STATISTICAL EVALUATION OF NO-REFERENCE IMAGE VISUAL QUALITY METRICS

Nikolay Ponomarenko(*), Oleg Eremeev(*), Vladimir Lukin(*), Karen Egiazarian(**)

(*) National Aerospace University, Kharkov, Ukraine
(**) Tampere University of Technology, Tampere, Finland

ABSTRACT

A task of no-reference visual quality metric verification is considered. A test set that contains 500 JPEG format images having different distortions is created. The results of experiments carried out by 316 volunteer observers to evaluate visual quality of images are presented. The experiments allowed obtaining mean opinion scores (MOS) based on averaging the evaluations. Several non-reference image visual quality metrics have been analyzed. Spearman and Kendall correlations between MOS and metrics values are calculated. It is shown that all analyzed metrics have not enough correspondence to human perception.

Index Terms— Image analysis, image processing

1. INTRODUCTION

No-reference visual quality metrics [1-4] are widely used in various applications of image and video processing, remote sensing, medical diagnostics, etc. First, they can be exploited for a preliminary estimation of image quality for image forming systems. A good example is an estimation of image quality inside a digital camera. Later, such an estimate can be used to inform a person taking this image that a formed image is of a non-satisfactory quality. In other applications, such an estimate can be incorporated into an automatic procedure of setting compression parameters of a lossy coder to be applied to a given image. Second, no-reference visual quality metrics can be used in content based image retrieval systems for image indexing and sorting.

There are also many other applications of no-reference visual quality, such as blind evaluation of noise characteristics [5], estimation of blur and contrast [6-9], blind evaluation of compression quality for jpeg images [10, 11], assessment of distortions and artifacts in images [12, 13] and so on.

Whilst for full reference visual quality metrics there are good image databases for metric verification (e.g., TID2008 [14], Live Database [15]), the design of no-reference metrics is more complicated because of absence of good representative image databases. This paper deals with description of a procedure we have used to form such a database and methodology to verify no-reference visual quality metrics. We have also analyzed different simple no-reference quality metrics in order to understand what image features (or distortions) are more important to human perception.

2. IMAGE SELECTION FOR THE TEST SET

There can be the following requirements to an image test set for no-reference metric verification:

1) The test set should include images with various characteristics containing homogeneous regions, details, textures (in particular, textures with different characteristics), etc.;
2) The test set has to contain all possible types of distortions typical for practice, such as blur, noise of different levels, brightness distortions, etc.;
3) Images are to be of different size in order to take into account subjective influence of an image size on its perception and quality evaluation by a human;
4) The test set should contain images with visual quality varying from very bad to very good and a histogram of such quality evaluation (in quantitative units) should be uniform enough;
5) The test set has to contain basic types of images that correspond to different modes of their forming; for example, it has to include portrait images and macro-mode images (characterized by blurred background that does not essentially influence quality perception), few composite images that contain added text, and others.

According to these requirements, 500 real life images in JPEG format have been finally selected from a considerably larger number of images. The image size varied from 300x400 (minimal size) to 400x600 pixels (maximal size). On one hand, such image size allows placing two images at computer monitor simultaneously without scaling them (this is needed for comparing their quality, see below). On the other hand, image area differed by up to 2 times, which allows to take into account the influence of image size while estimating its visual quality by a metric and analyzing this factor.

While selecting images for our test set, their visual quality has been preliminarily evaluated by experts using a scale with five gradations. To provide an appropriate representation, we have selected 50 images of «very bad»

quality, 75 of «bad», 100 of «middle», 125 of «good» and 150 images of «excellent» qualities. As it is seen, there is some non-uniformity: less bad than good quality images have been selected. The reason is that in practice there are less bad than good quality images, so we expect that it would be enough to have 50 «very bad» quality images for obtaining enough statistics. The test set includes images formed in good and bad illumination conditions, by portrait and macro modes, in night conditions. There are images with visible and practically invisible noise, sharp and blurred ones. The test set includes several artificially synthesized images (computer graphics) as well as composite ones.

For all images, their original format (JPEG) was kept. At the same time, we tried to avoid selecting images containing a lot of semantic information (human faces, text, etc.) able to attract special attention of observers and, thus, to distort evaluation of image visual quality.

