AUTOMATIC SCORING OF CDMAM USING A MODEL OF THE RECOGNITION THRESHOLD OF THE HUMAN VISUAL SYSTEM: R*

Gabriel Prieto, Margarita Chevalier, Eduardo Guibelalde

Dept. Radiología. Fac. Medicina. Universidad Complutense de Madrid.
28040 Madrid (Spain)

ABSTRACT

A software tool is presented for the automatic evaluation of the CDMAM phantom images that are currently used for the quality assessment of the image quality in mammography. This software tool is based on the use of the cross-correlation component of the index MS-SSIM*, R*, oriented to the *recognition threshold* of different image representations. Given an image sequence, whose images begin as unrecognizable and are gradually refined to include more information, the recognition threshold corresponds to first the image in the sequence in which an observer indentifies the content.

We have validated our software tool by comparing our readouts with those obtained by a total of 4 expert observers in the evaluation of 8 CDMAM images. The correlations obtained between both readout sets are better than 0.99 and the range of useful correlation comprises diameters from 0.16 mm to 2.0 mm of the gold disks inside the CDMAM.

Index Terms— Human visual system modeling, digital mammography, CDMAM phantom, quality control, MS-SSIM* index.

1. INTRODUCTION

The European Guidelines for the quality control of mammography [1] include minimum standards for the image quality of digital mammography systems. These standards are based on contrast-detail measurements using the CDMAM phantom images (version 3.4, Nijmegen) [2]. The observer task consists of determining the detail with minimum thickness (recognition threshold) which is able to detect for a series of gold disks, with different diameters and thicknesses, inserted into the phantom.

Scoring CDMAM images by observers has two major disadvantages. It is a consuming time task and the results are highly dependent from the observer. The most spread solution to solve these problems is the use of the CDCOM program [3], which automatically scores the CDMAM images. The results derived from the use of this software have demonstrated a bad correlation with those from human observers. Several approaches has been proposed to overcome this limitation [4] [5]. The resulting readouts could be fitted to psychometric curves and smoothing methods (some of them very complex) to adjust the CDCOM readouts to the human ones (here named processed CDCOM (PRCDCOM)). However, these methods cannot be applied to disks with diameters larger than 0.80 mm because of the lack of correlation between their results and those from the human observer [4].

To detect the disks in the CDMAM phantom image, CDCOM uses a perceptual Non Prewhitening Matched Filter model (NPWMF) [6] [7] which has been used in other programs [8]. In this perceptual model the human observer cannot prewhiten the correlated noise of the image.

In the perceptual visual theory proposed by Wang et al. [9] the Human Visual System (HVS) is considered to be highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity should be a good approximation of perceived image quality. A family of quality assessment algorithms has been developed based on this premise. They evaluate visual quality according to the resemblance of a test image's structural information to that of the reference image. This family includes Structural SIMilarity index (SSIM) [10], Multi-Scale Structural SIMilarity index (MS-SSIM) [11] and a modified version of MS-SSIM, MS-SSIM* [12]. SSIM and MS-SSIM measure similarity between two images. Both algorithms are oriented for the supra-threshold problem which invalidates them to be applied to the analysis of images with details in the limit of visibility. The cross-correlation component of MS-SSIM* (R*) avoids this problem since it is designed according to the recognition threshold problem.

CDMAM readout can be interpreted as a comparison among the phantom image and a reference image formed by a synthetic black background and white disks of different diameters. Our approach to perform this comparison is to use the R* index as a human perceptual model.

2. MATERIALS AND METHODS

The CDMAM phantom (Fig. 1) consists of an aluminum base with gold disks of different diameters and thicknesses. The aluminum base is attached to a polymethyl-methacrylate cover. The gold disks are arranged in a matrix of 16 rows by 16 columns, forming 205 cells. Within a row, the disk diameters are constant, with logarithmic increasing thickness. Within each column, the thickness of the disks is constant, but there is a logarithmic increasing of the diameter. Each cell contains two identical disks, one in the center and one in a randomly selected corner. The observer's task is to detect the position of the eccentric disks.

