World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:12, 2015

Subjective Versus Objective Assessment for Magnetic Resonance Images

Heshalini Rajagopal, Li Sze Chow, Raveendran Paramesran

***Abstract*—**Magnetic Resonance Imaging (MRI) is one of the most important medical imaging modality. Subjective assessment of the image quality is regarded as the gold standard to evaluate MR images. In this study, a database of 210 MR images which contains ten reference images and 200 distorted images is presented. The reference images were distorted with four types of distortions: Rician Noise, Gaussian White Noise, Gaussian Blur and DCT compression. The 210 images were assessed by ten subjects. The subjective scores were presented in Difference Mean Opinion Score (DMOS). The DMOS values were compared with four FR-IQA metrics. We have used Pearson Linear Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) to validate the DMOS values. The high correlation values of PLCC and SROCC shows that the DMOS values are close to the objective FR-IQA metrics.

***Keywords*—**Medical Resonance (MR) images, Difference Mean Opinion Score (DMOS), Full Reference Image Quality Assessment (FR-IQA).

I. Introduction

MAGNETIC RESONANCE IMAGING (MRI) is a non-invasive imaging modality that helps physician to diagnose and treat diseases. However, MR images are subjected to artifacts during acquisition, processing, transmission, and reproduction. This may lead to inaccurate diagnosis [1]. Therefore, image quality assessment (IQA) metric that is sensitive to these distortions need to be identified [2]. There are two categories of IQA, which are subjective and objective.

Subjective assessment is rated by human subjects based on their judgment on the image quality. It is regarded as the gold standard to evaluate MR images. However, this method is impractical as it is slow and time consuming [2]. Objective assessment is defined mathematically and more consistent. There are three types of objective assessment: Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA) and No-Reference IQA (NR-IQA). FR-IQA evaluates an image based on its reference image. The reference image should be a perfect image that is free from any distortion. RR-IQA evaluates an image based on the partial information of the reference image. The partial information is a set of features extracted from the reference image. NR-IQA does not use the reference image to evaluate the distorted image [2]. This method is more practical since information on reference image is normally unavailable. Several researches have evaluated MR images using FR-IQA [2], [3].

Kumar et al. performed a comparative analysis on various quality metric for MRI [2]. They measured the quality of the MR images distorted with different levels of blur, noise, compression and contrast using Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise (PSNR), Maximum Difference (MD), etc. Their study showed that SSIM outperfomed all the other metrics used in their study. However, the computational time for SSIM was large.

The goal of IQA is to model an objective assessment metric that correlates with subjective assessment [4]. Thus, to achieve this goal, several research groups have presented database on natural images. They performed experiments on subjective based FR-IQA on natural images. Examples of the database are: LIVE [4] and TID2008 [5]. These databases include the reference images, distorted images which were derived from reference images by applying various types of distortions and the subjective ratings. Kumar et al. performed similar subjective based FR-IQA study on MR images [3]. They evaluated compressed medical images subjectively and compared the subjective ratings which were presented in Mean Opinion Score (MOS) with PSNR and SSIM. They concluded that MOS correlates well with PSNR than SSIM. However, their studies were done on compressed MR images. Hence, we intend to create a database that considers more distortion types that may occur in MR images in real-world application.

In this study, subjective evaluation on 210 MR images which include 10 reference images and 200 distorted images is presented. The reference images were distorted by four types of distortions: Rician Noise, Gaussian White Noise, Gaussian Blur and DCT Compression. Ten subjects were assigned to evaluate the images. The subjective ratings were presented in Difference Mean Opinion Score (DMOS). Similar work was presented by [6], which compared the subjective DMOS with three objective FR-IQA metrics: Signal-to-Noise Ratio (SNR), PSNR and SSIM. We extend the work by comparing DMOS with four FR-IQAs: PSNR, SSIM, Noise Quality Measure (NQM) and Visual Information Fidelity (VIF). The DMOS values were validated using Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC).

