Abstract - In this paper we present new image quality database which consists of four degradation types: JPEG, JPEG2000, White noise and Gaussian blur. Results for five commonly used objective quality measures are compared using newly developed image database, as well as LIVE image database.

Keywords – image database, Mean Opinion Score, Jpeg, Jpeg2000, White noise, Gaussian blur, MSE, PSNR, SSIM, MSSIM, VIF, VSNR

I. 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 psycho-visual 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.

II. VCL@FER - NEW IMAGE DATABASE

The database consists of 575 images in total. 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 levels, depending on severity. 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 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 ($n1 \times n2$ pixels). It can be described as (without normalization):

\[
G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}}
\]

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 Rec. BT. 500-11 [8].

This paper is organized as follows. Section II describes new image database and subjective quality assessment methodology. Section III describes existing objective measures. In section IV explains performance measures used in comparing objective measures. Section V presents results and finally section VI gives the conclusion.

| 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         |
6 different sizes of Gaussian function were used to calculate blur degradation, using Irfanview [9].

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 bites per pixel (8 is without compression), using “kdu_compress” [10].

Testing was done on a study group of 118 people, non-experts, between 20 and 30 years of age. Each subject was given about 96 images to grade. The graded images were not grouped by types of degradation, nor were test subjects told about the types of degradation which to expect. Each image was rated between 16 and 36 times, 20 times on average. Method used in testing was Single-stimulus (SS) method, which uses numeric criteria scale with 100 grades, and was further compared to graph scale and scale ratio, as described in ITU-R BT. 1082 [11].

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

All the testing results and grades have been collected, with grades for each picture being averaged. According to ITU-R 500-11[2] 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 condition, one repetition and one time window within a test condition.

Afterwards, results for every observer were rescaled to the full (and same) range of 0-100, according to the (5):

\[
mos_{n,l} = \frac{100}{\max(r_n)-\min(r_n)} (r_{n,l}-\min(r_n))
\]

In (6) \(r_{n,l}\) represents grade that the \(n\)-th viewer has given for \(l\)-th image (including reference images), \(mos_{n,l}\) 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.

III. OBJECTIVE QUALITY MEASURES

To compare correlation results between subjective scores from our database and objective measures, we used six objective measures: MSE (Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity) [12], MSSIM (Multiscale SSIM) [13], VIF (Visual Information Fidelity) [14] and VSNR (Visual Signal-to-Noise Ratio) [15].

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

\[
PSNR = 10 \log_{10} \frac{255^2}{\text{MSE}}
\]

(6) \(a_{i,j}\) and \(b_{i,j}\) are corresponding pixels from original and tested image. \(x\) and \(y\) describe height and width of an image. In (7) 255 is maximum possible amplitude for 8-bit image.

SSIM a novel method for measuring the similarity between two images [12]. 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 with normalized Gaussian weighting function described in (1), with \(\sigma=1.5\)):

\[
h(n_1, n_2) = e^{-\frac{(n_1-n_2)^2}{2\sigma^2}}
\]

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:

\[
P_l > 2\% \text{ or } Q_n > 2\%
\]

Using (4), 2 observers were discarded. Recommendation ITU-R 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.
Visual Information Fidelity Criterion (VIF) [14] 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.

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

IV. PERFORMANCE MEASURES

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

Pearson's correlation coefficient is calculated according to:

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

In (15) \(x_i\) 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 (16):

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{y} = \frac{1}{n} \sum y_i$$

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2}$$

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 [18]:

$$Q(x) = b_1 \cdot \frac{1}{1 + e^{b_2(x-x_0)}} + b_3 \cdot x + b_4$$

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.

We used three different methods to find the best fitting coefficients: Trust-Region method [19], Levenberg-Marquardt method [20] and Gauss-Newton method [21].

Final method for finding coefficients for nonlinear regression was the one which computed better results for performance measures (lower RMSE (18) and higher Pearson's (17) correlation). An algorithm for optimizing coefficients \( b \) in (17) was developed. Firstly, set of 20 starting \( b \) parameters were checked to see which one gives best overall Pearson's correlation: \([b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}], b_{11}, b_{12}, b_{13}, b_{14}, b_{15}, b_{16}, b_{17}, b_{18}, b_{19}, b_{20}\), 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 lowest RMSE after nonlinear regression, for every optimization method and every starting parameter. At the end, same iterative algorithm was performed, where starting parameters for every degradation type were chosen as ending (best) parameters of all other degradation types.

