Modelling of Subjective Radiological Assessments with Objective Image Quality Measures of Brain and Body CT Images

Ilona A. Kowalik-Urbaniak[1(✉)], Jane Castelli[3], Nasim Hemmati[4], David Koff[3], Nadine Smolarski-Koff[3], Edward R. Vrscay[1], Jiheng Wang[2], and Zhou Wang[2]

[1] Department of Applied Mathematics, Faculty of Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
{iakowali@uwaterloo.ca,ervrscay,jiheng.wang}@uwaterloo.ca
[2] Department of Electrical and Computer Engineering, Faculty of Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
zhouwang@ieee.org
[3] Department of Radiology, McMaster University, Hamilton, ON L8S 4L8, Canada
jane.castelli@miircam.ca, koff@hhsc.ca, nadine.koff@realtimemedical.com
[4] Department of Diagnostic Imaging, Hamilton Health Sciences, McMaster University, Hamilton, ON L8S 4L8, Canada
nasim.hemmati@medportal.ca

Abstract. In this work we determine how well the common objective image quality measures (Mean Squared Error (MSE), local MSE, Signal-to-Noise Ratio (SNR), Structural Similarity Index (SSIM), Visual Signal-to-Noise Ratio (VSNR) and Visual Information Fidelity (VIF)) predict subjective radiologists' assessments for brain and body computed tomography (CT) images.

A subjective experiment was designed where radiologists were asked to rate the quality of compressed medical images in a setting similar to clinical. We propose a modified Receiver Operating Characteristic (ROC) analysis method for comparison of the image quality measures where the "ground truth" is considered to be given by subjective scores. The best performance was achieved by the SSIM index and VIF for brain and body CT images. The worst results were observed for VSNR.

We have utilized a logistic curve model which can be used to predict the subjective assessments with an objective criteria. This is a practical tool that can be used to determine the quality of medical images.

1 Introduction

Speed limitations of existing networks along with the explosive growth of image modalities with extremely high volume outputs have combined to make the issue of irreversible medical image coding one of the key considerations in the design of future PACS systems. Existing lossy image compression techniques are well suited for images where the only concern is visual quality.

© Springer International Publishing Switzerland 2015
M. Kamel and A. Campilho (Eds.): ICIAR 2015, LNCS 9164, pp. 3–13, 2015.
DOI: 10.1007/978-3-319-20801-5_1

As expected, increasing the degree of compression of an image leads to decreasing fidelity. The extent of allowable irreversible compression is dependent on the imaging modality and the nature of the image pathology and anatomy. Image compression often results in the distortion of the images and therefore creates the risk of losing or altering relevant diagnostic information.

If not implemented properly, distortions resulting from lossy compression can impede the ability of radiologists to make confident diagnoses from compressed medical images. However, defining the amount of accepted distortion is a complex task. For this reason, reliable image quality assessment methods are needed in order to achieve "Diagnostically Acceptable Irreversible Compression (DAIC)", defined in [13], which refers to an irreversible compression that has no effect on diagnostic task.

2 Image Quality Assessment

Many objective image quality metrics have been proposed in the last decade. Due to the wide variety of image types and applications, image quality assessment is not (yet) fully automatic and subjective approaches are still predominant [12]. How do we measure diagnostic quality?

There is no standard method to measure the quality of compressed medical images, however, three approaches are usually considered [3]:

1. Subjective image quality rating using psychovisual tests or questionnaires with numerical ratings.
2. Diagnostic accuracy is measured by simulating a clinical environment with the use of statistical analysis (e.g. Receiver Operating Characteristic (ROC)).
3. Objective quality measures such as the Mean Squared Error (MSE) and Structural Similarity (SSIM).

2.1 Objective Image Quality Methods

The existing objective image quality measures are not necessarily reliable measures of diagnostic quality for medical images. Moreover, compression ratios, generally used as pre-compression quality predictors, indicate a poor correlation with image quality and are image dependent [4]. According to Marmolin [10]: "MSE is not very valid as a quality criterion for pictures reproduced for human viewing and the improved measures could be derived by weighting the error in accordance with assumed properties of the visual system." Although MSE is shown to poorly correlate with visual quality, it should not be taken for granted that any perceptual object quality measure must be better. According to the relevant literature, SSIM and other objective measures show better performance than MSE for natural image/video content for consumer electronics applications based on subjective tests [16,17]. It cannot be assumed that an objective quality metric that performs well for natural images will ensure a superior diagnostic quality for medical images. In spite of these pitfalls, MSE and other objective

methods have been used in medical image quality assessment. Moreover, no objective model has been yet "established" for medical images, especially when using radiologists as subjects.

