Comprehensive Quantitative Image Quality Evaluation of Compressed Sensing MRI Reconstructions Using a Weighted Perceptual Difference Model (Case-PDM): Selective Evaluation, Disturbance Calibration and Aggregative Evaluation of Noise, Blur, Aliasing and Oil-Painting Artifacts

Jun Miao[1], Feng Huang[2], and David L. Wilson[1, 3]
[1]Department of Biomedical Engineering, Case Western Reserve University
[2]Invivo Corporation
[3]Department of Radiology, University Hospitals of Cleveland & Case Western Reserve University

ABSTRACT

The perceptual difference model (Case-PDM) is being used to quantify image quality of fast MR acquisitions and sparse reconstruction algorithms as compared to slower, full k-space, high quality reference images. To date, most perceptual difference models average image quality over a wide range of image degradations and assume that the observer has no bias towards any of them. Here, we create metrics weighted to different types of artifacts, calibrated to a human observer's preference, and then aggregate them to produce a comprehensive evaluation. The selective PDM is tuned using test images from an input reference image degraded by noise, blur, aliasing, or "oil-painting." To each artifact, responses of cortex channels in the PDM are normalized to be weights used for selective evaluation. A pair comparison experiment based on functional measurement theory was used to calibrate selective PDM score of each artifact to its measured disturbance. Test images of varying quality were from identical reference image degraded by one type of artifact. We found that human observers rated aliasing > blur > oil-painting > noise. In order to validate the new evaluation approach, PDM scores were compared to human ratings across a large set of compressed sensing MR reconstruction test images of varying quality. Human ratings (i.e. overall, noise, blur, aliasing, and oil-painting ratings) were obtained from a modified Double Stimulus Continuous Quality Scale experiment. For 3 brain images (transverse, sagittal, and coronal planes), averaged r values [comprehensive-PDM, noise-PDM, blur-PDM, aliasing-PDM, oil-painting-PDM] were [0.947±0.010, 0.827±0.028, 0.913±0.005, 0.941±0.016, 0.884±0.025]. We conclude the weighted Case-PDM is useful for selectively evaluating MR reconstruction artifacts and the proposed comprehensive PDM score can faithfully represent human evaluation, especially when demonstrating artifact bias, of compressed sensing reconstructed MR images.

Keywords: Perceptual difference model, Image quality, Image artifact, Noise, Blur, Aliasing, Oil-Painting, Compressed Sensing, Magnetic Resonance Imaging, DSCQS, FMT

1. INTRODUCTION

Magnetic Resonance Imaging (MRI) has been developed very fast since it was first proposed by Lauterbur and Mansfield in 1973. In 2003, there were approximately 10,000 MRI units worldwide, and approximately 75 million MRI scans were performed. MRI has many advantages of over other medical imaging modalities: it is very flexible and sensitive to a broad range of tissue properties; it is a non-invasive operation and applicable to people of almost any age.

Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment,
edited by David J. Manning, Craig K. Abbey, Proc. of SPIE Vol. 7627, 762709

Although researchers are able to analyze the original k-space data from the MR scanners, in most cases, only magnitude images are examined, in the way that is very similar to the traditional X-ray and CT imaging. Since images are the "final product" of MRI, appropriate methods to evaluate the image quality are very important. With the recent development of the fast MR imaging techniques, thousands of images could be easily generated by combining different parameters in MR acquisition and reconstruction. A convenient and accurate mathematic model to evaluate the MR image quality quantitatively would be very helpful for the MR researchers and radiologists.

