Chapter in *Encyclopedia of Biomedical Engineering*, Elsevier, 2018

Perceptual Quality Assessment of Medical Images

Author Information

Hantao Liu
School of Computer Science and Informatics
Cardiff University
E-mail: liuh35@cardiff.ac.uk

Zhou Wang
Department of Electrical and Computer Engineering
University of Waterloo
E-mail: zhou.wang@uwaterloo.ca

Abstract

Today, healthcare professionals are viewing medical images in a variety of environments. The technologies and methodologies used to acquire, process, store, transmit and display images vary, and consequently, the ultimate visual information received by the clinicians differs significantly in perceived quality. Visual signal distortions, such as various types of noise and artifacts arising in medical image acquisition, processing, compression and transmission, affect the perceptual quality of images and potentially impact diagnoses. To optimize clinical practice, we need to understand human perception of medical image quality in practical settings, and then use what is learned to develop useful solutions for improved image quality and better image-based diagnoses. This chapter focuses on the methodologies used to measure the perceptual quality of medical images using magnetic resonance (MR) image acquisition and computed tomography (CT) image compression as examples, where modern digital image processing technologies and statistical analysis approaches play important roles in helping with both subjective visual testing and objective quality predictions.

Keywords

Computed tomography, diagnostically lossless compression, ghosting artifact, image compression, image quality assessment, Kolmogorov-Smirnov statistic, magnetic resonance imaging, mean squared error, noise artifact, receiver operating characteristic, structural similarity index, structured artifact, unstructured artifact

Quality assessment in medical image acquisition

Quality degradation of medical images often starts at the acquisition stage. For example, magnetic resonance (MR) imaging in practice is vulnerable to a variety of artifacts, which degrade the perceived quality of images and, consequently, may cause inefficient and/or inaccurate diagnoses. Sources of artifacts in MR imaging include non-ideal hardware characteristics, intrinsic tissue properties and their possible changes during scanning, assumptions underlying the data acquisition and image reconstruction processes, and a poor choice of scanning parameters. To minimize or eliminate these artifacts, many correction procedures have been developed. These methods typically involve one or more of the following strategies: improvement of hardware and scanning protocols, optimization of scan parameters and pulse sequences, and advanced digital post-processing. Nonetheless, reducing artifacts in MR imaging is not straightforward, and in practice, it is still a challenge to achieve optimal image rendering from the user's point of view. One reason is that strategies dealing with one type of artifact may induce another. As a consequence, optimization of these strategies requires a comprehensive understanding of the relative annoyance of different types of artifacts to perceived image quality.

Progress has been made in studying the causes and characteristics of artifacts in MR images. The first step is to classify these artifacts so that they can be recognized from relevant features. In general, the artifacts may be classified into two categories – unstructured artifact as random noise, and structured artifact as any type of coherent artifact that represents the anisotropy of the spectral content of local structure of the object being scanned. Ghosting, which is a cross-talk artifact generating a lower-intensity double image, spatially shifted with respect to the original content, is one example of structured artifact. Random noise can be further classified into white noise and colored noise, according to its spectral density – white noise has a flat frequency spectrum, whereas the frequency spectrum of colored noise is not flat. A similar classification can be made for structured artifacts, i.e., we may make a distinction between a white structured artifact and a colored structured artifact. The example of ghosting – explained above – can be considered as a colored structured artifact with the same distribution in the frequency spectrum as the object being scanned. "White" ghosting – referred to as edge ghosting – can be obtained by making the frequency spectrum flatter, e.g., by adding the gradient of the originally scanned object as a double image to the original. Based on the analysis of the power spectral density of a thin-slice two-dimensional MR image, it can be shown that the power spectrum of the gradient of any line of an MR image is rather flat, and thus for the purpose of our study is approximated to be "white". Figure 1 illustrates the four types of artifacts on an exemplary MR image.

To what extent a given artifact presenting with a certain energy reduces the perceived image quality is a challenging problem that may be addressed by measuring the relative impact of four types of artifacts: a white unstructured artifact (i.e., white noise), a colored unstructured artifact (i.e., colored noise), a white structured artifact (i.e., edge ghosting), and a colored structured artifact (i.e., ghosting).

