

Chapter 2

Video Quality

Quality—you know what it is, yet you don't know what it is.

Robert M. Pirsig

Zen and the Art of Motorcycle Maintenance [256]

In video processing, often the subsequent question arises about the processed videos' quality and we face a dilemma similar to the protagonist in Pirsig's novel: we have a intuitive notion of the perceived videos' quality, yet find it difficult to provide an adequate description of its different aspects. Moreover, depending on the context and individual experience, different observers may arrive at different definitions of quality. Hence we need to agree on a generally accepted definition of video quality and corresponding methodologies that allow us not only to describe, but also to measure video quality. In this chapter, I therefore discuss the different concepts of quality. Starting with a general discussion on the possible definitions of quality, I review three quality concepts used in video processing: *Quality of Service (QoS)*, *Quality of Perception (QoP)* and *Quality of Experience (QoE)*. Based on these concepts, the definition of video quality used in this book is introduced, followed by an overview of methodologies and requirements to assess video quality subjectively.

2.1 What Is Quality?

Quality. A word intuitively used every day, yet elusive in its meaning. Before defining video quality, it is therefore useful to first review the possible meaning of the word *quality* itself. As we are aiming to describe the quality of video, we are interested in the interpretation and definition of quality with respect to a specific object or entity, in our case represented by the individual video sequences. Martens and Martens [201] suggest that there are at least four different common definitions of the quality of an object:

Definition 2.1 (*Quality as Qualitas*) Quality is the essential nature, inherent characteristic or property of an object.

The first definition considers quality as an intrinsic property of an object, providing us with an objective meaning of quality. Relating this definition to video, quality is represented by the objectively measurable features of a video. But this understanding of quality does not decide if all the measurable properties are relevant or not.

Definition 2.2 (*Quality as Excellence*) Quality is any character or characteristic which may make an object good, bad, commendable or reprehensible.

The second definition addresses the implicit relevance of the objects' properties by taking into account that human observers are evaluating these properties. It assumes a common understanding of the *excellence* or *goodness* of an object, of what is good and bad about an object. Again considered from the video perspective, this definition corresponds to the subjective assessment of a video's excellence or goodness by a human observer.

Definition 2.3 (*Quality as Standard*) Quality is the ability of a set of inherent characteristics of a object to fulfil requirements of interested parties.

The third definition considers quality as the degree to which an object's inherent characteristics fulfil given requirements. It combines aspects of the first definition with the second definition: inherent characteristics of an object are defined as descriptive quality criteria that are then used as specifications to assess the object's excellence, where the assessment need not be done necessarily by human observers. Translating this definition to video quality, it corresponds to the assessment of video with respect to the degree that the excellence criteria as defined by certain values of the features are satisfied.

Definition 2.4 (*Quality as Event*) Quality is not a thing, it is the event at which awareness of subject and object is made possible.

The fourth definition considers quality as something that is not only dependent on the object itself, but also on the event or occasion in which the object occurs and is perceived. Hence this definition takes into account the context in which an object is observed. This definition of quality therefore considers the interpretation of an object's intrinsic properties as dependent on the context, resulting in a differently subjectively perceived excellence of the object in different contexts. Martens and Martens call it therefore also the *lived quality* in [201]. Expressed in terms of the

above quality definitions, the *Quality as Qualitas* of the object leads to a different *Quality as Excellence* of the object depending on the context. For video quality, this definition can be understood to correspond to the subjective assessment of a video's excellence not only depending on the video itself, but also depending on the viewing conditions and the human observers' constitution, representing the context in which the video is viewed.

Each of these four definitions of quality can in principle be used as a foundation of a possible definition of video quality, depending on which aspect of quality we want to focus on. But as our aim is to assess how differently processed or distorted videos are perceived and experienced differently by human observers, the first definition of *Quality as Qualitas* is not suitable, as it only describes the intrinsic, objectively measurable properties of a video, represented by the features. Similarly, the third definition of *Quality as Standard* is also not suitable, as it uses only certain levels of the intrinsic, objectively measurable properties of a video to determine its goodness. Thus only the definitions of *Quality as Excellence* and *Quality of Event* are sensible foundations for a definition of video quality.

2.2 Quality of Service

Moving from the general consideration of the meaning of quality back to video processing, a commonly used quality concept associated with video quality is the *Quality of Service (QoS)* and one often used definition of QoS is given in the ITU recommendation ITU-T E.800 [121]:

Definition 2.5 (*Quality of Service—QoS*) Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.

Although aimed at telecommunications services, this definition can easily be adapted to our application area of video by substituting *telecommunication service* and *service* with *video*. Comparing this definition then to the general quality concepts in the previous section, QoS can be considered equivalent to *Quality as Excellence* in Definition 2.2, as it assesses the excellence of the object's characteristics with respect to a certain need of a user i.e. a human observer. This definition suggests that QoS could therefore be an adequate concept to describe video quality, as it implies an evaluation of the video's excellence by human observers.

