SIViP (2014) 8:1159–1168
DOI 10.1007/s11760-014-0654-3

ORIGINAL PAPER

IQM2: new image quality measure based on steerable pyramid wavelet transform and structural similarity index

E. Dumic · S. Grgic · M. Grgic

Received: 5 July 2013 / Revised: 8 May 2014 / Accepted: 25 May 2014 / Published online: 28 June 2014

Abstract In this paper, we present a new full-reference objective image quality measure—IQM2, based on structural similarity index and steerable pyramid wavelet transform. IQM2 is tested using different number of orientation kernels and seven subjective databases. Finally, IQM2 measure is compared with twelve commonly used full-reference objective measures. Results show that proposed IQM2 measure, using kernel with 2 orientations, provides good correlation with the results of subjective evaluation while keeping computational time lower than other similar performing objective measures.

Keywords Image quality · Image decomposition · Image databases · IQM2 measure · SSIM index · SPWT transform

1 Introduction

Objective image quality evaluation plays an important role in many image and video processing techniques, such as compression [1], interpolation, image processing [2], and watermarking [3] where evaluation method is based on image quality estimation. Quality of image can be evaluated using different measures. The best way to do that is by making a visual experiment under controlled conditions, in which human observers grade image quality [4]. Such experiments are time consuming and costly. Much easier approach is to use some objective measure that evaluates the numerical error between the original image and distorted image [5]. Every objective image quality measure has its aim to approximate the human quality perception (or human visual system, HVS) as much as possible, which means to correlate well with subjective measures (mean opinion score, MOS). Depending on the development type, objective measures can be based on bottom-up or top-down approaches. In bottom-up approach, underlying premise is that the sensitivities of the human visual system (HVS) are different for different aspects of the visual signal that it perceives. Unlike these models, top-down approach is not affected by assumptions about HVS models, but is motivated instead by the need to capture the loss of visual structure in the signal that the HVS hypothetically extracts for cognitive understanding. Some objective measures can be a combination of both approaches.

Objective quality measures according to the reference information they are using can be generally divided into three categories:

- full-reference image quality measures;
- reduced-reference image quality measures;
- no-reference image quality measures.

In this paper, we propose a new full-reference image quality measure, IQM2, which is based on top-down approach, and compare it with twelve commonly used full-reference objective measures. Except newly proposed IQM2 measure, we tested several other publicly available objective image quality measures, which were also usually tested and reported in other papers. These measures are as follows:

- MSE (mean squared error) [6];
- NAE (normalized absolute error) [6];
- SSIM (structural similarity) [7];
- MS-SSIM (multiscale SSIM, MSSIM) [8];
- VIF (visual information fidelity) [9];
- VIFP (pixel-based VIF) [9];
- VSNR (visual signal-to-noise ratio) [10];

E. Dumic (✉) · S. Grgic · M. Grgic
Department of Wireless Communications,
Faculty of Electrical Engineering and Computing,
University of Zagreb, Unska 3, 10000 Zagreb, Croatia
e-mail: emil.dumic@fer.hr

- NQM (noise quality measure) [11];
- CW-SSIM (complex wavelet SSIM) [12];
- IW-PSNR (information content weighted PSNR) [13];
- IW-SSIM (information content weighted SSIM) [13];
- MAD (most apparent distortion) [14].

Generally, full-reference image quality measures can be divided into several groups, some of which are explained in detail in [15]:

- pixel-based metrics (e.g., MSE, PSNR (peak signal-to-noise ratio), and NAE measures): these measures presume that every pixel in image is equally important;
- structural distortion based metrics (UIQI (Universal Quality Index) [16], SSIM, MS-SSIM, CW-SSIM): such metrics consider the assumption that the human visual system is good at extracting the structural information from the scene as well as local properties of an image. Newly proposed measure IQM2 belongs in this group;
- natural scene statistic models: NSS metrics (IFC (information fidelity criterion) [17], VIF, VIFP, IW-PSNR, and IW-SSIM measures): These algorithms capture regularities of natural images and try to quantify how these regularities are modified or lost when distortions occur. They can be combined with measure groups mentioned earlier (like IW-PSNR and IW-SSIM);
- multiple strategy-based metrics (VSNR and MAD measures): measures that attempt to model HVS in two separate stages: near-threshold and clearly visible distortions. Final measure is weighted combination of these two stages;
- other metrics (e.g., NQM, WSNR (weighted signal-to-noise ratio), PQS (Picture Quality Scale), etc.): measures based on contrast sensitivity function (NQM and WSNR) [18], PQS which takes into account the properties for both global features and localized disturbances [19].

Pixel-based metrics are generally the fastest to calculate, however, they have the lowest correlation with MOS. Structural distortion measures have better correlation than pixel-based metrics and are slower in calculation time. NSS metrics and multiple strategy-based metrics have in general best correlation with MOS and are, also in general, the slowest.

All tested measures can be calculated using Matlab and codes can be downloaded from the following locations: MSE, VIF, VIFP, VSNR, and NQM can be calculated using program "Metrix_mux" [20]; CW-SSIM measure can be downloaded from [21]; IW-PSNR, IW-SSIM, SSIM, and MS-SSIM measures can be downloaded from [22]; MAD measure can be downloaded from [23]. PSNR (peak signal-to-noise ratio) results are the same as MSE so they will not be shown. Same is for IW-MSE and IW-PSNR, but here we used IW-PSNR. IFC, UIQI, and WSNR results were also skipped because same authors proposed measures based on them (VIF, SSIM, and NQM, respectively). All measures (between tested ones) give best results when they are calculated from luminance component only, which means that all color images had to be transformed in gray scale images. However, incorporating color information is also an important issue, so some authors proposed measure's extension based on color perceptual model [24]. For measures that are using wavelet transforms, depending on the number of scales and filter length, images had to be rescaled firstly.

This paper is organized as follows. In Sect. 2, proposed image quality measure (IQM2) is described in detail. Section 3 explains performance measures used for comparison of the objective measures, as well as statistical significance tests. Section 4 summarizes seven tested subjective databases on which we tested objective quality measures. Section 5 compares different objective image quality measures with results of subjective assessment. Finally, Sect. 6 gives the conclusion.

