Journal of Electronic Imaging 11(2), 224–235 (April 2002).

Validation of a human vision model for image quality evaluation of fast interventional magnetic resonance imaging

Kyle A. Salem
Case Western Reserve University
Department of Biomedical Engineering
Cleveland, Ohio

Jonathan S. Lewin
Case Western Reserve University
University Hospitals of Cleveland
Department of Radiology
Department of Oncology
Cleveland, Ohio

Andrik J. Aschoff
Case Western Reserve University
University Hospitals of Cleveland
Department of Radiology
Cleveland, Ohio

Jeffrey L. Duerk
Case Western Reserve University
Department of Biomedical Engineering
University Hospitals of Cleveland
Department of Radiology
Cleveland, Ohio

David L. Wilson
Case Western Reserve University
Department of Biomedical Engineering
University Hospitals of Cleveland
Department of Radiology
Cleveland, Ohio
E-mail: dlw@po.cwru.edu

Abstract. *Perceptual difference models (PDMs) have become popular for evaluating the perceived degradation of an image by a process such as compression. We used a PDM to evaluate interventional magnetic resonance imaging (iMRI) methods that rapidly acquire an image at the expense of some anticipated image degradation compared to a conventional slower diagnostic technique. In particular, we examined MR keyhole techniques whereby only a portion of the spatial frequency domain, or k-space, was acquired, thereby reducing the time for the creation of image updates. We used a PDM based on the architecture of another visual difference-model and validated it for noise and blur, degrading processes present in fast iMRI. The PDM showed superior correlation with human observer ratings of noise and blur compared to the mean squared error (MSE). In an example application, we simulated four keyhole techniques and compared them to a slower, full k-space diagnostic acquisition. For keyhole images, the MSE gave erratic results compared to the ratings by interventional radiologists. The PDM performed much better and gave an A_z value >0.9 in a receiver operating characteristic analysis. Keyhole simulations showed that a single, central stripe acquisition, which sampled 25% of k-space, provided stable image quality within a clinically acceptable range, unlike three other keyhole schemes described in the literature. Our early experience shows the PDM to be an objective, promising tool for the evaluation of fast iMRI methods. It allows one to quantitatively make engineering decisions in the design of iMRI pulse sequences.* © *2002 SPIE and IS&T.* [DOI: 10.1117/1.1453412]

Paper 001002 received Jan. 8, 2001; revised manuscript received June 19, 2001; accepted for publication Sep. 5, 2001.
1017-9909/2002/$15.00 © 2002 SPIE and IS&T.

1 Introduction

Image quality is important in medical imaging because images are viewed by physicians for diagnosis, for planning of therapy, for application of therapy, and for assessment of therapy. Since the diagnostic task is often one of detecting a lesion, there is a long history in medical imaging of quantitatively measuring image quality as the capability to detect a target defect. Researchers use experimental methods such as the receiver operator characteristic[1] and forced choice,[2] and theoretical analyses using a variety of models of human detection.[3,4] Most often this has been done in projection x-ray and nuclear medicine imaging where ionizing radiation must be limited and quantum noise is often a factor.

We wish to optimize the image quality of interventional magnetic resonance imaging (iMRI) methods, and this presents several issues, some of which are different than those found in diagnostic imaging. First, fast iMRI images are used to guide minimally invasive procedures such as biopsies of diseased tissue[5] and therapy of focal tumors.[6] Optimization of these images is not amenable to traditional analyses because the task is no longer detection of pathology. During interventional procedures, the needle, the target tissue, and any critical structures to avoid, like arteries, must all be well above the detection threshold. Instead, as will be shown later, fast MR imaging methods cause considerable image differences compared to traditionally acquired images, such as blurred or missing needle tips and ghosted tissue boundaries. An appropriate image quality measure must capture all of these differences and simultaneously consider many image features. Second, it is desirable to measure image quality in the context of the task performed with the images. This is precisely what is done with detection analyses for diagnostic imaging. Unfortunately, the ability of a physician to perform an iMRI-guided task is probably not easily or reliably measured, even if one uses a controlled, simulated intervention or a simplified task such as guiding a needle to a target. Third, a metric useful for optimization should be a continuous function of imaging parameters. Fourth, because of the very many parameters associated with fast iMRI methods, a computational approach is highly desirable. These considerations led us to reject conventional detection analyses and some potential experiments in favor of a perceptual difference model (PDM) that is described later. To relate results to the ability to perform a task, we compare model outputs to visual scoring by radiologists highly skilled in iMRI-guided interventions.

There are a variety of technical solutions to speed iMRI imaging, and they typically degrade visualization of the target and needle compared to standard, slower conventional diagnostic acquisition. We are currently examining a popular method called keyhole imaging whereby a portion of the Fourier domain, or *k*-space,[7,8] is updated to create the next image in an image sequence (Fig. 1). The imaging rate is improved because fewer *k*-space samples are acquired for each new image. Compared to conventional full *k*-space acquisition, such images are typically degraded by blurring of the needle's tip and by changes in noise structure.[9,10]

A number of strategies and applications for keyhole sampling have been reported. Some methods include static stripe acquisition,[11] *k*-space sampling based on the energy content,[10] and random selection.[12] Other proposed methods change or rotate the frequency-encode direction, and they include rotated keyhole methods, phase read exchange keyhole (PHREAK) imaging, and rotated keyhole methods.[11,13,14] Various applications have been proposed, including dynamic contrast agent studies,[9] functional MRI,[15] active tip tracking,[16] and contrast-enhanced breast imaging.[17] Finally, in addition to keyhole, there are alternative spatial encoding schemes such as wavelets[18,19] and singular value decomposition.[20,21]

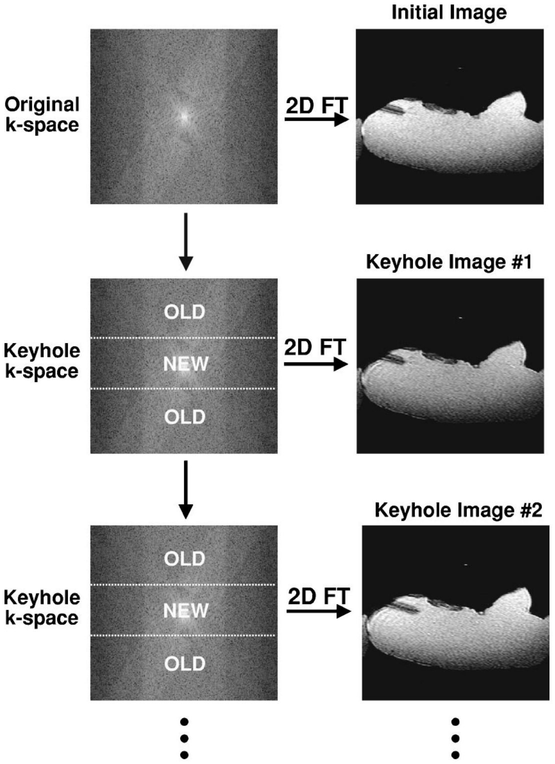


Fig. 1 Keyhole MR imaging process. The process starts with an initial full *k*-space acquisition and reconstruction using a 2D FT (top). Subsequent images are reconstructed from newly acquired *k*-space samples that replace some of the old ones. The sample-replace-and-reconstruct process is repeated for the entire keyhole sequence. Multiple strategies for sampling *k*-space are proposed (see the text).

Fast iMRI techniques have been analyzed and assessed in various ways. The modulation transfer function of the keyhole process was simulated by Spraggins to evaluate the change in maximum intensity of objects of various sizes.[9] Duerk *et al.* measured errors in the width and position of a simulated needle.[10] Others have employed anecdotal visual inspection and the absence of severe artifacts as a measure of image quality and usefulness.[14,16] It is also popular to compare MR images from new acquisition techniques or reconstruction methods with an accepted ‘‘gold standard’’ by calculating the mean square error between them.[20,22] No definitive method for assessing fast MR image quality has been determined.

Perceptual difference models should provide a good alternative for assessing image quality. It is reasonable to assume that the acquisition parameters (TR, TE, etc.) of conventional, slowly acquired MR images are optimized to give appropriate visualization of the needle, target tissue, and other tissues of interest. Hence, our goal is to create a fast iMRI image that best matches this optimal image, and the perceptual difference between the two images becomes our image quality figure of merit. Several researchers are using perceptual difference models to examine the differences between original images and images degraded from compression algorithms (see the work of Winkler[23] for a review). The degradation in keyhole imaging is very similar to that in compression, but the task is very different. We must consider the task of the interventional radiologist when assessing the quality of keyhole imaging. We are, therefore, compelled to validate the model for this new application and the clinical task at hand.

