Use of the cross-correlation component of the multiscale structural similarity metric (R* metric) for the evaluation of medical images

Gabriel Prieto,[a] Eduardo Guibelalde, and Margarita Chevalier
Department of Radiology, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

Agustín Turrero
Department of Statistics and Operations Research, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

(Received 26 July 2010; revised 9 June 2011; accepted for publication 10 June 2011; published 21 July 2011)

Purpose: The aim of the present work is to analyze the potential of the cross-correlation component of the multiscale structural similarity metric (R*) to predict human performance in detail detection tasks closely related with diagnostic x-ray images. To check the effectiveness of R*, the authors have initially applied this metric to a contrast detail detection task.
Methods: Threshold contrast visibility using the R* metric was determined for two sets of images of a contrast-detail phantom (CDMAM). Results from R* and human observers were compared as far as the contrast threshold was concerned. A comparison between the R* metric and two algorithms currently used to evaluate CDMAM images was also performed.
Results: Similar trends for the CDMAM detection task of human observers and R* were found in this study. Threshold contrast visibility values using R* are statistically indistinguishable from those obtained by human observers (F-test statistics: $p > 0.05$).
Conclusions: These results using R* show that it could be used to mimic human observers for certain tasks, such as the determination of contrast detail curves in the presence of uniform random noise backgrounds. The R* metric could also outperform other metrics and algorithms currently used to evaluate CDMAM images and can automate this evaluation task. [DOI: 10.1118/1.3605634]

Key words: MS-SSIM, model observer, mammography, image quality, CDMAM

I. INTRODUCTION

Image quality analysis plays a central role in the design of imaging systems for medical diagnosis. A great effort to develop meaningful metrics (lab and clinical), well correlated with imaging phantom studies and with clinical performance of the medical imaging systems has been made in the last few years. The final objective of these image quality metrics (IQM) is usually to design an algorithm able to score the perceived quality of a medical image. For phantom studies, the use of automatic tools that mimic the radiologist's point of view analyzing an x-ray image could avoid interobserver and intraobserver variability and minimize the great number of images, of observers and the great deal of time usually required to optimize the image acquisition parameters or to evaluate equipment or new technologies, for instance, by means of the receiver operating characteristic (ROC). So far only partial success has been achieved. The search for IQM that fully correlates with the quality perceived by the human visual system (HVS) and particularly with the radiologist's point of view is still an open question.

Certain widely used metrics such as the peak signal-noise ratio or the mean-squared error are very simple to calculate, but do not show a good correlation with the image quality perceived by human observers[1] and indeed they are not useful to deduce the capability of diagnostic equipment.[2] Other metrics closer to the actual performance of systems, such as the modulation transfer function, the noise power spectrum, the noise equivalent quanta, and the detection quantum efficiency describe much better the image formation process of the system and can be used not only to improve image quality but also to predict the observer response under the ideal observer model approach.[3] This model, based on the statistical theory of decision can only apply to simple tasks such as a "signal-known-exactly/background-known-exactly" ("SKE/BKE") detection task.[4] Moreover, the sensitivity of the ideal observer model is much higher than that of the human observer.

There are other models that have a better correlation with the human observer, which can also be applied to more complex tasks than SKE/BKE. These include mainly the channelized Hotelling observers, the nonprewhitening matched filter (NPW) and the NPW with an eye-filter.[5] However, for mammographic images, these models are not good predictors of human performance.[5]

There are other metrics such as the structural similarity (SSIM)[6] that have shown very good results mimicking the human performance in analyzing natural images in videos and still-images. These metrics are based on the perceptual visual theory proposed by Wang and Bovik[7] that considers the HVS highly adapted for extracting structural information from the scenes. A family of objective image quality assessment algorithms has been developed based on this premise.[6,8,9] They evaluate visual image quality by measuring the structural similarities between two images, one of them

being the reference one. This family includes the cross-correlation component of the multiscale structural similarity metric (R*),[9] that has been explicitly designed for recognition threshold tasks. Note that the radiologist's tasks usually use *reduced reference* or *no reference* metrics that require only a partial reference signal or none at all. However, in some specific situations, as the case presented in this paper, it is possible to model the "perfect image" and to use reference metrics to perform automatic tasks highly correlated with observer predictions.