II. Methodology

A. MR Images

Ten good MR images were selected from online database: Osirix DICOM Viewer MRI database [7]. The reference MR images were shown in Fig. 1. All the MR images are in

Heshalini Rajagopal, Li Sze Chow and Raveendran Paramesran are with the Department of Electrical, Faculty of Engineering, University of Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia (phone: +603-7967 5226; e-mail: heshalini@gmail.com, lschow@um.edu.my, ravee@um.edu.my).

grayscale and they were normalized to (0,255) for the ease of applying the same values of distortion across all the reference images. The image pixel sizes are stated below each image.

B. Image Distortion Types

Four types of distortions have been applied to the reference images. They are explained as follows:

i. Rician Noise: Rician Noise Probability Density Function (PDF) with standard deviation, σ_R was added to the images.
ii. Gaussian White Noise: Gaussian White Noise Distribution with standard deviation, σ_N was added to the images.
iii. Gaussian Blur: A square kernel window of size 3σ (rounded off) was used with Gaussian kernels (standard deviation, σ_{GB}) for blurring.
iv. Discrete Cosine Transform (DCT): Two Dimensional (2-D) DCT was applied to the image.

These four types of distortions were added to the images because they often occur in MR images. MR images are subjected to Gaussian Noise if the SNR is greater than 2; but subjected to Rician Noise if the SNR is lower than 2 [8]. MR images are subjected to Gaussian Blur if the MR images are exposed to the atmosphere for a long time [9]. DCT compression is a common technique used to compress the wide range of MRI information [10].

C. Test Methodology

The evaluation was done in an office environment with normal indoor illumination level. The images were displayed on 24-in LED monitor with a resolution of 1920 x 1080. Ten subjects (6 male, 4 female) with normal vision were assigned to evaluate the MR images. They are research scholars from Electrical Engineering department, age in between 22 to 35. Visual test for the subjects was done using Snellen Chart. Subjects sat at a distance of 760mm from the Snellen Chart during the vision test to test their near vision acuity. The image evaluation was done using Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method. Two sequences of images were displayed on the monitor where left side was always the reference image whereas the right side was the distorted images. The subject rated the images according to the difference between the two images displayed on the screen. They may rate either: Excellent (90), Good (70), Fair (50), Poor (30) and Bad (10). The numerical scores were not disclosed to the subjects to avoid bias judgment. The image evaluation took less than 20 minutes in average, although there was no time constrain.

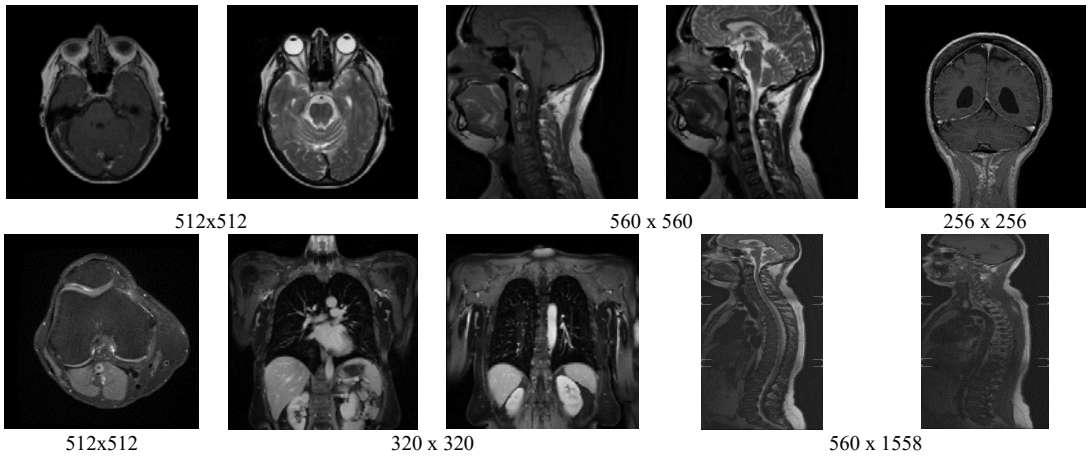

Fig. 1 Ten reference MR images

D. DMOS Calculation

The DMOS was computed by firstly, calculating the raw quality difference scores by m^{th} subject on the n^{th} image, D_{mn} using (1) [4]:

