Validation of no-reference image quality index for the assessment of digital mammographic images

Helder C. R. de Oliveira[a], Bruno Barufaldi[a], Lucas R. Borges[a], Salvador Gabarda[b], Predrag R. Bakic[c], Andrew D. A. Maidment[c], Homero Schiabel[a], Marcelo A. C. Vieira[a]

[a]Department of Electrical and Computer Engineering, University of São Paulo, São Carlos, Brazil
[b]Institute of Optics, Spanish Council for Scientific Research, Madrid, Spain
[c]Department of Radiology, University of Pennsylvania, Philadelphia, USA

ABSTRACT

To ensure optimal clinical performance of digital mammography, it is necessary to obtain images with high spatial resolution and low noise, keeping radiation exposure as low as possible. These requirements directly affect the interpretation of radiologists. The quality of a digital image should be assessed using objective measurements. In general, these methods measure the similarity between a degraded image and an ideal image without degradation (ground-truth), used as a reference. These methods are called Full-Reference Image Quality Assessment (FR-IQA). However, for digital mammography, an image without degradation is not available in clinical practice; thus, an objective method to assess the quality of mammograms must be performed without reference. The purpose of this study is to present a Normalized Anisotropic Quality Index (NAQI), based on the Rényi entropy in the pseudo-Wigner domain, to assess mammography images in terms of spatial resolution and noise without any reference. The method was validated using synthetic images acquired through an anthropomorphic breast software phantom, and the clinical exposures on anthropomorphic breast physical phantoms and patient's mammograms. The results reported by this no-reference index follow the same behavior as other well-established full-reference metrics, e.g., the *peak signal-to-noise ratio* (PSNR) and *structural similarity index* (SSIM). Reductions of 50% on the radiation dose in phantom images were translated as a decrease of 4dB on the PSNR, 25% on the SSIM and 33% on the NAQI, evidencing that the proposed metric is sensitive to the noise resulted from dose reduction. The clinical results showed that images reduced to 53% and 30% of the standard radiation dose reported reductions of 15% and 25% on the NAQI, respectively. Thus, this index may be used in clinical practice as an image quality indicator to improve the quality assurance programs in mammography; hence, the proposed method reduces the subjectivity inter-observers in the reporting of image quality assessment.

Keywords: no-reference image quality assessment, blind index, digital mammography, PSNR, SSIM.

1. INTRODUCTION

The most reliable method for image quality assessment (IQA) is by subjective evaluation[1]. Indeed, the mean opinion score (MOS) is considered one of the best methods of image quality measurement, but it requires several human observers to evaluate a large number of images. For digital mammography, the evaluation of mammographic images in clinical practice is a hard task, since it must be performed by experienced radiologists; furthermore, it is time-consuming and expensive[2].

The *peak signal-to-noise ratio* (PSNR) and the *mean squared error* (MSE) are widely used throughout the literature of image processing and many other signal-processing fields. PSNR is a logarithmic interpretation of MSE and there is a functional monotonicity between both measures. These objective measurements are not always in agreement with the human visual system and they are not feasible enough for most applications[3]. In order to overcome these drawbacks, new metrics such as the *structural similarity index* (SSIM)[4] were developed based on the human visual system. These indexes are capable of mimicking human visual perception, and some of them assume that it is possible to have a good approximation of the image without any kind of degradation. This class of metrics is named full-reference image quality assessment (FR-IQA)[1].

Approaches for FR-IQA require an ideal image without degradation (ground-truth) as reference to assess the quality of a degraded image. Studies have shown a satisfactory level of performance, as demonstrated by high correlations with

Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Matthew A. Kupinski, Proc. of SPIE Vol. 9787, 978713
© 2016 SPIE · CCC code: 1605-7422/16/$18 · doi: 10.1117/12.2217229

human subjective evaluation[3,5]. On the other hand, the dependency of a ground-truth limits the application domains of FR-IQA methods and it does not succeed in clinical practice for digital mammography, since it is not feasible to acquire clinical images without any degradation.

Given that a reference image in clinical practice does not exist, the only method that can be embedded into this application is based on no-reference image quality assessment (NR-IQA) or "blind" assessment[1,6]. The majority of blind indexes require deep knowledge of the image degradation behavior or computationally trained systems, which are based on human model observers[7,8].