Despite some criticisms of the SSIM family,[10] the R* metric shows some promising features that suggest the possibility of being successfully applied to medical image analysis tasks. As mentioned above, this family is designed and fully tested to analyze natural scenes, whose complexity is of the order or even greater than that of medical imaging. It has been successfully used for ensuring the quality and fidelity of the image in a large number of commercial and research applications. In particular, it surpasses most of the metrics currently used in the analysis of video and still image.[9] Moreover, some experiments prove that R* sensitivity for detecting image structures close to the perception threshold is analogous to that of human observers.[9]

To check the effectiveness of R*, we have initially applied this metric to a contrast detail detection task. For this, we developed an automatic evaluation tool based on the R* metric that was applied to score images of the CDMAM phantom.[11] Similarly to other authors,[12] we have made a comparison of our method with human-observer contrast-detail detection tasks as well as with other automatic evaluation algorithms based on the CDCOM software.[13,14]

II. THEORY

The R* metric belongs to the set of quality assessment (QA) algorithms that seek an objective evaluation of image quality consistent with subjective visual quality. These algorithms evaluate a test image X with respect to a reference image Y to quantify their similarity. In this sense, all of them (including R*) are signal known exactly (SKE) tasks. R* evaluates perceptual quality of the X image, referred to the test image Y, by computing a local spatial index, $r(x, y)$, that is defined[9] as follows:

X and Y being images to be compared (computed as matrixes of pixels) and $\mathbf{x} = \{x_i \mid i = 1, 2,\ldots, N\}$ and $\mathbf{y} = \{y_i \mid i = 1, 2,\ldots, N\}$ pairs of local square windows (computed as sub-matrixes of pixels) of X and Y, respectively, $\mathbf{x}$ and $\mathbf{y}$ are located at the same spatial position in both images. The index $r(x, y)$ is defined in terms of the pixel value standard deviations σ_x and σ_y at sub-matrixes $\mathbf{x}$ and $\mathbf{y}$ and the covariance σ_{xy} of $\mathbf{x}$ and $\mathbf{y}$:

$$r(x,y) = (\sigma_{xy})/(\sigma_x\sigma_y) \tag{1}$$

As can be seen, if sub-matrixes $\mathbf{x}$ and $\mathbf{y}$ cover the same object in the same location, r shows a maximum.

$r(x, y)$ takes values between -1 and 1. The closer the value of $r(x, y)$ to 1, the closer the similarity between sub-matrixes $\mathbf{x}$ and $\mathbf{y}$.

When the signal or the signal + background are uniform, σ_x or σ_y tend to be zero and the value of $r(x, y)$ is unstable. This is the case of sub-matrixes measured inside uniform reference signals, where all pixels take the same value and the variance is null. For these limits, the index calculation is made by supposing that $\sigma_x > 0$, and the sub-matrix $\mathbf{y}$ is uniform. Then, the variance of $\mathbf{y}$ is zero. Under these conditions, $\mathbf{x}$ does not correlate with $\mathbf{y}$, so the $r(x, y)$ value must be set to zero. When both sub-matrixes have equal variance, the $r(x, y)$ value must be set to 1. Thus, the alternative definition of the index is given as

$$r^*(x,y) = \begin{cases} 0 \text{ for } \sigma_x > 0 \text{ and } \sigma_y = 0, \text{ or } \sigma_y > 0 \text{ and } \sigma_x = 0 \\ 1 \text{ for } \sigma_x = \sigma_y = 0 \\ r(x,y) \text{ other} \end{cases} \tag{2}$$

As the model compares two images, the test (X) and the reference (Y), the sub-matrixes ($\mathbf{x}$, $\mathbf{y}$) are moved over X and Y and $r^*(x, y)$ values are calculated for each position. If X and Y contain the same object in the same location, $r^*(x, y)$ shows a maximum.

Detail perception depends, among other factors, on the resolution of the image and on the observer-to-image distance.[8] To incorporate M observer viewing distances, the algorithm simulates different spatial resolutions by iterative down-sampling in two steps: first, a low-pass filter is applied to reduce the bandwidth of the signal to avoid aliasing effects before the signal is down sampled, and second, the size of both images (reference and test) is reduced by a factor of 2, sub-sampling without any average (averaging is not needed after the low-pass filter is applied).

