Automatic intensity windowing of mammographic images based on a perceptual metric

Alberto Albiol
iTeam Research Institute, Universitat Politlècnica de Valéncia (Spain)

Alberto Corbi* and Francisco Albiol
Instituto de Física Corpuscular (IFIC), Universitat de València,
Consejo Superior de Investigaciones Científicas (Spain)

Purpose: Initial auto-adjustment of the window level WL and width WW applied to mammographic images. The proposed intensity windowing (IW) method is based on the maximization of the mutual information (MI) between a perceptual decomposition of the original 12-bit sources and their screen displayed 8-bit version. Besides zoom, color inversion and panning operations, IW is the most commonly performed task in daily screening and has a direct impact on diagnosis and the time involved in the process.

Methods: The authors present a human visual system and perception-based algorithm named GRAIL (Gabor-Relying Adjustment of Image Levels). GRAIL initially measures a mammogram's quality based on the MI between the original instance and its Gabor-filtered derivations. From this point on, the algorithm performs an automatic intensity windowing process that outputs the WL/WW that best displays each mammogram for screening. GRAIL starts with the default, high contrast, wide dynamic range 12-bit data, and then maximizes the graphical information presented in ordinary 8-bit displays. Tests have been carried out with several mammogram databases. They comprise correlations and an ANOVA analysis with the manual IW levels established by a group of radiologists. A complete MATLAB implementation of GRAIL is available at `https://github.com/TheAnswerIsFortyTwo/GRAIL`.

Results: Auto-leveled images show superior quality both perceptually and objectively compared to their full intensity range and compared to the application of other common methods like Global Contrast Stretching (GCS). The correlations between the human-stablished intensity values and the ones estimated by our method surpass that of GCS. The ANOVA analysis with the upper intensity thresholds also reveals a similar outcome. GRAIL has also proven to specially perform better with images that contain micro-calcifications and/or foreign X-ray-opaque elements and with healthy BI-RADS A-type mammograms. It can also speed up the initial screening time by a mean of 4.5 seconds per image.

Conclusions: A novel methodology is introduced that enables a quality-driven balancing of the WL/WW of mammographic images. This correction seeks the representation that maximizes the amount of graphical information contained in each image. The presented technique can contribute to the diagnosis and the overall efficiency of the breast screening session by suggesting, at the beginning, an optimal and customized windowing setting for each mammogram.

I. INTRODUCTION

Displaying radiological images, including digital mammograms, with good quality is essential in order to ensure a good diagnostic experience. The main problem when visualizing mammograms is that their full intensity range is usually much wider (normally 12 bits) than the typical 8 bits that ordinary diagnostic displays can handle [1]. Recent studies have demonstrated that although a human observer can perceive up to 900 shades of gray, most viewing applications only support 256 depth levels [2]. The 2014 *Digital Mammography* report [3] by the American Association of Physicists in Medicine (AAPM), the Society for Imaging Informatics in Medicine, and the American College of Radiology (ACR) establishes a minimum requirement of 8 bits for diagnostic displays. This document also claims that relatively few data have been reported to address possible advantages of higher bit-depth display devices.

In this context, it is quite common for 8-bit-based visualization software to select a window of intensity values (IW) that are linearly mapped to the display range, i.e., 0-255 levels. This step produces an unavoidable loss of information. The radiologist then usually performs one or a series of manual window adjustments by manipulating the window level and width (WL/WW), or more directly, the lowest and highest intensity limits. These elemental and very common operations are reviewed in Section II. Manual IW manipulation is typically carried out with the help of specific user interface controls (mouse dragging events or *ad hoc* sliders). Recommended WL/WW values can also be specified by the digital mammography system used when serializing the data of each image to a DICOM Service Object Class (SOP) Instance [4]. It is up to the

* Email: alberto.corbi@ific.uv.es; Address: Instituto de Física Corpuscular (IFIC), Catedrático José Beltran, 2, 46980 Paterna (Spain); The authors report no conflicts of interest in conducting this research.

Picture Archiving System (PACS) or external DICOM viewer whether or not to apply these stored contrast settings when each study or image is loaded for screening. Most viewers usually ignore these window values and just apply the default ones, which for a 12-bit mammogram correspond to WL = 2048 and WW = 4096 (equal to a range of intensities that goes from 0 to 4095). Also, Global Contrast Stretching (GCS), which is discussed in Section II, is often the default operation adopted by visualization software packages as a quick contrast enhancement strategy.

Even though IW is a very basic image enhancement technique during screening, it can represent an important step [5] towards achieving a correct diagnosis. In this scenario, image quality assessment techniques can play a crucial role in automatically selecting the windowing parameters that achieve the best image contrast and visualization. However, a characteristic problem in quality assessment is the subjectivity that is inherent to the process [6]. For this reason, defining an objective quality assessment method is essential in order to attain experiment reproducibility and diagnostic repeatability. Once this metric is specified, it is then possible to derive an optimizable cost function. Based on this minimization/maximization process, an appropriate IW setting for an image can be found and automatically assigned to it when it is first loaded in the radiologist's screen. As a side advantage, this pre-adjustment may also contribute to saving time by speeding up the screening and diagnostic process [7].

With more detail, the proposed IW method consists of an interplay of the concepts of Mutual Information (MI), Human Visual System (HVS) and Gabor filtering. We begin by computing the MI between the Gabor-filtered representations of both the original and the displayed images. Then, the WL/WW combination that makes this set of MI values maximum is iteratively sought. When the optimal WL/WW is eventually found, it can be applied to the 8-bit displayed mammogram. This optimized and contrast-stretched version is finally presented to the radiologist as a starting point in a conventional screening session. This method is called GRAIL (Gabor-Relying Adjustment of Image Levels). To test its suitability and advantages, a panel of radiologists was asked to manually and independently seek the best windowing setting for a set of digital mammograms representing a wide spectrum of clinical cases. The time taken to perform this adjustment was also measured.

II. MAMMOGRAM CONTRAST ENHANCEMENT TECHNIQUES

Here we review the most commonly used techniques related to mammographic image enhancement through contrast manipulation. As stated in Section I, IW represents a basic operation that is performed in everyday breast screening (and in almost all radiological disciplines). Its foundations are graphically represented in Fig. 1 and in the following equation:

$$j = \mathrm{IW}(i,a,b) = \begin{cases} 0 & \forall \quad i < a \\ 255 \times \frac{i-a}{b-a} & \forall \quad a \leq i \leq b \\ 255 & \forall \quad b < i \end{cases} \tag{1}$$

where i and j account for the intensities of the original 12-bit ($\mathcal{I}$) and the contrast-stretched 8-bit images ($\tilde{\mathcal{I}}$), respectively. From Eq. (1), it is easy to derive that a basic IW operation renders the lowest intensity pixels of $\mathcal{I}$ equal to black ($j = 0$) and the highest intensity ones equal to white ($j = 255$). These intensity thresholds are determined by the a and b parameters, respectively, which are usually manually established and modified by the health professional at the beginning of the screening session. However, they can also be automatically determined (as proposed in this work) or predefined by a group of presets (i.e., to highlight certain types of tissues or densities).