The two most relevant degradations found in digital mammographic images are noise and blur. Quantum noise is dominant in digital mammograms. It follows the Poisson distribution and it is dependent on the radiation dose[9]. The spatial resolution is a function of the detector size and the focal spot dimensions.

Gabarda *et al*. developed a blind image quality assessment through anisotropy[6] and visual perception, named Anisotropic Quality Index (AQI). This metric provides quantitative data that differ noise and blurring levels of natural scene statistics (NSS) images. In the current study, the authors suggest to modify and expand the original anisotropic quality index to be suitable enough for assessing the image quality in digital mammography.

This study aims to develop a computerized system to assess clinical images in digital mammography without a reference image, using a Normalized Anisotropic Quality Index (NAQI). To validate the method, several synthetic and clinical images of anthropomorphic breast phantoms were tested at different levels of blur and noise, followed by clinical mammograms. Thus, mammographic images can be assured, hence reducing subjectivity of human evaluations.

2. MATERIALS & METHODS

2.1 Mathematical background

According to Gabarda, *et al*., the anisotropy found in natural images is sensitive to both noise and blur. Images with lower quality, i.e., higher noise levels and lower spatial resolution, report lower anisotropy values. This property can be used as a base for a metric capable of measuring image quality[6].

The method is based on entropy histograms calculated using the generalized Rényi entropy. Let $P[n,k]$ be a discrete space-frequency distribution, the Rényi entropy calculated for this distribution (R_α) is given by:

$$R_\alpha = \frac{1}{1-\alpha}\log_2\left(\sum_n\sum_k P^\alpha[n,k]\right), \tag{1}$$

where n and k are spatial and frequency components, respectively, and α is a constant for measurement of space-frequency distributions.

The desired space-frequency distribution is given by the normalization performed with the windowed pseudo-Wigner distribution, as described in Eq. 2.

$$W_z[n,k] = 2\sum_{m=-N/2}^{\left({}^{N}\!/_{2}\right)-1} z[n+m]z^*[n-m]e^{-2i\left({}^{2\pi m}\!/_{N}\right)k}, \tag{2}$$

where n and k represent the time and frequency discrete variables, respectively, m is a shifting parameter and z^* is the complex conjugate of z. In order to account for the entropy direction, $z[n]$ is a 1D sequence of N gray values of pixels.

Thus, based on Eqs. 1 and 2, the Rényi entropy for a pixel located at the position n, described at the pseudo-Wigner domain, is given by:

$$R_3[n] = -\frac{1}{2}\log_2\left(\sum_{k=1}^{N} \breve{P}_n^3[k]\right) \tag{3}$$

Anisotropic Quality Index (AQI) metric, based on result of Eq. 3, is given by:

$$\bar{R}[\theta_s] = \sum_n R_3[n, \theta_s]/M \tag{4}$$

where M is the image size and $\theta_s \in [\theta_1, \theta_2, \dots, \theta_S]$ represents the S different orientations which are considered to measure the entropy.

For the proposed method, two parameters are given as input along with the degraded image. One of them is the window size (W) used during the calculation of the pseudo-Wigner distribution. The second is how many orientations (S) must be considered when calculating the entropy.

Dividing the AQI value by the expected Rényi entropy we can calculate a normalization of AQI (NAQI). Gabarda *et al.* emphasized that entropy presents a relationship with image quality[6]. However, in statistical dispersion of entropy, blur and noise have an inverse correlation. To overcome that problem, the current study uses the anisotropy with directional dependency, since this measurement decreases according to the amount of degradation. As mentioned previously, high spatial resolution and low noise degradation are important to assure good quality of mammograms.

2.2 Digital Phantoms

The preliminary assessment of the index proposed in this study was performed using an anthropomorphic breast software phantom developed at the University of Pennsylvania[10]. A reference image without noise and spatial resolution of 14 pixels/mm was generated to be used as the ground-truth, as seen in Figure 1 (A). The experiments were tested in ROI inside the breast to avoid bias from the background. The ROI size (2048 × 512 pixels) was selected to contain as much as possible tissue inside the breast, as seen in Figure 1 (B).

