New image-quality measure based on wavelets

Emil Dumic
Sonja Grgic
Mislav Grgic
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Wireless Communications
Unska 3/XII, HR-10000 Zagreb, Croatia
E-mail: emil.dumic@fer.hr

Abstract. *We present an innovative approach to the objective quality evaluation that could be computed using the mean difference between the original and tested images in different wavelet subbands. Discrete wavelet transform (DWT) subband decomposition properties are similar to human visual system characteristics facilitating integration of DWT into image-quality evaluation. DWT decomposition is done with multiresolution analysis of a signal that allows us to decompose a signal into approximation and detail subbands. DWT coefficients were computed using reverse biorthogonal spline wavelet filter banks. Wavelet coefficients are used to compute new image-quality measure (IQM). IQM is defined as perceptual weighted difference between coefficients of original and degraded image.* © *2010 SPIE and IS&T.* [DOI: 10.1117/1.3293435]

1 Introduction

Discrete wavelet transform (DWT) can be used in various image processing applications, such as image compression and coding.[1] In this paper, we examine how DWT can be used in image-quality evaluation, which has become crucial for the most image-processing applications. Quality of an image can be evaluated using different measures. The best way to do this is by making a visual experiment, under controlled conditions, in which human observers grade which image provides better quality. Such experiments are time consuming and costly. A much easier approach is to use some objective measure that evaluates the numerical error between the original image and the tested one. In the real world, there is no perfect way for an objective assessment of image quality.[2] The problem with the most objective measures is that objective measures need a reference (original) image to be able to grade the corresponding tested image, while human observers can grade image quality independently of a corresponding original image. Over the past years, there have been many attempts to develop models or metrics for image quality that incorporate elements of human visual system (HVS) sensitivity.[3,4] These metrics for quality assessment have limited effectiveness in predicting the subjective quality of real images. However, there is no current standard and objective definition of image quality.

Paper 09058SSPRRR received Apr. 29, 2009; revised manuscript received Nov. 25, 2009; accepted for publication Dec. 1, 2009; published online Jan. 25, 2010.

1017-9909/2010/19(1)/011018/19/$25.00 © 2010 SPIE and IS&T.

In our research, Watson's wavelet model[5] is used to incorporate HVS characteristics in image-quality measure. This model is based on direct measurement of the HVS noise visibility threshold for the specific wavelet decomposition level using a linear-phase CDF9_7 biorthogonal filter. Blank images with uniform gray level were decomposed, and afterward noise was added to the wavelet coefficients. After inverse wavelet transform, the noise visibility threshold in the spatial domain was measured by subjective experimentation at a fixed viewing distance. An experiment was conducted for each subband, and the visual sensitivity for that subband was then defined as the reciprocal of the corresponding visibility threshold. This model can be directly applied for the perceptual image compression by quantizing the wavelet coefficients according to their visibility thresholds. It is also extendable to image-quality assessment, as was done in Ref. 6, where a wavelet visible difference predictor was used to predict visible differences between original and compressed (or noisy) images. In this paper, we present a new way of using Watson's wavelet model for image-quality evaluation.

In order to investigate the effectiveness of objective measurements when evaluating and monitoring the picture quality, the work was carried out in the following three steps:

1. Objective measurements, including our own developed measure, were performed on the same set of picture sequences taken from an already-known image database.[7]
2. Subjective assessment results were taken from Ref. 7. The database includes subjective grades with calculated differential mean opinion score (DMOS) results. The main goal of these studies was to obtain subjective results that would be used in the third step for verification and comparison of objective measures.
3. The results of the objective assessments (step 1) and subjective measurements (step 2) were studied.

In our approach, original and distorted images are decomposed by DWT into approximation and detail subbands.[8] Difference of DWT coefficients between original and distorted images is computed over each subband separately, and then global quality measure is calculated. Objective

measure achieved in this way shows better correlation with subjective grades in comparison to traditional objective measures, such as peak signal-to-noise ratio (PSNR) or mean squared error (MSE). Results are also compared to other quality measures that take into account image-quality perception by HVS.

Results depend on type of an image (more or less details in image) as well as image resolution. Different wavelet filters[9] as well as different wavelet scales can be used to achieve good correlation results for the same type of image.

The paper is organized as follows. In Section 2, subjective image quality measure (IQM) is briefly presented. Section 3 explains some of the existing IQM. Section 4 explains the basics of DWT. In Section 5, we explained in detail how our proposed IQM is calculated. Section 6 compares different objective IQMs with results of subjective assessment. Finally, Section 7 draws the conclusion.

2 Subjective Image Quality Measure

To be able to compare several later-described objective methods, we used subjective quality results from Ref. 7. Subjective quality evaluation was based on ITU-R recommendation BT.500-11.[10] Details of the subjective testing can be found in Ref. 11. Briefly, they are as follows: 29 high-resolution 24 bits/pixel RGB color images (typically, 768×512) were degraded using five degradation types:

1. JP2K, JPEG2000 compression
2. JPEG, JPEG compression
3. WN, white noise in the RGB components
4. Gblur, Gaussian blur
5. Fastfading, transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model

Each of these 29 images had versions with seven to nine different qualities for JPEG and JPEG2000 and six images with different qualities for white noise, Gaussian blur, and fastfading. About 20–29 observers had to grade image quality on a continuous scale with five grades (bad, poor, fair, good, and excellent). In this way, observers evaluated total of 982 images, out of which 203 were reference and 779 degraded images. The experiments were conducted in seven sessions: two sessions for JPEG2000, two for JPEG, and one each for white noise, Gaussian blur, and fastfading transmission errors.

Raw scores for each subject were converted in difference scores between the test and reference images,

$$d_{i,j} = r_{\mathrm{iref}(j)} - r_{i,j}, \tag{1}$$

where $r_{\mathrm{iref}(j)}$ denotes the raw quality score assigned by the i'th subject to the reference image corresponding to the j'th distorted image and $r_{i,j}$ score for the i'th subject and j'th image. Difference scores were converted to Z scores

$$z_{i,j} = \frac{d_{i,j} - \bar{d}_i}{\sigma_i}, \tag{2}$$

where $\bar{d}_i$ is the mean of the raw score differences overall images ranked by the subject i and σ_i is the standard deviation. Z scores are used to make scores more equal, because each observer uses different part of grading scale. Finally, a DMOS value for each distorted image was computed by shifting Z scores to the full range (1–100).

3 Objective Image Quality Measures

In this paper, we examined several commonly used objective quality measures, which were applied to a luminance channel only, because they give better correlation results with subjective testing (by comparison to calculating objective measures using RGB components separately and then calculating mean of them), as follows:

1. MSE
2. PSNR
3. structural similarity (SSIM)
4. multiscale SSIM (MSSIM)
5. visual information fidelity (VIF)
6. visual signal-to-noise ratio (VSNR)
7. IQM (our proposed measure)

MSE represents the power of noise or the difference between original and tested images.

$$\mathrm{MSE} = \frac{\Sigma_i \Sigma_j (a_{i,j} - b_{i,j})^2}{x \cdot y}, \tag{3}$$

where $a_{i,j}$ and $b_{i,j}$ are corresponding pixels from the original and tested images, and x and y describe height and width of an image.

PSNR is the ratio between the maximum possible power of a signal and the power of noise. PSNR is usually expressed in terms of the logarithmic decibel

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{\mathrm{MSE}}, \tag{4}$$

where 255 is maximum possible amplitude for an 8-bit image.

SSIM is a novel method for measuring the similarity between two images.[12] It is computed from three image measurement comparisons: luminance, contrast, and structure. Each of these measures is calculated over an 8×8 local square window, which moves pixel-by-pixel over the entire image. At each step, the local statistics and SSIM index are calculated within the local window. Because the resulting SSIM index map often exhibits undesirable "blocking" artifacts, each window is filtered with a Gaussian weighting function (11×11 pixels). In practice, one usually requires a single overall quality measure of the entire image; thus, the mean SSIM index is computed to evaluate the overall image quality. The SSIM can be viewed as a quality measure of one of the images being compared, while the other image is regarded as perfect quality. It can give results between 0 and 1, where 1 means excellent quality and 0 means poor quality. Similar to SSIM, the MSSIM method is a convenient way to incorporate image details at different resolutions.[13] This is a novel image synthesis-based approach that helps calibrating the parameters (such as viewing distance) that weight the relative importance between different scales.

VIF criterion[14] quantifies the Shannon information that is shared between the reference and distorted images relative to the information contained in the reference image itself. It uses natural scene statistics modeling in conjunc-

tion with an image-degradation model and an HVS model. Results of this measure can be between 0 and 1, where 1 means perfect quality and near 0 means poor quality.

VSNR[15] operates in two stages. First, the threshold for distortions of a degraded image is determined to decide if it is below or above human sensitivity of error detection. This is computed using wavelet-based models of visual masking. If distortions are below the threshold, then the distorted image is assumed to be perfect (VSNR=∞). If the distortions are above the threshold, then a second stage is applied. Calculations are made on the low-level visual property of perceived contrast and the midlevel visual property of global precedence. These properties are used to determine Euclidean distances in distortion-contrast space of multiscale wavelet decomposition. Finally, VSNR is calculated from a linear sum of these distances. A higher VSNR means that the tested image is less degraded.

4 DWT

DWT refers to wavelet transforms for which the wavelets are discretely sampled. This can be done with multiresolution analysis of a signal.[8] Multiresolution analysis allows us to decompose a signal into approximations and details. These coefficients can be computed using various filter banks, such as Daubechies, Coiflets, or biorthogonal filters.[16–18]

Suppose we have a one-dimensional input signal $x(t)$. It can be decomposed into approximation and detail coefficients of the first level. Then we can also decompose approximation coefficients at the first level further into approximation and detail coefficients at the second level. This can be expressed as

$$x(t) = \sum_k a_0(k)\phi_{j,k}(t) = \sum_k a_1(k)\phi_{j-1,k}(t) + \sum_k d_1(k)\omega_{j-1,k}(t), \tag{5}$$

where a_0 are approximation coefficients at scale index j, a_1 approximation coefficients, and d_1 detail coefficients at scale index $j-1$ (analysis). Bases $\varphi_{j,k}(t)$ and $\omega_{j,k}(t)$ are the wavelet basis. These bases are used to decompose input signal. Because wavelets and scales at each index level are orthogonal, it can be shown that coefficients a_1 and d_1 can be expressed as

$$a_1(k) = \sum_n h_0(n-2k)a_0(n)$$

$$d_1(k) = \sum_n h_1(n-2k)a_0(n). \tag{6}$$

Equations (6) look like convolution, but there is a downsampling involved (by a factor of 2). h_0 and h_1 are accordingly scaling and wavelet filters. The decomposition of a signal into an approximation and a detail can be reversed. Similar expressions like (6) can be used, but we have to use upsampling and conjugate mirror filters.

