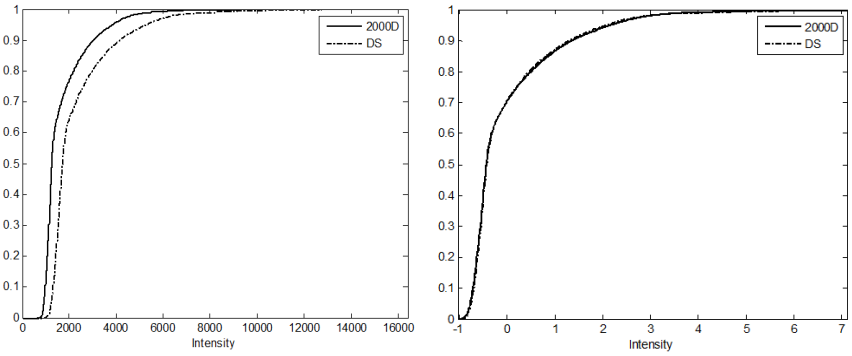

Figure 2. The cumulative distribution function (CDF) curves of the image intensity distribution within the breast region of the phantom image for the GE 2000D and DS systems. The left plot is for the original raw images and the right is for the z-scored raw images. Here, x-axis is image intensity and y-axis is the cumulative distribution value. The comparisons of CDF curves indicate that the image intensity between the two detectors may differ by a scale factor, which can be corrected by z-score normalization.

3.2 Texture feature comparison: system effects and imaging noise effects

In this section, the differences caused by the two systems and by the imaging noises are compared. In Figure 3(4), the texture images are generated from the original (z-score normalized) raw images respectively.

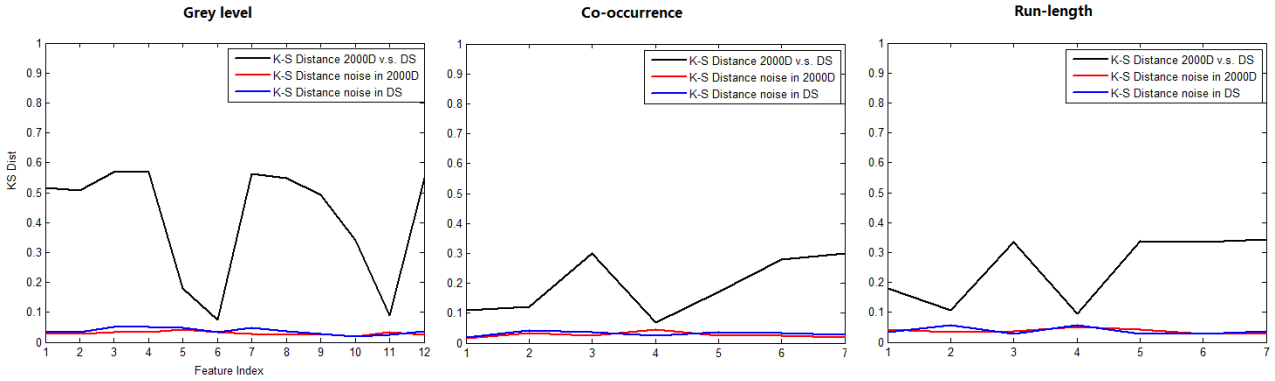

Figure 3. Initially, 26 texture feature images are generated for each raw phantom image. These 26 features are divided into the three sub-figures, left: grey-level (12 features); middle: co-occurrence (7 features, with the default offset length of 1 pixel); right: run-length texture features (7 features). The index is ordered as in Table 1, from top to below and from left to right, For each feature, the K-S distance between the two CDF curves of the feature distribution within the breast region in each detector is shown (black line). The x-axis is the index of the feature within each feature group, and the y-axis is the corresponding K-S distance. The black/red/blue curve stands for K-S distance due to different image systems, noise (e.g. quantum noise) in 2000D and noise in DS. (Results are based on the lattice window size of 63 pixels for texture feature extraction).