Simulation of MR Imaging Artifacts

To be able to vary the four types of artifacts – namely ghosting, edge ghosting, white noise, and colored noise – in a controlled process, they are simulated separately at different levels of energy, and then linearly added to the original image content, as illustrated in Figure 2. A benchmark energy level (BEL) is defined (illustrated as the energy level L5 in Figure 2). For an original image of size M×N (height × width) pixels with intensity of the simulated ghosting artifact $I_g(i, j)$ ($i \in [1, M]$, $j \in [1, N]$), its BEL is calculated as:

$$BEL = \sum_{i=1}^{M} \sum_{j=1}^{N} I_g(i, j)^2$$

The BEL is determined separately for each original image, and is defined by the amount of energy in a typical ghosting artifact for that particular content. As such, a ghosting artifact is always generated first in the simulation process. Based on the BEL defined for ghosting, the other levels of energy are determined by reducing the BEL successively by 20%, resulting in 0.8 × BEL for energy level L4, 0.6 × BEL for energy level L3, 0.4 × BEL for energy level L2, and 0.2 × BEL for energy level L1, respectively.

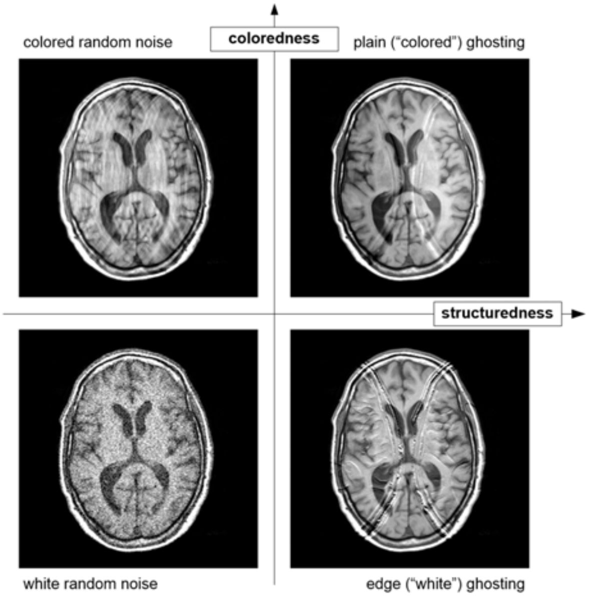


Fig. 1. Four types of artifacts in a typical MR image. The horizontal axis indicates the "structuredness" of the artifact: the two left quadrants refer to the unstructured artifacts (i.e. random noise), and the two right quadrants refer to the structured artifacts (i.e. ghosting). The vertical axis indicates the colored-ness of the artifact: the two top quadrants refer to the colored artifacts, and the two bottom quadrants refer to the white artifacts.

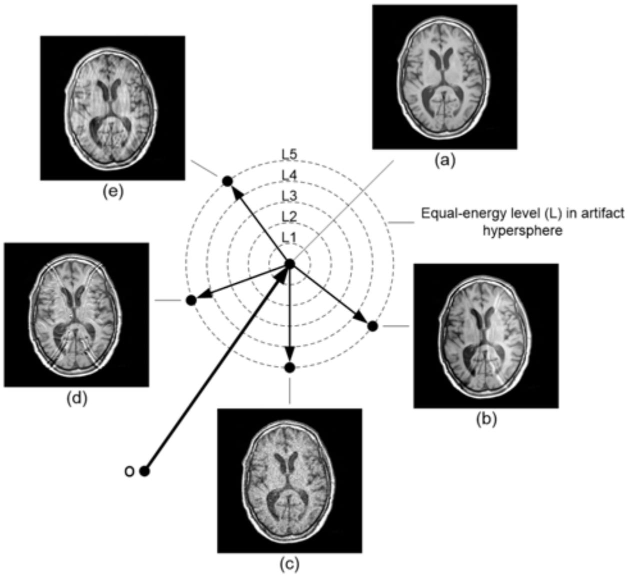


Fig. 2. Images with the same level of energy in the artifact added to the original image constitute a hypersphere in the image space: (a) original image, (b) image with ghosting, (c) image with white noise, (d) image with edge ghosting, and (e) image with colored noise. Five different levels of energy, i.e. L1, L2, L3, L4, L5, are used here.

Subjective Experiments

The goal of the subjective experiments is to quantitatively measure how the four types of artifacts, applied at the same energy level in the distortion, affect the perceived quality of MR images. To this end, perception experiments have been performed with clinical application specialists (mainly qualified clinical medical physicists). Here we investigate the relative impact of structured versus unstructured artifacts on the perceived quality of MR images. The experiments consists of two *parts*: one to compare images degraded with ghosting to those degraded with white noise, and the other to compare images degraded with edge ghosting to those degraded with colored noise.

