Ann. Telecommun. (2010) 65:3–17
DOI 10.1007/s12243-009-0146-6

Automated qualitative assessment of multi-modal distortions in digital images based on GLZ

Andrzej Głowacz · Michał Grega · Przemysław Gwiazda · Lucjan Janowski · Mikołaj Leszczuk · Piotr Romaniak · Simon Pietro Romano

Received: 23 October 2008 / Accepted: 24 November 2009 / Published online: 17 December 2009

Abstract This paper introduces a novel approach to a qualitative assessment of images affected by multi-modal distortions. The idea is to assess the image quality perceived by an end user in an automatic way in order to avoid the usual time-consuming, costly and non-repeatable method of collecting subjective scores during a psycho-physical experiment. This is achieved by computing quantitative image distortions and mapping results on qualitative scores. Useful mapping models have been proposed and constructed using the generalised linear model (GLZ), which is a generalisation of the least squares regression in statistics for ordinal data. Overall qualitative image distortion is computed based on partial quantitative distortions from component algorithms operating on specified image features. Seven such algorithms are applied to successfully analyse the seven image distortions in relation to the original image. A survey of over 12,000 subjective quality scores has been carried out in order to determine the influence of these features on the perceived image quality. The results of quantitative assessments are mapped on the surveyed scores to obtain an overall quality score of the image. The proposed models have been validated in order to prove that the above technique can be applied to automatic image quality assessment.

Keywords Image quality · Image distortion · MOS · Mean opinion score · GLZ · Generalised linear models · Quality metrics

1 Introduction

Nowadays, several processing and transmission operations are commonly applied to digital images. Examples could be compression that allows reduction of a size of images or transmission over a telecommunications network based on connectionless protocols. This may result in introducing image distortions and (in consequence) an imperfect reconstruction of the original image. As a result, mono-modal (e.g., noise or blur) or, rather, multi-modal distortions (e.g., combination of noise and blur) may be introduced. This paper presents a uniform approach allowing for independent quantitative assessment of isolated distortion types and mapping them onto qualitative scores representing both isolated distortions and overall quality. Most image quality evaluation systems provide only a single score representing overall image quality, while the proposed independent assessment allows for specifying a particular source of image degradation as well.

Image quality metrics can be classified using three orthogonal classification schemes: by the amount of reference information required to specify the quality, by the metric calculation method, and by the way the quality is expressed. If the amount of reference information required to specify the quality is taken into

A. Głowacz · M. Grega · L. Janowski (✉) ·
M. Leszczuk · P. Romaniak
Department of Telecommunications, AGH University of Science and Technology, Krakow, Poland
e-mail: janowski@kt.agh.edu.pl

P. Gwiazda
Telekomunikacja Polska R&D, Warsaw, Poland

S. P. Romano
Computer Science Department, Universita' degli Studi di Napoli Federico II, Naples, Italy

account, “full reference”, “reduced reference” and “no reference” scenarios can be specified.

If the metric calculation method is taken into account, then metrics include a plethora of possible scalar parameters based on algorithms ranging from simple data (pixel-to-pixel) comparisons up to sophisticated image analysis. Data metrics look at the fidelity of the signal without considering its content. Examples of such measures are: peak signal-to-noise ratio (PSNR), mean square error (MSE) and similar. There are also several metrics based on sophisticated image analysis. Image metrics treat the data as the visual information that it contains. These metrics include a wide range of possible scalar parameters of the human visual system (HVS) that analyses the spectrum of the digital image in order to reproduce human perception. As an example of a metric, authors of picture quality scale (PQS) [15] defined an overall measure combined from several error scalars. In this solution, however, no detailed information can be obtained on specific image distortions.

Image quality metrics can also be classified by the way the quality is expressed and, furthermore, into the qualitative or quantitative. Quantitative criteria are usually expressed by a numerical value that can be easily calculated. However, there are no straightforward mappings between those values and user experience. In order to find such mapping function, subjective tests have to be run. On the other hand, metrics trained upon results obtained in subjective experiments are commonly referred to as qualitative. These criteria are considered with graphical (e.g. Hosaka plots [6]), textual (e.g. mean opinion score—MOS [9]) or numerical measures (e.g. *R* value and MOS).

An example of a quality metric providing the overall quality score (MOS) is the perceptual evaluation of video quality (PEVQ) [16] based on ITU-T J.144 [11] and ITU-T J.247 [10]. It is designed to estimate the video quality degradation occurring through a network; however, it can be simplified to the image quality metric since it operates on a decompressed video stream (frames level).

The main idea of the presented approach is to develop a set of cross-distortion robust[1] algorithms for independent assessment of the selected image distortions. Assessment of any mono-modal distortion of an image quality is not a very challenging research issue when mono-modally distorted images are considered (only one type of distortion). The task becomes much more complex when an image is multi-modally distorted (e.g. both noised and blurred). As the final step of the presented approach, a mapping between automatically obtained quantitative values and qualitative responses of a simulated user has been assured, based on psychophysical experiments (subjective tests) previously executed.

There are two contributions introduced within the presented research. The first one is a set of the algorithms for independent quantitative assessment of selected image distortions being robust to cross-distortion influence. The second one is the mapping of quantitative metrics onto qualitative scores that allows for elimination of difficult-to-organise, inaccurate and resource-consuming subjective tests, while retaining their clarity.

The paper is structured as follows: the next section describes the methodology details for metrics and compensations. Section 3 presents subjective quality evaluation, and Section 4 presents user response mapping. In the third section, the results are validated. The fourth section presents the implementation, while the fifth section concludes the paper.

2 Quantitative metrics and compensations

This section presents details of the quantitative metrics for assessment of selected image distortions being resilient to cross-distortion influence.

2.1 Quantitative metrics

The authors have developed metrics for quantitative assessment of the following seven distortion types: contrast distortion, blur, granularity, geometry distortion, noise, colour distortion and gamma distortion. The output of each algorithm is in a different metric scale, e.g. 0…1 or −infinity…infinity, and it has a different meaning each time, e.g., brightness[2] or motion vector length; however, user response mapping functions convert the different outputs into a common measure scare—MOS. Please refer forwards to Sections 4 and 5, as well as to Fig. 13, for more details. Two proposed metrics are based on some well defined approaches, examples of similar blur and noise metrics can be found in [5] and [6] respectively. The remaining part of the metrics represents a novel approach. Telekomunikacja Polska R&D (Polish Telecom), the orderer of the work, has specified the distortion types (quality parameters) list. The choice was motivated and justified by HVS characteristics and quality parameters of existing metrics (PQS, PEVQ). Figure 1 presents the general methodology for assessment of particular distortion types (including compensations that allow elimination

[1] Being insensitive to other distortions introduced to the image.