There are many full reference image quality assessment algorithms proposed in the literature. A lengthy review of objective image quality was presented in [5]. Full reference methods are based on comparison between the original image and its distorted version. Among the most common ones are Mean Squared Error (MSE), Signal-to-Noise Ratio (SNR), Structural Similarity (SSIM) [15], Visual Signal-to-Noise Ratio (VSNR) [1] and Visual Information Fidelity (VIF) [14].

MSE is related to the L^2 distance between image functions. The MSE between the compressed image g and the original image f is given by

$$\mathrm{MSE}(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f(i,j)-g(i,j)\right)^2. \tag{1}$$

The SSIM index, introduced by Wang and Bovik [16], assumes that the HVS is highly sensitive to structural information/distortions (e.g. JPEG blockiness, "salt-and-pepper" noise, ringing effect, blurring) in an image and automatically adjusts to the non-structural (e.g. luminance or spatial shift, contrast change) ones [15]. Another assumption of the SSIM index is that images are highly structured and there exist strong neighbouring dependencies among the pixels, which the MSE totally ignores. The SSIM index measures the difference/similarity between two images by combining three components of the human visual system: luminance, $l(f,g)$, contrast, $c(f,g)$ and structure, $s(f,g)$.

The (local) SSIM is given by:

$$SSIM(f,g) = \left(\frac{2\mu_f\mu_g + C_1}{\mu_f^2+\mu_g^2+C_1}\right)\cdot\left(\frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2+\sigma_g^2+C_2}\right)\cdot\left(\frac{\sigma_{fg}+C_3}{\sigma_f\sigma_g+C_3}\right). \tag{2}$$

where μ is the mean, σ_f^2 is the variance, and σ_{fg} is the covariance. SSIM is computed over $m \times n$ pixel neighbourhoods. The (non-negative) parameters C_1, C_2 and C_3 are stability constants of relatively small magnitude, which are designed to avoid numerical "blowups", which could occur in the case of small denominators. In the special case $C_3 = C_2/2$, the following simplified, two-term version of the SSIM index is obtained:

$$SSIM(f,g) = \left(\frac{2\mu_f\mu_g + C_1}{\mu_f^2+\mu_g^2+C_1}\right)\left(\frac{2\sigma_{fg} + C_2}{\sigma_f^2+\sigma_g^2+C_2}\right). \tag{3}$$

For natural images there are recommended default values for these parameters [15]. On the other hand, the question of optimal values for these stability constants for medical images is still an open one. The smaller the values of these constants, the more sensitive the SSIM index is to small image textures such as noise. In our study the constants were adopted from our previous work [8,9], where we examined a range of stability constants in order to determine the

value(s) which are optimal for the assessment of diagnostic quality of medical images.

In this study, we also employ the local MSE. This score is computed in a similar manner as the SSIM index, i.e. for a local $m \times n$ pixel neighbourhood. The quality score is computed by averaging the local MSE scores. In this work we use 11×11 pixel neighbourhoods for the computation of SSIM and local MSE measures, which is the default parameter in the computation of the SSIM index [15].

SNR is a measure of quality that considers the MSE and the variance of the original signal. It is defined as follows,

$$SNR(f,g) = 10\log_{10}\frac{\sigma_f^2}{MSE(f,g).} \tag{4}$$

The result is measured in decibels. SNR is considered in the literature as a valid quality measure [3].

