Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 53(4), 344–354(2012)

Anđela Zarić, Nenad Tatalović, Nikolina Brajković, Hrvoje Hlevnjak, Matej Lončarić, Emil Dumić, Sonja Grgić

VCL@FER Image Quality Assessment Database

DOI 10.7305/automatika.53-4.241
UDK 004.65:004.932.2.05
IFAC 2.8; 1.1.8

Original scientific paper

In this paper we present new image quality database VCL@FER (http://www.vcl.fer.hr/quality/) which consists of four degradation types, 6 levels of each degradation and 23 different images (552 degraded images). It can be used in objective image quality evaluation, as well as to develop and test new image quality measures. Results for six commonly used full reference objective quality measures are compared using newly developed image database, as well as 6 other image databases.

Key words: Image database, Objective image quality measures, VCL@FER

VCL@FER – baza slika za procjenu kvalitete slike. VCL@FER baza slika nova je baza slika (http://www.vcl.fer.hr/quality/) koja se sastoji od 4 vrste izobličenja, 6 razina svakog izobličenja i 23 različite slike (ukupno 552 izobličene slike). Baza slika može se koristiti za usporedbu različitih objektivnih mjera kvalitete slike, kao i za razvoj novih objektivnih mjera. Uporabom nove baze te još šest dostupnih baza slika provedena je usporedba šest relevantnih objektivnih mjere kvalitete slike.

Ključne riječi: Baze slika, objektivne mjere kvalitete slike, VCL@FER

1 INTRODUCTION

Subjective image quality assessment (IQA) is based on subjective experiments in which image quality has been evaluated by human observers perceiving and ranking images [1]. The results of such experiments depend on psychological processes of perception. Though reliable, because it depends on psychovisual perception of the each individual that is assessing image quality, subjective method is expensive, difficult to design and time consuming to compute. Several critical factors of the human observers can influence on the final results of assessment such as environmental conditions, motivation and mood of the observers. On the other hand, subjective IQA allows a better understanding of the mechanisms underlying quality perception, providing useful information for the subsequent modeling phase.

Undoubtedly, there is a need for objective measures of image quality that correlate well with the results of subjective assessments. Objective IQA as mathematically defined measures, are more attractive because they are independent of viewing conditions, individual observers and usually have low computational complexity. Because of that objective IQA measure can be calculated easier and they measure the image quality automatically. The evaluation results should be statistically consistent with those of the human observers.

Assessment of image quality is an open problem today. In order to allow easier and less expensive testing of objective IQA algorithms and to define benchmarks, there exists the big necessity for a publicly available image quality assessment database that contain results of subjective experiments. In that way new objective IQA algorithms can be presented together with a standard and reliable validation.

Some of the publicly available image quality assessment databases are: A57 database [2], CSIQ database [3], LIVE database [4], IVC database [5], TID2008 database [6] and Toyama-MICT database [7]. All of them have various numbers of reference images, distorted images and distortion types, different number of human observers, and the type of images, Table 1.

Table 1. Characteristics of tested databases

Database	Source Images	Distorted Images	Distortion Types	Image Type	Observers
TID2008	25	1700	17	color	838
CSIQ	30	866	6	color	35
LIVE	29	779	5	color	161
IVC	10	185	4	color	15
Toyama-MICT	14	168	2	color	16
A57	3	54	6	gray	7

The goal of our study was to create image quality

assessment database that is based on subjective opinion of human observers, subsequent the work from the best known and most frequently used image quality databases, LIVE database [4]. Subjective evaluation has been conducted in accordance with Recommendation ITU-R BT.500-11 [8].

This paper is organized as follows. Section 2 describes new image database, its subjective quality assessment methodology, as well as six other publicly available image databases. Section 3 describes existing objective measures. Section 4 explains performance measures used in comparing objective measures. Section 5 presents results and finally section 6 gives the conclusion.

2 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 correlation with objective measures:

- VCL@FER (Video Communications Laboratory @ FER) [9],
- A57 (A57 database) [2],
- CSIQ (Categorical Image Quality Database) [3],
- LIVE (Laboratory for Image & Video Engineering) [4],
- IVC (Image and video-communication) [5],
- TID (Tampere Image Database 2008) [6],
- Toyama [7].

2.1 VCL@FER

In [9] we proposed new VCL@FER database. VCL@FER image database consists of 575 images. 23 of those images are original images, without any degradation. Each image has gone through 4 different types of degradation, and each type of degradation has been divided into 6 quality levels. First level of degradation represents mildest degradation, while sixth represents the most severe degradation.

The four types of degradation present in the database are: average white Gaussian noise (AWGN), Gaussian blur, JPEG2000 and JPEG.

AWGN degradation was calculated as a sum of original image and normally distributed pseudorandom numbers with 6 different standard deviations (which represents 6 degradation levels). It was calculated in Matlab.

Gaussian blur is calculated as filtering of an image with Gaussian function with different size ($n_1 \cdot n_2$ pixels). It can be described as (without normalization):

$$h(n_1, n_2) = e^{-\frac{n_1^2+n_2^2}{2\cdot\sigma^2}}. \quad (1)$$

6 different sizes of Gaussian function were used to calculate blur degradation, using Irfanview software [10].

JPEG degradation was performed using 6 different qualities (in the range 0-100), using Matlab. JPEG2000 degradation was performed so that the final size was 4, 1.5, 0.5, 0.25, 0.125 and 0.0625 bits per pixel (8 is without compression), using "kdu_compress" [11].

