Theoretical and Applied Informatics
ISSN 1896-5334
Vol. 26 (2014), no. 1, 2
pp. 39 - 67

The survey of subjective and objective methods for quality assessment of 2D and 3D images

Sebastian Opozda[a], Arkadiusz Sochan[a]

Institute of Theoretical and Applied Informatics
Polish Accademy of Science
ul. Bałtycka 5, Gliwice, Poland
{sebastian, arek}@iitis.pl

Abstract: Evaluating the image quality is a very important problem in image and video processing. Numerous methods have been proposed over the past years to automatically evaluate the quality of images in agreement with human quality judgments. The purpose of this work is to present subjective and objective quality assessment methods and their classification. Eleven widely used and recommended by International Telecommunication Union (ITU) subjective methods are compared and described. Thirteen objective method is briefly presented (including MSE, MD, PCC, EPSNR, SSIM, MS-SSIM, FSIM, MAD, VSNR, VQM, NQM, DM, and 3D-GSM). Furthermore the list of widely used subjective quality data set is provided.

1. Introduction

Distortion of digital images is a common problem that results, among others, from imperfections of acquisition system and data compression. Therefore, Image Quality Assessment (IQA) plays a central role in shaping most visual processing systems. Initially it was used to assess television systems mainly in terms of presentation and transmission quality. Now, it is present at all stages of data processing such as: signal acquisition, enhancements, synthesis, compression, transmission, storage, retrieval, reconstruction and presentation.

The main purpose of the quality measurement of pictures and videos is the evaluation of human comfort of perception called Quality of Experience (QoE). Both individual characteristic of the observer and technical properties of presentation have significant impact on the image quality. Therefore, there are two kinds of methods for quality assessment: subjective and objective. Subjective assessment methods require human observers to evaluate the tested material. The scores are analysed to determine the objective indicators for image quality, taking into account Human Visual System (HVS). Objective assessment is generally used to evaluate the quality of the distorted image with reference to the original image. These methods are used to measure the impact of technical factors on the perceptual image quality.

The IQA can be basically used for two purposes. Firstly, it can be used to evaluate an influence of technical parameters on image perception. The technical parameters are described in section 3.1. The second purpose is to evaluate the perception of image quality in strictly defined environment. This scenario enables to determine the relationship between the image quality perception and e.g. degradation of material or content type. In both cases, the test environment must meet the general viewing conditions. Many technical parameters, test environments, test methods and datasets used in image quality assessments are described in standards and recommendations issued by the International Telecommunication Union (ITU) [1], the European Broadcasting Union (EBU) [2] or Video Quality Experts Group (VQEG) [3].

This paper reviews the state of art in the field of the quality of image assessment. In this paper, we give an up-to-date review of technical details and measurement methodology for image quality. We provide brief description of subjective and objective methods of evaluating quality of experience in 2D and 3D raster form images. The paper is organized as follows: section 2 contains information about test materials for quality assessment. In section 3 the comparison of subjective methods and requirements for test environment are described. Section 4 is dedicated to objective measurements of image quality. Sections 5 and 6 contain summary and acknowledgments.

2. Datasets in image quality assessment

Proper preparation of test materials is a critical element of the quality assessment framework. The dataset should contain a source signal - usually a raw and undistorted data acquired from digital or analogue source, and processed material for evaluation. In some cases it is not possible to obtain an undistorted picture or video, e.g. in reconstruction process, therefore, distorted signal can be used as source signal. The source signal provides directly the reference signal for comparison. In order to avoid distortion of the source signal, it should be stored in digital format.

The testing dataset may include single images (static material), image sequences (dynamic material) or both. There are many types of image distortions affecting the quality of their perception. A test dataset should represent the widest spectrum of tested distortion and should be prepared for the specific assessment task. In other words, the dataset should have proper spatial and temporal characteristics for reliable results. Spatial perceptual Information (SI) [4] quantifies the complexity of the spatial details in static or dynamic material. The SI is computed as fallows. Image is filtered using Sobel filter and next a standard deviation over the pixels is computed. The SI for image sequences is the maximum value of calculated SI for each image. Temporal perceptual Information (TI) [4] indicates the temporal changes in image sequence. The measure of TI is computed as the maximum value of the standard deviation of the differences between pixels at the same location on two consecutive images. It is usually higher for high motion sequences

The presentation of a test materials should be reproducible while maintaining the same quality. For this reason, storing data in digital form is the best solution - the test material can used numerous times without loss of quality. The reference signal should have a maximum quality, and should comply with the relevant technical requirements (e.g. resolution, fps, colours, and correct focus) depending on the test type. For certain types of distortion like JPEG or JPEG2000 compression there are publicly available databases for image quality assessment. The most common are presented in Appendix A and described in ITU Recommendation [5].

3. Subjective assessment

Subjective assessment of picture quality is used in digital television to evaluate the influence of an image codec on picture quality as described in [6, 7]. It is also used in computer systems for evaluating the overall audio-visual quality of multimedia applications such as videoconferencing, storage and retrieval applications [4]. The subjective quality assessment is based on scores assigned by observers to presented images or sequences.

Human observers are the final recipients in most image and video applications thus a subjective evaluation is the most accurate and reliable way. However, subjective methods have some drawbacks:

- they cannot be used in real-time applications,
- the results are influenced by physical conditions and emotional state of the observers,
- their results depend on viewing conditions,
- they are time consuming and expensive.

3.1. General viewing conditions

Human eye can adapt to wide range of light intensity. Thus, the environment in which the quality assessment tests are performed must meet the relevant requirements e.g. room illumination or screen size can significantly affect the perception of image quality as described in [8]. Depending on requirements two basic test environments can be distinguished: the laboratory environment and the home environment (Tab. 1). The laboratory environment is intended to provide critical conditions to the test system, while home environment is intended to provide a means to evaluate quality at the consumer side.

Parameter	**Laboratory environment**	**Home environment**
Room illumination	low	*Not specified*
Environmental illuminance on the screen	*Not specified*	200 lux
Chromaticity of background	D65[1]	*Not specified*
Peak luminance	70-250 cd/m2	70- 200 cd/m2
Ratio of luminance of inactive screen to peak luminance	≤ 0.02	≤ 0.02
Ratio of luminance of background behind picture monitor to peak luminance of picture	≈ 0.15	*Not specified*
Display size	≥20"	*Not specified*

Tab. 1. General viewing conditions

Recommendation [8] specifies two possible criteria for the selection of the viewing distance and screen size: Design Viewing Distance (DVD) and Preferred Viewing Distance (PVD). The DVD is used in digital systems and it described as the distance at which two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye. The Tab. 2 reports the DVD, expressed in multiples of the picture's height and optimal horizontal viewing angle for representative sample of display resolution.

The viewing distance and screen sizes are to be selected in order to satisfy the PVD. The PVD is based on experimentally defined preferences of viewers. Generally PVD is represented as ratio of viewing distance in meters to screen height (H) in meters [9]. For SDTV and HDTV the PVD is shown in Fig. 1.

