Image Quality Assessment on CT Reconstruction Images: Task-specific vs. General Quality Assessment

Jianmei Cai, Xiaogang Chen, Wuhao Huang and Xuanqin Mou

***Abstract*—For a given projection data scanned from an object, the task of optimizing CT reconstruction, tuned with different methodologies and parameters, is to produce the image with the best possible quality in terms of that the reconstructed image seems to be well balanced between apparent noises and fine image structures, or has the most visibility of pathological signals for diagnostic purpose. The former concerns the issue of general quality assessment over the whole image, as can be evaluated by a general image quality assessment (IQA) metric, e.g., SSIM, while the latter concerns the visual perception of a specific pathological signal from its background, which is task-specific and usually assessed by an observer model, e.g., Hotelling model. In the context of medical X-ray CT reconstruction, the ultimate goal is to produce the image with enough pathological information for task-specific evaluations. While considering optimizing the used reconstruction algorithm by tuning parameters, the task-specific assessment will fail in use because the task, or possible pathological signal, is unknown before the image is well reconstructed. Alternatively, in this phase only a general IQA metric can be used for optimization. In the case of the two kinds of IQA tasks are tangled in the optimization for medical X-ray CT reconstruction, it's a very interesting problem that if the two kinds of quality assessment tasks perform consistently or inconsistently? In this paper, we develop some reference images composed by simulated lesions and computed tomography image backgrounds and four test image databases derived from reference images by adding noise. we experimentally investigate the difference of IQA among four general IQA metrics when they are used to evaluate the image quality in the database. Experimental results show that for most images, the involved IQA models give inconsistent evaluations. This discloses that there are a lot of works to do before IQA models can be used in algorithm optimization for medical X-ray CT reconstruction tasks.**

Keywords—general image quality assessment; task-specific image quality assessment; image reconstruction; subjective experiment; computed tomography

I. INTRODUCTION

Image quality assessment(IQA) have been a hot topic in various medical imaging modalities, including X-ray CT imaging. The ultimate goal of medical imaging is to make the image providing enough pathological information and better visibility for clinical diagnosis. Considering the damage of X-Ray medical imaging should obey the principle 'As Low As Reasonably Achievable' (ALARA) under any condition. As defined in [1], the ALARA principle dictates that all radiologists take steps to limit the radiation dose received by both staff and patients during any radiological procedure, whilst at the same time maintaining the diagnostic quality of the images obtained. Namely, adequate clinical information should be provided with the minimum radiation dose to the patient. The higher radiation dose given, the better image quality obtained. For balancing between image quality and radiation dose, image quality assessment have been particularly important for medical imaging. There are multiple approaches to evaluate image quality through the whole medical care. Fig. 1 illustrates four layers of medical care: physical layer, algorithm layer, diagnosis layer and retrieval layer, and image quality is evaluated differently in each layer. Different parameters set in each layer can lead different affects to image quality. As showed in Fig. 1, physical layer concerns the imaging system related to equipment and scan protocols, such as KVP and mAs, which can introduce noise into projection data. For a given projection data scanned from an object, medical images can be reconstructed by varied reconstruction algorithms which are related to corresponding parameters selection resulting in different degrees of smooth in the context of algorithm layer. The over-smoothed process will damage the image structures which can be evaluated by general quality over the whole image, namely general IQA. Diagnosis layer concerns the extent of pathological information can be explored by clinicians which is assessed by task-specific IQA. At last, retrieval layer involves the different phases of postoperative recovery for a patient as retrospective evaluation of medical images. As discussed above, the four layers should all obey the 'ALARA' principle. Different parameters set in different layer can cause different impact on image quality. Up to now, many literatures have discussed IQA over different layers for medical X-ray CT images.

In our context, we pay more attention on algorithm and diagnosis layer as for the engineering application of X-ray tomographic imaging. For a given projection data from an object, we just consider how to get a fine reconstructed image in terms of fine image structures and most visibility of pathological signals. The former concerns the issue of general quality assessment over the whole image in algorithm layer, which can be evaluated by general IQA, while the latter concerns discriminating a specific pathological signal from its background in diagnosis layer, referred as task-specific IQA. The ultimate goal of CT reconstruction is to provide enough pathological information for medical diagnosis. While optimizing reconstruction algorithm by tuning parameters, task-specific IQA cannot be directly used because of the unknown

This work was supported by the National Key Research and Development Program of China (No. 2016YFA0202003) and the National Natural Science Foundation of China (NSFC) (No. 61571359).