Subjective testing was done on a study group of 118 people, non-experts, between 20 and 30 years old. Each subject had about 96 images to grade. The graded images were not grouped by the types of degradation. Subjects did not know which type of degradation to expect. Each image was rated between 16 and 36 times. Method used in experiment was Single-stimulus (SS) method, which uses numeric criteria scale with 100 grades.

Test was done in a room without natural light, with electric illumination. Each monitor was pre-calibrated for such lightning. The length of test was around 19 minutes per observer. Software used for testing and grading image quality was developed for the purposes of the project.

All testing results and grades have been collected, with grades for each picture being averaged. According to ITU-R BT.500-11 [8] results of every observer should be compared with all others to see if they differ too much from the average value and discard them if they do. For SSQCE (Single Stimulus Continuous Quality Evaluation) two steps are required for screening of the observers. In our test configuration we had one test condition, one repetition and one time window within a test combination of test condition and sequence, so second step could be discarded. In the first step, we had 118 observers and 575 test images, while other parameters of the test were constants. Screening of the observers can be described as follows. For each time window (8s per image, reference or distorted) firstly was determined if distribution of the results were normal or not by using kurtosis β. β is defined as fourth central moment of the variable, which in our case can be described as:

$$\beta = \frac{m_4}{\sigma^4} = \frac{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}. \quad (2)$$

Process can be mathematically expressed as:

$$\begin{array}{l}\forall l \in L \text{ where } L = 575 \text{ stands for number of images} \\ \forall n \in N \text{ where } N = 118 \text{ stands for number of observers} \\ \left.\begin{array}{l}\text{if } u_{nl} \geq \bar{u}_l + 2\sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } u_{nl} \leq \bar{u}_l - 2\sigma_l \text{ then } Q_n = Q_n + 1\end{array}\right\} \text{for } 2 \leq \beta \leq 4 \\ \left.\begin{array}{l}\text{if } u_{nl} \geq \bar{u}_l + \sqrt{20}\sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } u_{nl} \leq \bar{u}_l - \sqrt{20}\sigma_l \text{ then } Q_n = Q_n + 1\end{array}\right\} \text{for } \beta \notin [2, 4].\end{array} \quad (3)$$

At the end, P and Q values were determined for every observer and if any of the values were greater than 2% of the number of tested images (575), observer was discarded:

$$\begin{array}{l}\frac{P_n}{L} > 2\% \text{ or } \frac{Q_n}{L} > 2\% \\ P_n \geq 12 \text{ or } \mathrm{Q}_n \geq 12 \rightarrow \text{ discard observer } n\end{array} . \quad (4)$$

Using (4), 2 observers were discarded. Recommendation ITU-R BT.500-11 proposes 0.2%, but in this case 55 observers would be discarded. If this ratio would be set to 1%, 17 observers would be discarded. However, correlation results between objective and subjective measures would be generally lower.

Afterwards, results for every observer were rescaled to the full (and same) range of 0-100, according to the:

$$\begin{array}{l}\forall n \in \{1, 116\} \\ \mathrm{MOS}_{n,l} = \frac{100}{\max(r_n) - \min(r_n)} \cdot (r_{n,l} - \min(r_n))\end{array} . \quad (5)$$

In (5) $r_{n,l}$ represents grade that the n-th viewer has given for l-th image (including reference images), $\mathrm{MOS}_{n,l}$ represents rescaled grades of the same viewer and r_n represents all grades of n-th subject. At the end, average MOS (Mean Opinion Score) grade was calculated for each of the distorted image as an arithmetic mean of all grades for each image.

2.2 A57 database

A57 database [2] consists of 6 types of degradation: quantization of the LH subbands of a 5-level DWT of the image using the 9/7 filters, additive Gaussian white noise, baseline JPEG compression of the image (using the standard quantization matrix), JPEG-2000 compression of the image (using the 9/7 filters and no visual frequency weighting), JPEG-2000 compression (using the 9/7 filters) with the dynamic contrast-based quantization algorithm, blurring by using a Gaussian filter.

Three natural images were used, obtained from Kodak database [12] with 6 described types of degradations and 3 degradation levels. Due to the limited number of images and limited number of human subjects, the A57 database is of limited statistical reliability. Results are presented as a MOS measure.

2.3 CSIQ database

CSIQ database [3] consists of 6 different types of degradation: JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink noise, Gaussian noise and Gaussian blurring. 30 original images were distorted with 5 degradation levels for all mentioned degradation types except contrast decrement degradation which had 3-4 degradation levels. 35 observers graded 866 degraded images in over 5000 subjective ratings and are reported in the form of DMOS (Difference MOS). All images were displayed on LCD monitors with resolution 1920x1080 pixels. The observers had to put in the row all degradation levels of the same image simultaneously. Horizontal distance between images showed similarity between them, e.g. smaller distance means that two images had more similar grades. At the end, DMOS results were calculated using difference scores between degraded and original subjective score (for each observer).

2.4 LIVE database

LIVE database [4] consists of 5 types of degradation: JPEG compressed images, JPEG-2000 compressed images, Gaussian blur, white noise and bit errors in JPEG2000 bit stream. 29 original images (out of which 24 were obtained from Kodak database [12]) were degraded using 5 mentioned degradation types and 7-9 degradation levels for JPEG and JPEG-2000 compression and 6 degradation levels for other degradation types. 20-29 observers tested and graded image degradations in 7 sessions, separately for each type of degradation. Observers graded 982 images in total, out of which 779 were degraded and other images were original. Similar actions were performed over each subjective grade like in our VCL@FER database, only DMOS results were calculated. At the end, DMOS (Difference MOS) results were scaled to the full range as described in (5) and averaged across all observers. Details can be found in [13].