[1] D65 – The Spectral Power Distribution (SPD) or chromaticity of white, representative of northern daylight and having a colour temperature of approximately 6504K.

Resolution	Aspect ratio	Optimal horizontal viewing angle	Design viewing distance (DVD)
720×483	4:3	11°	7 H
640×480	4:3	11°	7 H
720×576	4:3	13°	6 H
1024×768	4:3	17°	4,5 H
1280×720	16:9	21°	4,8 H
1400×1050	4:3	23°	3,3 H
1920×1080	16:9	31°	3,2 H
3840×2160	16:9	58°	1,6 H
7680×4320	16:9	96°	0,8 H

Tab. 2: Dependence on participants viewing distance and angle of the image resolution

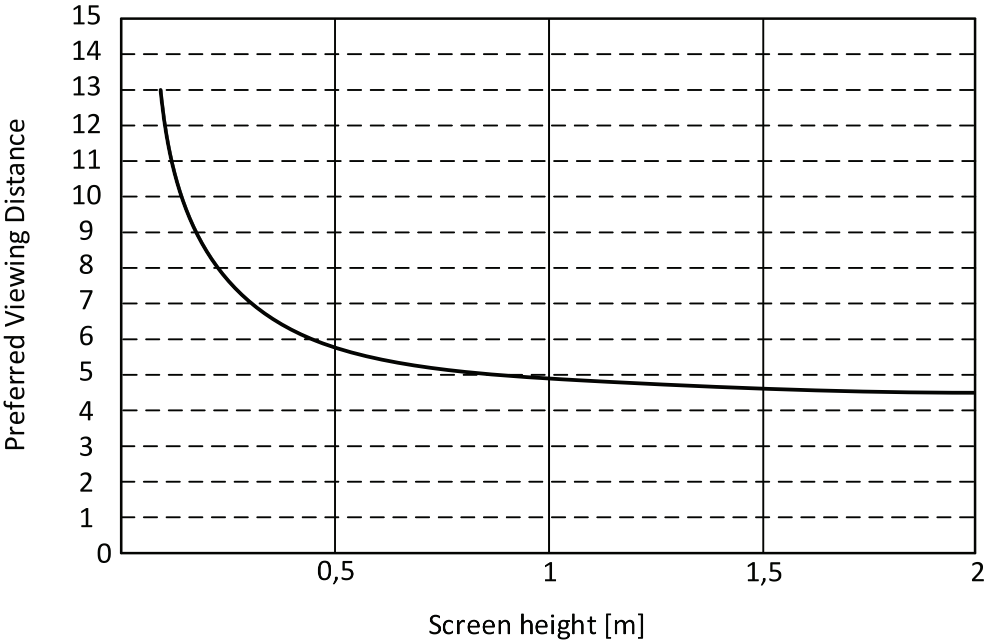

Fig. 1: Screen height relationship to PVD in H

3.2. Test participants

According to [4], test participants can be divided in two groups, based on their experience with the subject. Experts have general expertise with test methodology and image distortions, while non-experts represent typical understanding of a typical observer.

Prior to a session, the participants should be screened for normal visual acuity, colour vision, or in case of stereoscopic content for 3D depth perception. For the assessment for 2D images and sequences each observer should be screened for visual acuity (e.g. with Snellen charts), colour blindness (e.g. with Ishihara plates). Assessment of stereoscopic content needs additional tests. Stereovision is how each eye may see an object from different angles. Two points of view from different angles of observation allow the creation of a 3D image. Tests dedicated for stereoscopic vision are called stereopsis tests, and many of them "ask" the observer to identify the "rised" shape or letter to measure depth perception. The most popular and used tests are described in [10] and are as follows: Randot, Titmus Stereo Fly, and Frisby plates. More detailed analysis of the stereoscopic abilities of the observers is presented in [11].

The number of observers is strictly related to the test assessment type. According to [9] at least 15 observers should participate in the experiment. Some sophisticated tests may require only experts and in that case the participants group may be smaller (4-5 observers). As an example evaluating video codecs quality may require experts trained in the diagnosis of a specific distortion. Whereas, more extensive tests often require more non-expert assessors divided into several groups. Every participant must be familiar with the purpose of the test and any possible negative effects. This is especially important in case of assessment the quality of 3D materials where exposure to the stimuli can be uncomfortable. All the information for the participants should be described in the manual on paper or electronically, and the participants should be aware of them before tests.

3.3. Using reference signal

During the test, the observer assesses a series of images. They can be both original, not distorted images and processed images. The observer can also be informed about the type of the currently displayed signal, whether it is processed or original. Depending on this, we can distinguish three types of subjective assessment methods:

- with full reference – reference signal is present in data set and an assessor knows which signal is the reference one,
- with hidden reference – reference signal is present in data set but an assessor is not aware of which one is it,
- with no-reference – reference signal is not present in data set.

3.4. Index scores in image quality assessment

The observers score the test materials on a scale corresponding to their assessment of the quality-this is termed Mean Opinion Score (MOS). When a reference signal is used, additional indexes are available such as Difference Mean Opinion Score (DMOS) or Comparison Mean Opinion Score (CMOS). DMOS is the difference between reference and processed MOS. DMOS is computed using the fallowing formula:

$$DMOS = MOS(I_{proc}) - MOS(I_{ref}) + maxVal \qquad (1)$$

where I_{proc} is the processed image, I_{ref} is the reference image and $maxVal$ indicates maximum value on grading scale or continuous scale. In CMOS the evaluator observes the processed image, and the reference image, and make their assessment by comparing them.

Standard score, also called Z-score, can be used to compare observer's opinion about quality of images. The Z-score for i-th observer and j-th image can be computed using following equation:

$$z_{i,j} = \frac{DMOS_{i,j} - \bar{d}_\iota}{\sigma_i} \qquad (2)$$

where $\bar{d}_\iota$ is mean DMOS, and σ_i is a standard deviation. The $\bar{d}_\iota$ and σ_i are computed across all images that are rated by the i-th observer.

The variety of distortions that can occur in input signal excludes use of one simple judgements scale. These judgements can be divided into two main types: adjective categorical judgement and non-categorical judgement. The adjective categorical judgement is based on a set of predefined categories defined in semantic terms. The categories may reflect the existence of perceptible differences or extent of judgements as shown in Tab. 3. The observer assigns a category for each image, sequence or relation between signals. While in non-categorical judgement the observer can use an index on a scale that reflects its judged level on a specific dimension as presented in Fig. 2.

The most common scales include grading, continuous, and comparison scales. They are described in next subsections.

3.4.1. Grading scales

Scales used in quality assessment of images or sequences can have five, seven or eleven grades. The most common is the five-grade scale that is shown in Tab. 3.