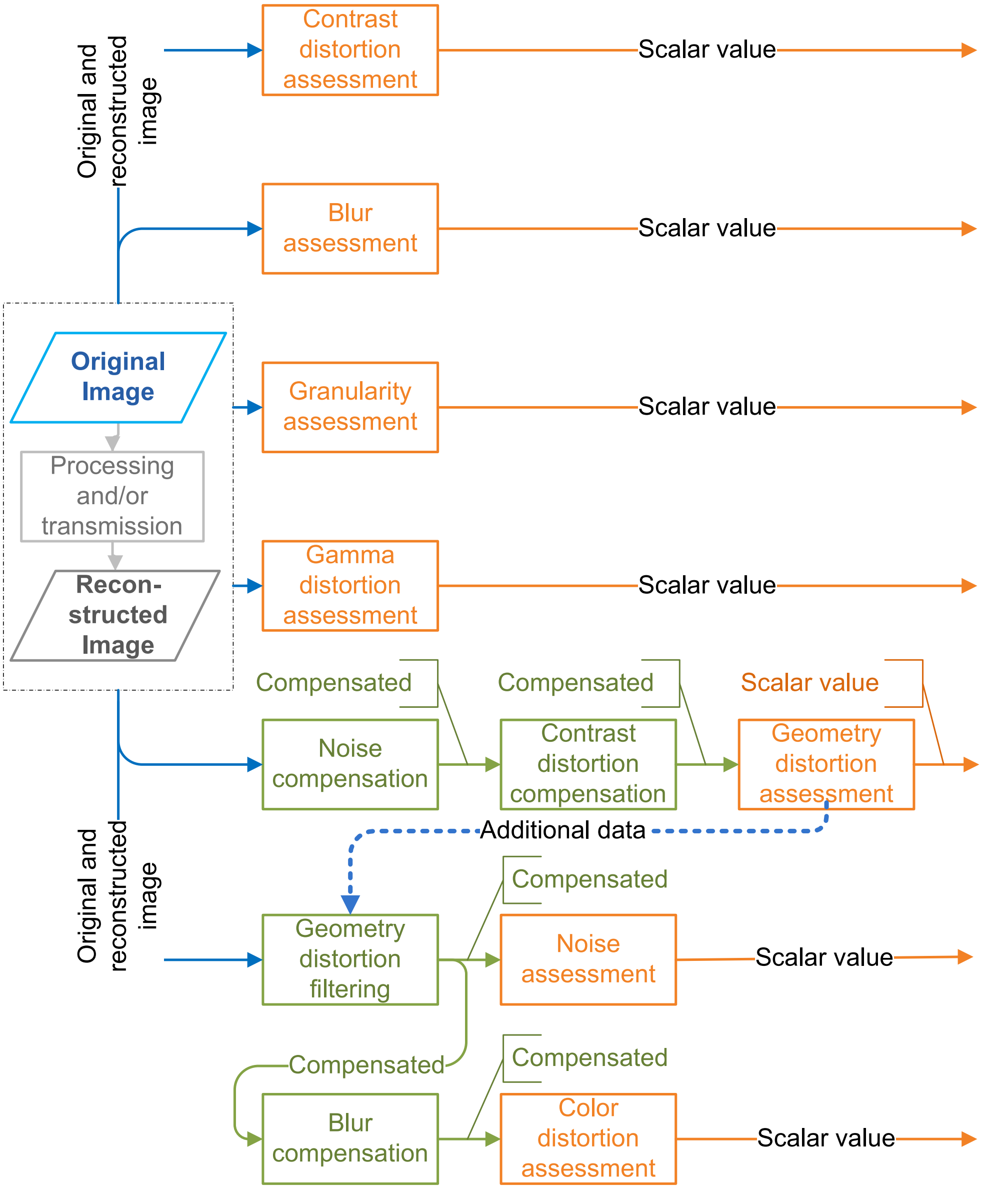


Fig. 1 Assessment methodology overview

of harmful influence of some distortions; the issue has been described in detail in Subsection 2.2).

2.1.1 Contrast distortion assessment

In order to calculate contrast distortion ConD, a method illustrated in Fig. 2 is used. The histograms of the original F and the reconstructed image $\widehat{F}$ are normalised. Afterwards, two pairs of the images (one pair consists of the images F and its normalised equivalent F^N) are compared using the PSNR metric defined by Eq. 2. The PSNR metric returns similarity levels in the decibel scale. The result of the PSNR values' subtraction stands for the comparison indicator (as a subtraction in the decibel scale is equal to a division in the linear scale). The experiment has proved that the applied approach assures insensitivity to any other type of image distortion.

$$\mathrm{ConD} = \mathrm{PSNR}\left(F, F^N\right) - \mathrm{PSNR}\left(\widehat{F}, \widehat{F}^N\right) \tag{1}$$

$$\mathrm{PSNR}\left(F, \widehat{F}\right) = 20 \log_{10}\left(\frac{F_{\max}}{\sqrt{\mathrm{MSE}\left(F, \widehat{F}\right)}}\right) \tag{2}$$

$$\mathrm{MSE}\left(F, \widehat{F}\right) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\left(F\left(j,k\right) - \widehat{F}\left(j,k\right)\right)\right]^2}{M \cdot N} \tag{3}$$

Assuming $F\left(j,k\right)$ as an original image luminance function, $\widehat{F}\left(j,k\right)$ as a reconstructed image luminance function, $F^N\left(j,k\right)$ as a normalised original image

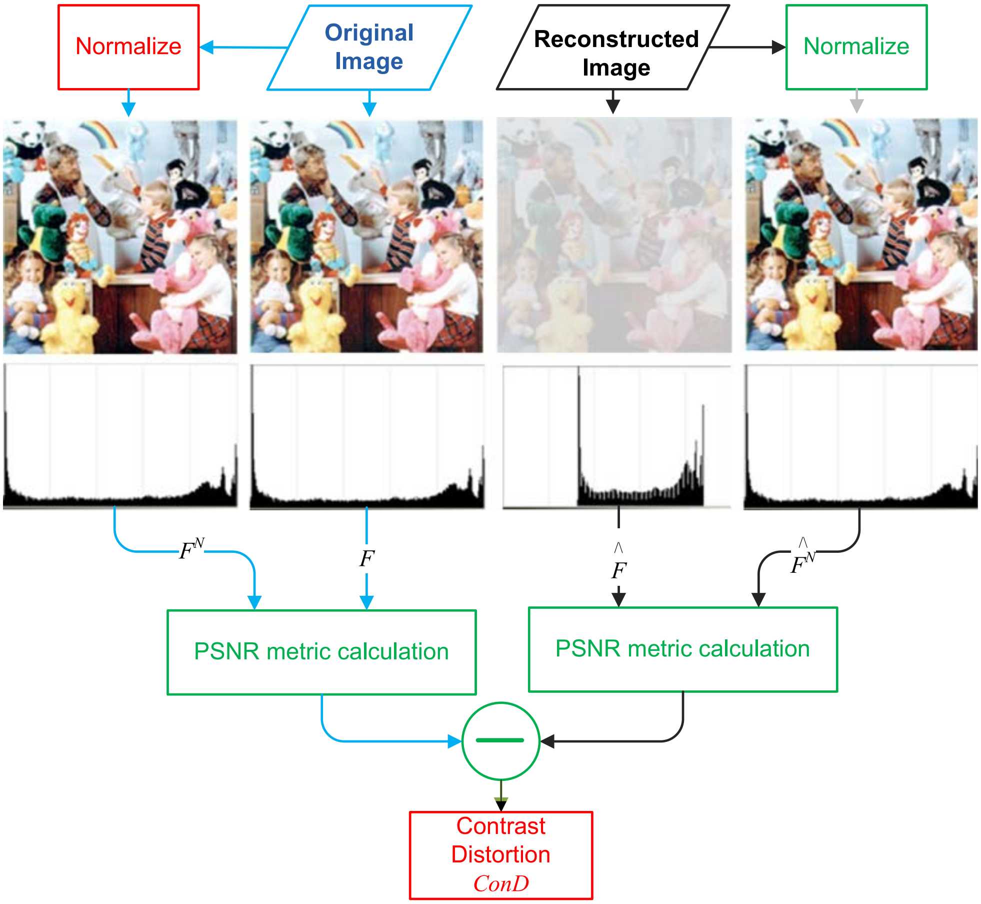


Fig. 2 Contrast distortion detection

luminance function, $\widehat{F}^N(j,k)$ as a normalised reconstructed image luminance function and $F_{\max}$ as a maximum luminance value, contrast distortion assessment algorithm ConD can be described by Eq. 1.

2.1.2 Blur assessment

The blur (also commonly referred to as a sharpness distortion) is one of the most significant factors that have an influence on the subjective opinion about the image quality. It is closely related to the amount of the details that an image can provide. The blur is defined as the shortest distance between the areas having different tones of colours (e.g. black and white). An edge detector seems to be an appropriate image blur indicator. The more edges detected on images, the better the image sharpness.

The first step to calculate an image blur is to convert both input images to gray scale (see Fig. 3). Afterwards, all the edges on the images are detected using the canny edge detection method (CED) [4]. The next step is to calculate power P of the images, which directly reflects an absolute edges amount. The result of the subtraction of the images' power defines the blur comparison value, which is returned as the output of the script. The experiment has proved that the applied approach assures insensitivity to any other type of image distortion.

$$B = P(F) - P\left(\widehat{F}\right) \tag{4}$$

$$P(F) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\mathrm{CED}\left(F(j,k)\right)\right]^2}{M \cdot N} \tag{5}$$

$$P\left(\widehat{F}\right) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\mathrm{CED}\left(\widehat{F}(j,k)\right)\right]^2}{M \cdot N} \tag{6}$$

The blur assessment algorithm B is defined by Eq. 4, provided that $P(F)$ and $P(\widehat{F})$ are correspondingly an original image power and a reconstructed image power.