2.5 IVC database

IVC database [5] consists of 10 original images with 5 types of degradation: JPEG, JPEG-2000, JPEG compression (luminance and chrominance), Gaussian blur and LAR (Locally Adaptive Resolution) compression. 15 subjects graded 185 degraded images with 2-5 degradation levels using DSIS subjective test method. Finally, MOS was calculated.

2.6 TID database

TID database [6] consists of 17 degradation types: additive Gaussian noise, additive noise in color components which is more intensive than additive noise in the luminance component, spatially correlated noise, masked

noise, high frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG compression, JPEG-2000 compression, JPEG transmission errors, JPEG-2000 transmission errors, non-eccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift) and contrast change.

Database consists of 25 original images (24 were obtained from Kodak database [12] and one is with patterns) and every degradation type had 4 degradation levels. This way a total of 1700 degraded images were graded. 838 subjects from three countries (Finland, Italy and Ukraine) graded 256428 image pairs with grades 0-9 on LCD or CRT 19" monitors with resolution 1152x864 pixels. Subjective screening was done according to the ITU-R BT.500-11 [8] and at the end MOS was calculated.

2.7 Toyama database

Toyama image database [7] consists of 2 degradation types: JPEG and JPEG-2000. 14 original images were used with 6 degradation levels, which give 168 degraded images. 16 subjects graded images on the scale 1-5 and in the end MOS was calculated.

3 OBJECTIVE QUALITY MEASURES

In this section six existing full-reference objective quality measures are described. Objective quality measures which were tested are: MSE (Mean Squared Error) [14], SSIM (Structural Similarity) [15], MS-SSIM (Multiscale SSIM) [16], VIF (Visual Information Fidelity) [17], IW-SSIM (Information Content Weighted SSIM) [18] and MAD (Most Apparent Distortion) [19]. MSE, and VIF can be calculated using Matlab program "Metrix_mux" [20]. SSIM, IW-SSIM and MS-SSIM measures were downloaded from [21] and MAD measure was downloaded from [22]. PSNR (Peak Signal to Noise Ratio) results are the same as MSE so these results will not be shown. All measures give best results when they're calculated from luminance component only, which means that all color images had to be transformed in grayscale images. For measures that are using transforms, depending on the number of scales and filter length, they had to be rescaled firstly.

3.1 MSE and PSNR

Mean Squared Error (MSE) measure [14] is defined as:

$$\mathrm{MSE} = \frac{\sum_i \sum_j (a_{i,j} - b_{i,j})^2}{x \cdot y}, \tag{6}$$

where in (6) *a* and *b* are original and distorted image. x and y are width and height of images.

Peak Signal to Noise Ratio (PSNR) measure [14] is defined as:

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{\mathrm{MSE}}. \tag{7}$$

3.2 SSIM

Structural Similarity index (SSIM) is a novel method for measuring the similarity between two images [15]. It is computed from three image measurement comparisons: luminance, contrast and structure. At each step, the local statistics and SSIM index are calculated within the local window. Because resulting SSIM index map often exhibits undesirable "blocking" artifacts, each window is filtered with normalized Gaussian weighting function (11×11 pixels) prior calculation of the three components mentioned earlier. Gaussian weighting function is described in (1) and afterwards normalized so that the sum of all filter values equals 1, Fig. 1.

In practice, one usually requires a single overall quality measure of the entire image, so 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 of perfect quality. It can give results between 0 and 1, where 1 means excellent quality and 0 means poor quality. It is calculated over 11x11 pixels from three components, luminance, contrast and structure (after being filtered):

$$\mathrm{SSIM\_}lum = \frac{2 \cdot \mu_x \cdot \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{8}$$

$$\mathrm{SSIM\_}cont = \frac{2 \cdot \sigma_x \cdot \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{9}$$

$$\mathrm{SSIM\_}struct = \frac{\sigma_{xy} + \frac{C_2}{2}}{\sigma_x \cdot \sigma_y + \frac{C_2}{2}}. \tag{10}$$

μ_x and μ_y are weighted means from original and degraded image, σ_x and σ_y are weighted variances from original and degraded image. σ_{xy} is similarly defined as weighted covariance between original and degraded image. C_1 and C_2 are constants defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$ where K_1 and K_2 are constants experimentally determined ($K_1 = 0.01$ and $K_2 = 0.03$). They help to improve stability of the measure when denominator is close to zero (for $K1$ and $K2$ equal to zero, this would be Universal Quality Index described in [23]). Final local SSIM measure is the product of (8), (9) and (10):

$$\mathrm{SSIM} = \frac{(2 \cdot \mu_x \cdot \mu_y + C_1)(2 \cdot \sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{11}$$

At the end, mean SSIM of all local SSIM values is calculated as their arithmetic mean.

In [24] it is suggested to downsample images prior calculating SSIM according to the:

$$F = \max(1, \mathrm{round}(\min(M, N)/256)). \tag{12}$$

In (13) $M \times N$ is image size. Then we average images over $F \times F$ pixels and downsample images F times in horizontal and vertical direction. Afterwards, SSIM measure is calculated according to (11). This SSIM was used later in comparison with other objective measures.

