

Internet Based Subjective Assessment of Image Quality Experiment

Tomasz Grzywalski, Adam Łuczak, and Ryszard Stasinski

Department of Electronics and Communications

Poznań University of Technology

Poznań, Poland

aluczak@multimedia.edu.pl, rstasins@et.put.poznan.pl

Abstract— In the paper an Internet based massive experiment devoted to subjective assessment of images quality is described. The investigations have been based on DSIS method with some modifications induced by the nature of Internet poll. Special measures have been undertaken for removing results given by incompetent experiment participants. The huge number of tests allowed for multi-aspect image quality evaluation: 12 deformation effects have been applied to 22 images. The results point at important problems linked with automation of image quality assessment: factors determined by high levels of human visual cognition (e.g. face deformation), surprisingly high impact of local distortions on scores for some tested images, finally, correct modeling of low-level effects, like masking.

Keywords—subjective and objective image quality assessment; DSIS; Internet poll

I. INTRODUCTION

Formulation of an objective measure for image quality evaluation that is fully compatible with human's assessment would be a breakthrough in testing of visual systems quality. The list of papers covering the subject would fill the space of this work, let us cite only few recent papers describing very different approaches [1], [2], [3]. The main problem with automatic image quality evaluation is that algorithms work well only for some particular types of image impairments, hence, it is still a need for research that is enhancing our comprehension how a human cognition of artifacts works.

In the paper the realization of an idea of multi-aspect while low cost and statistically meaningful subjective image quality assessment experiment is reported – by Internet poll. The approach necessitated adaptation of existing standards, and special measures needed for rejecting useless data, see sections 2 and 3. The provided in section 4 results of experiments show how important is image contents for recognition of an image deformation as strongly disturbing, or not that much. Probably the most interesting conclusion for researchers working on automatic systems is potentially high importance of impairments localized only in small parts of evaluated images.

II. INTERNET POLL

For obtaining valuable results participants of Mean Opinion Score (MOS) investigations should be found among ready to cooperate people having good eyesight. There is no lack of

such people on Internet, however, there are also few 'jokers' encouraged by their anonymity. Secondly, there are limits to the good will of those who want to cooperate, they are anonymous, too. Both factors complicate MOS investigations by Internet, nevertheless, the extra effort may be rewarded by collecting multi-aspect while statistically meaningful result database at minimal cost.

As there was not any standard concerning still image MOS assessment, the poll procedure was based on ITU-R standard for evaluating video sequences [4], method DSIS (Double-Stimulus Impairment Scale), variant I, with some modifications:

- A participant was informed about the goal and procedure of the test, and meaning of notes (defined in standard 5-note scale). There was no example of image evaluation, on the other hand, some hints how to optimize viewing conditions were given (including monitor calibration).
- The investigation consisted of 15 cycles, each divided into 4 periods: original image presentation (10s), 3 second break, distorted image show (10s), finally, unlimited time for undertaking decision (deviation from standard requirement).
- 13 image pairs were drawn by chance from a database of 22 images, 12 types of distortions were applied. Images did not repeat. For simplifying detection of 'jokers' two additional image pairs were always shown (in randomly chosen cycles): MAX, consisting of two originals, and MIN with a particularly heavily distorted copy, see section 3.
- All notes but those for MIN and MAX pairs were taken into consideration in statistical analysis. The MIN and MAX pairs were used for detecting flawed tests.

The database of 22 original images is provided in Table 3. As can be seen, important effort was undertaken to collect a set of widely differing genres of photographs. Table 1 contains definitions of impairments illustrated by deformed copies of selected images.

The poll was realized using an Internet page (now defunct). It was not necessary to log-in to the page, hence, great effort was done to minimize the chance that a participant would

evaluate the same sequence of images twice, this was done practically impossible for users of Internet Explorer and Mozilla FireFox. The investigation was blocked when image viewing window was too small, for guaranteeing proper timing a cycle would not start until both images from a pair were collected in computer memory. Finally, it was necessary to undertake special measures for avoiding multiple registrations of note records in the investigations database. Detailed analysis of data collecting process suggests that the described above safety mechanisms have worked perfectly.

Testing of the system started on November 20, 2007, first 10 days in well-controlled lab environment. The working Internet page was open to public since the end of February. Due to timing of advertising campaign on carefully chosen Internet fora all but 12 results were obtained till the beginning of May 2008. Till that time there were 713 correctly concluded visits to the page resulting in 9269 notes, summary of their statistical analysis is provided in Table 2. The least frequently viewed image-impairment combination was evaluated 30 times, more than enough for making its note statistically meaningful.

III. PREPROCESSING OF RESULTS

The first step in result processing was checking for tests credibility. Namely, the test was rejected if image MAX got less than 4 points or image MIN more than 2. The deteriorated image of MIN pair has had so poor quality that a note higher than 2 suggested that the participant wasn't very attentive during the test, at best, see Table 3, the last position. The MAX pair consisted of two originals, the image has not been particularly attractive. This choice has not been done without a reason, in this way participants who did not understood that the experiment consisted in image pair comparison (and not e.g. distorted image evaluation) have been eliminated. In this way 37 tests have been rejected, only 3 of them were indisputably malicious (both rejecting criterion met).