2.1.3 Granularity assessment

Two images showing the same object, having exactly the same resolution, can present diverse qualities – they can provide completely different amounts of details. The reason why it can happen is a decrease of the number of effective pixels contained in an image. In other words, the effective size of a single pixel on an

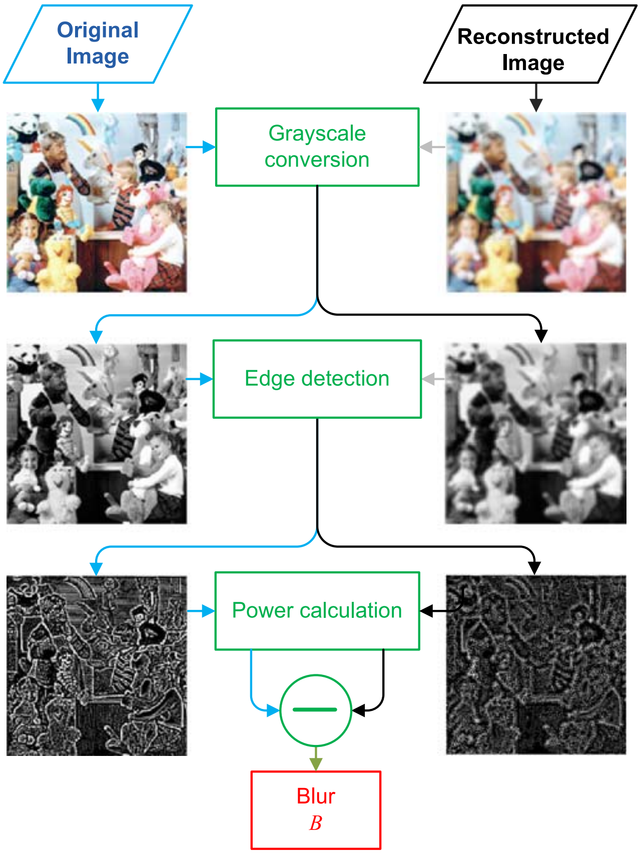


Fig. 3 Blur detection

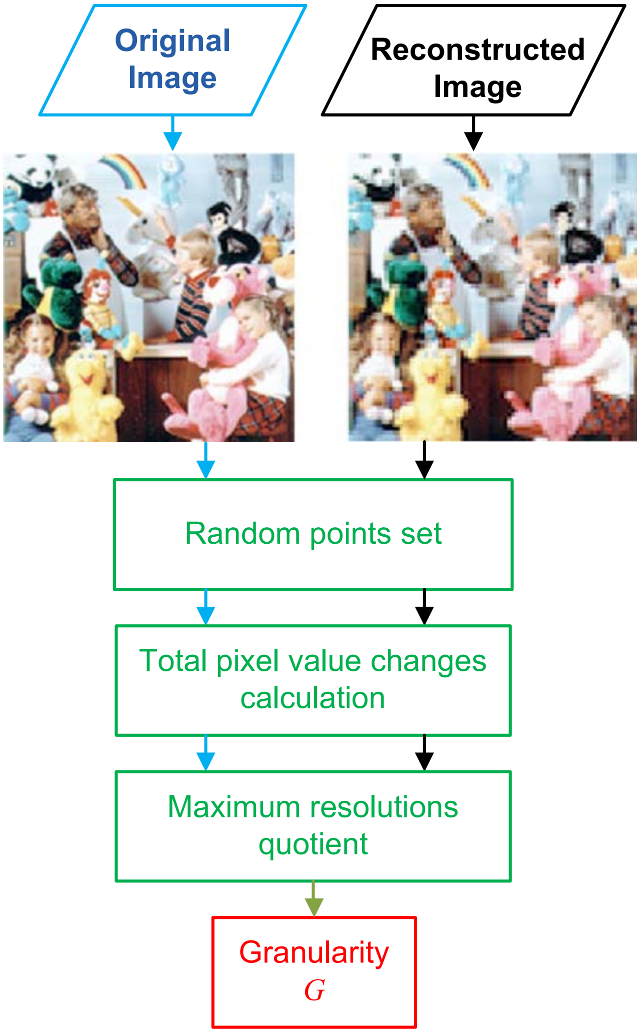

Original Image
Reconstructed Image
Random points set
Total pixel value changes calculation
Maximum resolutions quotient
Granularity
G

Fig. 4 Granularity detection

image can be significantly enlarged, which will result in higher granularity of an image. The type of distortion applies to the whole image.

Calculation of image granularity is performed in a few steps, as illustrated in Fig. 4. At the beginning, l random points are chosen from an image (L_l^{begin}). Starting from each point, the total number of pixel-changes PixCh_l is calculated for the vertical $\text{VL}(l)$ (see Eq. 9) and horizontal $\text{HL}(l)$ (see Eq. 10) lines (a pixel-change Ch appears when at least one of the R, G or B values is different from the previous one). Line length is described as $L_l^{\text{end}} - L_l^{\text{begin}}$ and was set to 100 pixels in the experiment (each L_l^{end} point is 100 pixels distant from corresponding L_l^{begin} point). The maximum number of pixel-changes for all the lines is the image resolution (see Eq. 8). It is possible that the real (maximum) value of the image resolution will not be found, as the whole image area is not being analysed. However, it is not a problem since the same lines are analysed on both images (original and reconstructed). As a result of the granularity comparison process, a quotient of the maximum found resolution for reference and distorted images is obtained. The experiment has proved that the applied approach assures insensitivity to any other types of image distortion.

$$G = \text{EffRes}(F)/\text{EffRes}\left(\widehat{F}\right) \tag{7}$$

$$\text{EffRes} = \max\left(\left(\text{PixCh}(\text{VL}(l)), \text{PixCh}(\text{HL}(l))\right)\right) \tag{8}$$

$$\text{PixCh}(\text{VL}(l)) = \sum_{i=L_l^{\text{begin}}}^{L_l^{\text{end}}} \text{Ch}(f(j,i), f(j,i+1)) \tag{9}$$

$$\text{PixCh}(\text{HL}) = \sum_{i=L_l^{\text{begin}}}^{L_l^{\text{end}}} \text{Ch}(f(i,k), f(i+1,k)) \tag{10}$$

$$\text{Ch} = \begin{cases} 1 & \text{when R or G or B value is} \\ & \text{different for two pixels} \\ 0 & \text{other cases} \end{cases} \tag{11}$$

The granularity assessment algorithm G is defined by Eq. 7, where $\mathrm{EffRes}(F)$ and $\mathrm{EffRes}(\widehat{F})$ are an effective resolution of the original image and the effective resolution of the reconstructed image, respectively.

2.1.4 Geometry distortion assessment

The geometric distortions may be introduced into the image during the analog processing stage. The analysis of geometric distortions is based on motion detection. Although the measurement is performed on still images, the original and reconstructed images are treated as the concurrent frames in order to apply the motion detection algorithm. The geometric distortion is treated as a movement between two frames. The motion detection algorithm is similar to motion estimation used in video compression in the MPEG standard applications. The algorithm is executed in several steps. In the first step, a set of uniformly distributed square blocks (n) is chosen on both original and distorted images. For each block from the original image, a similar image is searched for in the distorted image. The search is performed in a set radius (r) from the original location of the block, i.e., all blocks in this radius are analysed (see Fig. 5). The similarity here is understood as the smallest difference between the original block and all possible blocks within the set radius in the distorted image. The difference between blocks is calculated with the use of MSE. In the next step, for each pair of blocks—the one from the original image b and the one from the distorted ($\widehat{b}$) image—a motion vector ($\mathbf{V}$) is calculated. Moving vector is a vector defining vertical and horizontal distance that moved the original block. Finally, the total length of the movement vectors is taken into account in order to allow assessment of the geometrical distortion using the following formula:

$$\mathrm{GeD} = \frac{\sum_{i=1}^{n} |\mathbf{V}|}{n \cdot r} \tag{12}$$

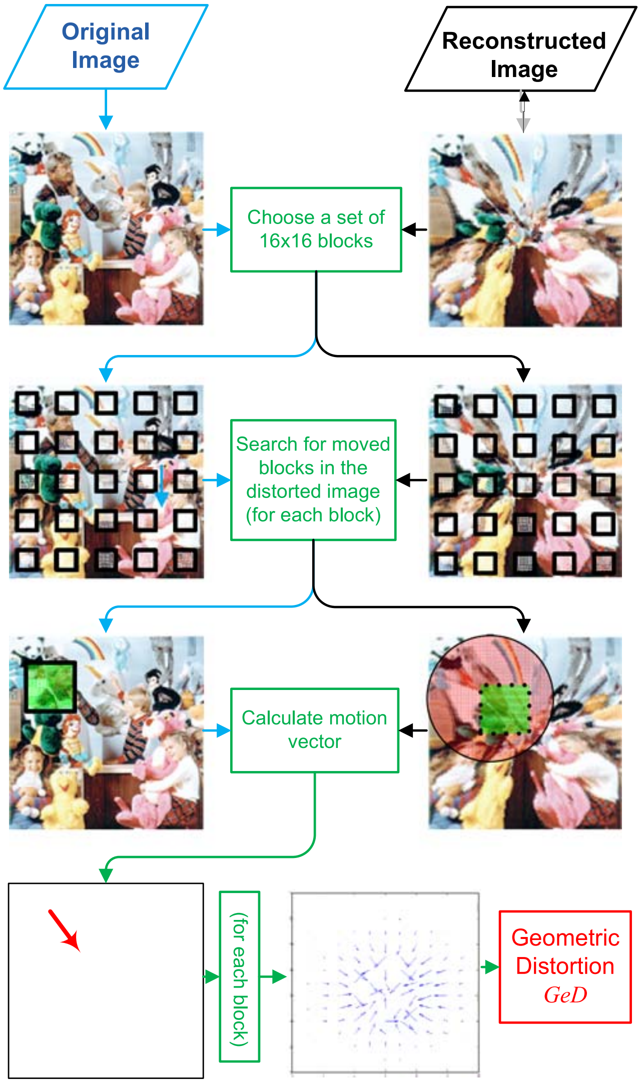


Fig. 5 Geometry distortion detection

2.1.5 Noise assessment

Noise assessment is roughly based on an idea of Hosaka plots. The algorithm starts with quad-tree image decomposition, and then noise is assessed in square pixel blocks divided into a couple of classes C_i, $i = 0, \ldots, n$ (usually where $n = 4$), thus in blocks beginning from 1×1, usually up to 16×16. Both noise and reconstruction inaccuracy parameters are represented by an equal number of Hosaka values: $\mathrm{DS}\,(C_i)$ (noise coefficients) and $\mathrm{DM}\,(C_i)$ (inaccuracy coefficients), accordingly ($\mathrm{DS}\,(C_0) \equiv 0$). The Hosaka plots are drawn at a polar chart, where one hemi-disk is related to DS and the second to DM. The shape of the Hosaka plot specifies if noise is introduced for details (represented by small blocks) or for larger, homogeneous areas (large blocks). Please refer to [6] for more details on Hosaka plots, which are not presented here due to space limits.

The distortion value N has been defined as being proportional to the area of the noise part of the Hosaka plot; thus, a sum of areas of $O\,(0,0)\,S_i(\theta_i, \mathrm{DS}\,(C_i))\,S_{i+1}\,(\theta_{i+1}, \mathrm{DS}\,(C_{i+1}))$ triangles, where point coordinates are given in a polar coordination system, and θ_i is an angle at which the $\mathrm{DS}\,(C_i)$ value has been presented

$$N = \sum_{i=0}^{n-1} \frac{1}{2} \sin|\theta_{i+1} - \theta_i|\,\mathrm{DS}\,(C_i)\,\mathrm{DS}\,(C_{i+1}) \tag{13}$$

Considering that $|\theta_{i+1} - \theta_i| \equiv \frac{\pi}{n-1}$, as well as the fact that the distortion value is not normalised, all constant values can be excluded, resulting altogether in a simplified N notation as

$$N = \frac{1}{2}\sin\frac{\pi}{n-1}\sum_{i=0}^{n-1}\mathrm{DS}\left(C_i\right)\mathrm{DS}\left(C_{i+1}\right)$$
$$\propto \sum_{i=0}^{n-1}\mathrm{DS}\left(C_i\right)\mathrm{DS}\left(C_{i+1}\right) \qquad (14)$$

Please consult Fig. 6 for the graphical algorithmic presentation of the noise level assessment algorithm. The output metric starts from 0 and is increasing along with squared noise distortions (the highest observed was around 200).

2.1.6 Colour distortion assessment

The colour distortion of an image is perceived by the quality of hue component representation in a hue-saturation-value (HSV) colour space. We consider all pairs of corresponding pixels in the original and distorted image. A difference histogram is created based on hue distortions of each pixel pair. During experiments, it has been found that certain types of distortion produce large peaks in a difference histogram (see Fig. 7).

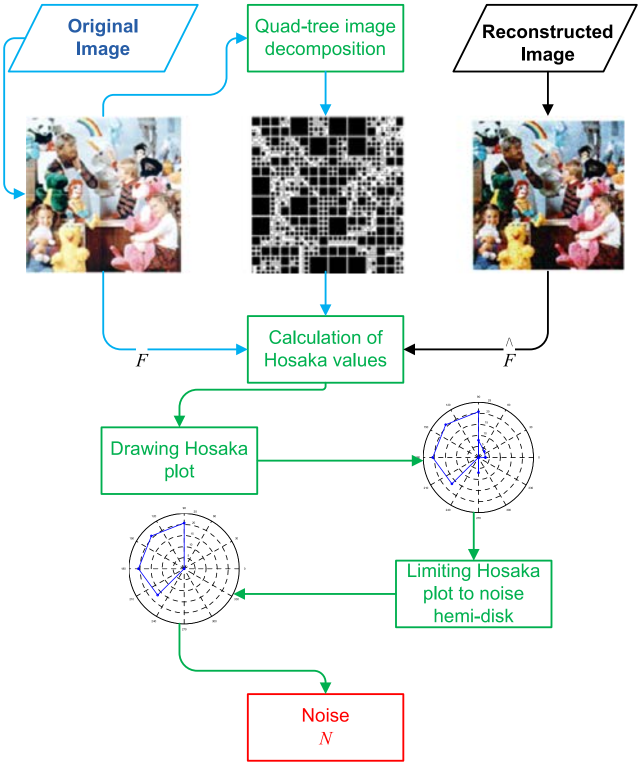

Original Image
Quad-tree image decomposition
Reconstructed Image
F
Calculation of Hosaka values
$\hat{F}$
Drawing Hosaka plot
Limiting Hosaka plot to noise hemi-disk
Noise
N

Fig. 6 Noise detection

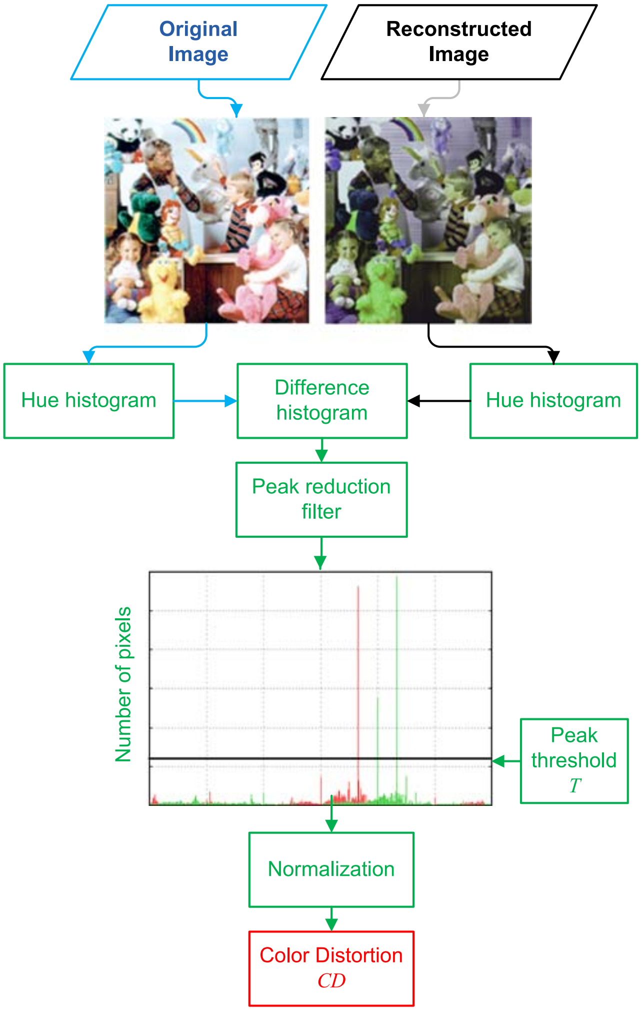


Fig. 7 Colour distortion algorithm

This is concerned especially with either contrast or geometric distortions. Elimination of these peaks is crucial for a reduction of a metric variance. Therefore, a peak threshold T is defined

$$T = \frac{M \cdot N}{360}. \qquad (15)$$

Differences lower than the threshold T can only be included into a difference histogram. Each hue is represented in the histogram by its number of pixels in the image. Threshold T is the average number of pixels per hue and, thus, depends on image size. The colour quality assessment is considered as the sum of all the hue differences between image histograms (measured in number of pixels) divided by total number of pixels. This division is carried out for measure normalisation. The output metric lies in the range $0\ldots1$ and is increasing along with colour distortions.