$$D_{mn} = s_{mref(n)} - s_{mn} \quad (1)$$

where s_{mn} is the raw score rated by m^{th} subject for n^{th} image, and $s_{mref(n)}$ is the raw quality score rated by the m^{th} subject for the reference image that correspond to the n^{th} distorted image. Next, D_{mn} was transformed to Z scores using (2) [4]:

$$Z_{mn} = \frac{D_{mn} - \overline{D_m}}{\sigma_m} \quad (2)$$

where $\overline{D_m}$ is the mean of the raw difference scores over all images evaluated by the m^{th} subject, and σ_m is the standard deviation. Then, Z_{mn} was scaled between 1 and 100 using (3). A higher image quality is represented by a lower DMOS value and vice versa.

$$DMOS = \frac{\overline{z_n} - min(\overline{z_n})}{max(\overline{z_n}) - min(\overline{z_n})} \times 100 \quad (3)$$

E. FR-IQA Metrics

The FR-IQA metrics used in this study are PSNR [11], SSIM [12], NQM [13] and VIF [14]. They are explained in Table III.

F. Logistic Regression

Logistic regression is used to construct nonlinear mapping between objective FR-IQA metrics and subjective DMOS

values. The objective FR-IQA (NQM, PSNR, SSIM and VIF) scores after regression, Q_p is calculated as (4) [4]:

$$Q_p = \beta_1 \left(\frac{1}{2} - \frac{1}{1+\exp(\beta_2 (Q-\beta_3))} \right) + \beta_4 Q + \beta_5 \quad (4)$$

where Q is the original objective FR-IQA scores, $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the regression model parameters.

G. Correlation Coefficient

Correlation coefficient is a statistical measure of the relationship between two datasets. In our study, we have used two prominent correlation coefficients: PLCC and SROCC.

PLCC is known as the Pearson product-moment correlation coefficient. It calculates linear correlation between two datasets. It is calculated using (5) [15]:

$$PLCC = \frac{\sum_i^n (x_i - \bar{x}) \sum_i^n (y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \sqrt{\sum_i^n (y_i - \bar{y})^2}} \quad (5)$$

where dataset x represents the DMOS values and dataset y represents the objective scores obtained from FR-IQA.

SROCC is used to measure statistical similarity between two datasets. It is the nonparametric version of Pearson coefficient used for ranked datasets. It measures the relationship between two datasets using a monotonic function. It is calculated using (6) [16]:

$$SROCC = \frac{1 - 6 \sum_{i=1}^n d_i^2}{(n^3 - n)} \quad (6)$$

where d_i is the difference between the ranks for each DMOS and objective scores obtained from FR-IQA data pair, and n is the total number of data pairs.

III. Results and Discussion

A. Results

Fig. 2 shows the scatter plots of DMOS versus standard deviations, σ_R, σ_N and σ_{GB} for Rician Noise, Gaussian White Noise and Gaussian Blur, respectively. Fig. 2 (d) shows the scatter plot of DMOS versus DCT compression rate. The higher the standard deviation for Rician Noise, Gaussian White Noise and Gaussian Blur, the poorer the image quality is. For DCT compression, the image quality gets poorer as the compression rate gets lower. Fig. 2 shows that the DMOS values increases with the distortion levels of the Rician Noise and Gaussian White Noise. However, Fig. 2 (c) shows that the DMOS values did not deviate much although the standard deviation of the Gaussian Blur increases. The same trend is seen in Fig. 2 (d) where there is a wide range of DMOS values regardless of the increasing DCT compression rate.

Fig. 3 shows the graph of DMOS versus NQM with the nonlinear curve fitting for the four types of distortions. The PLCC and SROCC values were also printed on the graph for each distortion.

Tables I and II show the PLCC and SROCC values respectively, between DMOS and the four FR-IQA metrics.

B. Discussion

According to [17], there is a high correlation between two datasets if their correlation coefficient value is larger than 0.68. Referring to Tables I and II, the DMOS and FR-IQA metrics have high correlation.