JM Cai and XQ Mou are with the Institute of Image processing and Pattern recognition, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China; XG Chen are with State Grid Zhejiang Fuyang Power Supply Company; WH Huang are with Hangzhou Power Supply Company of State Grid Zhejiang Electric Power Company. Corresponding author: XQ Mou (e-mail:xqmou@xjtu.edu.cn).

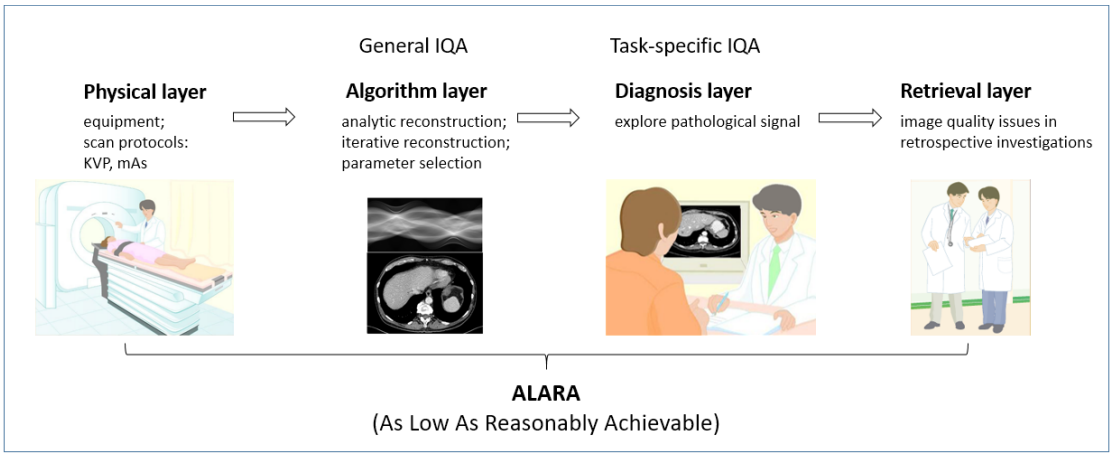


Fig. 1 X-ray CT images through various layers in the medical care

task before the reconstruction is completed. Alternatively, general IQA have been used to optimizing reconstruction algorithm, such as parameter selection. At the same time, task-specific IQA also have been used to predict human performance on medical diagnosis. Considering the different usage of two kinds of IQA for CT reconstruction, the problem comes out that if the two kinds of IQA perform consistently or inconsistently? This problem provides a new research topic.

In this paper, we investigated the difference between general IQA and task-specific IQA by organizing subjective experiments to evaluate CT image quality reconstructed by filter back projection (FBP) algorithm. Additionally, four general IQA models were used to evaluate the objective image quality. We analyzed the correlation of subjective task-specific scores and general scores as well as the consistency between subjective scores and objective scores. From the experimental results, subjective general scores don't have correlation with task-specific scores, similarly, objective general scores.

The rest of paper is organized as follows. Section II briefly introduces the formulation of general and task-specific IQA. Section III briefly describes experimental data and relevant set-up. Section IV illustrates the experimental results. Finally, section V gives some discussions and concludes this paper.

II. BACKGROUND

As mentioned in Section I, four layers will have different influence on image quality through the whole medical care. As for engineering purpose, image quality assessment focuses more on algorithm and diagnosis layer, which are related with general and task-specific IQA separately. In this section, we will introduce the formulation of general and task-specific IQA.

A. General IQA

In the algorithm layer, various algorithms and selected parameters will introduce kinds of distortions into an image reconstructed from a given CT projection, which will result in noise and over-smooth structures. In this circumstances, a strategy to ensure the optimal output image quality will be used to balance the distortion between apparent noises and fine image structures by tuning with different methodologies and parameters.

General IQA measures the perceptual difference between distorted images and reference images. Reference images are regarded as "perfect", while distorted images are distorted by acquisition, compression, and transmission etc. In this situation, general IQA can be used to select the optimal reconstructed CT image for balancing the CT structure distortion. Nowadays, general IQA models have been applied to optimize reconstructed algorithm such as selecting selecting regularization parameter tuned by blind image quality assessment (BIQA) on iterative CT reconstruction[5].