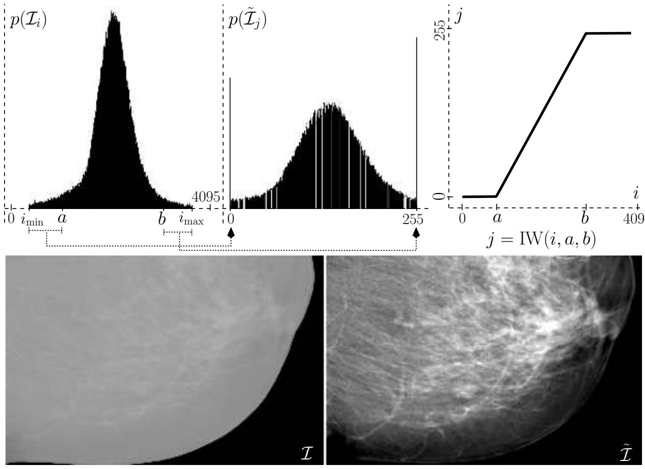

Figure 1. Top: histogram and functional representation of a basic IW operation. As an image histogram, intensities below and above the a and b thresholds are cumulatively transferred from the original image $\mathcal{I}$ to the 0 and 255 bins in the 8-bit windowed version $\tilde{\mathcal{I}}$, respectively. As an intensity transform function, an intensity value of j in $\tilde{\mathcal{I}}$ is the result of applying Eq. (1) to an intensity level i in $\mathcal{I}$. Bottom: an example of a raw mammographic image with low contrast and narrow dynamic range and next to it, the result of applying an appropriate IW operation.

The second most widely used contrast modification approach is Global Contrast Stretching (GCS). This technique [8] enhances the image from the luminance information of all of the pixels and is governed by the equation:

$$j = 255 \times \frac{i - i_{\min}}{i_{\max} - i_{\min}} \tag{2}$$

where $i_{\min}$ and $i_{\max}$ are the minimum and the maximum pixel intensities of the original image (which will very likely correspond to a number close to 0 and 4095, respectively, in a 12-bit mammogram). As with IW, in GCS, all pixel values are homogeneously and linearly altered given that a, b, $i_{\min}$ and $i_{\max}$ are global to the 12-bit source image. In Section V, we experimentally compare our proposed GRAIL method with GCS.

Local Range Modification (LRM) is a pseudo-linear variation of the aforementioned approach [9]. LRM estimates regional maximum and minimum values based on the interpolation of eight surrounding grid points.

From here, the rest of the available methodologies related to mammographic image enhancement (for which there are good reviews in the literature [10, 11]) involve deeper, more drastic, and non-linear histogram alterations [12]. We summarize the most relevant ones here.

In the weighted local differences approach [13], pixels related to micro-calcifications show a slightly higher luminance than that of their surrounding neighbors. This filter automatically gives more weight to rare combinations of gray levels, which often correspond to micro-calcifications.

Contrast Limited Adaptive Histogram Equalization (CLAHE) operates on small image tiles rather than on the entire image [14]. Each tile's contrast is enhanced independently so that the histogram of the output region approximately matches the histogram predefined by a distribution function. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. Also, in homogeneous areas, the contrast can be limited to avoid amplifying the noise that might be present in the image.

Unsharp Masking (UM) subtracts a low-pass filtered signal from the original image [15]. UM is used to improve the visual quality of images by emphasizing the high-frequency portions that contain fine details. UM also amplifies noise and over-enhances the steep edges. A variation of UM [16], Rational Unsharp Masking (RUM), uses a rational function operator to replace the high-pass filter. RUM is intended to enhance the details of images that contain low and medium sharpness without significantly amplifying noise or affecting the steep edges.

Adaptive Neighborhood [17] Contrast Enhancement (ANCE) first identifies a nearly homogeneous region surrounding each pixel being processed using a region-growing procedure. The visual contrast of the region is then computed by comparing the intensity of the region with the intensity of its surroundings. The region's contrast is selectively increased by modifying its intensity if some conditions related to the standard deviation of the region's background are met. This approach is applied sequentially at each pixel in order to enhance the contrast of all objects and features in the image.

Direct Image Contrast Enhancement (DICE) directly amplifies the vertical, horizontal, and diagonal sub-band components at different levels of the wavelet decomposition. It then reconstructs them to obtain the enhanced image [18].

Non-Linear Unsharp Masking NLUM [12] integrates a nonlinear filtering operation with the UM technique. The resulting image is then normalized and fused with the original image to get the final enhanced version. This version is usually sharper than its original.

HVS techniques have also been recently used for mammogram enhancement. HVS relies on the assumption that human observers pay more attention to details like structural information. For instance, Zhou *et al.* [19] have developed a HVS method based on the second derivative. A biologically-inspired algorithm [20] for micro-calcification cluster detection has been proposed by Lingurau *et al.* Other common operations applied to mammograms that use HVS characteristics [12] are gray tone function, addition, subtraction, scalar and bitmap multiplication.

Each of the aforementioned methods plays a specific role in applications related to image filtering and general image improvement. However, conventional contrast enhancement in radiology (including breast screening) still relies mostly in GCS, tissue type-based presets, and manual window adjustments.

III. METHODS AND MATERIALS

A. Traditional information theory-based image quality assessment

Some of the earliest methods applied to image quality determination are the Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) which have been widely used in the context of image coding due to their easy implementation. Another commonly used quality metrics related to information theory are entropy and MI.

Intuitively, we can consider that a displayed image has the *highest quality* when the maximum amount of relevant information from the source image is preserved. Thus, the concept of entropy [21] naturally emerges. Shannon borrowed this concept from physics as a measure of the amount of *uncertainty* of an information source. Entropy has been used in many image processing applications such as spatial registration [22]. In the case of digital images, entropy is computed as:

$$H(\mathcal{I}) = -\sum_{i=0}^{4095} p(\mathcal{I}_i) \cdot \log_2(p(\mathcal{I}_i)) \tag{3}$$

where $p(\mathcal{I}_i)$ is the probability distribution of the pixel intensities of image $\mathcal{I}$. Eq. (3) measures the sharpness and shape of the image histogram, which is indirectly related to the image texture. For instance, images with large homogeneous areas have a very peaked histogram (low entropy). On the other hand, images which rich textures have a lot of contrast and a flatter histogram, and thus a higher entropy. Although image entropy has been regularly used for bitmap quality quantification, it has also been proven to be easily fooled. For instance, the entropy of an image can be easily manipulated by simply adding white noise to it.

