Quality Assessment in Magnetic Resonance Images

Neelam Sinha & A.G. Ramakrishnan*

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

*Address all correspondence to Neelam Sinha, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India; neel.iam@gmail.com, ramkiag@ee.iisc.ernet.in

ABSTRACT: Assessing quality of medical images is critical because the subsequent course of actions depend on it. Extensive use of clinical magnetic resonance (MR) imaging warrants a study in image indices used for MR images. The quality of MR images assumes particular significance in the determination of their reliability for diagnostics, response to therapies, synchronization across different imaging cycles, optimization of interventional imaging, and image restoration. In this paper, we review various techniques developed for the assessment of MR image quality. The reported quality indices can be broadly classified as subjective/objective, automatic/semi-automatic, region-of-interest/non-region-of-interest–based, full-reference/no-reference and HVS incorporated/non-HVS incorporated. The trade-off across the various indices lies in the computational complexity, assumptions, repeatability, and resemblance to human perception. Because images are eventually viewed by the human eye, it is found that it is important to incorporate aspects of human visual response, sensitivity, and characteristics in computing quality indices. Additionally, no-reference metrics are the most relevant due to the lack of availability of a golden standard against which images could be compared. Techniques that are objective and automatic are preferred for their repeatability and to eliminate avoidable errors due to factors like stress, which arise in human intervention.

KEY WORDS: MR image, quality metric, subjective image quality, objective image quality, HVS-based metric, full reference metric, no-reference metric

I. INTRODUCTION

Assessment of image quality has been a very big challenge for the computer-vision community. It is especially true for medical images because they are generally monochromatic and noisy (e.g., ultrasound, computed tomography [CT], X-ray, and magnetic resonance [MR] images). Typically, each imaging modality consists of an electromechanical system, with its own acquisition characteristics, that manifest in the images generated. In MR imaging, data are acquired in the spatial frequency domain, which is Fourier transformed to obtain the desired image. The clinical popularity of MR imaging is due to its rich contrast for soft tissues and the availability of arbitrary imaging planes, which make it suitable for various applications in musculoskeletal imaging and neuroimaging. Additionally, unlike X-ray and CT, this imaging modality does not use any harmful, ionizing radiation. The most desired feature in medical images is to be able to clearly distinguish each structure based on its chemical and physical properties. This feature is available in MR imaging because the modality allows great flexibility to control the contrast among the various soft tissues by appropriately setting imaging parameters such as

ABBREVIATIONS

AF, autofocusing; **CNR,** contrast to noise ratio; **CSF,** contrast sensitivity function; **DCE-MRI,** dynamic contrast enhanced mri; **fMRI,** functional mri; **FOV,** field of view; **FWHM,** full width half maximum; **HVS,** human visual system; **INU,** image non-uniformity; **JND,** just noticeable differences; **MOS,** mean opinion score; **NSA,** null space analysis; **PD,** percentage distortion; **PDM,** perceptual difference model; **PSF,** point spread function; **QILV,** quality index based on local variance; **RF,** radio frequency; **RMSE,** root mean square error; **ROI,** region of interest; **SNR,** signal to noise ratio; **SSIM,** structural similarity index; **TE,** time to echo; **TI,** inversion time; **TR,** repetition time

repetition time (TR), time to echo (TE), flip angle, and inversion time (TI). However, it also has its unique set of artifacts that could potentially distort the image, leading to false conclusions. Hence, it is important to assess the quality of the images generated and penalize the presenting artifacts before utilizing it to draw conclusions.

II. SIGNIFICANCE OF QUALITY ASSESSMENT

The importance of quality metrics in medical imaging has been emphasized by Boone.[1] Image quality varies not only with changes in acquisition parameters but also across different sites, where they are acquired. This could be due to the differences in ambience, unintended coupling of the receiver coils with any other miscellaneous object in the neighborhood, variations in hardware (such as radiofrequency [RF] coil, Q-factor, passive shims, and center frequency) and RF transceiver variations. Hence, the quality of images obtained at one site may not match that acquired elsewhere. Thus, it is important to establish acceptable standards that can be applied across imaging systems. Because the MR acquisition system has been well studied, the imaging parameters and the procedures to evaluate the performance of clinical scanners are well established. The research on quality assurance reported by Price et al.[2] describes standard procedures to evaluate clinical scanners. In this paper, recommendations are made for acceptable phantom designs, materials, and analysis techniques. Imaging parameters such as resonance frequency, signal-to-noise ratio, image uniformity, spatial linearity, spatial resolution, slice thickness, and contrast-to-noise ratio are also described.

Assessment of image quality is important for several reasons, such as confidence in subsequent diagnostics,[3–6] determining the quality of motion correction in functional MR imaging (fMRI) time series analysis,[7] quantitative morphometric studies of brain images,[8] optimal phantom design for enhanced lesion detectability,[9] assessment of brain tumor response to therapies,[10] determining synchronization of cardiac cycles in cardiac imaging,[11] optimization of interventional MR imaging,[12] quality of image reconstruction in parallel imaging,[13] improvement of acquisition design,[14,15] and evaluating the quality of image fusion.[16]

III. IMPORTANT IMAGING PARAMETERS

Some of the important parameters defined in the standard MR literature[17,18] that directly affect the image quality are outlined below. They can be classified into those that are determined by the given imaging system and those that can be controlled by the operator. Parameters that are determined by the imaging system are as follows:

1. *Resonance frequency*: Resonance frequency is defined as that RF f, which matches the static B-field (B_0), according to the Larmor equation $f = (\gamma/2\pi)B_0$.
2. *Image uniformity*: This refers to the ability of the MR imaging system to produce a constant signal response throughout the scanned volume when the object being imaged has homogeneous MR characteristics.
3. *Signal-to-noise ratio (SNR)*: Signal is defined as the mean pixel intensity within a user-marked region of interest. Noise is defined as the standard deviation of the pixel intensity in the background region. SNR is the ratio of signal-to-noise energy.
4. *Spatial linearity*: This is used to describe the degree of geometrical distortion present in images obtained.
5. *Distortion*: This quantifies the error in the measured value of a distance between any two points in the image. Percentage distortion (PD) is defined as: PD = (Perceived Dimension – True Dimension)/ True Dimension expressed as a percentage. Distortion measurement may be performed between any two points within the field of view (FOV). For reliable measurements, it is recommended that the true dimension be greater than 10 pixels.