The source MR images used in the experiments are chosen to have high quality in terms of resolution, artifacts and signal-to-noise ratio. Three original MR images are selected: two images of a brain (i.e., referred to as "brain_1" and "brain_2") and one image of a liver (i.e., referred to as “liver”). The three source images are shown in Figure 3. Each source image is first distorted with ghosting at the energy level BEL, and subsequently, edge ghosting, white noise and colored noise are applied at the same energy level. The added energy (i.e. at BEL) of ghosting, edge ghosting, white noise and colored noise is then downscaled with factors of 4/5, 3/5, 2/5, and 1/5, respectively, resulting in four new energy levels for each artifact type. By doing so, each original image is distorted with 5 levels of simulated ghosting, edge ghosting, white noise and colored noise, respectively. Hence, the test database existed of 30 stimuli (i.e. 3 originals × 5 energy levels × 2 types of artifacts) *per part*, and so 60 stimuli in total.

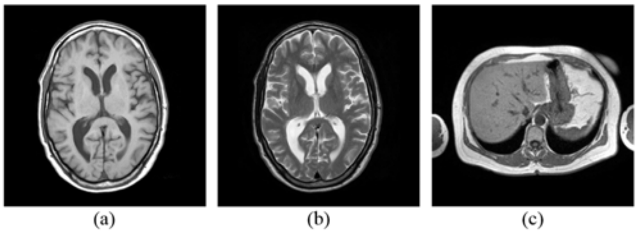


Figure 3. Source images: (a) “brain_1”, (b) “brain_2”, and (c) “liver”.

The experiments should be conducted in a controlled environment similar to a typical radiology reading room environment with low surface reflectance and approximately constant ambient light. No image adjustment (zoom, window level) should be allowed. The participants should be clinical scientists or application specialists of both genders. To score perceived quality a simultaneous-double-stimulus (SDS) method [15] may be used, for which the subjects are requested to score the quality for each stimulus in the presence of the original image as a reference. In scoring image quality, the two stimuli, i.e. the original at the left-hand side and the test stimulus at the right-hand side are displayed side by side on the same screen. The scoring scale ranges from 0 to 100, and included additional semantic labels (i.e. “Bad”, “Poor”, “Fair”, “Good” and “Excellent”) at intermediate points. Subjects are requested to assess the quality of the test stimulus with respect to the quality of the reference by moving the slider on the scoring scale.

Before the start of each experiment, written instructions about the procedure of the experiment (i.e. explaining the type of assessment, the scoring scale and the timing) are given to each subject. Subsequently, a set of ten images covering the same range of artifact annoyance as used in the actual experiment is presented in order to familiarize them with the impairments and the scoring scale. In a next step, six representative stimuli are shown one by one and the participant scores their quality on the scoring scale. The images used in this training part of the experiment should be independent of those used in the formal experiment. After training, the test stimuli are shown one by one in random order in a separate session. Each stimulus is shown once, and the participants are allowed to take as much time as they need to assess the quality of each stimulus.

Data Analysis

Data analysis is an important step in understanding the experimental results. First, a simple outlier detection and subject exclusion procedure is applied to the raw scores. An individual score for an image is considered to be an outlier if it is outside an interval of two standard deviations around the mean score for that image. All scores of a subject are rejected if more than 20 percent of the scores are outliers. After having applied the outlier removal and subject exclusion procedure, the scores of the remaining subjects are normalized towards the same mean and standard deviation using z-scores:

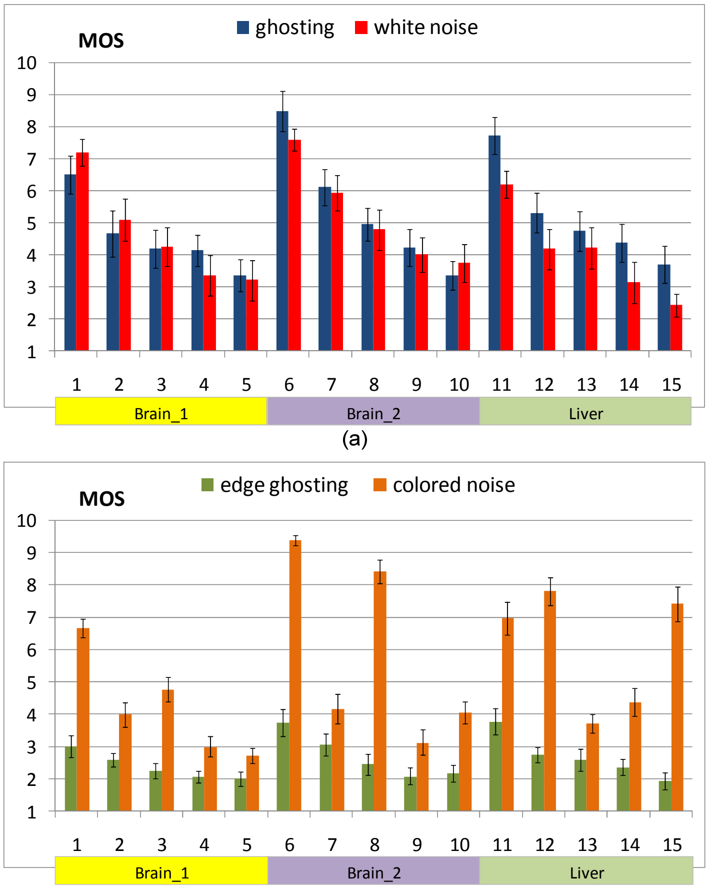
$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}$$