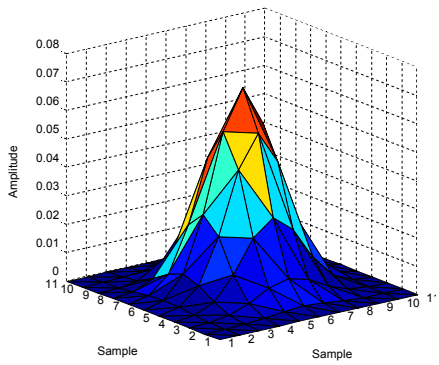


Fig. 1. Normalized Gaussian filter ($\sigma = 1.5$)

3.3 MS-SSIM

Similar like SSIM, MS-SSIM (Multiscale SSIM, MSSIM) method is a convenient way to incorporate image details at different resolutions [16]. This is a novel image synthesis-based approach which helps calibrating the parameters (like viewing distance) that weight the relative importance between different scales. MSSIM measure calculates SSIM on 5 scales. At every scale, image is averaged over 2x2 pixels and then SSIM contrast and structure is calculated according to (9) and (10) . On the next scale, image is downsampled by factor 2 in both directions. Only on the final, fifth scale, luminance component is also calculated according to (8). Final MSSIM is calculated as:

$$\mathrm{MSSIM} = \mathrm{SSIM\_}lum_5^{\alpha_5} \cdot \prod_{j=1}^{5} (\mathrm{SSIM\_}cont_j^{\beta_j} \cdot \cdot SSIM\_struct_j^{\chi_j}). \quad (13)$$

In (13) parameters α, β i γ are experimentally determined from LIVE image database:

$$\begin{aligned} \alpha_5 &= 0.1333 \\ \beta_j &= [0.0448 \;\; 0.2856 \;\; 0.3001 \;\; 0.2363 \;\; 0.1333] \\ \chi_j &= [0.0448 \;\; 0.2856 \;\; 0.3001 \;\; 0.2363 \;\; 0.1333] \end{aligned} \quad . \quad (14)$$

3.4 VIF

Visual Information Fidelity Criterion (VIF) [17] quantifies the Shannon information that is shared between the reference and the distorted images relative to the information contained in the reference image itself. It uses Natural Scene Statistics (NSS) modeling in concern 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. In first step, original and degraded images are transformed using SPWT [25]. VIF is then calculated from 5 parameters, out of which 2 are calculated from error image and 2 from only original image. Fifth parameter, noise variance is experimentally determined (0.4 in later comparison).

3.5 IW-SSIM

Prior calculating measures, original and degraded images are transformed using Laplacian pyramid transform on 5 scales [26].

Information-weighted SSIM (IW-SSIM, IWSSIM) is defined as [18]:

$$\begin{aligned} \mathrm{IW-SSIM} &= \prod_{i=1}^{M} (\mathrm{IW-SSIM}_j)^{\beta_j}, \\ \mathrm{IW-SSIM}_j &= \frac{\sum_i w_{j,i} \cdot c(x_{j,i}, y_{j,i}) \cdot s(x_{j,i}, y_{j,i})}{\sum_i w_{j,i}}, \\ &\qquad \mathrm{j} = 1, ..., \mathrm{M}-1, \\ \mathrm{IW-SSIM}_M &= \frac{1}{N_M} \sum_i l(x_{j,i}, y_{j,i}) \cdot c(x_{j,i}, y_{j,i}) \cdot \\ &\quad \cdot s(x_{j,i}, y_{j,i}). \end{aligned} \quad (15)$$

where x and y are i-th coefficients on j-th scale, β is defined in (14), M represents scale, w is weighted function and l, c and s are luminance, contrast and structure defined in (8), (9) and (10) accordingly.

3.6 MAD

Most Apparent Distortion (MAD) measure [19] is newly developed measure. After applying Contrast Sensitivity Function (CSF) described in [27], depending on the severity of the degradation, measure calculates one grade (detectability) and second grade (appearance) and finally combines them in one measure with nonlinear function, with parameters experimentally determined from A57 image database.

4 PERFORMANCE MEASURES

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

Pearson's correlation coefficient is calculated according to:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y}, \; \mathrm{i} = 1, ..., n. \quad (16)$$

In (16) x_i and y_i are grade values (x are objective grades and y are MOS), $\bar{x}$ and $\bar{y}$ are average grade values, and s_x and s_y are standard deviations, calculated by (17):

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i,$$
$$s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{17}$$
$$s_y = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

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 [13]:

$$Q(x) = b_1 \cdot \left(\frac{1}{2} - \frac{1}{1 + e^{b_2 \cdot (x - b_3)}} \right) + b_4 \cdot x + b_5. \tag{18}$$

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 quite 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 [19] somewhat different logistic function is proposed, with 4 parameters. We calculated Pearson's correlation using this function also:

$$Q(x) = \frac{b_1 - b_2}{1 + e^{\frac{x - b_3}{b_4}}} + b_2. \tag{19}$$

We used three different methods to find the best fitting coefficients: Trust-Region method, Levenberg-Marquardt method and Gauss-Newton method [28].

Final method for finding coefficients for nonlinear regression was the one which computed better results for performance measures (higher Pearson's correlation). An algorithm for optimizing coefficients b in (18) was developed. Firstly, set of 20 starting b parameters were checked to see which one gives best overall Pearson's correlation. For (18) $b_{1-5} = [i,\ i,\ i,\ i,\ i]$ and $[i,\ i+1,\ i+2,\ i+3,\ i+4]$, for $i \in \{1, 10\}$; for (19) $b_{1-4} = [i,\ i,\ i,\ i]$ and $[i,\ i+1,\ i+2,\ i+3]$, for $i \in \{1, 10\}$. Iterative algorithm for finding best b parameters was performed as long as difference between new and old Pearson's correlation was not under 0.0001. Best b coefficients were determined by the highest Pearson's correlation after nonlinear regression, for every optimization method and every starting parameter. At the end, same iterative algorithm was performed, where starting parameters for every image database were chosen as ending (best) parameters of all other image databases (for the same image quality measure).