Another investigation removed all tests differing too much from the average. Namely, the test deviation measure has been introduced:

$$D_n = \frac{1}{13} \sum_{i=1}^{264} (q_{i,n} - Q_i),$$

where $q_{i,n}$ is n -th participant measurement for i -th image pair, and Q_i is the mean measurement for the i -th image pair, if an image pair wasn't shown to a participant, the difference was set to zero (only 13 of 264 image pairs have been shown). It appeared that for 676 tests under consideration the mean of D_n values was 0.6624, while their standard deviation 0.1901. Due to some simple data model, if notes were taken in a completely random manner, the expected value of D_n would be close to 1.5, with deviation slightly smaller than 0.3. This observation prompted us to reject all tests for which D_n was greater than 1, which diminished the set by next 36 series. In contrast to MIN/MAX criterion this one probably affected results and led to slight decrease in credible measurements variance. On the other hand, we feel entitled to claim that the impact of non-cooperating participants on the statistical value of our results has been negligible, if any at all.

IV. RESULTS

MOS scores obtained for different distortions are summarized in Table 2. All scores have been higher than that for MIN pair (1.04), while few scores appeared to be higher than that for MAX pair (4.65). This is not very surprising as the MAX image hasn't been very appealing, while in fact the score 4.86 for effect 7 (RGB components normalization) has been linked with a pair of identical images (RGB components of image 5 have been already normalized).

Probably the most interesting results have been obtained for geometric deformations of images, both global (effect 6), and local (effect 10). In both cases the impact of distortion has been heavily dependent on an image contents. Two factors lowered the scores dramatically: clearly visible smooth lines, and presence of a human face. The factors refer to higher levels of human visual cognition, and this fact leads us to an important question, if automatic image quality assessment algorithms can take into account such effects, and if yes, where are their performance limits in this respect?

A brief analysis of several automatic image quality assessment methods shows that they seem blind to local image impairments (parameters averaged over the entire image are usually used). On the other hand, results for effects 2 (local Gaussian blur), 4 (black spot), and for commented above effect 10 show that they have quite important impact on MOS scores, which is particularly well visible when comparing those for local and global Gaussian blur (effect 1). What makes automatic evaluation of their impact even more difficult is their strong dependence on image contents (high variation of results).

Effects 3 and 5 (reduction of color number and high ISO noise) were particularly well visible in low-frequency image parts (wide smooth plains), hence, some images have been strongly affected by them, while some not. Their link with masking effect is quite clear.

Results for other effects have been almost image-independent and not very surprising. Relatively big variance of results for RGB components normalization (effect 7) reflected the fact that some images changed quite a bit after the normalization, while others not that much.

V. CONCLUSIONS

An Internet based massive experiment on subjective evaluation of images quality is described in the paper. The research has been based on modified DSIS method, followed by data preprocessing undertaken for removing unreliable test results. A multi-aspect image quality evaluation has been done: impact of 12 deformation effects on quality of 22 images has been measured. Some conclusions have been drawn concerning factors linked with high levels of human visual cognition (e.g. susceptibility to face deformation), and surprisingly high importance of local distortions on scores for some tested images. Further work will be aimed at testing coherence of obtained results with those given by some automatic image quality evaluation methods.

REFERENCES

- [1] H.R. Sheikh, A.C. Bovik, "Image information and visual quality," IEEE Trans. Image Proces., Vol.15, No. 2, pp. 430-444, 2006.
- [2] E. Girschtel, V. Slobodyan, J. Weissman, A. Eskicioglu, "Comparison of three full-reference color image quality measures," 18-th IS&T/SPIE Ann Symp. Electron. Imaging, Image Quality and Syst. Perform., Proc. SPIE 6059, 2006.
- [3] Wu Dong; Qian Yu, Zhang, C.N, Hua Li, "Image Quality Assessment Using Rough Fuzzy Integrals", Distributed Computing Systems Workshops - Supplements, ICDCSW 2007. pp. 1-5, 2007.
- [4] ITU-R, Recommendation BT.500-11, Methodology for subjective assessment of the quality of television pictures, 2003.

TABLE I. DEFORMATION EFFECTS APPLIED TO IMAGES





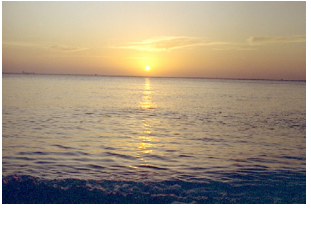


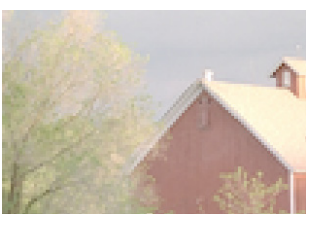

			
1. Gaussian blur	2. Gaussian blur applied to 9% of image area near center	3. Reduction of color number to 20	4. Black spot added, bottom-left corner
			
5. Digital high ISO noise	6. Geometric distortion	7. RGB components normalization	8. Slight modification of IHS parameters
			
9. Medium JPEG compression (quality=45)	10. Geometric distortion close to image center	11. Strong pixelization	12. Strong JPEG compression (quality=10)

TABLE II. MEAN, MINIMUM, MAXIMUM MOS SCORES, AND THEIR STANDARD DEVIATIONS FOR 12 DEFORMATION EFFECTS APPLIED TO TEST IMAGES

Effect	1	2	3	4	5	6	7	8	9	10	11	12
Minimum	1.78	1.57	1.42	3.10	2.65	3.53	2.76	3.97	3.27	1.69	1.29	1.17
Mean	2.25	2.36	2.22	3.77	3.34	3.24	4.23	4.32	3.82	3.85	1.64	1.74
Maximum	3.04	4.18	4.18	4.76	4.49	4.50	4.86	4.67	4.15	4.67	2.13	2.52
Std. dev.	0.34	0.70	0.60	0.58	0.50	0.97	0.44	0.20	0.23	0.78	0.21	0.33

TABLE III. IMAGES USED IN TESTS, THREE LAST ONES IN THE LAST ROW: MAX, MIN, AND MIN DISTORTED