3. PERFORMING EXPERIMENTS TO FORM MOS

To obtain MOS, observers have been offered different image pairs (double stimuli comparisons). Each time, for each pair of images, an observer had to choose an image of a better visual quality. As a result of each pair-wise comparison, such (better) image got one point, and a worse image got 0 points, respectively.

During one experiment, each image has appeared at monitor screen 11 times. Thus, in aggregate, it could get from 0 to 11 points. This determines the used range of possible values for formed estimates of visual quality.

To evaluate visual quality of all 500 images, a human needs to analyze 250 x 11 = 2750 image pairs. If a human spends 2…3 seconds for each comparison, this takes about 2 hours for each experiment. However, according to existing recommendations to methodology of image quality assessment experiments [16], a total time of one experiment should not exceed 20…30 minutes. If this limit is exceeded, participants start to feel tired and efficiency of such experiments radically drops down.

Keeping this in mind, each participant was offered only 60 images taken randomly from the total number of 500 images. Then, the number of analyzed pairs for each experiment participant was 30 x 11 = 330 pairs. For 2…3 seconds spent for each comparison, the total time was 10…15 minutes for each experiment. In fact, depending upon individual, the experiment time varied from 10 to 25 minutes.

The obtained quality estimates have been then averaged for all participants of our experiments (from 30 to 50 participants for each image in our database) and MOS has been calculated. Currently, 316 volunteers have participated in our experiments. Certainly, before starting they have been carefully instructed. Visualization and observation conditions varied in reasonable limits to be comfortable for each participant. Different monitors were used, both LCD and CRT, mainly 19" with the preset resolution 1152x864 pixels.

Validation of experimental data obtained has been carried out according to recommendations of the corresponding standard [16]. An estimate is considered abnormal if its value differs from the mean estimate by more than 2.33σ where σ is a standard deviation of the estimate obtained for all experiments. If for an experiment the number of such abnormal estimates is larger than 10%, then entire experiment is considered wrong. According to these rules, 14 experimental results have been rejected. Total percentage of rejected abnormal estimates was 6.3 %.

Distribution of MOS on images is presented in Fig.1 ("bad quality" images have smaller indices). Fig. 2 shows the histogram of MOS values (larger MOS corresponds to better visual quality).

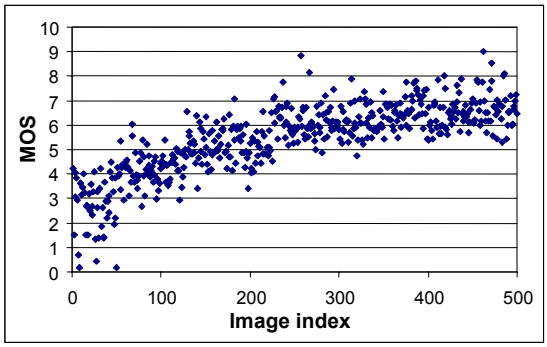

Fig. 1. MOS values for all images

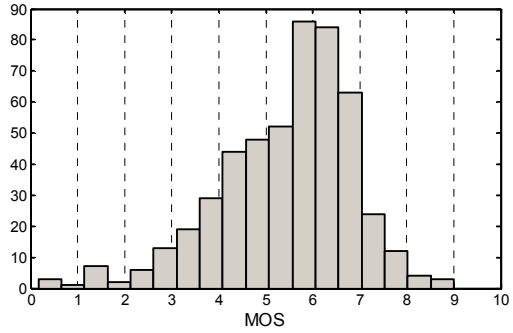

Fig. 2. Histogram of obtained MOS values

As it is seen in Fig. 1, there is an obvious correlation between preliminary estimation of image visual quality (carried out by few experts at the stage of test set forming) and the finally obtained MOS. However, the scatter-plot contains few outliers, i.e. preliminary estimates considerably differ from the corresponding obtained MOS. This indirectly approves a necessity to obtain average opinions on image visual quality in order to provide reliable MOS.

Averaged MOS and its variance for all images are equal to 5.46 and 1.17, respectively. Relative variance is

equal to 0.039. Quite small value of relative variance of MOS estimates indicates on their high confidence and their applicability in verification of visual quality metrics. Table 1 contains the aggregate information about the test set and subjective experiments.