While the basic architecture is the same,[24] a number of perceptual difference models have been used for a variety of evaluation tasks. The spatiotemporal model by van den Branden Lambrecht and Verscheure has been used to assess the image quality of digitally coded pictures and image sequences.[25] Lubin developed a different spatio-temporal model for the evaluation of image display quality known as the Sarnoff human vision discrimination model.[26,27] Jackson *et al.* applied the Sarnoff model to medical imaging in an attempt to use it to design x-ray image systems.[28] The model has also been applied to microcalcification detection in mammography.[29] Daly developed the visible differences predictor (VDP) for the development of image processing algorithms, imaging system hardware, and imaging media.[30]

In this article, we develop, validate, and use a modified PDM similar to the VDP proposed by Daly.[30] We examine the capability of the model to correlate with human observer ratings of image quality and to predict the clinical acceptability of keyhole images. As an example application, we quantitatively compare four different methods for keyhole imaging previously suggested for interventional work. These strategies evaluate the effectiveness of periodically sampling all of *k*-space, as opposed to the repeated sampling of only the central portion of *k*-space. We next describe methods for validating the PDM and for simulating and evaluating keyhole images with the PDM.

2 Methods

2.1 *Perceptual Difference Model*

The PDM used in this study has the same basic architecture as the VDP developed by Daly,[31] and it is described in detail in the Appendix. The PDM is a mechanistic model that represents the functional anatomy of the visual pathway (Fig. 2). It contains components that model the optics and sensitivity of the retina, the spatial contrast sensitivity function, and the channels of spatial frequency found in the visual cortex. The PDM accepts two input images, a ‘‘gold standard’’ full *k*-space image and a keyhole image. The output is a spatial map that shows the magnitude of the differences that a human observer will perceive between the two images. This map can be qualitatively used to detect regions of difference. It can also be averaged over the image or a region of interest (ROI) to give a scalar PDM error per pixel.

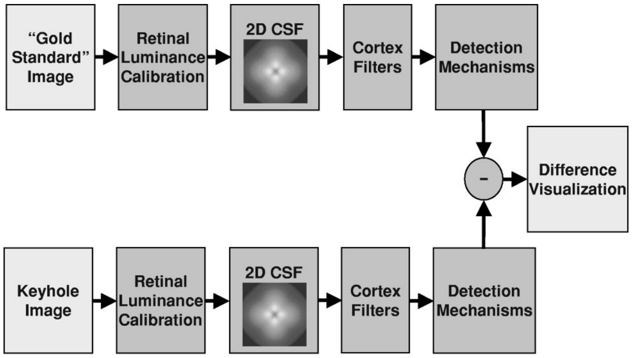


Fig. 2 Block diagram of the PDM. The output is a map showing the likelihood of a perceptual difference between the two input images. The PDM models major processes in the human visual system, described in the Appendix.

2.2 *Experiments with Degradation Due to Noise or Blur*

Two major degrading processes in keyhole MR imaging are additive noise and blur. We examined the capability of the PDM to quantify these degradations by a comparison to image quality ratings assigned by human observers in a manner similar to that in previous work by Martens and co-workers.[32,33]

In the noise experiment, a series of 72 image presentations (12 images, each with 6 different noise levels) were shown to five observers consisting of three image-processing experts and two board-certified radiologists. Each presentation consisted of a two-panel display. On the left was a low noise ‘‘reference’’ image, and on the right was a ‘‘test’’ image with added MR noise. The images were similar but not identical to fast iMRI keyhole images in later studies. MR noise is known to be Rician,[34] but, at the low signal-to-noise ratios experienced in keyhole imaging, its amplitude can be approximated by a Gaussian distribution.[35] Hence, white noise fields were created from a Gaussian distribution, and the noise standard deviation was varied from 0 to 80 gray levels, a range that exceeds that seen in fast interventional MR imaging. (Typical MR signal values range from 0 to 850 with a mean of 135 on a 0–4095 scale, and the maximum noise standard deviation is about 60 gray values in our clinical images.) The noise fields were added to high signal to noise ratio (SNR) images created through signal averaging, and maintained the image contrast and resolution of fast interventional MR images. Observers were instructed to score the quality of the test image on a scale of 0–100, with 100 being the worst quality (most noise) and 0 being the highest quality (least noise). They were told to scale their responses to the reference image assuming it had a score of 20. In a learning session, they were shown several images so as to gauge the extremes, and throughout testing the viewing distance was between 0.5 and 1 m. For each observer, the results were normalized by linearly mapping all responses between the highest and lowest scores, which were fixed at 1 and 0, respectively. This normalization accounts for interobserver differences in scoring range and offset. The same image pairs were also evaluated with PDM and mean squared error (MSE) metrics.

The capability of the PDM to measure blurring was evaluated in a similar manner. Test images were degraded with a set of ideal low-pass filters in the frequency domain that closely mimics the blur due to keyhole imaging. The filters were circular with six cutoff frequencies ranging from 0.05 to 0.5 cycles/pixel, retaining about 1%–80% of the *k*-space samples, respectively. The resulting 72 images were used in the same experimental protocol described for noise evaluation above. Observers were asked to score the image quality of the degraded image on a scale of 0–100, with 100 being the worst quality (most blur) and 0 being the highest quality (least blur), assuming the reference image received a score of 20. Observer results, normalized over each subject, were compared to both the PDM and MSE.

2.3 *Experiments With Keyhole Images*

Image quality must be determined within the context of the intended visual task. The above studies that separately correlate noise and blur with human observer image quality scoring should help provide confidence in the relative comparison of keyhole image scores. We next compare PDM scores to ratings by interventional radiologists of keyhole MR images that simultaneously include all degrading processes that result from keyhole imaging. Because our engineering interest is not only to create the fastest MR images possible but to also maintain image quality suitable for interventional procedures, the highly experienced iMRI radiologists were asked to determine the suitability of the images for interventional tasks. For each keyhole image, we asked the radiologists to give a binary rating, acceptable or unacceptable. This rating allowed us to determine a minimally acceptable PDM score.

Details of the study follow. Two board-certified interventional radiologists with extensive experience in iMRI [two of the present authors (J.S.L. and A.J.A.)] evaluated 100 keyhole images of varying quality created with the four keyhole techniques described later. Each keyhole image was paired with the corresponding gold standard, full *k*-space image and presented side by side in such a way that the keyhole image was always on the right. To assist in determination of needle tip accuracy, the radiologists were provided with a ‘‘matched’’ cursor that gave the same spatial location in both images, allowing them to compare the location of features, such as the needle tip position, between the gold standard and keyhole images. Without knowledge of the computer scoring, the radiologists classified the keyhole images by consensus voting as acceptable or unacceptable for clinical use. The radiologists examined images for severe artifacts, accurate representation of the device shaft and tip location, and sufficient signal to noise and contrast to noise characteristics. In certain instances, the radiologists preferred to use a marginally acceptable rating for images that might be used only in low risk interventions far from critical anatomy. The PDM and MSE scores were compared to classifications by the radiologists.

2.4 *Example Application to Keyhole Optimization*

Keyhole MR imaging was simulated from full *k*-space image acquisitions of needle insertions in *ex vivo* bovine liver. Images were acquired on a 0.2 T interventional MRI system (Magnetom Open, Siemens, Erlangen, Germany) using a True FISP (16/8/1/90, TR/TE/NSA/FA) imaging sequence that could be used to acquire *k*-space in actual keyhole imaging experiments. In each sequence, 20 images were acquired as a needle was inserted at an angle ≈30° above the horizontal axis over a distance of approximately 3–5 cm. Images were taken in the plane of the needle, and the needle was advanced approximately 1.0–2.5 mm between each frame. This image series was the gold standard by which the keyhole methods were evaluated. Each image of the 20 frame sequence was converted to *k*-space data by taking a two-dimensional Fourier transform (FT). The simulated keyhole imaging process was begun by sampling selected lines of *k*-space from the second image and using these new values to replace *k*-space data from the first image. A keyhole image was then constructed by taking the inverse Fourier transform of the *k*-space array consisting of new and old samples. This sample, replace, and inverse transform procedure was repeated for the rest of the image sequence (Fig. 1).