Another image fidelity measure that we consider in our work is the VSNR [2]. The VSNR is a low complexity method that considers near-threshold and suprathreshold properties of the HVS. There are two stages in the algorithm. In the first one, wavelet-based models for the computation of contrast thresholds for distortions detection are used in order to determine whether distortions are visible. Based on the outcome of the first step, if the distortions are below the threshold of detection, then no further computation is required and the distorted image is of perfect visual fidelity. In the case where distortions are "suprathreshold", a second step of the algorithm is applied. In the second step, two properties are computed: the perceived contrast of the distortions and the disruption of global precedence. Finally, VSNR is computed as follows,

$$VSNR(f,g) = 20\log_{10}\left(\frac{C(f)}{\alpha d_{pc} + (1-\alpha)(d_{gp}/\sqrt{2})}\right) \tag{5}$$

where $C(f)$ denotes the contrast of the original image f, $d_{pc} = C(E)$ is the perceived contrast of the distortions, $E = f - g$ is the distortion, d_{gp} is the global precedence and $\alpha \in [0,1]$ determines the relative contribution of d_{pc} and d_{gp}. A detailed explanation and equations required to compute the VSNR are presented in [2]. According to the author, the VSNR metric has relatively low computational complexity.

VIF is based on visual information fidelity that considers natural scene statistical information of images. A detailed description can be found in [14]. The idea is to quantify the statistical information that is shared between the original and distorted images using conditional mutual information.

$$\text{VIF(f,g)} = \text{Distorted Image Information / Reference Image Information.} \tag{6}$$

3 Design of the Subjective Experiment

A subjective experiment was designed in order to assess the global prediction of the image quality assessments being examined. The experiment employed

sixty CT slices - thirty neurological and thirty upper body images- obtained from Medical Informatics Research Centre at McMaster (MIIRC@M), Hamilton, Canada. These images were first windowed according to their default viewing parameters (window width and window centre) in order to reduce their bit-depth from 16 to 8 bits per pixel (bpp). The resulting 512×512 pixel, 8 bpp images were compressed at five different compression levels using both the JPEG and JPEG2000 compression algorithms. Preliminary visual observations were used to select the compression levels employed in the experiment. The range of compression ratios was intended to represent a wide variety of visual qualities, from no and barely noticeable to fairly noticeable distortion.

An image viewer was developed specifically for this study in order to provide an easy-to-use graphical interface for the radiologists. The viewer displayed a compressed image beside its uncompressed counterpart without zoom. The compressed images were presented randomly and independently to each subject. During the course of the experiment, each compressed image was presented to each radiologist with some repetitions, but without the radiologists' knowledge. The subjects were not made aware of the compression ratios or quality factors of the compressed images. Three buttons were placed at the bottom of the user interface: (1) not noticeable and acceptable, (2) noticeable and acceptable and (3) noticeable and not acceptable. A confirmation was requested before passing to the next stimulus. The experiment can be summarized as follows,

- The subjects were six radiologists including experienced radiologists as well as residents (McMaster University, Hamilton, ON, Canada).
- Types of pathologies: Based on previous findings by Koff *et al.* [6,7], the pathologies include subtle lesions in the liver for the body, and parenchyma and posterior fossa for brain. Subtle lesions include the following two types:
  - Very small lesions, limit in size, of less than 2 mm, but high contrast (calcifications) or low density (tiny cysts);
  - Subtle parenchymal alterations translating into subtle differences in density such as cerebral infarcts.
- The brain CT and body CT images used in the experiment were carefully chosen with the help of radiologists and contain pathological and normal cases (about 1/3 of images represented normal cases).
- Working environment: MIIRC@M office (Hamilton, ON, Canada); Eizo Radiforce monitor, 54 cm (21.3") display, with a 1200×1600 native resolution (3:4 aspect ratio) and a viewing size of 324.0×432.0 mm. Capable of displaying 10-bit colors.
- Compression levels: Five quality factors (JPEG input parameters) were chosen:
  - $[90, 65, 45, 20, 5]$ for brain CT images, $[90, 75, 55, 35, 10]$ for body CT images. First the images were compressed using JPEG algorithm, then using JPEG2000 with the corresponding compression ratios.
- Trial and main experiments:
  - Trial experiment included 6 images from each brain CT and body CT sets. These images were repetitions of images that were included in the main part of the experiment.