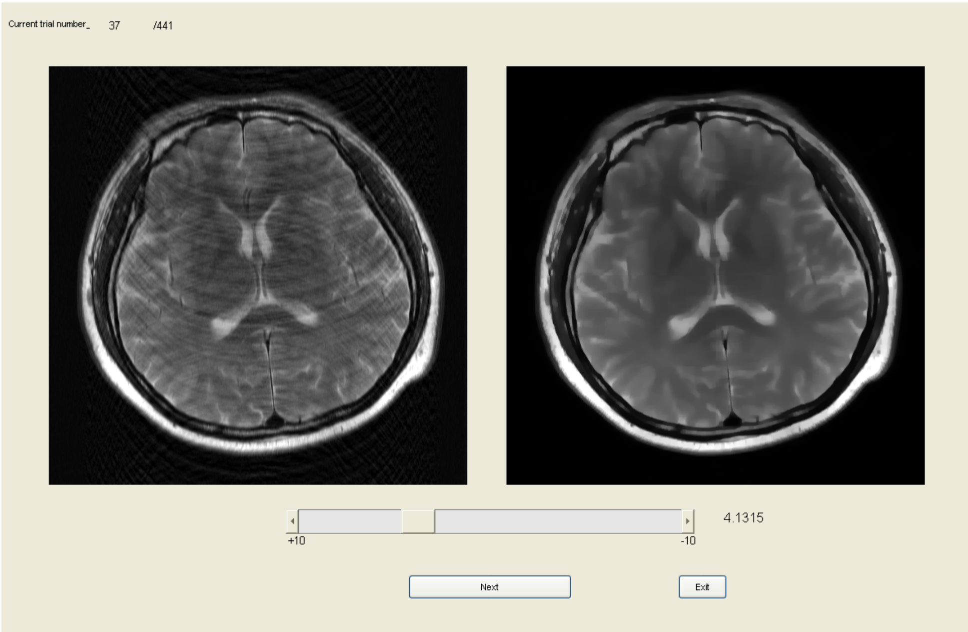
The perceptual difference model (Case-PDM) is being used to quantify image quality of fast, parallel MR acquisitions and reconstruction algorithms by comparing to slower, full k-space, high quality reference images. Case-PDM objectively, quantitatively evaluates image quality, and we have found it to be quite useful in investigations of keyhole, spiral, SENSE, and GRAPPA applications [1-4]. To date, most perceptual difference models average image quality over a wide range of image degradations. In this paper, we propose a method to tune PDM so that it can selectively evaluate MR reconstruction artifacts and to create a comprehensive PDM evaluation with human observer's bias toward different artifacts considered. We design human subject experiment to calibrate and validate the proposed method in the context of MR image quality evaluation.

2. METHODS

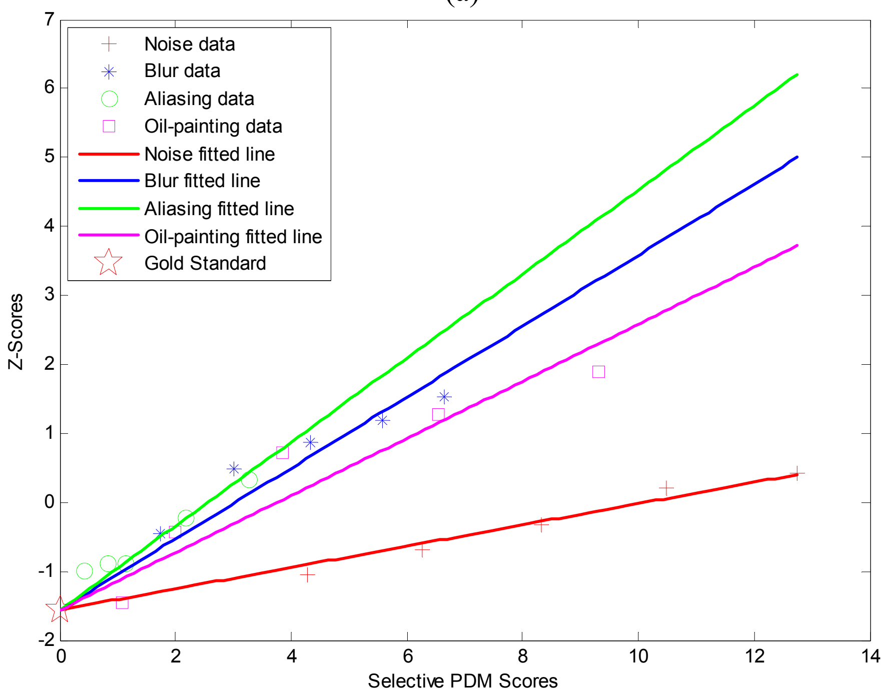
The selective PDM is tuned using test images from an input reference image degraded by noise, blur, aliasing, or oil-painting (which is a typical effect in total variation regulated compressed sensing reconstruction). To each artifact, responses of cortex channels in the PDM are normalized to be weights used for selective evaluation. Therefore, the Case-PDM output can be decomposed into noise, blur, aliasing, or oil-painting effect visual difference maps. A comparison experiment based on functional measurement theory (FMT) was used to calibrate selective PDM score of each artifact to its measured disturbance. In such experiment (GUI is shown in Fig. a), two images were presented side-by-side to a subject and he/she needs to rate how much one (left) is better than another (right). Test images of varying quality are from identical reference image degraded by one type of artifact. A calibration curve was then obtained by linearly fitting disturbances of test images into the corresponding selective PDM scores. And a comprehensive PDM score is created by adding up calibrated selective PDM scores of each artifact. A modified DSCQS (Double Stimulus Continuous Quality Scale) experiment was designed to obtain human subject ratings (i.e. overall, noise, blur, aliasing, and oil-painting ratings). The experiment covered 3 different brain images (transverse, sagittal, and coronal planes). Test images were reconstructed by a TV-L_1 compressed sensing algorithm [5]. In total, 120 test images covering a wide range of image quality were evaluated (40 test images for each of the 3 datasets). There were 2 human subjects participating in the experiment. Linear correlation coefficients were calculated from pairs of human ratings and PDM scores among the test images.