where r_{ij} and z_{ij} indicate the raw score and z-score of the *i*-th subject and *j*-th image, respectively. μ_i is the mean of the raw scores over all images scored by subject *i*, and σ_i is the corresponding standard deviation. These scores are averaged across subjects to yield a mean opinion score (MOS) for the *j*-th image, i.e.

$$MOS_j = \frac{1}{S} \sum_{i=1}^{S} z_{ij}$$

where *S* is the total number of subjects. To make the final scores easier to interpret, the resulting MOSs were linearly remapped to the range of [1, 10]. The MOSs and their corresponding error bars are illustrated in Figure 4. Figure 4(a) indicates that the difference in perceived quality between degradations with ghosting and white noise is in general small. Whether at the same energy level either ghosting or white noise mostly affects the overall quality tends to depend on the distortion level and image content. For the source image “liver”, the added white noise consistently results in a lower image quality than the added ghosting (see stimuli referred to as 11-15 in Figure 4(a)). A similar consistency, however, is not found for the two brain images, i.e. “brain_1” and “brain_2”.

Figure 4(b) shows that the quality of an MR image degraded by colored noise is consistently scored higher than that by edge ghosting. This suggests that the perceived quality is largely reduced when changing the signal distortion from unstructured colored noise to structured edge ghosting, even for the same level of energy in the distortion. In addition, we can observe a trend from the comparison of the four types of artifacts that when either ghosting, white noise or edge ghosting is added to a source image, the perceived image quality monotonously decreases with the energy in the distortion; this, however, is not the case for colored noise, for which the resulting quality may jump up and down as a function of distortion level.



(a)

(b)

Figure 4. The MOS resulting from the subjective image quality assessment: (a) images degraded with ghosting and white noise, and (b) images degraded with edge ghosting and colored noise. The numbers on the horizontal axis refer to the stimuli: numbers 1-5 for image “brain_1” with increasing level of distortion, numbers 6-10 for image “brain_2”, and numbers 11-15 for image “liver”. Each number corresponds to two bars; one for ghosting (or edge ghosting) and one for white noise (or colored noise), with each the same energy in the signal distortion. The error bars indicate the 95% confidence interval.

Statistical analysis is performed on the observed tendencies with an ANOVA (Analysis of Variance) per graph/part of the experiment separately. In each case, the perceived quality is selected as the dependent variable, the image content, artifact type and energy level as fixed independent variables and the participants as random independent variable. All 2-way interactions of the fixed variables are included in the analysis as well. The results for images degraded with ghosting and white noise are summarized in Table 1 (including F-statistic and its degrees of freedom and significance p-value), and show that image content, artifact type and energy level have a significant effect on perceived quality. On average, images affected with ghosting are scored higher in quality than images affected with white noise (<MOS> for ghosting = 5.05, <MOS> for white noise = 4.61). The post-hoc analysis on image content shows that the viewers score the image “brain_2” (<MOS>=5.31) on average statistically significantly higher than the images “brain_1” (<MOS>=4.59) and “liver” (<MOS>=4.59). Also the interaction between image content and artifact type is significant, which implies that the difference in quality between the two types of artifact is not the same for the three images.

The results for images degraded with edge ghosting and colored noise are summarized in Table 2, where all main effects and interactions appear to be highly statistically significant. Overall images degraded with colored noise (<MOS>=5.37) are scored higher in quality than images degraded with edge ghosting (<MOS>=2.59). The post-hoc analysis on the image content indicates that the image "brain_1" (<MOS>=3.04) received statistically significantly lower quality scores than the other two images (<MOS>=4.26 for "brain_2" and <MOS>=4.37 for "liver"). The interaction between image content and artifact type is caused by the fact that the quality difference between images is much larger for the colored noise artifact than for the edge ghosting artifact. The interaction between artifact type and energy level is significant since the quality monotonically decreases with increasing energy level for the edge ghosting artifact, but not for the colored noise. In the latter case, the perceived quality fluctuates with increasing energy level. This phenomenon also explains the significant interaction between image content and energy level.