The CDMAM phantom images used in this study have been downloaded from the European Reference Organisation for Quality Assured Breast Screening and Diagnostic Services (EUREF) site, section CDMAM Readout [13]. This set of 8 raw images was acquired with a Senograph 2000D system (General Electric) at 27 kVp, 125 mAs and with a detector resolution of 1 pixel per 100μm. The images were evaluated by four human expert observers and

their scores are available at the same website. In a second step we use the CDCOM and PRCDCOM programs for scoring the images.

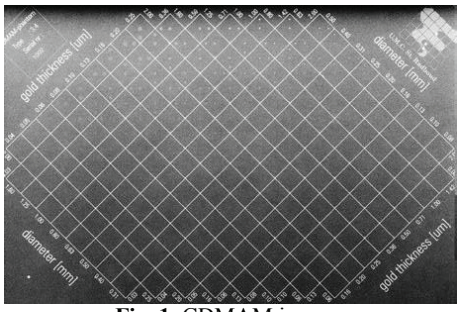

Fig. 1. CDMAM image.

2.1. Algorithm design

2.1.1. Determining the grid position.

In order to manage the disk information from the phantom images, it is necessary to accurately detect the position of the phantom grid. Several methods [7][8] have been applied to find this position. The method used in this work has been described in a previous work [14].

Using the known disk diameter and the position of the phantom grid, we can construct disk templates that fit the actual position of the disks at each of the four locations in a cell and apply the R* index.

2.1.2. Evaluation of the most likely position. MS-SSIM cross-correlation index: R*.*

SSIM specification [10]

Let $\mathbf{x} = \{x_i \mid i = 1, 2,\ldots,N\}$ and $\mathbf{y} = \{y_i \mid i = 1, 2,\ldots,N\}$ be two discrete non-negative signals that have been aligned, and let μ_x, σ_x^2 *and* σ_{xy} be the mean of **x**, the variance of **x**, and the covariance (cross-correlation) of **x** and **y**, respectively. σ_{xy} measures the tendency of **x** and **y** to vary together. Thus, this parameter is an indication of structural similarity. The mean, variance and cross-correlation comparison measures are given as follows:

$$m(\mathbf{x},\mathbf{y}) = (2\,\mu_x\,\mu_y + C_1)/(\mu^2_x + \mu^2_y + C_1) \qquad \text{(Eq. 1)}$$

$$v(\mathbf{x},\mathbf{y}) = (2\,\sigma_x\,\sigma_y + C_2)/(\sigma^2_x + \sigma^2_y + C_2) \qquad \text{(Eq. 2)}$$

$$r(\mathbf{x},\mathbf{y}) = (\sigma_{xy} + C_3)/(\sigma_x\,\sigma_y + C_3) \qquad \text{(Eq. 3)}$$

where *C*1, C2 and *C*3 are constants to avoid instability when $(\mu^2_x + \mu^2_y)$ or $(\sigma^2_x + \sigma^2_y)$ are very close to zero. The general form of the SSIM index between **x** and **y** is defined as:

$$SSIM(x, y) = [m(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot [v(\mathbf{x}, \mathbf{y})]^{\beta} \cdot r(\mathbf{x}, \mathbf{y})]^{\gamma} \qquad \text{(Eq. 4)}$$

where α, β and γ are parameters to define the relative importance of the components. Specifically in [10], α=β=1.

MS-SSIM specification [11].

The perception of image details depends on, amongst other things, the sampling density of the image and the distance from the image to the observer. A single scale method, as SSIM, may be appropriate only for specific settings. Wang et al. [11] proposed a Multi-Scale SSIM method for image quality. Taking the distorted and reference image as the input of the system, the algorithm iteratively applies a low-pass filter and downsamples the filtered image by a factor of 2 M-times. The overall evaluation of the Multi-Scale SSIM (MS-SIMM) is obtained by combining the measurement at different scales using:

$$MS\text{-}SSIM=[m_M(\mathbf{x},\mathbf{y})]^{\alpha M} \cdot \prod_{j=1}^{M} [v_j(\mathbf{x},\mathbf{y})]^{\beta j} \cdot [r_j(\mathbf{x},\mathbf{y})]^{\gamma j} \qquad \text{(Eq. 5)}$$

for $m(\mathbf{x},\mathbf{y})$, $v(\mathbf{x},\mathbf{y})$ and $r(\mathbf{x},\mathbf{y})$ as defined in Eq. (1, 2, 3) respectively. The exponents α, β, and γ are used to adjust the relative importance of different components at different scales.