Four keyhole sampling strategies were simulated based on methods reported by Duerk *et al.*[10] and by Parrish and Hu.[12] Each sampling method acquired 64 lines (25%) of *k*-space for each update. We define the line of *k*-space containing the dc value as line 0 with lines above and below defined as positive and negative, respectively. The four *k*-space update methods listed below were performed on a 256 pixel square image.

1. Standard. The central 64 lines, lines −32 through 31, of *k*-space were sampled for each frame over the entire image sequence.
2. Sequential. The first update was performed by selecting the central 64 lines of *k*-space (lines −32 through 31). The second update sampled 32 lines on either side of the previously updated area (lines −64 to −34 and 32 to 63). The third update sampled lines −96 to −65 and 64 to 95. The final update sampled the remaining lines, −128 to −97 and 96 to 127. This process was repeated, and all of *k*-space was updated every four images.
3. Rank-ordered. Assuming one has *a priori* knowledge of the final needle location and orientation, the energy in each line of the two-dimensional (2D) discrete FT (DFT) of a needle in this position was calculated. All lines were then ranked from highest to lowest in terms of energy. The first update sampled the 64 lines that contained the highest energy. The second update sampled the next 64 lines, ranked 65–128. The third and fourth updates sampled lines ranked 129–192 and 193–256, respectively. This process was repeated over the image sequence, and again, all of *k*-space was sampled every four updates. (The first 64 line updates matched those from the standard and sequential methods exactly.)
4. Random. This algorithm randomly selected 64 lines of *k*-space to update. The second update was performed by randomly selecting 64 of the remaining 192 lines. The third and fourth updates were performed similarly. The process was repeated for the image sequence ensuring that all 256 lines were updated every four images.

For all methods other than standard keyhole, each method samples all of *k*-space over any set of four chronological images. Each of the methods was simulated both with the read direction nearly parallel and nearly perpendicular to the needle insertion.

For each simulated keyhole sequence, the 20 simulated keyhole images were compared to their corresponding images in the full *k*-space gold standard sequence using both the PDM and MSE. Both metrics were applied to the entire image and an area inside a fixed ROI that encompassed the needle. This manually selected ROI was ≈5 cm in length and ≈1.5 cm in width, and encompassed the entire needle track. A total of 120 simulated images (3 original sequences, 2 read axis orientations, 20 images per sequence) were created for each of the 4 keyhole methods. To examine the development of image error over a sequence, PDM errors were plotted as a function of frame number. The averages and maximum values were also examined.

3 Results

3.1 *Correlation of Human Scoring With Degradation Due to Noise or Blur*

We compared PDM and MSE metrics to human scoring of image blur. In Fig. 3(a), PDM scores from the 72 image presentations (12 images each with 6 different levels of blur) are plotted versus human observer scores, normalized in the manner described in Sec. 2 and averaged over 5 observers. The results are highly correlated ($R^2=0.861$). The MSE is poorly correlated ($R^2=0.363$) with human observer ratings [Fig. 3(b)]. At a given filter level, the MSE responses varied considerably for similarly filtered images because of variation in the number or size of blurred structures in the images. In addition, the human observer responses and PDM error scores decreased with an increase in the radius of the filter roughly as the inverse of the square root of the radius (not shown).

Figure 4 shows the results from noise experiments. Once again observer responses were normalized and averaged over the five observers. Both the PDM and MSE scores are linearly related to human observer scoring. With an increase in noise standard deviation, the PDM, MSE and human error observer scores all increased as one minus the inverse of the cube root of σ (not shown).

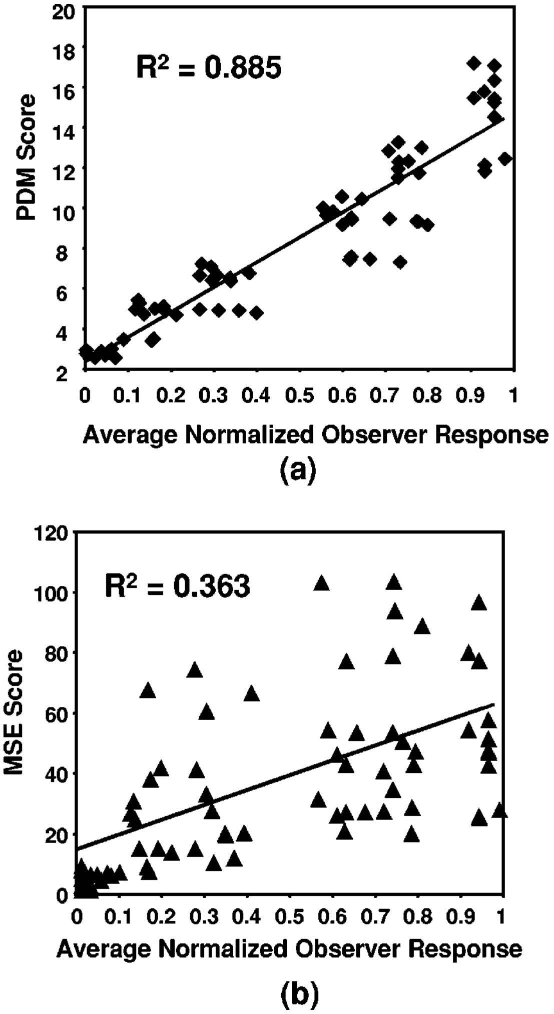


Fig. 3 Correlation of the PDM and MSE with human observer ratings of image quality in the presence of blur. Human observer ratings are averaged over five observers after subject normalization as described in text. PDM results in (a) show excellent linear correlation ($R=0.941$). The MSE in (b) does not correlate as well, giving $R=0.602$. The scatter of the MSE shows that using a nonlinear model to fit the data will not improve the correlation coefficient. To create the 72 blurred images, filters preserved from 0.7% to 78% of the original *k*-space data.

3.2 *Correlation with Clinical Acceptability*

For keyhole images of needle insertions, we compared image quality metrics with acceptability ratings by radiologists. We used 100 images with 25 from each of the four keyhole strategies described in Sec. 2.4. The panel had difficulty in identifying marginally acceptable images, stating they might be appropriate for some procedures but not others. Only 12 of 100 images were classified in this manner and their PDM and MSE scores overlapped both the acceptable and unacceptable values (Fig. 5). For further analyses, these images were deemed unacceptable. This is the most conservative approach; any borderline images are classified as unacceptable.

PDM and MSE scores are plotted against image classification in Fig. 5. Clearly, acceptable and unacceptable MSE scores [Fig. 5(b)] overlap more than do PDM scores [Fig. 5(a)]. One method by which to set a threshold is to use the mean of the average score from the set of acceptable images and the average score from the set of unacceptable images. (This threshold minimizes the total number of misclassification errors in the case of Gaussian distributions of equal variance and equal prior probability.[36]) The values, 11.41 for PDM and 9.25 for MSE, are used as a threshold for clinical acceptability and are plotted as solid horizontal lines. The PDM and MSE have similar sensitivities with these thresholds (68.2% and 65.9%, respectively), but the specificity of the PDM is 92.9% compared to only 51.8% for the MSE. Since it is highly undesirable for an image quality metric to give an acceptable rating when the image is unacceptable, this high specificity result for PDM is clearly superior. Other thresholds are possible. One is to set the threshold just below the score of the lowest unacceptable image, thereby ensuring that all unacceptable images

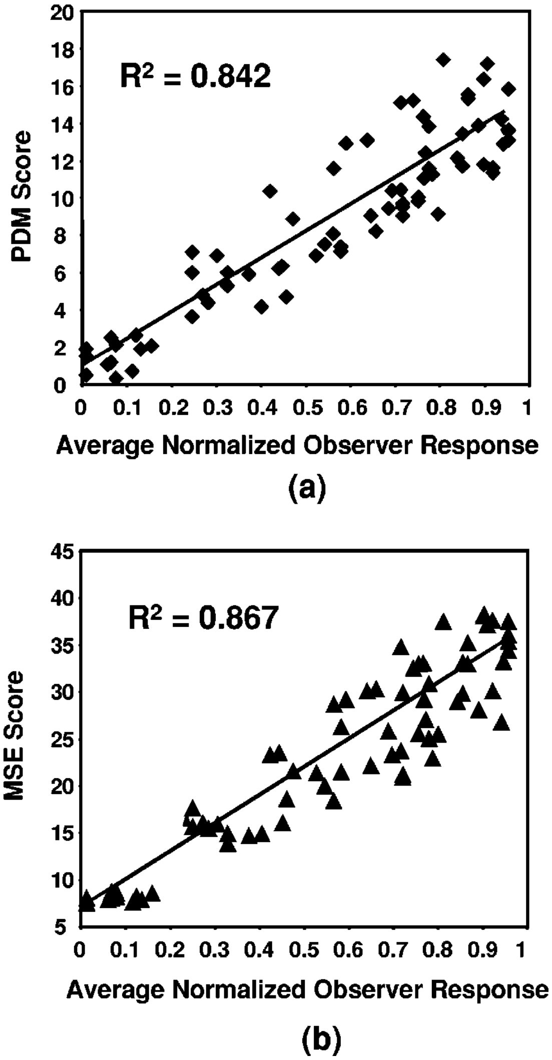


Fig. 4 Correlation of the PDM and MSE with human observer ratings of image quality in the presence of additive noise. Plots of the PDM vs human observer results in (a) show excellent linear correlation, $R=0.918$. MSE results in (b) show an equally linear relationship ($R=0.931$). A broad range of noise levels is used with standard deviations ranging from 0 to over 80 gray levels.

