Radiation Protection Dosimetry (2005), Vol. 117, No. 1–3, pp. 291–297
doi:10.1093/rpd/nci742 Advance Access published on February 7, 2006

QUALITY ASSURANCE (QA) PROCEDURES FOR SOFTWARE: EVALUATION OF AN ADC QUALITY SYSTEM

E. P. Efstathopoulos[1,*], O. Benekos[1], M. Molfetas[2], E. Charou[3], S. Kottou[4], S. Argentos[1] and N. L. Kelekis[1]
[1]Second Department of Radiology, Medical School, University of Athens, University General Hospital 'Attikon', Rimini 1, 124 62, Athens, Greece
[2]Medical Physics Department, 'Evangelismos' Hospital, Marasli 12, 111 11, Athens, Greece
[3]Institute of Informatics and Telecommunications, National Research Center 'Demokritos', 153 10, Athens, Greece
[4]Medical Physics Department, Medical School, University of Athens, 75 Mikras Asias, 115 27, Athens, Greece

Image viewing and processing software in computed radiography manipulates image contrast in such a way that all relevant image features are rendered to an appropriate degree of visibility, and improves image quality using enhancement algorithms. The purpose of this study was to investigate procedures for the quality assessment of image processing software for computed radiography with the use of existing test objects and to assess the influence that processing introduces on physical image quality characteristics. Measurements of high-contrast resolution, low-contrast resolution, spatial resolution, greyscale (characteristic curve) and geometric distortion were performed 'subjectively' by three independent observers and 'objectively' by the use of criteria based on pixel intensity values. Results show quality assessment is possible without the need for human evaluators, using digital images. It was discovered that the processing software evaluated in this study was able to improve some aspects of image quality, without introducing geometric distortion.

INTRODUCTION

Computed radiography (CR) imaging is based on storage phosphor detectors which are characterised by a large dynamic range, in comparison with the limited range of the output medium and viewing process. Image processing techniques in CR aim to manipulate image contrast, in such a way that all relevant image features are rendered to an appropriate degree of visibility, despite the restriction of viewing density range. Image processing can also contribute to improving the image quality by using techniques such as edge enhancement and multi-scale contrast enhancement, in addition to adjusting density, contrast and gradation of the whole image.

The main objective of the current work was the development of quality assurance (QA) procedures for image interpretation and processing software used in CR and the evaluation of the image quality improvement offered by such software[1]. The image quality of a radiograph can be determined by local contrast, spatial resolution, image noise, latitude and geometric distortion[2–4].

MATERIALS AND METHODS

Imaging conditions

Radiographs of a test object were obtained in a Siemens Tridoros 5S X-ray system used for general purpose radiography (Siemens AG, Munich, Germany) using a size 24 cm × 30 cm Agfa ADCC HR cassette (Agfa-Gevaert N.V., Mortsel, Belgium). The cassette used is dedicated for CR and contains a storage phosphor imaging plate. The phantom was placed in contact with the cassette on the table, with a focus detector distance of 100 cm. In order to ensure that cassette sensitivity was not altered through the exposures, the same cassette was used throughout. The exposures were carried out using measured kilovoltage ranging from 70 to 100 kV_p with a 1 mm copper filter. The phantom used was a Leeds Test Object type TOR-CDR[5] (Leeds Test Objects Ltd, West Yorkshire, UK). The test object was used at 70 kV_p, with 1 mm Cu added filtration.

Image digitisation

The image cassette was read in an Agfa ADC Solo^{TM} digitiser system (Agfa-Gevaert N.V., Mortsel, Belgium) reading sampling frequency of 9 pixels per mm and greyscale resolution of 12 bits per pixel[6]. Images containing 2570 × 2040 pixels were saved in the reader system's format. To make the images available to other workstations, they were exported in 12-bit intensity DICOM format. During this export procedure, the same export parameters were used for all the images. Each pixel's value is calculated by the MUSICA software, taking into account all the processing involved. Image data are not compressed when they are exported to 12-bit DICOM standard.