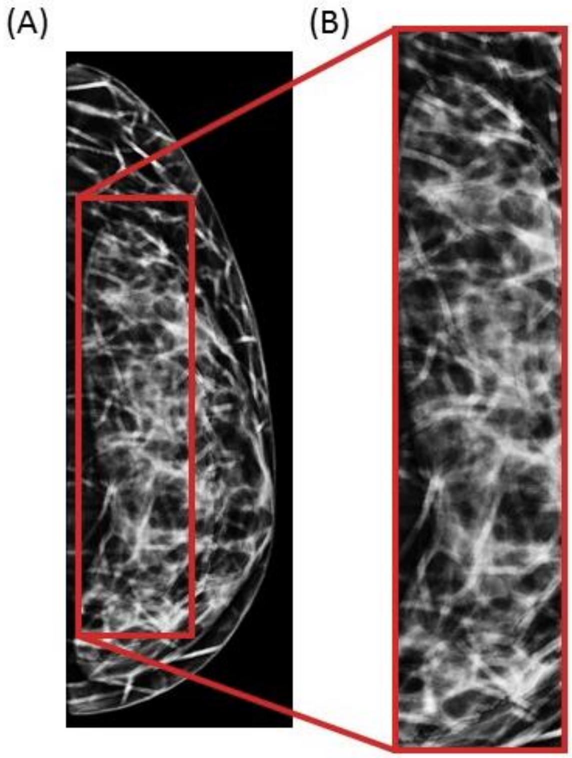

Figure 1 – (A) Example of a reference digital phantom image without noise and with high spatial resolution. (B) ROI used during the experiments.

The reference image (ground-truth) was modified using two different sources of degradation: blur and noise. Blur was introduced into the images using a convolution of Gaussian mask with standard deviations ranging from 1 to 6 in integer steps. This mask was created with the same size as the reference image.

Different levels of Poisson noise were added into the blurred images. Since the radiation dose affects the noise level of the image, higher noise degradation is generated by lower radiation exposures. Thus, the simulated radiation doses ranged from 100% to 50% of the standard radiation dose for mammography in steps of 10%. In total, 49 images were used to perform this experiment.

2.3 Physical Phantoms

After the preliminary assessment with the synthetic images, additional tests were performed using two different anthropomorphic breast physical phantoms. The first one was prototyped by CIRS, Inc. (Reston, VA) with a license from the University of Pennsylvania[10]. This phantom consists of six slabs, each containing simulated anatomical structures manufactured using tissue mimicking materials, based upon a realization of the breast software phantom developed at Penn. The phantom simulates a 450 mL breast, compressed to 5cm, with 17% volumetric breast density (excluding the skin). The second breast phantom used in this study is well known as "Rachel" (*Model 169, Gammex, RMI*, Madison, WI), and consists of tissue-mimicking materials disposed inside a PMMA case simulating the shape of a human breast [11].

The physical phantom images were acquired using the mammographic unit *Selenia Dimensions* (*Hologic Inc., Bedford MA*). This mammography system is equipped with an amorphous selenium (a-Se) detector layer with a thickness of 250µm and pixel pitch of 70µm. The exposures were acquired using W/Rh as target-filter combination.

For the phantom developed by the University of Pennsylvania, the following doses were acquired manually: 100% (AEC mode, with incident air kerma of 6.05 mGy, 160 mAs and 29 kVp), 87.5%, 75% and 50% of the AEC indent air kerma, achieved by decreasing the mAs. The estimation of the incident air kerma was taken from the DICOM header using tag {0040,8302} "Entrance Dose in mGy". We acquired 15 realizations of the AEC configuration (100%), and the average between ten of those images is assumed a good approximation of a noiseless image. The remaining five images were used for testing the objective metrics (PSNR, SSIM and NAQI). For the cohorts generated by dose reductions, we performed only five realizations (the ground-truth is not required). Figure 2 (A) shows an example of a standard dose acquisition of this phantom. The ROI (2200 × 480 pixels), shown in Figure 2 (B). It should be stressed that the ROI size was selected to maximize the area inside the breast and avoid bias from the image background.

The exposures on the "Rachel" phantom were also acquired manually. The reference dose (100%) was given by the AEC mode, with incident air kerma of 8.06 mGy, 140 mAs and 31 kVp. Similarly, a set of images was generated by consecutive reductions in the current-time exposure of 83%, 71% and 50% of the reference dose. The average between ten of 15 realizations was assumed as a good approximation of the ground-truth image and the remaining images were used for testing. To test the objective metrics, five images were acquired using the exposure factors with dose reductions. Figure 2 (C) shows an example of a reference dose acquisition of the Rachel phantom and Figure 2 (D) the ROI (1700 x 480) used for tests.