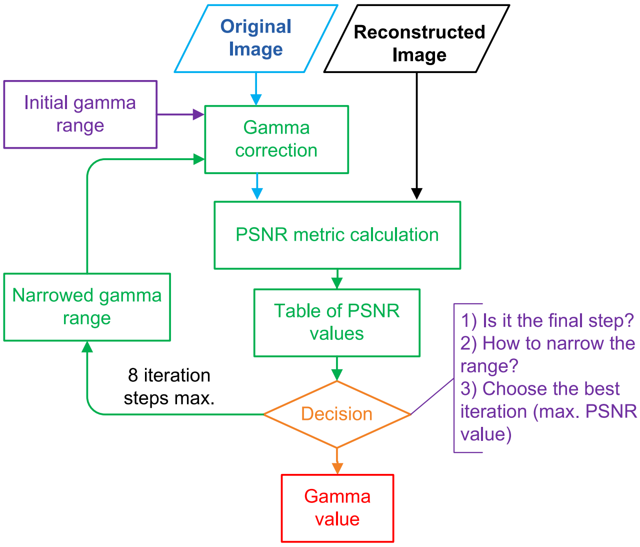


Fig. 8 Gamma distortion algorithm

2.1.7 Gamma distortion assessment

The gray-scale distortion of an image is mostly caused by the changes in a gamma level. This type of distortion can be successfully assessed by gradual degradation of the original image and its direct comparison to the distorted image. We use the following algorithm to assess gray-scale distortion of an image (Fig. 8).

First, we empirically determine the limits of gamma that are the levels of the highest perceivable darkening and brightening (selected range was between 0.3 and 2.3). Then, we degrade the original image by applying gamma distortion using two different levels selected from the range (middle points of each half, i.e. 0.8 and 1.8 in the first step). Calculation of PSNR value between each original image and the same distorted one allows to narrow initial gamma range (half of the gamma range with lower PSNR metric is discarded). This step is repeated a number of times. Each step leads to consecutive game range narrowing. At the end, the centre gamma value from the sub-range with the highest corresponding PSNR value is considered as the distortion level. It is a simple numerical method used to find the point within the continuous range, using a limited number of steps. We decided to use seven steps, which allows us to achieve accuracy equal to 2^{-7} of the initial range width.

2.2 Compensation

Assessment of any distortion of an image quality is not a very challenging research issue as far as mono-modally distorted images are considered (only one type of distortion). The task becomes much more complex when an image is multi-modally distorted (e.g. combining contrast distortion, blur and gamma distortion). According to performed research, only contrast distortion, blur and granularity assessment algorithms are insensitive to other types of distortions. This means that evaluations of some of the distortions cannot be performed properly when at least one additional algorithm appears (distortion disables proper calculation of the quantitative quality). In order to enable accurate assessment of a single distortion of a multi-modally distorted image, a number of compensations were applied. The authors considered each particular distortion metric to be insensitive (or fully compensated) to other distortions even if, for other maximum distortion values, the distortion being evaluated was not affected by one point in the MOS scale. Compensations allow to eliminate harmful influence of some distortions, based on improvement of the reconstructed image (applicable only for fully reversible distortions) or distorting the original image (for irreversible distortions).

2.2.1 Contrast distortion compensation

A compensation of the contrast distortion allows to calculate other quality metrics more precisely. Moderate losses in a contrast level are fairly reversible distortions, and quality of the reconstruction is acceptable from the user point of view. In order to compensate a contrast distortion, histograms of the reference and the reconstructed images are normalised (stretched to the maximum range).

2.2.2 Blur compensation

This type of distortion is irreversible; hence, the reconstructed image cannot be corrected without knowing the exact distortion model. The only possible solution is to apply the same level of blur for the original image. In order to assess parameters of distortion that should be applied on the original image, the numeric method of the minimal differences level between the original and the reconstructed image is used (see Fig. 9). A computation of differences is based on the blur evaluation methods. The whole process consists of eight steps, each step narrowing the range of the possible distortion parameters and giving better results. If we assume distortions ranging from 0 to 1, eight steps of this method allow us to assess distortion parameters with an error equal to 2%, in the worst case.

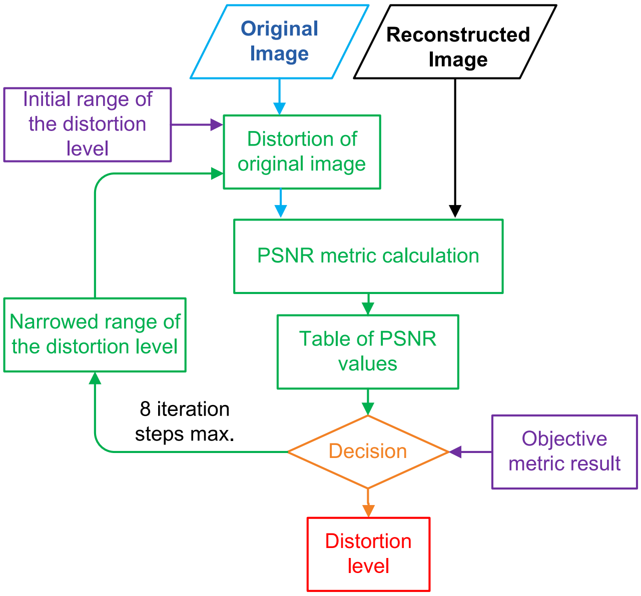


Fig. 9 Numeric method of blur compensation

2.2.3 Geometry distortion filtering

Once the geometric distortions are quantitatively assessed, they are filtered out in the proposed solution—thus, no areas affected are further processed and assessed. The filtering is based on results of assessment of geometrical distortions. Matching blocks from the original and the reconstructed images are passed for further assessments.

The criterion for qualifying a particular pair of blocks (b and $\hat{b}$) for further processing (sets B^F and $\widehat{B}^F$) is based directly on $|\mathbf{V}|$ (the length of its motion vector related to them). Only pairs of blocks having $|\mathbf{V}| = 0$ are qualified (Fig. 10).

$$B^F = \left\{ b \in B : |\mathbf{V}| \left(b, \hat{b} \right) = 0 \right\} \tag{16}$$

$$\widehat{B}^F = \left\{ \hat{b} \in \widehat{B} : |\mathbf{V}| \left(b, \hat{b} \right) = 0 \right\} \tag{17}$$

Two new images are created as a composition of original and reconstructed blocks that passed the filter. The blocks are aligned in single rows in each of the images, with horizontal bars having a height equal to the height of a single block and a width equal to a width of all qualified blocks. The new images are passed as a basis for further processing.

2.2.4 Noise compensation

For various distortion metrics, noise compensation procedures had to be applied. In most cases, the peak noise elimination filter was deployed [7]. Considering that each image noise compensation introduces changes in de-noised images as well, the compensation is applied to both the original and the reconstructed image (see Fig. 11).