The parameters that are controlled by the operator are as follows:

1. *Acquisition matrix*: Acquisition matrix (N_x,N_y) is the ordered pair of the number of data points acquired along k_x and the number of phase encodes acquired along k_y. The time needed for scan increases as the size of the acquisition matrix increases.
2. *FOV*: This is defined as the image area within the object to be scanned that is spatially encoded by the imaging system.
3. *Spatial resolution*: This is a measure of the capacity to clearly distinguish between objects when there is no significant noise contribution. Spatial resolution for MRI systems is typically limited by the pixel size of the acquisition matrix (resolution along X or Y = FOV along X or Y divided by the number of sampling points along X or Y).
4. *Slice thickness*: This is defined as the thickness of an imaged slice and is measured as the distance between the points at half the sensitivity of the maximum or full width at half maximum (FWHM) of a slice profile.
5. *Contrast-to-noise ratio (CNR)*: The ability to distinguish between various soft tissues is called CNR, which is defined as:

$$\text{CNR} = \text{abs}(\text{Signal}_1 - \text{Signal}_2)/\text{standard deviation of noise}$$

where Signal_1 and Signal_2 are the intensities of pixels (or voxels) from the two different regions being compared. The CNR depends on the contrast mechanism used and the spatial resolution. It is influenced by operator-dependent settings for imaging parameters, such as TR, TE, flip angle, TI, or the use of preparatory pulses.

III.A. Quantifying Imaging Characteristics

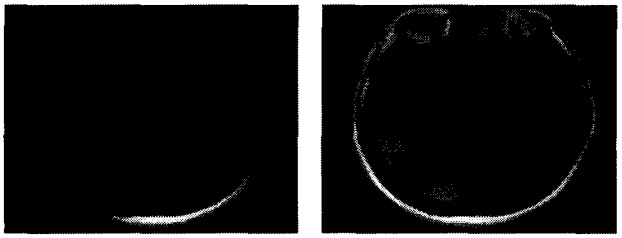
The most popular method of quantifying the inherent characteristics of an imaging system is through measures such as point spread function (PSF) and FWHM. Also, standard phantoms have been designed to study the geometrical accuracy and spatial resolution obtained through the imaging process.

PSF is the response of the imaging system when the input stimulus is a single point. The hypothetical point generally results in an extended image, characterizing the limitation of the imaging process. Knowledge of the PSF helps in predicting how any given object would be imaged by the imaging system, under assumptions of linearity. It is used in understanding and measuring the spatial resolution of the imaging system. The degree of spreading of the point object is the measure of the spatial resolution that can be achieved with the imaging system.

FWHM is commonly used to measure the quality of an imaging system and its spatial resolution, by denoting the significant width of a nonrectangular process. It is the width at half the maximum value of the peak and is commonly used for spectral lines and slice profiles. FWHM is measured by identifying the points on the signal curve, which are half the maximum value. The horizontal distance between these two points is called the FWHM.

Several standard phantoms are designed, each dedicated to test a certain aspect of the performance of the imaging system. Although the final goal is to image the human body, phantoms are used because the resulting images can be assessed. Also, if any drawback of the specific imaging technique is revealed, then it can be worked on. Phantoms can be repeatedly imaged, unlike human subjects. Phantoms can easily be standardized, and various imaging sites could concur on the quality of the expected image. Phantoms have to be designed with materials that constitute the human body in order to mimic the MR signal expected while imaging a human subject. Some of the constituent materials are aqueous paramagnetic solutions, pure gels of gelatin, agar, polyvinyl alcohol, silicone, polyacrylamide, and agarose.

Geometrically accurate phantoms include spheres and cylinders, some with grid-like structures and yet others with different symbols in different quadrants. The fidelity of reproduction is checked along with sizes, positions of the grid, and symbols. Corrective coefficients are determined to enable correction of subsequent patient images.



(a) (b)

FIGURE 1. Illustration of artifacts.[21] (a) Nonuniform coil sensitivity and (b) Gibbs ringing.

IV. ARTIFACTS IN MR IMAGES

Research reported by Ruan[19] and Patton et al.[20] extensively deals with the commonly encountered artifacts in MR images. The objective of their work was to develop consistent image quality and appropriate selection of imaging parameters. Additionally, both reports have outlined typical artifacts and corresponding remedial actions. Incorporation of such awareness in clinical practice leads to better imaging quality and assessment. These authors have also grouped MR imaging artifacts into two general categories, namely, hardware related and operator/patient related.

IV.A. Artifacts Caused by MR Imaging Hardware

Hardware-related artifacts are difficult to correct and may require the assistance of service personnel. The most common of these artifacts are as follows:

1. *RF noise*: Improper RF shielding could result in external noise being acquired along with the RF signal. This form of artifact could lead to structures such as a bright spot in the image.
2. B_0 *inhomogeneity*: Intensity distortions result from B_0 inhomogeneity. This is because the strength of B_0 affects the value of the parameter T2*, which dictates the decay rate of the signal.
3. *Coil sensitivity*: The image might look brighter in regions with greater coil sensitivity and vice versa, if the coil sensitivity profile is not uniform across the entire object being imaged.
4. *Gradient linearity*: If the object occupies a region outside the area where gradient linearity holds, then geometric distortion could occur.
5. *RF inhomogeneity*: An RF inhomogeneity leads to the presence of an undesired variation in signal intensity across the obtained image. The cause is either a nonuniform $B1$ field or a nonuniform sensitivity in a receive-only coil.

IV.B. Operator-/Subject-Related Artifacts in MR Images

The second category of artifacts consists of those that arise out of an undesired action by the operator or subject, which can be easily corrected with the right course of action.

Gibbs ringing: These artifacts manifest at regions of sudden discontinuities, as single-dimensional ripples. Gibbs ringing is seen in images when an insufficient acquisition matrix is used (Fig. 1). It means that adequate data are not acquired at high spatial frequencies. Solutions include use of a higher resolution imaging matrix and filtration methods.

Patient positioning: Appropriate positioning of the patient in the RF coil is important in order to obtain good image quality. The region of interest should always be positioned in the center of the coil because systems are optimized for imaging in this region.

Motion: Motion is the most prevalent source of MR imaging artifacts. Motion artifacts, also called ghosting, could be caused by the motion of the imaged object as a whole or a part of it during the imaging sequence, leading to blur. Figure 2 shows the degradation caused by motion artifact. To minimize artifacts due to respiratory motion, gated imaging is carried out for anatomies such as the abdomen. For cardiac imaging, electrocardiographic triggering is used, in which data collection is synchronized with the cardiac phase. Such synchronization enables

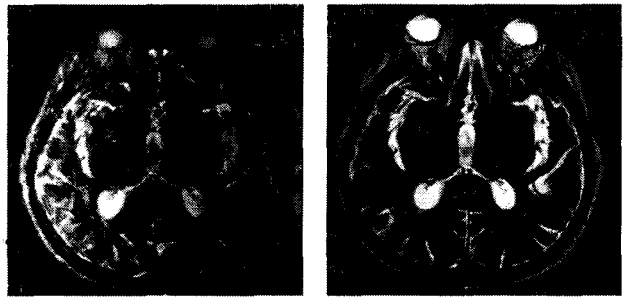


FIGURE 2. Illustration of motion artifact.[22] (a) Image with no motion compensation and (b) motion-compensated image.