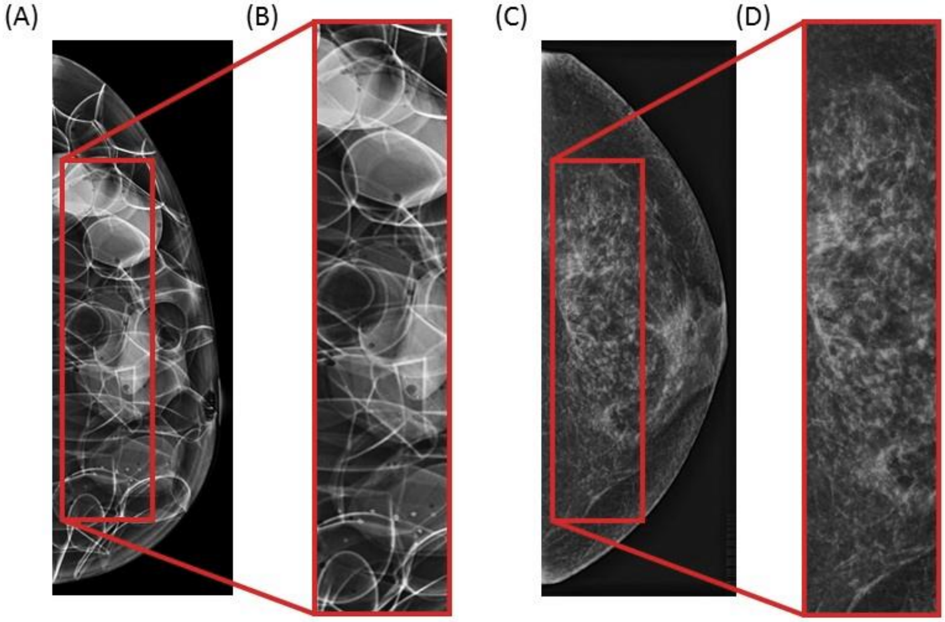

Figure 2 – Physical phantoms used during the experiments. (A) Anthropomorphic phantom from University of Pennsylvania. (B) ROI used during the experiments. (C) "Rachel" Anthropomorphic phantom. (D) ROI used during the experiments.

2.4 Clinical images

Finally, the last set of tests was performed using clinical mammograms (patient's exams). The images were selected from the American College of Radiology Imaging Network (ACRIN) PA 4006 study, conducted at the Hospital of the University of Pennsylvania. We analyzed two mammograms (one CC and one MLO) from two different patients. The CC image was acquired using the following exposure factors: 29 kVp, 151 mAs, W/Rh as target-filter combination, entrance dose of 5.33 mGy and the compressed breast thickness is 54 mm. The estimation of the mean glandular dose (MGD) was 1.43 mGy, calculated from the organ dose values held in the DICOM headers using tag {0040,0316} "Organ Dose". The second image (MLO view) was acquired using 31 kVp, 130 mAs, W/Rh as target-filter combination, entrance dose of 5.59 mGy and 61 mm for compressed breast thickness. For this image, the estimation for MGD was 1.40 mGy. Figure 3 shows the clinical images, along with the corresponding ROI used during the experiments.

To analyze the index behavior at different dose levels, images acquired from the same patient at different radiation doses were needed. However, the patients cannot be submitted to multiple exposures for clinical trials without IRB approval, due to radiation-related risks. To overcome this limitation, we simulate the dose reduction in clinical practice using a method presented in previous study[9].

The simulation method requires three images as input. The first one is the clinical image, acquired using a standard exposure factor, usually selected by the AEC mode. The second one must be a uniform image acquired using the exposure factors as close as possible from the clinical image. The last image must be acquired using reduced current-time exposure (mAs), at the dose level that we wish to simulate.

Thus, we simulated dose reduction using the clinical images shown in Figure 3. For the first patient, we simulated 92%, 80% and 53% of the standard dose and for the second patient we simulated 62% and 30% of the standard dose. Using the simulated images, we analyzed the behavior of the index at different radiation doses. We also performed additional tests to assess the dependency between the reported value of the index and the selected region of interest of the breast.

