Bio-Medical Materials and Engineering 24 (2014) 2865–2873
DOI 10.3233/BME-141105
IOS Press

Generalized relative quality assessment scheme for reconstructed medical images

Shaoze Wang[a], Yong Ding[a,*], Hang Dai[a], Dahong Qian[a], XinfengYu[b]and MinmingZhang[b]
[a]*Institute of VLSI Design, Zhejiang University, Hangzhou, Zhejiang 310027, China*
[b]*Radiology Department, The Second Affiliated Hospital, Medical School of Zhejiang University, Hangzhou, Zhejiang 310027, China*

Abstract. A generalized relative quality (RQ) assessment scheme is proposed here based on the Bayesian inference theory, which is reasonable to make use of full reference (FR) algorithms when the evaluation of the quality of homogeneous medical images is required. Each FR algorithm is taken as a kernel to represent the level of quality. Although, various kernels generate different order of magnitude, a normalization process can rationalize the quality index within 0 and 1, where 1 represent the highest quality and 0 represents the lowest quality. To validate the performance of the proposed scheme, a series of reconstructed susceptibility weighted imaging images are collected, where each image has its subjective scale. Both experimental results and a ROC analysis show that the RQ obtained from the proposed scheme is consistent with subjective evaluation.

Keywords: Relative quality, full reference, Bayesian inference, reconstructed image, subjective evaluation

1. Introduction

Medical images obtained by radiology techniques have been playing an important role in clinical examination, diagnosis of disease, as well as post-treatment period [1]. Different reconstruction strategies generate reconstructed images with different quality. To select the best reconstruction strategies, there is a need to evaluate the quality parameters of these reconstructed images based on the distortion level. Such quality parameters are often evaluated by radiologists, which is cumbersome and inaccurate. Therefore, it is necessary to develop computer-aided assessment (CAA) algorithm.

Till now, several researches related to CAA are proposed [2–5]. J.Q Liu et al. [6] compared seven commonly used quality assessment methods based on PET/CT reconstruction; M. Razaak et al. [2] presented broad categories of quality assessment metrics. These methods or metrics are affiliated with the full-reference (FR) methodology, such as mean square error (MSE), peak signal to noise ratio (PSNR), universal quality index (UQI) [1], spatial frequency measurement (SFM) [7], and structure similarity index metric (SSIM) [8]. These FR algorithms measure the degradation or difference between distorted images and original images, and thus, FR can give the relative quality (RQ) if and only if the original image is provided. But in reality, original images cannot be directly compared with

*Corresponding author: Yong Ding, Institute of VLSI Design, Zhejiang University, Hangzhou, Zhejiang 310027, China. Tel.: +86 0571 87951071; Fax: +86 87942486; E-mail: dingy@vlsi.zju.edu.cn.

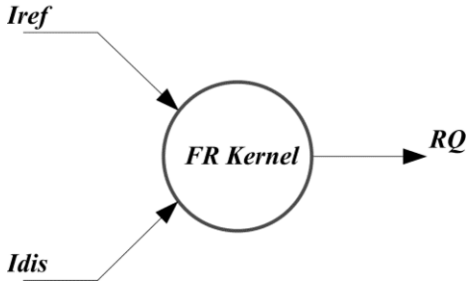

Fig. 1. Interface with FR kernel that inputs reference image and distorted image, and outputs RQ which reflects the distortion degree of the distorted image.

distorted images, as both kinds of images have different format, e.g. an original susceptibility weighted imaging (SWI) image is formed by one amplitude image and one phase image, whereas a reconstructed image is acquired by multiplication of both the amplitude and phase images. It is observed that all FR algorithms have the same limitation.

In this paper, a generalized relative quality assessment algorithm scheme is proposed. Due to the foundation of the Bayesian inference theory, FR algorithms can successfully calculate the RQ of distorted images even if original images are not accessible. The remaining part of this paper is arranged as follow. Section 2 reviews common FR quality methods and section 3 introduces the proposed model utilizing the Bayesian inference theory. Experiment and results are presented in Section 4. Discussions are provided in Section 5 and conclusion is presented in Section 6.