In image transform, we have two dimensions. Thus, we need to extend analysis of decomposition and reconstruction in two dimensions. We may do decomposition with separable wavelet transform, which is in fact one-dimensional convolution with subsampling by a factor of 2 along the rows and columns of image. Reconstruction is done reversely. This means upsampling by 2 and then convolution along the rows and columns. Decomposition and reconstruction at level j are shown in Fig. 1.

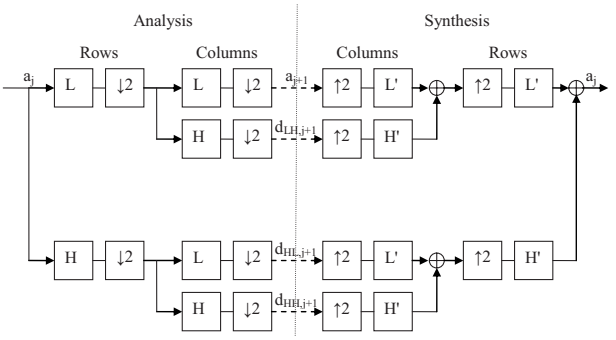

Fig. 1 Wavelet decomposition and reconstruction: *L*, low-pass analysis filter (from scaling function); *H*, high-pass analysis filter (from wavelet function); *L′* and *H′* are low- and high-pass reconstruction filters; *a* is approximation coefficient and *d* is detail coefficient; and ↓2 and ↑2 denote downsampling and upsampling by factor 2.

5 Proposed New Algorithm for Image Quality

Some of the existing objective measures described in Section 4 did not take into account HVS in the sense that the eye will see and grade image quality according to the type of error, as well as location of an error in subband space. Because of that, our method calculates image quality using wavelet decomposition and grades quality depending on the wavelet subband in which the error occurs. Experiments on image database[7] have shown that different types of image degradation produce different error distributions in wavelet subbands. For example, for JPEG and JPEG2000 compressed image errors will be placed in the higher wavelet subbands (HH subband, level 2 and higher) while images with Gblur and fastfading degradations will also have errors in lower subbands. White noise has equally distributed errors in all subbands.

In our research, we used two types of wavelet filters. The first filter, called CDF9_7[17] (nine coefficients in decomposition low-pass and seven in decomposition high-pass filters), is designed as a spline variant with less dissimilar lengths between low-pass and high-pass filters and has seen widespread use in image processing. The second filter, Coif22_14[18] (22 coefficients in decomposition low-pass and 14 in decomposition high-pass filters), has properties of antisymmetric biorthogonal Coiflet systems, whose filter banks have even lengths and linear phase. Coefficients of these wavelet filters are presented in Table 1. Figure 2 shows decomposition low-pass and high-pass wavelet filters.

All color images were first converted to gray-scale images by forming a weighted sum of the red (R), green (G), and blue (B) components:

$$Y = 0.2989\mathrm{R} + 0.5870\mathrm{G} + 0.1140\mathrm{B}. \tag{7}$$

In this way, we calculated errors only for luminance component (Y) in images. After converting the original and degraded images, the degraded image is subtracted from

Table 1 Coefficients of the used wavelet filters.

CDF9_7, lowpass	CDF9_7, highpass	Coif22_14, lowpass	Coif22_14, highpass
0.03782845550726	−0.06453888262870	−0.00006038691911	0.00249239584019
−0.02384946501956	0.04068941760916	−0.00007137535849	0.00294555229198
−0.11062440441844	0.41809227322162	0.00097545380465	−0.02160076866236
0.37740285561283	−0.78848561640558	0.00120718683898	−0.02777241079070
0.85269867900889	0.41809227322162	−0.00658124080240	0.09720345190957
0.37740285561283	0.04068941760916	−0.00932685158094	0.16200574375453
−0.11062440441844	−0.06453888262870	0.03683394176520	−0.64802297501813
−0.02384946501956		0.01809725255148	0.64802297501813
0.03782845550726		−0.14280042659266	−0.16200574375453
		0.07881441881590	−0.09720345190957
		0.73001880866394	0.02777241079070
		0.73001880866394	0.02160076866236
		0.07881441881590	−0.00294555229198
		−0.14280042659266	−0.00249239584019
		0.01809725255148	
		0.03683394176520	
		−0.00932685158094	
		−0.00658124080240	
		0.00120718683898	
		0.00097545380465	
		−0.00007137535849	
		−0.00006038691911	

the original image. The result is the difference image. It gives the same result as if we would subtract images in the wavelet domain, because wavelet transform is orthogonal at each level. After decomposing the difference image into three-level decomposition, the error distance in each wavelet subband can be computed using the following equation:

$$E = \left(\sum_i \sum_j |e_{i,j}|^k \right)^{1/k}. \quad (8)$$

In Eq. (8), $e_{i,j}$ are coefficients from the difference image, in the same subband. Factor k has experimentally been determined to give the best possible correlation results. When using Watson's wavelet model, it was 5. Weighting factors for level 3 decomposition are presented in Table 2, according to indexing of DWT bands (Fig. 3).

To improve the results achieved by Watson's model, we used the Coif22_14 wavelet filter, which gave a little better

Table 2 Weighting factors $w_{\lambda,\theta}$ for three-level CDF9_7 DWT, Watson's model.

	Level (λ)		
Orientation (θ)	1	2	3
1	—	—	0
2	0	14.68	12.71
3	0	28.41	19.54
4	0	14.69	12.71

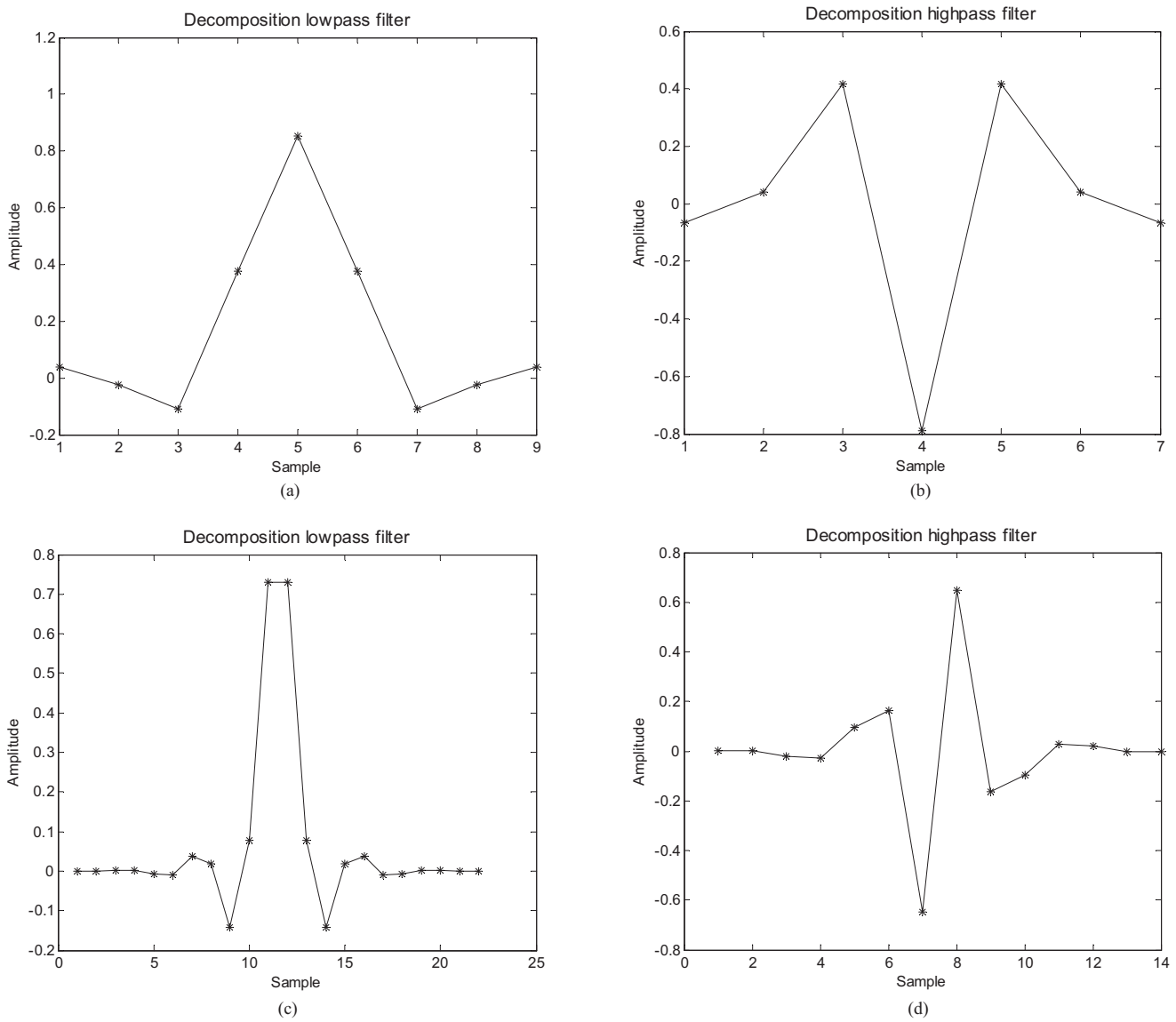

Fig. 2 Wavelet filters: (a) CDF9_7 decomposition low-pass filter, (b) CDF9_7 decomposition high-pass filter, (c) Coif22_14 decomposition low-pass filter, and (d) Coif22_14 decomposition high-pass filter.

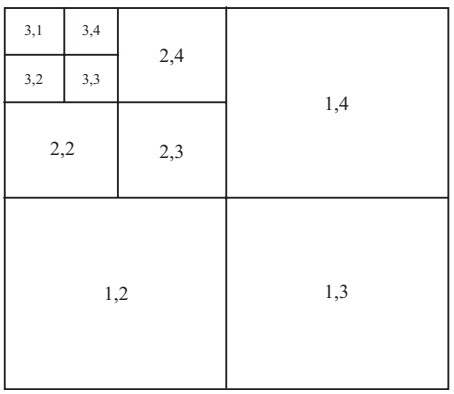

Fig. 3 Indexing of DWT bands. Each band is identified by a level and orientation (λ, θ). This example shows a three-level transform.