Mutual information [23] can also play a role in image quality assessment [24]. MI measures the reciprocal dependence between two variables. In other words, it quantifies the information obtained from the first variable through the second one. In mammography screening, MI has traditionally been used for image registration [25] and diagnosis through template matching [26, 27]. The MI value between a 12-bit source image and its 8-bit resampled version can be mathematically expressed as:

$$\mathrm{MI}(\mathcal{I},\tilde{\mathcal{I}}) = H(\mathcal{I}) + H(\tilde{\mathcal{I}}) - H(\mathcal{I},\tilde{\mathcal{I}}) \tag{4}$$

where $H(\mathcal{I})$ and $H(\tilde{\mathcal{I}})$ are the correspondent proper entropies of each image representation. The term $H(\mathcal{I},\tilde{\mathcal{I}})$ is defined as the joint entropy, which can be estimated with the 4096×256 bi-dimensional histogram of the intensities of $\mathcal{I}$ and $\tilde{\mathcal{I}}$:

$$H(\mathcal{I},\tilde{\mathcal{I}}) = -\sum_{i=0}^{4095}\sum_{j=0}^{255} p(\mathcal{I}_i \cap \tilde{\mathcal{I}}_j) \cdot \log_2(p(\mathcal{I}_i \cap \tilde{\mathcal{I}}_j)) \tag{5}$$

where $p(\mathcal{I}_i \cap \tilde{\mathcal{I}}_j) = p(\mathcal{I}_i) \cdot p(\tilde{\mathcal{I}}_j \mid \mathcal{I}_i)$ is the probability that corresponding pixels in $\mathcal{I}$ and $\tilde{\mathcal{I}}$ have intensities i and j, respectively. An easy and very common way of understanding the relation between these information entities is by a Venn diagram, which is shown in Fig. 2.

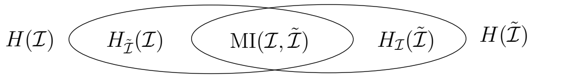
$H(\mathcal{I})$
$H_{\tilde{\mathcal{I}}}(\mathcal{I})$
$\mathrm{MI}(\mathcal{I},\tilde{\mathcal{I}})$
$H_{\mathcal{I}}(\tilde{\mathcal{I}})$
$H(\tilde{\mathcal{I}})$

Figure 2. Venn diagram showing the relation among each image entropy, conditional entropies $H_{\tilde{\mathcal{I}}}(\mathcal{I})$, $H_{\mathcal{I}}(\tilde{\mathcal{I}})$, and MI.

The diagram makes use of the term *conditional entropies* $H_{\tilde{\mathcal{I}}}(\mathcal{I})$ and $H_{\mathcal{I}}(\tilde{\mathcal{I}})$. The conditional entropies reflect the part of information in one image that cannot be explained when the other image is known. Following the previous example, if we add white noise to an image, its entropy grows because the conditional entropy is higher, but the MI between the original and the *corrupted* version remains the same.

Unfortunately, none of the aforementioned methods can measure by themselves the quality perceived by a human observer [28]. In contrast, techniques based on HVS have shown a better performance in image quality estimation [29, 30]. Even information-based quality assessment methods like entropy and MI assume that the image pixels are statistically independent (which is obviously a wrong principle) and they do not take into account how the visual cortex and human perception work.

In order to address this limitation, a new image quality metric that complements MI with a HVS approach is proposed in Section III B. This metric, based on Gabor filters, is significantly more faithful and consistent with the quality perceived by the human brain.

B. An image quality metric based on mutual information and Gabor filtering

Gabor filters are an excellent tool for texture analysis of images. In short, the responses of Gabor filters [31] correspond to those of single cells in the visual cortex. These cells extract contours and directional patterns. Gabor filters are commonly grouped in *banks* where each filter *captures* the image information in the vicinity of a frequency (f_m) and at a specific direction (θ_n). The output of each filter is then related to the contours of the image for a given f_m and θ_n. These filters have been extensively used in texture analysis and object classification [32] and have recently been proposed for image quality assessment [33]. They are also conquering a niche [34, 35] in computer-aided diagnosis (CAD).

Mathematically, a Gabor filter consists of a sinusoidal wave modulated by a Gaussian envelope. The impulse response of a complex Gabor filter for an image pixel x, y is defined as:

$$G_{f_m,\theta_n}(x,y) \;=\; e^{\left(\frac{-x'^2}{2\sigma^2}\right)} \cdot e^{(2\pi f_m x' \sqrt{-1})} \tag{6}$$

where $x' = x\sin\theta_n + y\cos\theta_n$, f_m is the m^{th} spatial frequency, and θ_n is its n^{th} orientation relative to the x-axis. In this work, we have initially used $N = 6$ different values for θ_n $(0, {}^{\pi}/_{6}, {}^{\pi}/_{3}, {}^{\pi}/_{2}, {}^{2\pi}/_{3}$ and ${}^{5\pi}/_{6})$ and $M = 3$ different values for f_m $({}^{1}/_{8}, {}^{\sqrt{2}}/_{8}$ and ${}^{1}/_{4})$, although other combinations of these parameters have been tested for the sake of completeness in Section V. The term $\sigma = {}^{1}/_{(2f_m)}$ represents the spatial deviation for each filter. We generate a total of 18 complex responses, some of which are shown in Fig. 3. Each image pixel $\mathcal{I}(x,y)$ is then linearly convolved to obtain a complex Gabor response ($\mathcal{R}_{f_m,\theta_n}$) with the expression:

$$\mathcal{R}_{f_m,\theta_n}(x,y) = \mathcal{I}(x,y) * G_{f_m,\theta_n}(x,y) \tag{7}$$

After an image has been filtered, the Gabor response for this same pixel is defined as:

$$R_{f_m,\theta_n}(x,y) = |\mathcal{R}_{f_m,\theta_n}(x,y)| \tag{8}$$

which corresponds to obtaining the amplitude of Eq. (7).