2 IQM2: new image quality measure

2.1 Previous work

In our previous work, we proposed measure IQM that was based on discrete wavelet transform (DWT) [25] and coefficients optimization at different subbands using particle swarm optimization algorithm (PSO) [26]. DWT was performed using Coiflet wavelet filter with 22-low-pass and 14-high-pass coefficients, Coif22_14. In [27], similar evaluations were performed, only here we used LIVE database as learning database and VCL@FER database as testing database. Because of that, somewhat different weighting factors were obtained. IQM measure was calculated as a product of weighted differences between original and degraded image at each subband [27]. It can be concluded that, when constructing image quality measure, much attention should be given to the weighting factors, because they can fit measure to the specific type of degradation or to the specific image database, instead of the HVS.

2.2 IQM2: new image quality measure

In previous work, we managed to have satisfying results for LIVE and VCL@FER databases. However, when testing other image databases (especially TID database with 17 degradation types), measure did not correlate well with MOS results. Because of that, we propose new IQM2 measure. Original and degraded images are firstly transformed using steerable pyramid wavelet transform (SPWT) with K orientations and maximal number of scales M and on each scale modified SSIM measure is calculated [28], with contrast and structure terms only. The steerable pyramid is a linear multiscale, multi-orientation image decomposition that provides a useful front end for image-processing and computer vision applications [29]. Kernel design is described in [30], and tool for the transform calculation can be downloaded from [31]. Other authors have also proposed framework for computing full-reference image quality scores in the discrete wavelet domain [32].

At every stage of the image decomposition, SSIM contrast (SSIM_cont) and structure (SSIM_struct) have been calculated, for passband coefficients only. High-pass and low-pass coefficients are not taken for calculation because it is not yet defined how to incorporate their information in the whole image measure (and to improve correlation with MOS). Because of this, it is useful to decompose an image to a maximum number of scales, M in (1).

Luminance component (which exist in original SSIM measure) could be eventually calculated only from the low-pass coefficients, because it is incorporated in these coefficients after SPWT transformation. However, calculating SSIM measure (or only parts from it: luminance, chrominance, or structure) from low-pass coefficients would reduce overall correlation with MOS, tested on TID database. In [33], it was shown that luminance component does not influence on correlation with subjective testing as much as contrast and structure. However, in [27], it was shown that luminance part should be in any case avoided, because it does not influence much in databases without luminance or contrast degradations, while in other, it can significantly reduce correlation. In the results section, modified SSIM measure (SSIMmod) will be also compared with other image quality measures, by using only contrast and structure components.

SSIM_cont represents comparison between contrasts of original and degraded images, where σ_x and σ_y are weighted standard deviations as an estimate of the signal contrast, (3). SSIM_struct considers image degradations as structural information in an image, and those attributes represent the structure of objects in the scene, independent of the average luminance and contrast [7], (4).

For filter used in our measure, maximum number of scales can be obtained by

$$M = \left\lfloor \log_2 \left(\frac{\min(I, J)}{D} \right) \right\rfloor + 1 \tag{1}$$

In (1), I and J are height and width of the original and degraded images, M is maximum number of scales for given filter, and D represents low-pass SPWT filter dimension [30] which depends on number of orientations K, Table 1. In the results section, it will be shown that the best correlation with MOS can be obtained using 1 or 2 orientations (Table 1). IQM2 measure is calculated according to the following formula:

Table 1 Filter dimensions D for different number of orientations K

K	1	2	4	6
D	13	17	17	9

$$\mathrm{IQM2} = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{SSIM\_cont}(m, k) \cdot \mathrm{SSIM\_struct}(m, k) \tag{2}$$

where M represents maximum number of scales, K number of orientations, and SSIM_cont and SSIM_struct are calculated according to the following:

$$\mathrm{SSIM\_cont} = \frac{2 \cdot \sigma_x \cdot \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3}$$

$$\mathrm{SSIM\_struct} = \frac{\sigma_{xy} + \frac{C_2}{2}}{\sigma_x \cdot \sigma_y + \frac{C_2}{2}} \tag{4}$$

In (3) and (4), σ_x^2, σ_y^2, σ_{xy}, and C_2 are similarly defined as in SSIM measure, only here they are calculated for each subband (m) and orientation (k) separately. σ_x^2 and σ_y^2 are weighted variances from original and degraded images within local window, and σ_{xy} is defined as weighted covariance between original and degraded images within local window. $C_2 = (Z_2 B)^2$ where Z_2 is constant experimentally determined ($Z_2 = 0.03$) and B is maximal amplitude. Also, normalized Gaussian filter was used with different window size (original SSIM uses 11×11 pixel window) with $\sigma = 1.5$.

When (3) and (4) for contrast and structure terms are inserted in (2), final IQM2 measure is given by:

$$\mathrm{IQM2} = \prod_{m=1}^{M} \prod_{k=1}^{K} \overline{\left(\frac{2 \cdot \sigma_{xy}(m, k) + C_2}{\sigma_x^2(m, k) + \sigma_y^2(m, k) + C_2} \right)} \tag{5}$$

Mean by the bar on top of (5) denotes that mean value of all local values (within window 5×5 pixels) which is calculated as their arithmetic mean in each subband separately. Algorithm for calculating IQM2 measure can be faster when SSIM contrast and structure at every scale is calculated simultaneously with decomposition. It does not need to perform full transformation and save coefficients, but rather filter, calculate modified SSIM on the according scale, and then downsample images. Also, using function “imfilter” from Matlab, advantage can be taken of the Intel integrated performance primitives (Intel IPP) library (which has to be present on the testing machine), thus accelerating its execution time.

IQM2 measure was tested using other wavelet decomposition types (DWT [25], Quincunx wavelet transform [34]) and calculating SSIM_cont and SSIM_struct at each scale of the decomposition. However, their correlation with MOS grades was lower so results were not shown.