Spearman's correlation coefficient [29] 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 in Eq. (16) over ranked variables. Rank of the sample in variable is its sorted location in a row. In the case of tied ranks, positions of all tied samples are calculated as an arithmetic mean of their ranks. If there are no any tied ranks, Spearman's correlation coefficient can be calculated simpler as:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}. \tag{20}$$

In (20) $d_i = x_i - y_i$ are differences between the ranks of each observation from the two variables being compared and n is the number of samples.

Kendall's rank correlation coefficient [30] is another performance measure which was used to compare objective and subjective measures. It measures the similarity of the orderings of the data when ranked by each of the quantities. All pairs of observations are ranked according to the first variable X (rank i) and then according to the second variable Y (rank j). Afterwards, every pair of observations from the first ranking is compared with all pairs of observations from the second ranking. Any pair of observations (x_i, y_i) and (x_j, y_j) are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$ (case of tied ranks), the pair is neither concordant nor discordant. Final correlation coefficient is calculated as (τ_b coefficient):

$$\tau_b = \frac{n_{concordant} - n_{discordant}}{\sqrt{\left(\frac{N \cdot (N-1)}{2} - \sum_{i=1}^{T} \frac{t_i (t_i - 1)}{2} \right)}} \cdot \frac{1}{\sqrt{\left(\frac{N \cdot (N-1)}{2} - \sum_{j=1}^{U} \frac{u_j (u_j - 1)}{2} \right)}}. \tag{21}$$

Table 2. Pearson's correlation for all objective quality measures and all image databases, using 5-parameter logistic function

	IWSSIM	MAD	MSE	MSSIM	SSIM	VIF
A57	0.90353	0.91079	0.69324	0.86039	0.80188	0.69899
CSIQ	0.91441	0.95067	0.81536	0.89974	0.86126	0.92775
LIVE	0.95219	0.96752	0.87305	0.94894	0.94488	0.95983
IVC	0.92306	0.92195	0.72145	0.91085	0.91194	0.90283
VCL@FER	0.91909	0.90531	0.8241	0.92319	0.91436	0.89548
TID	0.85791	0.83083	0.58495	0.84515	0.77317	0.80934
TOYAMA	0.92488	0.94068	0.64918	0.89274	0.8887	0.91629
mean	0.91358	0.91825	0.73733	0.89728	0.87089	0.87293
wt_mean	0.90017	0.89844	0.72386	0.8898	0.85092	0.87826

Table 3. Pearson's correlation for all objective quality measures and all image databases, using 4-parameter logistic function

	IWSSIM	MAD	MSE	MSSIM	SSIM	VIF
A57	0.90244	0.90591	0.66947	0.85734	0.80185	0.61604
CSIQ	0.90253	0.95019	0.80299	0.89718	0.85938	0.92526
LIVE	0.9425	0.96718	0.85823	0.94021	0.93835	0.95924
IVC	0.92285	0.92096	0.72065	0.91067	0.91165	0.90262
VCL@FER	0.91526	0.90506	0.81091	0.91831	0.90892	0.89234
TID	0.84882	0.83057	0.56892	0.84044	0.77153	0.80505
TOYAMA	0.9244	0.94062	0.62642	0.89201	0.88771	0.91367
mean	0.9084	0.91721	0.72251	0.89374	0.86849	0.85918
wt_mean	0.89191	0.89804	0.70945	0.88514	0.84796	0.8744

Table 4. Spearman's correlation for all objective quality measures and all image databases

	IWSSIM	MAD	MSE	MSSIM	SSIM	VIF
A57	0.87127	0.90139	0.61763	0.8415	0.80666	0.62228
CSIQ	0.92129	0.94665	0.8058	0.91364	0.87563	0.91945
LIVE	0.95665	0.96689	0.87556	0.95128	0.9479	0.96315
IVC	0.9125	0.91457	0.68844	0.898	0.90182	0.89637
VCL@FER	0.91633	0.90607	0.82465	0.92269	0.91125	0.88665
TID	0.85594	0.83401	0.5531	0.85418	0.77493	0.74907
TOYAMA	0.92024	0.93617	0.61319	0.88738	0.87938	0.90767
mean	0.90775	0.91511	0.7112	0.89552	0.87108	0.84923
wt_mean	0.9002	0.89826	0.70611	0.89553	0.85391	0.85068

Table 5. Kendall's correlation for all objective quality measures and all image databases

	IWSSIM	MAD	MSE	MSSIM	SSIM	VIF
A57	0.68462	0.72238	0.43007	0.64825	0.60629	0.45944
CSIQ	0.75287	0.79701	0.60836	0.7395	0.6907	0.75373
LIVE	0.81752	0.84213	0.68646	0.80445	0.79629	0.82701
IVC	0.73388	0.74061	0.52175	0.7203	0.72231	0.71581
VCL@FER	0.7372	0.72135	0.63614	0.74973	0.73315	0.69244
TID	0.66364	0.64451	0.40275	0.65685	0.57676	0.58605
TOYAMA	0.75366	0.78229	0.44428	0.70286	0.69394	0.7315
mean	0.73477	0.75004	0.53283	0.71742	0.68849	0.68085
wt_mean	0.72568	0.7313	0.53248	0.71652	0.67068	0.68671