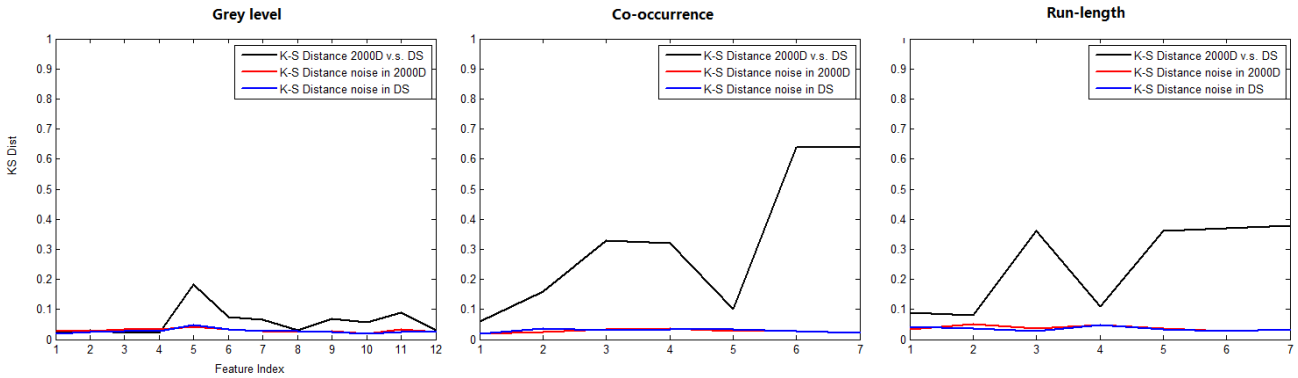

Figure 4. Compared to Figure 3, the three sub-figures are plotted for the same texture feature generated from the z-score intensity normalized raw images. The K-S distance is reduced towards the noise level. In the left plot, for grey-level histogram features, the K-S distance is close to the level of noise, which is supported by the observation from Figure 2 that a scale factor may exist in grey level intensity between the two detectors. Most of the features in the second and third group remain different from noise levels, implying that the detector effects on these features can not be alleviated by a simple z-score intensity transformation.
(Results are based on the lattice window size of 63 pixels for texture feature extraction)

3.3 Effects of offset length as a parameter in generating co-occurrence features

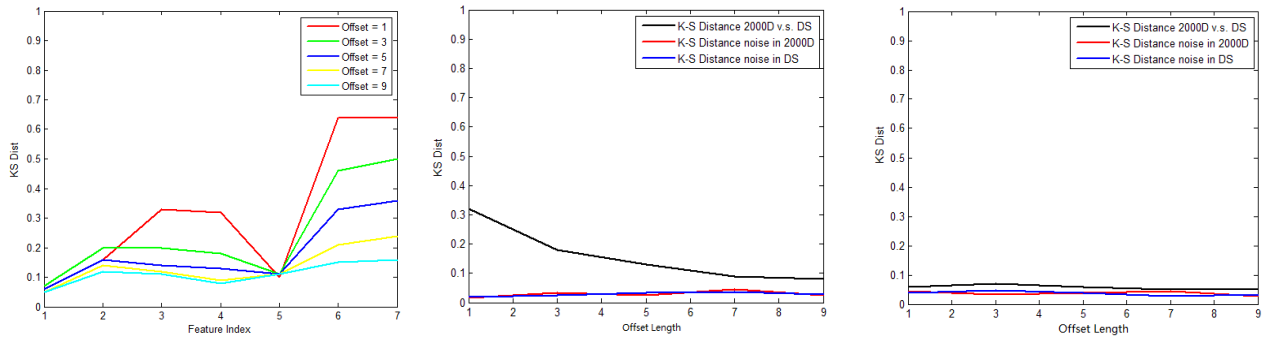

Figure 5. The left figure shows the effect of offset length on the K-S distance for all 7 co-occurrence features, the x-axis is the feature index; y-axis is the K-S distance. Color stands for different offset length when estimating gray-level spatial co-occurrence frequencies required for the computation of the GLCM, as indicated in the legend of the plot. Choosing two examples, the middle figure shows the effect of offset length choice on the co-occurrence feature entropy, and the right figure for feature cluster shade. x-axis is offset length; y-axis is the K-S distance. The black/red/blue curve stands for K-S distance between the two systems, noise in 2000D and noise in DS respectively. Features that are (possibly) affected by blurring(middle figure), effects of system differences drop as the offset is increased from 1-10 pixels; for features that are not affected, increasing offset length doesn't change the K-S distance too much, as a result will not change the statistical test result. (Results are based on the lattice window size of 63 pixels for texture feature extraction)