As showed in Fig. 2, we demonstrate the mechanism of general IQA. The distorted image is reconstructed by low doses of projection data, while the reference image is reconstructed by normal doses. In the situation of CT reconstruction, the general IQA can be divided into two types: full reference IQA and blind IQA according to the absence of reference image. Full reference IQA can get a quality map by point based error measure of distorted image and reference image following pooling over the whole image into a quality score, such as SSIM[2] and GMSD[3]. While blind IQA will extract local feature and statistics of distorted images following learned prediction model to get a quality score, such as BIQA[4]. The quality score can represent the image quality.

B. Task-specific IQA

In the diagnosis layer, the clinical purpose of medical CT reconstruction is to provide enough pathological information for medical diagnosis. In other words, the existed signals should be recognized as easy as possible from its surroundings.

Task-specific IQA concerns the visual perception of a specific pathological signal from its background, which is usually assessed by observer model. The specific signal (signal known exactly, SKE) should be discriminated from an exactly/statistically-known background (background known exactly/statistically, BKE/BKS) for an image. Up to now, a number of observer models have been proposed and applied to predict human performance on medical diagnosis. Wunderlich and Noo used several observer models to evaluate the influence of tube current modulation on lesion detectability in computed tomography images[6]. Miho et al. utilized channelized Hotelling observer to predict human performance in

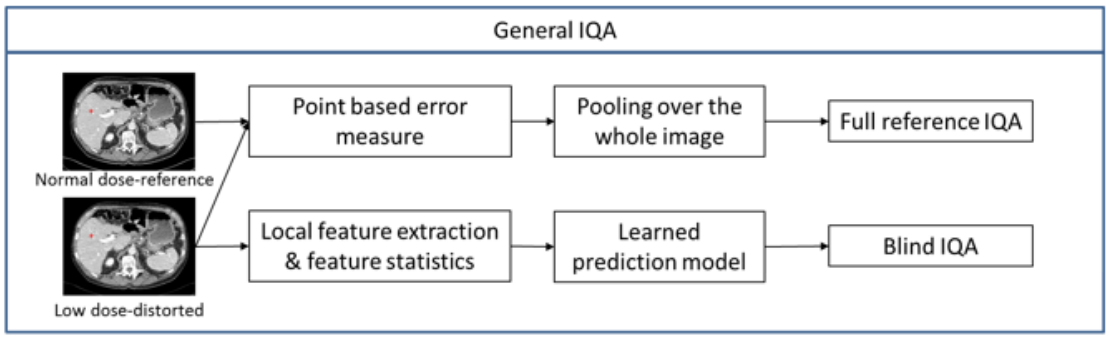


Fig. 2 General IQA model: FR IQA and BIQA

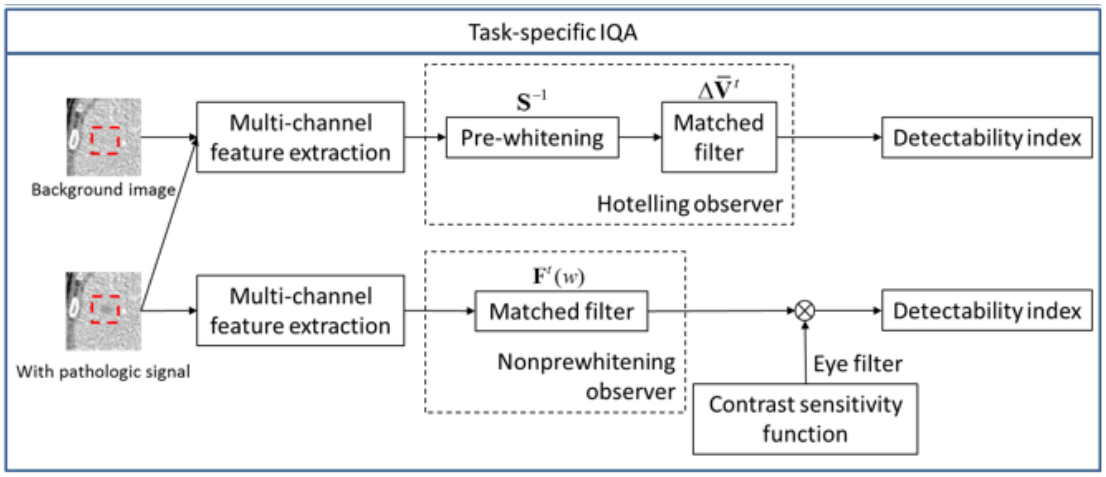


Fig. 3 Two Task-specific IQA Models: Hotelling Observer and NPWE

discriminating Alzheimer's dementia and controls by brain perfusion SPECT[7].