RMSE (Root Mean Square Error) is calculated as:

\[
\text{RMSE} = \sqrt{\frac{1}{n-k} \sum (x - y)^2} \tag{18}
\]

In (19) \( n \) is the number of tested images modified by a correction for degrees of freedom \((k=5\) in our case, we have 5 parameters in fitted function, Equation (18)), \( x \) is DMOS or MOS measure and \( y \) fitted objective measure after nonlinear regression.

Spearman's correlation coefficient [22] 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 (15) over ranked variables. Rank of the sample in variable is its rank. If there are no any tied ranks, Spearman's correlation coefficient can be calculated simpler as:

\[
\rho = 1 - \frac{6 \sum d_i^2}{n (n^2 - 1)} \tag{19}
\]

In (19) \( d_i = x_i - y_i \) 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 [23] 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 \( x_i = x_j = y_i = y_j \) (case of tied ranks), the pair is neither concordant nor discordant. Final correlation coefficient is calculated as \((\tau_b\) coefficient):

\[
\tau_b = \sqrt{\frac{n_{\text{concordant}} - n_{\text{discordant}}}{N(N-1)}}, \quad (20)
\]

In (20) \( N \) is the number of observations, \( t_i \) is the number of \( t \) similar samples of variable \( X \) at rank \( i \in \{1, T\} \). Similarly, \( u_j \) is the number of \( u \) similar samples of variable \( Y \) at rank \( j \in \{1, U\} \). In the case where there are no tied ranks, Kendall's correlation coefficient can be simplified and calculated as \((\tau_s\) coefficient):

\[
\tau_s = \sqrt{\frac{n_{\text{concordant}} - n_{\text{discordant}}}{N(N-1)}} \tag{21}
\]

V. RESULTS

Results for five commonly used objective quality measures are compared using newly developed VCL@FER image database, as well as LIVE image database. Subjective quality evaluation in LIVE database was based on ITU-R recommendation BT.500-11. Details of the subjective testing can be found in [24]. Briefly they are as follows: 29 high-resolution 24 bits per pixel RGB color images (typically 768x512) were degraded using five degradation types: JP2K - JPEG2000 compression, JPEG - JPEG compression, AWGN - Average white Gaussian noise, Gblur - Gaussian blur, Fastfading - transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model.

Each of these 29 images had versions with 7-9 different qualities for JPEG and JPEG2000 and 6 images with different qualities for white noise, Gaussian blur and Fastfading. About 20-29 observers had to grade image quality on a continuous scale with 5 grades ("Bad", "Poor", "Fair", "Good" and "Excellent"). In this way observers evaluated total of 982 images, out of which 203 were reference and 779 degraded images. The experiments were conducted in seven sessions: two sessions for JPEG2000, two for JPEG, and one each for white noise, Gaussian blur, and fastfading transmission errors. Outlier detection algorithm was used to screen observers if their results differ too much from the average results (4 subjects were rejected). Similar actions were performed over each subjective grade like in our VCL@FER database, only DMOS results were calculated using difference scores between degraded and original subjective score (for each observer),
Unlike our database which uses subjective raw scores directly. At the end, DMOS (Difference MOS) results were scaled to the full range as described in (5) and averaged across all observers.

Figures 2 and 3 show the results of each of the four performance measures (Pearson's, Spearman's, Kendall's correlation and RMSE), compared to LIVE database. Results for Pearson's correlation and RMSE are given after nonlinear regression. Results for PSNR measure are the same as the results for MSE, after nonlinear regression, so PSNR results were skipped. Legend used in Figs. 2-3 is shown in Fig. 2. Generally, for every measure from the left to the right performance measures are shown in the order of the degradation type: overall, AWGN, Blur, JPEG2000 and JPEG (Fastfading degradation in LIVE database was skipped, however it is included in overall results).