The imaging software used to identify, process, view and archive the acquired images is ADC

*Corresponding author: stathise@cc.uoa.gr

Quality System QS (Agfa-Gevaert N.V., Mortsel, Belgium). ADC QS provides image interpretation and processing options using image processing software called MUSICA, which is standard in QS software.

Image processing

Images were processed using MUSICA processing algorithms. The resulting images were evaluated with the same methods as the non-processed images in order to determine the quality performance of the software. The software offers a choice between various characteristic curves, depending on the diagnostic area of the image requiring enhancement[7–9]. All images were processed using a linear gradation curve and automatic window width and level. Each image was processed with the most commonly used parameters (in a scale of 0–6, Musi-contrast = 3, Latitude reduction = 3, Edge contrast = 3) and, in addition, with each individual algorithm separately.

Image evaluation

The images were evaluated by three independent observers, determining the high-contrast resolution (HCR) and low-contrast resolution (LCR) as well as the spatial resolution (SR). Two of the observers are medical physicists experienced in the evaluation of the test object used. The third observer was a trained radiologist with no previous experience in using the specific test object and without any prior knowledge of the study. Images were shown to the observers in a random order so that any bias would be removed. The average value of observers' scores was used in data analysis and a measure of inter-observer variability is given as the standard deviation of these scores. These measurements are referred to as 'subjective' measurements.

Additionally, images were evaluated using software developed especially for this purpose (i.e. an 'objective' evaluation). This software enabled the DICOM images to be read, rotated, regions of interest (ROIs) set and the mean intensity and standard deviation of a desired region in the image to be measured. Furthermore, we were able to create the intensity value profile of a line drawn in the image and use these profiles to evaluate spatial resolution.

The Leeds Test Objects TOR-CDR contains 17 LCR discs of 11 mm diameter each. The mean greyscale value of each disc was measured and compared with the mean greyscale of the background around each disc. The task in our case was to detect the presence of a disc-shaped object of known size and location in a noisy background. This is a signal-known-exact problem that has been modelled using statistical decision theory. The probability of correctly detecting the object is related to the absolute difference of mean greyscale values of the object and the background. The criterion whether each disc is clearly distinguishable from the region around it (defined as background) was considered to be that this difference should be >2 standard deviations of the background, establishing a level of 95% confidence that the two regions are different[10].

The test object also contains an area with 30 separate groups of bar patterns, each group comprising 5 bars and 4 spaces, giving 4.5 line pairs. The HCR was evaluated using the intensity profile of a line perpendicular to the line pairs of the HCR area of the phantom. A separate group was considered to be apparent if all five bars of the group were distinguishable. A bar was considered as distinguishable if its greyscale value was larger than the average of successive spaces, plus one standard deviation, establishing a 68% probability that the bar is distinguishable.

Another step of evaluation of spatial frequency was the measurement of the spatial frequency of each group and comparison with the nominal value that the manufacturer states for the test object. The total width of each group in pixels was measured, from which the spatial frequency can be calculated using Equation 1:

$$\mathrm{SR} = 4.5 \text{ line pairs} \times \mathrm{width(pixels)} \times \text{pixel spacing(pixels per mm)}. \quad (1)$$

The manufacturer of the Leeds test object provides a set of 10 circular details of diameter 5.6 mm, used to obtain sensitometry measurements of the image (normally, the optical density of these discs can be measured with a densitometer). The contrast value of each disc is given for beam conditions of 70 kV_p (constant potential) and 1 mm Cu filtration[5]. The contrast values are calculated as the ratio of the difference of greyscale value of the objects, minus the greyscale value of the background, divided by the background greyscale value, as shown in the following equation:

$$C_{\mathrm{object}} = \frac{GS_{\mathrm{object}} - GS_{\mathrm{background}}}{GS_{\mathrm{background}}}. \quad (2)$$

These are nominal values, however, and actual values depend on the shape of tube potential waveforms and are subject to manufacturing tolerances of ±5%. For the same beam conditions, the nominal contrast values of the large discs, used for LCR, and the small discs, used for HCR, are also provided. Using the software we developed, the contrast value of each disc was calculated (Equation 2) and the greyscale values of both the greyscale set of discs and the large LCR discs. Sensitometry measurements of this kind allow a direct comparison between the different algorithms provided in

MUSICA software concerning the changes each algorithm yields in the contrast of different objects.