These two steps are iteratively applied $M-1$ times. (The original size of the image is taken as the first viewing distance. There is no need for downsampling for $M = 1$) The overall cross-correlation multiscale structural similarity metric R^* value is obtained by combining measurement at different scales according to the following expression:

$$R^* = \prod_{M}^{j=1} r_j^*(x, y) \tag{3}$$

III. MATERIALS AND METHODS

The CDMAM phantom (version 3.4, Artinis, St. Walburg 4, 6671 AS Zetten, The Netherlands) consists of an aluminum base with a matrix of gold disks of varying thicknesses and diameters, which is attached to a PMMA cover. The discs are arranged in a matrix of 16 rows by 16 columns. Within a row, the disk diameter is constant, with logarithmically increasing thickness. Within a column, the disk thickness is constant, with logarithmically increasing diameter. Each cell in the matrix contains two gold disks each with the same diameter and thickness. The reference signal is the disk at the center of the cell and the test signal is the disk randomly located in one of the four quadrants. The imaging task can be identified as a four-alternative-forced choice (4AFC) task, since the observer has to detect the quadrant of each cell in which a disk appears to be present. This phantom

is widely used and fully tested for image quality assessment in mammography.

A set of eight raw CDMAM images (set #1) were downloaded from the European Reference Organization for Quality Assured Breast Screening and Diagnostic Services (EUREF) web site.[14] The images were obtained with a GE Senographe 2000D at 27 kVp, 125 mAs and with a resolution of 1 pixel per 100 μm. The CDMAM images were scored by four experienced human observers. Each observer scored two different images once. The observer readouts are available at the same website.

A second set of 20 images (set #2) was obtained with another CDMAM unit. In this case the images were acquired with a Sectra MicroDose LD30 at 32 kVp and 50 μm pixel size. Scoring was performed by a panel of seven experts. Six observers scored three different CDMAM images once. The seventh observer scored two different CDMAM images once. The experience of the observers interpreting mammograms was at least 3 yrs.

Both data sets were evaluated according to the methodology, and rules for CDMAM scoring published and described in the phantom manual.[11] According to this methodology, the purpose of each observation is to determine, for each disk diameter, the threshold gold thickness (the "just visible" gold thickness). So in every column (same diameter) the last correctly indicated eccentric disk has been determined. Finally, the nearest neighbors correction (NNC) rules[11] are applied to the image readouts for smoothing the edges among cells that were correctly and noncorrectly evaluated. According to these rules, for every score there are three possibilities:

- True: the eccentric disk was indicated at the true position (TP).
- False: the eccentric disk was indicated at a false position (FP).
- Not: the eccentric disk was not indicated at all.

and two main rules:

- A"True" needs two or more correctly indicated nearest neighbors to remain a "True".
- A "False" or "Not" disk will be considered as "True" when it has three or four correctly indicated nearest neighbors.

These two main rules have minor and specific exceptions for those disks that have only two nearest neighbors (at the edges of the phantom).

The software tools here presented are written as a JAVA computer algorithm and integrated program (plug-in) for the display and image processing IMAGEJ software.[15] All images are captured or defined in a gray scale of 16-bits, with pixel values from 0 up to 65535.

III.A. R* metric application to CDMAM scoring

The first task to manage the disk information from the phantom images is the accurate detection of the grid line images, which form the matrix in which gold disks are distributed. Although several methods have been applied to find the grid position,[13,16] we used here an algorithm[17] developed by ourselves, which has been successfully proven even when slight distortions of the images are present.[18]

Once the grid lines are detected, the second step to be followed is the accurate detection of the disks in each matrix cell. The algorithm looks for the gold disks around the four quadrants near the grid crossing points by analyzing the structural similarity among the cells in the phantom image (image X in the "Theory" section) and in a reference mask image of the disks (image Y in the "Theory" section). The reference or mask image is a perfect white disk, with a pixel value of 65 535, inserted into a black background (margin), with a pixel value of 0, whose size matched the disk diameter to be evaluated [Fig. 1(a)].

The technical specifications of the phantom give the nominal disk distances from the grid crossing points. However,