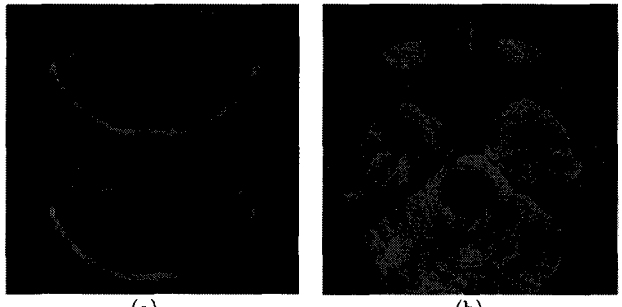


FIGURE 3. Illustration of (a) a wrap-around artifact[21] and (b) a susceptibility artifact.[23]

cardiac tissue to be located in a consistent position as each successive phase-encoding step is acquired, resulting in increased tissue signal intensity and decreased phase errors.

Chemical shift: Chemical shift artifacts are caused by the difference in Larmor frequency of fat and water. The artifact manifests itself as a mis-registration between fat and water pixels in an image. The effect is that fat and water spins in the same voxel are encoded as being located in different voxels. The magnitude of the effect is proportional to the magnitude of the B_0 field and inversely proportional to the sampling rate in the frequency-encoding direction. Fat suppression methods often eliminate visible artifacts.

Wrap-around artifacts: Wrap-around artifacts occur when the effective field of view is smaller than the region of interest. This implies that the sampling rate is less than the range of frequencies in the RF signal. An illustration of the artifact is shown in Figure 3a. The solution is to choose a larger field of view.

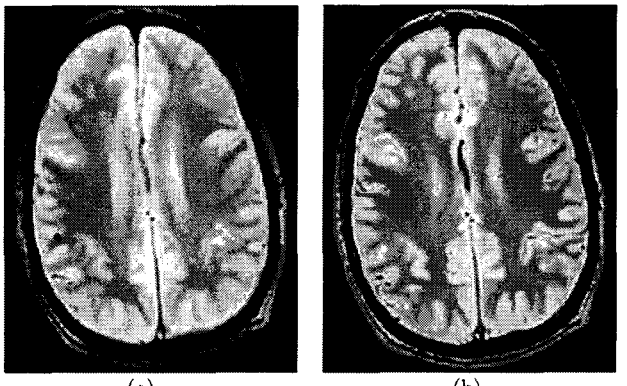
Susceptibility artifacts: These occur as the result of variations in the magnetic field strength at interfaces of tissues with different magnetic susceptibilities, such as air-tissue boundaries in nasal cavities. These artifacts manifest as bright and dark regions with spatial distortion of the surrounding anatomy, as shown in Figure 3b.

Partial volume artifacts: Partial volume artifacts are caused by the size of the image voxel. These occur when multiple tissue types are encompassed within a single voxel. For example, if a small voxel contains only fat or water signal, and a larger voxel might contain a combination of the two, the larger voxel possesses a signal intensity equal to the sum of those of water and fat, weighted by their relative amounts, present in the voxel. An illustration of the artifact is shown in Figure 4. It also occurs when structures are oriented obliquely to the imaging plane and when structures move in and out of a given section during image acquisition.

Zebra stripes: These can be observed along the periphery of gradient-echo images where there is an abrupt transition in magnetization at the air-tissue interface (Fig. 5). They are accentuated by aliasing that results from the use of a relatively small field of view. Solutions include expanding the field of view, using spin-echo pulse sequences or oversampling techniques to reduce aliasing.

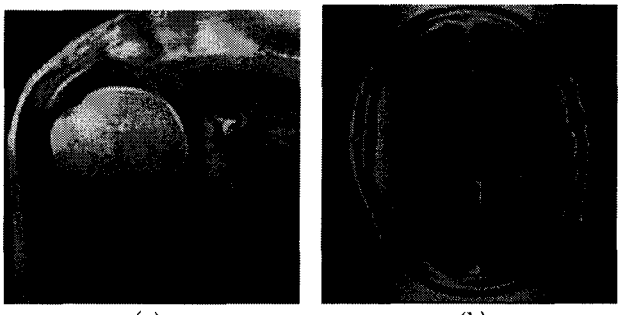
RF overflow: These artifacts impart a nonuniform, washed-out appearance to an image. This artifact occurs when the signal received by the scanner is too intense to be accurately digitized by the analog-to-digital converter. Auto-prescanning is usually performed before the actual scan in order to adjust the receiver gain to prevent this.

Flowing spin artifacts: This occurs when unsaturated spins in the blood enter the excited slice. It is characterized by bright signal in a blood vessel (artery or vein) in the particular slice. Usually, the signal is seen on more than one slice, fading with distance. This artifact can be confused with thrombosis, and with disastrous results.



(a) (b)

FIGURE 4. Illustration of partial volume artifact.[18] Image with (a) 10-mm resolution along *Y* and (b) 3-mm resolution along *Y*.



(a) (b)

FIGURE 5. Illustration of (a) a zebra stripes artifact[23] and (b) an overflow artifact.[23]

Zipper artifacts: These manifest as high-intensity lines across the MR image. The zipper artifacts that can be controlled easily are those due to RF entering the scanning room, when its door is open during acquisition of images. Solutions include identifying and removing external RF sources, ensuring that the door to the imaging room remains closed and verifying the integrity of the magnet room enclosure and associated seals.

Cross-excitation: This is caused by the imperfect shape of RF slice excitation profiles, leading to unintended excitation of adjacent tissues. This excitation might result in the saturation of such tissues and manifest as decreased signal intensity and contrast. One way to avoid this artifact is to introduce an intersection gap that is 10–50% of the prescribed section thickness. Another method is to acquire slices in interleaved manner.

Shading artifacts: These manifest as relatively reduced signal intensity across a portion of the image. Abnormalities contained in the shaded portion of the MR image may be obscured.

V. QUALITY METRICS IN MEDICAL IMAGING

As pointed out by Woodard and Carley-Spencer[24] and Video Experts Quality Group,[25] desirable properties of quality measures are that they accurately predict human judgments of distortion, be monotonically related to the level of distortion, and provide reasonably consistent results across a wide variety of images, along with mathematical tractability, simplicity, and computational efficiency.

Boone[1] states that although image interpretation is subjective, clinical applications such as tumor imaging require quantitative indexing. He predicts that the subjective interpretive environment will evolve toward the increased use of quantitative metrics for evaluating patient's health from images. He also comments that for the quantitative imaging environment to expand, medical physicists, physicians, researchers, and equipment vendors need to work collaboratively to develop quantitative protocols for imaging, scanner calibrations, and robust analytical software. Such a collaboration would lead to the routine inclusion of quantitative parameters in diagnosis and therapeutic assessment of human health. Most importantly, the report emphasizes that developing quantitative metrics would have a great impact on patient diagnosis.