3 Performance measures

3.1 Pearson's and Spearman's correlation

Each of the objective measures described earlier was graded using different performance measures: Pearson's correlation coefficient and Spearman's rank correlation coefficient.

Pearson's correlation coefficient is calculated as normalized covariance between 2 variables [35]. Because Pearson's correlation coefficient measures linear relationship between two variables, nonlinear regression should be done prior calculation of the correlation. The nonlinearity chosen for regression for each of the methods tested was a 5-parameter logistic function (a logistic function with an added linear term), as it was proposed in [36]:

$$Q(z) = b_1 \cdot \left(\frac{1}{2} - \frac{1}{1 + e^{b_2 \cdot (z - b_3)}}\right) + b_4 \cdot z + b_5 \tag{6}$$

However, this method has some drawbacks: firstly, logistic function and its coefficients will have direct influence on correlation (e.g., if someone chooses another function or even the same function with other parameters, results can be different). Another drawback is that function parameters are calculated after the calculation of the objective measures, which means that resulting parameters will be defined by the used image collection database. Different database can again produce different parameters. In [10], somewhat different logistic function is proposed, with 4 parameters. We calculated Pearson's correlation using this function also:

$$Q(z) = \frac{b_1 - b_2}{1 + e^{\frac{z - b_3}{b_4}}} + b_2 \tag{7}$$

We used three different methods to find the best fitting coefficients: Trust-region method, Levenberg–Marquardt method, and Gauss–Newton method [37].

Spearman's correlation coefficient [35] 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. Spearman's correlation coefficient is calculated like Pearson's correlation over ranked variables.

3.2 Statistical significance

To be able to test whether results between different objective quality measures are statistically significant, we used several hypothesis tests. Firstly, we calculated residuals between each observed quality measure (after nonlinear regression) and MOS or DMOS.

We performed first test, chi-square goodness of fit test to see whether residuals have Gaussian distribution [38].

Second test, *F* test, was performed on each of the two sets of calculated quality measure residuals [38] and calculated *P* value. Because in our case it relies on the hypothesis that in every case tested pairs of variables have normal distribution, chi-square test was performed before. Unfortunately, sometimes chi-square goodness of fit test failed, meaning that *F* test can give unreliable conclusion. In later comparison, we used significance level 10 % for two-tailed test (or 5 % for one-tailed test).

Because some of the residuals did not have Gaussian distribution, we also tested measures against Ansari–Bradley test [39]. In later comparison, we used significance level 10 % for two-tailed test (or 5 % for one-tailed). If the groups do not have the same medians, in [39], it is recommended to subtract the median in that case.

4 Subjective image databases

Subjective databases have important role in creating and testing new image quality measure. We used seven different image quality databases to determine parameters for new image quality measure and correlation with objective measures:

- VCL@FER (Video Communications Laboratory @ FER) [40]: 4 distortion types, 23 source images, 552 distorted images, 116 observers;
- A57 (A57 database) [41]: 6 distortion types, 3 source images, 54 distorted images, 7 observers;
- CSIQ (categorical image quality database) [42]: 6 distortion types, 30 source images, 866 distorted images, 35 observers;
- LIVE (Laboratory for Image & Video Engineering) [43]: 5 distortion types, 29 source images, 779 distorted images, 161 observers;
- IVC (Image and video communication) [44]: 5 distortions, 10 source images, 185 distorted images, 15 observers;
- TID (Tampere Image Database 2008) [45]: 17 distortion types, 25 source images, 1,700 distorted images, 838 observers;
- Toyama [46]: 2 distortion types, 14 source images, 168 distorted images, 16 observers.

Details about each database, tested image sizes, and specific degradation type can be found in their references.

Table 2 Spearman's correlation for IQM2 with different number of orientations K, 11 × 11 Gaussian filter

	K=1	K=2	K=4	K=6	*SSIM*	*SSIMmod*	*MSSIM*
A57	0.8463	**0.8555**	0.8471	0.8393	0.8067	0.8066	0.8415
CSIQ	0.9197	**0.9367**	0.929	0.8924	0.8756	0.9284	0.9136
LIVE	0.95	0.948	0.9466	0.943	0.9479	0.9478	**0.9513**
IVC	0.8915	0.8791	0.8742	0.8744	0.9018	**0.9029**	0.8980
VCL@FER	**0.9355**	0.9308	0.9297	0.9345	0.9113	0.9100	0.9227
TID	0.874	**0.881**	0.866	0.818	0.7749	0.8177	0.8542
TOYAMA	0.8863	0.8666	0.8583	0.8627	0.8794	0.8794	**0.8874**
MEAN	**0.9090**	0.9074	0.9007	0.8857	0.8711	0.8847	0.8955
WT_MEAN	0.9057	**0.9098**	0.9012	0.8750	0.8539	0.8813	0.8955

5 Results

5.1 IQM2 with different filter size and different number of orientation kernels

In this section, IQM2 measure will be tested using different number of orientations and have concluded that the best correlation with MOS was obtained with 2 orientations ($K = 2$). Best values for each database are put in bold. Spearman's correlation coefficient was calculated for each subjective database separately, as well as their mean and weighted mean (Table 2). As SSIM and MS-SSIM use 11 × 11 size of Gaussian filter, this block size was firstly tested, to be able to compare with these measures. Here, we also put results for Spearman's correlation of SSIM, SSIMmod, and MSSIM measures. Later, other filter sizes will be tested to determine optimal size.

When comparing correlation across multiple databases, it can be calculated as an arithmetic mean correlation or weighted arithmetic mean as proposed in [13]. Weighted arithmetic mean correlation is calculated as:

$$wt\_mean = \frac{\sum_{i=1}^{7}(w_i \cdot corr_i)}{\sum_{i=1}^{7} w_i}$$
$$w_i = \{54; 866; 779; 185; 552; 1{,}700; 168\} \qquad (8)$$

In (8), w_i are database sizes (A57, CSIQ, LIVE, IVC, VCL@FER, TID and TOYAMA accordingly). Weighted arithmetic mean is calculated because larger databases should have higher influence on final correlation results, because MOS results obtained using higher number of degraded images should better describe HVS. Larger database usually means higher number of distorted images with more observers; CSIQ database is exception with 866 degraded images and 35 observers (which probably graded images in more sessions).