To assess the quality of a displayed image, we use the MI between the Gabor decompositions of $\mathcal{I}$ and $\tilde{\mathcal{I}}$. From here, if we let R_{f_m,θ_n} and $\tilde{R}_{f_m,\theta_n}$ be the Gabor responses of $\mathcal{I}$ and $\tilde{\mathcal{I}}$ obtained with Eq. (8), respectively, we can measure the perceived quality as:

$$\mathcal{Q}(\mathcal{I},\tilde{\mathcal{I}}) = \sum_{n=1}^{N}\sum_{m=1}^{M} \mathrm{MI}(R_{f_m,\theta_n},\, \tilde{R}_{f_m,\theta_n}) \tag{9}$$

The value of $\mathcal{Q}$ is upper bounded by the entropy of the input image. Moreover, the pixels are not assumed to

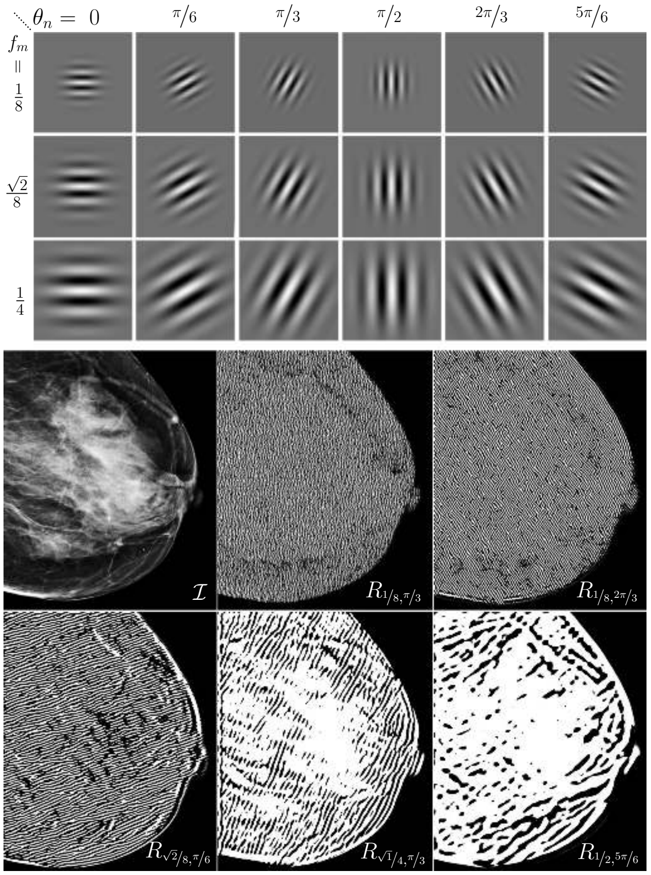

Figure 3. Top: real response of a sample Gabor filter bank $\mathrm{Re}(G_{f_m,\theta_n})$ generated with three frequencies ($M = 3$) and six orientations ($N = 6$). Bottom: Gabor features R_{f_m,θ_n} for several frequencies (f_m) and orientations (θ_n) applied to a mammogram, whose original instance $\mathcal{I}$ is also shown.

be independent (in contrast to conventional information-based methodologies, such as those reviewed in Section III A) because the statistical dependencies between pixels are taken into account by the Gabor filters. Eq. (9) is the proposed HVS function that will be maximized (as described in Section III C) to find the IW limits that assure the best quality when presenting 12-bit mammograms in 8-bit screening software and hardware.

C. Auto-adjustment of image intensity levels

As mentioned in the introduction, a common visualization problem in radiology is that the dynamic range of an image (i.e., a mammogram) is much wider than the dynamic range of current displays. A naive solution to this problem is to linearly map the minimum and maximum image intensity values to the 0-255 interval. This approach may lead to a low-contrast configuration or a deficient image visualization.

As a solution to this problem, practitioners usually set a manual visualization window that is defined by two 12-bit values a and b so that pixels whose intensity fits within these thresholds are linearly mapped to the correspondent 8-bit range. Values outside this range are clipped. The problem with this approach is that it requires human intervention and it is, therefore, very subjective and time-consuming. However, using the perceptual image quality metric related to Gabor filtering proposed in Section III B, it is possible to define an objective function $\mathcal{F}_{\mathcal{I}}(a, b)$ based on Eq. (9) that can be maximized via a and b. The optimization of $\mathcal{F}_{\mathcal{I}}(a, b)$ will in turn assure the highest MI possible between $\mathcal{I}$ and $\tilde{\mathcal{I}}$.

However, our preliminary experiments showed that this function has plenty of local maxima, and, for this reason, it is difficult to find the optimum range using a gradient-based approach. On the other hand, a thorough search for the best parameters is computationally very expensive because it depends quadratically on the image grayscale depth. Therefore, we have devised GRAIL, a hierarchical iterative algorithm (Fig. 4) that maximizes $\mathcal{F}_{\mathcal{I}}(a, b)$, optimizing the intensity threshold values a and b until convergence. For each loop iteration (tagged with the parameter k), we define a grid of low (A_k) and high search values (B_k). The spacing of this grid is defined by Δ, which starts with a predefined and relatively large custom size (300, 200, 100, etc.) and is reduced, for instance, by a tenth in each k iteration ($\Delta/10$). The range of search intensities for each k is determined by the previous values a_{k-1} and b_{k-1}, which at the beginning ($k = 0$) are set to $i_{\min}$ and $i_{\max}$, respectively. The algorithm stops when the intensities found in an iteration are equal to those in the previous one or when it reaches a predefined iteration limit (K). From the obtained a and b, we can easily derive $\mathrm{WL} = \frac{1}{2}(b - a)$ and $\mathrm{WW} = b - a$.

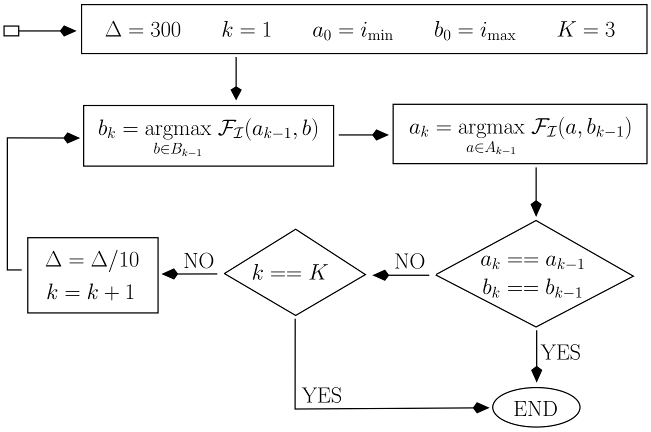
$\Delta = 300$ $k = 1$ $a_0 = i_{\min}$ $b_0 = i_{\max}$ $K = 3$
$b_k = \operatorname{argmax}_{b \in B_{k-1}} \mathcal{F}_{\mathcal{I}}(a_{k-1}, b)$
$a_k = \operatorname{argmax}_{a \in A_{k-1}} \mathcal{F}_{\mathcal{I}}(a, b_{k-1})$
$a_k == a_{k-1}$ $b_k == b_{k-1}$
NO
$k == K$
NO
$\Delta = \Delta/10$ $k = k + 1$
YES
YES
END

Figure 4. Operational diagram of GRAIL, the proposed algorithm that seeks the optimal a and b intensity levels that maximize the MI between a source 12-bit image ($\mathcal{I}$) and its 8-bit displayed version ($\tilde{\mathcal{I}}$). It does so by iteratively optimizing an objective function $\mathcal{F}_{\mathcal{I}}(a, b)$ based on Eq. (9).