Table 3 Weighting factors $w_{\lambda,\theta}$ for three-level Coif22_14 DWT, experimentally determined.

	Level (λ)		
Orientation (θ)	1	2	3
1	—	—	0
2	−0.41	1.1	−0.1
3	−1.8	3.1	0
4	−0.41	1.1	−0.1

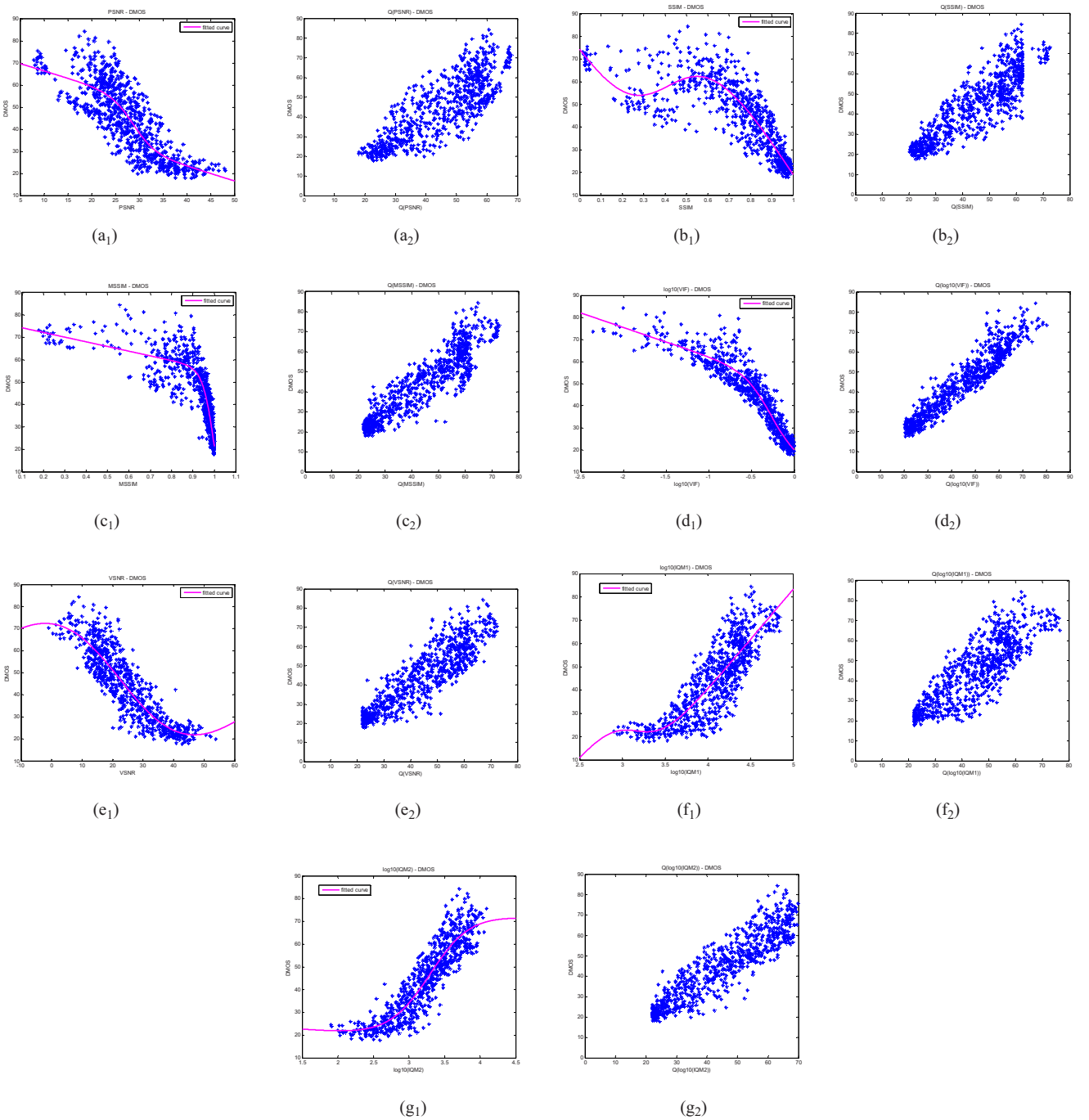

Fig. 4 Comparison of all 779 degraded images and objective quality measures with DMOS, before (index 1) and after (index 2) nonlinear fitting: (a) PSNR-DMOS, (b) ssim-DMOS, (c) MSSIM-DMOS, (d) VIF-DMOS, (e) VSNR-DMOS, (f) IQM1-DMOS, and (g) IQM2-DMOS.

optimization results than the CDF9_7 filter. For this filter, we have used the partial swarm optimization algorithm[19] to determine weighting factors for overall results (Section 6.2). We had 10 parameters to optimize for three-level decomposition (three factors for each level plus approximation factor). The main goal was to calculate as high a Pearson's correlation coefficient as possible for all 779 degraded images, before nonlinear regression. Weighting factors are given in Table 3. In this case, k was 2, also because with this parameter, we obtained the best overall optimization results. Factor k had to be assumed prior optimization because of overall calculation time. All three levels were used, disregarding only the approximation error. It should be noted that for calculating weighting factors, training and testing sets were both from the same image database (the LIVE image database). Using another image database, it is possible that weighting factors could have been calculated differently.

Final measure IQM is then calculated as

Table 4 Coefficient parameters for logistic function.

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	−23.25 (−33.94, −12.57)	0.4292 (0.2096, 0.6488)	28.71 (27.96, 29.45)	−0.6641 (−1.059, −0.2692)	61.49 (50.2, 72.79)
SSIM	−100.9 (−128.9, −72.8)	−7.904 (−9.698, −6.11)	0.4158 (0.4011, 0.4304)	−151.6 (−175.5, −127.8)	121.2 (111.1, 131.4)
MSSIM	−71.36 (−124.2, −18.48)	36.51 (23.82, 49.2)	1.002 (0.9657, 1.039)	−20.94 (−25.73, −16.14)	40.7 (13.17, 68.24)
$\log_{10}$(VIF)	−34.3 (−41.76, −26.83)	6.443 (4.845, 8.04)	−0.2692 (−0.3165, −0.2218)	−13.14 (−15.5, −10.78)	32.05 (30, 34.1)
VSNR	163 (−257.2, 583.1)	−0.07769 (−0.1624, 0.006981)	22.4 (20.95, 23.85)	1.432 (−3.402, 6.265)	15.04 (−93.66, 123.7)
$\log_{10}$(IQM1) (Watson's model)	−36.85 (−71.25, −2.446)	5.183 (1.016, 9.351)	3.168 (2.855, 3.481)	43.23 (38.4, 48.06)	−114.5 (−129.4, −99.57)
$\log_{10}$(IQM2) (experimentally determined)	57.36 (20.68, 94.04)	3.431 (2.048, 4.813)	3.292 (3.25, 3.335)	−2.56 (−17.8, 12.68)	55.09 (5.219, 105)

$$\mathrm{IQM} = \sum_{\lambda=1}^{3} \sum_{\theta=2}^{4} w_{i,j} E_{i,j}. \tag{9}$$

where w are weighting factors in the related subband and E is the error distance calculated according to Eq. (6). From Table 3, it can be seen that all subbands have to be included in the IQM2 measure except the approximation subband (3,1), but levels 1 and 3 have to be calculated using a negative weighting factor (experimentally, they give better results). Our experiments show that the best results for IQM1 measure (Watson's model) are obtained if we disregard level 1 (highest frequencies) and approximation error [from subband (3,1)] (see Table 2).

6 Results

6.1 *Performance Measures*

To be able to compare different IQMs and DMOS, we used several different measures of performance, as follows:

1. Pearson's product-moment correlation coefficient
2. root-mean-square error (RMSE)
3. Spearman's rank-order correlation coefficient

Pearson's product-moment correlation coefficient is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}, \quad i = 1, \ldots, n, \tag{10}$$

where, in Eq. (10), x_i and y_i are sample values (x are results for different objective measures and y are results for DMOS), $\bar{x}$ and $\bar{y}$ are sample mean, s_x and s_y are the standard deviation (calculated using $n-1$ in the denominator),

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i, \tag{11}$$

$$s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{12}$$

$$s_y = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{13}$$

Pearson's correlation reflects the degree of linear relationship between two variables, from −1 to 1, where 0 means that there is no relationship and ±1 means perfect fit.

RMSE is calculated as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n-k} (x-y)^2}. \tag{14}$$

where n is the number of tested images modified by a correction for degrees of freedom [$k=5$ in our case, we have five parameters in fitted function, Eq. (13)], x is DMOS measure, and y fitted objective measure after nonlinear regression.

Spearman's correlation coefficient is a measure of a monotone association that is used when the distribution of the data makes Pearson's correlation coefficient undesirable or misleading. Spearman's coefficient is not a measure of the linear relationship between two variables. It assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.[20]

6.2 *Overall Results: RMSE, Spearman's, and Pearson's Correlation*

Figure 4 shows a comparison between objective quality measures (MSE, PSNR, SSIM, MSSIM, VIF, VSNR, and IQM) and subjective quality measure (DMOS). SSIM, MSSIM, VIF, and VSNR were calculated using software from Ref. 21. We calculated Pearson's correlation coefficient before and after nonlinear regression. The nonlinearity chosen for regression for each of the methods tested was a five-parameter logistic function (a logistic function with an added linear term), as it was proposed in Ref. 22,

$$Q(x) = b_1\left(\frac{1}{2} - \frac{1}{1 + e^{b_2 \cdot (x - b_3)}}\right) + b_4 x + b_5. \quad (15)$$

However, this method has some drawbacks: First, the logistic function and its coefficients will have a direct influence on correlation (e.g., if someone chooses another function or even the same function with other parameters, the results can be quite different). Another drawback is that function parameters are calculated after the calculation of theobjective measures, which means that resulting parameters will be defined by the used image collection database. A different database can again produce different parameters. Coefficient parameters are given in Table 4.

As proposed in Ref. 22, the correlation coefficient is computed either by using measure directly or by its logarithm, whichever gave better correlation results and lower RMSE. By using this feature, MSE and PSNR give the same results if we compare $\log_{10}(\mathrm{MSE}) - \mathrm{DMOS}$ and $\mathrm{PSNR} - \mathrm{DMOS}$; thus, results for MSE will be excluded from further analysis.