In (21) N is the number of observations, t_i is the number of t similar samples of variable X at rank $i \in \{1, \mathrm{T}\}$. Similarly, u_j is the number of u similar samples of variable Y at rank$j \in \{1, \mathrm{U}\}$. In the case where there are no tied ranks, Kendall's correlation coefficient can be simplified and calculated as (τ_a coefficient), (22):

$$\tau_a = \frac{n_{concordant} - n_{discordant}}{\frac{N \cdot (N-1)}{2}}. \tag{22}$$

5 RESULTS

When comparing images across multiple databases, correlation can be calculated as an arithmetic mean or weighted arithmetic mean as proposed in [18]. Weighted arithmetic mean is calculated as:

$$wt\_mean = \frac{\sum_{i=1}^{7}(w_i \cdot corr_i)}{\sum_{i=1}^{7} w_i}$$
$$w_i = \{54,\ 866,\ 779,\ 185,\ 552,\ 1700,\ 168\}. \quad (23)$$

In (23) w_i are database sizes (A57, CSIQ, LIVE, IVC, VCL@FER, TID and TOYAMA accordingly). Mean and weighted mean correlations are shown on Fig. 2. Pearson's correlation is calculated after nonlinear regression.

Pearson's correlation (for 4 and 5 parameter fitting function), Spearman's and Kendall's correlations are presented in Tables 2-5.

From the Fig. 2 it can be concluded that objective measures from the best are in the following order:

Arithmetic mean

- Pearson's correlation, 5 parameters in fitting function: MAD, IWSSIM, MSSIM, VIF, SSIM, MSE
- Pearson's correlation, 4 parameters in fitting function: MAD, IWSSIM, MSSIM, VIF, SSIM, MSE
- Spearman's correlation: MAD, IWSSIM, MSSIM, SSIM, VIF, MSE
- Kendall's correlation: MAD, IWSSIM, MSSIM, SSIM, VIF, MSE

Weighted mean

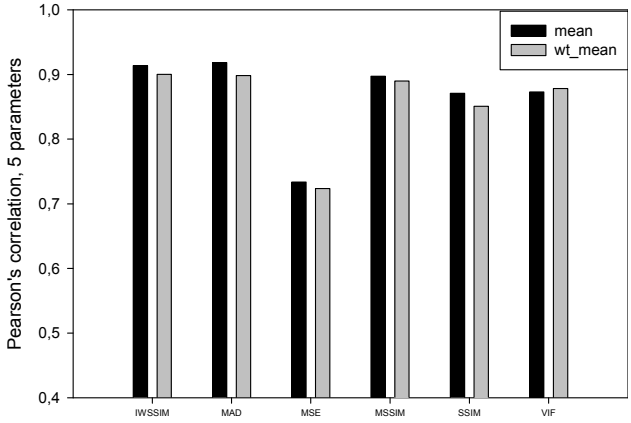
- Pearson's correlation, 5 parameters in fitting function: IWSSIM, MAD, MSSIM, VIF, SSIM, MSE
- Pearson's correlation, 4 parameters in fitting function: MAD, IWSSIM, MSSIM, VIF, SSIM, MSE
- Spearman's correlation: IWSSIM, MAD, MSSIM, SSIM, VIF, MSE
- Kendall's correlation: MAD, IWSSIM, MSSIM, VIF, SSIM, MSE

It can be concluded that best correlation results are obtained using IWSSIM or MAD measures, depending on the correlation type and fitting function (for Pearson's correlation smaller differences are possible). However, it should be noted that MAD measure has significantly higher calculation time than IWSSIM measure due to the Gabor filter calculation on 5 scales and 4 orientations.

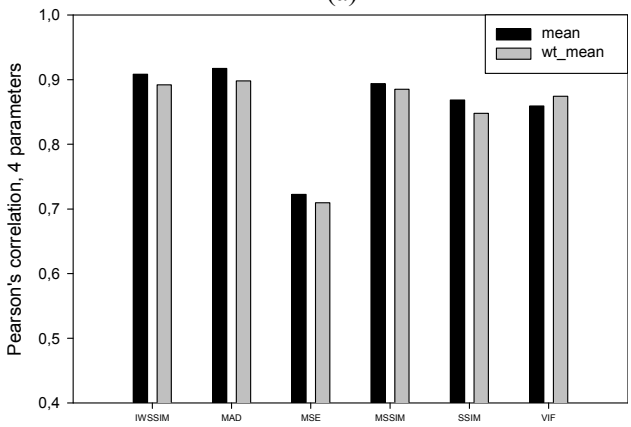
When comparing results from only VCL@FER database, best measure is MSSIM, however IWSSIM and MAD measures have similar correlations.