Most publications that propose new techniques for improved image quality[14,26,27] evaluate it by visual inspection. In most cases, the lack of visible artifacts is seen as a requisite to consider the quality of the image "good." Although there is no concurrence on what constitutes a good quality image, it is implicit that occurrences of artifacts must be penalized. Increasing advancement for acquisition of good quality images, with higher spatial/temporal resolution, makes it important to quantify the goodness of an image.

In the following section, we outline different categories of quality metrics, with their subdivisions

and instances of their reported applications on MR images. Each category is based on consideration for one of the important attributes of a quality metric. They are not mutually exclusive and a given quality metric could fit into each of these categories. However, each chosen quality metric is elaborated in only one of these categories to emphasize a particular attribute of the metric.

V.A. Subjective and Objective Measures

This classification of indices arises out of consideration or lack of it by human observers.

1. Objective Measures

Objective measures are most preferred because they are reproducible. They are generally the result of a well-thought-out algorithm that should incorporate rewards for fidelity in image reproduction and penalize artifacts. Such indices render computer-aided diagnosis easy because the model is well represented and results are predictable. However, the disadvantage is that outliers and/or special instances may be misinterpreted to be erroneous.

McGee et al.[28] propose several objective metrics from the gradient of the image, based on a model postulating that the ideal image should have areas of uniform brightness separated by sharp edges. The gradient of a good uniform image would be zero or very small over large areas and large in a relatively small number of pixels. Metrics developed by these investigators are based on image resolution, characteristics of the gray-scale histogram, contrast, and autocorrelation, which are derived from the pixel intensities and their statistical characteristics. They include gradient, Laplacian, entropy, cube of normalized intensities, fourth power of normalized intensities, normalized gradient squared, normalized gradient to fourth power, squared intensities, squared gradient, fourth power of gradient, gradient entropy, histogram mean, histogram standard deviation, histogram skewness, histogram kurtosis, histogram energy, histogram entropy, standard deviation, standard deviation of gradient and autocorrelation.

The intensity-based metrics at position (i, j), with intensity $g(i, j)$ are defined as follows:

$$\text{Gradient}_1 = \sum_{ij} \left| \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} * g \right|$$

$$\text{Gradient}_2 = \sum_{ij} \left| \begin{bmatrix} 1 \\ -1 \end{bmatrix} * g_{ij} \right|$$

$$\text{Laplacian}_1 = \sum_{ij} \left| \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} * g_{ij} \right|$$

$$\text{Laplacian}_2 = \sum_{ij} \left| \begin{bmatrix} -1 & -2 & -1 \\ -1 & -2 & -1 \\ -2 & -12 & -2 \end{bmatrix} * g_{ij} \right|$$

$$\text{Entropy} = -\sum_{ij} h_{ij} \ln[h_{ij}]$$

where,

$$h_{ij} = \frac{g_{ij}}{\sum_{ij} g_{ij}^2}$$

$$\text{Cube of normalized intensities} = \sum_{ij} \left(\frac{g_{ij}}{\sum_{ij} g_{ij}^2} \right)^3$$

$$\text{Fourth power of normalized intensities} = \sum_{ij} \left(\frac{g_{ij}}{\sum_{ij} g_{ij}^2} \right)^4$$

$$\text{Squared intensities} = \sum_{ij} \left(\frac{g_{ij}}{n} \right)^2$$

$$\text{Squared gradient} = \frac{\sum_{ij} \left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} * g_{ij} \right)^2}{n}$$

$$\text{Fourth power of gradient} = \frac{\sum_{ij}\left(\begin{bmatrix}1\\-1\end{bmatrix} * g_{ij}\right)^4}{n}$$

$$\text{Gradient entropy} = -\sum_{ij} h_{ij}\log_2[h_{ij}]$$

where,

$$h_{ij} = \frac{\left|\begin{bmatrix}1\\-1\end{bmatrix} * g_{ij}\right|}{\sum_{ij}\left|\begin{bmatrix}1\\-1\end{bmatrix} * g_{ij}\right|}$$

McGee et al.[28] evaluate the above metrics by autocorrecting MR images of the rotator cuff. MR imaging of the shoulder requires high spatial and contrast resolution and, thereby, long acquisition times, predisposing the images to degradation due to motion. The evaluated quality metrics are used to autocorrect any motion that might have occurred during imaging. The objective is to deduce motion during imaging by calculating a metric that reflects the image quality and searching for motion values that optimize this metric. Raw data from 164 clinical coronal rotator cuff exams acquired with interleaved navigator echoes are used. Changes in values of the metric before and after navigator-based adaptive motion correction are correlated with changes in the observer score using a least-squares linear regression model. On the basis of their analysis, the metric that exhibits the strongest relationship with observer ratings of MR shoulder images turns out to be the entropy of the one-dimensional gradient along the phase-encoding direction.

Lin et al.[29,30] have reported an autofocusing (AF) motion correction technique in high-resolution trabecular bone imaging, where image SNR is limited. Here, image sharpness and structural parameters are computed and compared after correction to determine the efficacy of the AF method. Image sharpness is computed as the opposite of image nonuniformity (INU) or blur. Sharpness is modeled as a high-pass process like the gradient. Raw data from 26 clinical three-dimensional wrist exams are motion corrected using AF for both in-plane rotation and translation. Changes in image metrics and subsequently computed structural parameters are used to gauge the performance of the AF algorithm, which is compared with translation-only navigator-corrected results. Results show that AF-generated images have higher image sharpness compared to the navigator echo technique. The average normalized gradient squared metric improves by 0.40%, 0.73%, and 0.84%, respectively, following translation-only navigator, translation-only AF, and combined rotation/translation AF. For all structural parameters, the rotation/translation AF results in an approximately two-fold greater change compared to the navigator technique.

2. Subjective Measures

Subjective measures typically involve human inspection. These measures, such as mean opinion score (MOS), have been in use for a long time and are often used as the benchmark in validating novel metrics.[24,31,32] However, visual inspection and the extent of details captured depend on the level of experience of the observer. Hence, the result may not only be nonreproducible but could also depend on factors such as the stress level of the observer. However, one great advantage is that because HVS characteristics are implicitly considered, it is likely to agree to a great extent with the radiologist's perception.