Figs. 2 (a) and (b) show that the subjects are able to differentiate the images subjected to different levels of distortions for these noises. This is because these noises can cause the low contrast object to be less visible [18], thus affect the visual quality of the MR images.

Figs. 2 (c) and (d) show that the DMOS values did not deviate much although the levels of the distortions increased for Gaussian Blur and DCT compression. This is because Gaussian Blur can only cause smalls objects and fine details to be less visible [19] whereas DCT compression gives very little quality loss in image [20]. Therefore, subjects were unable to differentiate images subjected to different levels of Gaussian Blur and DCT compressions. Nonetheless, we are unable to ensure that the reference MR images used in this study are perfect reference images, since there is no gold standard to confirm this. In fact, FR-IQA is not the best method to assess MR images. Therefore, NR-IQA is more suitable for evaluating MR images. Mortamet et al. performed an experiment to assess MR images without using reference image [1]. They proposed a fully-automatic method to measure image quality of three-dimensional (3D) structural MRI. They explored the fact that 30% to 55% of the MR image's area is occupied by air background. The image quality was assessed by measuring the air background of the images, which was used to detect any distortions. Two quality metrics were proposed in their study. The first quality metric used an atlas-based air background segmentation to extract the air-background region. The second quality metric was based on the noise distribution analysis. The overall results showed that both quality indices are effective. However, their second quality index model seems to be uncertain in cases for parallel-imaging.

Our study was motivated by the experiment done by [21] to evaluate natural scene using NR-IQA method which is known as Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [21]. They used the DMOS calculated by [4] to train a Support Vector Machine (SVM) regression model. Similarly, we have calculated DMOS from the subjective raw scores on MR images. In order to conform that our DMOS values are valid, we have compared them with four FR-IQA metrics and validated them with correlation coefficients, PLCC and SROCC.

From the high correlation coefficients (PLCC and SROCC) in Tables I and II, we can say that the DMOS values calculated from our subjective raw scores are valid and reliable. However, our database is still insufficient to train vector machine to develop a new NR-IQA. Thus, we are currently working on extending this database by using more MR images and distortion types that may occur on MR images in real-world application. Future study will create a larger database of more reference MR images with more distortion

types and to be compared with more FR-IQA metrics for better validation.