D. Test databases used

Here we describe the mammogram databases used in our research. We were interested in collections that hosted fully-digital 12-bit images acquired with relative modern equipment. We were also concerned about incorporating and reflecting a broad spectrum of features, densities, associated health statuses, presence or not of foreign elements, and different qualities in order to account for as many clinical scenarios as possible. All these sets of images are publicly available for research purposes and include the typical examination views: mediolateral oblique, craniocaudal, mediolateral and lateromedial. We have compiled a comprehensive test set consisting of 159 images. This compilation is big and heterogeneous enough for our statistical analysis but its size still remains adequate for the screening sessions performed by each of the corresponding human observers (subject tackled in Section III E).

UPMC breast tomography collection from the University of Pittsburgh Medical Center (UPMC). All of the images contain hamartomas, subtle cancers, lobular carcinomas, cysts, papillomas, invasive ductal carcinomas, atherosclerotic calcifications, radial scars, vascular calcifications, benign ducts, oil cysts and fat necrosis.

Society of Breast Imaging collection or SBI. Its database contains diagnostic images used during workshops and annual meetings. They mainly include calcifications and surgical clips.

Cancer Genome Atlas collection or TCGA Research Network, part of the National Cancer Institute (part of the National Institute of Health). Images mainly contain invasive carcinoma.

Integrating the Healthcare Enterprise or IHE. It is an initiative developed by healthcare professionals to improve the way computer systems in hospitals and clinics share information. The image collection is part of their MESA software package, that was engineered at the Mallinckrodt Institute of Radiology together with the Healthcare Information and Management Systems Society (HIMSS) and the Radiological Society of North America (RSNA).

Cancer Imaging Archive which holds an important breast diagnosis collection [36]. This compilation contains cases with high-risk normals, ductal carcinoma *in situ*, fibroids, and lobular carcinomas.

Task Group 18 or TG18 from the AAPM. This task force evaluates the performance of electronic display devices [37]. For this purpose, they have efficiently gathered a set of high quality images, including not only geometrical and grayscale patterns but also anatomical ones, such as the two wide dynamic range mammograms shown in Fig. 5.

According to the Breast Imaging Reporting and Data System (BI-RADS) from the ACR [38] and its four categories of breast density, our collected image set comprises: 33 mammograms associated to almost *entirely fatty* breasts (BI-RADS A), 36 representing *scattered fibroglandular* densities (BI-RADS B), 55 *heterogeneously dense* cases (BI-RADS C), and 35 images identifiable as *extremely dense* breasts (BI-RADS D). Also, 29 instances contain foreign X-ray-opaque elements (i.e., surgical staples, fiducial markers, etc.) which reveal some sort of surgical intervention. Implants appear in 7 images. Around half of the mammograms were generated from a craniocaudal angle, the rest were obtained with lateral protocols such as mediolateral, oblique or lateromedial. Finally, 59 images contain high-density elements, including the aforementioned foreign items, makers, calcifications and other abnormalities.

E. Comparison with windowing settings determined by human observers

In order to test the suitability of the proposed windowing mechanism presented in Section III C, we gathered a focus group of ten radiologists. The selection of the members of this group has allowed us to leverage more than 120 years of cumulative knowledge and experience in mammography diagnosis. Each radiologist was asked to manually establish the IW of the same mammogram dataset presented in Section III D. An intuitive and simple tool developed in MATLAB presented each image, one after another. When each image was shown, the observer was allowed to manually and freely modify the lowest (a) and highest (b) intensity limits of the image histogram (as shown in Fig. 1). No further diagnosis steps or reporting phases were required.

During the examination, half of the mammograms (even images) were already intensity-windowed with the settings obtained by GRAIL. The other half (odd instances) were shown in their full range (FR), as would have normally been presented by ordinary screening software. This has also allowed us to measure the mean time spent on the initial IW operation and compare this time interval between non-contrast-stretched images and the pre-windowed ones. As stated in Section I, this step (usually performed through manual mouse gestures) is very common during radiological examinations and is therefore very interesting to measure how our technique can contribute to reducing mammogram screening time, especially in setups involving legacy acquisition equipment and/or 8-bit diagnostic workstations.

This verification experiment involving the participation of human observers extends our preliminary results [39]. Specifically, we have correlated the a and b values (for each image) determined by each member of the focus group with those derived by GRAIL. A repeated-measures ANOVA statistical analysis (with SPSS v16) has also been carried out with the differences between the b values derived by each method (GRAIL, GCS and the full range) and those proposed by each radiologist. Our null hypothesis (H_0) is that the mean value is the same for these differences ($\mu_{\text{GRAIL}} = \mu_{\text{GCS}} = \mu_{\text{FR}}$). We initially perform a multivariate contrast test and then, after rejecting H_0, we carry out a Bonferroni adjustment [40] for multiple comparisons. The reason for choosing the maximum intensity thresholds (b) for this analysis is that they may contribute more to revealing the presence of injuries, calcifications and surgical elements.

IV. RESULTS

In this section, we summarize the results obtained after comparing the values reported by GRAIL with the measurements performed by the panel of radiologists presented in Section III E.

Before addressing this comparison and as an example, Fig. 5 shows a specific application of the optimization process described in this text. It consists on two mammograms (part of the anatomical patterns set from the TG18 group introduced in Section III D) whose levels were automatically windowed with GRAIL. The unanimous opinion from the aforementioned focus group is that these specific stretching examples (as well as the majority applied to the rest of the images in the mammogram datasets used) proved to show a very good contrast, sharpness, and brightness.