RESULTS

Low-contrast resolution

LCR values of images obtained at 6.4 mA s vs. the processing algorithm for various kilovoltage values are shown in Table 1.

High-contrast resolution

HCR values of images obtained at 6.4 mA s vs. the processing algorithm for various kilovoltage values are shown in Table 2. HCR scores were not taken using the 'objective' method owing to the small size of high-contrast discs.

Spatial resolution

SR values of images obtained at 6.4 mA s vs. the processing algorithm for various kilovoltage values are shown in Table 3. The calculated SR for each line pair group was plotted against the nominal SR value and the plot is shown in Figure 1.

Sensitometry

In Figures 2 and 3, contrast values for all LCR and greyscale discs are shown. The contrast value of each disc was calculated as described above. In these figures, it is shown how each algorithm results in manipulation of object contrast depending on the object contrast in the original images.

Geometric distortion

Vertical and horizontal diameters of a circular area of the phantom were measured for all the images obtained and their ratio (r) was calculated as an indication of the geometric distortion, with a value of 1.0 corresponding to no geometric distortion. The average of this ratio for all the images (nine total images) is 0.999 with a standard deviation equal to 0.002.

DISCUSSION

LCR was not affected by image processing for 'subjective' evaluation at 70 kV_p, except for minor differences (5%), as shown in the first column of Table 1. The same is the case for data at 80 kV_p, with the maximum difference of 12% for the edge

Table 1. LCR scores (number of visible discs) vs. processing algorithm for various kilovoltage values at 6.4 mA s.

LCR vs. algorithm	70 kV_p		80 kV_p		100 kV_p	
	'Subjective' score (s.d.)	'Objective' score	'Subjective' score (s.d.)	'Objective' score	'Subjective' score (s.d.)	'Objective' score
Original	13.3 (0.6)	11.0	14.0 (0.0)	11.0	14.7 (0.6)	8.0
Edge contrast	13.0 (0.0)	4.0	12.3 (0.6)	5.0	13.3 (0.6)	5.0
Latitude reduction	13.0 (0.0)	11.0	14.3 (0.6)	11.0	15.0 (1.0)	9.0
Musi-contrast	13.0 (0.0)	11.0	14.3 (0.6)	10.0	15.3 (1.5)	8.0
Noise reduction	12.7 (0.6)	11.0	14.0 (1.0)	13.0	13.7 (0.6)	8.0
Commonly used parameters	13.0 (0.0)	7.0	14.0 (0.0)	8.0	14.3 (0.6)	8.0

The values of 'subjective' assessment are the mean values of the observers' scores with their standard deviation

Table 2. HCR scores (number of visible small discs) vs. processing algorithm for various kilovoltage values at 6.4 mA s.

HCR vs. algorithm	70 kV_p 'Subjective' score (s.d.)	80 kV_p 'Subjective' score (s.d.)	100 kV_p 'Subjective' score (s.d.)
Original	12.3 (0.6)	12.3 (0.6)	12.0 (1.0)
Edge contrast	12.7 (0.6)	12.7 (1.5)	12.3 (1.2)
Latitude reduction	12.3 (0.6)	12.3 (0.6)	12.0 (1.0)
Musi-contrast	12.7 (0.6)	12.7 (1.2)	12.3 (0.6)
Noise reduction	12.0 (0.0)	12.3 (0.6)	12.0 (0.0)
Commonly used parameters	12.3 (0.6)	12.7 (0.6)	12.7 (0.6)

The values of 'subjective' assessment are the mean values of the observers' scores with their standard deviation

Table 3. Spatial resolution values (number of visible bar groups) vs. processing algorithm for various kilovoltage values at 6.4 mA s.