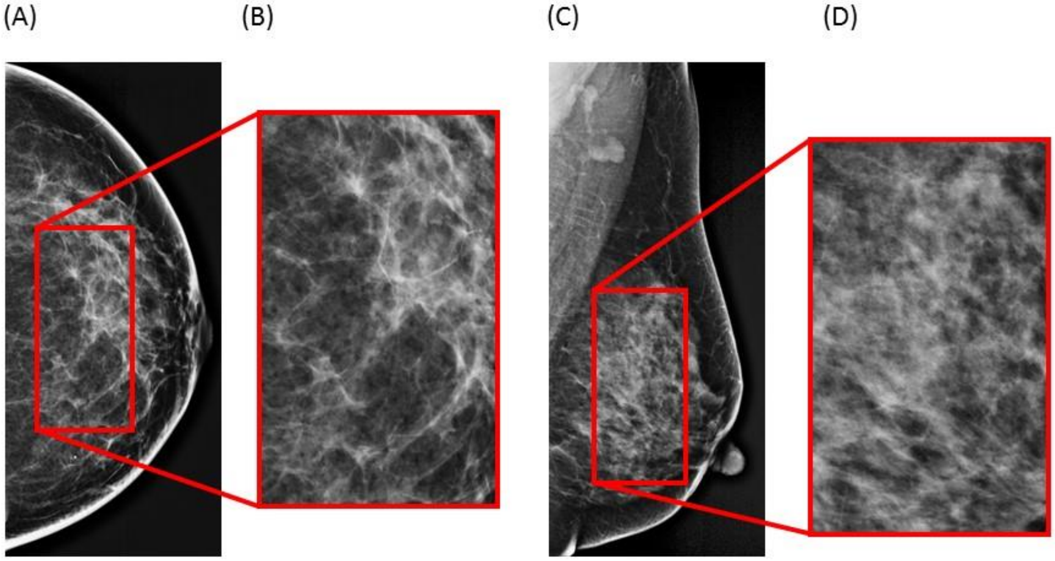

Figure 3 – Clinical cases analyzed during the experiments. (A) CC view of patient A. (B) ROI selected from patient A. (C) MLO view of patient B. (D) ROI selected from patient B.

2.4 Data analysis

To assess objectively the performance of the proposed index, the NAQI was calculated for images with different levels of degradation and compared to other well-established QA metrics, such as the *peak signal-to-noise ratio* (PSNR) and *structural similarity index* (SSIM). NAQI parameters were set to $W = 16$ pixels and $S = 4$ orientations for all experiments conducted in this work[6]. The PSNR value is reported in decibels (dB) and it is given by the ratio between noise and true signal. Higher reported values for the PSNR can be interpreted as better image quality. The SSIM is based on human visual perception and can report values ranging from -1 to +1, with higher values meaning better similarity between the assessed image and the ground-truth.

3. RESULTS

3.1 Digital Phantoms

The results indicate that the proposed blind index decreases as the image is more degraded. The PSNR and the SSIM follow the same behavior, however, they need a reference image to be calculated. As predicted, the highest values are achieved by images with less degradation. Figure 4 shows the surface fitting, using cubic interpolation, for these metrics among images corrupted by noise and blur.

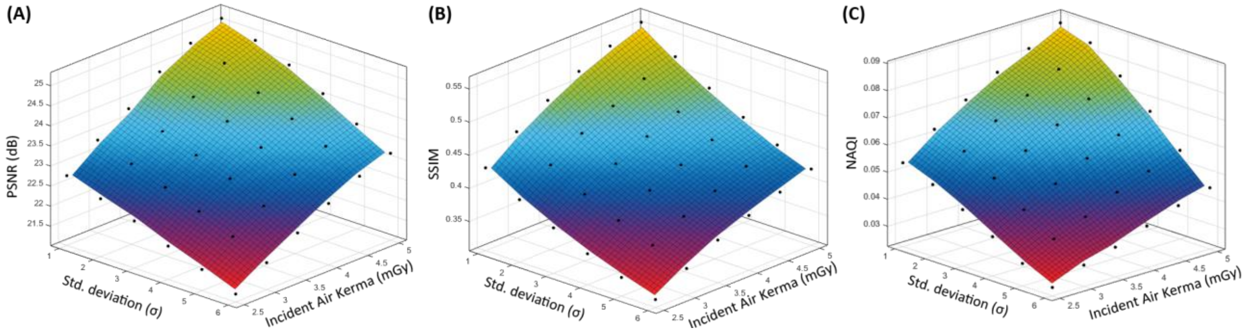

Figure 4 – (A) PSNR, (B) SSIM and (C) NAQI as function of the incident air kerma in mGy (varying noise) and standard deviation of a Gaussian mask (modifying blur) for the synthetic phantom images.