From the results in Table 2, it can be concluded that the best correlation depends on number of kernel orientations, for same image database. However, best approximation of the HVS gave kernel with only one (LIVE, IVC, VCL@FER, and TOYAMA image databases) or two orientations (A57, CSIQ, and TID). Higher number of orientations (4 and 6) gave in general somewhat lower results (4 orientation kernel gives nearly equal results, however, it is slower) which means that HVS can be best approximated (in IQM2 measure) using only one or two orientations. When calculation time and weighted mean correlation was taken into account, we concluded that optimum number of orientations K was 2. We have chosen two orientations for later comparison because it gave better correlation in databases with higher number of degradation types than one orientation kernel, while being only slightly slower.

Table 3 Weighted average mean of Spearman's correlation for IQM2 with different number of orientations K and $S \times S$ Gaussian filter

	$K = 1$	$K = 2$	$K = 4$	$K = 6$
$S = 3 \times 3$	0.9074	0.9121	0.9065	0.8818
$S = 5 \times 5$	0.9079	**0.9128**	0.9067	0.8822
$S = 7 \times 7$	0.9075	0.9122	0.9050	0.8799
$S = 9 \times 9$	0.9065	0.9109	0.9031	0.8770
$S = 11 \times 11$	0.9057	0.9098	0.9012	0.8750

In Table 3, weighted average mean, for all image databases (according to (8)), was calculated for different number of orientations and different size of Gaussian filter S to determine optimal filter block size.

For Gaussian filter size 5 × 5 pixels, separate results for every image database are presented in Table 4. Here, calculation times for all orientations are given. Computer configuration which was used for calculating was Intel Q6600 @2400 MHz, 4 GB RAM, Windows Vista 64 with Matlab program. Mean time was calculated for all degraded images in TID database (1,700 images with resolution 512 × 384 pixels). Converting to gray scale was not taken in calculation time.

5.2 IQM2 robustness to rotational transformation

We have also tested proposed IQM2 measure to the rotational transformation to check which orientation gives best results. Tested database was TID database and angle of rotation was set to 0–1° in step of 0.1°. Rotation was done on degraded images and rotated image was calculated using bicubic interpolation. Because size of the image after rotation will not be the same as in the original image (black borders will appear), we cropped maximal area out of the rotated image, which does not have the black border and compared such cropped and rotated image with its original cropped image (without rotation). An algorithm for finding highest correlation with MOS results was used, where original cropped image was determined. Results are presented in Fig. 1. From

Table 4 Spearman's correlation for IQM2 with different number of orientations K and $S = 5 \times 5$ pixels Gaussian filter

	$K = 1$	$K = 2$	$K = 4$	$K = 6$
A57	0.83815	**0.83964**	0.83235	0.82481
CSIQ	0.92014	**0.93766**	0.92961	0.89086
LIVE	**0.95299**	0.95064	0.9503	0.94848
IVC	**0.89358**	0.88169	0.8797	0.88115
VCL@FER	**0.93822**	0.93497	0.93392	0.93708
TID	0.87645	**0.88547**	0.87545	0.83237
TOYAMA	**0.8948**	0.87288	0.86645	0.87402
MEAN	**0.90205**	0.90042	0.8954	0.88411
WT_MEAN	0.90799	**0.91289**	0.90668	0.88221
Calculation time (ms)	**113.9**	188.2	338.4	460.9

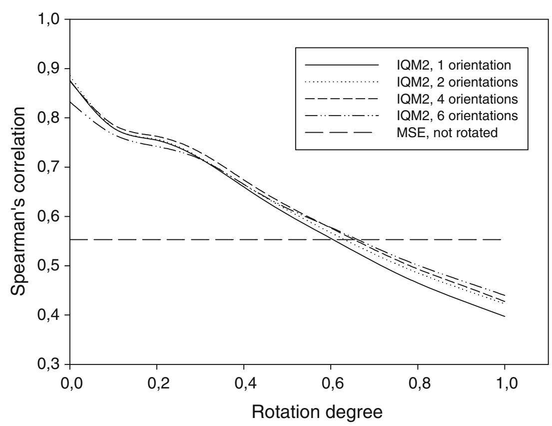

Fig. 1 Spearman's correlation with MOS for TID database and IQM2 measure with different orientation kernels, Gaussian filter size 5 × 5 pixels

the figure, it can be concluded that IQM2 measure is highly rotation dependent, because after rotation of 0.6° its correlation with MOS becomes lower than correlation of MSE measure (with not rotated images). For lower angles, IQM2 measure gives best correlation with 4 orientation kernels and for higher angles IQM2 with 6 orientation kernels. However, results are pretty similar and correlation drops considerably for angle degrees higher than 0.6°, so it would be useful to use some preprocessing technique to align images before using IQM2 measure, if it is known that images are not aligned.

5.3 Pearson's and Spearman's correlation, statistical significance

A total of 14 full-reference objective quality measures (including newly presented IQM2 with 2 orientations) will be compared using 7 existing subjective image quality databases described in Section IV. They will be compared using Pearson's and Spearman's correlation and statistical significance of the correlation results will be checked using F test and Ansari–Bradley test. Pearson's correlation, after nonlinear regression, from the best image quality measure is mentioned in Table 5, for 5-parameter fitting function. Measures in italics have statistically insignificant variance difference with IQM2 measure according to the F test, while underlined measures have statistically insignificant dispersion with IQM2 measure according to the Ansari–Bradley test. Maximum normalized median difference (median difference divided by MOS range) was 0.0492 (in TOYAMA database), so we subtracted medians from the tested pairs of measures. IQM2 measure is put in bold to emphasize its place.