SR vs. algorithm	70 kV_p		80 kV_p		100 kV_p	
	'Subjective' score (s.d.)	'Objective' score	'Subjective' score (s.d.)	'Objective' score	'Subjective' score (s.d.)	'Objective' score
Original	16.0 (0.0)	18	16.0 (0.0)	18	16.0 (0.0)	17
Edge contrast	16.3 (0.6)	18	15.7 (1.2)	18	16.3 (0.6)	17
Latitude reduction	15.7 (0.6)	18	15.3 (0.6)	18	16.0 (0.0)	17
Musi-contrast	16.0 (0.0)	17	15.7 (0.6)	19	16.3 (0.6)	18
Noise reduction	15.3 (0.6)	17	15.3 (0.6)	18	16.0 (0.6)	17
Commonly used parameters	17.0 (0.0)	19	17.0 (0.0)	18	17.3 (0.6)	18

The values of 'subjective' assessment are the mean values of the observers' scores with their standard deviation

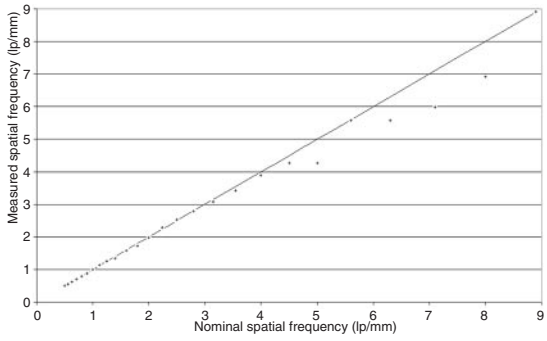

Figure 1. Calculated vs. nominal spatial frequency (lp mm^{-1}). Nominal values are shown in the horizontal axis while the calculated ones are on the vertical axis. The measured values for each line pair group are shown with a cross. The straight line represents the ideal situation in which calculated and nominal values would be equal.

contrast enhancement (row 2, Table 1). Scores show a greater variability at 100 kV_p, where there is a deterioration of 9.5% for edge contrast enhancement and a minor improvement of 4.5% for musi-contrast (last column, Table 1).

Scores of 'objective' measurements show a greater variability depending on the processing algorithm. For all tube potentials, LCR scores decrease with edge contrast enhancement as much as 64% (row 2, Table 1). Latitude reduction and musi-contrast show only minor differences in scores at all tube potentials. Image noise reduction improves the score by a degree of 18% at 80 kV_p, but does not affect the score at other tube potentials.

'Commonly used parameters' were chosen because they were the default set of parameters used in clinical practice. As it becomes apparent, one should make a careful choice of settings depending on the image characteristic (i.e. LCR) to be improved. The deterioration caused by the edge enhancement algorithm (row 2, Table 1) is the result of an increase in the standard deviation of the selected ROI, forcing the criteria not to be attained, even though the local contrast between the disc and the background had been improved[11–13].

As far as spatial resolution is concerned, application of edge contrast enhancement results in a small overall improvement for 70 and 100 kV_p, and a small decrease for 80 kV_p in the case of 'subjective' scores (row 2, Table 3). The same situation is observed for latitude reduction and musi-contrast enhancement, where only small differences of $\sim$4% can be seen. In contrast, noise reduction seems to decrease the scoring, except at 100 kV_p where no difference is observed. 'Objective' scores of edge contrast enhancement and latitude reduction are exactly the same as the non-processed image scores for all tube potentials, while musi-contrast enhancement seems to improve scoring, and noise reduction scores are the same as the original image except for 70 kV_p (17 instead of 18 bar groups). The combination of edge enhancement with latitude reduction and multiscale contrast amplification in the commonly used parameters slightly improved SR scoring[14].

In Figure 1, measured spatial frequency of each bar group is plotted against nominal spatial frequency value given by the manufacturer. For low frequencies, measured and nominal values are linear, but as frequency increased a deviation from linearity is observed. This deviation is due to the limiting factor of pixel resolution and subsequent pixel spacing does not allow the measurement of small objects with great accuracy.

As shown in Table 2, no statistically significant differences between original and processed image scores are observed in HCR scores (P-values ≥ 0.05, Mann–Whitney test).