Value	Impairment scale	Quality scale	Comfort scale
5	Imperceptible	Excellent	Very Comfortable
4	Perceptible, but not annoying	Good	Comfortable
3	Slightly annoying	Fair	Mildly Comfortably
2	Annoying	Poor	Uncomfortably
1	Very annoying	Bad	Extremely uncomfortably

Tab. 3: Examples of five-grade scale

3.4.2. Continuous scales

This scale is used when the viewer scores presented image, sequence or relation between presented materials using a point on the line drawn between two semantic labels. The scale may include additional labels at intermediate points for reference. The distance from an end of the scale is taken as the index. The descriptions from grade scales can be used as labels, as shown in Fig. 2.

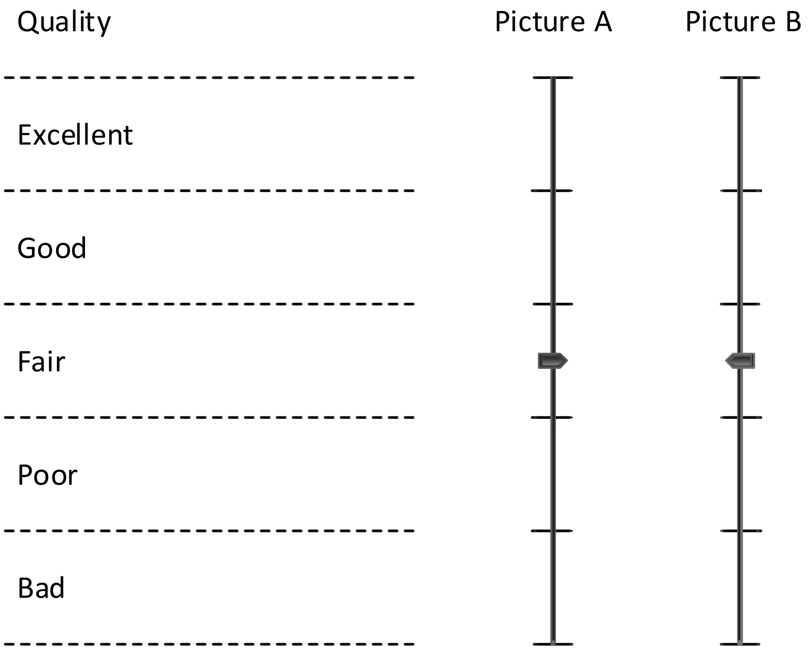

Fig. 2: Example of continuous scale

3.4.3. Comparison scale

The comparison is used to evaluate the relation between two stimuli. Typically, the pair of images are shown one after another. The stimuli may be also presented simultaneously. The participant compares two stimuli and assess the difference between them. An example of comparison scale is shown in Tab. 4.

Value	Score
-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

Tab. 4: Example of comparison scale

Method	Score	Material presentation			Reference signal	Suitable for many observers
		Stimulus	Simultaneous	Repetition		
DSIS	5-grade impairment	Double	No	Optional	Yes	Yes
DSCQS	Continuous	Double	No	Optional	Yes	Optional
SS	5-grade quality, 11-grade quality, non-category judgement	Single	No	No	Optional	Yes
SSMR	5-grade quality, 11-grade quality, non-category judgement	Single	No	Yes	Hidden	Yes
SC	7-grade comparison, non-category judgement	Double	Yes	Optional	Yes	Yes
SSCQE	Continuous	Single	No	Optional	No	Yes
SDSCE	Continuous	Double	Yes	Optional	Yes	Yes

SAMVIQ	Continuous	Single	No	Yes	Yes, hidden	No
ACR	5-grade quality	Single	No	No	Optional	Yes
ACR-HR	5-grade quality	Single	No	No	Hidden	Yes
DCR	5-grade impairment	Double	No	Optional	Yes	Yes
DCR-SP	5-grade impairment	Double	Yes	Optional	Yes	Yes
PC	7-grade comparison, non-category judgement	Double	No	Optional	Yes	Yes
PC-SP	7-grade comparison, non-category judgement	Double	Yes	Optional	Yes	Yes

Tab. 5 Comparison of subjective quality assessment methods

3.4.4. Subjective assessment methods

There are many subjective methods for assessing the quality of both static and dynamic 2D images. These methods are well known and documented in many standards and recommendations for picture quality assessment described in e.g. [4, 7, 11, 9].

Method	**Score**	**Material presentation**			**Reference signal**	**Suitable for many observers**
		Stimulus	**Simultaneous**	**Repetition**		
DSIS	5-grade impairment	Double	No	Optional	Yes	Yes
DSCQS	Continuous	Double	No	Optional	Yes	Optional
SS	5-grade quality, 11-grade quality, non-category judgement	Single	No	No	Optional	Yes

SSMR	5-grade quality, 11-grade quality, non-category judgement	Single	No	Yes	Hidden	Yes
SC	7-grade comparison, non-category judgement	Double	Yes	Optional	Yes	Yes
SSCQE	Continuous	Single	No	Optional	No	Yes
SDSCE	Continuous	Double	Yes	Optional	Yes	Yes
SAMVIQ	Continuous	Single	No	Yes	Yes, hidden	No
ACR	5-grade quality	Single	No	No	Optional	Yes
ACR-HR	5-grade quality	Single	No	No	Hidden	Yes
DCR	5-grade impairment	Double	No	Optional	Yes	Yes
DCR-SP	5-grade impairment	Double	Yes	Optional	Yes	Yes
PC	7-grade comparison, non-category judgement	Double	No	Optional	Yes	Yes
PC-SP	7-grade comparison, non-category judgement	Double	Yes	Optional	Yes	Yes

Tab. 5 presents a comparison of subjective assessment methods.

3.4.5. Double Stimulus Impairment Scale

The Double Stimulus Impairment Scale (DSIS or the “EBU method”) implies that two sequences with the same content are shown to the participant [9]. At first, the unimpaired image is shown (the reference phase), and then impaired one (the test condition phase). There are two variants of the structure of presentation. The materials can be presented either once or twice. The presentation time should be 10s each, separated with 2s of silence/grey image. The voting should be performed after test condition phase. Voting time should be less than or equal 10s.

3.4.6. Double Stimulus Continuous Quality Scale

The Double Stimulus Continuous Quality Scale (DSCQS) method is based on assessing a pair of pictures [9]. Depending on the number of observers, this method has two variants. In the first variant, the observer is able to switch between two signals until each signal is assessed. Evaluation time is 10s. In the second variant, the material is presented to many observers simultaneously. In this case the image pair is shown once or several times, with each repetition presented for the same amount of time. The presentation scheme looks as follows: For still pictures, it is a 3-4s sequence with five repetitions. For moving pictures, it is a 10s sequence with two repetitions. The results can be recorded during or after the last repetition.

3.4.7. Single Stimulus

In the Single Stimulus (SS) method the assessor provides an index of the entire presentation of images or sequences [9]. The material can contain only the test data or can include some reference images or sequences. Materials with the same impairment are presented only once in the test session.

There are three phases for a typical assessment trial: a mid-grey adaptation screen presented for 3s, a stimulus displayed for 10s and a mid-grey post-exposure screen displayed for 10s. The voting scores are collected either during display of the stimuli or during the post-exposure stage.