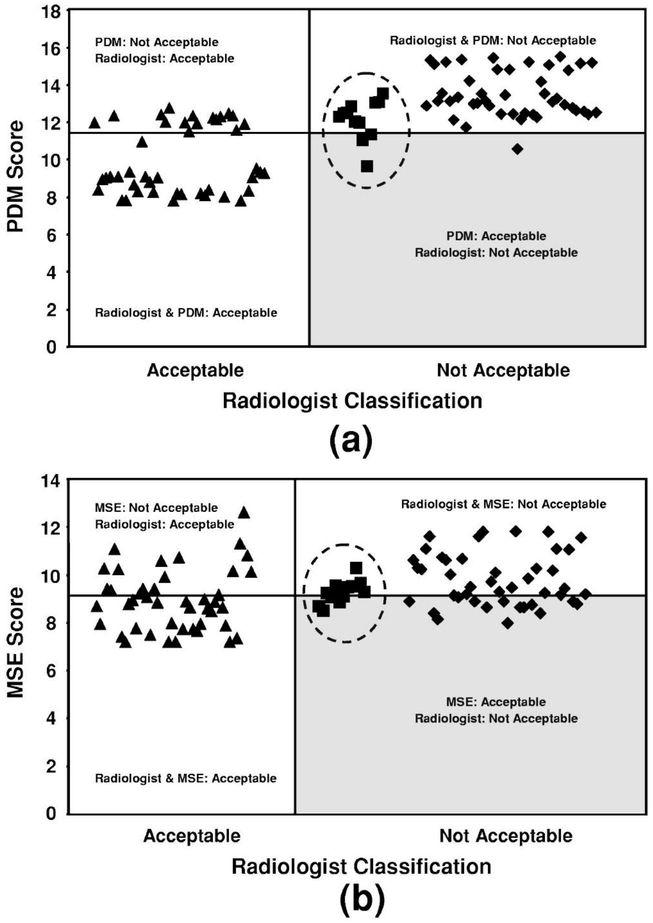


Fig. 5 PDM (a) and MSE (b) scores compared to clinical acceptability determined by a panel of interventional radiologists. Marginally acceptable scores are circled, and, in our conservative analysis, they are lumped with the unacceptable group (see the text). Data points represent the average error inside the region of interest and are horizontally displaced in chronological order for display purposes only. The thresholds were 11.41 and 9.25 for the PDM and MSE, respectively. The PDM shows a good separation between acceptable and unacceptable scores and a low number of classification errors. MSE scores overlap more than the PDM ones, resulting in a large number of classification errors.

are rejected by the metric. This creates specificities equal to 100% and sensitivities of 61.4% and 31.8% for the PDM and MSE, respectively.

ROC analysis provides a method for determining the capability of scalar image quality metrics to predict clinical acceptability without setting a threshold (Fig. 6). This analysis assumed the radiologists' binary rating (acceptable or unacceptable) was ground truth. The PDM was superior to the MSE ($p<0.01$), with A_z values (area under the ROC curve) of 0.9397 and 0.7046, respectively [Fig. 6(a)]. ROC curves were also generated with data from a previous experiment using a different method of keyhole imaging, in which the keyhole stripe size and orientation were varied[13] [Fig. 6(b)]. Again, the PDM was superior to the MSE in predicting clinical acceptability with A_z values of 0.975 and 0.888, respectively. The absolute PDM scores in both studies had a similar range, from 1.5 to 21.3 in the previous study and from 2.6 to 28.3 in the current study despite being different techniques. Additionally, the one technique that was analyzed in both studies, standard one fourth of k-space, had remarkably stable PDM scores of 6.1 and 6.85 in the current and previous experiments, respectively. The thresholds of the two studies were 11.41 and 8.4 for the current and previous studies, respectively. Some of the difference can be attributed to the sampling of the images used to set the threshold. Nevertheless, either threshold could be applied to either set of data to produce similar ROC results and similar conclusions about keyhole sampling.

The MSE sometimes produced erratic results over an image sequence (Fig. 7). PDM and MSE results from a standard keyhole sequence sampling the central 25% of k space are plotted versus the frame number (top panel). In all images, some ringing at the tissue boundary and blurring of the needle tip were seen, yet all were classified as clinically acceptable. The PDM graph is very stable, showing little to no change in perceptual image quality over the sequence. The MSE results, however, show large variations

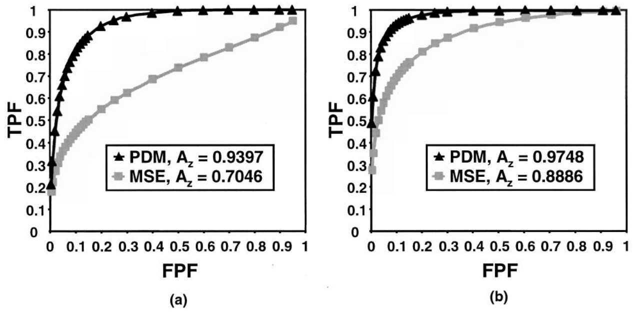


Fig. 6 ROC curves comparing the capability of the PDM and MSE to predict image acceptability. Using classification by radiologists as ground truth, ROC curves from MSE and PDM scores were created by defining acceptable as a lower score on a continuous scale. Data from the study shown in Fig. 5 are shown in (a). The PDM outperforms the MSE, with an area under the ROC curve that is significantly different ($p<0.01$). In (b), ROC curves are shown using data from a previous study with another keyhole method whereby the stripe size and orientation were varied (after Ref. 13). Again, PDM outperforms the MSE. The data points plotted were created through the maximum-likelihood estimation of the parameters for the ROC curve and were calculated directly from input data without binning using ROCKIT version 0.9 B (Ref. 48).

over the sequence. Visual inspection of frames 4, 8, and 14 shows little difference in image quality even though frame 8 has a MSE value more than twice that of frame 4. Inspection of image values reveals a small change in average image brightness between the frame 8 full *k*-space image and the keyhole image that accounts for much of the MSE error. The PDM is insensitive to such shifts and agrees much better with qualitative visual inspection. This is primarily due to the low frequency attenuation of the contrast sensitivity function component in the PDM model. MSE is not considered further in this article.

3.3 *Example Application to Keyhole Optimization*

As an example application, we now compare in Figs. 8 and 9 standard keyhole imaging to the alternative methods discussed in Sec. 2.4. A total of 480 keyhole images was analyzed; they consisted of 3 sequences of 20 frames each and 4 keyhole strategies each simulated with 2 different read axis orientations. Because one fourth of *k*-space was acquired for each update, the effect of read axis orientation was not as large as when fewer samples were obtained.[13] As a function of the frame number, the results showed similar trends, with a small improvement for the acquisition in which the read axis was closest to perpendicular to the long axis of the needle. Hence, we averaged results for the two orientations to account for this variation. We reached similar conclusions to the averaged results when the two read axis orientations were examined separately. The results for the three sequences were similar and again the results were averaged. The averaged PDM scores are shown over 20 frames for both the whole image [Fig. 8(a)] and inside the ROI [Fig. 8(b)]. The horizontal dotted line shows the threshold for clinical acceptability determined in Fig. 5. Images with high PDM scores contained severe blurring or did not accurately show the position of the needle in an image update. The PDM analysis shows a clear advantage of standard keyhole imaging over the other three methods. It provides stable, acceptable results throughout the entire

