Image Quality Assessment - Comparison of Objective Measures with Results of Subjective Test

Andela Zaric, Matej Loncaric, Dijana Tralic, Maja Brzica, Emil Dumic, Sonja Grgic
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
andela.zaric@fer.hr; matej.loncaric@fer.hr

***Abstract* - Image quality assessment is a challenging task that is traditionally approached by computational models. To maintain, control, and enhance the quality of images, it is important for image acquisition, management, communication, and processing systems to be able to identify and quantify image quality degradations. A great deal of effort has been made in recent years to develop objective image quality metrics that correlate with perceived quality measurement. To find the right solution to this problem we need measures to estimate quality and amounts of degradation compared with original image to determinate optimal quality of the coded picture. This paper provides comparison of subjective and objective picture quality for three different distortions, each made with four different levels of distortion.**

***Keywords* - JPEG, Blur, Gaussian Noise, SSIM, MSE, MOS**

I. INTRODUCTION

Traditionally, image quality has been evaluated by human subjects. This method, though reliable, is expensive and too slow for real-world applications. So we need to rely on objective image quality assessment, where the goal is to provide computational models that can automatically predict perceptual image quality. In this paper we made a correlation between subjective and objective test results, picture quality assessment criteria, subjective and objective methods and metrics, testing procedures.

II. OBJECTIVE IMAGE QUALITY ASSESSMENT

People are the ultimate beneficiaries of most visual applications and as such can most accurately assess the image quality. Image quality tests carried out with human observers, called subjective tests, provide most accurate assessment of image quality. But the aim is to develop objective measures because subjective tests consume a lot of time to carry out and are expensive compared to computable algorithms that can quickly assess image quality.

When designing an objective measure it is needed to take into consideration the characteristics of human visual system (HVS) but then the algorithm can be very complex. Therefore a compromise should be found between the accuracy of the measure and its calculation complexity.

In design of such measure following information may be used: available information about the original image, information about the distortion that occurred on the images and information about the characteristics of the HVS. In accordance with that information, objective measures can roughly be divided into these three categories:

- full-reference (FR) image quality measures - based on the difference between original and distorted image,
- reduced-reference (RR) image quality measures - quality of distorted image evaluated based on information extracted from original image,
- no-reference (NR) image quality measures - based on the measurement of image distortion at the place of receipt without any knowledge about the original [1].

In addition to this classification objective image quality measures can be divided by their purpose on the measures that are designed for specific applications and on general measures which do not imply the type of distortion. Third possible division is based on two different approaches to describing the HVS. The first approach would be to study and simulate each component of HVS separately and then connect them in one model to fully describe HVS. Another approach is to observe the HVS from the outside, and describe it in terms of relationship between inputs and outputs (i.e. black box model). Designing measures that correspond to HVS is more complex than designing FR measure, but these NR measures give similar grades as an average observer would give.

A. Mean Square Error - MSE

The simplest and oldest objective measure for evaluating image quality is the Mean Square Error (MSE). Unfortunately it is still commonly used despite the perceived shortcomings. The main reason of it's shortcomings it that characteristics of HVS are not included in the MSE model and the knowledge of features of real images is not also included. But it is a very simple measure to compute.

Let x and y represent two images:

$$\vec{x} = \{x_i \mid i = 1,2,\ldots, N\} \quad (1)$$

$$\vec{y} = \{y_i \mid i = 1,2,\ldots, N\} \quad (2)$$

where N is the number of pixels of each image, and **x** is the original and **y** distorted image. This notation treats the image as one-dimensional vector and does not take into consideration the arrangement of pixels in the image, nor their correlation. The MSE is defined as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \quad (3)$$

Figure 1 demonstrates why the MSE is poor in evaluating image quality. It shows the original image "Girl" that was distorted by these procedures: changed luminance, changed contrast, added impulse noise, added Gaussian noise, added blur and used JPEG compression. Values of MSE measures are listed below each distorted image and is clearly seen that this measure is very poorly correlated with perceived image quality. All MSE grades are very close to each other and the quality is not similar at all.