2. Common FR quality methods

As mentioned above, common FR quality assessment methods include MAE, PSNR, UQI, SFM, and SSIM et al. All of them need both an original image and a distorted image, denoted by I_{ref} and I_{dis}, respectively. For medical images, the proposed method assumes that three conditions should be satisfied:

- I_{ref} and $I_{\text{dis}}$are both obtained with the same imaging principal. Therefore, they have the same imaging type, e.g. X-ray or SWI.
- The content of I_{ref} and I_{dis} should be similar. Both focus on the same area such as brain or lung.
- Images should be calibrated according to the same standard.

Once images satisfy these conditions, they can be evaluated using each of the common algorithms. These FR algorithms provide an interface between two images. Figure 1 illustrates the interface that inputs I_{ref} and I_{dis}, and outputs an index. Inside the interface, there are three paradigms of FR kernel: pixel-based (MSE, PSNR), block-based (UQI, SSIM), and hybrid-based (SFM). RQ of I_{dis} can be presented by a FR kernel:

$$\mathrm{RQ}(\mathrm{I_{dis}}) = \mathrm{FR}\left(I_{ref}, I_{dis}\right) \tag{1}$$

where the details of FR (I_{ref}, I_{dis}) depend on which algorithm researchers select. Table 1 sums up the closed-form expression of common FR algorithms [2].

Table 1
Instances of FR kernel: closed-form expressions of MSE, PSNR, UQI, SSIM, and SFM

Paradigm	Algorithm	Expression
Pixel-based	MSE	$\mathrm{RQ}(\mathrm{I_{dis}}) = \frac{1}{MN}\sum_{x=1}^{N}\sum_{y=1}^{M}\left[I_{dis}(x,y) - I_{ref}(x,y)\right]^2$
	PSNR	$\mathrm{RQ}(\mathrm{I_{dis}}) = 20\log\frac{255}{\sqrt{\frac{1}{MN}\sum_{x=1}^{N}\sum_{y=1}^{M}\left[I_{dis}(x,y) - I_{ref}(x,y)\right]^2}}$
Block-based	UQI	$\mathrm{RQ}(\mathrm{I_{dis}}) = \sum_{k=1}^{n}\frac{4\sigma_{X_kY_k}\overline{X_k}\cdot\overline{Y_k}}{\left(\sigma_{X_k}{}^2+\sigma_{Y_k}{}^2\right)\left(\overline{X_k}^2+\overline{Y_k}^2\right)}$
	SSIM	$\mathrm{RQ}(\mathrm{I_{dis}}) = \frac{1}{n}\sum_{k=1}^{n}\frac{(2\overline{X_k}\overline{Y_k}+C_1)(2\sigma_{X_kY_k}+C_2)}{\left(\overline{X_k}^2+\overline{Y_k}^2+C_1\right)\left(\sigma_{X_k}{}^2+\sigma_{Y_k}{}^2+C_2\right)}$
Hybrid-based	SFM	$\mathrm{RQ}(\mathrm{I_{dis}}) = \sqrt{\left(\sum_{x=1}^{N}\sqrt{\sum_{y=1}^{M}\left[I_{dis}(x,y) - I_{ref}(x,y)\right]^2}\right)^2 + \left(\sum_{y=1}^{M}\sqrt{\sum_{x=1}^{N}\left[I_{dis}(x,y) - I_{ref}(x,y)\right]^2}\right)^2}$

3. Bayesian inference

Based on the Bayesian inference, FR algorithms are still able to evaluated RQ of distorted images even though original images are missing.