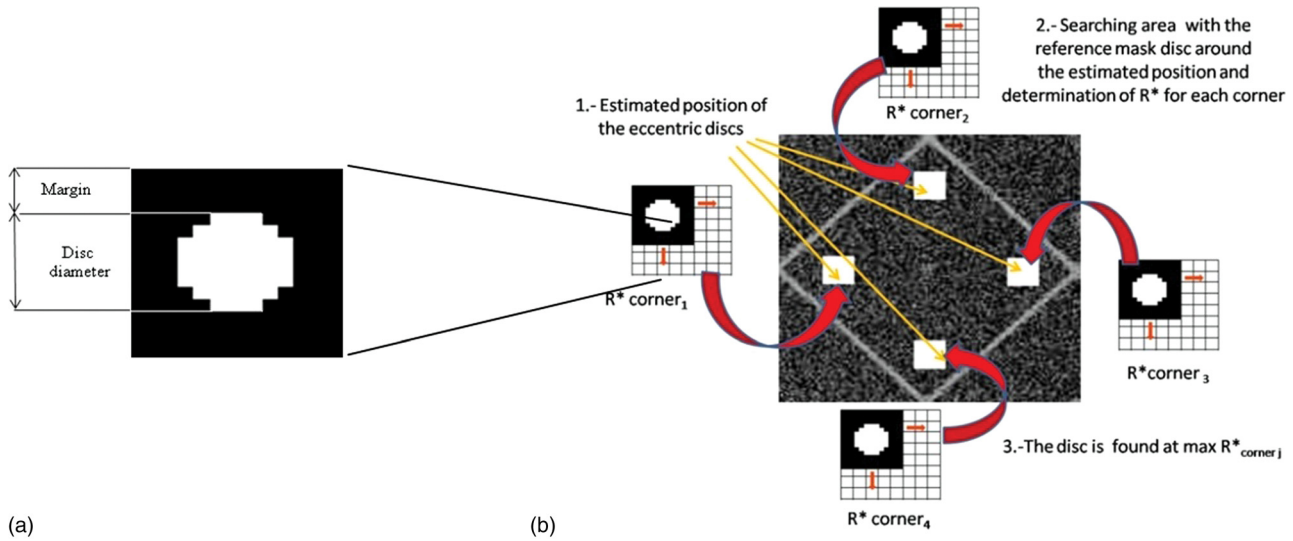
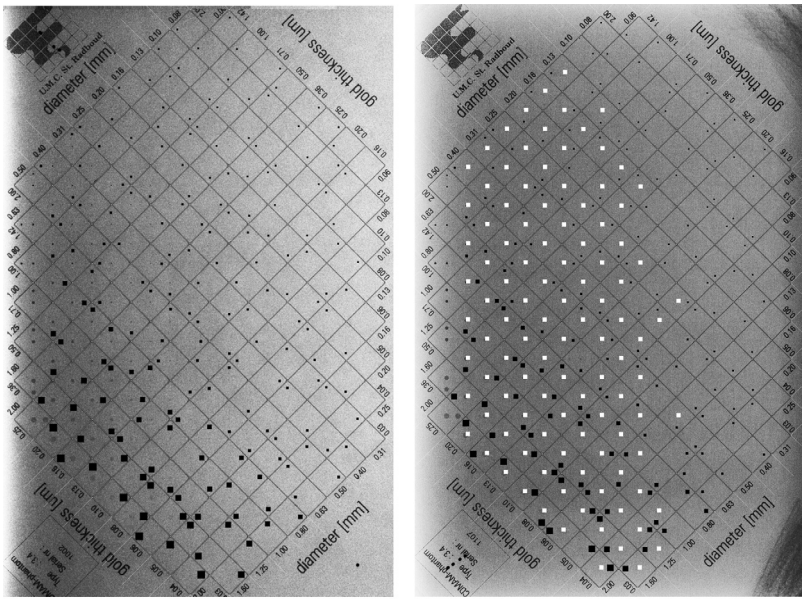


FIG. 1. Searching methodology. (a) Reference or mask image (b) Steps followed to search for the quadrant with the maximum R* i.e., most probably position of the eccentric disk.



(a) (b)

FIG. 2. (a) Graphical layout showing the predicted eccentric disk positions (black squares). (b) Graphical layout showing the correctly found eccentric disks. If the algorithm has found the eccentric disk, a white central square appears.

due to the manufacturing process, these distances can vary from unit to unit. Therefore, the R* value is calculated at 25 different positions (5×5) around the expected location of the disks at each cell quadrant [see Fig. 1(b)]. The maximum value of R* was adopted as the R* value for this cell quadrant. Then, the maximum value of the R* derived from the four quadrants determines the most probable position of the eccentric disk at each cell.

In the present work, the value of R* is calculated according to Eq. (3) where the value of M has been set to be a maximum of 5 for set #1 and of 6 for set #2, since after 5 and 6 (respectively) downsizing steps by a factor of 2, even the details of the largest disks disappear. (The larger disks of the CDMAM have a diameter of 2 mm. That means 20 pixels for set #1 and 40 for set #2, with resolutions of 100 and 50 μm per pixel, respectively. After 5 and 6, (respectively) downsizing by two, the diameter of these disks is less than 1 pixel and disappears in the image.)