It is obvious that MSE does not take into consideration the way people perceive images, and the characteristics of the HVS. This is because it is based on a l_p norm (any norm that uses the absolute difference between elements). These are the assumptions made when using an l_p norm (that result in describe behaviour of MSE):

- assumption that the perceived image quality is independent of the spatial distribution of pixels ,
- assumption that the perceived image quality is independent of the error signal, meaning if the same error signal was added on two different images distorted pictures' quality would be the same,
- assumption that the perceived image quality is determined by amplitude and not the sign of error signal,
- assumption that all the pixels of an image are equally important to image quality [1].

None of these assumptions are grounded when examining quality of natural images because such images are highly structured. That means that arrangement and structure of pixels in the image carry out the most information about the objects and directly affects the perceived image quality. And the correlation between the error signal and the original image significantly affects the perceived quality (if the error signal is very similar to the image then the average observer would not perceive this error like in a case where there is a large difference between the image and signal error).

B. Structural Similarity Index - SSIM

As it was already concluded, natural images are highly structured, and pixels in an image are mutually dependent and that dependence carries information about the shape of objects in the picture. The basic idea is that the average observer perceives image quality according to the objects in the image. So measuring the structural similarity between images can well approximate observer's perceived image quality.

To make the concept of structural similarity applicable inimage quality measures it must be defined which distortions are structural and which are non-structural, and how they differ.

Figure 1. MSE and SSIM for different distortions

The simplest definition would be that structural distortions have an effect on shapes of objects in the image ant the non-structural distortions do not [1]. For e.g. distortions made by JPEG compression are structural and luminance shift is not structural. The index of structural similarity or SSIM is an implementation of the above concept in spatial domain and is function of the two images, or vectors, which can be denoted with **x** and **y** (like in (1) and (2)). One image is considered to be the reference and it is of perfect quality, while the second is distorted and the SSIM is computed to grade it. In accordance with previous conclusions about structural distortions the algorithm for calculating the SSIM can be broken down into three parts, which separately compare luminance, contrast and structure [1]. These three indexes are combined into an overall index. All the indexes are computed for an local window, rather than for entire image because statistical characteristics of image are not the same on every part of the image and the distortions also do not have to be the same on whole picture.

Another improvement of this method is the application of the Gaussian weighting function in the window that is slightly bigger than needed to get a smoother transition between the values of SSIM's and thus gets an index map, which is then combined into a global SSIM grade. The SSIM evaluates an image with scores from 0 to 1, where 0 is the lowest quality and 1 the highest (meaning that the picture is identical to the original). In Figure 1 is shown that SSIM better describes the real, subjective, image quality, unlike MSE, but not perfectly (for e.g. blurred image has a similar grade as the one with Gaussian noise and it is apparent that there is a significant difference in quality).

III. SUBJECTIVE IMAGE QUALITY ASSESSMENT

Subjective experience of the image quality is one of the central factors affecting the user experience and acceptance. The current methodologies for subjective assessment of the quality of television images are given in ITU-R Recommendation BT. 500-11. It describes in details the environmental settings, monitor settings, test and assessment procedure for standardized viewing trials.

According to tests methods, there are two common classes of assessments:

a. quality assessments - it establish the performance of systems under optimal conditions;
b. impairment assessments - it establish the ability of systems to retain quality under non-optimal conditions that relate to transmission or emission.

One of most used procedures for subjective quality evaluation is double-stimulus impairment scale (DSIS) method, also used in this work. It is a cyclic method in which viewer is firstly presented with an unimpaired reference, and then with the same picture impaired. In this work, the second picture is distorted by adding Gaussian noise, blur or using JPEG compression.

Measurement was taken in controlled environment on two monitors, and the same content was shown on each monitor to three observers. The distance between monitors and observers was equal to 3H, where H represents height of picture on monitor. Group of 16 non-expert young observers, who were not directly concerned with television picture quality as part of their usual work, were participated in assessment. Before each measurement, observers were introduced to the assessment method, distortion types, grading scale and timing trough set of training images. Type of display was 24" TFT s-pva with resolution of 1920x1080 @60 pixels.