  - Main experiment: Number of images: 306 brain CT, 306 body images (30 different images compressed at five compression ratios using JPEG and JPEG2000 algorithms including 6 repetitions added at the end of the sequence).
- Repetitions were included in the trial experiment as well as at the end of the first part of the main experiment.
- Duration of the experiment: The number of images was adjusted to the time limitation of the experiment. Expected duration of the experiment: 60 min.
  - Trial experiment including explanation of the task: 10 min.
  - Main experiment, Part 1:
    * brain CT: 25 min, body CT: 25 min.

4 Data Analysis

4.1 Modified ROC Analysis

The Receiver Operating Characteristic (ROC) curve is a common tool for visually assessing the performance of a classifier in medical decision making [11]. ROC curves illustrate the trade-off of benefit (true positives, TP) versus cost (false positives, FP) as the discriminating threshold is varied. In our experiment, we employed a three-level subjective rating of image quality by radiologists: (1) not noticeable and acceptable, (2) noticeable and acceptable and (3) noticeable and not acceptable. Since we are concerned about diagnostic quality, we combine the images that fell into the two classes during subjective assessments: (1) not noticeable (distortions) and acceptable, (2) noticeable (distortions) and acceptable. By doing so, we now have a binary rating scale: acceptable and unacceptable images. At this point, we must clarify that due to the nature of the problem we are investigating, our definitions of FP and TP differ from those normally applied for the purposes of medical diagnosis. In this study, we wish to examine how well different "image quality indicators" compare to the subjective assessments of image quality by radiologists. As such, we must assume that the "ground truth" for a particular experiment, i.e., whether or not a compressed image is acceptable or unacceptable, is defined by the radiologist(s). From this ground truth, we measure the effectiveness of each image quality indicator in terms of FP, TP, etc. This leads to the following definitions of P, N, TP, FP, etc.:

P= FP+TP total positives (acceptables) and **N= TN+FN** total negatives (unacceptables): these refer to radiologists' subjective opinions, which represent the True Class. On the other hand, **P'** and **N'** belong to the Hypothesis Class which, in our experiment, corresponds to a given objective image quality assessment method.

TP (true positives): images that are acceptable to both radiologists and a given quality assessment method.
TN (true negatives): images that are unacceptable to both radiologists and a given quality assessment method.

FN (false negatives): images that are acceptable to radiologists but unacceptable to a given quality assessment method.
FP (false positives): images that are unacceptable to radiologists but acceptable to a given quality assessment method.

The Acceptability/unacceptability of a given quality measure is defined with respect to the discrimination threshold s' associated with the method, where $0 \leq s' \leq 1$. Using this threshold value s', FPR and TPR are computed. Each threshold value s' generates a point on the ROC curve which corresponds to the pair of values **(FPR, TPR) = (1-SP, SE)**, where SP denotes specificity and SE denotes sensitivity, i.e.,

FPR (false positive rate) $= FP/N = 1 - SP$ (specificity)
TPR (true positive rate) $= TP/P = SE$ (sensitivity)
FNR (false negative rage) $= FN/N = 1 - SE$
TNR (true negative rate) $= TN/N = SP$.

4.2 ROC Analysis Results

ROC curves were computed corresponding to SSIM, MSE, SNR, VIF, VSNR for body and brain CT images; they are shown in Fig. 1 and 2. Table 1 shows the Area Under the (ROC) Curve (AUC) scores corresponding to each of the objective measures studied. The largest AUC corresponds to the SSIM index. The second best method is the VIF measure. We observed that MSE, MSE local, SNR and VSNR have smaller AUC. Moreover, the AUC scores from ROC analysis are larger for brain CT images than for body CT images with respect to all objective image quality measures considered. This indicates that the studied objective measures can predict the subjective assessments of radiologists of brain CT images with better accuracy than those corresponding to body CT images.

According to the presented modified ROC analysis, SSIM index provides the closest match with radiologists' subjective assessments. The second best measure is the VIF. A worse performance is observed for MSE, local MSE, SNR and

Table 1. AUC scores resulting from ROC analysis corresponding to objective quality measures and subjective radiologists' assessments for brain and body CT images.