Before deducting the RQ expression using Bayesian inference, several variables are necessary to be defined: Group $\{I_{\mathrm{ref}}, I_{\mathrm{dis}}{}^1, I_{\mathrm{dis}}{}^2, \ldots I_{\mathrm{dis}}{}^{\mathrm{N}}\}$ denotes a series of distorted medical images. There are N distorted images and one reference image. All the distorted images are required to meet the three above mentioned conditions. In addition, some more incidents are stated. Incident R represents that the reference image I_{ref} exists, and with highest quality without doubt. Secondly, incident $\mathrm{D^i}$ represents that the quality of the distorted image $I_{\mathrm{dis}}{}^{\mathrm{i}}$ has also the highest quality.

Now the RQ of a distorted image $I_{\mathrm{dis}}{}^{\mathrm{i}}$ can easily be expressed by $\mathrm{P(D^i|R)}$, where the P(*) represents the probability. When the original image is not accessible (P(R)=0),the form of RQ is changed to $\mathrm{P(D^i|R')}$. According to Bayesian inference theory, there is

$$\mathrm{P}(\mathrm{D_i}|\mathrm{R}') = \frac{\mathrm{P}(\mathrm{R}'|\mathrm{D_i})\mathrm{P}(\mathrm{D_i})}{\sum_{\mathrm{j}=1}^{\mathrm{N}}\mathrm{P}(\mathrm{R}'|\mathrm{D_j})\mathrm{P}(\mathrm{D_j})} \tag{2}$$

where $\mathrm{P(R'|D_i)=1-P(R|D_i)=1}$, therefore, Eq. (2) can be further simplified to:

$$\mathrm{P}(\mathrm{D_i}|\mathrm{R}') = \frac{\mathrm{P}(\mathrm{D_i})}{\sum_{\mathrm{j}=1}^{\mathrm{N}}\mathrm{P}(\mathrm{D_j})} \tag{3}$$

Eq. (3) indicates that RQ of I_{dis}^{i} is relevant to the other distorted images, which means that each distorted image can be regarded as the original image with the equal probability:

$$P(D_i) = \sum_{j=1}^{N} P(D_i|D_j)P(D_j) = \frac{1}{N}\sum_{j=1}^{N} P(D_i|D_j) \tag{4}$$

Therefore, Eq. (3) can be implemented by

$$P(D_i|R') = \frac{\sum_{j=1}^{N} P(D_i|D_j)P(D_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} P(D_i|D_j)P(D_j)} = \frac{\frac{1}{N}\sum_{j=1}^{N} P(D_i|D_j)}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} P(D_i|D_j)} = \frac{\sum_{j=1}^{N} P(D_i|D_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} P(D_i|D_j)} \tag{5}$$

For each $P(D_i|R')$, the denominator remains the same and the $P(D_i|D_i)$ always equals to unit. In particular, some distorted medical images may include several slices, say, *M* slices. Considering all these elements, the RQ of I_{dis}^{i} is modified:

$$RQ(I_{dis}^{i}) = \sum_{j\neq i}^{N}\sum_{k=1}^{M} P(D_{i,k}|D_{j,k}) \tag{6}$$

where $P(D_{i,k}|D_{j,k})$ is indeed a FR process. Therefore, the final RQ expression is given by

$$RQ(I_{dis}^{i}) = \frac{1}{N-1}\frac{1}{M}\sum_{j\neq i}^{N}\sum_{k=1}^{M} FR\left(I_{ref}^{j,k}, I_{dis}^{i,k}\right) \tag{7}$$

where1/[(N-1)M] is taken to give an averaged RQ.

4. Experimental results

4.1. Testing images

Test images used for validation of this proposed algorithm are SWI images which were processed on the basis of the original magnitude and phase images. Reconstruction process involves high pass filter with the block size of 64×64, followed by consecutive 4 times phase multiplication and 4-slices minimum intensity projection. Three echoes with time echo (TE) of 23.144 ms, 29.192 ms, and 35.24 ms were added and averaged, respectively. Therefore, the process generates five reconstructed images, i.e. swi23 (TE=23.144 ms), swi29 (TE=29.192 ms), swi35 (TE=35.24 ms), swia (added echoes), and swiw (averaged echoes). Each SWI image consists of nine slices. There are 8 testing series labeled from No. 012 to No. 019. Each series contains 5 SWI images or 45 slices. Figure 2 illustrates the slices in the No. 019 series. 9 slices of each SWI image are arranged in columns and there are five rows denoting five SWI images. All these images have been subjectively evaluated using double-blind review. The subjective scale ranges from 1 to 5. Scale of 5 denotes the best quality and scale of 1 the worst.