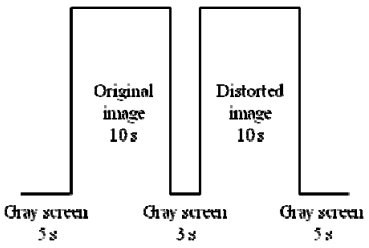

Figure 2. Image presentations

Pairs of pictures (original and distorted) were arranged into sequence of pictures, where each one had different distortion type and different distortion value. Between two images with same content, a gray screen was shown for 3 seconds. One pair of images was shown once in process of assessment and observers were asked to grade the second image, keeping in mind the first one. After watching one pair of images, gray screen was shown again and observers had 5 second to make their opinion on image quality. Figure 2 show image arrange and duration of their presentation.

Duration of one assessment, which included 48 pairs of images, was about 22.5 minutes. Complete test, with duration of training set included, lasted about 25 minutes.

The five-grade impairment scale was used in assessment (Table 1), and presenters explained meaning of the grades to the observers. Voting template contained table with list of presented images and blank field for grades.

TABLE I. FIVE-GRADE IMPAIRMENT SCALE

Grade	Description
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

A. Results of Subjective Tests

MOS (Mean Opinion Score) for distorted images was calculated using this formula:

$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{4}$$

where u_{ijkr} is score given by observer i, for test condition j, image k, repetition r and N is the number of observers.

We used scale 1-5 for 5 different degradation levels, so normalization to the same standard deviation for each observer wasn't necessary (all observers used in general whole scale). Finally, a scaled MOS value for each distorted image was computed by shifting MOS scores to the full range (0 to 100). Linear transformation is given as:

$$transformed\_MOS = -25 \cdot (MOS - 5) \tag{5}$$

before averaging across all subjects. MOS is transformed because it gives little better correlation results after nonlinear regression (described in next subsection). This is represented as scaled MOS in Table 2. SSIM and MSE objective measures are described before in Section 2.

Original pictures are shown in Figure 3. Distortions made on each image were adding Gaussian white noise, JPEG compression and blurring. Also each distortion on each picture was made in four levels as shown in Table 2. For Gaussian white noise number in column Level of distortion stands for the variance of noise (for every image the mean value was zero). Numbers in same column for JPEG distortion are the quality, or the level of compression, when the image is compressed. Range of quality goes form 0 to 100 where 0 is the lowest quality, maximal compression, and 100 the best quality or no compression at all. For human observer changes in quality are visible for quality lower than 40 and that was the reason for choosing levels of distortions noted in Table 2. Both Gaussian white noise distortion and JPEG distortion were added to original images using Matlab command imnoise. Images were blurred using open source program Gimp and levels of degradation are actually radius of blurring in pixels.

IV. RESULTS

A. Performance measures

To be able to compare different image quality measures and DMOS, we used two different measures of performance:

- Pearson's product-moment correlation coefficient;
- Spearman's rank-order correlation coefficient.