3.4.8. Single Stimulus with Multiple Repetition

The Single Stimulus with Multiple Repetition (SSMR) method is based on the SS method [9]. The main difference is that the pictures or sequences are presented three times. The test session is divided into three presentations, each of them including all the pictures or sequences to be tested with no repetition. A given picture or sequence cannot be located in the same position in the other presentations and cannot be located immediately before the same picture or sequence in the other presentations. The scores assigned to the pictures and sequences are computed as the mean score of the data acquired from the second and third presentations.

3.4.9. Stimulus Comparison

The Stimulus Comparison (SC) method consists in showing the pair of images or sequences simultaneously [9]. The observer provides an index of the relation between the two stimuli. The assessment trial generally proceeds as in single stimulus cases.

3.4.10. Single Stimulus Continuous Quality Evaluation

The Single Stimulus Continuous Quality Evaluation (SSCQE) is based on long sequences and it takes into account temporal variations of quality in digital transmission [9]. The assessment is continuous during presentation of the sequences. This continuous evaluation is carried out by means of sliders. There is no reference for anchoring the subjective assessment and it is not well suited for quality assessments that require a high sensitivity to distortions.

3.4.11. Simultaneous Double Stimulus for Continuous Evaluation

The Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method was developed for digital systems with lossy compression [9]. It consists in showing the pair of sequences simultaneously. The continuous assessments is carried out using sliders with values ranging from 0 to 100. At the beginning the training session is conducted. The main test starts when the viewer assigns maximum values only for undistorted or slightly distorted sequences.

3.4.12. Subjective Assessment of Multimedia Video Quality

The Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ) described in [12] has been designed for multimedia content. In this method, the viewer has access to several versions of a sequence. The different versions are selectable randomly and the viewer can stop, review and modify the score of each version of a sequence. The content of the sequence has to be homogeneous and should contain a wide range of spatial and temporal perceptual information. The hidden reference is mandatory but an explicit reference can be used as well. The image quality is assessed on a multimedia screen with slider mark on a scale. The grading scale is continuous and is divided in five equal portions. The presentation time for a sequence is typically in the range of 10 to 15s. The assessor may choose the order of tests and correct their votes, as appropriate.

3.4.13. Absolute Category Rating

The Absolute Category Rating (ACR) presented in [13] is a category judgement method providing no explicit reference design for multimedia applications. The test sequences are presented one at a time in random order and rated independently on a category scale. The presentation time of the test material should be about 10s but may be reduced or increased according to the content of the test material. The voting time should be less than or equal 10s. The hidden reference may be also used in this method. This version is called Absolute Category Rating with Hidden Reference (ACR-HR).

3.4.14. Degradation Category Rating

The Degradation Category Rating (DCR) described in [13] is a subjective video quality assessment method for multimedia applications. The test sequences are presented in pairs: the first stimulus presented in each pair is always the reference, while the second stimulus is the test data. The presentation time of each stimulus should be separated with 2s silence/grey image. The voting time should be less than or equal 10s. In this method stimuli may also be displayed simultaneously. This variant is named Degradation Category Rating with Simultaneous Presentation (DCR-SP).

3.4.15. Pair Comparison

The method of Pair Comparison (PC) implies that the test sequences are presented in pairs [13]. Each pair consists of the same sequence presented through different systems with different impairment. All the pairs of sequences should be displayed in all the possible combinations. The observer evaluates, which element in a pair is preferred in the context of the test scenario. The time pattern for this method is the same as for DCR method.

3.5. Subjective assessment methods in multimedia applications

Functionalities specified for computer systems cause development of new methods of image quality assessment. Their main domain of application are: Object-Based Evaluation (OBE) [13] and recognition tasks. The OBE assessment is realized in two stages. In first stage the whole image is assessed with ACR or DCR method. In second stage Video Objects (VOs) are extracted and object oriented assessment is conducted. Each of VOs is presented separately on a grey background and then assessed. The OBE is used for assessment of spatial scaling algorithms (used in H.264/SVC).

Subjective quality assessment methods used for recognition tasks are designed to assess the recognition of a specific target in an image or sequence for a specific task [14]. Examples of tasks include:

- human identification (including facial recognition),
- object identification,
- alphanumeric identification.

The dataset should span multiple scenarios taking into account different lightning conditions, different objects of interests, or small changes in scene.

3.6. 3D Image Quality

The main requirement of stereoscopic imaging is the presentation of at least two views of the same scene from two horizontally aligned points of view. The views can be synthetic (computer generated images, CGI) or recorded using cameras. The displacement or difference in the apparent position of an objects along two different lines of sight is called parallax or image disparity. Perception of depth depends on the parallax value. If the value is too small, the 3D image is viewed as a flat plane. If the value is too high, the 3D scene cannot be perceived.

We can distinguish the following effects, depending on stereo camera setup and placement of objects on the recorded scene: The positive parallax effect, when objects in stereoscopic image is perceived behind the screen. The negative parallax effect, when the objects are perceived in front of the screen. The Zero parallax effect, when objects are perceived directly on the screen plane.

The parallax of test materials should be in the range of comfort zone as shown in Fig. 3. The comfort zone should be between ±0.2D (dioptres) for negative parallax and ±0.3D (dioptres) for positive parallax based on [11]. For 1920x1080 screen resolution these values corresponds approximately ±2% and ±3% of screen parallax, defined as a ratio of the image disparity value on the screen to its horizontal size. Objects that are beyond that comfort zone can cause viewer discomfort and annoyance.

There are many factors that are characteristic for stereoscopic systems such as:

- depth resolution – spatial resolution in depth direction,
- depth motion – motion or movement reproduction along depth direction,
- puppet theatre effect – objects are perceived as unnaturally large,
- gigantism – object are perceived as unnaturally large,
- cardboard effect – object perceived stereoscopically are unnaturally thin,
- alignment effect – cameras misalignment,
- stereo window violation – object perceived in front of screen are clipped by screen frame,
- ghosting or cross talk – incomplete isolation of the left and right image channels.

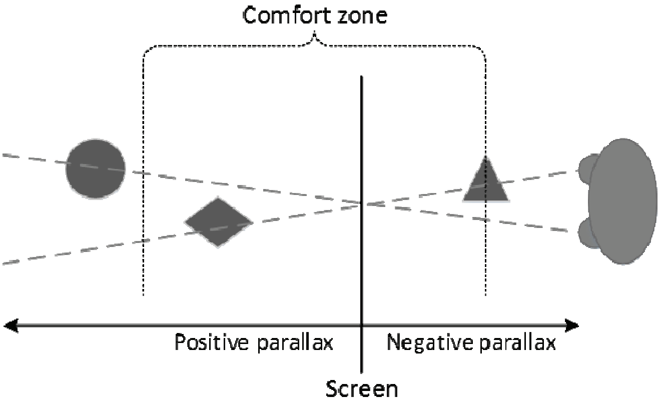

Fig. 3: Viewer comfort zone

All of these factors affect the quality of perception of stereoscopic materials and can be the subject of study. The ITU-R has released a recommendation [11] focused on picture quality, depth quality and visual comfort. The VQEGV also conducts research designed to assess 3D subjective video quality. SS, SDCQS, SC, and SSCQE were adapted from 2D to 3D

The length of the test session may be similar to 2D evaluation (about 20 -40 min). If the viewing material is known to be potentially uncomfortable (e.g. too big parallax or its fast and rapid changes), test duration should be shortened. The comfort scale can be used for evaluation of discomfort of viewers during test presentation.