Table 1. Test set and subjective experiments

№	Main characteristics	Value
1	Number of images	500
2	Number of experiments carried out	316
3	Methodology of visual quality evaluation	Pair-wise sorting (choosing the best visual quality between two considered images)
4	Number of elementary evaluations of image visual quality in experiments	86900
5	Scale of obtained estimates of MOS	0..11
6	Variance of estimates of MOS	1.17
7	Normalized variance of estimates of MOS	0.039

4. VERIFICATION OF DIFFERENT NO-REFERENCE VISUAL QUALITY METRICS

Fig. 3 presents the block diagram of database use for estimation of quality metric efficiency (adequateness). As it is seen, efficiency of a metric is characterized by a rank correlation between MOS values and metric values. As rank correlation factors, it is possible to use Spearman and Kendall correlations [17].

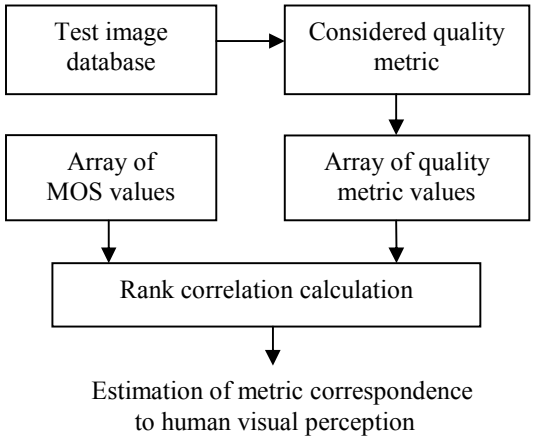

Fig. 3. Block diagram of verification process for metrics of image visual quality

We have analyzed here some basic no-reference metrics that correspond to evaluation of different image quality distortions (or image features). One example of such a distortion is a blur. Therefore, we have included in the study of a good metric of its blind evaluation (JNB) [18].

Another distortion is a level of blocking artifacts caused by a lossy JPEG compression. Therefore, we have included in our analysis a good metric for evaluation of such distortions (NR JPEG) [19].

One should not forget that sizes of compressed JPEG-images can be use as quite good estimates of their visual quality. Small size indicates, most likely, about a poor visual quality and vice versa. Therefore, we have included in our analysis such a metric as a file size.

Other important features relevant to the assessment of image visual quality are image details. As a measure to evaluate image details here we have used the average of local variance in the image that is calculated in blocks of 3x3 pixels. We have analyzed two versions of such metric. The first one is calculated on the component Y in YCbCr color space (local variance for Y component), while the second one is calculated on both Cb and Cr color components (local variance for Cb and Cr components).

Also, a robust estimate of a range of pixel values of an image have been analyzed here as a metric of the image visual quality (range of pixel values for Y component). This metric is calculated as a difference between values of a pixel with 99.5% index (quantile) in ordered sample of image pixels (in ascending order) and pixel with 0.5% index.

Finally, we have analyzed JPEG blockiness metric (NRBM) [20] and squared gradient (SG) [21].

Table 2 contains values of Spearman and Kendall rank correlation coefficients between MOS and considered metrics.

Table 2. Correlations with MOS of considered metrics

Metric	Spearman correlation	Kendall correlation
Local variance for Cb and Cr components	0.598	0.422
File size	0.575	0.410
Local variance for Y component	0.419	0.291
JNB	0.315	0.218
Range of pixel values for Y component	0.298	0.204
NR JPEG	0.186	0.126
NRBM	0.130	0.092
SG	0.513	0.356