TABLE II. COMPARISON OF MOS, MSE AND SSIM MEASURES

Im.	Distortion	Amount of degradation	Scaled MOS	SSIM	MSE
Im1	Gauss Noise	0.0002	0	0.9809	5.76
		0.003	45.3125	0.8170	85.06
		0.01	65.6250	0.6473	273.45
		0.03	75	0.4709	757.09
	Blur	1.0	4.68750	0.9250	107.03
		1.5	10.9375	0.8547	197.00
		2.0	26.5625	0.7962	269.59
		3.5	43.7500	0.6618	437.25
	JPEG	10	51.5625	0.7253	280.48
		15	48.4375	0.7767	217.99
		20	39.0625	0.8093	180.65
		25	34.3750	0.8337	153.00
Im12	Gauss Noise	0.0002	4.68750	0.9710	5.56
		0.003	54.6875	0.7889	80.35
		0.01	65.6250	0.6412	256.74
		0.03	75	0.4951	718.68
	Blur	1.0	14.0625	0.9130	271.83
		1.5	34.3750	0.8340	477.60
		2.0	45.3125	0.7687	631.62
		3.5	56.2500	0.6270	942.77
	JPEG	10	71.8750	0.7427	557.93
		15	50	0.7940	442.49
		20	39.0625	0.8293	358.72
		25	29.6875	0.8554	293.67
Im16	Gauss Noise	0.0002	6.2500	0.9447	5.84
		0.003	60.9375	0.5585	85.83
		0.01	73.4375	0.2956	279.18
		0.03	59.3750	0.1364	789.62
	Blur	1.0	3.1250	0.9858	2.22
		1.5	9.3750	0.9740	3.66
		2.0	17.1875	0.9632	5.31
		3.5	35.9375	0.9349	9.80
	JPEG	10	79.6875	0.8299	30.70
		15	67.1875	0.8727	20.24
		20	64.0625	0.8969	15.32
		25	48.4375	0.9117	12.58
Im3	Gauss Noise	0.0002	9.3750	0.9592	5.59
		0.003	54.6875	0.6806	79.40
		0.01	64.0625	0.4673	254.99
		0.03	75	0.2970	709.08
	Blur	1.0	4.6875	0.9624	39.64
		1.5	25	0.9277	71.97
		2.0	28.1250	0.8971	98.94
		3.5	46.8750	0.8206	164.99
	JPEG	10	60.9375	0.7936	139.09
		15	59.3750	0.8386	107.77
		20	40.6250	0.8653	90.26
		25	40.6250	0.8832	78.04

Pearson's product-moment correlation coefficient is calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y}, \mathrm{i} = 1,..., n \tag{6}$$

where in Equation (6) x_i and y_i are sample values, (x are results for different objective measures and y are results for DMOS), $\bar{x}$ and $\bar{y}$ are sample mean, s_x and s_y are standard deviation (calculated using n - 1 in the denominator), Eq. (7-9):

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{7}$$

$$s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{8}$$

$$s_y = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{9}$$

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

Spearman's correlation coefficient is a measure of a monotone association that is used when the distribution of the data makes Pearson's correlation coefficient undesirable or misleading. Spearman's coefficient is not a measure of the linear relationship between two variables. It assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables [2].

B. Results - overall and for each type of degradation

Figures 4 and 5 show comparison between objective quality measures (MSE, SSIM) and subjective quality measure (MOS), for all degradations together, before and after linearization. We calculated Pearson's correlation coefficient before and after nonlinear regression. 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 [3]:

$$Q(x) = b_1 \cdot (\frac{1}{2} - \frac{1}{1 + e^{b_2 \cdot (x - b_3)}}) + b_4 \cdot x + b_5 \tag{10}$$

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. Coefficient parameters are given in Table 3.

Like proposed in paper [3], correlation coefficient is computed either by using measure directly or by its logarithm whichever gave better correlation results. By using this feature, MSE and PSNR give the same results if we compare $\log_{10}$(MSE) - MOS and PSNR - MOS, so results for PSNR were excluded from analysis.

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

- Trust-Region method [4];
- Levenberg-Marquardt method [5];
- Gauss-Newton method [6].

Final method for finding coefficients for nonlinear regression was the one which computed better results for performance measures (higher Pearson's and Spearman's correlation).

For each graph in Figs. 4 and 5 it is calculated overall Pearson's and Spearman's correlation coefficients, as well as for each type of degradation separately. They are presented in Fig. 6. Coefficients are compared with already known image database [7] and which correlation between MSE, SSIM and DMOS is calculated in [8]. In Fig. 6 gray bars denote our correlation after nonlinear regression and black bars correlation for other image database after nonlinear regression.

However, it should be noted that database [7] uses 5 different degradation types, while we use only 3 of them.

From the Fig. 6 it can be concluded that in general correlation between MSE and MOS is lower than SSIM and MOS, which follows results given in [8] (black bars in Fig. 6). Only for JPEG compression in our case SSIM measure gave poor results in comparison with MSE (Fig. 6 (g) and (h)). This could be because degraded JPEG images did not have bigger differences in degradation level (in Table 2, Image 3 had same MOS for JPEG degradation, quality 20 and 25), which also shows SSIM measure is not relevant for minor differences in degradation.