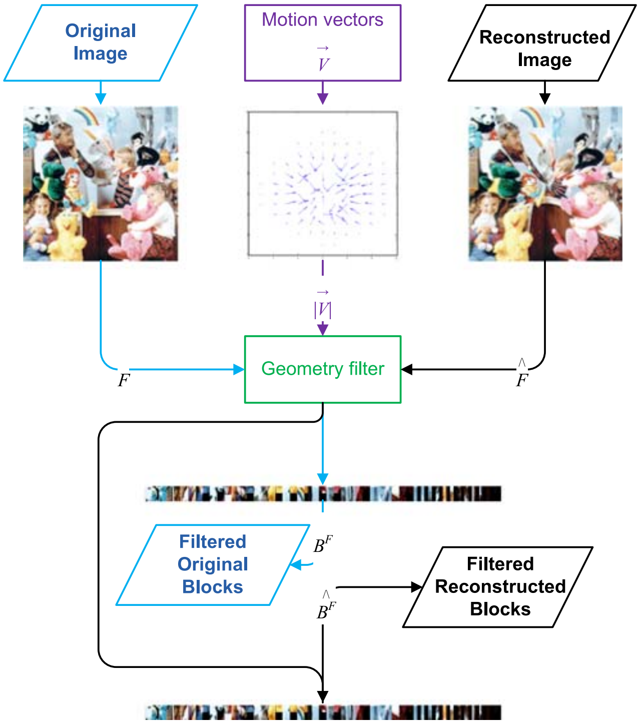


Fig. 10 Geometry distortions filtering

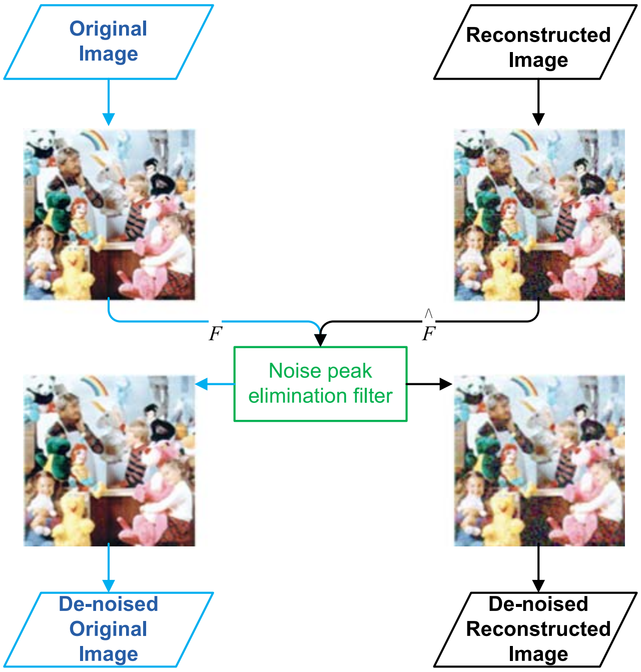


Fig. 11 Noise compensation

The primary function of the noise eliminating filter is to smooth image objects without losing information about edges and without creating unnecessary image structures. The key assumption is to replace every pixel tagged as a noise pixel with the neighbouring pixels' values. A pixel is qualified as noise only if it has the maximal or the minimal values within a pixel window—$w\,(j,k)$.

$$w\,(j,k) = \begin{bmatrix} F\,(j-r,k-r)\ \ldots\ F\,(j+r,k-r) \\ \vdots \\ F\,(j-r,k+r)\ \ldots\ F\,(j+r,k+r) \end{bmatrix} \quad (18)$$

It is possible to specify a neighbouring pixels' proximity radius, thus the width of the neighbourhood (r). In most of the cases, $r = 2$ gave the best results. The noise filtering is done by applying a digital filter to the noisy reconstructed image. The digital filter is based on computing weighted averages of target pixel colour components.

3 Subjective quality evaluation

The main motivation of subjective trials was to collect opinion scores (OSs) regarding quality of reconstructed images in order to determine a mapping function between quantitative quality (output of the algorithms for quantitative quality assessment) and quality perceived by a typical user. Subjective OSs allow for constructing models, eliminating the necessity of involvement of human testers in further tests.

3.1 Methodology

General provision for the subjective assessment of the quality is presented in [8]. According to the recommendation, subjective tests of image quality should be conducted on the diverse and numerous groups of subjects (testers). For all the reconstructed images, a number of subjective scores (OS) should be collected. In order to assess how strongly few distorted parameters influence the perceived quality, each test session should include evaluation of both mono- and multi-modally distorted images.

The test required the double stimulus method with five-level impairment grading and the absolute image quality assessed. Hence, DSIS [8] was used as a basic methodology for subjective tests, with one minor change. Assessment of the image quality did not refer to the distortion level, but to the absolute image quality (just as described in the double stimulus continuous quality scale methodology [8]). Therefore, the presented methodology is a combination of both approaches based on the double stimulus, and it is similar to the ACR methodology [13]. The applied modifications eliminate error being a result of transition between distortion level and image quality, which is required as the final result of the image distortions' assessment process.

3.2 Test-set

Subjective tests were performed upon a test material (test-set) prepared using a software distortion tool designed in the scope of the research. The distortion tool allowed application of all types of considered distortion aspects (seven isolated quality aspects).

One image from the standardised digitised image set [12] was chosen as a base to create the whole test-set (see Fig. 12). The image presents variegated content and seems to be representative for colour images. The

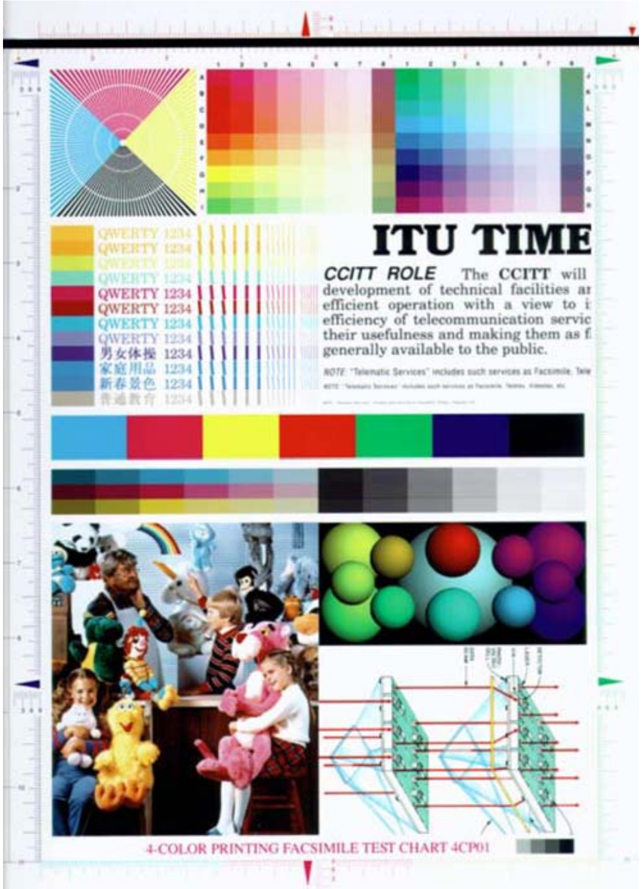

QWERTY 1234
ITU TIME
CCITT ROLE The CCITT will development of technical facilities ar efficient operation with a view to i efficiency of telecommunication servic their usefulness and making them as f generally available to the public.
NOTE: "Telematic Services" includes such services as Facsimile, Tele
4-COLOR PRINTING FACSIMILE TEST CHART 4CP01

Fig. 12 Base test image

test-set included several images generated with a distortion tool: 94 distorted mono-modally and 330 distorted multi-modally.

3.3 Subjective tests

Subjective tests included approximately 250 trials (testers that where mostly students, i.e. their ages were between 19 and 25) overall. Each evaluation trial consisted of 60 random images (generated separately for each trial) chosen from the whole test-set (424 images), which eliminated error being a result of the order of fixed images. Within these 60 images, in each test, 12 were mono-modally distorted and 48 were multi-modally distorted. The number of images in one trial was limited by the human capability to give reliable answers in a continuous period of time (about 15 min). As a result of subjective tests, about 2,400 OSs for mono- and 9,600 OSs for multi-modally distorted images have been collected.

4 User-response mapping

The next research goal was to find a function mapping the quantitative image distortion levels to the qualitative user responses. In Section 2.1, different distortion assessment algorithms have been described. Nevertheless, the value obtained for each algorithm does not predict the qualitative user response. Therefore, a function mapping the assessment values to the qualitative user responses had to be found.

At first look, the quantitative result of an assessment and a mean qualitative user response could be mapped using a regression algorithm. Nevertheless, the basic assumption of the regression algorithm is that the response distribution can be approximated by a normal distribution. As the users could choose only one of the five answers, the obtained answer distribution cannot be approximated by the normal distribution. The reason is a symmetry of the normal distribution (around the mean value) that cannot be guaranteed, as distribution of testers' responses reveals a skewness. Moreover, the verbal description used in the DSIS is easy to understand for people, but it has no clear mathematical meaning. Therefore, in [9], mapping the verbal answers to numbers is proposed. As a consequence, the numbers are only a convention and the analysed variable (response) is of the ordinal type [1].

The ordinal variables are variables for which an ordering relation can be defined but a distance measure cannot be defined. The OSs have an order relation because "Poor"[2] is better than "Bad", but worse than "Fair". Nevertheless, a distance between answers "Excellent" and "Fair" or "Good" and "Poor" cannot be found. Everyone has their own measure of these differences. Therefore, modeling the ordinal answers in the same way as strictly numerical data is a common mistake [1].

In order to model the ordinal answers properly, more general models than simple regression models have to be used [1]. The generalisation of the regression model is the generalised linear model (GLZ). The recommended approach is the GLZ which, in the presented study, is supported by an ordinal multinomial distribution and the *logit* link function.[3]

Note that there are five possible answer; therefore, the user response distribution is a discrete distribution. As a consequence, the GLZ model is the probability of each possible answer from "Excellent" to "Bad" computed as a function of the distortion assessment algorithm value.

