KAPITEL 7 / *CHAPTER 7* [7]
THE ANALYSIS OF SUBJECTIVE METRICS AND EXPERT METHODS FOR IMAGE QUALITY ASSESSMENT

DOI: 10.30890/2709-2313.2023-18-03-031

Introduction

The development of the digital age is characterized by a non-linear accumulation of data volumes. For effective data processing, it is necessary to increase the hardware capabilities and develop high-speed algorithms. When implementing high-speed algorithms, the completeness of presenting the features of real entities is lost. For example, when graphically rendering an object, its lighting features may not be accurately presented [1]. Since the visual channel of information perception [2] is one of the most important, it is necessary to develop new effective algorithms and models of computer graphics. The developed highly efficient models and methods should provide the smallest possible loss of object visualization quality. Special metrics are used to determine the level of visualization quality. The advantage of subjective methods and metrics for assessing the quality of formed images is their closeness to human perception. Therefore, it is important to systematize subjective methods and metrics for image quality assessment.

7.1. The Current State of Subjective Image Quality Assessment

There are two classes of metrics for testing the quality of object visualization: subjective and objective. Objective quality metrics are calculated using the data extracted from the images. The test image is compared with the reference image. The subject of comparison can be differences in color intensities, structural patterns, noise levels. The most popular objective metrics [3] of image quality are RMSE, MNSE, PSNR, UIQ, SSIM. Subjective image quality metrics are assessed by humans. Usually, experts with the relevant knowledge and experience or future users of graphics systems give quality scores. If experts are used, the metrics are called expert metrics. The two main types of subjective image quality metrics are single-stimulus and double-stimulus metrics. Single-stimulus metrics [4] are used when only the test image is shown for evaluation. Double-stimulus metrics [4] are used when a test and a reference image are shown for comparison. The main single-stimulus subjective quality metrics of images

[7] ***Authors:*** *Romanyuk Olexandr N., Romanyuk Oksana V., Titova Nataliia V., Zavalniuk Evgeny K., Romanyuk Sergey O.*

[4,5] are MOS, SSCQE. The main double-stimulus subjective quality metrics [4, 5] are DMOS, DSIS, DSCQS. In addition to metrics, special methods of expert evaluation are used to subjectively determine image quality. The methods of ranking, preferences, pair-wise comparison, sequential comparison, direct evaluation [6] are used to organize a set of images by quality, and calculate weight coefficients of quality characteristics. Commission and Delphi methods can be used to discuss the level of image quality in special circumstances. Another important direction is to determine for what percentage of people the quality of scene visualization was acceptable. For this $\theta -$Acceptability and Acceptance metrics are used.

In the paper [7] (Hosfeld et al. 2016), an analysis of single-stimulus quality metrics and quality acceptability metrics was carried out. No information is provided about double-stimulus metrics and expert methods. In the paper [8] (Kumar et al. 2014), only the main single-stimulus metrics are considered. The work [6] (Velychko et al. 2015) provides an analysis of general expert evaluation methods that can be used for image quality assessmnet. The ITU-R BT 500.9 standard [5] describes only single-stimulus and double-stimulus quality metrics for television images. Therefore, there is a need for a more complete analysis of subjective metrics and methods for image quality assessment.

7.2. The Overview of Subjective Metrics and Methods for Image Quality Assessment

MOS (Mean Opinion Score) [8] is a metric for subjective assessment of image quality. A panel of experts examines the image and assigns a quality rating on a discrete scale, usually containing five categories. The given grades are translated into a numerical scale of values on the interval (table 1) $[1,5]$.

Table 1 – The correspondence of scores and image quality levels

Quality Level	Bad	Poor	Fair	Good	Excellent
Score	1	2	3	4	5

At the end, the mean value of grades is calculated according to the formula [8]

$$MOS = \frac{1}{n} \cdot \sum_{i=1}^{n} s(i),$$

where i – expert number, n – number of experts, $s(i)$ – image quality score from the $\mathrm{i-th}$ expert.