Evaluation of HCR was not performed 'objectively', owing to difficulties concerning the very small size of high-contrast details not allowing ROI to be drawn. Line profiles can be used to make an

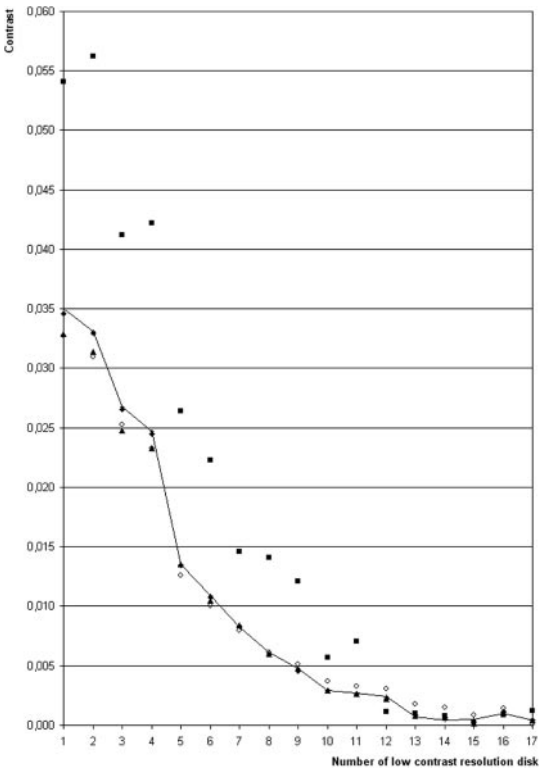

Figure 2. Contrast values of LCR discs for an image processed with different algorithms. Disc number is shown in horizontal axis and the corresponding contrast value in the vertical axis. Signs used are a cross with a continuous line for the original image, closed squares for musi-contrast enhancement, closed triangles for edge contrast enhancement, open diamonds for latitude reduction and closed diamonds for noise reduction.

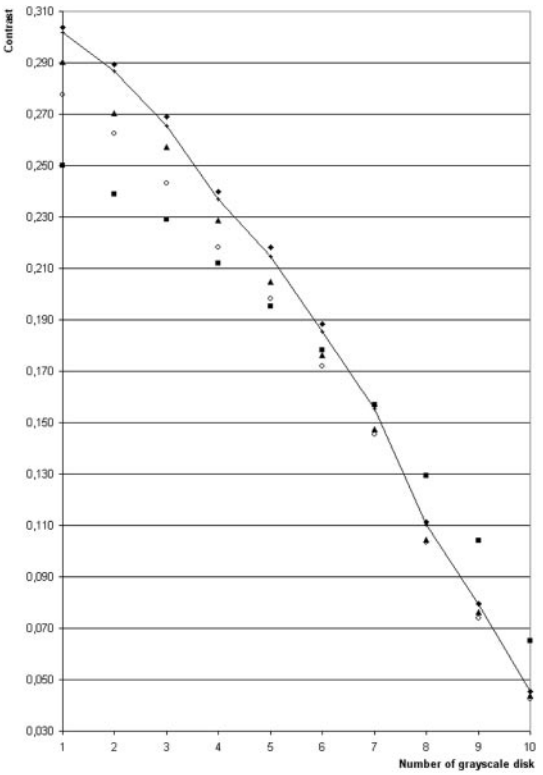

Figure 3. Contrast curves of greyscale discs for different processing algorithms. Disc number is shown in horizontal axis and the corresponding contrast value in the vertical axis. Signs used are a cross with a continuous line for the original image, closed squares for musi-contrast enhancement, closed triangles for edge contrast enhancement, open diamonds for latitude reduction and closed diamonds for noise reduction.

evaluation, but the precise position of the details must be known. Although this is not a problem when the user can clearly see the details on the screen, it is not possible to know where to take the profiles for details not clearly seen, making the 'objective' evaluation no better than the 'subjective' one.

These results indicate that there may be an improvement using edge contrast enhancement and musi-contrast amplification, but no firm conclusion may be drawn until more work is done, especially in the direction of developing a kind of 'objective' evaluation method for HCR.