6 CONCLUSION

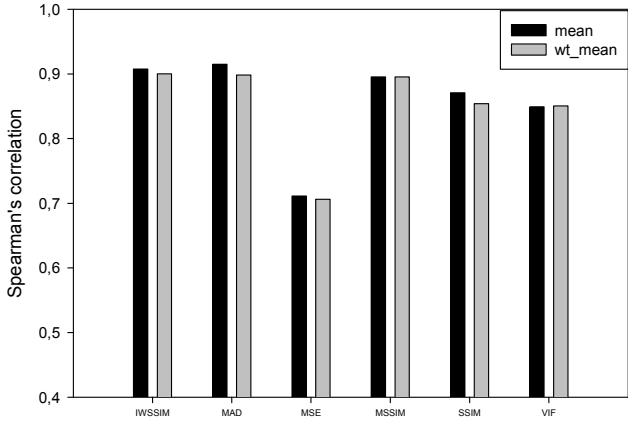
In this paper we presented newly developed image quality database VCL@FER and compared six objective measures using this database, as well as six other publicly available databases. Results show that best results are obtained for IWSSIM and MAD measures. In our database MSSIM shows best correlation results, however MAD and



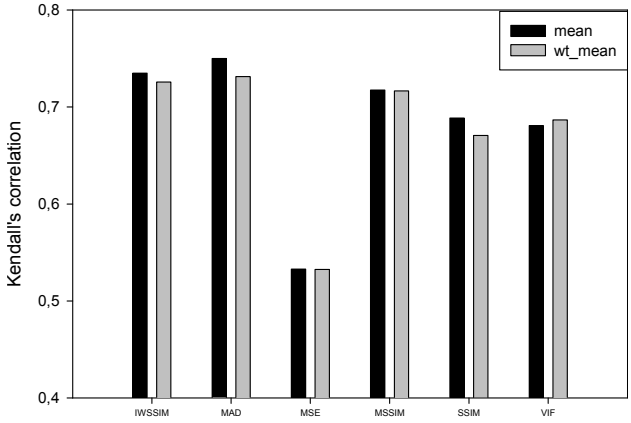
(a)



(b)



(c)



(d)

Fig. 2. Mean and weighted mean correlations for all databases: (a) Pearson's correlation, 5 fitting parameters, (b) Pearson's correlation, 4 fitting parameters, (c) Spearman's correlation, (d) Kendall's correlation

IWSSIM perform near equally well.

It can be concluded that VCL@FER database can be used to test other objective measures (like reduced-reference or no-reference measures) as well as to develop and test new image quality measures.

ACKNOWLEDGMENT

The work described in this paper was conducted under the research project "Picture Quality Management in Digital Video Broadcasting" (036-0361630-1635) supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] Lin Zhang, Dept. Computing, The Hong Kong Polytechnic University "Research on Image Quality Assessment", http://www4.comp.polyu.edu.hk/˜cslinzhang/IQA/IQA.htm

[2] Damon M. Chandler, Sheila S. Hemami, Online Supplement to "VSNR: A Visual Signal-to-Noise Ratio for Natural Images Based on Near-Threshold and Suprathreshold Vision", http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[3] E. C. Larson, D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", Journal of Electronic Imaging, 19 (1), 2010., http://vision.okstate.edu/index.php?loc=csiq

[4] H.R. Sheikh, Z.Wang, L. Cormack, A.C. Bovik, "LIVE Image Quality Assessment Database Release 2", http://live.ece.utexas.edu/research/quality/subjective.htm

[5] Patrick Le Callet, Florent Autrusseau: Subjective quality assessment IRCCyN/IVC database, http://www.irccyn.ec-nantes.fr/ivcdb/

[6] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics", Advances of Modern Radioelectronics, Vol. 10, str. 30-45, 2009., http://www.ponomarenko.info/tid2008.htm

[7] MICT (Media Information and Communication Laboratory) Image Quality Evaluation Database: http://mict.eng.u-toyama.ac.jp/mictdb.html

[8] ITU-R BT.500-11 "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union/ITU Radiocommunication Sector, January 2002.

[9] A. Zaric, N. Tatalovic, N. Brajkovic, H. Hlevnjak, M. Loncaric, E. Dumic, S. Grgic, " VCL@FER Image Quality Assessment Database", Proceedings of the 53rd International Symposium ELMAR-2011, pp., 14-16 September 2011., http://www.vcl.fer.hr/quality/

[10] Irfanview software: http://www.irfanview.com/

[11] JPEG2000 coder: http://www.kakadusoftware.com/

[12] Kodak Lossless True Color Image Suite: http://r0k.us/graphics/kodak/

[13] H.R. Sheikh, "Image Quality Assessment Using Natural Scene Statistics," Ph.D. dissertation, University of Texas at Austin, 2004.

[14] S. Grgic, M. Grgic, M. Mrak, "Reliability of Objective Picture Quality Measures", Journal of Elect. Engineering, Vol. 55, No. 1-2, 2004., pp. 3-10.

[15] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Trans. on Image Proc., Vol. 13, No. 4, 2004., pp. 600-612.

[16] Z. Wang, E.P. Simoncelli, A.C. Bovik, "Multiscale structural similarity for image quality assessment", 37th Proc. IEEE Asilomar Conf. on Signals, Systems and Computers, Vol. 2, 2003., pp. 1398-1402.

[17] H.R. Sheikh and A.C. Bovik, "Image information and visual quality", IEEE Trans. Image Processing, Vol. 15, No. 2, 2006., pp. 430-444.

[18] Z. Wang, Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment", IEEE Trans. on Image Processing, Vol. 20, No. 5, 2011., pp. 1185-1198.

[19] E. C. Larson, D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", Journal of Electronic Imaging, Vol. 19, No. 1, Article ID 011016, 2010., pp. 1-22.

[20] Visual Quality Assessment Package Version 1.1, http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[21] https://ece.uwaterloo.ca/˜z70wang/research/iwssim/

[22] http://vision.okstate.edu/mad/

[23] Z. Wang and A. C. Bovik, "A universal image quality index", IEEE Signal Processing Letters, Vol: 9 No: 3, 2002., pp. 81-84.