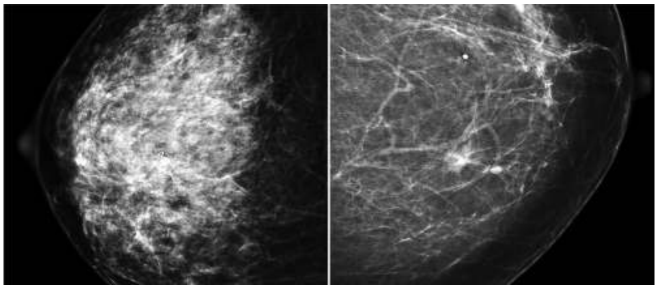

Figure 5. Mammograms TG18MM2 and TG18MM1. Both have been windowed with a and b values derived with GRAIL. They are normally displayed with a starting WL/WW of 2048/4096 (FR), but their intensity range can still be automatically focused to an optimal one thanks to GRAIL.

Table I shows the correlation of the IW parameters manually set by the human observers with those obtained with the proposed methodology (as well as GCS and the full range). This table also shows the mean time spent by each expert to adjust the intensity limits for images that were already pre-windowed with GRAIL and those that were presented with their full intensity range.

The multivariate contrasts test reveals that $\mu_{\text{GRAIL}} \neq \mu_{\text{GCS}} \neq \mu_{\text{FR}}$, with $F(2, 1549) = 1920.584$ and $P > 0.10$ and H_0 is then rejected. In order to further verify that the means are different, we apply a Greenhouse-Geisser analysis [41]. From this test, we obtain $F(1.676, 2598.166) = 2457.491$ and $P < 0.001$, also rejecting the Mauchly's sphericity hypothesis [42]. In the Bonferroni pairwise comparison, we derive that: $\mu_{\text{GRAIL}} - \mu_{\text{GCS}} = 601 \pm 11$ and $\mu_{\text{GRAIL}} - \mu_{\text{FR}} = 629 \pm 10$ (with $P < 0.001$ for both). Specifically, $\mu_{\text{GRAIL}} = -72 \pm 7$, $\mu_{\text{GCS}} = -673 \pm 9$ and $\mu_{\text{FR}} = -701 \pm 8$.

Rad.	Correlation						Time (s)	
	GRAIL		GCS		FR		w/	w/o
	a	b	a	b	a	b		
1	0.91	0.83	0.73	0.41	0.13	0.23	15	18
2	0.91	0.77	0.79	0.57	0.01	0.29	9	13
3	0.91	0.72	0.74	0.51	0.00	0.12	8	14
4	0.92	0.78	0.75	0.47	0.01	0.26	10	13
5	0.90	0.72	0.74	0.55	0.00	0.30	16	16
6	0.87	0.64	0.81	0.49	0.01	0.26	9	9
7	0.49	0.62	0.42	0.50	0.12	0.32	14	18
8	0.81	0.70	0.84	0.52	0.07	0.27	18	21
9	0.86	0.70	0.78	0.43	0.07	0.20	7	17
10	0.50	0.71	0.34	0.41	0.12	0.19	9	11

Table I. Correlation coefficients between the manually (by each radiologist) established IW parameters (min a and max b intensities) and the estimated ones with GRAIL, GCS and the full range (FR). The table also shows the average time taken (in seconds) by each human observer in adjusting the appropriate WL/WW when the image was pre-stretched with GRAIL (w/) and when not (w/o).

Fig. 6-top shows the increase in the mutual information obtained with GRAIL relative to that derived with the application of GCS. We have divided the set of images into two blocks: those with minor or no presence of microcalcifications or external Roentgen radiation-opaque elements, and those with the presence of such high density elements. Fig. 6-bottom shows a similar information but this time for the difference in the max intensity level (b) derived by both methods.

V. DISCUSSION

Proper contrast stretching is a key step during breast screening. Although recently generated mammograms have been produced with modern equipment and already display relatively good intensity histograms, radiologists and physicians usually need or prefer to further modify/customize windowing settings. In some other cases, mammographic images need profound IW adjustments in order to be of any value to the diagnosis, as the example shown in Fig. 1. During the first moments of examination and before zooming in and out on regions of interest, radiologists take valuable seconds seeking the min and max intensity values that maximize the initial visual and radiological information.

In this paper, we have presented an innovative method

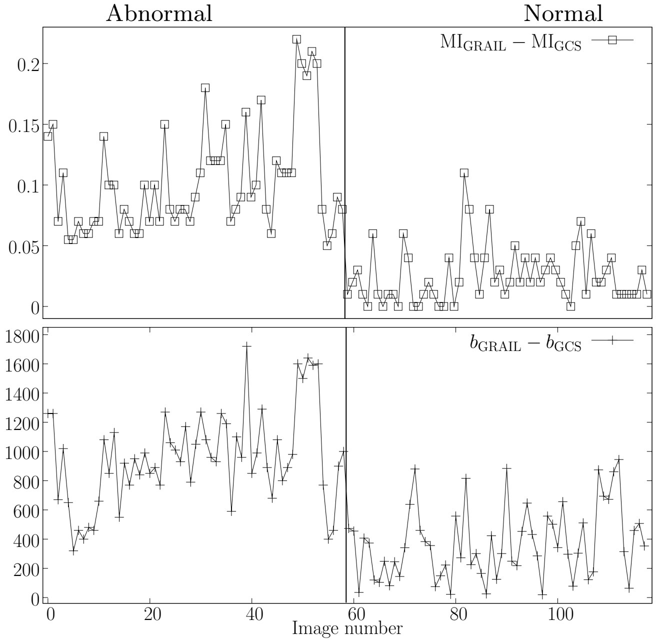

Figure 6. Top: mutual information difference between the windowed representations obtained with GRAIL and GCS for 59 instances with abnormalities (images with calcifications and/or external X-ray opaque elements) and 59 *clean*/normal ones. Bottom: difference in the b intensity threshold determined by each method and for the same set of images.

Different number of frequencies (M) and orientations (N)							
$\Delta = 300$ $K = 3$ s = 1	M	3		5		7	
	N	4	8	4	8	4	8
	c	0.85 0.74	0.89 0.81	0.87 0.74	0.90 0.74	0.85 0.73	0.86 0.72
	t	8	16	14	21	22	31

Different image downscale factors (s) in %							
$\Delta = 300$ $N = 6$ $M = 3$ $K = 3$	s	90	75	50	33	25	20
	c	0.89 0.81	0.91 0.76	0.89 0.72	0.89 0.67	0.88 0.63	0.89 0.62
	t	12	11	8	7	6	5

Different search grids (Δ) and max iterations (K)							
$M = 3$ $N = 6$ s = 1	Δ	200		300		400	
	K	1	3	1	2	1	3
	c	0.88 0.74	0.93 0.77	0.90 0.79	0.90 0.79	0.93 0.69	0.86 0.77
	t	12	15	8	10	7	9

Table II. Results of the tests with other GRAIL configurations. The c coefficients express the correlation (for the a and b thresholds, separated by a new line) of the IW values derived by GRAIL (for each test image) with the mean of the IW settings stablished by the panel of radiologists. Time (t) is expressed in seconds and accounts for the mean algorithm execution time when windowing each of the 159 mammograms.