The constants C_i in Eq. (1, 2, 3) were introduced to avoid instability when either $(\mu^2_x + \mu^2_y)$ or $(\sigma^2_x + \sigma^2_y)$ are very close to zero. This behavior occurs at the recognition threshold level (take into account that the CDMAM scoring problem is critical around the recognition threshold level). An alternative version of MS-SSIM, named MS-SSIM*, was proposed by Rouse et Hemami [12], where the positive constants C_i have been set to zero. Besides, they found that the cross-correlation component of MS-SSIM*, R*, shows a strong correlation with the recognition thresholds.

To calculate R*, suppose that $\sigma_x > 0$, and the patch **y** is constant. Then, the variance of **y** is zero. Under this scenario, y does not correlate with **x**, so the structure component must be set to zero. When both patches have equal variance and C3 = 0, the structure component must be set to one. The alternative structure component is given as

$$r^*(\mathbf{x},\mathbf{y}) = \begin{cases} 0 & \sigma_x > \sigma_y = 0 \; or \; \sigma_y > \sigma_x = 0 \\ 1 & \sigma_x = \sigma_y = 0 \\ r(\mathbf{x},\mathbf{y}) & else \end{cases}$$

for $r(\mathbf{x}, \mathbf{y})$ as defined in Eq. (3) with C3 = 0. **x** and **y** are patches of the complete images **X**, **Y**, so the overall evaluation of R* is obtained by combining the measurement at different scales using:

$$R^* = \prod_{j=1}^{M} r_j^*(\mathbf{X},\mathbf{Y})$$

In this work, $M = \log_2$ (maximum disk diameter in pixels + 1). The maximum disk diameter is 20 pixels for the CDMAM images used at this work. Therefore M has a maximum value of 5.

What we are measuring with R* is structure. We decided to compare structures at their maximum difference. The test image we use for each gold disk is a white circle with a diameter equal to the disk diameter (in pixels) that we are evaluating. Around this circle, we have created a black border with a margin equal to the resolution in pixels of the CDMAM for 300 μm (3 pixels in our test images). (Figure 2)

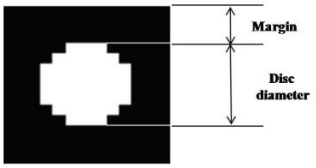

Fig. 2. Reference image.

Due to the CDMAM manufacturing process, the cells of the matrix are not "true" squares but trapezoids. Moreover we have found that disk images are located out of their expected centres (along the actual diagonals) up to 4 or 5 pixels. It is necessary to

fix a safe region around the theoretical position of the eccentric discs to find them. This safe region is +/- 5 pixels for the phantom images used in this work.

We estimate R* index inside each cell corner for 36 different positions around the nominal position of the expected disk (Figure 3). The maximum value of R* for each position inside each corner was chosen as the R* value of this corner. Finally, the corner selected was that with the maximum value of R* inside each cell. Following this procedure we get a matrix with the detected disks, that is compared with the true disk positions.

Finally, we compare R* results with those from human, CDCOM and PRCDCOM by correlating the threshold thickness for each disk diameter.

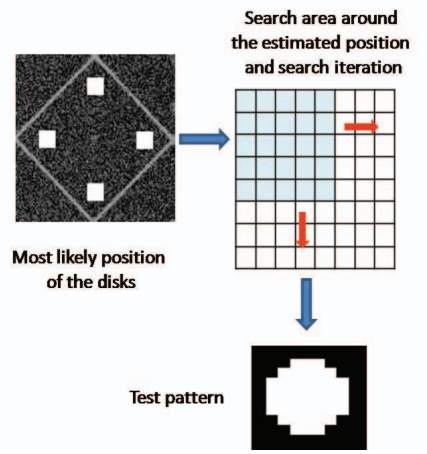

Fig. 3. Search method.