In addition to the five-category scale, scales of other dimensions can be used. It is believed that the dimensionality of the scale [9] does not significantly affect the assessment results. The metric is simple and widely used for quality assessment, but the same mean value of image quality scores may correspond to significantly different statistical distributions [7]. Another problem is the difficulty of choosing the limit value of quality [9]. The choice of the MOS threshold value may depend on the type of images and the commercial purpose of the assessment. The mean value of the ratings does not reflect the percentage of experts who gave the image at least a satisfactory rating. The metric, which considers only the assignment of image quality categories by experts, is called Absolute Category Rating (ACR) [4].

DMOS [4] (Difference Mean Opinion Score) is a modification of MOS and lies in comparing the quality score of the test image relative to the reference image. DMOS is calculated using the formula

$$DMOS = \frac{1}{n} \cdot \sum_{i=1}^{n} rs(i) - s(i),$$

where $rs(i)$ – reference image quality score from the $\mathrm{i-th}$ expert.

The Z-metric is the normalization of image quality scores. It is calculated according to the formula [8]

$$Z = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{s(i) - \overline{s}(i)}{\sigma(i)},$$

where $\overline{s}(i)$ – mean image quality score from the $\mathrm{i-th}$ expert, $\sigma(i)$ –standard deviation for the $\mathrm{i-th}$ expert's scores.

For DMOS, when calculating the Z-metric [4], instead of image quality values, quality scores differences are taken into account.

The $\theta-$Acceptability (A_θ) metrics [7] shows the probability that the image quality exceeds the threshold value. It is calculated according to the formula [7]

$$\mathrm{A}_\theta = \frac{1}{n} \left| \{ s(i) \geq \theta : i = 1,\dots,n \} \right|,$$

where θ – minimal score of acceptable quality, n – number of experts.

The $\theta-$Acceptability metric is related to the Acceptance metric [7]. The image quality scores are replaced by 1, if their values are not less than the limit. Otherwise, scores are replaced by 0. The metric is the case of $\theta-$Acceptability with $\theta = 1$. The value of the Acceptance metric is calculated using the formula [7]

$$\hat{f}_1 = \frac{1}{n} \sum_{i=1}^{n} \delta_{s(i),1},$$

where $\delta_{U_i,1}$ – Kronecker delta (equals 1 if $s(i) = 1$, otherwise equals 0).

The "Good-Or-Better" (GoB) [7] metric shows the percentage of satisfied feedbacks on the quality of the object relative to the subjective score. A scale of objective quality values is used (scale values can be translated into the MOS scale using the formula [7]). A quantile is selected from the scale, at which the percentage of experts satisfied with the quality of the object is 50. The quantile is selected based on the results of a series of tests. During the quality evaluation of audio recordings, for example, the quantile is 60. Then, the quantile value is used to calculate GoB using the formula [7]

$$GoB = 100E(\frac{R-60}{16}),$$

where $E(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$, R – objective object quality score, 60 – value of quantile.

Similarly, the metric of the percentage of unsatisfied feedbacks regarding the quality of the object "Poor-Or-Worse" is calculated according to the formula [7].

$$PoW = 100E(\frac{45-R}{16}).$$

This method was developed for evaluating the audio quality. However, the method can be modified for other tasks [7], in particular, image quality assessment. For this, it is necessary to make a number of changes. It is necessary to choose an objective coefficient of image quality, normalize the value of the coefficient to the scale $[0,100]$, calculate the value of quantiles using experimental results and modify the formula of the dependence between MOS and the objective coefficient. As a result, the dependence between the values of MOS and GoB/PoW will be obtained.

DSIS (Double Stimulus Impairment Scale) [5] is used to evaluate the level of image impairment relative to the reference image. For about 10 s, experts are shown a reference image on the screen. A gray background is displayed for 3 seconds. For 10 seconds, an image is displayed for evaluation. After that, experts are given 5 – 11 s to assess image impairment [5] (figure 1). A reference image for comparison is recalled by experts. In another variant of the method, the evaluation is performed during repeated exposure of the test and reference images.