Figure 2(a) shows the predicted disk location at each cell in a test image. Black squares are located at the quadrant with the maximum value of the R^*_j metric ($j = 1,..,4$). In Fig. 2(b) the white squares show the quadrants containing disks correctly identified (TP) by the algorithm (hits). Finally, the NNC rules were applied to the image readouts.

To compare the perception threshold of the R* algorithm and the human observer, Pearson correlation coefficients were calculated by comparing the thickness threshold for every image and for every diameter from the human observer and from R*, that is, this analysis was performed over the scatter plot of both variables (thickness and disk diameter) for the whole set of images.

The relationship between thickness and disk diameter was investigated by means of regression analyses in the two experiments. Comparisons of the models were carried out through the R^2 statistic. To overcome heterogeneity of variance, thickness data were log transformed.

III.B. Comparison with other methods

For comparison purposes, the sets of CDMAM images were also automatically evaluated by using two algorithms. The first one is CDCOM program,[13] which is a freely available[14] algorithm currently used for automatic evaluation of the CDMAM phantom. The second evaluation program, here named PRCDCOM, performs a smoothing and fitting of the readout matrixes produced by the CDCOM program following the procedures described by Young *et al.*[19] The threshold values derived with the two automatic methods were compared with those resulting from our algorithm.

The comparison of threshold values derived from the CDCOM and PRCDCOM methods with the values resulting from our algorithm and from the human observer was carried out through regression analyses. The obtained models were linearized and the comparison of the regression lines was studied by analyzing appropriate ANOVA tables. Statistical analyses were performed using SPSS® and STATGRAPHICS® statistical packages.

IV. RESULTS AND DISCUSSION

IV.A. R* metric application to CDMAM scoring. Threshold thickness calculations

Figure 3 shows, in a log–log graphic, the average threshold thickness for disk diameters ranging from 0.10 to 2.00 mm obtained by the experienced human observers (HO) and by applying the R* algorithm to the same sets of data.

IV.A.1. Correlation analysis

A strong linear relationship was observed between the thickness thresholds obtained from R* and human observers {Pearson coefficients $r = 0.9249$ in set #1 [Fig. 3(a)] and $r = 0.8922$ in set #2 [Fig. 3(b)]}.

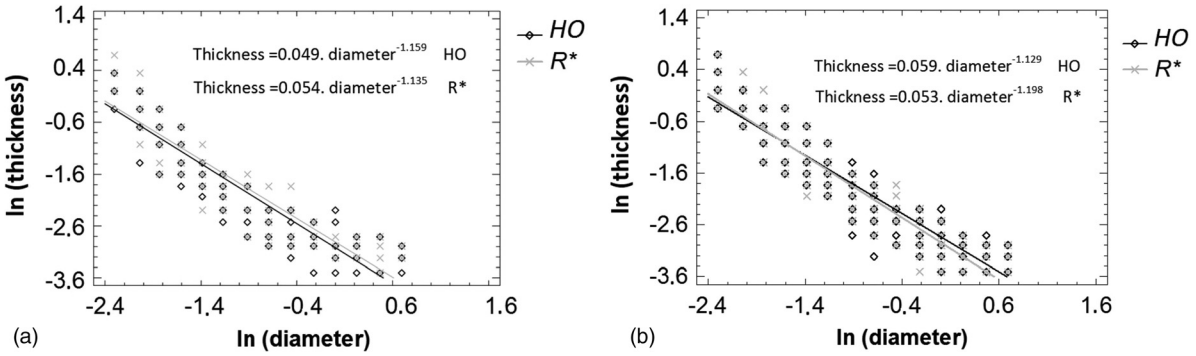


FIG. 3. Average threshold thickness as a function of the diameter from human observer (HO) and R* for set #1 (a) and set #2 (b).