Prieto et al.[31] have proposed a subjective measure based on just noticeable differences (JND), called JND scanning (JNDS). The JND[33] is defined as the maximum threshold below which distortions are not perceived in a given pixel. Hence, the JND profile is a binary image that gives the locations of pixels, where the difference between the two images exceeds a given threshold (where the difference becomes noticeable). The plot in Figure 6a shows the threshold for a given contrast. In their paper, the authors propose utilizing JND to measure how the image differences disappear as the contrast between the two

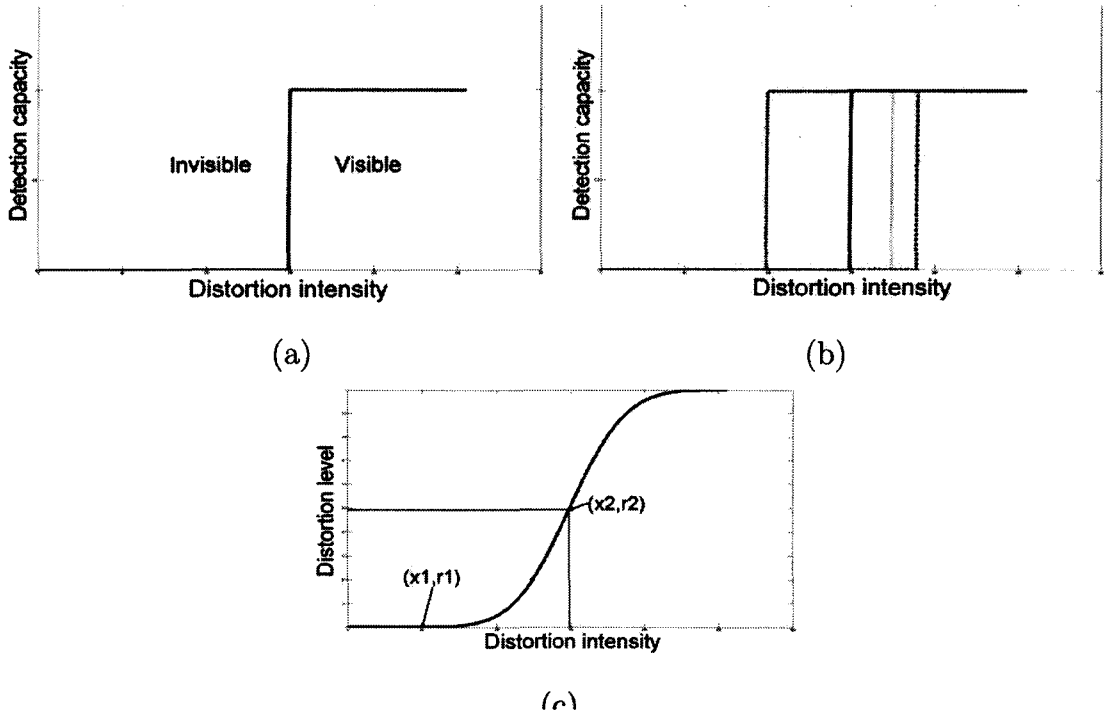


FIGURE 6. Plots related to JND.[31] (a) The JND threshold is a step function. (b) The JND threshold moves when the contrast changes. (c) Integration of all possible thresholds is depicted.

images decreases. Initially, the intensity differences that disappear correspond to the least distorted pixels, and as the process proceeds, the intensity differences corresponding to the most distorted pixels disappear. The JND threshold is a step function for each pixel. As the image contrast changes, the threshold moves horizontally. The plot in Figure 6b shows the varying thresholds for changing levels of contrast. The integration of all of the possible thresholds is a ramp, which is illustrated in Figure 6c. This ramp should be interpreted as the probability with which the distortion in a pixel is observed for all possible levels of contrast.

The probability of distortion for each pixel is assigned as the ratio of number of pixels with probability one to the number of pixels with probability zero, obtaining the JNDS index. To validate the metric, the authors compare it against MOS obtained from five volunteers on a data set of 150 brain images. The correlation with human perception was graded on a scale of 1 (worst) to 5 (best). The results are shown as plots in which the proposed metric showed better correlation with human perception compared to root mean square error (RMSE).

Gardner et al.[32] utilize human observers to detect significant errors in the images obtained, compared against the sensitivity of an automated test. The automated image analysis system for MR imaging with the sensitivity of trained human observers is used to evaluate images that were intentionally degraded. Degradation involved decreased SNR, which caused distortion, and increased slice thickness and separation. It was found that the human observers were able to detect a 6–13% reduction in SNR and distortions of >15% in human images. They were unable to identify 40% increases in the slice thickness. On the other hand, automated analysis of test object images was able to detect all image degradations at the minimum levels applied. Hence, it was concluded that visual analysis of clinical images is not sufficient to decide on the image quality. Quantitative image-quality analysis that reliably predict the degraded image

quality are necessary to detect subtle abnormalities in clinical images.

V.B. Region-of-Interest (ROI)–Based Versus Non-ROI–Based Indices

This categorization is based on the image area considered in declaring the quality of the image.

1. ROI-Based Measures

SNR and CNR are the most popular ROI-based image metrics. This class of metrics is useful in compression because the area of interest needs to be compressed without incurring any loss in quality, while the rest of the image can undergo lossy compression, without affecting the diagnostic quality of the image. In applications such as dynamic contrast-enhanced MRI (DCE-MRI), ROI-based measures are important because the enhancement curves are obtained from a very specific region. Hill et al.[34] have proposed improvement in the quality of breast images by reducing the variability of the enhancement curves at the pixel level yielding more reliable uptake and washout phases.

2. Non-ROI–Based Measures

The non-ROI–based measures take the entire image into account while deciding on the quality. This is especially true in cases with supervised classifiers such as those that train neural networks in deciding "good" quality against "bad."[35] Also, metrics that consider global statistical characteristics fall into the same category. Aja-Fernandez et al.[36] have proposed a new index based on the assumption that critical structural information of an image is coded in its local variance distribution. Quality index based on local variance (QILV) between images I and J is computed as

$$QILV(I,J) = \frac{2\mu_{V_I}\mu_{V_J}}{\mu_{V_I}^2 + \mu_{V_J}^2} \cdot \frac{2\sigma_{V_I}\sigma_{V_J}}{\sigma_{V_I}^2 + \sigma_{V_J}^2} \cdot \frac{\sigma_{V_I V_J}}{\sigma_{V_I}\sigma_{V_J}}$$

where μ_v is the mean of the local variance and σ_v is the standard deviation of the local variance. They claim that the stationary properties of the images are considered in this metric. The first term in the expression compares the mean of the local variance distributions of the two images, while the second term compares the standard deviation of local variances. Spatial coherence is introduced in the third term. The authors claim that the above metric represents global statistics of local variances of images. In order to validate the performance of the proposed metric, the authors tested it with MR images corrupted with Rician noise and found that the metric not only detects noise but also the associated blur when the images were de-noised using Wiener filtering. The popular clinical measure RMSE and several statistical metrics developed by McGee et al.[28] are also non-ROI–based measures.

V.C. Full-Reference/No-Reference Metrics

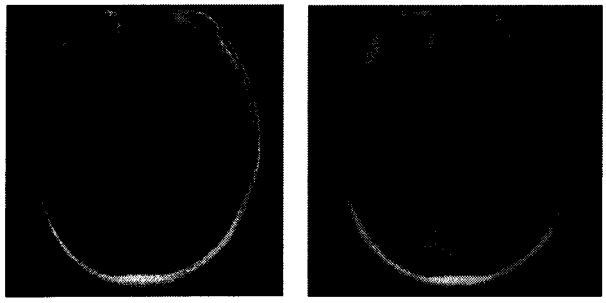
This categorization is based on the availability of the gold standard with respect to which the quality of the image in question can be compared.