4. Objective measurements

The digitalization of audio and video materials led to development of objective quality measurement methods. Storing the data in digital form provides unambiguous identification and repeated representation of the test material without deterioration of the quality. The objective measurements are generally used for evaluation of influence of the coding system and the transmission channel on quality of multimedia data presentation. They can be used to benchmark image processing algorithms or to monitor image quality in quality control systems. Proper use of an objective measurement requires subjective assessment as a reference evaluation. This enables to determine the appropriate measure for a specific type of distortion i.e. image blur, movement blur, edge business, false contouring, granular noise, jerkiness, blockiness, dirty windows effect.

The main purpose of objective quality measurements is to design the mathematical models that are able to automatically and accurately evaluate the quality of an image.

The objective measurements can be grouped depending on use of reference image and use of HVS.

4.1. Using reference signal

Based on the availability of a reference image, objective methods can be classified into three categories presented in Fig. 4.

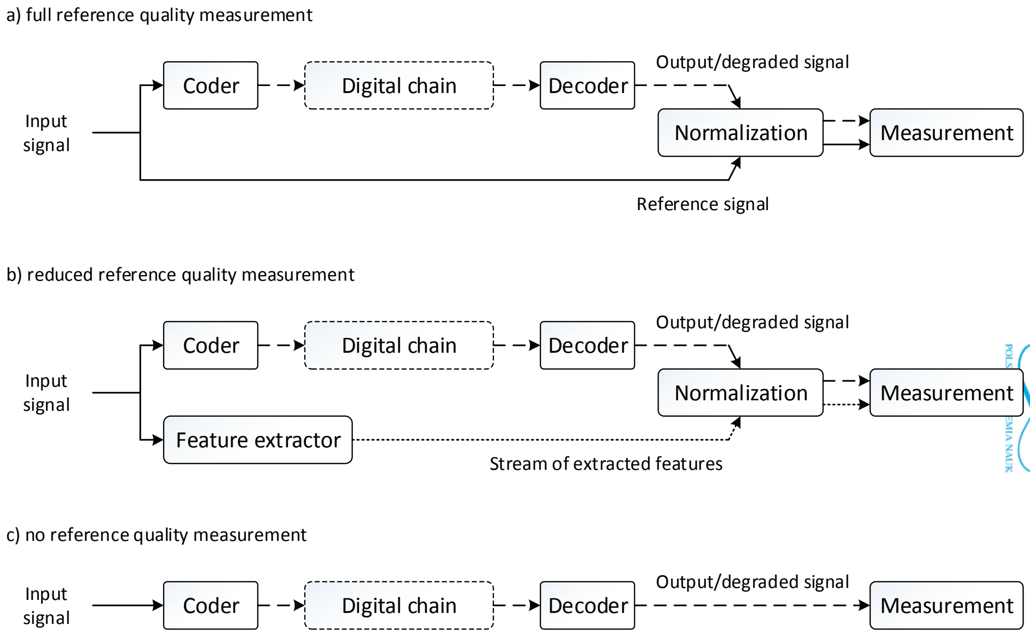

Fig. 4: Use of reference signal in objective quality measurement

4.1.1. Full Reference

The method with Full Reference signal (FR) presented in Fig. 4a evaluates the performance of the systems by making a comparison between the undistorted signal at the input on the system, and the degraded signal at the output of the system as proposed in [15, 16, 17]. The application scope of these metrics include compression or watermarking.

4.1.2. Reduced Reference

The method with Reduced Reference signal (RR) presented in Fig. 4b evaluates the performance of the system by making a comparison between features extracted from the undistorted signal at the input of the system, and features extracted from the degraded signal at the output of the system [18, 19, 20].

4.1.3. No Reference

Quality measurement with the No Reference signal method (NR) presented in Fig. 4c is a more difficult task in comparison to FR and RR. The quality evaluation is solely based on the processed image and reference image is not available. These methods are used to evaluate the quality of the particular type of known distortions e.g. for measure of smooth video playback or tilting effect, also known as blockiness, in video stream.

4.2. Human Visual System

The HVS model is used to deal with biological and psychological processes of human sight. Such model is used to simplify the behaviours of very complex human vision system.

4.2.1. Methods without HVS

These methods use well known statistical error measurements or correlation coefficients for images comparison. In the following subsection the most common methods are described. For simplification in subsequent formulas we introduce the following notation:

$$\sum_{(i,j)}^{(m,n)} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \qquad (3)$$

4.2.1.1. Mean Square Error

The Mean Square Error (MSE) denotes the difference between the reference and processed image. MSE for $m \times n$ monochrome image I_r and processed image I_p is calculated with the following equation:

$$MSE = \frac{1}{m \cdot n} \sum_{(i,j)}^{(m,n)} \left[I_r(i,j) - I_p(i,j)\right]^2 \qquad (4)$$

MSE is simple and computationally inexpensive method, defines the energy of an error signal. MSE satisfies properties like symmetry, differentiability and convexity and is widely used for optimization and assessment in wide range of signal processing applications. Unfortunately some of the physiological and psychological characteristics of the HVS are not considered by this measure. MSE is independent of temporal or spatial relationship between samples of the reference image. For colour images MSE is the sum over all squared value differences divided additionally by three

4.2.1.2. Peak Signal-to-Noise-Ratio

MSE is often converted to peak signal-to-noise-ratio (PSNR). The PSNR (in dB) is defined as:

$$PSNR = 10\, log_{10}\left(\frac{{MAX_I}^2}{MSE}\right) \qquad (5)$$

where MAX_I is denotes maximum possible pixel value of the image, dynamic range of pixel intensities, e.g. when pixels are represented using n-bits per sample MAX_I is equal $2^n - 1$. For colour images the PSNR can be alternatively reported against each channel in colour space YC_bC_r [21] or HSL [22].

4.2.1.3. Mean absolute Difference

The Mean absolute Difference (MD) is widely used in signal processing applications to investigate the similarity between two vectors. It is calculated as:

$$MD = \frac{1}{m \cdot n} \sum_{(i,j)}^{(m,n)} \left| I_r(i,j) - I_p(i,j) \right| \qquad (6)$$

MD is computationally very efficient and is often used in real-time applications.