(thoroughly reviewed in Section III C) to assess the intensity window that maximizes the visual information when displaying a mammogram. The proposed technique is in turn based on Gabor filters and the human visual system. The mean computing time taken by GRAIL is 12 seconds for 1 Mpx images on a 2,8 GHz Intel Core i5 computer. In a more realistic environment, a GRAIL implementation could run as a parallel background task that previously seeks the proper IW for each image before being displayed for screening. This windowing information could either be calculated and applied to the image in real time (i.e., based on the daily radiological work-list) or be stored in a separate database or in the same DICOM SOP Instance (as explained in Section I). If the appropriate intensity thresholds are calculated in advance and are ready to be applied when each mammogram is loaded, a radiologist can save valuable seconds of screening time.

As stated in Section III B, in this work Gabor filters were configured with a specific set of frequencies and orientations, but we have also experimented with other arrangements, shown in Table II. For the sake of completeness, we have run tests with downsized versions of high-resolution images, fewer iterations, and larger search grids, obtaining similar outcomes. This fact may enable the optimization phase (shown in Fig. 4) to run significantly faster (with very little effect in the performance) if less computational demanding configurations are carefully chosen.

In general, images with outlying calcifications or foreign radio-opaque elements have proven to be the ones for which our approach (compared against GCS) delivers the best results, calculated with Eq. (2). Fig. 6-top shows, for instance, how the MI between 12-bit and 8-bit images ($\mathrm{MI}_{\mathrm{GRAIL}} - \mathrm{MI}_{\mathrm{GCS}}$) increases by 0.05 bits in this type of scenario (relative to the MI obtained with GCS). The difference in the b intensity threshold (which, as stated in Section III E, better accounts for the presence of high density elements) derived by GRAIL and GCS (Fig. 6-bottom) is also more significant in the case of abnormal images. Specifically, the presented GRAIL technique works better when high density corpuscles are present in the image. One of the main goals of mammogram screening is the diagnosis of the existence of these calcifications and other types of high-density corpuscles and abnormalities. In this context, GRAIL represents an important contribution at the beginning of the examination process and clearly surpasses in effectiveness other more conventional methods like GCS. Regarding mammograms linked to healthy breasts, GRAIL performs slightly better than GCS for BI-RADS A+B than with BI-RADS C+D images (we obtain a $\mathrm{MI}_{\mathrm{GRAIL}} - \mathrm{MI}_{\mathrm{GCS}}$ of 0.03 bits). BI-RADS A+B mammograms usually have associated more peaked histograms centered in the range $i = 1000 \leftrightarrow 2400$ and generally have lower contrast, which GRAIL is able to better enhance over the application of GCS.

Finally, the focus group of radiologists unanimously reported an *out-of-the-box*, superior, and balanced visibility of microcalcifications, breasts margins, and Cooper's ligaments (as well as other criteria defined in the AAPM TG18 [43] report) in the images that had been automatically windowed with GRAIL. As shown in Table I, we obtain good correlations between the intensity limits derived by GRAIL and those decided by the members of the aforementioned panel of experts. These correlations are superior to using GCS. The ANOVA analysis shows that there is a significative difference between our method and GCS, as well as between GRAIL and the application of the full range, while the discrepancy between radiologists and each IW method (GRAIL, GCS and FR) is not statistically relevant.

A complete MATLAB and GNU/Octave-compatible implementation (including helper subroutines [44]), and the results of the panel of radiologists are available at `https://github.com/TheAnswerIsFortyTwo/GRAIL`.

VI. CONCLUSIONS

Mammographic images usually have wide 12-bit dynamic ranges associated to them that have to be adapted and rescaled for 8-bit displays. During the first moments of screening, radiologists spend a non-negligible amount of time performing ordinary contrast stretching adjustments to maximize the presence of all sorts of elements, from normal soft tissue areas to calcifications and other abnormal high-density corpuscles. The main novelty of this research work is the introduction of a new methodology (GRAIL) to automatically adjust the limiting intensity levels of mammograms (i.e., their window level and width). Our method, which is based on the human visual system and Gabor filtering, establishes an objective image quality metric and, based on this quantification, GRAIL is able to automatically seek the best intensity stretching values by maximizing the mutual information between the original bitmap sources and their displayed version. GRAIL surpasses the versatility of other common windowing techniques like global contrast stretching, as has been demonstrated in our experiments involving the collaboration with a panel of radiologists and the results obtained thereof. The proposed approach can be added to the diagnosis workflow, allowing radiologists to start screening sessions with better initial window adjustments and then keep on modifying them as needed.

ACKNOWLEDGMENTS

This work has the support of IST S.L., University of Valencia (CPI15170), Consolider (CPAN13TR01), MINETUR (TSI1001012013019) and IFIC (Severo Ochoa Centre of Excellence SEV20140398). The authors would also like to thank C. Bellot M.D., M. Brouzet M.D., C. Calabuig M.D., J. Camps M.D., J. Coloma M.D., D. Erades M.D., Mr. ˜ V. Gutierrez, J. Herrero M.D., Dr. I. Maestre, Dr. A. Ñeco M.D., C. Ortola M.D., A. Rubio M.D., Dr. R. Sánchez, Dr. F. Sellers, A. Segura M.D., and the Spanish Cancer Association (AECC) for their effort, participation, counseling and commitment in this research study. The authors report no conflicts of interest in conducting the research.

[1] A. D. Maidment, R. Fahrig, and M. J. Yaffe, "Dynamic range requirements in digital mammography," *Medical Physics*, vol. 20, no. 6, pp. 1621–1633, 1993.

[2] T. Kimpe and T. Tuytschaever, "Increasing the number of gray shades in medical display systems - How much is enough?" *Digital Imaging*, vol. 20, p. 422, 2007.

[3] ACR, AAPM, and SIIM, "Practice parameter for determinants of image quality in digital mammography," 2014.

[4] D. S. Committee, "PS3.3 Information Object Definitions," NEMA, Tech. Rep., 2015.