TABLE I

RESULTS OF THE ANOVA FOR EXPERIMENT 1 TO EVALUATE THE EFFECT OF GHOSTING AND WHITE NOISE ON THE DIFFERENT IMAGES

	F-value	Df	p
Image content	5.82	2	0.003
Artifact type	18.81	1	<0.001
Energy level	50.50	4	<0.001
Participant	41.71	13	<0.001
Image content x Artifact type	3.00	2	0.051
Artifact type x Energy level	0.163	4	0.957
Image content x Energy level	0.566	8	0.806

TABLE II

RESULTS OF THE ANOVA FOR EXPERIMENT 1 TO EVALUATE THE EFFECT OF EDGE GHOSTING AND COLORED NOISE ON THE DIFFERENT IMAGES

	F-value	Df	p
Image content	20.39	2	<0.001
Artifact type	386.30	1	<0.001
Energy level	38.50	4	<0.001
Participant	38.13	14	<0.001
Image content x Artifact type	12.02	2	<0.001
Image content x Energy level	10.89	8	<0.001
Artifact type x Energy level	10.74	4	<0.001

In conclusion, by investigating the relative impact on perceived image quality of four distortion types (i.e. ghosting, edge ghosting, white noise and colored noise), we find that the impact of the artifacts on image quality strongly depends on the specific content of the MR image. When neglecting this dependency (i.e. interactions with energy level and image content), in general "unstructured" artifacts deteriorate quality less than "structured" artifacts. This study provides an example of how insights about perceptual image quality may be gained in the context of medical image acquisition, and findings in such studies may be embedded in real-world MR imaging systems to optimize the image rendering to the perception of users.

Quality assessment in medical image compression

With the explosive growth of medical digital image data being acquired every day, the medical communities have gradually recognized the need of effective methods of storing and transmitting medical images of large volumes. There has also been a general acknowledgement that lossless compression techniques, with low rates of data compression, are no longer enough to achieve the adequate compression efficiency desired in practice. Therefore, it is necessary to consider substantially higher compression rates using lossy compression methods. However, lossy compression involves loss of information and possibly visual quality, thus it becomes essential to be able to determine the degree to which a medical image can be compressed before its diagnostic quality is compromised.

Currently, the radiological community has not yet accepted a single standard methodology for the quality assessment of medical images. Recommended compression ratios (CR) for various modalities and anatomical regions have been published. To date, these recommendations have been based on experiments in which radiologists subjectively score the diagnostic quality of compressed images. There are several drawbacks of such an approach. First, subjective testing is time-consuming and expensive. As a result, only a small set of images can be properly assessed in a typical test. Second, it is difficult to rely on subjective testing in the design and optimization of automated image compression and transmission systems. Therefore, it is highly desirable to find computational models and algorithms that can be used for objectively assessing the diagnostic quality of compressed medical images.

In the literature of image processing and image compression, the quality of an image are characterized objectively in several ways. Practitioners most often employ the mean squared error (MSE) and its close relative, peak signal-to-noise ratio (PSNR), even though MSE and PSNR are known to correspond poorly to visual quality. A more recent image fidelity measure, the structural similarity index (SSIM), measures the similarity between two images by combining three components of the human visual system – luminance, contrast and structure. Assuming one of the images being compared is of pristine quality, the SSIM result has been shown to be a much more reliable prediction of perceived image quality of the other image in a wide range of applications. While SSIM is becoming popular in many other image processing fields, its accuracy and reliability in assessing medical images are yet to be thoroughly tested.

In this section, we discuss a study that examines whether compression ratio (CR), MSE, quality factor (QF, the input parameter to JPEG compression algorithm), and SSIM actually serve as reliable indicators of the diagnostic quality of medical images. This is done by comparing the quality predictions of these metrics with subjective tests that involve radiologists.

Subjective Experiment

A subjective experiment is designed in order to assess the quality prediction performance of the image quality assessment being examined. The experiment employs five neurological and five upper body slices extracted from the Cancer Imaging Archive. These images are first windowed according to their default settings to reduce their bit-depth from 16 to 8 bits per pixel (bpp). Each of the resulting 512 x 512 pixel, 8 bpp images are compressed at five compression ratios using both the JPEG and JPEG2000 compression algorithms. Preliminary visual observations are used to select the compression ratios employed in the experiment. An image viewer software is constructed specifically for this study in order to provide an easy-to-use graphical interface for the radiologists. The viewer displays a compressed image beside its uncompressed counterpart. The compressed images are presented in random order. During the course of the experiment, each compressed image is presented twice to each radiologist, but without the radiologists' knowledge. The subjects are not made aware of the compression ratios or quality factors used to generate the compressed images. Two buttons were placed at the bottom of the user interface: acceptable and unacceptable. In the experiment, the radiologist subjects are instructed to flag an image as unacceptable in the case they believe there is any noticeable distortion that could have any impact on diagnostic tasks. The experiment is designed to last about one hour for each radiologist subject.