3.4 Features selected based on K-S test

As one example to show the proposed workflow, in Table 2, the K-S distance and K-S test results are shown for analysis based on texture feature generated using window size of 63 pixels. The first column is the feature class, the second is the feature notation as explained in Table 1. The third column is the K-S test based on the texture features generated from the original raw image; the fourth column is based on features generated from the z-scored raw images. As an additional step for studying the effects of parameters, the last column, we generate the co-occurrence features for z-scored images using the offset length of 5, 7, 9 pixels. In each cell of the table, the number indicates the K-S distance, features considered not significantly different (i.e. $p>0.05$) are denoted with a '*' . Those are deemed as robust features that can be chosen for the texture feature analysis. The notation '-' in the cell means the information is not applicable.

Table 2. List of texture features selected for texture feature analysis.
(Results are based on the lattice window size of 63 pixels for texture feature extraction)

Feature Class	Feature	Original	Z-score	Offset (pixel)		
				5	7	9
Grey level histogram	5TH	0.35	0.03*	-		
	5THM	0.35	0.03*	-		
	95TH	0.38	0.02*	-		
	95THM	0.38	0.02*	-		
	ETP	0.18	0.18	-		
	KTS	0.06*	0.08*	-		
	MAX	0.40	0.06*	-		
	MEAN	0.36	0.03*	-		
	MIN	0.36	0.07*	-		
	STD	0.31	0.06*	-		
	SKEW	0.05*	0.09*	-		
	SUM	0.36	0.03*	-		
Co-occurrence	CSD	0.07*	0.06*	0.06*	0.05*	0.05*
	COR	0.15	0.16	0.16	0.14	0.12*
	ENG	0.39	0.33	0.14	0.12*	0.11*
	CETP	0.36	0.32	0.13*	0.09*	0.08*
	HCOR	0.08*	0.10*	0.11*	0.11*	0.11*
	INT	0.32	0.44	0.33	0.21	0.15
	IDM	0.33	0.45	0.36	0.24	0.16
Run-length	GLN	0.18	0.09*			
	HGLRE	0.11*	0.08*	-		
	LRE	0.33	0.36	-		
	LGLRE	0.10*	0.11*	-		
	RLN	0.34	0.36	-		
	RP	0.33	0.37	-		
	SRE	0.34	0.38	-		

4. DISCUSSION

There are many challenges in the process of the texture feature standardization across different imaging systems. As a preliminary evaluation, our results in this study reveal that it's important to understand the physics of imaging before extracting texture features for analysis, implying that the proper choice of parameter in generating texture features is important. The analysis process in this work can be potentially generalized as a way to standardize texture features used in mammographic texture analysis across different imaging systems in applications such as breast cancer risk assessment. Considering the transition to future clinical applications, one of the limitations of the current work is that the limited number of detectors compared. There's only one breast phantom used in our study, and as GE 2000D, GE DS are manufactured from the same vendor, the differences might be relatively smaller when compared across various vendors. In order to make the work more promising for broader applications, it would be important to include additional commonly used digital mammography systems for comparison, and ideally additional imaging phantoms for analysis.

For our study, results were generated using a window size of 63 pixels. In fact, window size is also one parameter having effects on texture feature analysis. Using different window sizes to extract the local texture features of the image constructs the multi-resolution description of the parenchymal patter. Part of our ongoing work is to study the optimal choice of the window size in texture analysis study. Smaller versus larger window size could potentially capture more versus less information of the DM image, however maybe more sensitive to system differences. It will be worthwhile to study the balance between the bias introduced by system differences and the performance of feature discrimination.