We used the following three different methods to find the best fitting coefficients:

1. Trust–Region method[23]
2. Levenberg–Marquardt method[24,25]
3. Gauss–Newton method[26]

The final method for finding coefficients for nonlinear regression was the one that computed better results for performance measures (lower RMSE and higher Pearson's and Spearman's correlation).

For each graph in Fig. 4, it is calculated overall Pearson's and Spearman's correlation coefficients, as well as RMSE. They are presented in Fig. 5. When calculating correlation coefficients, those that are calculated before nonlinear regression are denoted on the figures with black bars and, after nonlinear regression, with gray bars. RMSE is calculated after nonlinear regression.

6.3 *Separate Results: RMSE, Spearman's, and Pearson's Correlation*

In this section, we examine how well each objective measure fits only one specific type of degradation, before and after nonlinear regression used in the previous section. Results for coefficient parameters for logistic function are presented in Tables 5–9 for different types of degradation. RMSE, Spearman's, and Pearson's correlation parameters for each type of degradation are given in Figs. 6–10. When calculating correlation coefficients, those that are calculated before nonlinear regression are denoted on figures with black bars and, after nonlinear regression, with gray bars. RMSE is calculated after nonlinear regression.

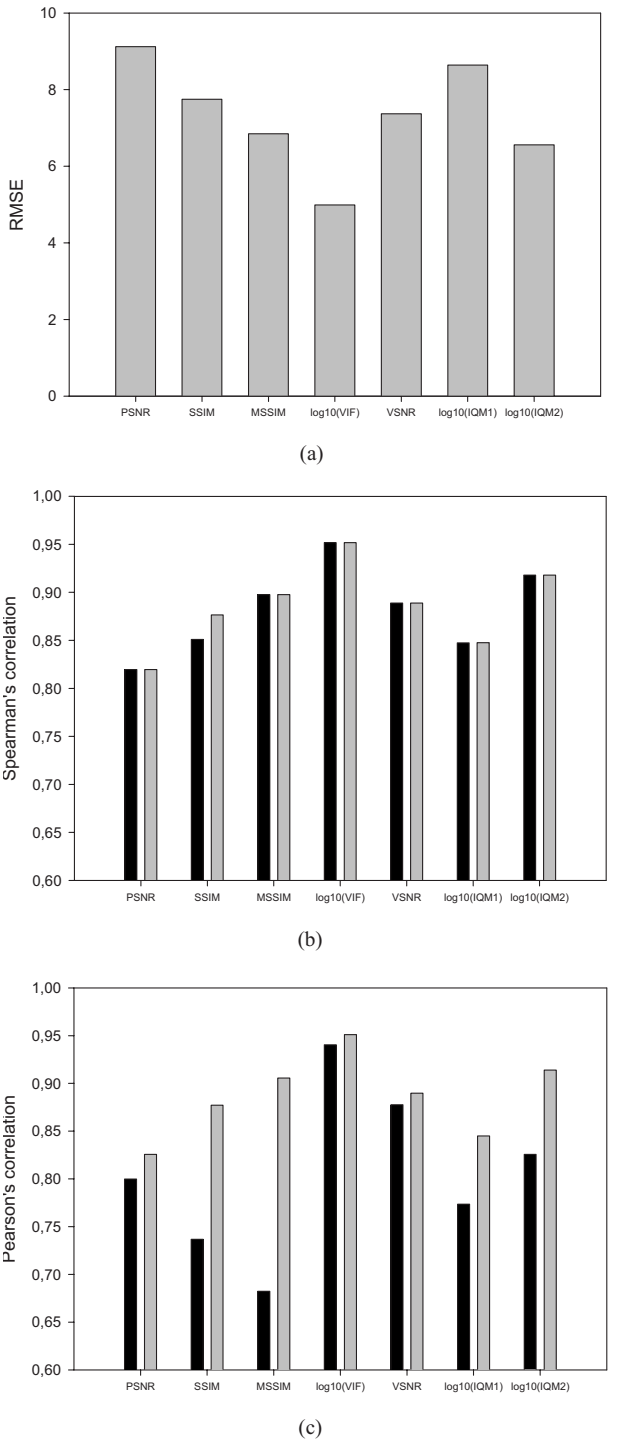

Fig. 5 Comparison of RMSE, Spearman's, and Pearson's correlation coefficient, for all 779 images in database: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression.

6.4 *Statistical Significance and Hypothesis Testing*

To be able to test whether results in Sections 6.1 and 6.2 are statistically significant, we used two hypothesis tests. First,

Table 5 Coefficient parameters for logistic function, for JP2K degradation (169 images).

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	−85.99 (−282.8, 110.8)	0.1779 (−0.03855, 0.3943)	29.08 (27.32, 30.84)	0.8319 (−3.767, 5.431)	23.19 (−111.5, 157.9)
SSIM	−66.5 (−186.5, 53.54)	8.855 (−2.241, 19.95)	0.947 (0.7301, 1.164)	−18.42 (−78.2, 41.35)	45.25 (−36.21, 126.7)
MSSIM	−1407 (−7415, 4600)	−5.532 (−14.62, 3.554)	0.7748 (0.7548, 0.7947)	−1928 (−7078, 3223)	1561 (−2414, 5537)
$\log_{10}$(VIF)	−33.86 (−51.62, −16.09)	−7.047 (−11.5, −2.591)	−1.036 (−1.141, −0.9308)	−61.88 (−69.97, −53.78)	0.8691 (−9.274, 11.01)
VSNR	−57.11 (−99.66, −14.56)	0.1801 (0.08609, 0.2742)	25.31 (24, 26.62)	0.1407 (−0.7327, 1.014)	43.57 (20.81, 66.33)
$\log_{10}$(IQM1)	93.37 (−51.97, 238.7)	3.145 (0.5175, 5.772)	3.98 (3.906, 4.054)	−18.46 (−77.99, 41.06)	120.9 (−115.4, 357.3)
$\log_{10}$(IQM2)	81.3 (−24.99, 187.6)	2.994 (0.7852, 5.203)	3.206 (3.136, 3.275)	−12.72 (−53.21, 27.78)	87.43 (−41.65, 216.5)

we calculated residuals between each observed quality measure (after nonlinear regression) and DMOS. For each residual set, the p value was calculated, which is the probability in statistical hypothesis testing, under the assumption of the null hypothesis, of observing the given statistic or one more extreme. The result is called statistically significant if it is unlikely to have occurred by chance. The lower the p value is, the less likely the result (in this case, the result is rejecting the null hypothesis); thus, the more significant the result is, in the sense of statistical significance. This means that for $p \leqslant 0.05$, the null hypothesis can be rejected at the 5% significance level (or with 95% confidence). Of course, significance level could be determined differently (e.g., 1% or 10%), in which case results would have been different.

We performed the first test, chi-square goodness-of-fit test to see if residuals have Gaussian distribution.[27] The Chi-square test has, in our case, a default null hypothesis that the data in vector **x** are a random sample from a normal distribution with mean and variance estimated from **x**, against the alternative that the data are not normally distributed with the estimated mean and variance. The result is 1

Table 6 Coefficient parameters for logistic function, for JPEG degradation (175 images).

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	−57.86 (−202.9, 87.18)	0.2477 (−0.1072, 0.6026)	29.55 28, 31.1	0.4373 (−4.12, 4.995)	30.41 (−106.1, 166.9)
SSIM	−95.51 (−202.1, 11.06)	9.035 (2.005, 16.07)	0.8969 (0.8161, 0.9777)	46 (−39.08, 131.1)	−6.375 (−92.89, 80.14)
MSSIM	−2197 (−14050, 9653)	−6.249 (−18.67, 6.171)	0.8262 (0.8158, 0.8366)	−3360 (−15100, 8375)	2838 (−6850, 12530)
$\log_{10}$(VIF)	−51.37 (−76.86, −25.87)	6.975 (3.684, 10.27)	−0.2966 (−0.3462, −0.2471)	5.384 (−10.26, 21.03)	41.65 (36.18, 47.13)
VSNR	−375.7 (−3222, 2470)	0.0833 (−0.16, 0.3266)	28.7 (27.77, 29.63)	5.642 (−31.11, 42.4)	−119.6 (−1177, 937.6)
$\log_{10}$(IQM1)	50.8 (−9.193, 110.8)	5.867 (1.1, 10.63)	3.971 (3.919, 4.024)	−5.984 (−46.62, 34.66)	66.57 (−93.45, 226.6)
$\log_{10}$(IQM2)	1564 (-2.869×10^4, 3.182×10^4)	1.148 (−6.701, 8.996)	3.154 (3.108, 3.2)	−398.7 (−6018, 5220)	1300 (-1.642×10^4, 1.902×10^4)

Table 7 Coefficient parameters for logistic function, for white noise degradation (145 images).

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	7.673 (5.76, 9.586)	−1.739 (−3.895, 0.4184)	11.58 (10.42, 12.75)	−1.407 (−1.465, −1.348)	79.12 (77.89, 80.36)
SSIM	−342.4 (−3564, 2879)	−2.727 (−12.47, 7.016)	0.5188 (0.4929, 0.5447)	−260.3 (−1632, 1112)	177.5 (−532.4, 887.3)
MSSIM	−739.9 (−41390, 39910)	29.33 (0.2676, 58.39)	1.118 (−0.9231, 3.159)	−40.59 (−43.36, −37.83)	−289.8 (−20620, 20040)
$\log_{10}$(VIF)	−357.4 (−10320, 9607)	3.605 (−1.918, 9.128)	0.7482 (−8.48, 9.976)	−20.89 (−26.84, −14.94)	−140.5 (−5114, 4833)
VSNR	7.893 (5.826, 9.96)	−1.48 (−3.817, 0.8567)	11.18 (10.03, 12.33)	−1.08 (−1.142, −1.019)	70.22 (69.04, 71.41)
$\log_{10}$(IQM1)	7.584 (6.056, 9.112)	84.46 (−203.8, 372.7)	4.628 (4.595, 4.661)	28.17 (26.98, 29.37)	−68.67 (−73.9, −63.43)
$\log_{10}$(IQM2)	7.788 (6.249, 9.327)	230.1 (−684.3, 1144)	3.749 (3.71, 3.788)	28.14 (26.92, 29.37)	−43.67 (−47.95, −39.4)

if the null hypothesis can be rejected at the 5% significance level and 0 if the null hypothesis cannot be rejected at the 5% significance level. Results of the chi-square test are presented in Table 10.