1. Full-Reference Metrics

These metrics assume the availability of the golden standard image against which a test image is to be compared. Although in principle this approach is not suitable for medical images, it is extensively used because typical characteristics of medical images remain broadly constant, among large masses of the population.

The most popular, clinical full-reference metric is perhaps RMSE. It is computationally simple and gives a fair idea of the deviation from the ideal. However, as pointed out by Prieto et al.,[31] two images with the same RMSE can be perceptually very dissimilar as shown in Figure 7.

Prieto et al.[31] have proposed a technique called null space analysis to quantify degradation of an image. The objective is to measure loss in the perceptually relevant information while transforming the reference image A, into another image D. This



(a) (b)

FIGURE 7. Perceptually different images with the same RMSE (RMSE = 11).[21]

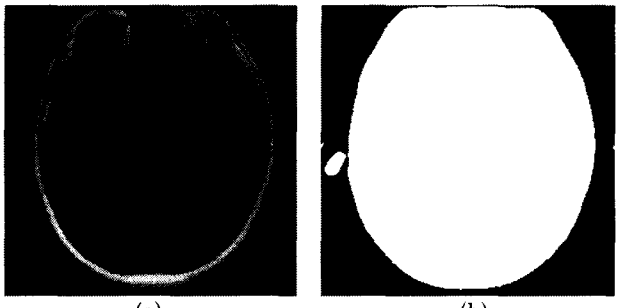
is accomplished by finding a linear transformation T, from A to D that maximizes its null space dimension. Because T is a linear transformation with $N(N-1)$ degrees of freedom (N is the number of pixels), T is selected such that its null space has the maximum allowable dimension of $(N-1)$. Thus, $T = De^*/\langle e, A\rangle$, where e is a reference image that determines a space that is orthogonal to the null space of T. Here, e is taken as the average of A and D. A can be split as $A = A_K + A_I$, where A_K and A_I are the lost and preserved information of the image after the transformation T, respectively. A_K is weighted by the contrast sensitivity function (CSF)[37] to match the perceptual significance of the lost information $\phi = W^{-1}.CSF.W$ where W is the Fourier transform. Finally, the index null-space analysis (NSA) is computed as the energy of ϕ. To validate the metric, the authors compared it against MOS obtained from five volunteers on a data set of 150 brain images. The correlation with human perception was graded on a scale of 1 (worst) to 5 (best). The results were shown as plots in which the proposed metric showed better correlation with human perception compared to RMSE.

2. No-Reference Metrics

These metrics do not assume the availability of the gold standard with respect to which a given image can be assessed. They quantify quality inherent in the single image. This is especially relevant in the medical imaging scenario because each subject is unique. Also, for the same subject, the same anatomy imaged once might look drastically different when imaged at another point in time.

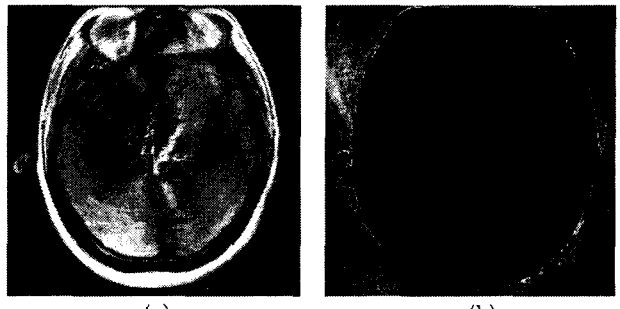
Woodard and Carley-Spencer[24] assess the effectiveness of no-reference image quality measures. They artificially distorted images with two levels of either additive Gaussian noise or intensity nonuniformity created from a linear model. A total of 239 different quality measures were defined and used to discriminate images for the type and level of distortion. A data set of 1001 MR images recorded from 143 subjects was used for their evaluation. Statistical analysis was carried out. No-reference quality measures that characterize the spread of energy and information in spatial and/or spatial-frequency domains, the localization, flatness, and Tsallis entropy-based measures were used. It was found that analysis of variance (ANOVA) identified two families of quality measures that were most effective: one based on natural scene statistics[38] and the other, the JPEG quality measure,[39] originally developed to measure distortion caused by image compression. Woodard and Carley-Spencer[24] report that measures from both of the families reliably discriminated between undistorted images, noisy images, and images distorted by intensity nonuniformity. According to Woodard and Carley-Spencer,[24] the best quality measures were sensitive only to the distortion category and were not significantly affected by other factors. The authors report that the results were encouraging enough that several quality measures were being incorporated in clinical scenarios.

Neuroinformatics researchers at MITRE[40] have added varied types of distortion, such as radio frequency irregularities of the scanner, electromagnetic interference from other equipment, artifacts due to patient head movement, and inherent properties of the tissue being scanned. A variety of metrics that measure an array of properties, including entropy, energy localization, and underlying image structure, are utilized. They claim that their experiments have shown that a subset of the metrics and combinations thereof can discriminate distorted from nondistorted images and even discriminate between levels of distortion. They found that undistorted, noisy, and



(a) (b)

FIGURE 8. Image with no degradation.[21] (a) Structural brain image and (b) segmented air background.



(a) (b)

FIGURE 9. Degraded image. (a) Structural brain image and (b) segmented air background showing visible degradation.

INU distorted images evidently cluster into different regions.

V.D. Automated/Semi-Automated Methods

This classification is based on whether manual intervention is needed.

1. Automated Methods

These methods are typically designed to suit a certain class of images. They exploit the characteristics of images within the class and also incorporate *a priori* knowledge about them. An example is the method reported by Mortamet et al.[41] to determine the quality of structural brain images. This method exploits the fact that the air background occupies above 30% of the total image area in structural brain images. The image quality is determined by examining the air-background region. The air-background region is assumed to provide sufficient information to detect image degradation from several sources, such as patient head movement, residual magnetization from incomplete spoiling, blurring, and ghosting, an illustration of which is shown in Figures 8 and 9.

Mortamet et al.[41] propose two quality indices. The first quality metric uses an atlas-based algorithm to extract the air-background region. In the air-background region, artifacts are isolated by means of morphological operations. One quality index is computed as the proportion of artifactual voxels relative to the background size. This quality measure reflects the presence of artifacts and relies on artifact delineation and clustered property of artifactual voxels. The second quality metric is based on the noise distribution analysis. The measure of goodness of fit of Rayleigh distribution to the noise histogram is chosen as a quality metric because the presence of artifacts enlarges the right tail of the noise intensity distribution to higher intensity values, and hence increases its skewness. The data used for validating the indices comprised 749 T1-weighted structural head MRI images from 188 subjects, which are obtained from the ADNI database.[42] As per the report, ANOVA could capture significant differences between the low- and high-quality groups. The discriminative performances of the proposed indices appeared to be independent of magnetic field strength or software-hardware combinations used by the MRI systems.