Results of sensitometry measurements shown in Figures 2 and 3 reveal the way each algorithm alters the contrast of different objects. Noise reduction does not alter object contrast, and edge enhancement deteriorates contrast of high-contrast discs, as seen in Figure 3 for discs 1–8 but no such effect is observed for lower contrast objects. While musi-contrast enhancement decreases contrast for high-contrast objects (as shown for discs 1–6, Figure 3), this imposes no problem in image evaluation since these discs can easily be differentiated from their background. On the contrary, it manages to increase contrast of low-contrast discs as shown in discs 1–11 (Figure 2). For lower contrast discs, contrast values do not show any clear difference (discs 12–17, Figure 2). Latitude reduction decreases object contrast of high-contrast discs, but as object contrast falls, no differences can be observed.

Results of musi-contrast show that the algorithm improves contrast for a number of discs as long as the difference between a disc and its background is not too small. This comes at the cost of decreasing the contrast of clearly visible discs, which is acceptable.

Variability between different observers is given in Tables 1–3 as the standard deviation of the average value. From these values, the coefficient of variation

(COV) can be calculated. For LCR assessment (Table 1), COV values range between 0 and 10% with a mean value of 3.4% and in only three cases are greater than 5% showing a good agreement between observers. For HCR assessment (Table 2), COV values range between 0 and 12.1% with a mean value of 5.5%. For SR assessment (Table 3), COVs range from 0 to 7.4% (mean = 2.2%) showing a very good agreement between observers.

'Subjective' scores between original and processed images were also evaluated using the Mann–Whitney test. The *P*-values calculated are >0.05 except for edge contrast enhancement in LCR at 80 kV_p (*P*-value of 0.03) and for 'commonly used parameters' in SR at all tube potential values. Therefore, differences in most cases are not statistically significant and no safe conclusion can be drawn whether changes introduced by processing algorithms in test object images can be observed. Further work in combination with clinical studies is needed in order to safely decide if evaluation of processing algorithms can be performed with test object using LCR, SR and HCR assessments with observers. Further work is also needed to decide if evaluation software is more accurate and sensitive than analysis using human observers making evaluation procedure with test objects possible.

LCR 'subjective' scores obtained by the observers compared with 'objective' scores differed significantly with *P*-values <0.001 for all kV_ps (Mann–Whitney test). The same is true for SR where 'subjective' and 'objective' scores show a statistically significant difference ($P \leq 0.001$ for 70 and 80 kV_p, $P = 0.002$ for 100 kV_p, Mann–Whitney test). These differences are due to the fact that the criteria used for the 'objective' assessment have not been set in accordance with the scores of the observers. This can be addressed by changing the selected threshold, making it less or more strict so that scores obtained agree with observers' scores. This issue should be addressed in further work.

Another conclusion concerning the sensitivity of each evaluation scheme in changes of image characteristics can be drawn by comparing the 'objective' and 'subjective' scores. Differences between original and processed images in LCR 'objective' scores (range between 4 and 13, Table 1) are larger than corresponding differences in 'subjective' measurements (*P*-values ≤ 0.05 except for 80 kV_p with *P*-value = 0.03, Mann–Whitney test) indicating that 'objective' evaluation method is more sensitive to LCR changes, thus allowing evaluation even when differences between images cannot be detected by the observers. SR measurements ('subjective' scores range between 15.3 and 17.3, 'objective' scores range between 17 and 19) do not indicate any advantage in favour of any method. In the case of 'commonly used parameters', both methods show a significant improvement except for 80 kV_p where 'objective' method does not detect any difference. However, 'objective' scores for musi-contrast algorithm at 80 and 100 kV_p show an increase (from 18 to 19 and 17 to 18, respectively) while 'subjective' scores are not statistically different from original images (*P*-value ≥ 0.05, Mann–Whitney test). Further work is needed in order to conclude whether 'objective' evaluation method is more sensitive than evaluation with human observers.

Finally, no geometric distortion was observed, indicating that MUSICA software does not alter the geometric characteristics of acquired images.