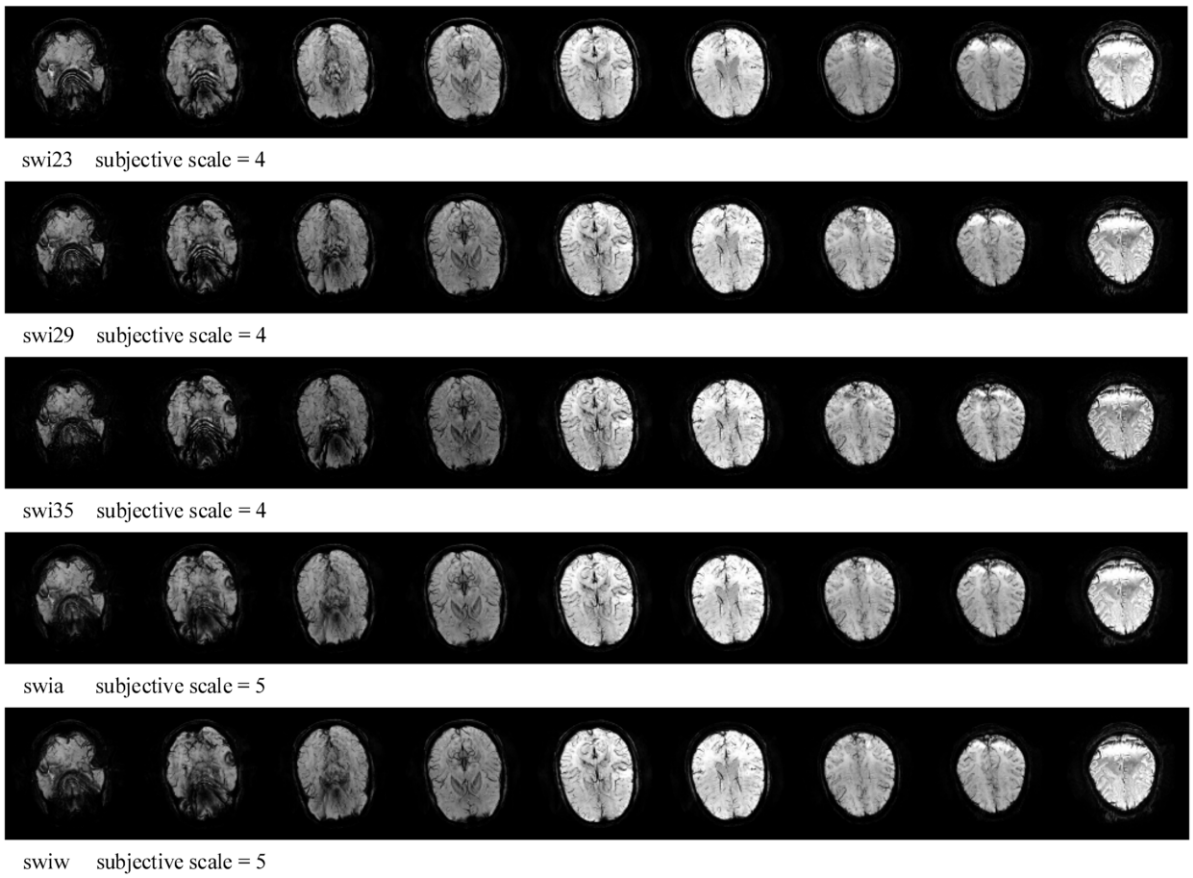

Fig. 2. The five SWI images in No.019 series: each image contains 9 slices which is arrayed in columns. Five SWI images placed from top to bottom are swi23, swi29, swi35, swia, and swiw. Subjective evaluation scales are also provided. Scale 5 denotes the best quality and scale 1 denotes the worst.