[24] https://ece.uwaterloo.ca/˜z70wang/research/ssim/

[25] E. P. Simoncelli, W. T. Freeman, "The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation", 2nd IEEE International Conference on Image Processing, Vol. 3, 1995., pp. 444-447.

[26] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code", IEEE Trans. Communications, Vol. 31, 1983., pp. 532–540.

[27] J. Mannos, D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images", IEEE Trans. Inf. Theory, Vol. 20, No. 4, 1974., pp. 525-535.

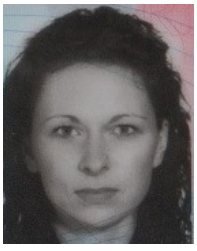
[28] Nocedal, Jorge, Wright, Stephen, "Numerical optimization", New York: Springer, 1999.

[29] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data", Proceedings of the MAT TRIAD 2007 Conference, Bedlewo, Poland, 2007.

[30] M. Kendall, "A New Measure of Rank Correlation", Biometrika 30 (1-2), 1938., pp. 81-89.

Anđela Zarić received the BSc degree in Electrical Engineering and Information Technology and the MSc degree (*summa cum laude*) in Information and Communication Technology from Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2009 and 2011, respectively. During her Master studies, from 2009 to 2011, she has been involved in a project on Image Quality Assessment at the Video Communications Laboratory, Faculty of Electrical Engineering and Computing, University of Zagreb. She is now working toward her PhD degree at Instituto de Telecomunicacoes – Instituto Superior Tecnico, Technical University of Lisbon, Portugal. Her current research interests include UWB antennas, reconfigurable antennas, ranging and localization techniques.

Nenad Tatalović was born on 29th of May, 1987 in Zagreb, where he attended the X. gymnasium from 2002 until 2006. He finished Bachelor's degree in the Computing field and Masters' degree in the field of Information and Communication Technology from Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2009 and 2011, respectively. Since finishing college, he works for Ericsson Nikola Tesla in Zagreb, first as a IMS Solution Integrator from 2011 until July of 2012, and then as a Software Designer from 2012 onwards.



Nikolina Brajković was born in Zagreb, 1986. She received the B.Sc. and M.Sc. degrees in Information and communication technologies from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 2009 and 2012 respectively. She is currently an intern in NATO Headquarter, Brussels, Belgium.

Hrvoje Hlevnjak was born on 19th of September, 1987 in Zagreb where he finished X. gymnasium between 2002 and 2006. After finishing high school he enrolled to the Faculty of Electrical Engineering and Computing, at University of Zagreb, where he got his Bachelor's degree in Computing in 2009, and Master's degree in Information and communication technology. After finishing college he worked as Information security specialist at Mack IT, and is currently working as a Software developer at Lemax.

Matej Lončarić was born in Zagreb, 1987. He received the M.Sc. degree in Information and Communication Technology from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia in 2012. He currently works on first private commercial television station, Nova TV, as New Media Producer. His research interests include digital television, Picture Quality Measurements, Live Video Streaming, Internet Communications, Web production, Project Development and Organization.

Emil Dumić was born in Zagreb, 1985. He received the B.Sc. and Ph.D. degrees in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 2007 and 2011 respectively. He is currently a Ph.D. at the Department of Wireless Communications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His research interests include image interpolation, wavelet transforms, image quality and digital television.

Sonja Grgić was born in Vukovar, Croatia. She received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 1989, 1992 and 1996, respectively. Since 1989 she is with Faculty of Electrical Engineering and Computing, where presently she is Full Professor at the Department of Wireless Communications. She is author or co-author of 4 book chapters, 18 papers published in scientific journals and more that 120 papers published in conference proceedings. She was editor of 8 international conference proceedings. She has participated in 10 domestic and international scientific projects. Her research interests include digital television, picture quality assessment, video compression, wavelet analysis for image processing and compressive imaging. Prof. Grgic is active member of SMPTE, EURASIP, KoREMA and Elmar. Since 1998 she is a member of the Croatian Academy of Engineering (HATZ). She received the silver medal "Josip Lončar" from the Faculty of Electrical Engineering and Computing in Zagreb for an outstanding Ph.D. thesis and annual award "Rikard Podhorsky" for year 2006 from HATZ.

AUTHORS' ADDRESSES

Anđela Zarić, M.Sc.
Instituto de Telecomunicacoes,
Instituto Superior Tecnico,
Technical University of Lisbon,
Avenida Rovisco Pais 1, PT-1049-001 Lisboa, Portugal
email: andela.zaric@gmail.com

Nenad Tatalović, M.Sc.
Ericsson Nikola Tesla,
Krapinska 45, HR-10000 Zagreb, Croatia
email: flamelash@gmail.com

Nikolina Brajković, M.Sc.
NATO Headquarter Consultation,
Command and Control Staff,
Blvd Leopold III, BE-1110 Brussels, Belgium
email: brajkovic.nikolina@gmail.com

Hrvoje Hlevnjak, M.Sc.
Lemax,
Radnička cesta 1a, HR-10000 Zagreb, Croatia
email: hrvoje.hlevnjak@gmail.com

Matej Lončarić, M.Sc
Nova TV,
Remetinečka cesta 139, HR-10000 Zagreb, Croatia
email: matej.loncaric@gmail.com

Emil Dumić, Ph.D.
Prof. Sonja Grgić, Ph.D.
Department of Wireless Communications,
Faculty of Electrical Engineering and Computing,
University of Zagreb,
Unska 3, HR-10000, Zagreb, Croatia
email: emil.dumic@fer.hr, sonja.grgic@fer.hr

Received: 2012-04-05
Accepted: 2012-09-08