The second test, the F test, was performed on each of the two sets of calculated quality measure residuals. Because in our case it relies on the hypothesis that in every case, tested pairs of variables have normal distribution, the chi-square test was performed before (Table 10). Unfortunately, sometimes chi-square goodness of the fit test failed, meaning that the F test can give an unreliable conclusion.

The F test has the default null hypothesis that two independent samples, in the vectors **x** and **y**, come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances (two-tailed test).[27] One-tailed test is also possible, where the null hypothesis is the same as in the two-tailed test (variances are equal), but the alternative is that the variance for the first variable is better (lower) than the variance for the second (left-tailed test) or the variance for the first variable is worse (higher) than the variance for the second (right-tailed test).

Table 8 Coefficient parameters for logistic function, for Gaussian blur degradation (145 images).

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	434 (−20830, 21700)	−0.07344 (−1.537, 1.39)	20.2 (−27.33, 67.73)	4.614 (−226.5, 235.7)	−29.52 (−4389, 4330)
SSIM	−259.4 (−4309, 3790)	−3.59 (−30.7, 23.52)	0.1306 (−4.927, 5.188)	−185 (−759.6, 389.7)	85.01 (−1171, 1341)
MSSIM	−3791 (−28980, 21390)	−3.239 (−10.92, 4.438)	0.7378 (0.7188, 0.7567)	−3068 (−16210, 10080)	2332 (−7358, 12020)
$\log_{10}$(VIF)	−28.27 (−39.26, −17.29)	7.385 (4.08, 10.69)	−0.2916 (−0.3598, −0.2233)	−21.7 (−25.96, −17.44)	29.97 (27, 32.94)
VSNR	123.8 (−93.73, 341.3)	−0.1246 (−0.2455, −0.0037)	16.23 (12.73, 19.73)	0.9105 (−2.445, 4.266)	41.56 (−9.934, 93.06)
$\log_{10}$(IQM1)	−123.4 (−569.8, 323)	4.285 (−5.601, 14.17)	3.725 (3.046, 4.405)	134.2 (−11.13, 279.5)	−477.1 (−954, −0.345)
$\log_{10}$(IQM2)	54.97 (−102.3, 212.2)	4.969 (−3.347, 13.28)	3.401 (3.296, 3.506)	2.933 (−94.71, 100.6)	41.05 (−290.7, 372.8)

Table 9 Coefficient parameters for logistic function, for Fastfading degradation (145 images).

Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
PSNR	220.9 (−2089, 2531)	−0.09387 (−0.5022, 0.3145)	23.98 (18.44, 29.52)	2.572 (−29.61, 34.76)	−10.15 (−772.9, 752.6)
SSIM	−27.02 (−49.52, −4.509)	19.2 (1.607, 36.8)	0.9193 (0.8415, 0.9971)	−38.91 (−49.31, −28.52)	67.26 (52.54, 81.98)
MSSIM	−2895 (−512000, 506200)	24.88 (−7.32, 57.07)	1.185 (−6.175, 8.544)	−35.64 (−49.02, −22.26)	−1360 (−255900, 253200)
$\log_{10}$(VIF)	−31.63 (−52.34, −10.93)	7.958 (2.845, 13.07)	−0.1506 (−0.3028, 0.00166)	−14.66 (−17.03, −12.29)	27.34 (18.83, 35.85)
VSNR	20.7 (−3.19, 44.59)	−0.2728 (−0.6424, 0.0968)	14.53 (9.837, 19.22)	−0.8506 (−1.288, −0.4135)	69.52 (61.84, 77.2)
$\log_{10}$(IQM1)	118.9 (−224.8, 462.6)	2.627 (−1.669, 6.922)	4.421 (4.071, 4.771)	−16.84 (−120.9, 87.19)	130.7 (−345.6, 607)
$\log_{10}$(IQM2)	220 (−1004, 1444)	1.27 (−1.942, 4.483)	3.553 (3.062, 4.045)	−32.44 (−249.1, 184.3)	170 (−620.3, 960.2)

Results are presented in Tables 11–16 for each type of degradation and for all degradations. The result is "−" if the null hypothesis (variances are equal) cannot be rejected at the 10% significance level for the two-tailed test or 5% significance level for the one-tailed test; results for one-tailed and two-tailed tests will be then equal for the null hypothesis, because p in the one-tailed test (−5%) is one-half of the two-tailed test (−10%).

Letter s means that the null hypothesis can be rejected at the 5% level, and the variance for tested residual in a row is better (lower) than tested residual in a column (left-tailed test), and L if the null hypothesis can be rejected at the 5% level and variance for tested residual in a row is worse (higher) than the tested residual in a column (right-tailed test). For each tested residual pair, the p value for the two-tailed test is written in Tables 11–16. For the one-tailed test, p is one-half (or 1 minus one-half, depending on the test, if it is right or left tailed) of the two-tailed test.

From Tables 11–16, it can be concluded that objective measures have variances from the highest to the lowest in this order (different brackets refer to statistically indistinguishable variances of measures), as follows:

1. JP2K: PSNR − SSIM − (MSSIM − VIF − VSNR − IQM1 − IQM2)
2. JPEG: PSNR − (IQM1 − [IQM2) − {SSIM] − MSSIM − VIF − VSNR}
3. WN: (SSIM − VSNR) − [PSNR − MSSIM − VIF − IQM1 − IQM2]
4. Gblur: PSNR − (SSIM − IQM1 − IQM2) − VSNR − MSSIM − VIF
5. Fastfading: (PSNR − [VSNR) − IQM1 − {MSSIM − IQM2] − SSIM} − VIF
6. Overall: PSNR − IQM1 − (SSIM − VSNR) − [MSSIM − IQM2] − VIF

6.5 *Computational Complexity*

Results of the average time required to calculate each of these measures is given in Table 17. The average time is calculated over the entire database (982 images), with an average size of 768 × 512 pixels. MSE and PSNR are calculated using Eqs. (1) and (2) directly and all other measures, except IQM1 and IQM2, using software from Ref. 21. IQM1 and IQM2 are calculated using Matlab.m files. DWT for IQM measures was calculated using software from Ref. 28. The same computer configuration was used for calculating all objective measures: AMD Athlon64 X2 4200 MHz, 4 GB RAM, Windows Vista 64. It is probably possible to speed up algorithms using the MEX-compiler from C/C++ or Fortran source code instead of Matlab.m files.

6.6 *Discussion of the Results*

In Sections 6.2 and 6.3, we tested objective measures using three performance measures: RMSE, Pearson's, and Spearman's correlation coefficient. The significance of these results are tested in Section 6.4. Because IQM2 measure always gave similar or better results than IQM1, in further analysis we will compare IQM2 measure with other ones.

Although our measure uses multiscale wavelet decomposition, some other measures are also based on a similar idea. VSNR also uses 9/7 CDF wavelet for weighting different scales in both the first and second stages of its measurement. VIF also uses wavelet decomposition (steerable pyramid decomposition with six orientations) to compare information that is shared between tested and reference images in order to quantify information fidelity relative to the information content of the reference image. Our IQM directly uses differences in wavelet scales to determine the final grade based on weighting factors, unlike these other

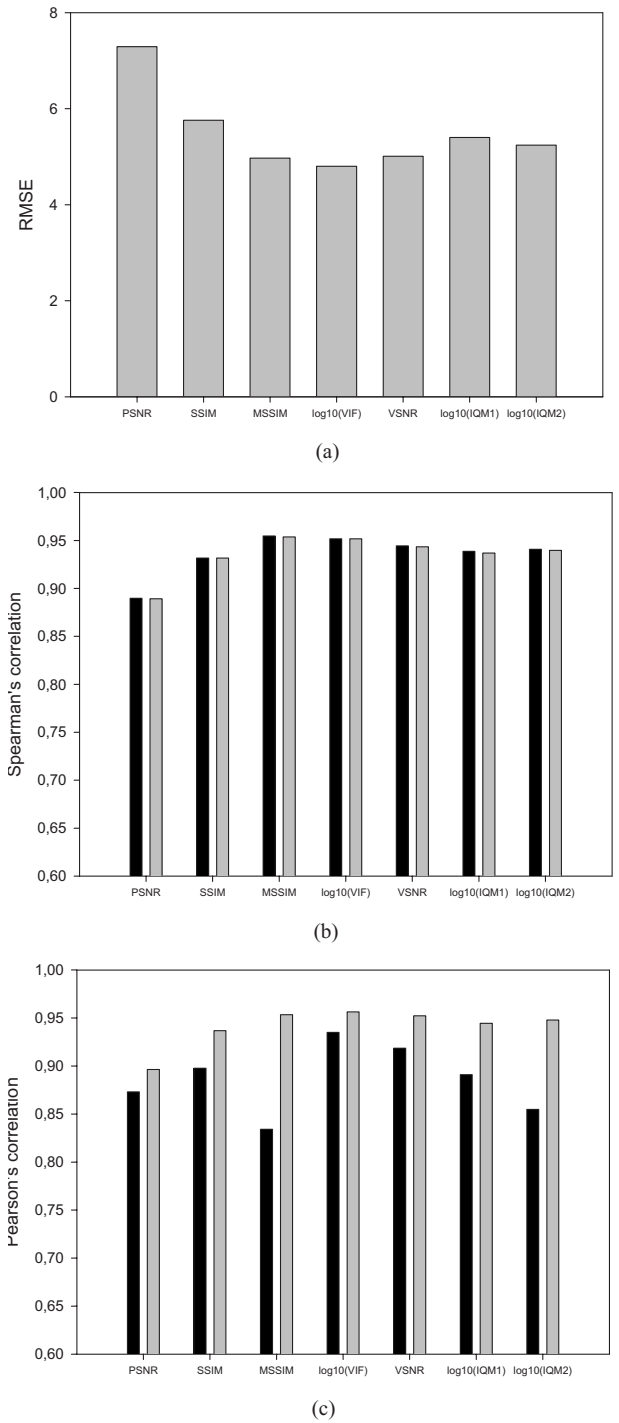

Fig. 6 Comparison of RMSE, Spearman's, and Pearson's correlation coefficient, for JP2K degradation: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression

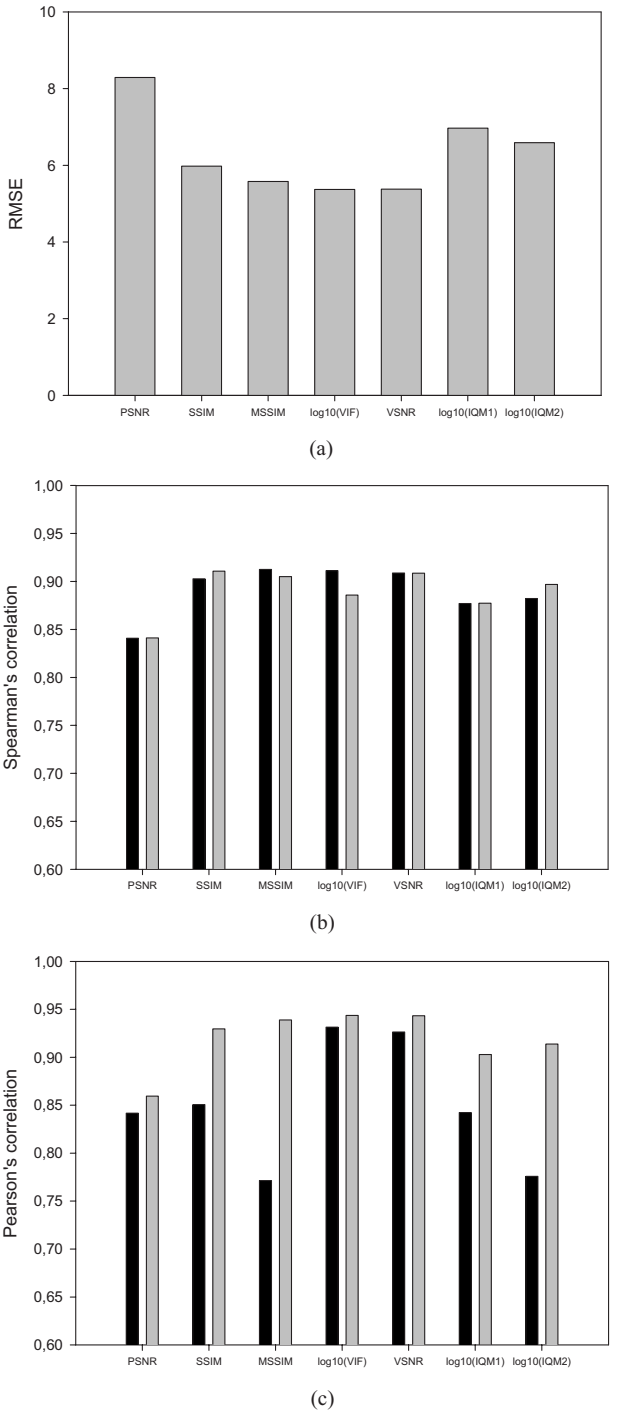

Fig. 7 Comparison of RMSE, Spearman's, and Pearson's correlation coefficient, for JPEG compression: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression.

measures that use much more complicated calculations with not always better results (for VSNR measure) in our experiment. VIF measure always outperformed our IQM measure, but tests have been made only on the image database[7] with fitting the function described in Eq. (15). Reference 15, which describes VSNR results, uses their own database and a slightly different fitting function, and claims to be better than VIF.

In Section 6.2, we tested overall results (all 779 images). Generally, the best results were obtained using the VIF objective quality measure. Our measure IQM2 gave the sec-

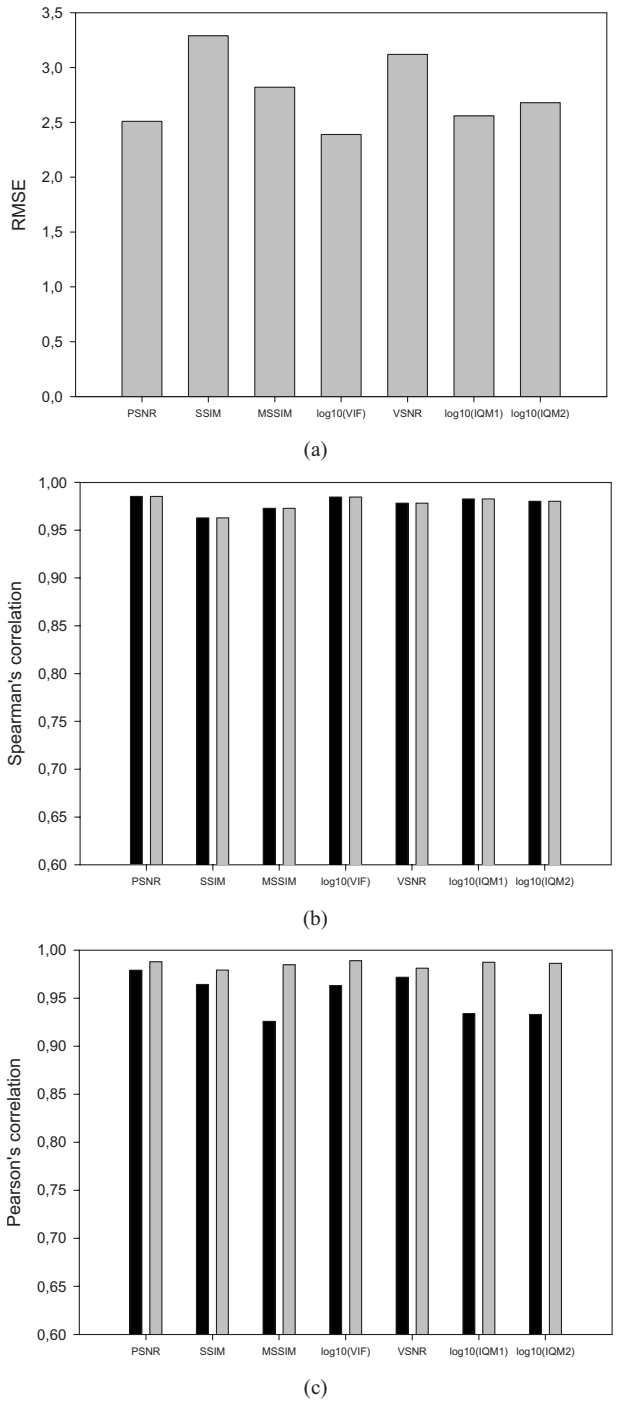

Fig. 8 Comparison of RMSE, Spearman's, and Pearson's correlation coefficient, for WN degradation: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression.

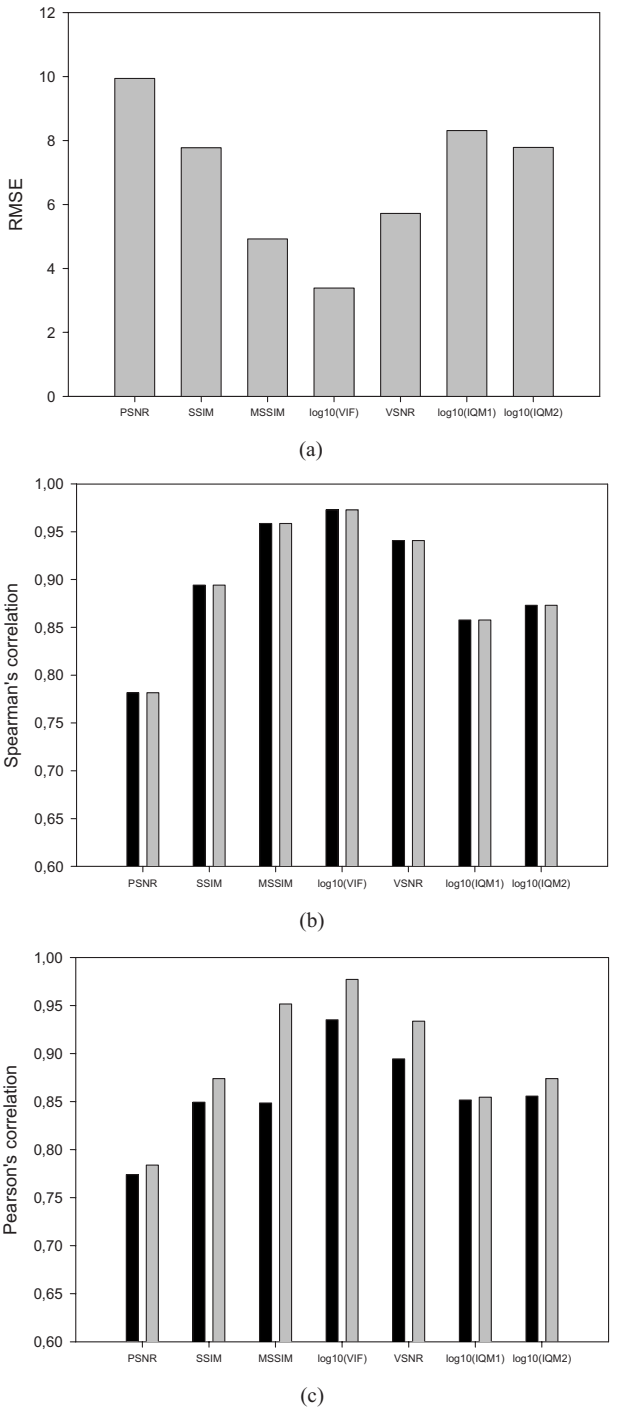

Fig. 9 Comparison of RMSE, Spearman's, and Pearson's correlation coefficient, for Gblur degradation: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression.

ond best results, somewhat better than MSSIM. After them, VSNR gave somewhat better results than SSIM, then our first proposed measure IQM1, and finally, MSE and PSNR gave the worst results. This order applies for all performance measures, which means each one of them follows the other ones. From Section 6.4, we can see that VIF gives significantly better results than other measures. IQM2 gives statistically similar results with MSSIM and significantly better than all other quality measures. Unfortunately, from Table 11 it can be seen that SSIM, MSSIM, VIF, and IQM2 do not have normal distribution (tested using chi-square test at 5% significance level) so it is questionable whether the F test results are confident.