The main advantages of the GLZ model in comparison to the linear regression are as follows:

- The user response distribution is found (for linear regression, only a mean value is known)
- It is not necessary to assume that the OSs are normally distributed (for linear regression, the OSs have to be approximately normally distributed).

As a result of using the GLZ model, the user response distribution is estimated as a function of the assessment values, i.e. we estimate

$$P(Y \le i) = \frac{\exp\left(\alpha_i + \beta_1 x + \cdots + \beta_k x^k\right)}{1 + \exp\left(\alpha_i + \beta_1 x + \cdots + \beta_k x^k\right)}, \tag{19}$$

where Y is an OS, α_i is a coefficient different for each OS value and β_j is an assessment polynomial coefficient the same for each assessment. Different polynomial orders can be considered. Moreover, more complicated functions of more than one assessment value can be used, too.

The distribution could be used to compute the MOS for each distortion. Nevertheless, during the research, it has been found that some distortions have the answer distribution where the mean value is around 3 and the most probable answer is 4 or 2 but never 3.

[2] The testers chose an answer described by words "Excellent", "Good", "Fair", "Poor" or "Bad". Moreover, the meaning of each single word was more precisely described according to recommendation [9].

[3] The GLZ can model different distributions and different nonlinear transformations of the distributions. The nonlinear transformations are called link functions.

For example, the distribution was 10% “Excellent” (5), 35% “Good” (4), 20% “Fair” (3), 25% “Poor” (2) and 10% “Bad” (1). In such a case, the MOS can give a corrupt feeling that the greatest number of users see the image as “Fair”. Therefore, we decided to introduce and to use the most probable opinion score (MPOS) as a measure of users’ responses.

Performing the research project for a commercial company, we were focused on a practical implementation of the obtained results. Therefore, we focused on finding out how the largest group of users behaves. Such functionality is another reason to use the MPOS measure instead of the MOS measure.

The detailed algorithm of the analysis of the user’s answers was as follows:[4]

1. Obtained data have been cleaned, i.e. if a tester has given a far better answer (i.e. at least two levels) for a worse image than for a better one, all tester’s answers have been removed. Such cleaning has to be done since some testers scored the pictures only in order to finish the test. They did not think about real picture quality.
2. The cleaned data obtained for a single distortion have been split up in order to obtain a training set and a test-set. The reason for dividing the data set into two sets is as follows. We were looking for a general function mapping a distortion assessment algorithm value on a distribution of the OSs. The mapping function is general if it predicts not only the distribution of the OSs that were used to estimate the mapping function parameters. Therefore, instead of repeating tests after finding the mapping function, the collected data set was divided into a training set and a test-set. The test-set was used to test if the mapping function correctly estimates the data that were not used to estimate the mapping function. In case of single distortion, a mapping function was a polynomial of distortion metric. Therefore, different mapping functions are different polynomials of assessment value (see Eq. 19).
3. The results obtained for a single distortion have been used to find a mapping function for a particular distortion, and the procedure has been repeated for all distortions.
4. Since it is possible to propose numerous different mapping functions (based on the GLZ modeling) the best one has to be chosen. The Schwartz information criterion (SIC) [3] was used as a criterion of comparing different mapping functions. Note that the mapping function is a GLZ model. Since the GLZ model is a statistical model, it is possible to compute a measure of the fitting goodness using R^2. Therefore, the SIC is one possible alternative fitting goodness measure.
5. The obtained GLZ model distribution and a test-set distribution have been compared on the basis of the Pearson χ^2 test [2]. If the obtained distributions were different from the test-set distribution (according to the Pearson χ^2 test), another model was analysed.

The final mapping function describes an OS probability as a function of a particular distortion assessment algorithm value. Note that, for the same distortion assessment algorithm value, the OS probability is different for each answer (1, 2, 3, 4 or 5). Therefore, five different probability functions of the distortion assessment algorithm value represent the final result obtained for a single distortion. Note that a distortion assessment unit does not influence the obtained results since we estimate β_j coefficients (see Eq. 19). Each such coefficient has unit revers to the assessment unit; therefore, the whole polynomial is unit-less.

We cannot describe the exact form of the obtained results (i.e. the estimated coefficients) since the research was done for TPsa and, as such, is their property. Some more details about the estimation methodology can be found in [14].

The obtained probabilities are computed with confidence intervals. The model answer is the MPOS; therefore, the highest drop probability has to be found. Since the confidence intervals of two probabilities can overlap, we could consider a crossing value. The crossing value could be marking the obtained MPOS or adding a noninteger value. Nevertheless, such a value would make the system more complicated to interpret. Therefore, we did not specify the intervals where probabilities’ confidence intervals overlap.

The final user response mapping is represented by seven different functions. The functions map the distortion assessment algorithm value on the five-level OS scale. Note that each function maps single distortion. Separately, a function mapping all seven distortion assessment algorithm values on the five-level OS scale has been found. Therefore, the final result of GLZ modeling was a set comprising eight functions. The first seven describe the distribution of the user OS for a single distortion. The last function describes the OS distribution for an image affected by multi-modal distortion.

[4] Since, the paper size is limited, it is not possible to explain all details. Nevertheless, we believe that presented steps are sufficient to implement the same methodology in another research.

5 Results validation

The goal of this research was to find a correlation between the automatically obtained qualitative scores and the user OSs about the images. In Section 4, mapping the assessment values on the OS distribution were proposed. Moreover, in Section 4, the reason for using MPOS instead of MOS is presented.

The function mapping the qualitative assessment values on the MPOS value is called the MPOS metric. Since eight different mapping functions have been estimated, eight different MPOS metrics are found. Note that seven of them map the MPOS for one of the seven mono-modal distortions. The eighth MPOS metric maps the MPOS for the multi-modally distorted image. This special MPOS metric has been called the complex MPOS metric. Moreover, we proposed the worst, i.e. the minimum, of all seven single distortion MPOS metrics to use as an alternative multi-modal distortion metric. This metric has been called the minimum MPOS metric. The analysis scheme and the obtained results are shown in Fig. 13

The accuracy of the obtained metrics is computed as an *answers difference*, i.e. the difference between the sample mode (the most frequent value) and a metric answer. Note that the negative values indicate that a metric overestimates the image quality and the positive answers indicate that a metric underestimates the image quality. For example, if the difference is −3, it means that the metric answer was 5 or 4 for the image for which the sample mode answer (the most frequent answer) was 2 or 1, respectively. This notation is used in Fig. 14.

In Fig. 14, frequencies of a difference between the answer that has been chosen by most testers and the metric answer have been presented. Figure 14a–g presents the accuracy of the single distortion metrics, i.e. the metrics considering only a mono-modal distortion. The accuracy of the metrics for mono-modal distortion metrics has been compared with the images distorted by the same distortion.

Figure 14h and i present the accuracy of the complex MPOS metric and minimum MPOS metric, respectively. The comparison for the metrics for multi-modal distortions has been performed for all images, including those multi-modally distorted.

From Fig. 14a–g, it can be seen that some distortions are very well predicted, such as contrast distortion (Fig. 14a). The others, for example, granularity (Fig. 14c), reveal a much higher variance. Figure 15 shows the metric and the user answers (the most probable answers) of the granularity on a single plot.

An interesting observation is that the high variability of the answers' difference is not necessarily a result of the metric inaccuracy. Note that the metric response has to be monotonic since, for a more distorted picture, its quality cannot be better. Nevertheless, the users' responses vary for the increasing distortion level (the solid line in Fig. 15). The user answers can vary for numerous different reasons. Note that different images can be scored by different persons and their feelings can be different.