Hotelling observer and nonprewhitening observer are the two major types of observer model. As revealed in Fig. 2, negative samples referred as background image and positive samples with pathological signal are obtained from reconstructed images following multi-channel feature extraction. As for Hotelling observer, pre-whitening and matched filter is processed by training negative and positive samples. In the context of nonprewhitening observer, matched filter can be obtained directly from samples. By multiplying the contrast sensitivity function based on human visual system, we can get detectability index. The detectability index can represent the quality of image sets.

III. THEORY

In this section, we briefly describe four well-known full reference IQA models, SSIM, FSIM, GMSD and NLOG-MSE, that we use in this paper. We set the parameters of IQA models to the default values mentioned in [2], [8], [3], [9]. For more details, the reader is suggested to consult references [2], [8], [3], [9] for more details.

A. Structural Similarity Index (SSIM)

The Structural Similarity Index(SSIM) proposed by Wang et al. is a measure of structural similarity between the reference image and distorted image composed by three components, luminance, contrast, and structure. The overall SSIM Index can be described in formula (1).

$$SSIM(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^{\,2} + \mu_y^{\,2} + C_1)(\sigma_x^{\,2} + \sigma_y^{\,2} + C_2)} \quad (1)$$

$\mathbf{x}$ and $\mathbf{y}$ is the reference image and distorted image respectively. μ describes the mean intensity, and σ is defined as the standard deviation of image while C_1 and C_2 prevent the denominator from getting very closely to zero.

B. Feature-Similarity(FSIM)

The feature-similarity Index(FSIM) proposed by Zhang et al. employs the phase congruency(PC) and the image gradient magnitude(GM) as its features. PC is a dimensionless measure of the significance of a local structure as the primary feature in FSIM. GM provides contrast information as the secondary feature in FSIM as a supplement. After obtaining the local quality map, PS is used again as a weighting function to derive a signal similarity score. Thus, FSIM index can be calculated as the following two stages.

In the first stage, the local similarity map of PC and GM is computed separately as showed in formula (2).

$$\begin{cases} S_{PC}(\mathbf{x},\mathbf{y}) = \dfrac{2PC_{\mathbf{x}} \cdot PC_{\mathbf{y}} + T_1}{PC_{\mathbf{x}}^{\,2} + PC_{\mathbf{y}}^{\,2} + T_1} \\ S_G(\mathbf{x},\mathbf{y}) = \dfrac{2G_{\mathbf{x}} \cdot G_{\mathbf{y}} + T_2}{G_{\mathbf{x}}^{\,2} + G_{\mathbf{y}}^{\,2} + T_2} \end{cases} \quad (2)$$

$PC_{\mathbf{x}}$ and $PC_{\mathbf{y}}$ represents the PC map of the reference image and distorted image respectively. Similarly, $G_{\mathbf{x}}$ and $G_{\mathbf{y}}$ is the gradient magnitude of the reference image and distorted image. T_1 and T_2 are positive constants to increase the stability of S_{PC} and S_G.

Then S_{PC} and S_G are combined to get the similarity S_L of the reference image and distorted image defined as

$$S_L(\mathbf{x},\mathbf{y}) = [S_{PC}(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [S_G(\mathbf{x},\mathbf{y})]^{\beta} \tag{3}$$

In the second stage, the similarity is pooled into a signal similarity score as described in formula (4).

$$FSIM(\mathbf{x},\mathbf{y}) = \frac{\sum_{\Omega} S_L(\mathbf{x},\mathbf{y}) \cdot PC_m}{\sum_{\Omega} PC_m} \tag{4}$$

Where $PC_m = \max(PC_{\mathbf{x}}, PC_{\mathbf{y}})$ weights the importance of $S_L(\mathbf{x},\mathbf{y})$ in the overall similarity between $\mathbf{x}$ and $\mathbf{y}$.

C. Gradient Magnitude Similarity Deviation(GMSD)

The Gradient Magnitude Similarity Deviation(GMSD) proposed by Xue et al. develops a novel pooling strategy after obtaining the pixel-wise gradient magnitude similarity(GMS) between the reference image and distorted image that captures local perceptual quality. This index reflects the range of distortion severities in an image.