Objective Quality Metrics

Let *f* denote an *M* × *N* digital image and *g* its compressed counterpart. The standard measure of error between *f* and *g* is the Mean Squared Error (MSE), defined as

$$MSE\,(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[f(i,j)-g(i,j)]^2 ,$$

where M and N are the height and width of the images, respectively. The MSE essentially defines a distance between *f* and *g*. The more distortion there is in the compressed image *g*, the higher the MSE value (If *f* and *g* are identical, then MSE = 0). The SSIM index between two images *f* and *g* is obtained by computing the following three terms, including the mean

$$\mu_f = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} f(i,j) ,$$

the variance

$$\sigma_f^2 = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[f(i,j)-\mu_f]^2 ,$$

and the covariance

$$\sigma_{fg} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [f(i,j) - \mu_f][g(i,j) - \mu_g] .$$

These terms are combined as follows to compute the SSIM index between images *f* and *g*:

$$SSIM(f,g) = \left(\frac{2\mu_f \mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \right) \left(\frac{2\sigma_f \sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \right) \left(\frac{\sigma_{fg} + C_3}{\sigma_f \sigma_g + C_3} \right) .$$

The SSIM index ranges between -1 and 1. It measures the similarity between *f* and *g*. The closer *f* and *g* are to each other, the closer SSIM is to the value 1. If *f* and *g* are identical, then SSIM = 1. The (non-negative) parameters C_1, C_2 and C_3 are stability constants of relatively small magnitude. For natural images, there are some recommended default values, and the question of optimal values for medical images is still an open one. The smaller the values of these constants, the more sensitive the SSIM index is to small image textures such as noise. Note that in the special case $C_3 = C_2/2$, the following simplified, two-term version of the SSIM index is obtained:

$$SSIM(f,g) = \left(\frac{2\mu_f \mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \right) \left(\frac{2\sigma_{fg} + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \right) .$$

When structural fidelity is of the main concern, one may focus on only the second term, i.e., the structure term. Note that MSE is an error measure (the lower the better quality) while SSIM is a similarity measure (the higher the better quality). In order to be able to compare their results more conveniently, we define the following quantity,

$$SMSE(f,g) = 1 - \frac{MSE(f,g)}{D} ,$$

where *D* is a constant. Using this definition, we now have that if *f* and *g* are “close”, then both SSIM and SMSE are near 1.

Another variation in the computation of SSIM is the local SSIM, for which one can employ any or all of the above formulas to compute SSIM values between corresponding local windows of two images. When such local windows slide pixel by pixel across the image, one obtain an SSIM quality map between two images *f* and *g* on a pixel-by-pixel basis – the map reveals local image similarities/differences between images *f* and *g*. A total SSIM score may then be computed by averaging over all the local SSIM values.

Data Analysis

Data analysis here mainly aims to examine how well different “image quality indicators”, e.g., compression ratio, MSE, quality factor, SSIM, compare to the subjective assessments of image quality by radiologists. The receiver operating characteristic (ROC) curve is a common tool for assessing the performance of a classifier in medical decision making. ROC curves illustrate the trade-off of true positives versus false positives as the discriminating threshold is varied. For our purpose, we perform ROC analysis by assuming the "ground truth" as whether or not a compressed image is acceptable or unacceptable by radiologists, and by defining the following terms:

- P (total positive) = FP + TP and N (total negative) = TN + FN: These refer to radiologists' subjective opinions, which represent the True class. On the other hand, P0 and N0 belong to the Hypothesis class which, in our experiment, corresponds to a given quality assessment method, i.e., SSIM, MSE, quality factor, and compression ratio.
- TP (true positives): images that are acceptable to both radiologists and a given quality assessment method.
- TN (true negatives): images that are unacceptable to both radiologists and a given quality assessment method.
- FN (false negatives): images that are acceptable to radiologists but unacceptable to a quality assessment method.
- FP (false positives): images that are unacceptable to radiologists but acceptable to a given quality assessment algorithm.