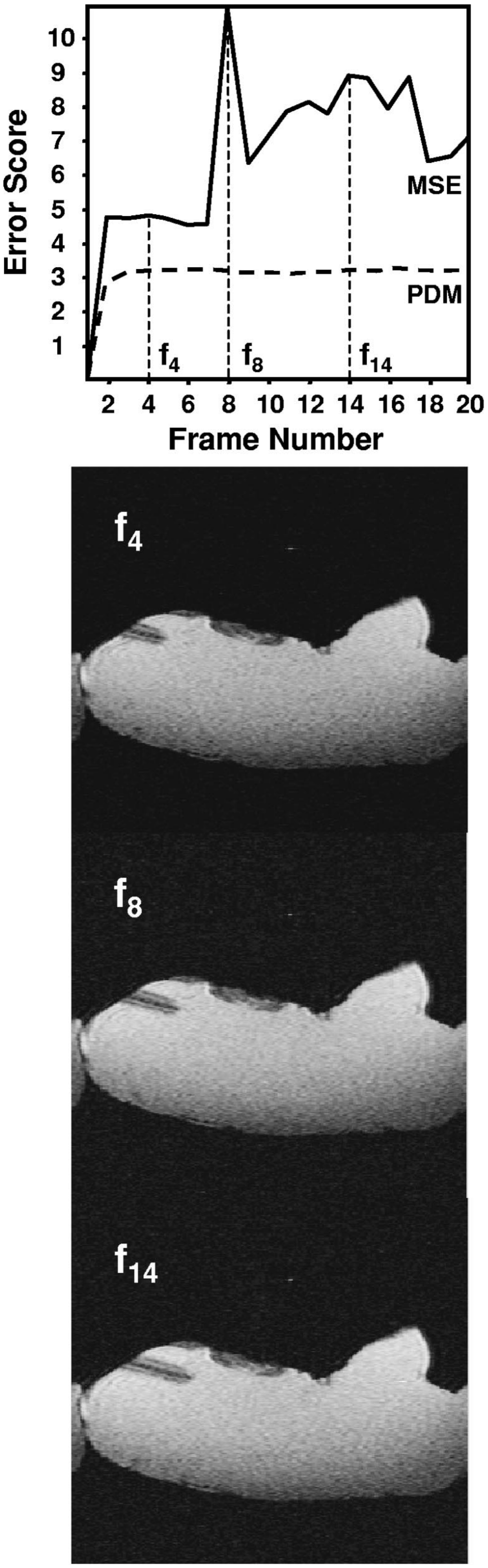
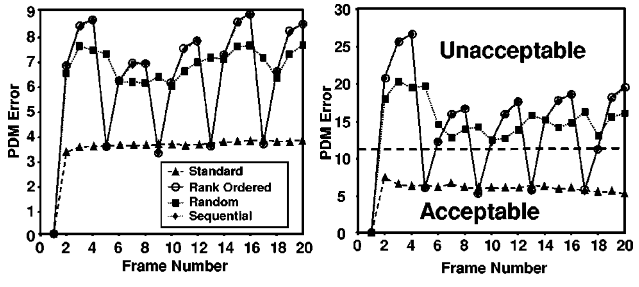


Fig. 7 Over a sequence, MSE scores vary erratically compared to the PDM. Plots of MSE and PDM scores (top) over a keyhole sequence show the highly variable nature of MSE compared to the stability of the PDM. Three keyhole images (f4, f8, and f14 corresponding to frames 4, 8, and 14, respectively) show the needle advancing slightly between each frame. All images are visually very similar and were classified as acceptable by radiologists, but f8 has a very poor MSE score.



(a) Whole image

(b) Inside ROI

Fig. 8 The PDM error is plotted over 20 frames for the keyhole image sequences. The PDM error is shown for the entire image (a) and inside the needle region of interest (b) for the four keyhole methods. Each data point represents the average of three different needle insertions each with two different read axis orientations. A low score is obtained when an image is perceptually similar to the original. Inside the ROI, the standard sequence is the only method with all images below the threshold specified by radiologists (dotted line).

sequence while the other methods show cyclical error patterns. Although a change in the read axis orientation from vertical to horizontal caused a small increase in the PDM error, all scores for the standard acquisition remained in the acceptable range. Other methods were less affected by a change in read axis orientation. The average and maximal errors also show the standard keyhole method to be superior (Fig. 9). The three other methods have nearly the same average.

4 Discussion

The results show that PDM is a promising method with which to measure iMRI image quality and that it is better correlated with radiologists' ratings of clinical acceptability than the standard method of MSE. First, PDM shows highly linear correlations with human observer ratings for images degraded by blur and noise, the principal features of degradation in keyhole MR imaging (Figs. 3 and 4). The broad range of degradations in these experiments should exceed those expected for practical fast iMRI methods. Second, when actual keyhole images that have a wide range of quality are evaluated, the PDM does a very good job separating acceptable and unacceptable images. That is, it is possible to choose a threshold value that selects those images deemed unacceptable by radiologists with high sensitivity (68.2%) and specificity (92.9%) as shown in Fig. 5, and the areas under the ROC curves are high, 0.975 and 0.9397 (Fig. 6). Together, these results validate the PDM to be an accurate measure of relevant image degradation and show that it reflects the suitability of images for clinically relevant tasks. To our knowledge, the PDM is the first image quality measure that has been correlated with a radiologist's evaluation of suitability for performing an iMRI task.

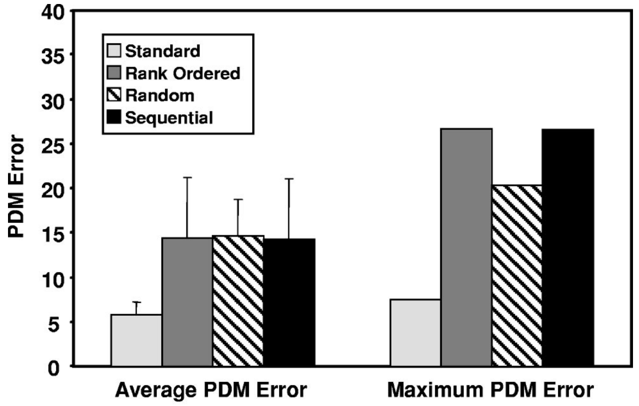


Fig. 9 Average and maximum PDM errors inside the region of interest show the standard keyhole sequence to be superior to the other simulated methods. While the random *k*-space selection method has an equivalent average PDM error, its maximum is considerably lower than the rank-ordered and sequential methods.

Mean square error performs poorly. Its most disturbing failure is the erratic frame-to-frame behavior (Fig. 6) that is not seen with the PDM. Additionally, MSE performs poorly compared to the PDM in the ROC analysis with keyhole images; the ROC area was 0.9397 for the PDM versus 0.7046 for the MSE. Although it does well with added image noise, MSE correlates poorly with human observer estimates of image quality in the presence of blur (Fig. 3). The reason is that the MSE score of a blurred image is related to the number and strength of the edges in the image and it gives very different responses for different images. Similar conclusions about the inappropriateness of MSE have been reached with regard to the image quality of compression algorithms.[27,37]

PDM scores accurately measure differences due to keyhole imaging by simultaneously capturing multiple degrading processes. The PDM error scores in the keyhole images are sometimes as high as 25 and exceed the maximal score of ≈18 in the most degraded images in the blur and noise experiments. This can be attributed to more complex image errors that occur due to keyhole processing. The combination of a blurred or misrepresented needle tip and ghosted tissue boundaries compounds the error beyond the individual effect of blur or noise. Finally, very high PDM scores are calculated inside the ROI where errors are concentrated, while scores of the blur and noise images are averaged over the entire image. The PDM not only captures complex image differences but also identifies their location.

While the good correlation of PDM scores with the assessment by radiologists suggests that the PDM is a useful measure of image quality for fast iMRI studies, one must be careful about generalizing results. The PDM might not be a good measure of absolute image quality for different types of clinical images and processing techniques. In a comprehensive review of image compression studies,[38] perceptual difference model scores were not always reliable when comparing vastly different images or processing techniques. Nevertheless, many in the image compression community consider such models to be rather reliable when comparing similar images and similar processing. Similarly, we believe the ordering of similar keyhole techniques over the same image sequences to be quite valid, and evidence of this is given by the excellent agreement with ratings of clinical task acceptability by radiologists for two different keyhole-processing techniques (Fig. 6). However, additional validation will be required before a PDM score from a liver study is compared to that in a brain, or a PDM score for keyhole is compared to that from wavelet or singular value decomposition imaging. Likewise, additional study is necessary before an acceptability threshold from simulated phantom results is used to select methods for *in vivo* patient imaging in which additional image errors might be present. Nevertheless, our results show that a threshold can be robust when comparing similar images. That is, a single threshold could be applied very well to the two different studies reported in Fig. 6. Although we should be cautious when applying a threshold for acceptability, relative PDM scoring can be successfully applied to a variety of engineering parameter optimizations. It can be used to determine the best method from a group.