The GMS map is computed as follows:

$$GMS(\mathbf{x},\mathbf{y}) = \frac{2m_{\mathbf{x}} \cdot m_{\mathbf{y}} + c}{m_{\mathbf{x}}^2 + m_{\mathbf{y}}^2 + c} \tag{5}$$

where $m_{\mathbf{x}}$ and $m_{\mathbf{y}}$ are the gradient magnitude images.

Then this model computes the standard deviation of the GMS map in formula (6)

$$GMSD(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GMS(\mathbf{x},\mathbf{y}) - GMSM)^2} \tag{6}$$

where GMSM is the average pooling to the GMS map as Gradient Magnitude Similarity Mean.

D. Normalized LOG- mean square error(NLOG-MSE)

This Normalized LOG-mean square error(NLOG-MSE) proposed by Xue et al. is comprised of two basic steps: the linear Laplacian of Gaussian (LOG) filtering and the nonlinear normalization. In the first step, the input image is decorrelated by the linear filter LOG and transformed into a local frequency domain. Then the nonlinear normalization further reduces the redundancy between the transform domain coefficients which can be computed as:

$$r = \frac{W}{\sqrt{W^2 \otimes g + c_1}} \tag{7}$$

Where g is a Gaussian kernel and W is the transform coefficients.

Based on the above steps of decorrelation, the full reference IQA model is calculated in formula (8).

$$NLOG - MSE(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(r_{\mathbf{x}} - r_{\mathbf{y}})^2 \tag{8}$$

Where n is the total number of pixels in the given image.

IV. METHODS

Traditionally, image quality is evaluated by human beings, referred as subjective image quality assessment, which is regarded as the ground truth of objective IQA. Thus in this paper, we will compare general IQA with task-specific IQA by organizing subjective experiments and testing on several objective IQA models which have been exhibited in section III. And more details for experiments are demonstrated below.

A. Materials

In our experiments, the experimental images are divided into reference images and test images. The reference images are composed by simulated lesions and computed tomography (CT) image backgrounds. The image size is 512 by 512 pixels. The CT image backgrounds are reconstructed by filter back projection (FBP) algorithm at regular dose are obtained from two negative cases, L109 and L333, of the training database supported by Low Dose Grand Challenge from Mayo Clinic. The profile of simulated lesions is created by formula (9).

$$s(r) = C_t \cdot [1 - (\frac{r}{D})^2]^{\tau} \tag{9}$$

The C_t is the contrast which is set for six levels. r is the radial distant and D is the radius of the lesion. τ is set as 1.5 which describes the shape of the lesion. The reference images are composited by backgrounds and simulated lesions by formula (10). $b(r)$ represents the CT image background.

$$g(r) = b(r) \cdot [1 - s(r)] \tag{10}$$

We got three slices from L109 and L333 separately as background images and set C_t for six levels. Then we can get 18 reference images finally.

The test images are derived from the reference images by adding photons and then iterative reconstruction with various regularization parameters using GPU. The photons generated by Poisson distribute are varied from 2×10^4 to 8×10^4. And the regularization parameters are ranged from 0 to 0.025. We created 1008 test images with and without simulated lesions and randomly divided images into four non-overlapping groups.

B. Subjective experiments

As stated before, we created totally 1008 test images and 18 reference images. Test images were randomly divided into four non-overlapping groups evaluated by human observers. Detection time and viewing distance were without restrictions. In each test, the test image was displayed randomly on the right screen as well as its corresponding reference image was displayed on the left screen, such as showed in Fig.4 (a) and (b). Considering to explore the difference between task-specific IQA and general IQA, two assignments were designed in the

subjective experiment. First, subjects were required to recognize the simulated lesion whose information was absolutely demonstrated in the reference image. If the test image was not considered as the existence of lesion, it was scored as "0", otherwise, subjects needed to evaluate the detectability degree in five gradations with 1-5. "1" referred to the worst, on the contrary, "5" was the most detectable. Second, for general IQA, subjects needed to take more perceptual detail information of the whole test image into account. Similarly, five gradations were used with 1-5. "1" referred the worst, while "5" was the best perceptual quality. Finally, we obtained two kinds of subjective score for every subject. Then we averaged all subjective evaluations for each test images, called mean opinion score (MOS). Therefore, there were two types of MOS, which we named task-MOS and general-MOS separately. Because of the limited scale 1-5, the subjective scores had small variance. The maximal general-MOS is ranged from 3.4286 to 4.4285 in overall four databases, while for the maximal task-MOS is 3.8571 to 4.4286.