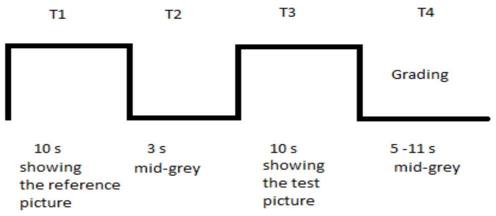
T1
T2
T3
T4
Grading
10 s showing the reference picture
3 s mid-grey
10 s showing the test picture
5 -11 s mid-grey

Figure 1 –DSIS calculation steps (one of the variants)

The impairment evalution scale, as a rule, includes five categories [5] (table 2).

Table 2 – The correspondence of scores and image impairment level

Impairment level	Very annoying	Annoying	Slightly annoying	Perceptible, but not annoying	Imperceptible
Score	1	2	3	4	5

After the evaluation, the mean value of the image impairment score is calculated.

The DSCQS metric (Double-Stimulus Continuous Quality Scale) [5] lies in using a continuous interval of values to evaluate the quality of images. A pair of images are displayed on the screen for experts. One is reference, the other is modified. Experts do not know which image is the reference image (experts may subconsciously increase the score of the reference image). The score is provided for both images. A scale of continuous values is used, which is divided into 5 levels [5] (figure 2). Marked continuous scale values are transformed into range [0,100] values. The difference of images scores is calculated.

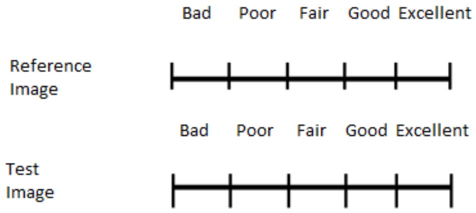

Figure 2 – Continuous scale of DSCQS scores

SSCQE (Single-Stimulus Continuous Quality Evaluation) [5] is an analogue of DSCQS for quality evaluation without direct comparison with a reference image. The metric represents a television viewing situation when only one stimulus is present.

JND (Just Noticeable Difference) [9] metric lies in distinguishing noticeable differences between test and reference images. Experts compare the reference and test images. If the difference is highlighted by a threshold percentage of experts, it is considered "noticeable" (minimum level of noticeability is 1 JND). A higher value of JND in this case means a worse quality [9] of the test image.

DAM [9] (Diagnostics Acceptability Metrics) include assessment of image quality acceptability in several directions. For example, for medical images, the presence of artifacts and the clarity of object contours can be evaluated [10]. Based on several aspects, a general value of the acceptability of image is calculated. The metric was originally developed to diagnose the quality of sound processing.

The ranking method can [6] be used when comparing the quality of images in a set and calculating coefficients of image quality characteristics. A table is formed (table

3), where rows represent images, columns represent experts. The expert assigns rank 1 to the highest quality image, other images receive higher ranks accordingly.

Table 3 – Example of table for images ranking

Image/Expert	Expert 1	Expert 2	Expert 3
Image 1	1	2	1
Image 2	2	1	2
Image 3	3	3	3

The lower the average rank, the higher the quality of the image.

When evaluating the quality of the image, its characteristics can be highlighted, for example, the accuracy of reproduction of the epicenter zone and attenuation of the ball's glow. It is necessary to assign weighting coefficients to the characteristics.

In one variant of the method, instead of assigning a rank, experts evaluate the importance of each characteristic on a scale [1,10]. The weight of the image characteristic is calculated according to the formula [6]

$$w_j = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{P_{ji}}{\sum_{j=1}^{m} P_{ji}},$$

where j – image characteristic number, i – expert number, n – number of experts, m – number of image characteristics, P_{ji} – the value of the $\mathrm{j-th}$ characteristic importance from the $\mathrm{i-th}$ expert.

Based on the calculated weights and characteristic values, a general value of image quality is formed. The quality value of the image from one expert is calculated according to the formula

$$\sum_{j=1}^{m} w(j) \cdot s(j),$$

where $w(j)$ – normalized on scale $[0,1]$ weight of the $\mathrm{j-th}$ image characteristic, $s(j)$ – value of the $\mathrm{j-th}$ characteristic.