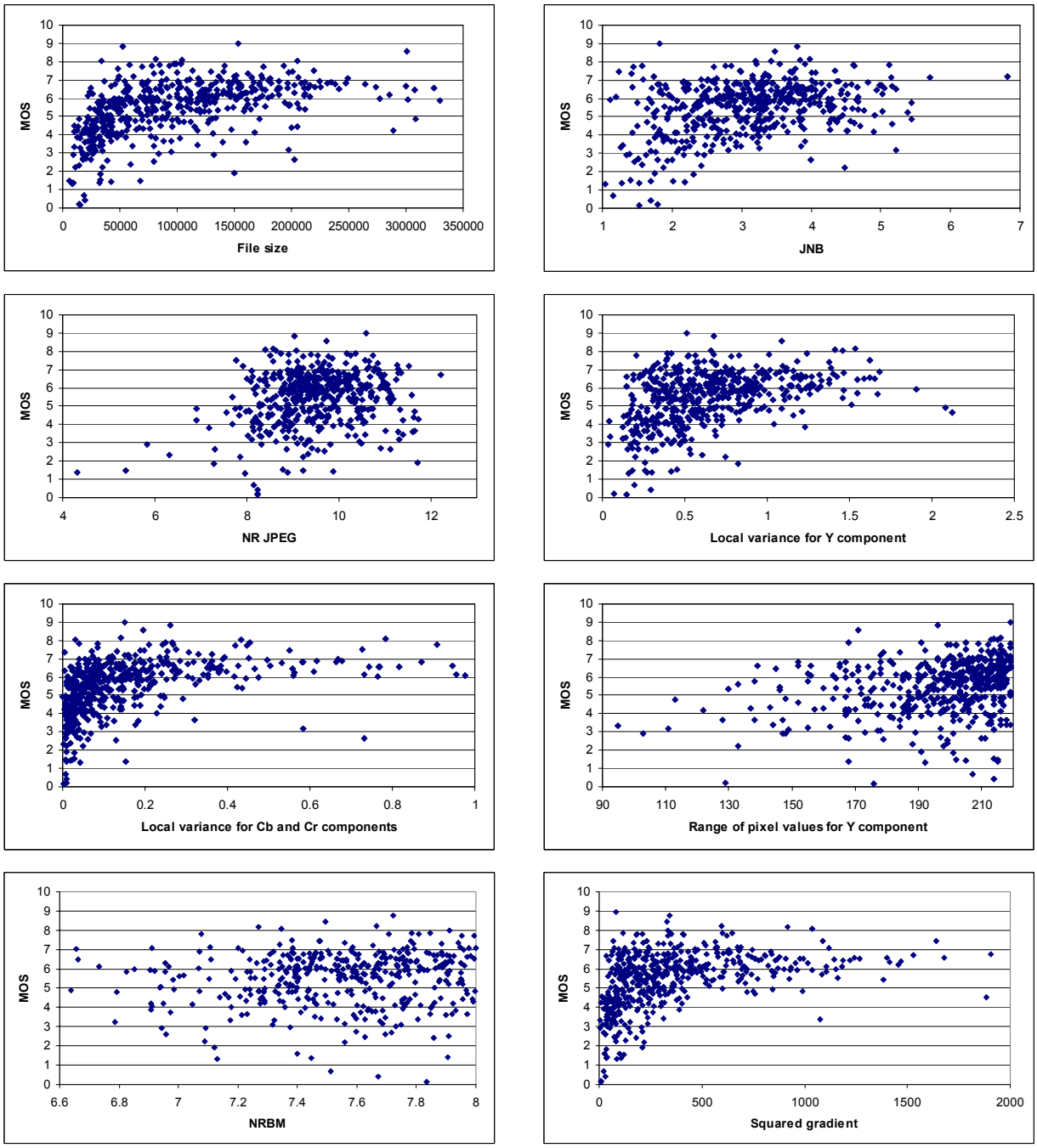

Fig. 4. Scatterograms of dependences between MOS and different no-reference metrics

Let us give a few comments to the data obtained. First of all the above metrics correlations with the human perception are far from being satisfactory. This illustrates that the task of evaluating no-reference image visual quality is not an easy one. This problem might be solved by taking into account simultaneously many simple metrics, using, e.g. a neural network.

It is also clear that it is ineffective to analyze only Y component (brightness), since the best metric in Table 2 is one that takes into account local variance just in color components.

It is obvious that test sets like LIVE database or TID2008 can not be used for evaluation of no-reference metrics. These test sets contain too many images of the same type of distortions and too many images that are very similar

(have the same etalon image). It is necessary to create a huge test sets of real life images that contain many (tens, hundreds) different types of distortions as well as their combinations.

Often in practice real systems of evaluation of no-reference image visual quality may be deceived by artificially embedding noise in images. For example, it may be done for decreasing of value of blind blur estimation on such image. Such nuances must be taken into account in the development of sets of test images for evaluation of no-reference image visual quality metrics.