The values of all these terms are determined by a discrimination threshold that vary between 0 and 1. Each threshold value generates a point on the ROC curve which corresponds to a pair of specificity (SP) and sensitivity (SE) values given by

FPR (false positive rate) = FP/N = 1 − SP (specificity)
TPR (true positive rate) = TP/P = SE (sensitivity)

Let the threshold value changes from 0 to 1, we obtain an ROC curve for each image quality model being tested. Performance measures are then computed based on the ROC curves.

Area Under Curve (AUC) Test: The AUC can be computed by integrating the ROC curve using the trapezoidal rule. Larger AUC values correspond to better performance. It is possible that two ROC curves cross. In this special situation, one method might demonstrate better performance for some threshold values whereas another method behaves better for other values. In this case, a single AUC may not be the best performance predictor.

Kolmogorov-Smirnov (KS) Test: Given two cumulative probability distributions $P_1(x)$ and $P_2(x)$, their KS statistic is defined as

$$KS\,(P_1, P_2) = \sup_x \,|\, P_1(x) - P_2(x)\,|\,.$$

In our study, P_1 and P_2 are the cumulative distributions of positive and negative radiologists' responses, respectively. The larger the difference between the two distributions, the better the performance of a given model. For a given threshold s' in [0, 1], we have the following relations:

Cumulative Probability Distribution of negatives = TN/(TN + FP) = 1 – FPR;
Cumulative Probability Distribution of positives = FN/(FN +TP) = 1 - TPR.

Thus, the KS statistic translates to

$$KS = \sup_x |TPR(x) - FPR(x)| .$$

A related idea is the Youden index. For a given discriminating threshold *s*, the Youden index is given by *Y*(*s*) = *TPR*(*s*) − *FPR*(*s*). Now suppose that the maximum value of *Y*(*s*) occurs at $s = s_0$. Then we have

$$KS = Y(s_0) .$$

The data points accumulated in the subjective experiment include the two image types (brain CT and body CT) and the two compression methods (JPEG and JPEG2000). Table III shows the AUC comparison results of QF, CR, MSE and SSIM for all images as well as the breakdown results for JPEG images only, JPEG2000 images only, brain CT images only, and body CT images only. Such analysis in terms of image types is particularly important since different classes of images compressed by different compression algorithms possess different characteristics which may yield different types and levels of compression artifacts. In all cases, the highest AUC values correspond to the SSIM index quality measure. In Table IV, the analysis is further split into four cross-cases of JPEG-Brain CT images, JPEG-body CT images, JPEG2000-brain CT images, and JPEG2000-body CT images, respectively. Once again, the SSIM index quality measure consistently yields the highest AUC values. This suggests that of the four quality measures under comparison, SSIM performs the best in modeling radiologists' subjective assessments of compressed images when the AUC is used as a performance indicator.

TABLE III

AUC PERFORMANCE COMPARISON OF CR (COMPRESSION RATIO), QF (QUALITY FACTOR, FOR JPEG COMPRESSION ONLY), MSE AND SSIM

Quality Model	JPEG	JPEG2000	Brain CT	Body CT	All
QF	0.9401	-	-	-	-
CR	0.8372	0.7573	-	-	-
MSE	0.9101	0.8691	0.8524	0.9226	0.8900
SSIM	**0.9485**	**0.9330**	**0.9447**	**0.9389**	**0.9471**

TABLE IV

BREAKDOWN AUC PERFORMANCE COMPARISON OF CR (COMPRESSION RATIO), QF (QUALITY FACTOR, FOR JPEG COMPRESSION ONLY), MSE AND SSIM

Quality Model	JPEG-Brain CT	JPEG-Body CT	JPEG2000-Brain CT	JPEG2000-Body CT
QF	0.7424	0.9332	-	-
CR	0.6818	0.8926	-	-
MSE	0.7424	0.8749	0.8859	0.8750
SSIM	**0.7828**	**0.9492**	**0.9204**	**0.9577**

Table V reports the KS analysis results when all applicable images are used for evaluation. As expected, the KS statistical results, i.e. the degree of separation between acceptable and unacceptable images to radiologists, are consistent with those of the AUC analysis. The largest KS value is achieved by SSIM (81%), followed by JPEG QF (78%) and MSE (64%), and CR yields the lowest KS value (60%). Notice that CR and QF are applied to JPEG compressed images only, while MSE and SSIM are applied to all images employed in the subjective test.