Fig. 13 Overview of the metrics analysis scheme

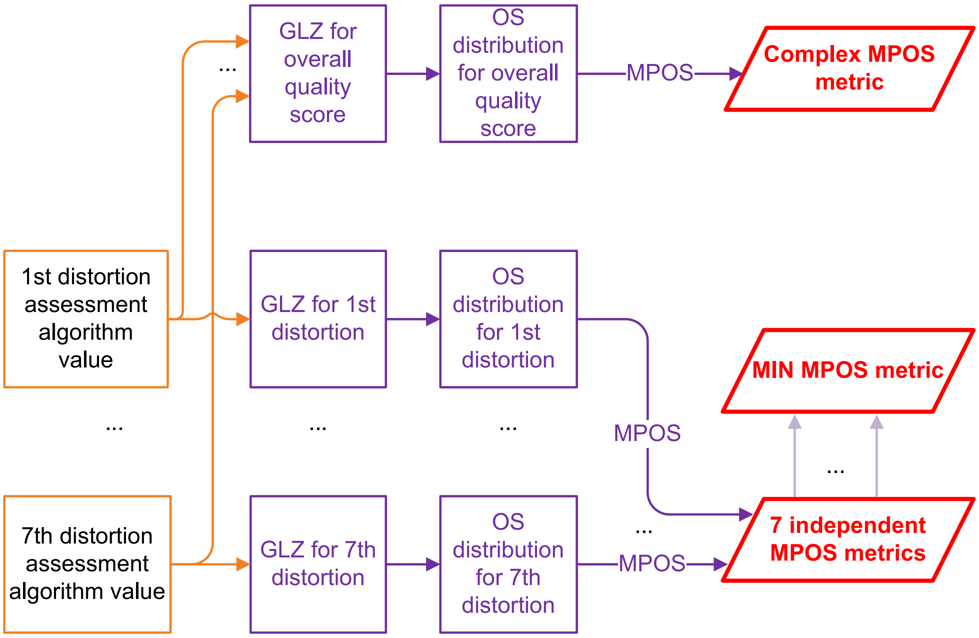
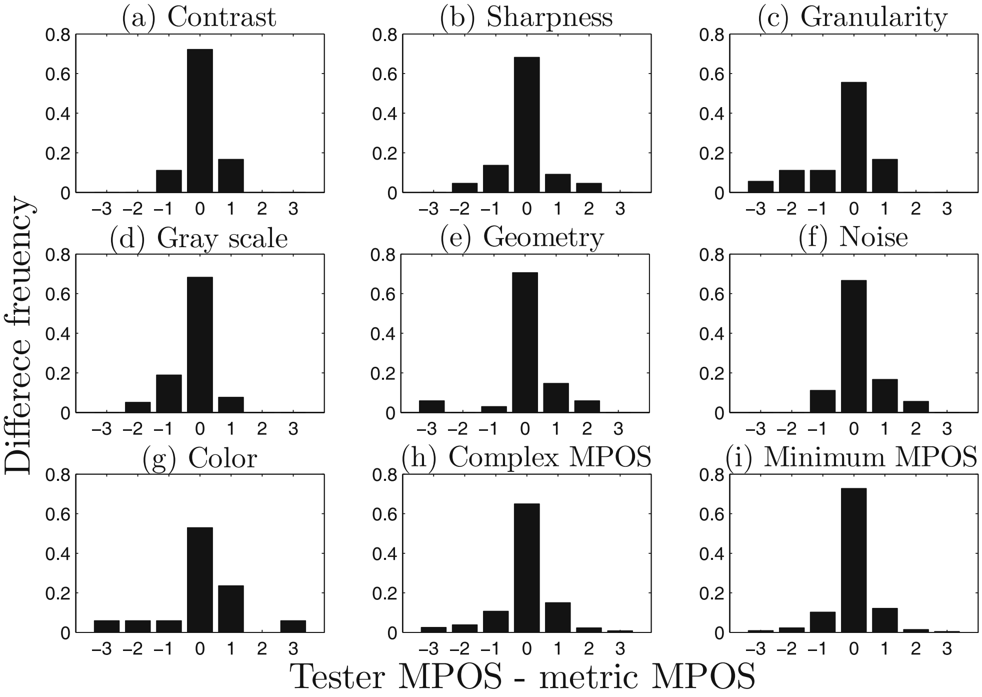

GLZ for overall quality score
OS distribution for overall quality score
MPOS
Complex MPOS metric
...
1st distortion assessment algorithm value
GLZ for 1st distortion
OS distribution for 1st distortion
MIN MPOS metric
MPOS
...
...
...
...
7 independent MPOS metrics
7th distortion assessment algorithm value
GLZ for 7th distortion
OS distribution for 7th distortion
MPOS

Fig. 14 The frequency of the answers difference for different metrics. **a** Contrast. **b** Sharpness. **c** Granularity. **d** Gray scale. **e** Geometry. **f** Noise. **g** Color. **h** Complex MPOS. **i** Minimum MPOS



An error not higher than −1 is obtained for 95% for both metrics for multi-modally distorted images. Moreover, both complex MPOS metric (Fig. 14h) and MIN MPOS metric (Fig. 14i) are accurate for almost 75% of the answers. The complex MPOS metric takes into consideration influences of different quality assessment values. Nevertheless, the complex MPOS metric is not much better than a simple minimum of the seven single distortion metrics. It shows that probably the most important from a tester point of view is the worst distortion level. Since the accuracy of the minimum MPOS metric is similar to the more complicated model, we implemented this solution as simpler and, therefore, more predictable.

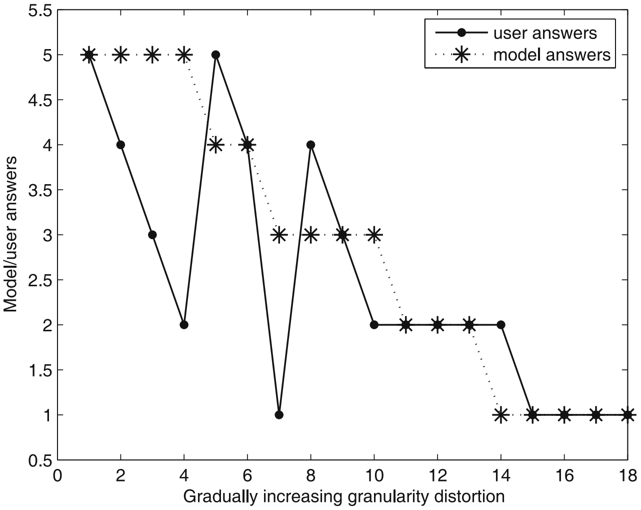


Fig. 15 Comparison between metric and the user answers obtained for the granularity distortion

6 Summary

The paper presented a solution to the problem of an automated evaluation of a subjective image quality. The authors designed a software tool for an evaluation of image quality implemented in the form of a Perl-based software package. The two-step procedure allows a user to compare a pair of images and receive information regarding the qualitative scores of a distorted image.

In the first step, seven types of distortion, covering possible image artefacts well, are being determined numerically. For numerical evaluation of some of the distortions, the basic functions of the ImageMagick software were found to be useful. For other cases, the algorithms were implemented by the authors from scratch. The problem of a mutual cross-distortion influence was identified as well and dealt with by compensation algorithms successfully.

During the second step, a mapping function is employed that transforms numerical distortion measures into scores equal or satisfactorily close to ones given by humans assessing the quality of the same image. The shape of the mapping function, together with its statistical credibility, was investigated and tuned with the sophisticated GLZ techniques based on results of extensive subjective tests for a reference image.

Acknowledgements The work presented in this paper was supported in part by Telekomunikacja Polska S.A. and the grants funded by EC (CONTENT FP6-0384239, INDECT FP7-218086) and Polish MNiSW (PBZ-MNiSW-02/II/2007 and N 517 4388 33).

References

1. Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
2. Aguirre-Torres V, Rios-Curil A (1994) The effect and adjustment of complex surveys on chi-squared goodness of fit tests: some Montecarlo evidence. In: Proceedings of the survey research methods section, pp 602–607
3. Bierens HJ (2004) Introduction to the mathematical and statistical foundations of econometrics. Cambridge University Press, Cambridge
4. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
5. Farias MCQ, Mitra SK (2005) No-reference video quality metric based on artifact measurements. In: IEEE international conference on image processing, ICIP 2005, vol 3, III - 141–4
6. Hosaka K (1986) A new picture quality evaluation method. In: Proc international picture coding symposium, pp 17–18
7. Imme M (1991) A noise peak elimination filter, CVGIP: graph. Models Image Process 53(2):204–211
8. ITU-T (1998) Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11
9. ITU-T (1996) Methods for subjective determination of transmission quality. Recommendation ITU-T P.800
10. ITU-T (2008) Objective perceptual multimedia video quality measurement in the presence of a full reference. Recommendation ITU-T J.247
11. ITU-T (2004) Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. Recommendation ITU-T J.144
12. ITU-T (1998) Standardized digitized image set. Recommendation ITU-T T.24
13. ITU-T (1999) Subjective video quality assessment methods for multimedia applications. Recommendation ITU-T P.910
14. Janowski L, Papir Z (2009) Modeling subjective tests of quality of experience with a generalized linear model. In: Proc QoMEX 2009
15. Miyahara M, Kotani K, Algazi VR (1998) Objective picture quality scale (PQS) for image coding. IEEE Trans Commun 46(9):1215–1226
16. OPTICOM GmbH (2007) Perceptual evaluation of video quality. http://www.opticom.de/technology/pevq.html