4.2. Testing results

To test the performance of the proposed method, three common FR algorithms (PSNR, SFM, and SSIM) are used as the FR kernel. Table 2 lists the RQs which are obtained through Eq. (7). It should be noted that different algorithms generate RQ in different levels, e.g. RQs calculated by PSNR are almost more than 20 while RQs calculated by SSIM are less than 1. The difference is dominated by different equations they adopt, shown in Table 1.

5. Discussion

The first observation is that, for PSNR and SSIM, high index indicates high RQ, whereas for SFM, low index denotes high RQ. Figure 3 plots RQs in the form of SSIM index in each series. It is apparent from plot that swia and swiw have the better quality than the others, which is consistent with subjective evaluation.

Table 2

RQ values with FR kernel based on PSNR, SFM, and SSIM, respectively. 012-019 denotes the eight series. It should be noted that different algorithms generate RQ in different levels, which dominated by different equations, shown in Table 1

		012	013	014	015	016	017	018	019
PSNR	swi23	23.18	23.46	25.19	23.19	26.02	23.62	22.50	23.10
	swi29	25.12	24.98	27.01	24.59	28.09	23.62	23.97	26.14
	swi35	24.19	23.76	25.74	16.64	26.83	24.20	22.84	23.63
	swia	35.71	33.61	36.92	28.53	35.45	32.37	24.39	30.79
	swiw	35.71	33.59	36.92	28.81	35.48	32.32	24.52	31.01
SFM	swi23	89905	74172	64313	153814	64069	78878	138748	122939
	swi29	58893	60158	47360	136256	40757	100841	110241	68226
	swi35	71481	68458	53707	416351	62782	80681	126020	117296
	swia	42874	38599	33757	138605	32059	50298	138510	65826
	swiw	42770	39276	33894	129959	31364	51860	127467	59836
SSIM	swi23	0.82	0.77	0.84	0.81	0.87	0.84	0.81	0.82
	swi29	0.84	0.80	0.86	0.84	0.89	0.84	0.84	0.86
	swi35	0.82	0.78	0.84	0.76	0.88	0.84	0.81	0.83
	swia	0.92	0.90	0.93	0.90	0.94	0.93	0.91	0.93
	swiw	0.92	0.90	0.93	0.91	0.94	0.93	0.90	0.92

Secondly, for each FR function, the probability that a reconstructed image has the best quality, denoted by $P(I_{dis}^{i})$, can be obtained through

$$P\left(I_{dis}^{i}\right) = \begin{cases} \dfrac{RQ\left(I_{dis}^{i}\right)}{\sum_{i=1}^{N} RQ\left(I_{dis}^{i}\right)}, & FR=PSNR,SSIM \\ \dfrac{\sum_{i=1}^{N} RQ\left(I_{dis}^{i}\right) - RQ\left(I_{dis}^{i}\right)}{\sum_{i=1}^{N}\left(\sum_{i=1}^{N} RQ\left(I_{dis}^{i}\right) - RQ\left(I_{dis}^{i}\right)\right)}, & FR=FSM \end{cases} \tag{8}$$

Table 3 presents the probabilities in all series. It is obvious that for any FR kernel and any series, the probability always indicates that swia and swiw have the better quality, which exhibits the robustness of the proposed scheme.

Finally, when probabilities for all images are obtained, a threshold is set to make a simple two-class classifier to determine which image has the best quality. According to subjective scaling, images name swia and swiw have the best quality, which is the target of the classifier. On the other hand, because each of the five images has one fifth probability to be the best, this classification should be supervised, which means that the threshold must be no less than 1/5. In the ROC analysis [9], four thresholds (0.20, 0.21, 0.22, and 0.23) are set for the classifier. Figure 4 plots the ROC curves and the area under curve (AUC) is measured. It can be observed that when threshold is between 0.20 and 0.21, the classifier has the right rate above 92%. Even though the AUC drops to 0.65 when threshold is beyond 0.22, all curves are above the gray line, in other words, all the four classifiers are meaningful.