Since the four groups were conducted independently, regarded as four databases, D1, D2, D3, and D4, we measured the consistency between the results of IQA models with subjective scores on each database by three sores: the Spearman rank order correlation coefficient (SROCC), the Pearson correlation coefficient (PCC) and the root mean squared error (RMSE). The correlation coefficient which was as close to 1 as possible demonstrated a better consistency. Meanwhile, RMSE approaching 0 demonstrated higher prediction accuracy.

V. RESULTS

For illustrating the difference between task-MOS and general-MOS, we firstly select two test images in Fig.4 as an example with different task-MOS and general-MOS. Fig.4 (a) is the reference image marked with a red rectangle of the simulated lesion. Fig.4 (b) and (c) are test images with different regularization parameters and photons. Fig.4 (b) contains more noise by adding photons, and Fig.4 (c) is much smoothed with greater iterative regularization parameters. The task-MOS of Fig.4 (b) and (c) is 3.4286 and 4 respectively, that means Fig.4 (b) has the better subjective lesion detectability in the opinion of human observers. While general-MOS is 3.2857 and 2.7143 separately, which implies Fig.4 (c) has better global image quality than Fig.4 (b). At the same time, we calculate the SSIM index for the two test images, 0.7228 and 0.6828 for each, which is consistency with the general-MOS.

Table I and Table II show the performance of general IQA models compared with general-MOS and task-MOS respectively. The top general IQA models for three indexes are highlighted in bold font. From Table I, clearly SSIM performs well in four databases that SROCC are all above 0.6, while the others have a poor performance in D1 or D2. Nevertheless, the results listed in Table I imply that general IQA models can predict human performance on overall image quality in some extent. As for Table II, the results show the performance of general IQA models with task-MOS. Apparently, SROCC which is greatly lower than 1 shows that general IQA models cannot predict human performance on task-specific image quality absolutely. In some extent, the results listed above demonstrate that general IQA models can be applied to evaluate the general perceptual difference on medical images instead of dealing with task-specific assignment. Additionally, the SROCC between task-MOS and general-MOS of four databases is 0.1646, 0.1556, 0.2434 and 0.1764, which also reveals inconsistency between task-specific IQA and general IQA.

VI. DISCUSSION AND CONCLUSION

In this paper, we have discussed two types of image quality assessments, referred to task-specific IQA and general IQA. By organizing subjective experiments on the evaluation of lesion detectability and general perceptual image quality, we obtained task-MOS and general-MOS separately. Then we analyzed the correlation coefficient between objective IQA models and subjective scores, showed in Table I and Table II, which can infer the inconsistency between task-specific IQA and general IQA at evaluating image quality. The results show that the image quality evaluation in algorithm layer are different from that in diagnosis layer. The optimized output image don't have the best diagnostic perceptual for clinicians. By investigating the relationship of general IQA and task-specific IQA, the optimization strategies of CT reconstruction need to take the tasks of two layers into account.

ACKNOWLEDGMENT

The real data of the experiment was based on the train database of Low Dose CT Grand Challenge organized by Mayo Clinics, so the authors would like to acknowledge Dr. Cynthia McCollough, the Mayo Clinic, the American Association of Physicists in Medicine, and grants EB017095 and EB017185 from the national Institute of Biomedical Imaging Bioengineering.

REFERENCES

[1] A. L. Baert, Encyclopedia of Diagnostic Imaging. Springer Berlin Heidelberg, 2008:60-60.

[2] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.

[3] Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2013). Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 23(2), 684-95.

[4] Xue, W., Mou, X., Zhang, L., Bovik, A. C., & Feng, X. (2014). Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. Image Processing IEEE Transactions on, 23(11), 4850-62.