TABLE III. COEFFICIENT PARAMETERS FOR LOGISTIC FUNCTION

Degradation type	Measure	b_1 (95% confidence bounds)	b_2 (95% confidence bounds)	b_3 (95% confidence bounds)	b_4 (95% confidence bounds)	b_5 (95% confidence bounds)
Overall	MSE	-121.5 (-416.8, 173.8)	3.447 (-2.979, 9.873)	1.864 (1.685, 2.042)	76.32 (-35.8, 188.4)	-96.99 (-300.9, 106.9)
	SSIM	-6402 (-3.25e+005, 3.122e+005)	-1.283 (-24.46, 21.9)	0.3427 (-0.237, 0.9224)	-2039 (-6.719e+004, 6.312e+004)	766.1 (-2.068e+004, 2.221e+004)
Gaussian noise	MSE	-8490 (-7.169e+007, 7.167e+007)	1.871 (-68.87, 72.61)	5.863 (-4628, 4640)	45.46 (-145.2, 236.1)	-4273 (-3.585e+007, 3.584e+007)
	SSIM	-6475 (-2.692e+005, 2.563e+005)	-1.124 (-18.59, 16.34)	0.149 (-1.331, 1.629)	-1772 (-4.722e+004, 4.367e+004)	327.4 (-4105, 4760)
Blur	MSE	-42.22 (-85.15, 0.7026)	65.12 (-2.107e+021, 2.107e+021)	1.401 (-6.363e+018, 6.363e+018)	32.4 (9.607, 55.2)	-25.98 (-61.73, 9.776)
	SSIM	278.4 (-5.014e+004, 5.07e+004)	4.987 (-356.8, 366.8)	0.7247 (-0.8857, 2.335)	-443.4 (-3.84e+004, 3.751e+004)	364.3 (-2.674e+004, 2.747e+004)
JPEG	MSE	-9624 (-8.898e+005, 8.705e+005)	0.7848 (-24.01, 25.58)	1.893 (1.783, 2.003)	1833 (-1.112e+005, 1.149e+005)	-3415 (-2.174e+005, 2.106e+005)
	SSIM	-11.29 (-44.41, 21.83)	-2129 (-2.189e+006, 2.184e+006)	0.8688 (-2.606, 4.344)	-97.05 (-356.3, 162.2)	134.7 (-85.9, 355.3)

Figure 3. Tested images

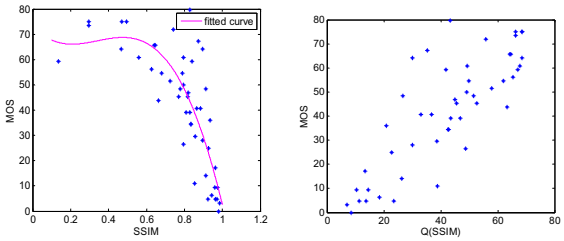

Figure 4. Comparison of all degraded images and $\log_{10}$(MSE) measure with MOS subjective measure, before (a) and after (b) nonlinear fitting

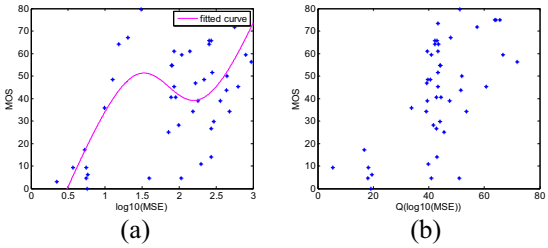

Figure 5. Comparison of all degraded images and SSIM measure with MOS subjective measure, before (a) and after (b) nonlinear fitting