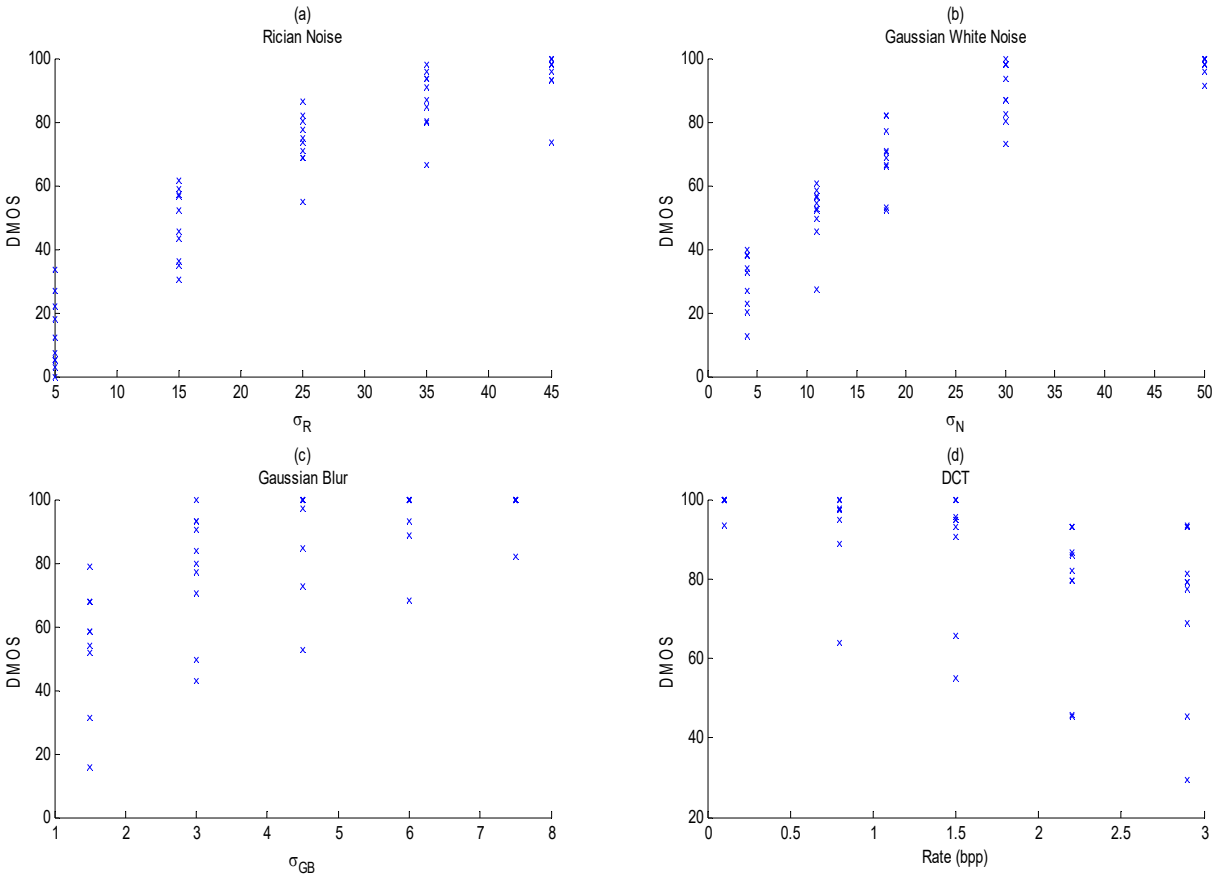

Fig. 2 DMOS values versus standard deviation: (a) σ_R of Rician Noise (b) σ_N of Gaussian White Noise (c) σ_{GB} of Gaussian Blur (d) DMOS values versus compression rate (bpp) of DCT Compression

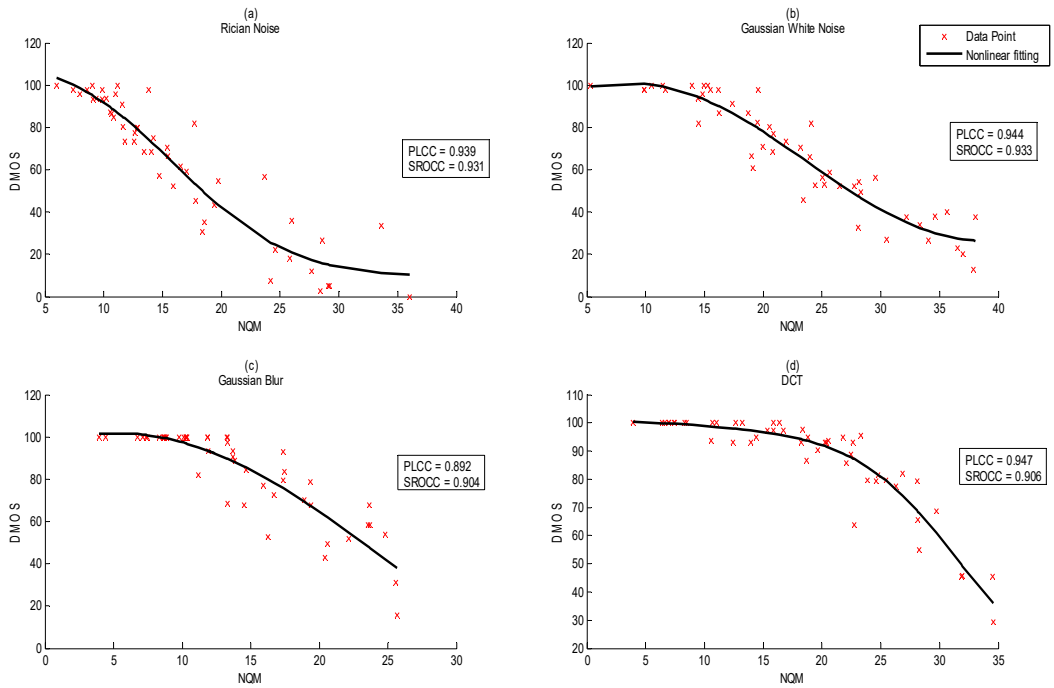

Fig. 3 DMOS versus NQM for (a) Rician Noise (b) Gaussian White Noise (c) Gaussian Blur (d) DCT Compression