The main advantage of the method of ranking images by quality level is the simplicity of implementation.

The preference method [6] is similar to the ranking method and can similarly be used for quality assessment and determination of characteristic coefficients. To determine the coefficients, each expert assigns an importance rank to the characteristics of the image (1 is the least important). The weight of a separate characteristic is calculated according to the formula [6]

$$w_j = \sum_{i=1}^{n} w_{ji} ,$$

where j – number of characteristic, i –expert number, n – number of experts, w_{ji} – importance rank of the $\mathrm{j-th}$ characteristic from the $\mathrm{i-th}$ expert.

Kendall's concordance coefficient (Kendall's W) [11] is used to establish the level of agreement between groups of ranks assigned by experts. The range of coefficient values is $[0,1]$. The coefficient is calculated according to the formula

$$\frac{12 \cdot \sum_{j=1}^{m} (R_j - \overline{R})}{n^2 (m^3 - m)} ,$$

where j – number of image (or characteristics), n – number of experts, m – number of images (characteristics), R_j – the sum of expert ranks for one image (characteristic), $\overline{R}$ – mean value of rank sums R_j.

The direct evaluation method [6] is used when experts have complete information about the image. Based on experience, experts directly assign weights of a defined interval to characteristics.

The methods of quality characteristics coefficient selection include the method of sequential comparison [6]. Image characteristics are sorted in the direction of decreasing importance. The first characteristic is assigned a coefficient of 1. The other characteristics are assigned a value between 0 and 1. The sum of the coefficients of the characteristics except the first is calculated. The expert decides what ratio of inequality (greater than, less than, equal to) the first coefficient should have in relation to the calculated sum. According to the adopted decision, the value of the first coefficient is updated. In the next step, the procedure is repeated, but the value of the last coefficient is not taken into account. The final value of the first coefficient is taken when compared with the sum of the following two coefficients. The values of the following coefficients are calculated similarly. The advantage of the method is the ability to carefully consider the values of the weights. The disadvantage is considerable complexity.

The basis of the method of pair-wise comparisons [6] is the construction of a special matrix. The matrix cells (table 4) record the results of a pair-wise comparison by image quality experts. For example, 1 means that the expert preferred the image, 0 means no preference.

After the experts fill in the matrices, a final matrix is formed with the sums of values for each cell. The quality of the image is expressed by the sum of the values in the row corresponding to it.

Table 4 – Example of pair-wise comparison matrix

	Image 1	Image 2	Image 3
Image 1	-	1	1
Image 2	0	-	1
Image 3	0	0	-

To set the weights of image characteristics, their importance is compared pair-wise. The coefficients of characteristics are calculated according to the formula [6]

$$w_j = \sum_{i=1}^{n} \frac{f_{ji}}{C},$$

where j – number of characteristic, i –expert number, n – number of experts, C – number of possible judgements $m(m-1)/2$ (m – number of characteristics), f_{ji} – the frequency of the $\mathrm{j-th}$ characteristic preferences from the $\mathrm{i-th}$ expert.

The method is characterized by high assessment accuracy.

The Delphi method [6] is a group expert method, among the possible applications of which is the assessment of image quality. The method includes surveying experts on image quality, analyzing responses, including significantly different opinions, and re-surveying. The assessment is completed when the group of experts reaches a generally acceptable score agreement. The advantages of the method are the possibility of independent expression of opinion and combination of views of experts from different directions. The disadvantage is a long discussion of the scores. The method is suitable for use in relatively complex cases of quality determination, for example, when testing new methods of visualization of medical images. Also, the method can be applied to select quality metrics, for example, to evaluate an anatomical image. A simpler method is the commission method, which lies in group discussion of quality and formation of a single opinion. The disadvantage is the possible suppression of individual thoughts.

Conclusions

The paper provides an overview of the main subjective metrics and expert methods for image quality assessment. Four areas of application of subjective image metrics can be distinguished: quality assessment, impairment level assessment, difference selection, acceptance level assessment. Expert methods allow to determine the highest quality image from the set and estimate the weights of the image quality characteristics.