5. CONCLUSIONS

The paper describes the test set of images intended for efficiency analysis of no-reference visual quality metrics. The already obtained MOS values possess rather high statistical confidence in the sense of correspondence to image perception and quality evaluation by humans. In the future, we plan to take into account new experiments and to correct MOS values. It is demonstrated how the created test set can be easily exploited in quality metric verification. It is shown that simple quality metrics have very low correlation with MOS. To develop a good metric it is necessary to take into account several such metrics simultaneously, which can be done, for example, by using neural networks.

6. REFERENCES

[1] H. R. Sheikh, A. C. Bovik, L. K. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000". *IEEE Transactions on Image Proc.*, Vol. 14 (11), pp. 1918-1927, 2005.

[2] S. Suresh, R. Venkatesh Babu, and H.J. Kim, "No-reference image quality assessment using modified extreme learning machine classifier", *Applied Soft Computing*, Vol. 9 (2), pp. 541-552, 2009.

[3] M. Jung, D. Léger, and M. Gazalet, "Univariant assessment of the quality of images", *J. Electronic Imaging*, pp. 354-364, 2002.

[4] J. Redi, P. Gastaldo, R. Zunino, "Hybrid Neural Systems for Reduced-Reference Image Quality Assessment". *ICANN 2009*, pp. 684-693, 2009.

[5] V. Lukin, N. Ponomarenko, S. Abramov, B. Vozel, K. Chehdi, "Improved noise parameter estimation and filtering of MM-band SLAR images", *Proceedings of the Sixth International Kharkov Symposium "Physics and Engineering of Millimeter and Sub-Millimeter Waves"*, Vol. 1, pp. 439-441, 2007.

[6] P. Marziliano, F. Dufaux, S. Winkler, "A no-reference perceptual blur metric", *IEEE 2002 International Conference on Image Processing*, Vol. 3, pp. 57-60, 2002.

[7] P. Marziliano, F. Dufaux, S. Winkler, "Perceptual Blur and Ringing Metrics: Application to JPEG2000", *Signal Processing: Image Communication*, Vol. 19, №.2, pp. 163-172, 2004.

[8] J. E. Caviedes, S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis", *Proceedings of ICIP2002*, Vol.3, pp. 53-56, 2002.

[9] R. Ferzli, L.J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)", *IEEE Trans. on Image Proc*, Vol. 18, Issue 4, pp. 717-728, 2009.

[10] Y. Horita, S. Arata, T. Murai, "No-reference image quality assessment for JPEG/JPEG2000 coding", *Proceeding of EUSIPCO*, Vol.2, pp.1301-1304, 2004.

[11] I. Hontsch, L.J. Karam, "Adaptive image coding with perceptual distortion control", *IEEE transactions on image processing*, Vol.11, № 3, pp. 213-222, 2002.

[12] M.C.Q. Farias, S.K.Mitra, "No-reference video quality metric based on artifact measurements", *IEEE International Conference on Image Processing*, Vol. 3, pp. 141-144, 2005.

[13] A.C. Bovik, S. Liu, "DCT-domain blind measurement of blocking artifacts in DCT-coded images", *Proceedings of the Acoustics, Speech, and Signal Proc.*, Vol.3, pp. 1725-1728, 2001.

[14] 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, pp. 30-45, 2009.

[15] H.R. Sheikh, M.F. Sabir, F.C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms", *IEEE Trans. on Image Proc.*, Vol. 15, no. 11, pp. 3441-3452, 2006.

[16] Methodology for Subjective Assessment of the Quality of Television Pictures Recommendation BT.500-1. – Geneva: ITU, 2002, 48 p.

[17] M.G. Kendall, "The advanced theory of statistics", Vol. 1, London: Charles Griffin & Company limited, 1945, 457 p.

[18] R. Ferzli, L.J. Karam, "A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB)", *IEEE Transactions of Image Processing*, Vol. 18, Issue 4, pp. 717-728, 2009.

[19] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," *IEEE International Conference on Image Processing*, Vol. 1, pp. 477-480, 2002.

[20] R.V. Babu, A. Bopardikar, A. Perkis, "A perceptual no-reference blockiness metric for JPEG images", I, pp. 455-460, 2004.

[21] A.M. Eskicioğlu, P.S. Fisher, "Image Quality Measures and Their Performance", *IEEE Transactions on Communications*, 43(12), pp. 2959-2965, 1995.