3. RESULTS AND DISCUSSION

Figure 4 shows the threshold thickness/diameter curve for human observer readouts (HO) and R* readouts, with error bars of 1 sem. R* is not significantly different from human readouts by taking into account the sem error. The average threshold thicknesses for the different disk diameters together with the deviations of R* from human readouts are summarized in Table 1. These deviations are apparently at random which can be due to the stochastic behaviour associated with the perception process.

Figure 5 shows the threshold thickness/diameter curve for all the considered methods. Although all methods exhibit a good correlation with the HO readouts (Pearson`s coefficient of 0.99), the behavior is completely different. CDCOM systematically is over-detecting the discs and PRCDCOM systematically is under detecting the disks. R* threshold thickness oscillate around the HO curve although it seems to be a slight trend to under detect the disks with diameters smaller than 0.20 mm. For this range of diameters the deviations of R* index from human readouts are larger than the PRCDCOM results.

We need to analyze the cause of the higher deviations of R* for these small diameters. We have run this test with another two sets of images with several results. The first set comprises images with a resolution of 75μ/pixel. Deviations of R* from human observer were 9% and the fitting was excellent, even for small diameters. The second set comprises images with a resolution of 50μ /pixel. Deviations of R* from human observer were 16%, but the fitting for small diameters was almost equal to those showed in this study.

There are several parameters, such as disc margin, low-pass filter values, evaluation window (**x, y**), etc. that can be adjusted and likely might improve the R* performance for small diameters. This adjustment will constitute our next lines of research.

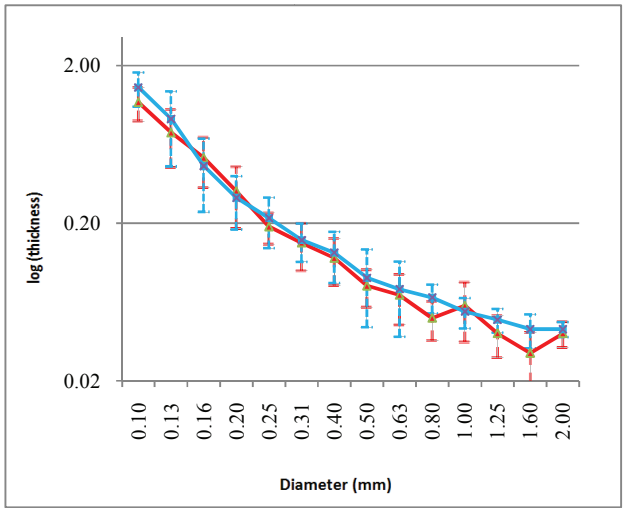

Fig. 4. 1 standard-deviation error bars for human observer readouts (HO) and R* readouts.

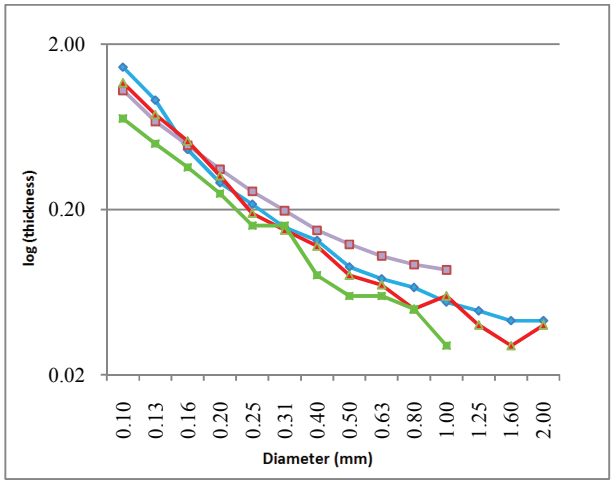

Fig. 5. Contrast-detail curve for CDMAM images.
Pearson's coefficient HO/R*=0.99
Pearson's coefficient HO/CDCOM=0.99
Pearson's coefficient HO/PRCDCOM =0.99

Table 1. Average values of threshold thickness (recognition threshold) from: human observer (HO), R* index, CDCOM and PRCDCOM readouts. The deviations (%) of the automated methods from human readouts are also included.