3.2 Physical Phantoms

The analysis of the physical phantom images was performed using the average of five realizations at each radiation dose. Reported values of NAQI are plotted in Figure 5 and Figure 6. Unlike the digital phantom images, the noise and blur were not simulated in this experiment. The acquisitions contain both degradations, but only the noise level varies due to the dose reduction. Figure 5 shows the results obtained by the objective metrics using the Penn physical phantom. As mentioned previously, lower dose levels result in lower values of PSNR, SSIM and NAQI, indicating lower image quality. It should be stressed that the NAQI does not require a reference image to assess the image quality. The error bars indicate the standard deviation among five different realizations of each dose, which is minimal in both metrics.

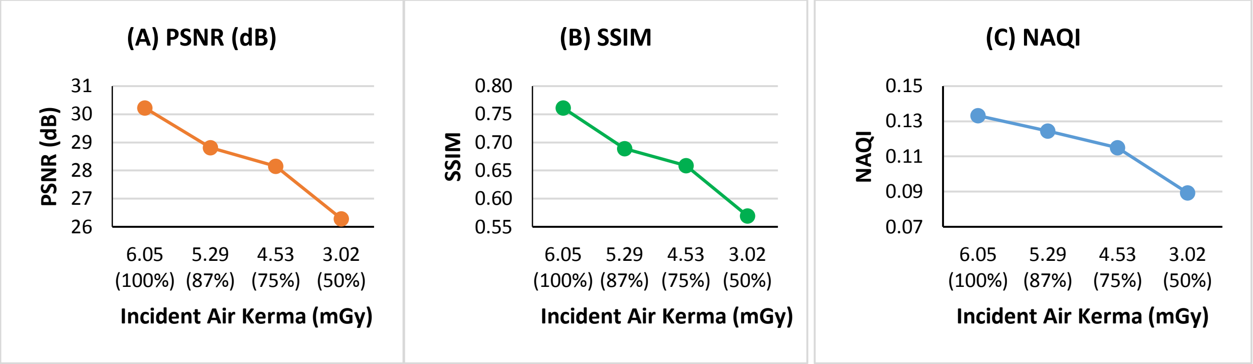

Figure 5 – (A) PSNR, (B) SSIM and (C) NAQI as function of the incident air kerma in mGy for the Penn physical phantom images acquired using the Hologic system.

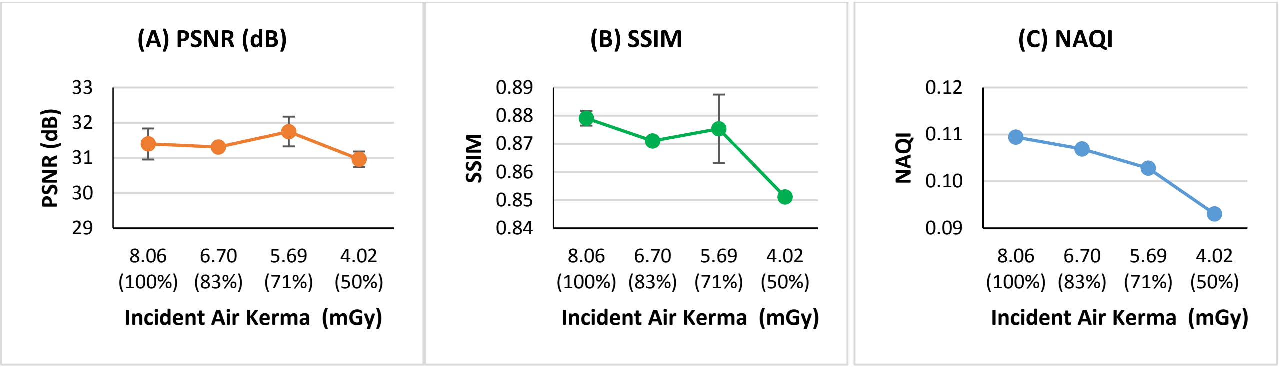

Figure 6 – (A) PSNR, (B) SSIM and (C) NAQI as function of the incident air kerma in mGy for the "Rachel" phantom images acquired using the Hologic system.