4.2.1.4. Pearson correlation coefficient

The Pearson Correlation Coefficient (PCC) is used in pattern recognition, image processing and statistical analysis. For $m \times n$ monochrome image I_r and processed image I_p PCC is defined as:

$$PCC = \frac{\sum_{(i,j)}^{(m,n)} \left[(I_r(i,j) - \bar{I_r})\left(I_p(i,j) - \bar{I_p}\right) \right]}{\sqrt{\sum_{(i,j)}^{(m,n)} (I_r(i,j) - \bar{I_r})^2} \sqrt{\sum_{(i,j)}^{(m,n)} \left(I_p(i,j) - \bar{I_p}\right)^2}} \qquad (7)$$

where $\overline{I_r}$ and $\overline{I_p}$ are means intensity of images I_r andI_p respectively. PCC often fails to find differences in images that are widely disparate also often fails to detect missing objects within an image. The advantages and disadvantages of the PCC are described in [23].

4.2.1.5. Spearman Rank Correlation Coefficient

The Spearman Rank Correlation Coefficient (SRCC) is the Pearson correlation coefficient based on intensity ranks, instead intensity values. The intensity rank is the position of an intensity value, if all intensity values of the image were ordered. SRCC allows to measure any monotonic dependencies between images and is computed according to:

$$SRCC = 1 - \frac{6\ \sum_{(i,j)}^{(m,n)} d_{i,j}}{(m \cdot n)((m \cdot n)^2 - 1)} \tag{8}$$

where $d_{i,j}$ is the difference between the $I_r(i,j)$ rank and $I_p(i,j)$ rank.

4.2.2. HVS methods

Most HVS models in image processing use three basic properties of human vision:

- frequency sensitivity determines eye sensitivity to various spatial frequencies,
- luminance sensitivity measure the effect of the detectability threshold of noise on a constant background,
- masking effect determines the visibility of one signal in the presence of another signal.

The most common measures using HVS model are presented in the following subsections.

4.2.2.1. Edge PSNR

The Edge PSNR metric described in ITU-R recommendation [20] is based on the observation that degradations in regions around edges are very disturbing for human observers. This metric evaluates the PSNR only for these pixels that have been classified to belong to an edge region. The classification can be done using an edge detection operators [24]. An edge detection can be classified as follows:

- gradient edge detectors such as: Sobel operator, Prewitt's operator and Robert operator,
- Laplacian of Gaussian (LoG),
- Gaussian edge detectors such as: Canny operator.

4.2.2.2. Structural SIMilarity index

The Structural SIMilarity index (SSIM) method [25] assumes that human observers is highly adapted to process structural information from a scene and attempts to measure the change in this information between reference and processed image. This method defines image degradation as perceived change in structural information.

The structure of the objects in a scene is independent of local luminance and contrast. Finally the similarity measures the similarities of three elements of the image patches: the similarity $l(I_r, I_p)$of the local patch luminances (brightness values), the similarity $c(I_r, I_p)$ of the local patch contrasts, and the similarity $s(I_r, I_p)$ of the local patch structures:

$$SSIM(I_r, I_p) = l(I_r, I_p) \cdot\ c(I_r, I_p) \cdot\ s(I_r, I_p) \tag{9}$$

$$SSIM(I_r, I_p) = \left(\frac{2\mu_r\mu_p + C_1}{{\mu_r}^2 + {\mu_p}^2\ + C_1}\right)\cdot\left(\frac{2\sigma_r\sigma_p + C_2}{{\sigma_r}^2 + {\sigma_p}^2\ + C_2}\right)\cdot\left(\frac{\sigma_{rp} + C_3}{\sigma_r\sigma_p + C_3}\right) \tag{10}$$

where μ_r and μ_p are the local sample means of I_r and I_p, and σ_r and σ_p are the local sample standard deviations of I_r and I_p, and σ_{rp} is the sample cross correlation of I_r and I_p after removing their means. The C_1, C_2 and C_3 are small positive constants that stabilize each term, so that near-zero sample means, variances, or correlations do not lead to numerical instability.

4.2.2.3. Multi-Scale Structural SIMilarity index

The Multi-Scale Structural SIMilarity index (MS-SSIM) presented in [26] supplies more flexibility than SSIM by incorporating the variations of viewing conditions. This algorithm iteratively preforms the low-pass filtering and downsampling (by factor of 2) for the reference image and processed image. The highest scale as Scale M is obtained after $M - 1$ iterations. The MS-SSIM index is calculated using the following equation:

$$MS - SSIM(I_r, I_p) =\ [l_M(I_r, I_p)]^{\alpha_M} \cdot \prod_{j=1}^{M}[c_j(I_r, I_p)]^{\beta_j}[s_j(I_r, I_p)]^{\gamma_j} \tag{11}$$

The $c_j(I_r, I_p)$ and $s_j(I_r, I_p)$ denotes the contrast comparison and the structure comparison at the j-th scale. The luminance comparison is computed only at Scale M.

4.2.2.4. Feature SIMilarity index

The Feature SIMilarity index (FSIM) presented in [27] is low-level feature-based image quality metric. Two kinds of features are used: the High Phase Congruency (PC) and the Image Gradient Magnitude (GM). PC is used as a primary feature to extract highly informative features from image. The similarity measure for PC features is defined as follows:

$$S_{PC}(x) = \frac{2PC_r(x) \cdot PC_p(x) + T_{PC}}{PC_r^2(x) + \ PC_p^2(x) + T_{PC}} \tag{12}$$

where x is a point in 2D image space, PC_r and PC_p are 2-D PC map computed for reference and processed images, and T_{PC} is a positive constant to increase the stability. The 2-D PC map is computed using Kovesi method [28]. GM is used as the secondary feature to take into account contrast of the image. The similarity measure for GM is defined as follows:

$$S_{GM}(x) = \frac{2G_r(x) \cdot G_p(x) + T_{GM}}{G_r^2(x) + \ G_p^2(x) + T_{GM}} \tag{13}$$

where T_{GM} is a positive constant depending on the dynamic range of GM values, G_r and G_p represents gradient magnitude of reference image and processed image respectively. GM is computed along horizontal and vertical directions using gradient operators (e.g. Prewittt, Sobel, Scharr).

The similarity of reference and processed image is defined as follows:

$$S_L = [S_{PC}(x)]^{\alpha} \cdot [S_{GM}(x)]^{\beta} \tag{14}$$

where α and β are parameters used to adjust the influence of PC and GM features.

The FSIM is defined by:

$$FSIM = \frac{\sum_{x \in \Omega} S_L\,(x) \cdot PC_m(x)}{\sum_{x \in \Omega} PC_m(x)} \tag{15}$$

$$PC_m = max\ (PC_r(x), PC_p(x)) \tag{16}$$

where Ω is the whole image spatial domain, S_L is similarity measure between I_r and I_p. Detailed information can be found in [27].