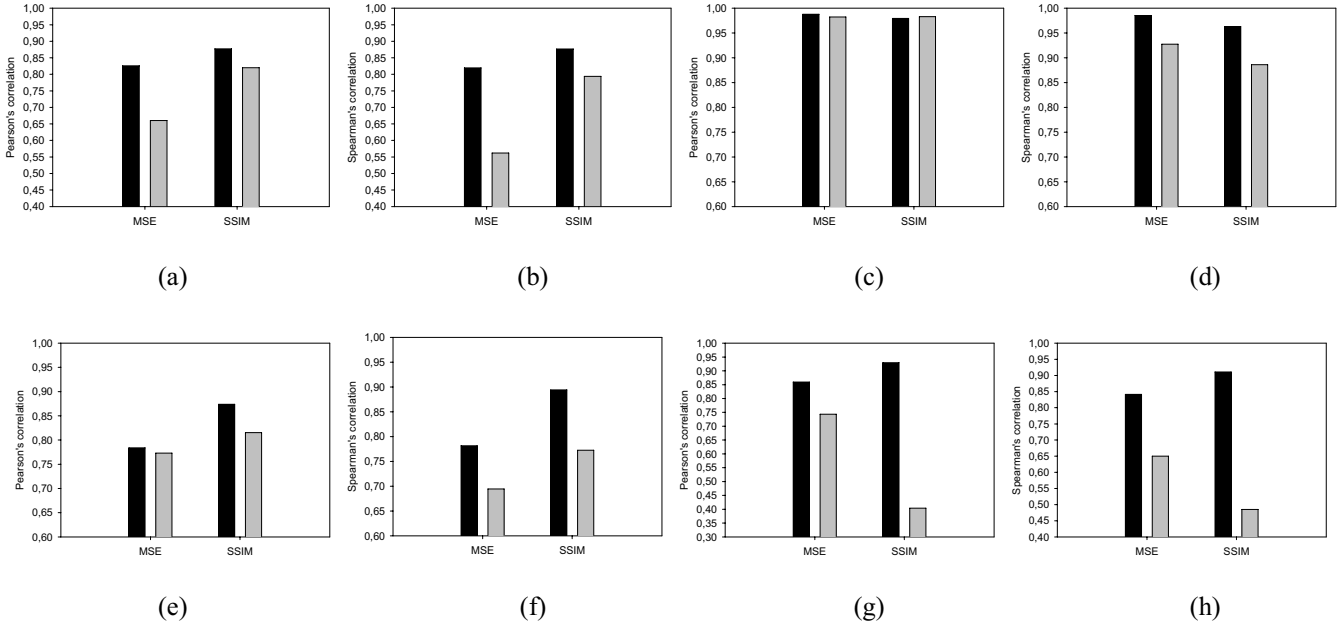

Figure 6. Pearson's and Spearman's correlation of MSE and SSIM measure with MOS after nonlienar regression; gray bars denote our correlation after nonlinear regression and black bars correlation for other image database after nonlinear regression: (a) Pearson's correlation for all degraded images, (b) Spearman's correlation for all degraded images, (c) Pearson's correlation for Gaussian noise degradation, (d) Spearman's correlation for Gaussian noise degradation, (e) Pearson's correlation for blur degradation, (f) Spearman's correlation for blur degradation, (g) Pearson's correlation for JPEG degradation, (h) Spearman's correlation for JPEG degradation

In Fig. 6 correlation is lower for our database (in comparison with database [7] which uses 5 degradation types) probably because we used for now only 4 images and only 3 types of degradation. Increasing the number of test images could result in more consistent correlation of objective and subjective measures. From Fig. 6, (a) and (b) it can be concluded that SSIM measure has in general better correlation with MOS. In Fig. 6, (c) and (d), it can be seen that MSE gives excellent results for Gaussian noise degradation, Spearman's correlation is somewhat better for MSE than for SSIM measure. Fig. 6, (e) and (f) show us that SSIM has better correlation for blur degradation.

V. CONCLUSION

In this paper we tested and calculated subjective quality measure (MOS) for three different distortions, each made by four different amount of degradation. Afterwards we compared it with two common objective quality measures, MSE and SSIM. Results show that in general SSIM provides better correlation with subjective measures, although it can give poor results if test images have minor degradation (JPEG degradation). MSE objective measure gives good results for Gaussian noise degradation. Future research could include more test images with more degradation levels, more degradation types as well as other objective measures.

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] Z. Wang, A.C. Bovik, Modern Image Quality Assessment. Morgan & Claypool Publishers, USA, 2006

[2] J. Hauke, 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

[3] 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, November 2006

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

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

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

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

[8] E. Dumic, S. Grgic, M. Grgic, "New Image-quality Measure based on Wavelets", Journal of Electronic Imaging, Vol. 19, No. 1, Article ID 011018, January - March 2010