The calculated indexes for the "Rachel" phantom are plotted in Figure 6. In this case, the relation between incident air kerma, PSNR and SSIM was not easily defined. The PSNR and SSIM indexes reported higher standard deviation among the five realizations and the mean value of the indexes presented a slight increase when the entrance dose decreased. The justification for such behavior requires further analysis; however, we have the hypothesis that this behavior can be related to how the Rachel phantom is built. We believe that the noise due to milling artifacts can overpower the decrease of the signal to noise ratio due to dose reduction in this phantom, therefore making it difficult to identify an objective relation between dose reduction and the reported quality index. The NAQI values respected the previous behavior, presenting lower values for reduced-dose images.

3.3 Clinical images

The final experiment was performed using clinical images from two patients. Different radiation doses were obtained using a method for simulating dose reduction in clinical images[9]. Since it is not possible to obtain a good approximation of the ground-truth image, other well-established metrics such as PSNR and SSIM are not suitable for clinical images. Figure 7 reports the NAQI index calculated at different radiation levels.

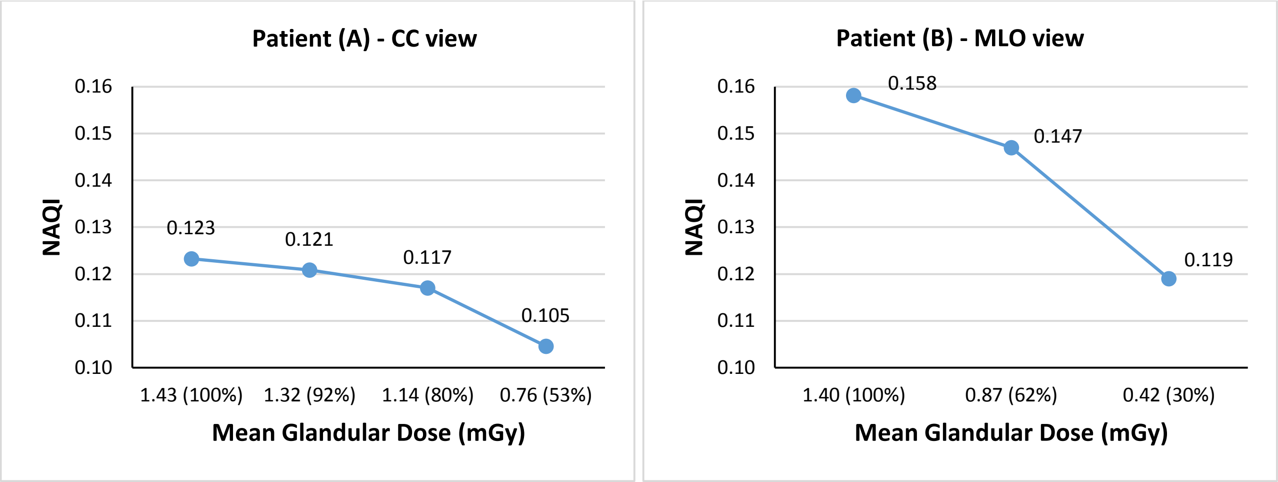

Figure 7 – NAQI calculated for both patients at different mean glandular doses.

Additional tests were performed to assess the dependency between the reported NAQI values and the selected ROI. In this experiment, we selected different ROI sizes for both patients with fixed radiation dose of 100% (AEC mode). Figure 8 shows the calculated metric for both patients.