Objective quality measure	AUC (brain CT)	AUC (body CT)
SSIM	0.9924	0.9618
MSE (local)	0.9899	0.9326
MSE	0.9892	0.9366
SNR	0.9896	0.9351
VNSR	0.9662	0.9400
VIF	0.9916	0.9571

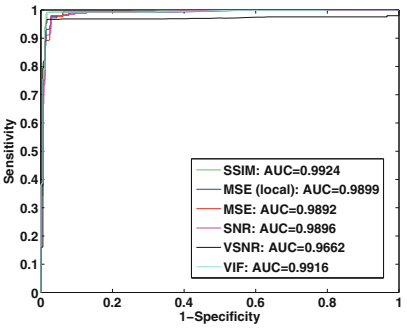

Fig. 1. ROC curves corresponding to SSIM, MSE, local MSE, SNR, VSNR and VIF for brain CT images(Color figure online).

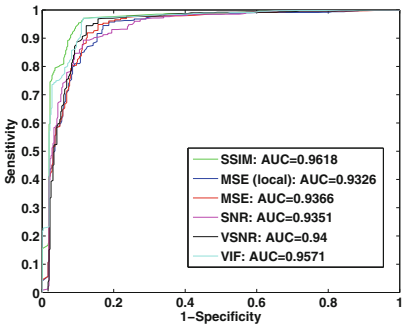

Fig. 2. ROC curves corresponding to SSIM, MSE, local MSE, SNR, VSNR and VIF for body CT images(Color figure online).

VSNR. According to our previous work, where we studied both local and global image quality [9], SSIM also showed a better correspondence with subjective assessments of image quality.

4.3 Logistic Curve Model

For medical images, the goal is to find an objective image quality measure that can best predict the subjective radiologists' assessments. The performance of an objective quality measure to predict subjective scores is usually measured by means of a curve-fitting model. First MSE values are plotted against mean scores of subjective assessments, then a curve (e.g. polynomial spline, quadratic, exponential, logistic) is fitted to the resulting points [3].

In this study we used a logistic curve model as a proposed predictor of subjective radiologists' assessments corresponding to the studied objective quality measures. In order to take into account the variability in the subjective quality assessment of compressed medical images, a logistic cumulative probability distribution is assumed to model the decision of a radiologist to either accept or not accept an image at a given objective score. A robust curve-fitting is performed on the plot of the average subjective score over all the radiologists as a function of the objective score.

Given x_1, the predicted value (SSIM, VIF) and y_1, the average subjective score, we determine the parameters a and b of the logistic function

$$y = 1/(1 + \exp(-a * x + b))$$

that produce the least weighted square error, with weighting according to the bisquare method.

A threshold can be selected so that the cumulative probability distribution model represents the desired level of confidence that the quality of the compressed

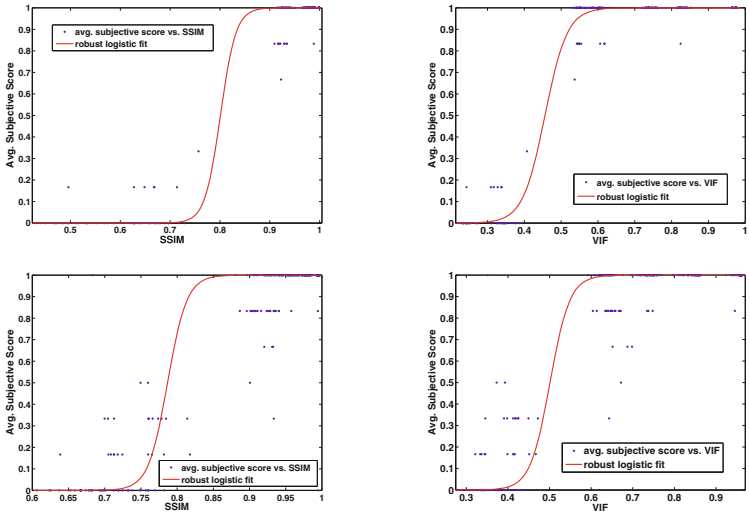

Fig. 3. Logistic curves corresponding to SSIM for (top left) brain, (bottom left) body CT images and VIF for (top right) brain, (bottom right) body CT images.

image is diagnostically acceptable. For example, if one requires a 99 % confidence, the recommended threshold has to be selected at the value for which the fitted logistic curve is at 0.99. Figure 3 shows logistic curves fitted with LAD Regression corresponding to SSIM and VIF for brain and body CT images.