5. CONCLUSION

In this study, we compare the texture feature from two GE Digital mammography (DM) systems using a physical breast phantom, and propose a general process for texture feature standardization across different devices. Our results show that the Cumulative Distribution Function (CDF) curves for raw image intensity values between the two systems reveal a scaling pattern. The z-score normalization and proper choice of offset length in generating co-occurrence features can help alleviate the effects introduced by system differences. These results are validated using the Kolmogorov-Smirnov (K-S) test on the CDF curves at significance level $p=0.05$. Features not significantly ($p>0.05$) affected by the two systems can be selected as robust for any further studies performing digital mammogram texture analysis using these systems. Further work is underway to consider more detectors from different vendors for comparison purposes. We also plan to take into consideration several additional parameters associated with texture feature generation, in order to provide a more generalized feature standardization scheme for digital mammography texture analysis.

ACKNOWLEDGEMENTS

This work was supported in part by the American Cancer Society (ACS) Research Scholar Grant (RSGHP-10-108-01-CPHPS), the National Institutes of Health/National Cancer Institute Research Grants (1U54CA163313-01, R01-CA161749-01A1) and by a Breast Cancer Alliance Young Investigator Research Grant.

REFERENCES

[1] Smith, K.L., Issacs, C., "Management of women at increased risk for hereditary breast Cancer,". Breast Disease 27, 51–67 (2006).

[2] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., Mulvihill, J.J., "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, ", Natl. Cancer 81(24), 1879–1886 (1989).

[3] Boyd, N.F., Guo H., Martin L.J., Sun L., Stone J., Fishell E,. et al. "Mammographic density and the risk and detection of breast cancer, " New England Journal of Medicine. 356(3):227-36(2007).

[4] McCormack, V.A., dos Santos Silva, I., "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis, " Cancer Epidemiology, Biomarkers & Prevention 15, 1159-1169(2006)

[5] Manduca, A., Carston, M.J., Heine, J.J., Scott, C.G., Pankratz, V.S., Brandt, K.R., et al., "Texture Features from Mammographic Images and Risk of Breast Cancer, " Cancer Epidemiol. Biomarkers Prev. 18(3), 837–845 (2009)

[6] Li, H., Giger, M.L., Huo. Z, Olopade. O.I, Lan. L, Weber. B.L., et al. " Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location," Medical Physics 31(3), 549-555(2004)

[7] Huo, Z., Giger, M.L., Zhong, W., Cumming, S., Olopade. O.I, " Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection, " Medical Physics 27(1), 4-12(2000)

[8] Haralick, R.M., Shanmugam, K., Dinstein, I.H., "Textural features for image classification, " IEEE Transactions on Systems, Man and Cybernetics 3, 610–621 (1973)

[9] Galloway, M.D., "Texture classification using gray level run length, " Computer Graphics and Image Processing 4, 172–179 (1975)

[10] Caldwell, C.B., Yaffe, M.J., "Development of an anthropomorphic breast phantom, " Medical Physics 17(2), 273–280 (1990)

[11] Zheng, Y., Keller, B.M., Wang, Y., Tustison, N., Song, G., Bakic, P.R., Maidment, A.D. , Conant, E.F., Gee, J.C., Kontos, D., "A Fully-Automated Software Pipeline for Parenchymal Pattern Analysis in Digital Breast Images: Toward the Translation of Imaging Biomarkers in Routine Breast Cancer Risk Assessment, " Quantitative Imaging Reading Room, the 97th Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA) 2011, Chicago, IL (software exhibit). (2011)

[12] Smirnov, N., "Tables for estimating the goodness of fit of empirical distributions, " Annals. of Mathematical Statistics 19, 279-281(1948)

[13] Rachev, S.T., [Probability Metrics and Stability of Stochastic Models], John Wiley & Sons (1991)

[14] Wang, Y., Keller, B.M., Zheng, Y., Acciavatti, R.J., Gee, J.C., Maidment A.D.A., Conant, E.F., and Kontos D., "A Phantom Study for Assessing the Effect of Different Digital Detectors on Mammographic Texture Features, "Digital Mammography/IWDM, 604-610(2012)