From the results of the F test, it can be seen that IQM2 measure gives best (and statistically significant) results in TID image database. Also in VCL@FER database, it has statistically insignificant variance with NQM measure. IQM2 measure has statistically lower correlation in A57 database only from VSNR measure. However, VSNR parameters are based on this database with only 54 degraded images so this result is of limited statistical reliability. In CSIQ, IVC, LIVE, and TOYAMA databases, some measures have statistically better correlation than IQM2 measure according to the F test.

When comparing results with Ansari–Bradley test, IQM2 measure gives best (and statistically significant) results in TID image database and statistically insignificant variance when comparing with best measure in VCL@FER, CSIQ, and IVC image databases. IQM2 measure is second in A57, which is also of limited statistical reliability. In LIVE and TOYAMA databases, some measures have statistically better correlation than IQM2 measure according to the Ansari–Bradley test. These results would be also the same with medians not subtracted from tested pairs of databases, when comparing IQM2 measure with all other objective measures.

It can be concluded that F test gives good results, although some of the measures did not have Gaussian distribution. Also, IQM2 has normal distribution only in A57 and TOYAMA image databases, so it should be compared using Ansari–Bradley test.

Main difference between F test and Ansari–Bradley test is because in F test, equal or less objective quality measures have statistically insignificant variance difference, when comparing with Ansari–Bradley test. In [39], it is said that relative efficiency of the Ansari–Bradley test, when comparing with F test, is about 0.608, when samples have normal distribution.

Pearson's correlation is shown in Tables 6 and 7. It can be seen that fitting function has influence on final Pearson's correlation results. Five-parameter fitting function gave usually higher correlation than 4-parameter, however, results are nearly similar. Spearman's correlation for all tested objective

Table 5 Pearson's correlation (5-parameter fitting function) and statistical significance, from the best image quality measure (from left)

A57	VSNR	*MAD*	*IWSSIM*	*IWPSNR*	***IQM2***	*MSSIM*	*NQM*	VIFP	SSIM	SSIMmod	VIF	MSE	NAE	CWSSIM
CSIQ	MAD	***IQM2***	*VIF*	SSIMmod	IWSSIM	VIFP	MSSIM	SSIM	CWSSIM	IWPSNR	MSE	VSNR	NAE	NQM
LIVE	MAD	VIF	VIFP	*IWSSIM*	*MSSIM*	***IQM2***	*SSIM*	*SSIMmod*	IWPSNR	VSNR	NQM	CWSSIM	MSE	NAE
IVC	IWSSIM	MAD	*SSIMmod*	*SSIM*	*MSSIM*	*IWPSNR*	*VIF*	***IQM2***	CWSSIM	NQM	VIFP	VSNR	MSE	NAE
VCL@FER	*NQM*	***IQM2***	MSSIM	IWPSNR	IWSSIM	SSIM	SSIMmod	MAD	VIF	VIFP	VSNR	CWSSIM	MSE	NAE
TID	**IQM2**	IWSSIM	MSSIM	MAD	SSIMmod	VIF	SSIM	VIFP	CWSSIM	VSNR	IWPSNR	NQM	MSE	NAE
TOYAMA	MAD	IWSSIM	VIF	*MSSIM*	*NQM*	*SSIM*	*SSIMmod*	***IQM2***	*VSNR*	*IWPSNR*	*VIFP*	CWSSIM	MSE	NAE

Table 6 Pearson's correlation for all objective quality measures, using 5-parameter logistic function

	CWSSIM	IQM2	IWPSNR	IWSSIM	MAD	MSE	MSSIM	NAE	NQM	SSIM	SSIMMOD	VIF	VIFP	VSNR
A57	0.38298	0.88977	0.8975	0.90353	0.91079	0.69324	0.86039	0.68135	0.82697	0.80188	0.80109	0.69899	0.8025	**0.95021**
CSIQ	0.85085	0.92996	0.82188	0.91441	**0.95067**	0.81536	0.89974	0.75362	0.74354	0.86126	0.91525	0.92775	0.90438	0.80053
LIVE	0.9041	0.94622	0.93293	0.95219	**0.96752**	0.87305	0.94894	0.83629	0.91294	0.94488	0.9447	0.95983	0.95964	0.92287
IVC	0.86159	0.89509	0.90548	**0.92306**	0.92195	0.72145	0.91085	0.63137	0.8498	0.91194	0.91259	0.90283	0.82294	0.80342
VCL@FER	0.83686	0.93474	0.92121	0.91909	0.90531	0.8241	0.92319	0.8061	**0.94293**	0.91436	0.9133	0.89548	0.89205	0.87835
TID	0.73143	**0.88657**	0.66636	0.85791	0.83083	0.58495	0.84515	0.40856	0.61353	0.77317	0.80976	0.80934	0.76825	0.68175
TOYAMA	0.81829	0.87667	0.85018	0.92488	**0.94068**	0.64918	0.89274	0.56154	0.88929	0.8887	0.88869	0.91629	0.84725	0.8705
MEAN	0.76944	0.90843	0.85651	0.91358	**0.91825**	0.73733	0.89728	0.6684	0.82557	0.87089	0.88362	0.87293	0.85672	0.84395
WT_MEAN	0.80485	**0.9123**	0.79894	0.90017	0.89844	0.72386	0.8898	0.62536	0.75972	0.85092	0.87608	0.87826	0.85202	0.79047

Table 7 Pearson's correlation for all objective quality measures, using 4-parameter logistic function