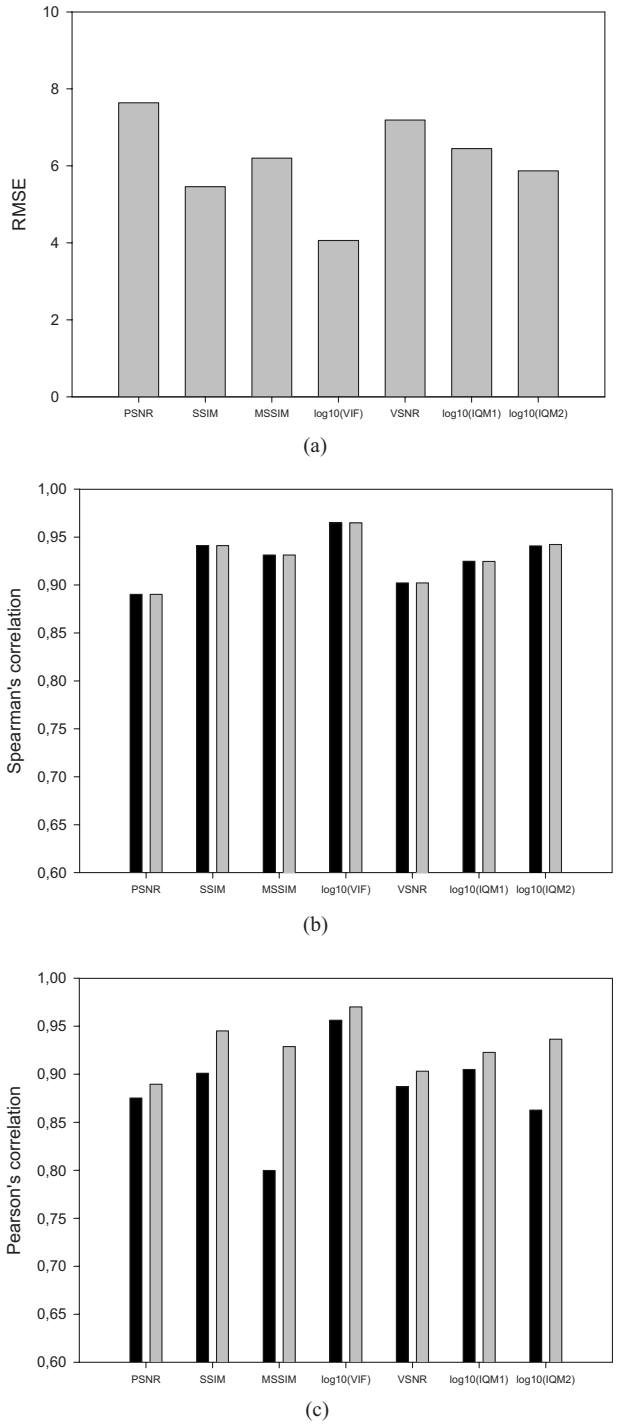

Fig. 10 Comparison of RMSE, Spearman's and Pearson's, correlation coefficient, for Fastfading degradation: (a) RMSE after nonlinear regression, (b) Spearman's correlation: black bars denote results before and gray bars after nonlinear regression, and (c) Pearson's correlation: black bars denote results before and gray bars after nonlinear regression.

As said previously, all performance measures rely on calculated coefficients for nonlinear fitting, which means that maybe it is possible to choose different coefficients that would yield a different conclusion. Another problem is that to be able to compare two or more different sets of images, they should be realigned to have the same distribution of subjective quality, so results cannot be taken strictly justified.[22] Anyway, such a comparison also has advantages that allows greater resolution in statistical analysis due to a larger number of tested images and also shows if a quality measure of one distortion type is consistent with another (Fig. 11). This means that measure should grade images equal if they have the same DMOS result but different types of degradation.

Section 6.3 calculates performance measures for each type of degradation separately. Generally, here again the VIF measure gives similar (JP2K and WN) or much better results (JPEG, Gblur, and fastfading) than all other quality measures. Our measure IQM2 gives similar and statistically indistinguishable results in JP2K and WN like VIF, yet still has good results in fastfading degradation (similar to MSSIM and SSIM, only worse than VIF). Also, from Table 10 it can be seen that in degradations JP2K and WN, nearly all measures have a statistical distribution similar to normal; thus, the F test can be assumed to be accurate. Only the fastfading degradation image set failed the chi-square test for SSIM, MSSIM, and VIF measure. When testing the JPEG test set of images, IQM2 gave the worst results than SSIM, MSSIM, VIF, and VSNR (yet, statistically indistinguishable to SSIM), but the chi-square test in JPEG test images failed on all quality measures (except MSE and PSNR). In the Gaussian blur test images, again IQM2 gave similar results to SSIM and worse than MSSIM, VIF, and VSNR. The chi-square test passed in this case for all measures except SSIM, which means that the F test can be taken to be accurate.

It can be also noted from Tables 7 and 13 that for white noise, PSNR (and subsequently MSE if we calculate its logarithm) gives similar results like the other much better quality measures tested in this paper.

Our quality measure gives very good results, given the simplicity of its idea. However, it is still not as good as some other algorithms like VIF. Anyway, from Table 17 it can be seen that such complex measures (such as VIF) are time consuming and cannot be used in applications where time is of importance (at least not without much optimization).

One example that shows how each objective measure (before nonlinear regression) grades the same image with similar DMOS grades and different types of degradation is shown in Table 18. Accordingly, Figs. 11(b)–11(f) show an error estimation for the same image in different wavelet subbands for each type of degradation. Error estimation is based on the difference image, which is decomposed with three decomposition levels. Results shown in Fig. 11 represent absolute values of wavelet coefficients amplified eight times to achieve better error visibility. It can be seen that different degradation types have errors in different subband spaces. The upper left corner of Figs. 11(b)–11(f) is generally rather bright because it shows the approximation coefficients difference, which are calculated using only a decomposition low-pass filter; thus, they represent the "average" difference values, unlike other subbands. From Figs. 11(b)–11(f), it can be concluded that probably better correlation results would have been obtained if we used an adaptive algorithm, e.g., that will calculate or choose weighting factors according to the type of the degradation.

Table 10 Chi-square test: A hypothesis result (H) of 0 means that related residual has normal distribution, 1 means that it does not have normal distribution, at the 5% significance level.

	JP2K		JPEG		WN		Gblur		Fastfading		All	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	0	0.2770	0	0.2489	0	0.6600	0	0.7428	0	0.0624	0	0.0529
SSIM	0	0.4161	1	0.0049	0	0.5910	1	0.0369	1	0.0065	1	8×10^{-6}
MSSIM	0	0.5307	1	0.0176	0	0.0969	0	0.3548	1	0.0022	1	0.0055
VIF	0	0.9932	1	3×10^{-7}	0	0.4555	0	0.2912	1	0.0287	1	1×10^{-8}
VSNR	0	0.2807	1	0.0162	1	0.0362	0	0.0651	0	0.4365	0	0.0691
IQM1	0	0.1139	1	0.0172	0	0.7610	0	0.4972	1	0.0260	0	0.2338
IQM2	0	0.4869	1	2×10^{-5}	0	0.7983	0	0.4090	0	0.1414	1	0.0085

Table 11 F test, JP2K degradation.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	L	0.0023	L	9×10^{-7}	L	1×10^{-7}	L	2×10^{-6}	L	1×10^{-4}	L	2×10^{-5}
SSIM	S	0.0023	—	1	L	0.0570	L	0.0194	L	0.0739	—	0.4106	—	0.2243
MSSIM	S	9×10^{-7}	S	0.0570	—	1	—	0.6621	—	0.9071	—	0.2789	—	0.4899
VIF	S	1×10^{-7}	S	0.0194	—	0.6621	—	1	—	0.5798	—	0.1288	—	0.2598
VSNR	S	2×10^{-6}	S	0.0739	—	0.9071	—	0.5798	—	1	—	0.3339	—	0.5661
IQM1	S	1×10^{-4}	—	0.4106	—	0.2789	—	0.1288	—	0.3339	—	1	—	0.6944
IQM2	S	2×10^{-5}	—	0.2243	—	0.4899	—	0.2598	—	0.5661	—	0.6944	—	1

Table 12 F test, JPEG degradation.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	L	2×10^{-5}	L	3×10^{-7}	L	2×10^{-8}	L	2×10^{-8}	L	0.0228	L	0.0026
SSIM	S	2×10^{-5}	—	1	—	0.3695	—	0.1557	—	0.1682	S	0.0433	—	0.1987
MSSIM	S	3×10^{-7}	—	0.3695	—	1	—	0.6009	—	0.6302	S	0.0036	S	0.0293
VIF	S	2×10^{-8}	—	0.1557	—	0.6009	—	1	—	0.9668	S	6×10^{-4}	S	0.0070
VSNR	S	2×10^{-8}	—	0.1682	—	0.6302	—	0.9668	—	1	S	7×10^{-4}	S	0.0079
IQM1	S	0.0228	L	0.0433	L	0.0036	L	6×10^{-4}	L	7×10^{-4}	—	1	—	0.4606
IQM2	S	0.0026	—	0.1987	L	0.0293	L	0.0070	L	0.0079	—	0.4606	—	1

Table 13 F test, WN degradation.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	S	0.0014	—	0.1646	—	0.5513	S	0.0092	—	0.8130	—	0.4495
SSIM	L	0.0014	—	1	L	0.0687	L	2×10^{-4}	—	0.5472	L	0.0030	L	0.0143
MSSIM	—	0.1646	S	0.0687	—	1	L	0.0474	—	0.2221	—	0.2486	—	0.5258
VIF	—	0.5513	S	2×10^{-4}	S	0.0474	—	1	S	0.0014	—	0.4053	—	0.1767
VSNR	L	0.0092	—	0.5472	—	0.2221	L	0.0014	—	1	L	0.0178	L	0.0639
IQM1	—	0.8130	S	0.0030	—	0.2486	—	0.4053	S	0.0178	—	1	—	0.6031
IQM2	—	0.4495	S	0.0143	—	0.5258	—	0.1767	S	0.0639	—	0.6031	—	1

Table 14 F test, Gblur degradation.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	L	0.0034	L	4×10^{-16}	L	0	L	1×10^{-10}	L	0.0331	L	0.0035
SSIM	S	0.0034	—	1	L	7×10^{-8}	L	0	L	3×10^{-4}	—	0.4221	—	0.9940
MSSIM	S	4×10^{-16}	S	7×10^{-8}	—	1	L	1×10^{-5}	S	0.0706	S	8×10^{-10}	S	7×10^{-8}
VIF	S	0	S	0	S	10^{-5}	—	1	S	9×10^{-10}	S	4×10^{-24}	S	4×10^{-21}
VSNR	S	1×10^{-10}	S	3×10^{-4}	L	0.0706	L	9×10^{-10}	—	1	S	10^{-5}	S	3×10^{-4}
IQM1	S	0.0331	—	0.4221	L	8×10^{-10}	L	4×10^{-24}	L	10^{-5}	—	1	—	0.4264
IQM2	S	0.0035	—	0.9940	L	7×10^{-8}	L	4×10^{-21}	L	3×10^{-4}	—	0.4264	—	1

Table 15 F test, Fastfading degradation.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	L	7×10^{-5}	L	0.0127	L	2×10^{-13}	—	0.4584	L	0.0422	L	0.0016
SSIM	S	7×10^{-5}	—	1	—	0.1283	L	4×10^{-4}	S	0.0011	S	0.0473	—	0.3947
MSSIM	S	0.0127	—	0.1283	—	1	L	6×10^{-7}	S	0.0789	—	0.6416	—	0.5021
VIF	S	2×10^{-13}	S	4×10^{-4}	S	6×10^{-7}	—	1	S	3×10^{-11}	S	5×10^{-8}	S	10^{-5}
VSNR	—	0.4584	L	0.0011	L	0.0789	L	3×10^{-11}	—	1	—	0.1959	L	0.0154
IQM1	S	0.0422	L	0.0473	—	0.6416	L	5×10^{-8}	—	0.1959	—	1	—	0.2559
IQM2	S	0.0016	—	0.3947	—	0.5021	L	10^{-5}	S	0.0154	—	0.2559	—	1

Table 16 F test, all degradations together.