Ø mm	HO	R* Index	Deviation(%) R*/HO	CDCOM	Deviation(%) CDCOM/HO	PRCDCOM	Deviation(%) PRCDCOM/HO
0.10	1.17	1.45	24	0.71	39	1.05	10
0.13	0.75	0.92	23	0.50	33	0.68	9
0.16	0.52	0.46	12	0.36	31	0.49	6
0.20	0.32	0.29	9	0.25	22	0.35	9
0.25	0.19	0.22	16	0.16	16	0.26	37
0.31	0.15	0.16	7	0.16	7	0.20	33
0.40	0.12	0.13	8	0.08	33	0.15	25
0.50	0.08	0.09	13	0.06	25	0.12	50
0.63	0.07	0.08	14	0.06	14	0.10	57
0.80	0.05	0.07	40	0.05	0	0.09	80
1.00	0.06	0.06	0	0.03	50	0.09	50
1.25	0.04	0.05	25	NA	NA	NA	NA
1.60	0.03	0.04	33	NA	NA	NA	NA
2.00	0.04	0.04	0	NA	NA	NA	NA
Average			***16***		***25***		***33***

4. CONCLUSIONS

The application of R* to the evaluation of radiological images is a novel approach and can provide a reliable alternative to human evaluation. This approach has showed to be a valid tool for the automatic readout of images with uniform backgrounds as the CDMAM images. Results show a very strong correlation for all diameters from 0.16 mm to 2.0 mm, without any kind of additional adjustment. Besides, R* outperforms de facto standard programs applied to CDMAM evaluation.

5. REFERENCES

[1] R. Van Engen, K. C. Young, H. Bosmans, M. Thijssen. The European protocol for the quality control of the physical and technical aspects of mammography screening. Part I and II. In: European Guidelines for Breast Cancer Screening, 4th edition. Luxembourg:European Commission, 2006.

[2] K.R. Bijkerk, M.A.O. Thijssen, Th. J. M. Arnoldussen, "Modification of the CDMAN Contrast-Detail Phantom for Image Quality Evaluation of Full-Field Digital Mammography Systems", *Proceedings of IWDM 2000*, pp.663-640, Yaffe, M. ed, Medical Physics Publishing, Madison, Toronto, 2000.

[3] R. Visser and N. Karssemeijer, "CDCOM Manual: software for automated readout of CDMAM 3.4 images". Note: CDCOM software, manual and sample images are posted at www.euref.org 24th January 2009.

[4] K. C. Young, J. H. Cook, J. M. Oduko, H. Bosmans, "Comparison of software and human observers in reading images of the CDMAM test object to assess digital mammography systems", *Medical Imaging 2006*: Physics of Medical Imaging. Edited by Flynn, Michael J.; Hsieh, Jiang. Proceedings of the SPIE, Volume 6142, pp. 39-51, 2006.

[5] K. C. Young, A. Alsager, J. M. Oduko, H. Bosmans, B. Verbrugge, T. Geertse, R. Van Engen, "Evaluation of software for reading images of the CDMAM test object to assess digital mammography systems". *Proc. SPIE* 6913, 69131C, 2008.

[6] R. F. Wagner, D. G. Brown, M. S. Pastel, "Application of information theory to the assessment of computed tomography", *Med. Phys*. **6**, 83–94, 1979.

[7] W.J.H. Veldkamp et al., "The value of scatter removal by a grid in full field digital mammography", *Med. Phys*. 30, 1712-1718, 2003.

[8] R. Rico, S.L. Muller, G. Peter "Automatic scoring of CDMAN a dose study". *Proc. SPIE* 5034, pp. 164-173, 2003.

[9] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult," in Proc. IEEE Int. Conf. Acoust. , Speech, and Signal Processing, vol. 4, (Orlando), pp. 3313–3316, May 2002.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment, in *Proc. of the 37th IEEE Asilomar Conf. on Sig*., Sys. and Comp., (Pacific Grove, CA), Nov. 2003.

[12] D. M. Rouse and S. S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM", *Proc. SPIE* Vol. 6806, Human Vision and Electronic Imaging 2008.

[13] CDMAM Readout at www.euref.org, 14th February 2009.

[14] G. Prieto, M. Chevalier, E. Guibelalde, "A CDMAM Image Phantom Software Improvement for Human Observer Assessment". E.A. Krupinski (Ed.): *IWDM 2008*, LNCS 5116, pp. 181–187, Springer-Verlag Berlin Heidelberg 2008.