	CWSSIM	IQM2	IWPSNR	IWSSIM	MAD	MSE	MSSIM	NAE	NQM	SSIM	SSIMMOD	VIF	VIFP	VSNR
A57	0.37494	0.8885	0.89575	0.90244	0.90591	0.66947	0.85734	0.67583	0.81605	0.80185	0.801	0.61604	0.80235	**0.95017**
CSIQ	0.85073	0.92632	0.82188	0.90253	**0.95019**	0.80299	0.89718	0.74865	0.74221	0.85938	0.91236	0.92526	0.9043	0.80052
LIVE	0.90318	0.93879	0.9329	0.9425	**0.96718**	0.85823	0.94021	0.82974	0.91283	0.93835	0.9383	0.95924	0.95935	0.92276
IVC	0.86104	0.89486	0.90548	**0.92285**	0.92096	0.72065	0.91067	0.62355	0.84888	0.91165	0.91232	0.90262	0.82292	0.80274
VCL@FER	0.83674	0.92993	0.92113	0.91526	0.90506	0.81091	0.91831	0.80399	**0.94282**	0.90892	0.90756	0.89234	0.89205	0.87746
TID	0.72259	**0.87868**	0.66612	0.84882	0.83057	0.56892	0.84044	0.37598	0.60963	0.77153	0.80817	0.80505	0.74813	0.68175
TOYAMA	0.81809	0.87635	0.85008	0.9244	**0.94062**	0.62642	0.89201	0.53736	0.88928	0.88771	0.8877	0.91367	0.84725	0.87048
MEAN	0.76676	0.90478	0.85619	0.9084	**0.91721**	0.72251	0.89374	0.65644	0.8231	0.86849	0.88106	0.85918	0.85376	0.8437
WT_MEAN	0.80101	**0.90645**	0.7988	0.89191	0.89804	0.70945	0.88514	0.60869	0.75771	0.84796	0.87292	0.8744	0.844	0.7903

Table 8 Spearman's Correlation for all objective quality measures

	CWSSIM	IQM2	IWPSNR	IWSSIM	MAD	MSE	MSSIM	NAE	NQM	SSIM	SSIMmod	VIF	VIFP	VSNR
A57	0.33148	0.83964	0.87619	0.87127	0.90139	0.61763	0.8415	0.56484	0.79778	0.80666	0.80662	0.62228	0.76854	**0.93588**
CSIQ	0.8412	0.93766	0.83106	0.92129	**0.94665**	0.8058	0.91364	0.75965	0.74116	0.87563	0.92839	0.91945	0.88068	0.81095
LIVE	0.9025	0.95064	0.93278	0.95665	**0.96689**	0.87556	0.95128	0.8367	0.9093	0.9479	0.94783	0.96315	0.96179	0.92713
IVC	0.85791	0.88169	0.89976	0.9125	**0.91457**	0.68844	0.898	0.60877	0.83431	0.90182	0.90285	0.89637	0.81091	0.79927
VCL@FER	0.83477	0.93497	0.9166	0.91633	0.90607	0.82465	0.92269	0.79464	**0.94359**	0.91125	0.91001	0.88665	0.89185	0.87261
TID	0.65859	**0.88547**	0.68234	0.85594	0.83401	0.5531	0.85418	0.32525	0.62359	0.77493	0.81771	0.74907	0.65389	0.70488
TOYAMA	0.81531	0.87288	0.8475	0.92024	**0.93617**	0.61319	0.88738	0.51649	0.8871	0.87938	0.87943	0.90767	0.84789	0.86082
mean	0.74882	0.90042	0.85518	0.90775	**0.91511**	0.7112	0.89552	0.62948	0.81955	0.87108	0.88469	0.84923	0.83079	0.84451
wt_mean	0.77266	**0.91289**	0.80586	0.9002	0.89826	0.70611	0.89553	0.58808	0.76153	0.85391	0.8813	0.85068	0.80153	0.801

measures and all databases are presented in Table 8. Weighted mean (wt_mean) was calculated according to the (8). Best values in Tables 6, 7, and 8 are put in bold.

Table 9 gives calculation times for all objective measures. Computer configuration which was used for calculating all objective measures: Intel Q6600 @2400 MHz, 4 GB RAM,

Table 9 Calculation time for all objective measures

	CWSSIM	IQM2	IWPSNR	IWSSIM	MAD	MSE	MSSIM	NAE	NQM	SSIM	SSIMmod	VIF	VIFP	VSNR
t(ms)	2,887	188.2	657.1	657.6	47,640	4.5	132.7	4.3	373.6	25.7	25.1	1321	166.2	55.8

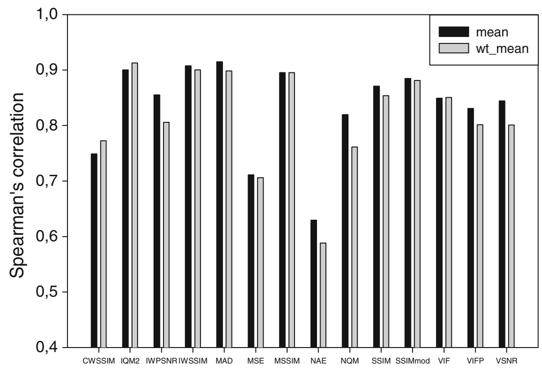

Fig. 2 Mean and weighted mean Spearman's correlation for all databases

Windows Vista 64 with Matlab program. Mean time was calculated for all degraded images in TID database (1,700 images with resolution 512×384 pixels). Converting to gray scale and scaling images if needed were not taken in calculation time. MEX files were used where possible (for SPWT transform and MAD measure).

Mean and weighted mean Spearman's correlation is shown on Fig. 2 and is calculated according to (8). When comparing correlation in each database separately, Tables 6, 7, 8, MAD measure performs best among all tested measures, because it has the highest correlation in 3 different databases using Pearson's correlation and 4 databases using Spearman's correlation. When comparing Pearson's and Spearman's correlation and arithmetic mean, correlation is the best for MAD, IWSSIM, and then IQM2 measure. When database sizes are taken into account and weighted mean is calculated, correlation is the best for IQM2 for all tested correlations. Second one is MAD measure for Pearson's correlation with 4-parameter fitting function (which its authors propose in [10]) and third one is IWSSIM for mentioned correlations. IWSSIM measure is second one for Pearson's correlation with 5-parameter fitting function (which its authors propose in [13]) and Spearman's correlation and MAD measure is third one for these correlations. MSSIM was fourth best measure according to all correlation types and mean values (average and weighted).