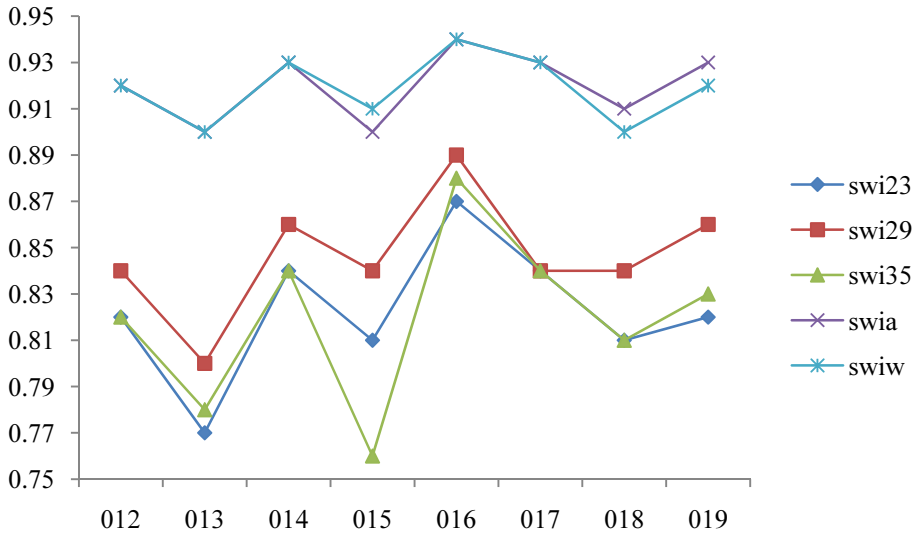

Fig. 3. RQs from SSIM kernel: horizontal coordinate denotes the series number form No. 012 to No. 019; vertical coordinate marks the RQs in the form of SSIM index. The index is between 0 and 1 and it represents a better quality when it is closer to 1. Therefore, swia and swiw have the better quality than the others, which is consistent with subjective evaluation.

Table 3

Probabilities that a reconstructed image has the best quality

		012	**013**	**014**	**015**	**016**	**017**	**018**	**019**
PSNR	swi23	0.1611	0.1683	0.1660	0.1905	0.1713	0.1735	0.1903	0.1715
	swi29	0.1746	0.1792	0.1780	0.2020	0.1850	0.1735	0.2028	0.1941
	swi35	0.1681	0.1704	0.1696	0.1367	0.1767	0.1778	0.1932	0.1755
	swia	0.2481	0.2411	0.2432	0.2343	0.2334	0.2378	0.2063	0.2286
	swiw	0.2481	0.2410	0.2432	0.2366	0.2336	0.2374	0.2074	0.2303
SFM	swi23	0.1765	0.1839	0.1810	0.2106	0.1807	0.1956	0.1959	0.1792
	swi29	0.2019	0.1964	0.1992	0.2151	0.2059	0.1805	0.2070	0.2107
	swi35	0.1916	0.1890	0.1924	0.1432	0.1821	0.1944	0.2008	0.1825
	swia	0.2150	0.2156	0.2138	0.2145	0.2153	0.2153	0.1960	0.2121
	swiw	0.2150	0.2150	0.2136	0.2167	0.2161	0.2142	0.2003	0.2155
SSIM	swi23	0.1898	0.1855	0.1909	0.1919	0.1925	0.1918	0.1897	0.1881
	swi29	0.1944	0.1928	0.1955	0.1991	0.1969	0.1918	0.1967	0.1972
	swi35	0.1898	0.1880	0.1909	0.1801	0.1947	0.1918	0.1897	0.1904
	swia	0.2130	0.2169	0.2114	0.2133	0.2080	0.2123	0.2131	0.2133
	swiw	0.2130	0.2169	0.2114	0.2156	0.2080	0.2123	0.2108	0.2110