IV.A.2. Regression analyses

The logarithm of thickness decreased with the disk diameter. The scatter plot suggests fitting a log–log model, that is, the approach is to consider a linear relationship among log-transformed variables; Figs. 3(a) and 3(b) show the results of these fits. The values of the R^2 statistic were 0.8624 and 0.8360 for HO and R*, respectively, in set #1 and 0.9020 and 0.8978 for HO and R*, respectively, in set #2. The statistical comparison of both regression lines shows no significant differences between them in set #1, according to the F-test statistics for the hypotheses of equality of intercepts and parallelism ($p = 0.1439$ and $p = 0.7117$, respectively) . The same comparison in set #2 shows results slightly nearer to statistical significance ($p = 0.085$ and $p = 0.065$), but always greater than statistical significance values ($p > 0.05$). These results suggest that R* could be used as a surrogate of the human observer with no evidence of statistical difference.

IV.B. Comparison with other methods

We have to point out at this juncture the range of validity for the CDCOM and the PRCDCOM programs. According to their developers,[13,19] these algorithms can only be applied to disk diameters equal to or smaller than 1.00 mm, so the graphics in Figs. 4(a) and 4(b) have been reduced from a maximum of 2.00 mm to a maximum of 1.00 mm to compare the four methods in the same range of experimental data.

The regression model which provides the best fit to the data derived from the four analyzed methods is the multiplicative or log–log model. Results are different for sets #1 and #2.

For set #1 [Fig. 4(a)], the four models fit quite well to the data (all R^2 statistics are greater than 0.93). The four regression lines are parallel with significant differences only between PRCDCOM and CDCOM threshold values (F-test statistic: $p = 0.0256$). Regarding the comparisons of intercepts, there are significant differences between the CDCOM method and the remaining ones (F-test statistics: all $p < 0.003$ for the equality of intercepts hypothesis). According to these results, PRCDCOM and R* could be adequate surrogates of the HO, but not CDCOM.

Similar results were found for set #2 [Fig. 4(b)] for the log–log model, (all R^2 statistics are greater than 0.96). In this case, the test for parallelism shows statistically significant differences between the CDCOM method and HO and R* methods ($p = 0.0002$ and $p = 0.0058$, respectively) and also between the PRCDCOM method and HO method (F-test statistic: $p = 0.011$). Regarding the comparisons of intercepts, there are significant differences between the CDCOM method and the remaining ones (all $p < 0.0002$).

Regarding Figs. 3(a) and 3(b), R* is valid for a larger range of diameters (up to 2.00 mm) than CDCOM and PRCDCOM with no statistically significant difference from the HO readouts.

According to these results, R* could be an adequate surrogate of the HO, but not PRCDCOM or CDCOM.

V. CONCLUSIONS

These results show that the R* metric can be used to mimic human observers for certain tasks, such as the

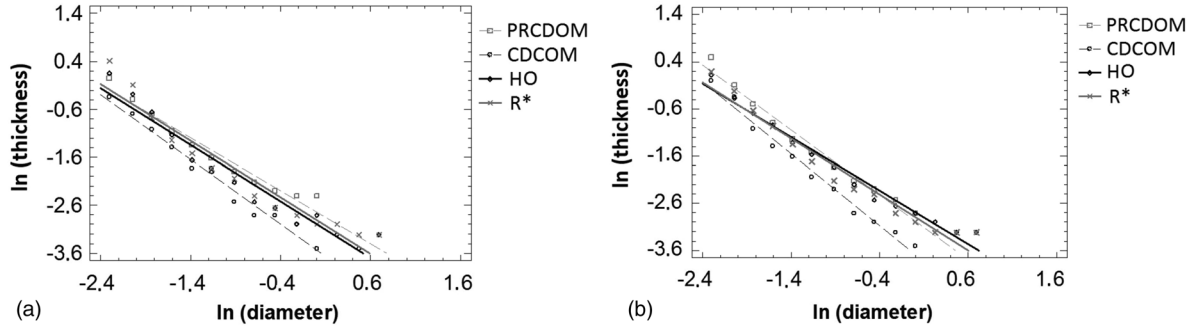


FIG. 4. Average threshold thickness as a function of the diameter from PRCDCOM, CDCOM, R*, and human observer (HO) for set #1 (a) and set #2 (b).

determination of contrast detail curves in the presence of uniform random noise backgrounds . The reliability of the results has been ensured by the similar threshold thickness obtained for each diameter by both observers, R* metric and HO, showing that both present a similar response independently of the signal, with no statistically significant difference.

Despite the fact that more samples and experiments should be carried out, the algorithm here designed based on R* metric could outperform other currently used metrics and algorithms used to evaluate CDMAM images, such as CDCOM and PRCDCOM and could be applied to the same range of disk diameters as the HO.