[5] E. D. Pisano, J. Chandramouli, B. M. Hemminger, D. Glueck, R. E. Johnston, K. Muller, M. P. Braeuning, D. Puff, W. Garrett, and S. Pizer, "The effect of intensity windowing on the detection of simulated masses embedded in dense portions of digitized mammograms in a laboratory setting," *Journal of Digital Imaging*, vol. 10, no. 4, pp. 174–182, 1997.

[6] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[7] S. Börjesson, M. Håkansson, M. Båth, S. Kheddache, S. Svensson, A. Tingberg, A. Grahn, M. Ruschin, B. Hemdal, S. Mattsson *et al.*, "A software tool for increased efficiency in observer performance studies in radiology," *Radiation protection dosimetry*, vol. 114, no. 1-3, pp. 45–52, 2005.

[8] S. I. Sahidan, M. Y. Mashor, A. S. W. Wahab, Z. Salleh, and H. Ja'afar, *Local and Global Contrast Stretching For Color Contrast Enhancement on Ziehl-Neelsen Tissue Section Slide Images*. Springer Berlin Heidelberg, 2008.

[9] B. Senthilkumar and G. Umamaheswari, "Combination of novel enhancement technique and fuzzy C means clustering technique in breast cancer detection," *Biomedical Research (India)*, vol. 24, no. 2, pp. 252–256, 2013.

[10] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K.-H. Ng, "Computer-aided breast cancer detection using mammograms: a review," *IEEE Reviews in Biomedical Engineering*, vol. 6, pp. 77–98, 2013.

[11] A. Papadopoulos, D. I. Fotiadis, and L. Costaridou, "Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques," *Computers in Biology and Medicine*, vol. 38, no. 10, pp. 1045–1055, 2008.

[12] K. Panetta, Y. Zhou, S. Agaian, and H. Jia, "Nonlinear unsharp masking for mammogram enhancement," *IEEE Transactions on Information Technology in Biomedicine*,

vol. 15, no. 6, pp. 918–928, 2011.

[13] Y. Kabbadj and M. M. Himmi, "Detection of Microcalcification in Digitized Mammograms Using Weighted Local Differences and Local Contrast," vol. 6, no. 131, pp. 6533–6544, 2012.

[14] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[15] J. Rogowska, K. Preston, and D. Sashin, "Evaluation of digital unsharp masking and local contrast stretching as applied to chest radiographs." *IEEE transactions on biomedical engineering*, vol. 35, no. 10, pp. 817–27, 1988.

[16] G. Ramponi and A. Polesel, "Rational unsharp masking technique," *Journal of Electronic Imaging*, vol. 7, no. 2, pp. 333–338, 1998.

[17] R. M. Rangayyan, L. Shen, Y. Shen, J. E. L. Desautels, H. Bryant, T. J. Terry, N. Horeczko, and M. S. Rose, "Improvement of sensitivity of breast cancer diagnosis with adaptive neighborhood contrast enhancement of mammograms," *IEEE Transactions on Information Technology in Biomedicine*, vol. 1, no. 3, pp. 161–170, 1997.

[18] J. Tang, X. Liu, and Q. Sun, "A direct image contrast enhancement algorithm in the wavelet domain for screening mammograms," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 74–80, 2009.

[19] Y. Zhou, K. Panetta, and S. Agaian, "Human visual system based mammogram enhancement and analysis," *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, no. 3, pp. 229–234, 2010.

[20] M. G. Linguraru, K. Marias, R. English, and M. Brady, "A biologically inspired algorithm for microcalcification cluster detection," *Medical Image Analysis*, vol. 10, no. 6, pp. 850–862, 2006.

[21] E. N. Kirsanova and M. G. Sadovsky, "Entropy approach in the analysis of anisotropy of digital images," *Open Systems & Information Dynamics*, 2002.

[22] S. K. S. Fan and Y. C. Chuang, "Entropy-based image registration method using image intensity difference on overlapped region," *Machine Vision Applications*, 2011.

[23] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *Digital Imaging*, 2007.

[24] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[25] M. A. Wirth, J. Narhan, and D. W. Gray, "Nonrigid mammogram registration using mutual information," in *Medical Imaging 2002*. International Society for Optics and Photonics, 2002, pp. 562–573.

[26] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious Jr, and C. E. Floyd Jr, "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," *Medical Physics*, vol. 30, no. 8, pp. 2123–2130, 2003.

[27] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Medical Physics*, vol. 34, no. 1, pp. 140–150, 2007.

[28] T. Samajdar and M. I. Quraishi, *Information Systems Design and Intelligent Applications: 2nd International Conference INDIA 2015*. Springer India, 2015, ch. Analysis and Evaluation of Image Quality Metrics.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[30] L. K. Choi, T. Goodall, and A. C. Bovik, "Perceptual Image Enhancement Keywords," pp. 1–37.

[31] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biol. cybernetics*, vol. 61, p. 103, 1989.

[32] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using Gabor filters," *Pattern Recognition*, vol. 30, no. 2, pp. 295–309, 1997.

[33] E. Vazquez-Fernandez, A. Dacal-Nieto, F. Martin, and S. Torres-Guijarro, "Image analysis and recognition," in *ICIAR International Conference*, 2010.

[34] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 312–348, 2007.

[35] S. Jaeger, A. Karargyris, S. Antani, and G. Thoma, "Detecting tuberculosis in radiographs using combined lung masks," in *Conference on Engineering in Medicine and Biology Society*, 2012.

[36] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The Cancer Imaging Archive: maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[37] Task Group 18 Imaging Informatics Subcommittee, "Report no. 03: Assessment of display performance for medical imaging systems," AAPM, Tech. Rep., 2005.

[38] A. C. of Radiology Committee, "Bi-rads atlas 5th edition," ACR, Tech. Rep., 2014.

[39] A. Albiol, A. Corbi, and F. Albiol, "Measuring X-ray image quality using a perceptual metric," in *Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, 2016.

[40] Y. Hochberg and Y. Benjamini, "More powerful procedures for multiple significance testing," *Statistics in medicine*, vol. 9, no. 7, pp. 811–818, 1990.

[41] H. Keselman and J. C. Keselman, "The analysis of repeated measures designs in medical research," *Statistics in medicine*, vol. 3, no. 2, pp. 185–195, 1984.

[42] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 204–209, 1940.

[43] E. Samei, A. Badano, D. Chakraborty, K. Compton, C. Cornelius, K. Corrigan, M. J. Flynn, B. Hemminger, N. Hangiandreou, J. Johnson *et al.*, "Assessment of display performance for medical imaging systems: executive summary of AAPM TG18 report," *Medical Physics*, 2005.

[44] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, "CloudID: trustworthy cloud-based and cross-enterprise biometric identification," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7905–7916, 2015.