IV. Conclusion

We have presented a database of 210 MR images which contains 10 reference images which were distorted with four types of distortions: Rician Noise, Gaussian White Noise, Gaussian Blur and DCT Compression to produce 200 distorted images. The database also includes the DMOS values calculated from the subjective raw score. High correlation coefficients (PLCC and SROCC) values show that the DMOS values are close to the FR-IQA metrics used in this study. Hence, the DMOS is valid and reliable. NQ-IQA is more practical than FR-IQA method in assessing MR images since perfect reference MR images are not always available. Thus, the DMOS that we obtained from this study are applicable for our future study to model a new NR-IQA method.

TABLE I
PLCC between DMOS and FR-IQA Metrics after Nonlinear Regression

Distortion	NQM	PSNR	SSIM	VIF
Rician Noise	0.939	0.959	0.947	0.953
Gaussian White Noise	0.944	0.949	0.929	0.943
Gaussian Blur	0.892	0.763	0.845	0.857
DCT	0.947	0.805	0.836	0.809

TABLE II
SROCC between DMOS and FR-IQA Metrics

Distortion	NQM	PSNR	SSIM	VIF
Rician Noise	0.931	0.937	0.927	0.919
Gaussian White Noise	0.933	0.930	0.909	0.923
Gaussian Blur	0.904	0.787	0.824	0.835
DCT	0.906	0.849	0.779	0.870

Appendix

Let *r(x,y)* represents the reference image and *t(x,y)* represents the distorted image. n_x and n_y are the size of the image in pixels across *x* and *y* dimensions. Both *r(x,y)* and *t(x,y)* should have the same size.

TABLE III
Formulas for the FR-IQA Metircs used in This Study

FR-IQA Metrics	Description
PSNR [11]	Ratio of peak signal power to average noise power. $PSNR = 10log_{10}\left[\frac{max(r(x,y))^2}{\frac{1}{n_x \cdot n_y}\sum_1^{n_x}\sum_1^{n_y}[r(x,y)-t(x,y)]^2}\right]$
SSIM [12]	Captures the loss in the structure of the image. $SSIM = \frac{(2\mu_r\mu_t + C_1)(2\sigma_{rt} + C_2)}{(\mu_r^2 + \mu_t^2 + C_1)(\sigma_r^2 + \sigma_t^2 + C_2)}$ where μ_r and μ_t are the mean intensity for the reference and distorted images respectively; σ_r and σ_t are the standard deviation for the reference and distorted images respectively; σ_{rt} is estimated as: $\sigma_{rt} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i - \mu_r)(t_i - \mu_t)$ where $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ where *L* is the dynamic range of the pixels values (i.e. 255 for 8-bit grayscale images, as in our case), $K_1 = 0.01$ and $K_2 = 0.03$.
NQM [13]	A measure of additive noise. It is designed based on Peli's contrast pyramid. $NQM(dB) = 10\ log_{10}\left(\frac{\sum_x\sum_y O_S^2(x,y)}{\sum_x\sum_y\left(O_S^2(x,y) - I_S(x,y)\right)^2}\right)$ where $O_S^2(x,y)$ and $I_s(x,y)$ represent the simulated versions of the model restored image and the restored images, respectively.
VIF [14]	Measures image information by computing two mutual information quantities from the reference and distorted images. $VIF = \frac{\sum_{j\in subbands} I(\vec{C}^{N,j};\vec{T}^{N,j}\|s^{N,j})}{\sum_{j\in subbands} I(\vec{C}^{N,j};\vec{R}^{N,j}\|s^{N,j})}$ where the subbands of interest are summed over, and $\vec{T}^{N,j}$ represent the subband in the test image, $\vec{R}^{N,j}$ represent the subband in the reference image, $\vec{C}^{N,j}$ represent *N* elements of the RF C_j that describes the coefficient subband *j*, and so on.