	PSNR		SSIM		MSSIM		VIF		VSNR		IQM1		IQM2	
	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value	H	*p* value
PSNR	—	1	L	7×10^{-6}	L	2×10^{-15}	L	0	L	4×10^{-9}	—	0.1320	L	0
SSIM	S	7×10^{-6}	—	1	L	5×10^{-4}	L	0	—	0.1591	S	0.0027	L	3×10^{-6}
MSSIM	S	2×10^{-15}	S	5×10^{-4}	—	1	L	0	S	0.0397	S	10^{-10}	—	0.2205
VIF	S	0	S	0	S	0	—	1	S	6×10^{-27}	S	2×10^{-50}	S	4×10^{-14}
VSNR	S	4×10^{-9}	—	0.1591	L	0.0397	L	6×10^{-27}	—	1	S	10^{-5}	L	0.0010
IQM1	—	0.1320	L	0.0027	L	10^{-10}	L	2×10^{-50}	L	10^{-5}	—	1	L	2×10^{-14}
IQM2	S	0	S	3×10^{-6}	—	0.2205	L	4×10^{-14}	S	0.0010	S	2×10^{-14}	—	1

7 Conclusion

In this paper, we proposed a new IQM based on DWT and Watson's model of noise visibility in different wavelet subbands. We examined how different objective measures correlate with subjective DMOS measure and presented two new objective measures. Our IQMs take into account properties of the HVS and provide better correlation with DMOS than some other quality measures. Proposed IQM could be considered as a good starting point for evaluation and a fair comparison of different types of image degradation, especially in applications where image-quality evaluation should be performed in real time. Although the results for VIF measure are slightly better than for our proposed IQM2 measure, computational time for IQM2 takes 1/25th of the time of the VIF calculation.

Further experiments could also include their own subjective testing and DMOS measurement in controlled conditions. In such a way, comparison and correlation could be computed more accurately regarding testing conditions, illumination, type of the display, viewing distance, etc. This way it would also be possible to use training and testing images for the optimization algorithm from different databases, thus removing the possibility of overfitting weighting factors to the LIVE image database only.

Also, IQM could be computed adaptively, depending on the type of the degradation. On the basis of IQM and properties of wavelet domain, the development of new no reference IQM could be considered as well.

Table 17 Average time required to calculate each measure.

Measure	Time (s)
MSE	0.0051
PSNR	0.0052
SSIM	0.1620
MSSIM	0.3593
VIF	8.1602
VSNR	0.9645
IQM1	0.2079
IQM2	0.3227

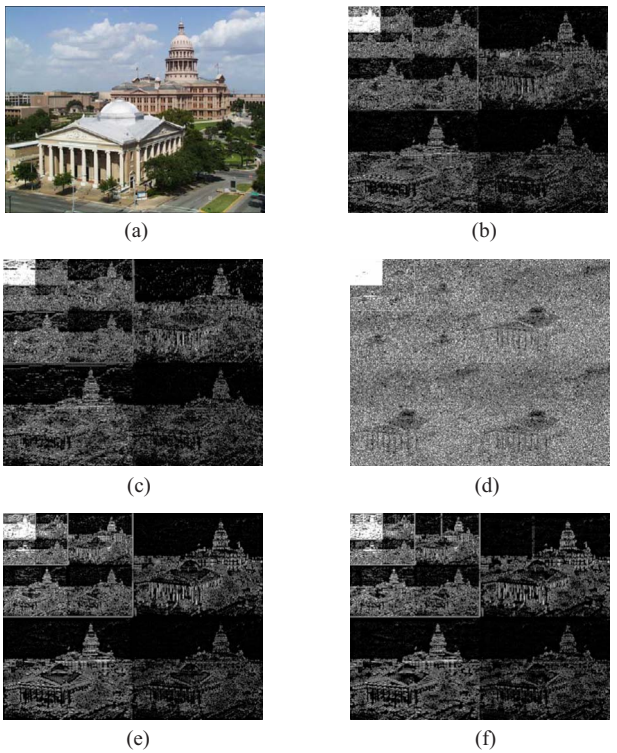

Fig. 11 Error estimation using three wavelet decomposition levels of the same image ("churchandcapitol.bmp" from the image database) with similar DMOS results and different degradation types: (a) original image, (b) JP2K compression, (c) JPEG compression, (d) white noise, (e) Gaussian blur, and (f) Fastfading.

Table 18 Objective quality measures and DMOS for image "churchandcapitol.bmp."

	JP2K	JPEG	WN	Gblur	Fastfading
DMOS	68.9113	79.632	75.678	82.289	76.971
MSE	282.5659	284.4560	8776	1177	1500
PSNR	23.6196	23.5907	8.6981	17.4241	16.3706
SSIM	0.7098	0.6986	0.0352	0.4678	0.4443
MSSIM	0.9020	0.8913	0.2275	0.5840	0.4343
VIF	0.1849	0.2228	0.0286	0.0299	0.0082
VSNR	17.3080	17.8571	5.8241	7.2399	5.6962
IQM1	20300	18650	61660	34730	40950
IQM2	3894	4040	7532	6.555	7616

Acknowledgment

The work described in this paper was conducted under the research projects: "Picture quality management in digital video broadcasting" (Grant No. 036-0361630-1635), and "Intelligent Image Features Extraction in Knowledge Discovery Systems" (Grant No. 036-0982560-1643), supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

1. S. Grgic, M. Grgic, and B. Zovko-Cihlar, "Performance analysis of image compression using wavelets," *IEEE Trans. Ind. Electron.* **48**(3), 682–695 (June 2001).
2. Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality," ⟨http://www.vqeg.org/⟩ (Sept. 2008).
3. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–444 (Feb. 2006).
4. T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed., pp. 939–959, Academic Press, New York (2005).
5. A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.* **6**(8), 1164–1175 (Aug. 1997).
6. A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Process.* **5**(8), 717–730 (May 1999).
7. H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2," http://live.ece.utexas.edu/research/quality.
8. S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, New York (1999).
9. N. Sprljan, S. Grgic, and M. Grgic, "Selection of biorthogonal filters for wavelet image compression," *Proc. 10th Int. Workshop on Systems, Signals, and Image Processing (IWSSIP'03)*, Prague, pp. 48–52 (Sept. 10–11 2003).
10. ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Int. Telecommun. Union/ITU Radiocommun. Sector (Jan. 2002).
11. H. R. Sheikh, "Image quality assessment using natural scene statistics," Ph.D. dissertation, University of Texas at Austin (May 2004).
12. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (April 2004).
13. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *37th Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA (Nov. 2003).
14. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–434 (2006).
15. D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.* **16**(9), 2284–2298 (Sept. 2007).
16. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Process.* **1**(2), 205–220 (April 1992).
17. A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.* **45**(5), 485–560 (1992).
18. D. Wei, H. T. Pai, and A. C. Bovik, "Antisymmetric biorthogonal coiflets for image coding," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, vol. 2, pp. 282–286 (Oct. 1998).
19. J. Kennedy and R. C. Eberhart, "Particle swarm optimization," *Proc. of IEEE Int. Conf. on Neural Networks*, vol. IV, pp. 1942–1948 (1995).
20. J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data," *Proc. of MAT TRIAD 2007 Conf.*, Bedlewo, Poland (Mar. 2007).
21. Visual Quality Assessment Package Version 1.1, Available at ⟨http://foulard.ece.cornell.edu/gaubatz/metrix_mux/⟩.
22. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**(11), 3440–3451 (Nov. 2006).
23. J. J. Moré and D. C. Sorensen, "Computing a trust region step," *SIAM J. Sci. Comput. (USA)* **4**(3), 553–572 (1983).
24. K. Levenberg, "A method for the solution of certain problems in least squares," *Q. Appl. Math.* **2**, 164–168 (1944).
25. D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.* **11**, 431–441 (1963).
26. J. E. Dennis Jr., "Nonlinear least-squares," in *State of the Art in Numerical Analysis*, D. Jacobs, Ed. pp. 269–312, Academic Press, New York (1977).
27. D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 3rd ed., Wiley, Hoboken, NJ (2003).
28. Wavelet Toolbox v2.10, available at ⟨http://www.sprljan.com/nikola/matlab/wavelet.html⟩.

Emil Dumic received his BSc in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 2007. He is currently a PhD student at the Department of Wireless Communications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His research interests include image interpolation, wavelet transforms, and digital satellite television.

Sonja Grgic received her BSc, MSc, and PhD in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 1989, 1992, and 1996, respectively. She is currently a full professor there in the Department of Wireless Communications. Her research interests include television signal transmission and distribution, picture quality assessment, and wavelet image compression. She has had more than 120 scientific papers published in international journals and conference proceedings.

Mislav Grgic received his BSc, MSc, and PhD in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 1997, 1998, and 2000, respectively. He is currently an associate professor there in the Department of Wireless Communications. His research interests include multimedia communications and image processing. He has had more than 100 scientific papers published in international journals and conference proceedings.