There are important observations from the quantitative image quality PDM measurements applied to the keyhole simulations. First, the standard keyhole sequence is the only option that provides acceptable image quality throughout the sequences. While others prefer this method based on visual inspection[11] and simulation of the modulation transfer function of the keyhole process,[9] only the PDM has quantitatively shown the standard sequence to be appropriate for clinical applications. Second, simulations show that changes in read axis orientation that do not perfectly align with the needle cause only small differences in image quality at this sampling level, one fourth of *k*-space. Since the sampling in this study is both at 30° and 60° from parallel to the needle track, we expect the simulation results to estimate the effect of not aligning the needle in any fashion. One might infer that the additional engineering required to orient the read axis might be unnecessary for the standard acquisition strategy. However, with smaller keyhole stripes and perfect alignment with the needle, the PDM shows a much greater effect of orientation.[13] Additionally, while previous keyhole research has shown perfect alignment of the read axis perpendicular to the long axis of the needle significantly increased image quality sampling one fourth of *k*-space,[11] the PDM is the first method that quantitatively describes the perceived advantage achieved through alignment. Third, compared to the standard method, there is only a very small improvement in image quality produced every fourth image by the rank-ordered and sequential methods. From this finding, we can conclude that sampling the high frequency regions of *k*-space provides little improvement in perceived image quality for iMRI applications. These three results show the capability of the PDM to aid in making engineering decisions.

PDM error averaged over an entire sequence can be misleading, and one must also consider worst-case results. For example, catastrophic artifacts in a single image may render an entire sequence unacceptable. Yet, the sequence error score could be well under the acceptability threshold if the single high error score is averaged with low image error scores from the rest of the sequence. This is the case for sequential and rank-ordered methods that produce images of widely varying quality. In fact, only one in every four images gives a clinically acceptable score, while the others do not. When images having the highest PDM scores were examined, we found that they contained high levels of blur and inaccurate needle tip representations for new image updates. Obviously, one does not want any such artifactual images during an intervention. The maximum score gives a more conservative approach for evaluating imaging methods.

The perceptual difference model allows fast, efficient evaluation of multiple keyhole methods. For example, we evaluated over 1500 images throughout the course of this study. The ability to quickly determine the effect of various sequence parameters on image quality provides the image engineer with valuable information about trade-offs and compromises of a growing number of novel imaging techniques. It is hoped that further optimization will result in more efficient interventions that decrease patient risk and possibly the cost of the procedure. We believe this to be the first effort at using quantitative image quality modeling to evaluate fast interventional MR imaging. Our early experience shows the PDM to be an objective, promising tool for the evaluation of fast iMRI methods.

Appendix

The perceptual difference model is designed to mimic human visual response. Although it is based on the architecture of Daly's VDP,[30] many major features of our PDM are implemented using other descriptions from the literature, and we consider it sufficiently different to necessitate a full accounting of it here. The PDM attempts to quantitatively describe the visual differences between two images when viewed by a human observer. As shown in Fig. 2, the model has two inputs, a reference image, $I_1(x,y)$, and a processed or degraded ''test'' image, $I_2(x,y)$. The output, $O(x,y)$, is a spatial map that shows the magnitude of perceived image differences. It can be averaged over the entire image or over a region of interest to give a scalar measure of degradation.

The first block in the model, retinal luminance calibration, changes a digital image into values that can be interpreted as visual inputs. Display calibration parameters are used along with viewing distance and ambient light levels to convert gray scale values into observed luminance values. It has been established that the human visual system responds nonlinearly to input luminance, and models include a local amplitude nonlinearity,[30] a logarithmic model,[39] and various power laws, such as a cube root model.[40,41] In the PDM, we implement the cube root model, proposed by Mannos and Sakrison[41] and by McCann *et al.*[40] This is also consistent with the heavily used CIELAB and CIELUV models of luminance.[42,43] The luminance, *u*, is converted by applying $f(u)=u^{0.33}$ to each pixel in image space. A two-dimensional Fourier transform converts both input images to spatial frequency space, where viewing distance is used to convert spatial frequencies on the display, in units of cycles per millimeter, to a measure of subtended visual angle, in cycles per degree. In these units, the Nyquist frequency of the images ranged from approximately 12 to 22 cycles/deg depending on the actual viewing distance. We continue to refer to this as spatial frequency space.

The next block filters the input by the human spatial frequency contrast sensitivity function (CSF). The CSF response is due to the optics of the eye and neural processing. We use the rotationally invariant function described by Mannos and Sakrison[41] given below where *f* is the spatial frequency in cycles per degree,

$$\mathrm{CSF}(f)\equiv 2.6(0.192+0.114f)\exp[-(0.114f)^{1.1}]. \qquad \text{(A1)}$$

The CSF is normalized to produce a peak value of 1.0 at

$f = 8.0$ cycles/deg and a zero frequency intercept at 0.5 cycles/deg. This function has been shown to be in close agreement with a number of published measurements taken with sinusoidal gratings.[41] The CSF multiplies each input image in the frequency domain.

The next block of the PDM mimics the spatial frequency channels in the human visual system. It is well known from neurophysiology[44] and psychophysics[45] that the human visual system has specialized cell pathways that are selective for certain spatial frequencies and orientations. We implement a version of the Cortex transform, first described by Watson,[46] with modifications by Daly.[30] As described below, the Cortex transform combines two sets of filters, spatial frequency selective difference of mesa (DOM) filters and orientation selective fan filters, to produce 31 spatial frequency and orientation selective channels.

Each DOM filter is the difference of two, two-dimensional low-pass mesa filters with different cutoff frequencies, where the transition region is modeled with a Hanning window.[30] That is, the kth DOM filter is described as

$$\mathrm{DOM}_k(f) = \mathrm{mesa}(f)\big|_{f_{1/2}=2^{-(k-1)}} - \mathrm{mesa}(f)\big|_{f_{1/2}=2^{-k}}, \tag{A2}$$

where

$$\begin{aligned}\mathrm{mesa}(f) &= 1.0, \qquad \text{for } f < f_{1/2} - \frac{\omega}{2},\\ &= \frac{1}{2}\left[1+\cos\left(\frac{\pi(f - f_{1/2} + \omega/2)}{\omega}\right)\right],\\ &\qquad \text{for } f_{1/2} - \frac{\omega}{2} < f < f_{1/2} + \frac{\omega}{2},\\ &= 0.0, \quad \text{for } f > f_{1/2} + \frac{\omega}{2}.\end{aligned} \tag{A3}$$

In this manner, the five DOM filters can be described by their half-amplitude frequency, $f_{1/2}$, and their transition width, ω. The filter that encompasses the origin and the lowest frequencies in the image is called the baseband filter. Instead of using a mesa filter, which results in ringing, Daly proposed a truncated Gaussian described by

$$\begin{aligned}\mathrm{base}(f) &= \exp[-(f^2/2\sigma^2)], \quad \text{for } f < f_{1/2} + \frac{\omega}{2},\\ &= 0.0, \quad \text{for } f \geqslant f_{1/2} + \frac{\omega}{2},\end{aligned} \tag{A4}$$

and

$$\sigma = \frac{1}{3}\left(f_{1/2} + \frac{\omega}{2}\right), \tag{A5}$$

where $f_{1/2} = 2^{-K}$.

The Gaussian is truncated at 3σ. We use the filter set below that consists of six radial filters with $K = 6$.

$$\begin{aligned}\mathrm{DOM}_k(f) &= \mathrm{mesa}(f)_{f_{1/2}=2^{-(k-1)}} - \mathrm{mesa}(f)_{f_{1/2}=2^{-k}},\\ &\qquad \text{for } k = 1, K-2,\\ &= \mathrm{mesa}(f)_{f_{1/2}=2^{-(k-1)}} - \mathrm{base}(f)_{f_{1/2}=2^{-k}},\\ &\qquad \text{for } k = K-1.\end{aligned} \tag{A6}$$

We use $\omega = (2/3)*f_{1/2}$ so that the transition width of the filter depends on the half-amplitude frequency.