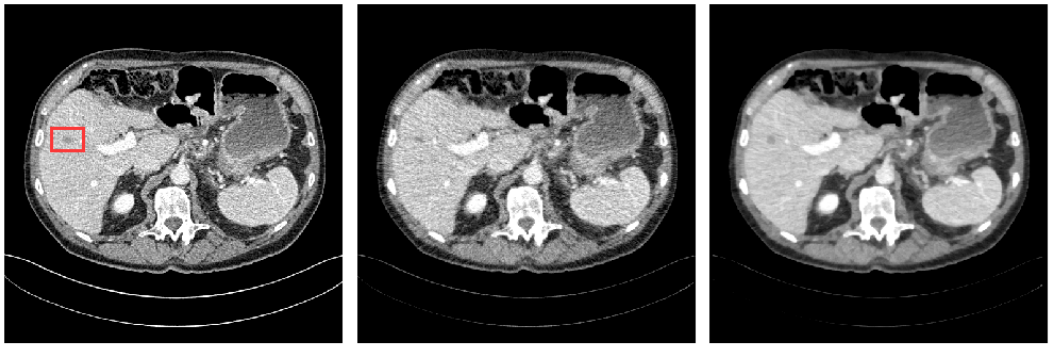
[5] Bai, T., Mou, X., Xue, W., Yan, H., & Jiang, S., B. Iterative CT Reconstruction with Regularization Parameter Tuned by Blind Image Quality Assessment. In 13th Fully3D meeting, 2015, pp. 124-127.

[6] Wunderlich, A., & Noo, F. (2008). Evaluation of the impact of tube current modulation on lesion detectability using model observers. Engineering in Medicine and Biology Society, 2008. Embs 2008. International Conference of the IEEE (Vol.2008, pp.2705-2708). IEEE.

[7] Shidahara, M., Inoue, K., Maruyama, M., Watabe, H., Taki, Y., & Goto, R., et al. (2006). Predicting human performance by channelized hotelling observer in discriminating between alzheimer's dementia and controls using statistically processed brain perfusion spect. Annals of Nuclear Medicine, 20(9), 605-13.

[8] Zhang, L., Zhang, D., Mou, X., & Zhang, D. (2011). Fsim: a feature similarity index for image quality assessment. *Image Processing IEEE Transactions on, 20*(8), 2378-2386.

[9] Xue, W., & Mou, X. (2014). Image quality assessment with mean squared error in a log based perceptual response domain. *IEEE China Summit & International Conference on Signal and Information Processing* (pp.315-319). IEEE.



(a) reference image

(b) test image1, task-mos=3.4286, general-mos=3.2857, ssim=0.7228

(c) test image2, task-mos=4, general-mos=2.7143, ssim=0.6828

Fig. 4. Comparison of task-MOS and general-MOS. (a) Reference image. (b) Test image1. (c) Test image2.

Table I Performance of general IQA models with general-MOS

	D1			D2			D3			D4		
	srocc	pcc	rmse	srocc	pcc	rmse	srocc	pcc	rmse	srocc	pcc	rmse
SSIM	0.6049	0.6205	0.2898	**0.6523**	**0.7080**	**0.3766**	0.7251	0.7275	0.3659	**0.8329**	**0.8608**	**0.3058**
FSIM	0.5427	0.5455	0.3097	0.4355	0.4786	0.4682	0.8408	0.8351	0.2933	0.7461	0.7765	0.3786
GMSD	0.5217	0.5262	0.3143	0.3092	0.3470	0.5001	**0.8510**	**0.8554**	**0.2762**	0.6476	0.6482	0.4575
NLOG-MSE	**0.6464**	**0.6523**	**0.2801**	0.5682	0.6095	0.4228	0.7857	0.8002	0.3198	0.7467	0.7685	0.3845

Table II Performance of general IQA models with task-MOS

	D1			D2			D3			D4		
	srocc	pcc	rmse	srocc	pcc	rmse	srocc	pcc	rmse	srocc	pcc	rmse
SSIM	**0.0988**	**0.0980**	**1.0805**	0.1170	0.0836	1.2068	0.0077	0.0010	0.9906	0.0412	0.0615	1.1084
FSIM	0.0729	0.0746	1.0827	**0.1221**	**0.1219**	**1.2021**	0.0342	0.0157	0.9905	0.0416	0.0604	1.1085
GMSD	0.0029	0.0041	1.0857	0.0940	0.0786	1.2073	0.0013	0.0192	0.9904	0.0002	0.0011	1.1105
NLOG-MSE	0.0295	0.0196	1.0855	0.0181	0.0142	1.2110	**0.0892**	**0.1193**	**0.9835**	**0.0647**	**0.0632**	**1.1083**