2. Semi-Automated Metrics

In semi-automated metrics, manual intervention is not completely eliminated. It may be needed for purposes such as marking a region of interest or giving intensity bounds or to determine sizes of smoothening masks. The most clinically popular among these is SNR, in which the user needs to select the region of interest and the background.

V.E. HVS-Based Versus Non-HVS–Based Methods

1. HVS-Based Methods

Methods based on the characteristics and limitations of the human visual system are highly relevant in any scenario, more so for medical images because they are finally inspected and interpreted by the human eye. The human visual system is very complex and although it has been well studied, it has not been adequately modeled analytically. However, attempts to emulate human visual perception continue to dominate computer vision-based applications.

Sinha et al.[43,44] utilize a HVS-based metric called the Structural SIMilarity (SSIM) index[45] to assess the quality of rapid MR imaging. The authors use the SSIM index to compare the closeness of the reconstructed image to the original. The SSIM index penalizes loss in structural correlation, intensity, and contrast. SSIM at location (p, q) is given by

$$SSIM(p,q) = [l(p,q)]^{\alpha}[c(p,q)]^{\beta}[s(p,q)]^{\gamma}$$

where l, c, and s reward similarities in intensity, contrast, and structure, respectively. In circumstances in which luminance, contrast, and structure are given the same weight, α, β, and γ are chosen to be equal to 1; else they are customized based on the application. A small local neighborhood is chosen to determine the corresponding point-wise attributes. Here, intensity factor is given by

$$l(p,q) = \frac{2\mu_p\mu_q}{\mu_p{}^2 + \mu_q{}^2}$$

where μ_p and μ_q are the average intensities in the neighborhoods of the two points being compared. The contrast comparison function is given by

$$c(p,q) = \frac{2\sigma_p\sigma_q}{\sigma_p{}^2 + \sigma_q{}^2}$$

where σ_p and σ_q are the respective standard deviations. Standard deviation σ is the square root of variance and is an unbiased estimate of the signal contrast. The structural comparison function is given by

$$s(p,q) = \frac{\sigma_{pq}}{\sigma_p\sigma_q}$$

which turns out to be the correlation (inner product) between the structures being compared. The structural comparison is carried out only after intensity subtraction and variance normalization. The SSIM score ranges from 0 to 1, where 0 represents the worst and 1 stands for the best quality. The index is developed based on the assumption that the structural information of the image is most critical in determining the perceived quality of the image. This metric is generally used for image/video quality assessment. The SSIM index was used on five MR data sets comprising reduced data undersampled along Cartesian, spiral, and radial trajectories. The authors claim that the performance of the metric correlates well with visual inspection.

Salem and Wilsof[12] propose a perceptual difference model (PDM) that incorporates various components of the human visual system. Properties such as maximum and minimum luminance values, physical pixel size, and viewing conditions (e.g., viewing distance and nonlinear gray-level perception) are incorporated in the model. To assess the frequency domain properties, the images are filtered by the human spatial frequency CSF. This is followed by processing using Cortex transform and contrast calculation, thus covering multiple facets of human visual perception. The PDM attempts to quantitatively describe the visual differences between two images when viewed by a human observer. The PDM model processes two inputs, a reference image and a degraded "test" image. The output is a pixel-by-pixel map showing the magnitude of perceived image differences. These differences are then combined over the entire image or a region of interest to give a scalar PDM score. To validate their model for clinical utility, a panel of two board-certified interventional radiologists were asked to evaluate a series of keyhole MR images, a k-space subsampling technique that introduces image blur. Seventy-two images were randomly selected from a number of simulated keyhole sequences; these images covered a broad range of image quality. The first eight images were used as training images. Each keyhole image was shown beside the corresponding "gold standard" full k-space image. Without any knowledge of the PDM or MSE scores, the two radiologists reached

a consensus and classified the keyhole images as acceptable, marginally acceptable, or unacceptable for clinical use. The results have validated the usage of PDM as a quality metric that mimics human perception satisfactorily.

2. Non-HVS–Based Methods

Most techniques that are still clinically popular, such as SNR, CNR, RMSE, and PSNR, are all non-HVS based because they do not incorporate characteristics of the human visual system. Most of these are computationally simple and give a fair idea of how good an image is. However, these techniques can easily be deceiving because they do not reward or penalize images holistically. They, however, have the advantage of being objective and reproducible.

VI. CONCLUSION

In this review, we have examined the necessity of developing quantitative metrics for MR imaging. This study deals with the various classes of image quality metrics used for assessing MR images. Various applications, advantages, and disadvantages of quality metrics of MR images have been evaluated under subjective/objective, automatic/semi-automatic, region-of-interest/non-region-of-interest–based, full-reference/no-reference, and HVS-incorporated/non-HVS classes. Results from various laboratories have been summarized, and it is concluded that automated, objective indices that incorporate HVS and are no-reference metrics are most preferred.

REFERENCES

1. Boone JM. Radiological interpretation 2020: Toward quantitative image assessment. Med Phys. 2007;34(11): 4173–9.
2. Price RR, Axel L, Morgan T, Newman R, Perman W, Schneiders N, Selikson M, Wood M, Thomas SR. Quality assurance methods and phantoms for magnetic resonance imaging. Med Phys. 1990;17(2):287–95.
3. Akella NS, Twieg DB, Mikkelsen T, Hochberg FH, Grossman S, Cloud GA, Nabors LB. Assessment of brain tumor angiogenesis inhibitors using perfusion magnetic resonance imaging: quality and analysis results of a phase I trial. J Magn Reson Imaging. 2004;20:913–22.
4. Morris E. Breast cancer imaging with MRI. Radiol Clin North Am. 2002;40(3):443–66.
5. Hermier M, Nighoghossian N, Derex L, Berthezène Y, Blanc-Lasserre K, Trouillas P, Froment JC. MRI of acute post-ischemic cerebral hemorrhage in stroke patients: diagnosis with T2*-weighted gradient-echo sequences. Neuroradiology. 2001;43(10):809–15.
6. Beyersdorff D, Taymoorian K, Knösel T, Schnorr D, Felix R, Hamm B, Bruhn H. MRI of prostate cancer at 1.5 and 3.0 T: comparison of image quality in tumor detection and staging. AJR Am J Roentgenol. 2005;185(5): 1214–20.
7. Freire AR, Mangin JF. What is the best similarity measure for motion correction in fMRI time series? IEEE Trans Med Imaging. 2002;21(5):108–21.
8. Fennema-Notestine C, Ozyurt IB, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, Jernigan TL, Fischl B, Segonne F, Shattuck DW, Leahy RM, Rex DE, Toga AW, Zou KH, Brown GG. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum Brain Mapp. 2006;27:99–113.
9. Chen BSHH. Quality assessment of high spatial resolution for MRI. http://ric.uthscsa.edu/personalpages/lancaster/DI2Projects2003/QAMRI.pdf. Accessed 2010 Feb 9.
10. Mulkern RV, Forbes P, Dewey K, Osganian S, Clark M, Wong S, Ramamurthy U, Kun L, Poussaint TY. Establishment and results of a magnetic resonance quality assurance program for the pediatric brain tumor consortium. Acad Radiol. 2008;15(9):1099–110.
11. Nijm GM, Swiryn S, Larson AC, Sahakian AV. Evaluation of image quality metrics for comparison of synchronization algorithms for cardiac cine MRI. Proc IEEE ICIP. 2008 Oct 12-15; San Diego, California. p. 2260–3.
12. Salem KA, Wilsof DL. A human vision model for the objective evaluation of perceived image quality applied to MRI and image restoration. Proc SPIE. 2002;4791:180–91.
13. Jiang Y, Huo D, Wilson DL. Methods for quantitative image quality evaluation of MRI parallel reconstructions: detection and perceptual difference model. Magn Reson Imaging. 2007;25(5):712–21.
14. Duyn JH, van Gelderen P, Li TQ, de Zwart JA, Koretsky AP, Fukunaga M. High-field MRI of brain cortical substructure based on signal phase. Proc Natl Acad Sci USA. 2007;104(28):11796–801.
15. Rutt BK, Lee DH. The impact of field strength on image quality in MRI. J Magn Reson Imaging. 1996;1:57–62.
16. Piella G, Heijmans H. A new quality metric for image fusion. Proc IEEE ICIP. 2003;3:173–6.
17. Nishimura D. Principles of magnetic resonance imaging. Palo Alto (CA): Stanford University; 1996.