Six orientation selective fan filters use a Hanning window described as a function of angular degrees, θ, in the spatial frequency domain. The equation for fan filter l is

$$\begin{aligned}\mathrm{fan}_l(\theta) &= \frac{1}{2}\left[1+\cos\left(\frac{\pi[\theta - \theta_c(l)]}{\theta_\omega}\right)\right],\\ &\qquad \text{for } [\theta - \theta_c(l)] \leqslant \theta_\omega,\\ &= 0.0, \quad \text{for } [\theta - \theta_c(l)] > \theta_\omega,\end{aligned} \tag{A7}$$

$$l = (1, \ldots, L),$$

where $L = 6$, θ_ω is the angular transition width, and $\theta_c(l)$ is the center orientation angle given by $\theta_c(l) = (l-1)\theta_\omega - 90$. The transition width, θ_ω, is set equal to the angular spacing between filters and defined as $\theta_\omega = 30$ in the current implementation.

Finally, the 31 cortex channel filters are formed as the polar separable product of the DOM and fan filters,

$$\begin{aligned}\mathrm{cortex}_{k,l}(f,\theta) &= \mathrm{DOM}_k(f)\,\mathrm{fan}_l(\theta),\\ &\qquad \text{for } k = 1, \ldots, K-1, \; l = 1, \ldots, L,\\ &= \mathrm{base}(f), \quad \text{for } k = K,\end{aligned} \tag{A8}$$

where particular filters are denoted by k and l. A sum of all the cortex filters gives a constant frequency response equal to 1.0, showing that the process of dividing the frequency domain into channels is lossless. Fourier domain information for each image from the previous block is multiplied by each of the 31 filters and then the inverse Fourier transform is performed to produce 31 images, $I_{\mathrm{ch}}(x,y)$, each representing the information in one visual channel.

The next block converts the 31 spatial domain images from luminance contrast values. We implement local band-limited contrast as defined by Peli.[47] We believe this is the appropriate implementation by which to combine the cube root nonlinearity, defined above, with the local contrast defined by Peli. This method calculates contrast separately for each spatial frequency channel since human contrast sensitivity is highly dependent on spatial frequency. Additionally, to address variations in the local mean level across the image, it calculates contrast at each point in the image individually. The PDM is particularly well suited to this method because it is channelized. For every $I_{\mathrm{ch}}(x,y)$, there is a corresponding mean luminance image, $m(x,y)$. This is defined as the image that contains all spatial frequency energy below the DOM filter used to create I_{ch}. The local band-limited contrast, calculated for each of the 31 images, is then defined as

$$c(x,y)=\frac{I_{\mathrm{ch}}(x,y)}{m(x,y)}, \tag{A9}$$

where $m(x,y)>0$. For the baseband, $c(x,y)$ is calculated by dividing $I_{\mathrm{ch}}(x,y)$ by the mean luminance of the entire image.

Finally, the contrast units are combined using a ''Qnorm'' described by Lubin.[27] For each of the $n=31$ bands, contrast information, $c[I(x,y)]$, exists for each input image, I_1 and I_2. We can calculate a Qnorm distance measure as

$$O(x,y)=\left(\sum_{i=1}^{31}\{c_i[I_1(x,y)]-c_i[I_2(x,y)]\}^Q\right)^{1/Q}, \tag{A10}$$

where the inputs are I_1 and I_2, and Q is 2.4, as suggested by Lubin. The result is a spatial map of the distance values, $O(x,y)$, a direct measure of the perceptual differences at the given location. The measure is then averaged over the entire image or a region of interest to give the PDM error score.

Acknowledgments

This work was supported by a Whitaker fellowship [awarded to one of the authors (K.A.S.)] and by NIH Grants Nos. R01-HL48918 [awarded to another author (D.L.W.)] and R01-CA81431-01A1 [awarded to a third author (J.L.D.)]. One author (A.J.A.) was supported by a grant from the German Research Foundation (Grant No. DFG As 116-1).

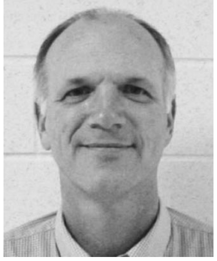
References

1. A. E. Burgess, ''Comparison of receiver operating characteristic and forced choice observer performance measurement methods,'' *Med. Phys.* **22**, 643–655 (1995).
2. P. Xue, C. W. Thomas, G. C. Gilmore, and D. L. Wilson, ''An adaptive reference/test paradigm with applications to pulsed fluoroscopy perception,'' *Behav. Res. Methods Instrum. Comput.* **30**(2), 332–348 (1998).
3. C. K. Abbey and F. O. Bochud, ''Modeling visual detection tasks in correlated image noise with linear model observers,'' in *Handbook of Medical Imaging*, Vol. 1 Physics and Psychophysics, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter, SPIE, Bellingham, WA (2000).
4. 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*, Vol. 1. Physics and Psychophysics, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter, SPIE, Bellingham, WA (2000).
5. J. S. Lewin, J. L. Duerk, V. R. Jain, C. A. Petersilge, C. P. Chao, and J. R. Haaga, ''Needle localization in MR-guided biopsy and aspiration: Effects of field strength, sequence design, and magnetic field orientation,'' *Am. J. Roentgenol.* **166**(6), 1337–1345 (1996).
6. J. S. Lewin, C. F. Connell, J. L. Duerk, Y. C. Chung, M. E. Clampitt, J. Spisak, G. S. Gazelle, and J. R. Haaga, ''Interactive MRI-guided radiofrequency interstitial thermal ablation of abdominal tumors: Clinical trial for evaluation of safety and feasibility,'' *J. Magn. Reson. Imaging* **8**(1), 40–47 (1998).
7. D. B. Tweig, ''The *k*-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods,'' *Med. Phys.* **10**(5), 610–621 (1983).
8. T. R. Brown, B. M. Kincaid, and K. Ugurbil, ''NMR chemical shift imaging in three dimensions,'' *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3523–3526 (1982).
9. T. A. Spraggins, ''Simulation of spatial and contrast distortions in keyhole imaging,'' *Magn. Reson. Med.* **31**(3), 320–322 (1994).
10. J. L. Duerk, J. S. Lewin, and D. H. Wu, ''Application of keyhole imaging to interventional MRI: A simulation study to predict sequence requirements,'' *J. Magn. Reson. Imaging* **6**(6), 918–924 (1996).
11. K. Hwang, J. S. Lim, M. Wendt, E. M. Merkle, J. S. Lewin, and J. L. Duerk, ''Improved device definition in interventional magnetic resonance imaging using a rotated stripes keyhole acquisition,'' *Magn. Reson. Med.* **42**, 554–560 (1999).
12. T. Parrish and X. Hu, ''Continuous update with random encoding (CURE): A new strategy for dynamic imaging,'' *Magn. Reson. Med.* **33**(3), 326–336 (1995).
13. K. A. Salem, J. S. Lewin, A. J. Aschoff, J. L. Duerk, and D. L. Wilson, ''Optimization of keyhole interventional MRI imaging using a human vision model,'' *Magn. Reson. Med.* (submitted).
14. M. Busch, A. Bornstedt, M. Wendt, J. L. Duerk, J. S. Lewin, and D. Gronemeyer, ''Fast 'real time' imaging with different *k*-space update strategies for interventional procedures,'' *J. Magn. Reson. Imaging* **8**(4), 944–954 (1998).
15. J. H. Gao, J. Xiong, S. Lai, E. M. Haacke, M. G. Woldorff, J. Li, and P. T. Fox, ''Improving the temporal resolution of functional MR imaging using keyhole techniques,'' *Magn. Reson. Med.* **35**(6), 854–860 (1996).
16. M. Wendt, M. Busch, R. Wetzler, Q. Zhang, A. Melzer, F. Wacker, J. L. Duerk, and J. S. Lewin, ''Shifted rotated keyhole imaging and active tip-tracking for interventional procedure guidance,'' *J. Magn. Reson. Imaging* **8**(1), 258–261 (1998).
17. J. E. Bishop, G. E. Santyr, F. Kelcz, and D. B. Plewes, ''Limitations of the keyhole technique for quantitative dynamic contrast-enhanced breast MRI,'' *J. Magn. Reson. Imaging* **7**(4), 716–723 (1997).
18. J. B. Weaver, Y. Xu, D. M. Healy, and J. R. Driscoll, ''Wavelet-encoded MR imaging,'' *Magn. Reson. Med.* **24**(2), 275–287 (1992).
19. L. P. Panych, P. D. Jakab, and F. A. Jolesz, ''Implementation of wavelet-encoded MR imaging,'' *J. Magn. Reson. Imaging* **3**(4), 649–655 (1993).
20. J. L. Duerk, D. H. Wu, Y. C. Chung, Z. P. Liang, and J. S. Lewin, ''A simulation study to assess SVD encoding for interventional MRI: Effect of object rotation and needle insertion,'' *J. Magn. Reson. Imaging* **6**(6), 957–960 (1996).
21. G. P. Zientara, L. P. Panych, and F. A. Jolesz, ''Dynamically adaptive MRI with encoding by singular value decomposition,'' *Magn. Reson. Med.* **32**(2), 268–274 (1994).
22. B. Girod, ''What's wrong with mean-squared error?'' in *Digital Images and Human Vision*, edited by A. B. Watson, MIT Press, Cambridge, MA (1993).
23. S. Winkler, ''Issues in vision modeling for perceptual video quality assessment,'' *Signal Process.* **78**, 231–252 (1999).
24. H. R. Wilson and J. R. Bergen, ''A four mechanism model for threshold spatial vision,'' *Vision Res.* **19**, 19–32 (1979).
25. C. J. van den Branden Lambrecht and O. Verscheure, ''Perceptual quality measure using spatio-temporal model of the human visual system,'' in *Proc. SPIE* **2668**, 450–461 (1996).
26. J. Lubin, ''Sarnoff JND vision model: Algorithm description and testing,'' (1997), ftp://ftp.its.bldrdoc.gov/dist/ituvidq/old2/jrg003.rtf
27. J. Lubin, ''A visual discrimination model for imaging system design and evaluation,'' in *Vision Models for Target Detection and Recognition*, edited by E. Peli, World Scientific, River Edge, NJ (1995).
28. W. B. Jackson, M. R. Said, D. A. Jared, J. O. Larimer, J. L. Gille, and J. Lubin, ''Evaluation of human vision models for predicting human-observer performance,'' *Proc. SPIE* **2708**, 64–73 (1997).
29. J. P. Johnson, J. Lubin, E. A. Krupinski, H. A. Peterson, H. Roehrig, and A. Baysinger, ''Visual discrimination model for digital mammography,'' *Proc. SPIE* **3663**, 253–263 (1999).
30. S. Daly, ''The Visual Differences Predictor: An algorithm for the assessment of image fidelity,'' in *Digital Images and Human Vision*, edited by A. B. Watson, MIT Press, Cambridge, MA (1993).
31. S. Daly, ''Method and apparatus for determining visually perceptible differences between images,'' U.S. Patent No. 906603 (1995).
32. V. Kayargadde and J.-B. Martens, ''Perceptual characterization of images degraded by blur and noise: Experiments,'' *J. Opt. Soc. Am. A* **13**(6), 1166–1177 (1996).
33. J.-B. Martens and L. Meesters, ''Image dissimilarity,'' *Signal Process.* **70**, 155–176 (1998).
34. R. M. Henkelman, ''Measurement of signal intensities in the presence of noise in MR images,'' *Med. Phys.* **12**(2), 232–233 (1985).
35. A. Macovski, ''Noise in MRI,'' *Magn. Reson. Med.* **36(3)**, 494–497 (1996).
36. R. N. McDonough and A. D. Whalen, *Detection of Signals in Noise*, 2nd ed., Academic, San Diego, CA (1995).
37. S. Westen, R. Lagendijk, and J. Biemond, ''Optimization of JPEG colour image coding using a human visual system model,'' *Proc. SPIE* **2657**, 370–381 (1996).
38. Video Quality Experts Group, ''Video quality experts group subjective test plan'' (1998).
39. C. F. Hall and E. L. Hall, ''A nonlinear model for the spatial characteristics of the human visual system,'' *IEEE Trans. Syst. Man Cybern.* **7**(3), 161–170 (1977).