5 Conclusions

The task of achieving Diagnostically Acceptable Irreversible Compression (DAIC) of medical images is a complex one. It involves tuning technology with radiological subjective responses/preferences. In this work, we compared the performances of some of the most popular image quality measures (MSE, SNR, SSIM, VSNR, VIF) based on experimental data collected in an experiment involving radiologists' subjective assessment of image quality. The experiment involved a global quality assessments of brain and body CT images at several compression ratios. Six radiologists evaluated compressed images as acceptable (with and without noticeable distortions) or unacceptable as compared to their uncompressed counterparts. An ROC analysis indicates that SSIM demonstrated the best performance, i.e., it provides the closest match to the radiologists' assessments. The worst performance was observed for the VSNR quality measure.

We have utilized a logistic curve model, which can be used to predict the subjective assessments with an objective criteria. This is a practical tool that can be used to determine the quality of medical images. The optimal quality score can be selected so that the cumulative probability distribution model represents the desired level of confidence that the quality of the compressed image is diagnostically acceptable.

Our current work involves developing advanced techniques of choosing a threshold for compression using the most popular quality measures.

Acknowledgements. We thank Prof. Paul Marriott, Department of Statistics and Actuarial Sciences, University of Waterloo for valuable advice with regard to the statistical design of our experiments. This research was supported in part by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (ERV and ZW).

References

1. Chandler, D.M., Hemami, S.S.: VSNR: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans. Image Process. **16**(9), 2284–2298 (2007)
2. Chandler, D.M., Lim, K.H., Hemami, S.S.: Effects of spatial correlations and global precedence on the visual fidelity of distorted images. In: Human Vision and Electronic Imaging XI, vol. 6057, February 2006
3. Cosman, P.C., Gray, R.M., Olshen, R.A.: Evaluating quality of compressed medical images: Snr, subjective rating, and diagnostic accuracy. In: 82 (ed.) Proceedings of the IEEE, vol. 6, pp. 919–932, June 1994
4. Fidler, A., Likar, B.: What is wrong with compression ratio in lossy image compression? Radiology **245**(1), 299 (2007)
5. A. George and S. J. Livingston. A survey on full reference image quality assessment algorithms. IJRET: Int. J. Research Eng. Technol. 2(12), December 2013
6. Koff, D., Bak, P., Brownrigg, P., Hosseinzadeh, D., Khademi, A., Kiss, A., Lepanto, L., Michalak, T., Shulman, H., Volkening, A.: Pan-canadian evaluation of irreversible compression ratios (lossy compression) for development of national guidelines. J. Digit. Imaging **22**(6), 569–578 (2009)
7. Koff, D., Shulman, H.: An overview of digital compression of medical images: Can we use lossy image compression in radiology? CARJ **57**(4), 211–217 (2006)
8. Kowalik-Urbaniak, I.A.: The quest for 'diagnostically lossless' medical image compression using objective image quality measures. Ph.D. thesis, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1 (2014)
9. Kowalik-Urbaniak, I.A., Brunet, D., Wang, J., Vrscay, E., Wang, Z., Koff, D., Koff, N., Wallace, B.: The quest for 'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images. In: Medical Imaging : Image Perception. Observer Performance, and Technology Assessment **9037**, 2014 (2014)
10. Marmolin, H.: Subjective MSE measures. IEEE Trans. Syst. Man and Cybern., SMC **16**(3), 486–489 (1986)
11. Metz, C.E.: Basic principles of ROC analysis. Semin. Nucl. Med. **8**, 282–298 (1978)
12. Nait-Ali, A., Cavaro-Menard, C.: Compression of Biomedical Images and Signals. Wiley, London (2008)
13. European Society of Radiology: (ESR). Usability of irreversible image compression in radiological imaging. Insights into. Imaging **2**(2), 103–115 (2011)
14. Sheikh, H.R., Bovik, A.C., de Veciana, S.G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Proces. **14**(12), 2117–2128 (2005)
15. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? - a new look at signal fidelity measures. IEEE Signal Proc. Mag. **26**(1), 98–117 (2009)

16. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Proces. **13**(4), 600–612 (2004)
17. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Proces. **20**(5), 1185–1198 (2011)