CONCLUSION

The purpose of this study was to determine a procedure for the quality assessment of image processing software for CR with use of existing test objects. Measurements of HCR, LCR, SR and geometric distortion were performed 'subjectively' by three independent observers and 'objectively' using software especially developed for this purpose. The performance of commercial image processing software was evaluated using these methods.

One of the key issues in this study is the applicability of the method used for evaluation of image processing software, which is developed for clinical purposes and may not perform well in test objects. Even though in most cases no significant differences were observed between original and processed images, some changes in specific cases were observed and therefore no safe conclusion can be drawn about the possibility of evaluation image processing algorithms using test objects.

This study did not aim to directly evaluate the processing software and conclude whether it passes an assessment, but rather to use the results of this evaluation in order to investigate whether image processing results in a measurable change of image characteristics. Furthermore, a prerequisite of software assessment is the establishment of a protocol and determination of threshold limits for the results that are yet to be done.

Further work is needed for the following:

(1) Make improvements on the software we have developed and create a tool for evaluation of HCR, which was not performed 'objectively' in current work.
(2) Establish a connection between the results of the assessment of test objects and performance in clinical studies. For this purpose, the use of a clinical phantom emulating clinical cases is being considered.
(3) Establish criteria for the measurements made by the assessing software that correlate to measurements made by the observers.

(4) Evaluate software made by other manufacturers for different purposes.
(5) Recommend quantitative acceptance criteria for satisfactory performance of software based on the evaluation results of different software in several institutions.

REFERENCES

1. Kamm. *The quality of digital X ray images*. Medica Mundi **38**, 111–116 (1992).
2. Moores, B. M., Stieve, F. E., Eriskat, H. and Schibilla, H. *Technical and physical parameters for quality assurance in medical diagnostic radiology: tolerances, limiting values and appropriate measuring methods*. British Institute of Radiology Report 18 (1989).
3. Samei, E., Seibert, J. A., Willis, C. E., Flynn, M. J., Mah, E. and Kunk, K. L. *Performance evaluation of computed radiography systems*. Med. Phys. **28**(3), 361–371 (2001).
4. Doi, K. *Basic imaging properties of radiographic systems and their measurement*. In: CG Orton, editor. Progress in Medical Radiation Physics (NY: Plenum Press) pp. 194–198 (1985).
5. Instructions for the use of Leeds Test Object type TOR-CDR. Medical Physics Department, University of Leeds, UK.
6. For Computed Radiography in general, the home page is: Available at http://www.agfa.com/en/he/solutions/radiology/cr/index.jsp.
7. Moores, B. M., Wall, B. F., Eriskat, H. and Schibilla, H. *Optimization of image quality and patient exposure in diagnostic radiology*. British Institute of Radiology Report 20 (1989).
8. Kheddache, S., Denbratt, L. and Angelhed, J. E. *Digital chest radiography—optimizing image processing parameters for the visibility of chest lesions and anatomy*. Eur. J. Radiol. **22**, 241–245 (1996).
9. Schaefer, C. M. *et al. Interstitial lung disease. Impact of postprocessing in digital storage phosphor imaging.* Radiology, **178**, 133 (1991).
10. Wells, P. N. T. *Scientific Basis of Medical Imaging* (Edimburgh;New York: Churchill Livingstone) (1982).
11. Rimkus, D. and Bailey, N. A. *Quantum noise in detectors*. Med. Phys. **10**(4), 470–471 (1983).
12. Ishida, M. *et al. Digital image processing: effect on detectability of simulated low contrast radiographic patterns.* Radiology **150**, 569–575 (1984).
13. Giger, M. L., Doi, K. and Metz, C. E. *Investigation of basic imaging properties in digital radiography. 2. Noise Wiener spectrum*. Med. Phys. **11**(6), 797–805 (1984).
14. Loo, L. D., Doi, K. and Metz, C. E. *Investigation of basic imaging properties in digital radiography. 4. Effect of unsharp masking on the detection of simple patterns*. Med. Phys. **12**, 209–214 (1985).