40. J. J. McCann, S. P. McKee, and T. H. Taylor, ''Quantitative studies in retinex theory: A comparison between theoretical predictions and observer responses to the 'color Mondrian' experiments,'' *Vision Res.* **16**, 445–458 (1976).
41. J. L. Mannos and D. J. Sakrison, ''The effects of a visual fidelity criterion on the encoding of images,'' *IEEE Trans. Inf. Theory* **IT-20**(4), 525–536 (1974).
42. R. C. Carter and E. C. Carter, ''CIE L*u*v* color-difference equations for self-luminous displays,'' *Color Res. Appl.* **8**, 252–253 (1983).
43. CIE, *Colorimetry*, 2nd ed., Publication No. 15.2 (1986).
44. D. H. Hubel and T. N. Wiesel, ''Receptive fields, binocular integration and functional architecture in the cat's visual cortex,'' *J. Physiol. (London)* **160**, 106–154 (1962).
45. C. F. Stromeyer and B. Julesz, ''Spatial-frequency masking in vision: Critical bands and spread of masking,'' *J. Opt. Soc. Am. A* **62**, 1221–1232 (1972).
46. A. B. Watson, ''The cortex transform: Rapid computation of simulated neural images,'' *Comput. Vis. Graph. Image Process.* **39**, 311–327 (1987).
47. E. Peli, ''Contrast in complex images,'' *J. Opt. Soc. Am. A* **7**(10), 2032–2040 (1990).
48. C. E. Metz, B. A. Herman, and C. A. Roe, ''Statistical comparison of two ROC curve estimates obtained from partially-paired datasets,'' *Med. Decis Making* **18**, 110–121 (1998).

Kyle A. Salem is a PhD candidate in the Department of Biomedical Engineering at Case Western Reserve University, where he received his BSE in biomedical engineering in 1997. His thesis research focuses on the development of human visual system perceptual difference models for the evaluation of novel interventional MRI techniques. His other interests include the use of multiple imaging modalities to track drug transport *in vivo*, image processing, and image system optimization. He was awarded a fellowship by The Whitaker Foundation to fund his current research, and plans to complete his degree in the spring of 2002.

Jonathan S. Lewin received his undergraduate degree in chemistry from Brown University in 1981 and his Doctor of Medicine from Yale University in 1985. Following his internship in pediatrics at Yale–New Haven Hospital and residency in diagnostic radiology at University Hospitals of Cleveland, he completed a magnetic resonance imaging research fellowship at the Siemens Research and Development Center in Erlangen, Germany, a neuroradiology fellowship at the Cleveland Clinic Foundation, and additional training in head and neck radiology at the Pittsburgh Eye and Ear Hospital. Currently, he is professor and vice chairman for research and academic affairs at the Department of Radiology at Case Western Reserve University, with secondary appointments in the Departments of Oncology and Neurological Surgery, and is Director of the Division of Magnetic Resonance Imaging at the University Hospitals of Cleveland.



David L. Wilson is an associate professor of biomedical engineering and radiology at Case Western Reserve University. Previously, he worked for Siemens Medical Systems at locations in New Jersey and Germany. One of his interests is the use of quantitative image quality methods for engineering optimization of interventional x-ray fluoroscopy and MRI imaging. Another area of research is image analysis methods such as 3D registration and segmentation for interventional MRI guided, minimally invasive treatments of cancer. In particular, he is developing computer and animal experimental methods to compare the tissue response from histology to interventional MRI thermal ablation images and measurements of temperature history. Dr. Wilson has a record of NIH grant support and over 90 publications including over 40 refereed publications.

Jeffrey L. Duerk is a professor of radiology and biomedical engineering at Case Western Reserve University (CWRU) and the director of physics research at the Department of Radiology at University Hospitals of Cleveland. He received his BS and MS degrees in electrical engineering from Purdue University (1981) and Ohio State University (1983), respectively. After completion of his PhD in biomedical engineering at Case Western Reserve University in 1987, he worked as a clinical scientist for Picker International before joining the faculty at CWRU. His current research interests include the development of new methods for interventional MRI in cancer and cardiovascular disease. Dr. Duerk has over 80 refereed publications and a record of research grant support.

Andrik J. Aschoff received his Doctor of Human Medicine degree from the University of Ulm, Germany, in 1994. Following residencies in both internal medicine and diagnostic radiology at the University of Ulm, he completed a magnetic resonance imaging research fellowship at Case Western Reserve University in Cleveland, OH. He has also received additional training in neuro- and musculoskeletal radiology at RKU, Germany. Currently, he is an assistant professor in the Department of Diagnostic Radiology at University Hospitals of Ulm. He has published over 30 peer reviewed articles and is currently a member of the International Society of Magnetic Resonance in Medicine, the Radiological Society of North America, and the German Roentgen Ray Society.