4.2.2.5. Most Apparent distortion

The Most Apparent Distortion (MAD) index was presented by Larson and Chandler on 2010 in [29]. It models and uses two strategies employed by the HVS: detection-based strategy for high-quality images containing near-threshold distortions, and appearance-based strategy for low-quality images containing suprathreshold distortions. For detection a simple spatial-domain model of local visual masking is used. This takes into account the contrast sensitivity function, luminance and contrast masking with distortion-type-specific adjustments. The appearance-based strategy, first uses a log-Gabor filter for decomposition of the reference and processed image, and then compute the local statistical difference map. The MAD index is computed by taking a weighted geometric mean of the detection-based d_{detect} and appearance-based d_{appear} qualities described in [29]. The final index is computed as follows:

$$MAD = (d_{detect})^{\alpha}\left(d_{appear}\right)^{1-\alpha} \qquad (17)$$

where the weight parameter $\alpha \in [0,1]$ is determined based on the amount of overall level of distortions.

The MAD measure is relatively high computationally complex and memory consuming and doesn't detect colour distortions.

4.2.2.6. Visual Signal-to-Noise Ratio

The Visual Signal-to-Noise Ratio (VSNR) [30] via a two-stage approach operates on near-threshold and suprathreshold properties of human vision. In the first stage the visual detectability of the distortions is determined via wavelet-based models of visual masking and visual summation. The algorithm terminates if the distortions are below threshold of detection and the image is found to be perfect visual fidelity. Otherwise, a second stage is applied. In the second stage the low-level property of perceived contrast and the mid-level property of global precedence are used. The properties are modelled as Euclidean distances in distortion-contrast space. The VSNR in decibels is given by:

$$VSNR = 20\,log_{10}\left(\frac{C(I_r)}{\alpha d_{pc} + (1-\alpha)\frac{d_{gp}}{\sqrt{2}}}\right) \qquad (18)$$

where $C(I_r)$ denotes root-mean-sqared (RMS) contrast of the reference image, d_{pc} denote a measure of the perceived contrast of the distortions, d_{gp} denotes a measure of the extent to which global precedence has been disrupted, and the parameter $\alpha \in [0,1]$ determines the relative contribution of each measures. Detailed description of algorithm and is presented in [30].

The VSNR can accommodate different viewing conditions and is efficient in computation complexity and memory requirements.

4.2.2.7. General Video Quality Model

The general purpose Video Quality Model (VQM) presented in [31] for video systems. The VQM, including calibration, it's a complete automated objective video quality measure. The VQM calculation is based on extracting perception-based features and computing video quality parameters. The calibration includes spatial alignment, valid region estimation, gain and level offset calculation, and temporal alignment. The VQM consists of the following linear combination of the seven parameters given in [31]:

$$\begin{aligned} VQM = -0.2097 \cdot si_l + 0.5969 \cdot hv_l + 0.2483 \cdot hv_g + 0.0192 \cdot c_s - 2.3416 \\ \cdot si_g + 0.0431 \cdot ct + 0.0076 \cdot c_e \end{aligned} \qquad (19)$$

where:

- si_l parameter detects a decrease or loss of spatial information (e.g. blurring),
- hv_l parameter detects a shift of edges from horizontal and vertical orientation to diagonal orientation,
- hv_g parameter detects a shift of edges from diagonal to horizontal and vertical (e.g. tilting or blocking artefacts),
- c_s parameter detects changes in the spread of the distribution of two-dimensional colour samples,
- si_g parameter measures improvements to quality that result from edge sharpening or enhancements,
- ct parameter identifies moving-edge impairments (e.g. edge noise),
- c_e parameter detects severe localized colour impairments (e.g. produced by digital transmission error).

The VQM was standardized by ANSI in 2003 (ANSI T1.801.03-2003) and is also included in ITU Recommendation [32].

4.2.2.8. The Distortion measure

The Distortion Measure (DM) uses the lowpass Contrast Sensitivity Function (CFS) and the Distortion Transfer Function (DTF) described in [33]. To model HVS perception the DM penalizes low frequency more heavily than high frequency distortions and it is defined as follows:

$$DM = \sum_{(i,j)}^{(m,n)} \left|[1 - DTF(i,j)]CFS(i,j)\right| \qquad (20)$$

4.2.2.9. Noise Quality Measure

The Noise Quality Measure (NQM) [33] is a measure of the additive noise and is based on Peli's contrast pyramid presented in [34]. The NQM takes into account:

- variation in the local luminance mean,

- variation in contrast sensitivity with distance, image dimensions, and spatial frequency,
- contrast masking effects,
- contrast interaction between spatial frequencies.

The NQM in dB is defined as follows:

$$NQM = 10\, log_{10}\left(\frac{\sum_{(i,j)}^{(m,n)}\left(I_r(i,j)\right)^2}{\sum_{(i,j)}^{(m,n)}\left(I_r(i,j) - I_p(i,j)\right)^2}\right) \tag{21}$$

4.3. 3D Objective image Quality

All of the quality evaluation methods designed for 2D images are also applicable to 3D images. Additional information according to depth space perception in HVS should be also considered. The example of quality measure dedicated to 3D images is the 3D Gradient Magnitude Similarity (3D-GSM) presented in [35]. The 3D-GSM was elaborated for stereoscopic images based on Gradient Magnitude Similarity deviation (GMS) described in [36]. The 3D volume is constructed from stereoscopic images across different disparity spaces and calculate pointwise 3D gradient magnitude similarity along three horizontal, vertical and viewpoint directions. The 3D-GMS is defined as:

$$\text{3D-GSM} = \frac{1}{n \cdot m}\sum_{d}\sum_{(i,j)}^{(m,n)}\frac{2m_0(i,j,d)\cdot m_d(i,j,d) + C_4}{{m_0}^2(i,j,d) + {m_d}^2(i,j,d) + C_4} \tag{22}$$

where $C_4$is a constant to avoid the denominator being zero, $m_0(i,j,d)$ and $m_d(i,j,d)$ are the 3D gradient magnitudes of the reference and processed 3D volumes described in [35].

5. Discussion

The subjective quality assessment methods for 2D pictures and sequences are well known and described. They are used for evaluating the quality of experience in many standards and recommendations. The quality of experience for 3D materials is usually also carried out using methods designed for 2D materials. The subjective methods are simple to implement for this purpose but the test procedures are time-consuming and expensive. In contrary, objective methods of 3D quality assessment require additional research especially in the field of spatial depth perception and visual comfort.

The development of measures for assessing the quality of image perception is not a simple task. The knowledge about the type and characteristics of the occurring distortions in an image or an image sequence is essential. While simple, objective measures can be used to evaluate a particular type of distortion (e.g. PSNR), whereas measures, which take into account the HVS, require a more complex approach. These measures may consist of several simple objective measures and usually require additional steps like preparation of test datasets, preparation of test environment, conducting a set of subjective tests and performing objective tests, which complicates the whole process.

An interesting idea would be to create an objective measure to evaluate the 3D scene composition on the basis of its stereo pair presentation. This measure would verify the quality of immersion of observer in the 3D scene. Such measure would be based on structure and composition of presented scene, e.g. position, size and occlusion of objects. It would therefore take into account not only disparity space but also more complex features of depth perception according to HVS like comfort zone, space geometry, frustum culling and air geometry.