5.4 Discussion of the results

When comparing correlation in each database separately, Tables 6, 7, 8, MAD measure performs best among all tested measures, because it has the highest correlation in 3 different databases using Pearson's correlation and 4 databases using Spearman's correlation. VSNR, IQM2, and NQM have best correlation in only one database, each of them, while IWSSIM has best correlation in one database using only Pearson's correlation.

From the correlation results, it can be seen that IQM2 measure has statistically significant lower variance in TID database which has 17 different degradation types, so it should be used in imaging systems where different degradation types can occur. Different degradation types have different impact on IQM2 correlation, however, in databases with more degradation types IQM2 performs well. In TOYAMA database (with JPEG and JPEG-2000 degradation types only), MAD, IWSSIM, and VIF measure gave better results than IQM2 measure, so these measures should be used rather than IQM2 measure, if computational time is not taken into account.

Calculation time (complexity) is also lowest for IQM2 measure in comparison with IWSSIM (about 3.5 times faster) and MAD measure (about 250 times faster). However, it is probably possible to speed up these competitive measures also. For even faster calculation time, IQM2 with 1 orientation kernel can also be used, with nearly similar correlation results, while being about 6 times faster than IWSSIM and 420 times faster than MAD measures.

It can be also seen that nearly all measures (except VSNR and NQM) have best correlation results in LIVE database (IQM2, MSSIM, IWSSIM higher than 0.95 and VIF and MAD higher than 0.96), probably because in LIVE database DMOS results were at the end realigned for each of the 7 independent sessions (each session had only 1 degradation type) of subjective evaluation. Realignment in fact changes subjective grades to be more similar in different types of degradation. However, this was made after initial subjective evaluations, which may affect HVS modeling. If one session of evaluation contains only 1 degradation type, it can be expected that HVS will, at least toward the end of session evaluation, be more adjusted to the specific distortion type (same conclusion was drawn in [45]). In our database, VCL@FER, observers graded all types of distortion in one session, so this type of realignment was not needed. TID database also used all degradation types in 1 session. Also, we calculated MOS results, not DMOS in VCL@FER database. This may explain generally lower correlation in our database than in LIVE, for nearly all objective measures.

When degraded images were firstly rotated in TID database, IQM2 measure showed that it is highly rotational dependent and its correlation with MOS drops significantly for angles higher than 0.6°, so some preprocessing technique should be firstly used, if it is known that images are not aligned. CW-SSIM could be then also used instead, because it showed rotational invariance for even higher degrees of rotation, as well as translation [12].

When comparing statistical significance among best measures (MAD, IQM2, IWSSIM):

- Comparing with MAD measure, IQM2 is (according to Ansari–Bradley test, Table 5) statistically better in 2 databases (including TID database), statistically the same in 3 databases (however, it has lower absolute correlation in these databases) and statistically worse in 2 databases
- Comparing with IWSSIM measure, IQM2 is (according to Ansari–Bradley test, Table 5) statistically better in 3 databases (including TID database), statistically the same in 3 databases (however, it has lower absolute correlation in these databases), and statistically worse in 1 database

Generally, it can be concluded that, when constructing new image quality measure, much attention should be given to the tested and fitted database. Because of that, we used seven different databases to check whether measure performs well on each of them. Our previous measure (IQM) performed well on LIVE and VCL@FER databases. However, we optimized coefficients according to the databases and not HVS. Other example which can be also seen is VIF measure that performs very well in LIVE (on which it was firstly tested, [9]) and TOYAMA databases, while having lower correlation results in A57 and TID databases (where its correlation with MOS drops considerably). MAD (multiple-based strategy metric) and IWSSIM (NSS model) measures also have somewhat lower correlation in TID database, while having best or nearly best correlation in some other databases. MSSIM, which is structural distortion based metric, shows equal correlation with IWSSIM and even better correlation than MAD in TID database.

Attention should be given to the optimized coefficients of the metrics, because it is apriori unknown how much these coefficients are influenced by, HVS or database, or even specific degradation type. All constructed metric should be tested on different databases and afterward conclusion can be made about how well they fit HVS. More free (or optimized) parameters in measure construction do not necessarily mean higher correlation with MOS. For example, our measure IQM2 could have somewhat higher correlation in TID database by using exponents after calculating SSIM_cont, (3), and SSIM_struct, (4), at each scale, e.g., to have some scales with higher influence on final measure (optimization could be made using, e.g., PSO). However, this would lead to the lower correlation in other. IQM2 was firstly tested on TID database because of the highest number of degradation types in this database. However, we tried not to introduce any overfitting parameter but rather to use methods which already shown good properties in image processing: SPWT which is translation and rotation invariant multistage decomposition and with aliasing eliminated between subbands; combined with modified SSIM measure which also shows good properties with HVS in its basic implementation. Because IQM2 shows good properties in all tested image databases, without having specific fitting parameter calculated from specific database, it can be expected to show good correlation with MOS on some new database, not tested yet. Results from VCL@FER database (which was developed for this purpose) also confirm this conclusion, having better correlation than IWSSIM and MAD measure.

6 Conclusion

In this paper, we proposed novel image quality measure—IQM2. Original and degraded images are firstly decomposed using SPWT with different number of orientations and maximal number of scales. At every stage of the image decomposition, SSIM contrast and structure has been calculated, for pass band coefficients only. IQM2 measure was calculated as the combination of SSIM contrast and structure values. Results show that the best correlation was obtained using 2 orientations (1 also gives nearly similar results).

When comparing correlation in each database separately, Tables 6, 7, 8, MAD measure performs best among all tested measures, because it has the highest correlation in 3 different databases using Pearson's correlation and 4 databases using Spearman's correlation. In comparison with other best performing measures with similar correlation results (IWSSIM, MAD), IQM2 has lower computational time (less complexity) so it should be used instead of them, when calculation time is important. Particularly, it is 250 times faster than best overall performing measure—MAD.

If we consider overall correlation results, we can conclude that IQM2 works very well for different image contents and degradation types, having the best weighted average correlation (both Pearson's and Spearman's correlations, Tables 6, 7, 8). IQM2 measure also gives the best (and statistically significant) results in TID image database, probably because it does not have specific fitting parameters calculated from specific database. It is important to mention that TID database has 17 different degradation types, so it should be used in imaging systems where different degradation types can occur.