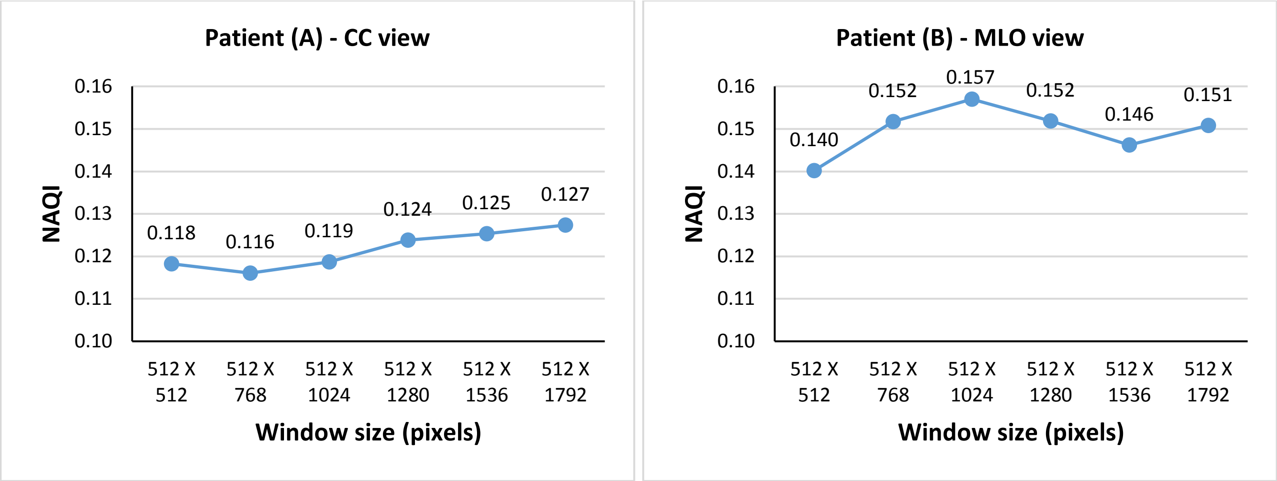

Figure 8 – NAQI calculated for both patients using different ROI sizes.

Noteworthy, the graph reported on Figure 8 uses the same Y scale as in Figure 7. As expected, changes on the ROI size resulted on small variations of the index, and no specific trend can be noticed on the results. However, since in this preliminary experiment we considered an extremely limited number of patients, it is not possible to draw any conclusions on whether this difference is statistically significant or not. Therefore, future work may include a statistical analysis of how the ROI selection influences the NAQI, and how to select the size and position of such ROI.

4. CONCLUSION

Although the best method for assessing the quality of degraded images is performed by human readers, this task turns out to be cumbersome and subjective. The current study provides an IQA approach that evaluates blur and noise blindly, since the reference image is not available in clinical practice. Thus, this index may be used as an image quality indicator, reducing the subjectivity inter-observers, regardless a large number of readers.

ACKNOWLEDGMENTS

The project described is supported by CAPES grant #99999.014175/2013-04 (Methods for Quality Control in Digital Mammography), NIH grant #U54-CA163313-04 (Penn Center for Innovation in Personalized Breast Screening), ACRIN PA 4006 (Comparison of Full-Field Digital Mammography with Digital Breast Tomosynthesis Image Acquisition) and São Paulo Research Foundation (FAPESP) grant #2013/18915-5.

REFERENCES

[1] Wang, Z., Bovik, a. C., Modern image quality assessment, Synth. Lect. Image, Video, Multimed. Process. **2**(1) (2006).

[2] Barufaldi, B., Lau, K. C., Schiabel, H., Maidment, A. D. A., “Computational assessment of mammography accreditation phantom images and correlation with human observer analysis,” SPIE Med. Imaging, Orlando, FL (2015).

[3] Wang, Z., Bovik, A. C., “Mean Squared Error : Love It or Leave It ?,” IEEE Signal Process. Mag. **26**(January), 98–117 (2009).

[4] Wang, Z., Bovik, A. C., Sheikh, H. R., Member, S., Simoncelli, E. P., Member, S., “Image Quality Assessment : From Error Visibility to Structural Similarity,” IEEE Trans. Image Process. **13**(4), 1–14 (2004).

[5] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process. **13**(4), 600–612 (2004).

[6] Gabarda, S., Cristóbal, G., “Blind image quality assessment through anisotropy.,” J. Opt. Soc. Am. A. Opt. Image Sci. Vis. **24**(12), B42–B51 (2007).

[7] Liu, L., Dong, H., Huang, H.., Bovik, A. C., “No-reference image quality assessment in curvelet domain,” Signal Process. Image Commun. **29**(4), 494–505 (2014).

[8] Mittal, A., Moorthy, A. K., Bovik, A. C., “No-reference image quality assessment in the spatial domain.,” IEEE Trans. Image Process. **21**(12), 4695–4708 (2012).

[9] Borges, L. R., Oliveira, H. C. R. de, Nunes, P. F., Vieira, M. A. C., “Method for inserting noise in digital mammography to simulate reduction in radiation dose,” SPIE Med. Imaging **9412**, 94125J, Orlando, FL (2015).

[10] Bakic, P. R., Zhang, C., Maidment, A. D. A., “Development and characterization of an anthropomorphic breast software phantom based upon region-growing algorithm.,” Med. Phys. **38**(6), 3165–3176 (2011).

[11] Caldwell, C. B., Yaffe, M. J., “Development of an anthropomorphic breast phantom,” Med. Phys. **17**(2), 273–280 (1990).