6. Acknowledgements

The work was partially supported by the project INNOTECH-K2/IN2/50/182645/NCBR/12 financed by National Science for Research and Development (A.Sochan) and by the project NN516405137 (decision DEC-2011/03/D/ST6/03753) financed by National Science Centre (S.Opozda).

Appendix A

In this appendix, we provide the list of datasets commonly used in quality image assessment:

- “A57” database described in “Online Supplement to “VSNR: A Visual Signal-to-Noise Ratio for Natural Images Based on Near-Threshold and Suprathreshold Vision”. Available at: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
- Subjective quality assessment - IVC database with 10 original images and 235 distorted images generated from 4 different processing algorithms (JPEG, JPEG2000, LAR coding, Blurring). Available at: http://www2.irccyn.ec-nantes.fr/ivcdb/
- Tampere image database TID2008 is intended for evaluation of full-reference image quality assessment metric. TID2008 contains 25 reference images and 1700 distorted images (17 types of distortion with 4 levels of distortions). Available at: http://www.ponomarenko.info/tid2008.htm
- LIVE Image Quality Assessment Database Release 2 contains scores from human subjects (DMOS) for a number of images distorted with different distortion types

(JPEG, JPEG2000, Gaussian blur, white noise, bit errors in JPEG2000 bit stream). Available at: http://live.ece.utexas.edu/research/quality/.

- MICT Image Quality Evaluation Database contains subjective scores for a number of images distorted with JPEG and JPEG2000 codec. Available at: http://mict.eng.u-toyama.ac.jp/mictdb.html.
- The CSQI image database consists of 30 original images, each is distorted using six types of distortion at four to five levels of distortion. The database contains 5000 subjective ratings reported in the form of DMOS. Available at: http://vision.okstate.edu/?loc=csiq.
- Wireless Image Quality (WIQ) database contains 7 reference images, 80 distorted images (JPEG, JPEG2000), and subjective scores for all images. Available at: http://www.bth.se/tek/rcg.nsf/pages/wiq-db.

References

[1] “International Telecommunication Union (ITU),” [Online]. Available: http://www.itu.int/.

[2] “European Broadcasting Union (EBU),” [Online]. Available: http://www3.ebu.ch/home.

[3] “Video Quality Experts Group (VQEG),” [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

[4] *ITU-T Recommendation P.911 Subjective audiovisual quality assessment methods for multimedia applications,* 1998.

[5] *ITU-R Recommendation BT.1210 Test materials to be used in subjective assessment,* 2012.

[6] *ITU-R Recommendation BT.1129 Subjective assessment of standard definition digital television (SDTV) systems,* 1998.

[7] *ITU-R Recommendation BT.710 Subjective assessment methods for image quality in high-definition television,* 1998.

[8] *ITU-R Recommendation BT.2022 General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays.,* 2012.

[9] *ITU-T Recommendation BT.500 Methodology for the subjective assessment of the quality of television pictures,* 2012.

[10] G. T. J. A. Pratt-Johnson, Management of Strabismus and Amblyopia: A Practical Guide, Thieme, 2001.

[11] *ITU-R Recommendation BT.2021 Subjective methods for the assessment of stereoscopic 3DTV systems.,* 2012.

[12] *ITU-R Recommendation BT.1788 Methodology for the subjective assessment of video quality in multimedia applications,* 2007.

[13] *ITU-R Recommendation P.910 Subjective video quality assessment methods for multimedia applications,* 2008.

[14] *ITU-R Recommendation P.912 The dataset should span multiple scenarios taking into account different lightning conditions, different objects of interests or small changes in scene.,* 2008.

[15] *ITU-R Recommendation BT.1907 Objective perceptual video quality measurement techniques for broadcasting applications using HDTV in the presence of a full reference signal,* 2012.

[16] *ITU-R Recommendation BT.1866 Objective perceptual video quality measurement techniques for broadcasting applications using low definition television in the presence of a full reference signal,* 2010.

[17] *ITU-R Recommendation J.247 Objective perceptual multimedia video quality measurement in the presence of a full reference,* 2008.

[18] *ITU-R Recommendation BT.1908 Objective video quality measurement techniques for broadcasting applications using HDTV in the presence of a reduced reference signal,* 2012.

[19] *ITU-R Recommendation BT.1867 Objective perceptual visual quality measurement techniques for broadcasting applications using low definition television in the presence of a reduced bandwidth reference,* 2010.

[20] *ITU-R Recommendation BT.1885 Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a reduced bandwidth reference,* 2011.

[21] *ITU-R Recommendation BT.601 Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios,* 2011.

[22] G. H. Joblove and D. Greenberg, "Color Spaces for Computer Graphics," *SIGGRAPH Comput. Graph.,* vol. 12, no. 3, pp. 20-25, #aug# 1978.

[23] K. Yen, E. K. Yen, R. G. Johnston, R. G. Johnston and P. D, *The Ineffectiveness of the Correlation Coefficient for Image Comparisons.*

[24] G. Padmavathi, P. Subashini and P. K. Lavanya, "Performance Evaluation of the Various Edge Detectors and Filters for the Noisy IR Images," in *Proceedings of the 2Nd WSEAS International Conference on Sensors, and Signals and Visualization, Imaging and Simulation and Materials Science*, Stevens Point, Wisconsin, USA, 2009.

[25] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on,* vol. 13, no. 4, pp. 600-612, April 2004.

[26] Z. Wang, E. Simoncelli and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, 2003.

[27] L. Zhang, D. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *Image Processing, IEEE Transactions on,* vol. 20, no. 8, pp. 2378-2386, Aug 2011.

[28] P. Kovesi, "Image features from phase congruency," *Videre: Journal of computer vision research,* vol. 1, pp. 1-26, 1999.

[29] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality

assessment and the role of strategy," *Journal of Electronic Imaging,* vol. 19, no. 1, pp. 011006-011006-21, 2010.

[30] D. Chandler and S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *Image Processing, IEEE Transactions on,* vol. 16, no. 9, pp. 2284-2298, Sept 2007.

[31] M. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *{IEEE} Transactions on Broadcasting,* vol. 50, no. 3, pp. 312-322, #sep# 2004.

[32] *ITU-T Recommendation J.149 Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM),* 2004.

[33] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans and A. C. Bovik, "Image quality assessment based on a degradation model," *Image Processing, {IEEE} Transactions on,* vol. 9, no. 4, pp. 636-650, 2000.

[34] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A,* vol. 7, no. 10, pp. 2032-2040, #oct# 1990.

[35] S. Wang, F. Shao, F. Li, M. Yu and G. Jiang, "A Simple Quality Assessment Index for Stereoscopic Images Based on 3D Gradient Magnitude," *The Scientific World Journal,* vol. 2014, 2014.

[36] W. Xue, L. Zhang, X. Mou and A. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," 2014.