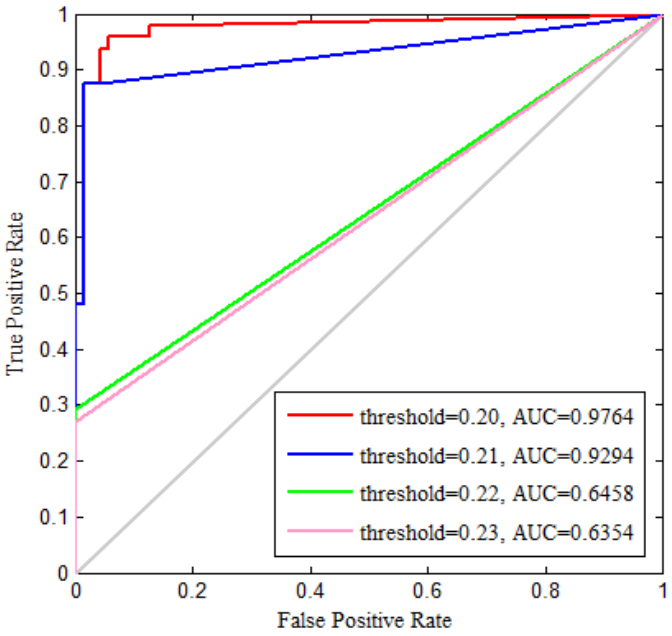

Fig. 4. ROC analysis of proposed algorithm scheme: four thresholds are set and it can be seen that all the four ROC curves are above the gray line, which means that AUCs are all more than 0.5.

6. Conclusion

The proposed generalized RQ assessment scheme is indeed a no-reference method, but it can embed any FR kernels to quantify the quality of medical images using the RQ index. Experimental results show that the RQ assessment scheme is a very powerful for evaluating the quality of reconstructed images. In practice, it can be helpful to determine which reconstruction strategy is the best.

Acknowledgement

Authors would like to acknowledge supports from National Natural Science Foundation of China (81271530/H1801), and Zhejiang Provincial Natural Science Foundation (LY14F020028).

References

[1] D. Dragan and D. Ivetic, Quality evaluation of medical image compression: What to measure, 2010 8th International Symposium on Intelligent Systems and Informatics (SISY), 2010, 37–42.
[2] M. Razaak and M.G. Martini, Medical image and video quality assessment in e-health applications and services, 2013 IEEE 15th International Conference on e-Health Networking, Applications & Services (Healthcom), 2013, 6–10.
[3] A.A. Nakhaie and S.B. Shokouhi, No reference medical image quality measurement based on spread spectrum and discrete wavelet transform using ROT processing, 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE), 2011, 121–125.
[4] W. Zhou and A.C. Bovik, Mean squared error: Love it or leave it? A new look at signal fidelity measures, IEEE Signal Process. Mag. **1** (2009), 98–117.

[5] V. Khieovongphachanh, K. Hamamoto and S. Kondo, Study on objective quality measurement for medical ultrasonic echo image compression, 2009 9th International Symposium on Communications and Information Technology, 2009, 1131–1135.
[6] J. Liu, L. Ma, J. He et al., A comparative study of assessment methods for medical image quality, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), 2012, 131–134.
[7] J. McElvain, S.P. Campbell, J. Miller and E.W. Jin, Texture-based measurement of spatial frequency response using the dead leaves target: Extensions, and application to real camera systems, in: Digital Photography Vi2010, F. Imai, N. Sampat and F. Xiao, eds., Washington D.C., USA, 2010, pp. 1–11.
[8] D. Brunet, E.R. Vrscay and W. Zhou, On the mathematic properties of the structural similarity index, IEEE Trans. Image Process **4** (2012), 1488–1499.
[9] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters **8** (2006), 861–874.