3. RESULTS AND CONCLUSION

For each kind of artifact, disturbance measured from the FMT experiment is linearly fitted to its selective PDM score with the relationship $Y = a*X + b$, where Y denotes disturbance in terms of z-score, X denotes selective PDM score, a denotes slope, and b denotes disturbance of reference image (for the same reference image, b is identical for all artifacts). Figure b is a result from a subject rating one of the datasets. In all datasets, we found that 4 artifact's disturbances are biased by human observers as aliasing > blur > oil-painting > noise. From inter-subject analysis, we found there is a small variation among their calibration curves. For the transverse brain dataset, the ratio of standard deviation of a's to averaged a's is 8.5%, 17.5%, 36.8%, and 3.4% for calibration curves of noise, blur, aliasing, and oil-painting, respectively. A comprehensive PDM evaluation in terms of z-score can be obtained by adding up all the transformed selective PDM scores from the corresponding calibration curve. We found that both the comprehensive PDM score and the selective PDM score have good correlation with human ratings. Human ratings (i.e. overall, noise, blur, aliasing, and oil-painting ratings) were obtained from a modified DSCQS experiment. For 3 brain images (transverse, sagittal, and coronal planes), averaged r values [comprehensive-PDM, noise-PDM, blur-PDM, aliasing-PDM, oil-painting-PDM] were [0.947±0.010, 0.827±0.028, 0.913±0.005, 0.941±0.016, 0.884±0.025].



(a)



(b)

Figure (a) is a GUI for FMT experiment. Subject needs to rate how much left image (aliasing artifact shown in image) is better than right image (oil-painting artifact shown in image) by giving a scalar number from -10 to 10. (b) is a plot for calibration curves. For the same PDM score, artifact disturbances are biased as aliasing > blur > oil-painting > noise.

We conclude the weighted Case-PDM is useful for selectively evaluating MR reconstruction artifacts and the proposed comprehensive PDM score can faithfully represent human evaluation of compressed sensing reconstructed MR images.

4. ACKNOWLEDGMENTS

This work was supported under NIH grant R01 EB004070 and the Research Facilities Improvement Program Grant NIH C06RR12463-01. We thank the subjects for participating in the experiments.

REFERENCES

1. K.A.Salem, J.S.Lewin, A.J.Aschoff, J.L.Duerk, and D.L.Wilson, "Validation of a human vision model for image quality evaluation of fast interventional magnetic resonance imaging," *Journal of Electronic Imaging*, vol. 11, pp. 224-235, 2002.
2. D.L.Huo, K.A.Salem, Y.Jiang, and D.L.Wilson, “Optimization of Spiral MRI using a Perceptual Difference Model”, *International Journal of Biomedical Imaging*, vol. 2006.
3. D.L.Huo, D.Xu, Z.P.Liang, and D.Wilson, "Application of perceptual difference model on regularization techniques of parallel MR imaging," *Magnetic Resonance Imaging*, vol. 24, pp. 123-132, 2006.
4. J. Miao, D. Huo and D.L. Wilson, “Quantitative image quality evaluation of MR images using perceptual difference models,” *Medical Physics*, Vol. 35, pp. 2541-2553, 2008.
5. J. Yang, Y. Zhang and W. Yin, "A Fast TVL1-L2 Minimization Algorithm for Signal Reconstruction from Partial Fourier Data", Technical Report, TR08-27, CAAM, Rice University.