Acknowledgment

This research was funded by BKP grant (BK053-2014) from the University of Malaya. We like to thank all the volunteers involved in the subjective assessment in this study.

References

[1] B. Mortamet, M. a. Bernstein, C. R. Jack, J. L. Gunter, C. Ward, P. J. Britson, R. Meuli, J. P. Thiran, and G. Krueger, "Automatic quality assessment in structural brain magnetic resonance imaging," *Magn. Reson. Med.*, vol. 62, pp. 365–372, 2009.

[2] R. Kumar and M. Rattan, "Analysis Of Various Quality Metrics for Medical Image Processing," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 11, pp. 137–144, 2012.

[3] K. Bindu, A. Ganpati, and A. K. Sharma, "A Comparative Study of Image Compression Algorithms," *Int. J. Res. Comput. Sci.*, vol. 2, no. 5, pp. 37–42, 2012.

[4] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *Image Process. IEEE Trans.*, vol. 15, no. 11, pp. 3441–3452, 2006.

[5] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, 2008, pp. 403–408.

[6] L. S. Chow and H. Rajagopal, "Comparison of Difference Mean Opinion Score (DMOS) of Magnetic Resonance Images with Full-Reference Image Quality Assessment (FR-IQA)," in *Image Processing, Image Analysis and Real-Time Imaging (IPARTI) Symposium*, 2015.

[7] "MR images from Osirix DICOM Viewer." (Online). Available: http://www.osirix-viewer.com/datasets/. (Accessed: 20-Jan-2015).

[8] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magn. Reson. Med.*, vol. 34, no. 6, pp. 910–914, 1995.

[9] A. Debnath, H. M. Rai, C. Yadav, and A. Bhatia, "Deblurring and Denoising of Magnetic Resonance Images using Blind Deconvolution Method," *Int. J. Comput. Appl.*, vol. 81, no. 10, pp. 7–12, 2013.

[10] R. L. de Queiroz, "DCT approximation for low bit rate coding using a conditional transform," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, vol. 1, pp. 237–240.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Process. IEEE Trans.*, vol. 13, no. 4, pp. 600–612, 2004.

[13] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and a C. Bovik, "Image quality assessment based on a degradation model.," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, 2000.

[14] H. R. Sheikh and A. C. Bovik, "Image information and visual quality.," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.
[15] X.-K. Song, *Correlated data analysis: modeling, analytics, and applications*. Springer Science & Business Media, 2007.
[16] T. D. Gautheir, "Detecting Trends Using Spearman's Rank Correlation Coefficient," *Environmental Forensics*, vol. 2, no. 4. Taylor & Francis, pp. 359–362, 2001.
[17] R. Taylor, "Interpretation of the Correlation Coefficient: A Basic Review," *J. Diagnostic Med. Sonogr.*, vol. 6, no. 1, pp. 35–39, 1990.
[18] A. Vibhakar, M. Tiwari, and J. Singh, "Performance Analysis for MRI Denoising using Intensity Averaging Gaussian Blur Concept and its Comparison with Wavelet Transform Method," *Int. J. Comput. Appl.*, vol. 58, no. 15, pp. 21–26, 2012.
[19] M. Ertas, I. Yildirim, M. Kamasak, and A. Akan, "An iterative tomosynthesis reconstruction using total variation combined with non-local means filtering.," *Biomed. Eng. Online*, vol. 13, no. 1, p. 65, 2014.
[20] A. B. Watson, "Image Compression Using the Discrete Cosine Transform," *Math. J.*, vol. 4, no. 1, pp. 81–88, 1994.
[21] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *Image Process. IEEE Trans.*, vol. 21, no. 12, pp. 4695–4708, 2012.