VI. CONCLUSION

From the Figs. 3-4 it can be concluded that in general best overall measure as well as for degradation types separately in LIVE database is VIF. However, in VCL@FER database MSSIM outperforms VIF measure when comparing all images. VIF gives best performance measures for all degradation types separately in VCL@FER database, except for MSE which outperforms VIF for AWGN degradation, for Spearman's and Kendall's correlation. In general, results in both image databases give similar conclusions about objective quality measures, only our image database gives lower correlation results because we used MOS measure, unlike LIVE which uses DMOS measure. DMOS measure is more useful for comparing full-reference objective measures. However, MOS measure can be compared with real life situations where reference image is generally unknown. Also, LIVE database consists of 5 different degradation types, unlike VCL@FER which has four. All this can have influence on the final conclusion about which objective measure correlates better. When comparing different performance measures generally better correlation of one type means lower RMSE and better correlation of other types. However, smaller differences are possible for Pearson's correlation because of the fitting algorithm and optimization coefficients $b$ in (17).

In future research, it is possible to include other new developed objective measures, as well as other publicly available image databases, so that comparison between objective measures has even better statistical reliability. Statistical significance between variances of the difference between objective (after nonlinear regression) and subjective measures can also be evaluated, thus showing if correlation results are statistically significant or not.

---

**Figure 1.** Legend used in Figs. 2-3

**Figure 2.** Pearson's correlation and RMSE (after nonlinear regression) for LIVE and VCL@FER image databases
Objective measures

MSE  MSSIM  SSIM  VIF  VSNR

Spearman's correlation, LIVE

0.4  0.5  0.6  0.7  0.8  0.9  1.0

Objective measures

MSE  MSSIM  SSIM  VIF  VSNR

Spearman's correlation, VCL@FER

0.4  0.5  0.6  0.7  0.8  0.9  1.0

Objective measures

MSE  MSSIM  SSIM  VIF  VSNR

Kendall's correlation, LIVE

0.4  0.5  0.6  0.7  0.8  0.9  1.0

Objective measures

MSE  MSSIM  SSIM  VIF  VSNR

Kendall's correlation, VCL@FER

0.4  0.5  0.6  0.7  0.8  0.9  1.0

Figure 3. Spearman's and Kendall's rank order correlation for LIVE and VCL@FER image databases

ACKNOWLEDGMENT

The work described in this paper was conducted under the research projects: "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/~csinzhang/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/VC database, http://www.irccyn.ec-nantes.fr/ivecb/

[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://micr.ece.u-toyama.ac.jp/mitcdb.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] Irfanview software: http://www.irfanview.com/

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

[11] BT.1082-1: "Studies toward the unification of picture assessment methodology " (1990)

[12] 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, 608-612 (April 2004)

[13] Z. Wang, E.P. Simoncelli, A.C. Bovik, "Multi-scale structural similarity for image quality assessment", in 37th Proc. IEEE Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA (Nov. 2003)

[14] H.R. Sheikh and A.C. Bovik, "Image information and visual quality", IEEE Trans. Image Processing, Dec. 2003

[15] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images", IEEE Transactions on Image Processing, Vol. 16, Issue 9, 2284-2298 (September 2007)

[16] https://ece.uwaterloo.ca/~z70wang/research/ssim/

[17] Z. Wang and A. C. Bovik, "A universal image quality index", IEEE Signal Processing Letters, Vol: 9 No: 3 (March 2002)

[18] H.R. Sheikh, M.F. Sabir, A.C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms", IEEE Transactions on Image Processing, Vol. 15, Issue 11, 3440-3451 (2006)

[19] J.J. Moré and D.C. Sorensen, "Computing a Trust Region Step," SIAM Journal on Scientific and Statistical Computing, Vol. 3, 553-572 (1983)

[20] K. Levenberg, "A Method for the Solution of Certain Problems in Least Squares," Quart. Appl. Math. Vol. 2, 164-168 (1944)

[21] J.E. Dennis, Jr., "Nonlinear Least-Squares," State of the Art in Numerical Analysis, ed. D. Jacobs, Academic Press, 269-312 (1977)

[22] 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 (March 2007)

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

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