These results demonstrate the possibility of applying the R* metric to the medical imaging area of research applying adequate experimental conditions and methodology.

ACKNOWLEDGMENTS

The authors want to acknowledge the collaboration of Secträs Image Processing Department, especially Björn Svensson for his collaboration delivering images and human observer readouts for this experiment. They would like to thank the following people for their initial comments and opinions which encouraged us to deepen our knowledge of the possibilities and limitations of the metric considered here: David A. Clunie, Sheila S. Hemami, Elizabeth A. Krupinski, Wayne S. Rasband, David M. Rouse, and Zhou Wang. They also would like to thank Michael P. Kennedy for his help reviewing our English grammar and syntax.

[a]Author to whom correspondence should be addressed. Electronic mail: gprietor@med.ucm.es

[1]B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision,* edited by A. B. Watson (MIT, Cambridge, MA, 1993), pp. 207–220.

[2]A. E. Burgess, "The Rose model, revisited," J. Opt. Soc. Am. A **16**, 633–646 (1999).

[3]K. J. Myers, "Ideal observer models of visual signal detection," in *Handbook of Medical Imaging, Physics and Psycophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE, Bellingham, WA, 2000), Vol. 1, pp. 558–592.

[4]H. H. Barrett, K. J. Myers, and R. F. Wagner, "Beyond signal detection theory," application of optical instrumentation in medicine XIV and Picture Archiving and Communications (PACS IV) for medical applications, Newport Beach, CA. *Proceedings of the Society of Photo-optical Instrumentation Engineers,* (Bellingham, WA, 1986), Vol. 626, pp. 231–239.

[5]M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "A practical guide to model observers for visual detection in synthetic and natural noisy images," in *Handbook of medical imaging, physics, and psychophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE, Bellingham, WA, 2000), Vol. 1, pp. 593–626.

[6]Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. Image Process. **13**, pp. 600–612 (2004).

[7]Z. Wang and A. C. Bovik, "Why is image quality assessment so difficult?," IEEE Trans. Acoust., Speech, Signal Process. **4**, 3313–3316 (2002).

[8]Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Groove, Ca, IEEE (2003), pp. 529–554.

[9]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 **6806**, 680615 (2009).

[10]R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image, and Video Processing* (Springer, London, 2009), Vol. 51, pp. 81–91.

[11]K. R. Bijkerk, M. A. Thijssen, and T. H. Arnoldussen, "Manual CDMAM-Phantom Type 3.4" (translation from the Dutch by S. van Woudenberg), University Medical Centre Nijmegen, July 2000.

[12]R. M. Gagne, B. D. Gallas, and K. J. Myers, "Toward objective and quantitative evaluation of imaging systems using images of phantoms," Med. Phys. **33**, 83–95 (2006).

[13]N. Karssemeijer and M. A. O. Thijssen, "Determination of contrast-detail curves of mammography systems by automated image analysis," in *Digital Mammography*, edited by K. Doi, R. Giger, R. M. Nishikawa, and R. A. Scmidt (Elsevier, Amsterdam, 1996), pp. 155–160.

[14]R. Visser and N. Karssemeijer, "CDCOM Manual: software for automated readout of CDMAM 3.4 images," CDCOM software, manual, and sample images are posted at www.euref.org, Last accessed June 2010.

[15]W. S. Rasband, ImageJ, U. S. National Institutes of Health, Bethesda, MD, http://rsb.info.nih.gov/ij/plugins/index.html 1997–2007, Last accessed June 2011.

[16]R. Rico, S. L. Muller, and G. Peter, "Automatic scoring of CDMAN a dose study," Proc. SPIE, **5034**, 164–173 (2003).

[17]G Prieto, M. Chevalier, and E. Guibelalde, "CDMAM image phantom software improvement for human observer assessment," in *Lecture Notes in Computer Science 5116 Digital mammography*, edited by E. A. Krupinski (Springer-Verlag, Berlin, Heidelberg 2008), Vol. 5116, pp. 181–187.

[18]G. Prieto, M. Chevalier, and E. Guibelalde, "A software tool to measure the geometric distortion in x-ray image systems," Proc. SPIE, **7622**, p. 173 (2010).

[19]K. C. Young, J. J. H. Cook, J. M. Oduko, and H. Bosmans, "Comparison of software and human observers in reading images of the CDMAM test object to assess digital mammography systems," Proc. SPIE, **6142**, 614206 (2006).