TABLE V
KS PERFORMANCE COMPARISON OF CR (COMPRESSION RATIO), QF (QUALITY FACTOR, FOR JPEG COMPRESSION ONLY), MSE AND SSIM

Quality Model	Test Images	K-S Statistics
QF	JPEG only	77.65%
CR	JPEG only	59.68%
MSE	All	64.40%
SSIM	All	**81.09%**

In summary, using both the AUC (area under ROC curve) and the KS (Kolmogorov-Smirnov) analyses, the current results indicate that compression ratio (CR) demonstrates the poorest performance of the four quality measures being examined. Quality factor (QF) provides moderately reasonable quality predictions on JPEG images, but since it is a parameter used to control JPEG compression, it is applicable to JPEG compressed images only and will not generalize to other compression methods or other types of image distortions. Furthermore, MSE performed inconsistently as an indicator of visual/diagnostic quality. Finally, among the four image quality measures, SSIM shows the best performance, i.e. SSIM provides the closest match to the subjective assessments by the radiologists.

Summary and Remarks

With the rapid growth of digital image acquisition, processing, transmission and display technologies in medical imaging field, it has become ever more important to understand how such technologies affect the perceived image quality, which may have strong impact on the diagnostic values of these images. This chapter provides a basic introduction of the methodologies that have been used to measure the perceptual quality of medical images using image acquisition and image compression as examples. The processing and measurement of medical images are often different from those of typical natural images, because medical imaging applications often involve significant domain knowledge that needs to be fully understood and taken into consideration in both the quality assessment and data analysis processes. These are clearly exemplified in the MR image acquisition and CT image compression applications elaborated in the current chapter.

Quality assessment of medical images is still at a fast evolving stage, and by no means has this chapter provided a comprehensive coverage of all the problems and methodologies in the field. It is worth noting that promising technologies, such as the SSIM index as a novel objective image quality model discussed in this chapter, is gaining significant attention recently. The potential benefits of developing, validating and deploying such objective image quality assessment methods are not limited to monitoring the acquisition, storage, communication, processing, and display of medical images for quality control purposes, but also to the optimal design of novel medical imaging methods and systems that could deliver even better image quality in more cost-effective ways than what we have in the current systems.

Further Readings

Barish M. A. and Jara H. (1999). Motion artifact control in body MR imaging. Magnetic Resonance Imaging Clinics of North America7, 289-301.

Clark J. A. and Kelly W. M. (1988). Common artifacts encountered in magnetic resonance imaging. Radiologic Clinics of North America 26, 893-920.

Hedley M. and Yan H. (1992), Motion artifact suppression: a review of post-processing techniques. Magnetic Resonance Imaging 10, 627-635.

ITU-R Recommendation BT.500-11 (2002). Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland.

Kowalik-Urbaniak I., Brunet D., Wang J., Vrscay E. R., Wang Z., Koff D. A. , Koff N. , and Wallace B. (2014). The quest for 'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images. SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, San Diego, CA.

Krupinski E. A. (2000). The importance of perception research in medical imaging. Radiation Medicine 18, 329-334.

Liu H., Klomp N. and Heynderickx I. (2010). A No-Reference Metric for Perceived Ringing Artifacts in Images. IEEE Trans. on Circuits and Systems for Video Technology 20, 529-539.

Mirowitz S. A. (1999). MR imaging artifacts. Challenges and solutions. Magnetic Resonance Imaging Clinics of North America 7, 717-732.

Pusey E., Lufkin R. B., Brown R. K., Solomon M. A., Stark D. D., Tarr R. W. and Hanafee W. N. (1986). Magnetic resonance imaging artifacts: mechanism and clinical significance. Radiographics 6, 891-911.

Reeder S. B., Altar E., Bolster B. D. and McVeigh E. R. (1997). Quantification and Reduction of Ghosting Artifacts in Interleaved Echo-Planar Imaging. Magnetic Resonance Imaging 38, 429–439.

Smith T. B. and Nayak K. S. (2010). An overview of MRI artifacts and correction strategies. Imaging in Medicine 2, 445-457.

van Overveld I. (1995). Contrast, noise, and blur affect performance and appreciation of digital radiographs. Journal of Digital Imaging 8, 168-179.

Wang Z. and Bovik A. C. (2009). Mean squared error: love it or leave it? - A new look at signal fidelity measures. IEEE Signal Processing Magazine 26, 98-117.

Wang Z. and Bovik A. C. (2006), Modern Image Quality Assessment, in syntheses lectures on Image, Video and Multimedia Processing, Morgan & Claypool Publishers.

Wang Z., Bovik A. C., Sheikh H. R., and Simoncelli E. P. (2004). "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing 13, 600-612.

Willis C. E., Thompson S. K., Shepard S. J. (2004). Artifacts and misadventures in digital radiography. Applied Radiology 33.