Future research will include optimization of the filter kernel for possibly better correlation. Next step will be the work on development of the reduced-reference objective quality measure which could be used in real applications with limited bandwidth (e.g., where it is impossible to use reference image).

References

1. Ponomarenko, N.N., Lukin, V.V., Egiazarian, K.O., Lepisto, L.: Adaptive visually lossless JPEG-based color image compression. Signal Image Video Process. **7**(3), 437–452 (2013)
2. Veerakumar, T., Esakkirajan, S., Vennila, Ila: Recursive cubic spline interpolation filter approach for the removal of high density salt-and-pepper noise. Signal Image Video Process. **8**(1), 159–168 (2014)
3. Sahraee, M.J., Ghofrani, S.: A robust blind watermarking method using quantization of distance between wavelet coefficients. Signal Image Video Process. **7**(4), 799–807 (2013)
4. ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union/ITU Radiocommunication Sector (2002)
5. 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/ (2008)
6. Grgic, S., Grgic, M., Mrak, M.: Reliability of objective picture quality measures. J. Electr. Eng. **55**(1–2), 3–10 (2004)
7. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
8. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: 37th Proc. IEEE Asilomar Conf. on Signals, Systems and Computers, vol. 2, pp. 1398–1402 (2003)
9. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
10. 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)
11. Damera-Venkata, N., Kite, T., Geisler, W., Evans, B., Bovik, A.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. **9**(4), 636–650 (2000)
12. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: a new image similarity index. IEEE Trans. Image Process. **18**(11), 2385–2401 (2009)
13. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011)
14. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron. Imaging **19**(1), 1–22 (2010). Article ID 011016
15. Soundararajan, R., Bovik, A.C.: Survey of information theory in visual quality assessment. Signal Image Video Process. **7**(3), 391–401 (2013)
16. Wang, Z., Bovik, A.C.: Universal image quality index. IEEE SP Lett. **9**, 81–84 (2002)
17. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Process. **14**(12) (2005)
18. Damera-Venkata, N., Kite, T., Geisler, W., Evans, B., Bovik, A.: Image quality assessment based on a degradation Model. IEEE Trans. Image Process. **9**(4) (2009)
19. Miyahara, M., Kotani, K.: Objective picture quality scale (PQS) for image coding. IEEE Trans. Commun. **46**(1) (1998)
20. Visual Quality Assessment Package Version 1.1, http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
21. http://dailyburrito.com/projects/cssim_index_multi.m
22. https://ece.uwaterloo.ca/~z70wang/research/iwssim/
23. http://vision.okstate.edu/mad/
24. He, L., Gao, X., Lu, W., Li, X.: Image quality assessment based on S-CIELAB model. Signal Image Video Process. **5**(3), 283–290 (2011)
25. Dumic, E., Grgic, S., Grgic, M.: New image-quality measure based on wavelets. J. Electron. Imaging **19**(1) (2010). Article ID 011018.
26. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proc. IEEE Int'l. Conf. on Neural Networks, IV, pp. 1942–1948 (1995)
27. Dumic, E., Grgic, S., Grgic, M.: Improved image quality measure through particle swarm optimization. In: 18th International Conference on Systems, Signals and Image Processing IWSSIP-2011, pp. 167–172 (2011)
28. Dumic, E., Bacic, I., Grgic, S.: Simplified structural similarity measure for image quality evaluation. In: 19th International Conference on Systems, Signals and Image Processing IWSSIP-2012, pp. 456–461 (2012)
29. Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: a flexible architecture for multi-scale derivative computation. In: 2nd IEEE International Conference on Image Processing vol. 3, pp. 444–447 (2012)
30. Karasaridis, A., Simoncelli, E.P.: A filter design technique for steerable pyramid image transforms. Proc. ICASSP **4**, 2387–2390 (1996)
31. Matlab Pyramid Tools, http://www.cns.nyu.edu/~eero/STEERPYR/
32. Rezazadeh, S., Coulombe, S.: A novel discrete wavelet transform framework for full reference image quality assessment. Signal Image Video Process. **7**(3), 559–573 (2013)
33. Rouse, D.M., Hemami, S.S.: Understanding and simplifying the structural similarity metric. In: IEEE Int. Conf. on Image Process. (ICIP), pp. 1188–1191 (2008)
34. Feilner, M., Van De Ville, D., Unser, M.: An orthogonal family of Quincunx wavelets with continuously adjustable order. IEEE Trans. Image Process. **14**(4), 499–510 (2005)
35. Hauke, J., Kossowski, T.: Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. In: Proceedings of the MAT TRIAD 2007 Conference, Bedlewo, Poland (2007)
36. Sheikh, H.R.: Image Quality Assessment Using Natural Scene Statistics, Ph.D. Dissertation, University of Texas at Austin (2004)
37. Jorge, Nocedal, Stephen, Wright: Numerical Optimization. Springer, New York (1999)
38. NIST/SEMATECH "e-Handbook of Statistical Methods", http://www.itl.nist.gov/div898/handbook/
39. Ansari, A.R., Bradley, R.A.: Rank-sum tests for dispersions. Ann. Math. Stat. **31**(4), 1174–1189 (1960)
40. http://www.vcl.fer.hr/quality/
41. http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
42. http://vision.okstate.edu/index.php?loc=csiq
43. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: LIVE Image Quality Assessment Database Release 2. http://live.ece.utexas.edu/research/quality/subjective.htm
44. Le Callet, P., Autrusseau, F.: Subjective quality assessment IRCCyN/IVC database, http://www.irccyn.ec-nantes.fr/ivcdb/
45. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008—a database for evaluation of full-reference visual quality assessment metrics. Adv. Modern Radioelectron. **10**, 30–45. http://www.ponomarenko.info/tid2008.htm/
46. MICT (Media Information and Communication Laboratory) Image Quality Evaluation Database: http://mict.eng.u-toyama.ac.jp/mictdb.html