18. Hornak JP. The basics of mri, interactive learning software [textbook on the Internet]. 2010. http://www.cis.rit.edu/htbooks/mri/. Accessed 2010 Feb 9.
19. Ruan C. MRI artifacts: mechanism and control [monograph on the Internet]. http://ric.uthscsa.edu/personalpages/lancaster/DI2Projects2003/MRIArtifacts.pdf. Accessed 2010 Feb 9.
20. Patton JA, Kulkarni MV, Craig JK, Wolfe OH, Price RR, Partain CL, James AE Jr. Techniques, pitfalls and artifacts in magnetic resonance imaging. Radiographics. 1987;7(3):505–19.
21. PULSAR [homepage on the Internet]. Parallel imaging utilizing localized surface-coil acquisition and reconstruction. http://www.ece.tamu.edu/~jimji/pulsarweb/downloads.htm. Accessed 2010 Aug 19.
22. Southern MRI [homepage on the Internet]. Motion artifact image. http://www.southernmri.org. Accessed 2010 Feb 9
23. MRITUTOR [website on the Internet]. MRI artifacts. http://www.mritutor.org/lectures/artifacts508.ppt. Accessed 2010 Feb 9
24. Woodard JP, Carley-Spencer MP. No-reference image quality metrics for structural MRI. Neuroinformatics. 2006;4(3):243–62.
25. VPEG [homepage on the Internet]. The Video Experts Quality Group. http://www.its.bldrdoc.gov/vqeg/. Accessed 2010 Aug 19.
26. Forbes KP, Pipe JG, Karis JP, Heiserman JE. Improved image quality and detection of acute cerebral infarction with PROPELLER diffusion-weighted MR imaging. Radiology. 2002;225(2):551–5.
27. Yeh EN. Advanced image reconstruction in parallel magnetic resonance imaging: constraints and solutions [dissertation]. Cambridge, MA: Massachusetts Institute of Technology; 2005.
28. McGee KP, Manduca A, Felmlee JP, Riederer SJ, Ehman RL. Image metric-based correction (autocorrection) of motion effects: analysis of image metrics. J Magn Reson Imaging. 2000;11(2):174–81.
29. Lin W, Ladinsky GA, Wehrli FW, Song HK. Image metric-based correction (autofocusing) of motion artifacts in high-resolution trabecular bone imaging. J Magn Reson Imaging. 2007;26(1):191–7.
30. Lin W, Ladinsky GA, Wehrli FW, Song HK. Combined rotational/translational motion correction using autofocusing for high-resolution trabecular bone images. Proceedings of the 14th Annual Meeting of ISMRM; 6–12 May 2006; Seattle, Washingon; 2006. p. 1702.
31. Prieto F, Guarini M, Tejos C, Irarrazaval P. Metrics for quantifying the quality of MR images. Proceedings of the 17th Annual Meeting of ISMRM; 2009 Apr 18–24. 2009. Honolulu, Hawaii: 2009. p. 4696.
32. Gardner EA, Ellis JH, Hyde RJ, Aisen AM, Quint DJ, Carson PL. Detection of degradation of magnetic resonance (MR) images: comparison of an automated MR image-quality analysis system with trained human observers. Acad Radiol. 1995;2(4):277–81.
33. Chou CH, Li YC. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. IEEE Trans Circuits Syst. 1995;5(6):467–76.
34. Hill A, Mehnert A, Crozier S, Leung C, Wilson S, McMahon K, Kennedy D. Dynamic breast MRI: Image registration and its impact on enhancement curve estimation. Conf Proc IEEE Eng Med Biol Sci. 2006;1:3049–52.
35. Babu RV, Suresh S, Perkis A. No-reference JPEG-image quality assessment using GAP-RBF. Signal Processing. 2007;87(6):1493–503.
36. Aja-Fernandez S, Estepar RSJ, Alberola-Lopez C, Westin CF. Image quality assessment based on local variance. Conf Proc IEEE Eng Med Biol Sci. 2006;1:4815–8.
37. Peli J. Test of a model of foveal vision by using simulations. J Opt Soc Am A Opt Image Sci Vis. 1996;13(6):1131–8.
38. Srivastava A, Lee B, Simoncelli E, Zhu S. On advances in statistical modeling of natural images. J Math Imag Vis. 2003;18(1):17–33.
39. Wang Z, Sheikh H, Bovik A. No-reference perceptual quality assessment of JPEG compressed images. Proc IEEE ICIP. 2002;1:477–80.
40. MITRE [homepage on the Internet]. http://neuroinformatics.mitre.org/imagequality.html. Accessed 2010 Feb 9.
41. Mortamet B, Bernstein MA, Jack CR Jr, Gunter JL, Ward C, Britson PJ, Meuli R, Thiran JP, Krueger G; Alzheimer's Disease Neuroimaging Initiative. Automatic quality assessment in structural brain magnetic resonance imaging. Magn Reson Med. 2009;62(2):365–72.
42. ADNI [database on the Internet]. Alzheimers Disease Neuroimaging Initiative ADNI database. http://www.adni-info.org. Accessed 2010 Feb 9.
43. Sinha N, Saranathan M, Ramakrishnan KR, Suresh S. Automatic quality assessment in structural brain magnetic resonance imaging. Proc IEEE ICIP. 2007;3:149–52.
44. Sinha N. Strategies for rapid MR imaging [dissertation]. Bangalore, India: Indian Institute of Science; 2008.
45. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.