Unfortunately, QoS in practical use is quite different from the above definition and is focused exclusively on the evaluation of objectively measurable signals degraded by processing or distortions [176]. It can therefore be considered as a pure fidelity measure. Hence QoS in reality is equivalent to *Quality as Qualitas* in Definition 2.2 and thus not suitable for a definition of video quality. Depending on the application

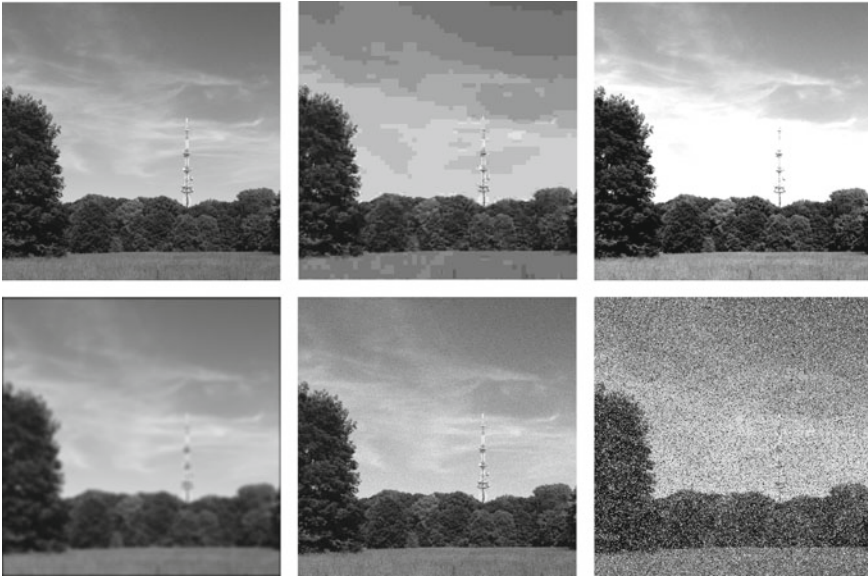


Fig. 2.1 QoS is insufficient: all images have the same MSE, yet clearly different visual quality depending on the influence of the distortion type on human perception

area and network type, different specifications for this interpretation of QoS exist as discussed in detail by Stankiewicz et al. [312].

MSE and PSNR—An example why QoS is not sufficient The inadequacy of QoS due its common interpretation as *Quality as Qualitas* for the description of *Quality as Excellence* can be demonstrated impressively on the example of the most popular QoS metric in image and video processing, the ubiquitous *mean squared error (MSE)* and the closely related *peak-signal-to-noise ratio (PSNR)*. Both provide a signal fidelity measurement with the average squared error between the original or unprocessed signal and the distorted or processed signal, where in our case the signal is represented by the pixels of an image or video frame.

Due to its simple definition and straightforward calculation, the MSE and PSNR are often used to describe video quality, thus equating QoS to video quality. The difference between distorted and undistorted image as represented by the MSE and PSNR, however, does not reflect the human quality perception adequately, as is illustrated in Fig. 2.1, where all distorted images have the same MSE, but clearly the perceived visual quality varies widely between the images. These shortcomings of the MSE and consequently the PSNR are well-known facts [73, 340, 372] and are discussed in detail by Wang and Bovik [339].

2.3 Quality of Perception

One concept less frequently used at least explicitly is the *Quality of Perception* (QoP), representing the *Quality as Excellence* in Definition 2.2. It addresses the shortcomings of QoS by using the subjectively perceived goodness of a distorted or processed video sequence as a description of its quality [55]. One possible definition of QoP is described in the ITU recommendation ITU-T E.800 by the *Quality of Service Experience* (QoSE) [121]:

Definition 2.6 (*Quality of Service Experienced—QoSE*) A statement expressing the level of quality that customers/users believe they have experienced.

QoSE focuses on a service's *level of quality* as perceived by the human observer and is thus describing the subjective detectability of quality changes caused by processing or distortions [55]. Note, that even though the above definition mentions *experience*, it focuses not on the overall experience itself but rather on a service's *level of quality* as the experience. Using QoP as a definition of video quality, we can then express video quality as the *Quality of Excellence*.

Perception and consequently QoP, however, can not be separated completely from the context or event. Or expressed differently, can perceiving be separated from experiencing? Although the desired separation may be possible to a certain degree in the design of psycho-visual experiments by using randomised patterns to examine perceptual properties of the human visual system, the laboratory environment still provides a specific context. Considering the determination of the perceptual quality of video in general, each subjective evaluation is performed in a specific environment and using a specific methodology. QoP is therefore always including the context of the evaluation. Thus in practical use QoP is implicitly equivalent to an interpretation of video quality as *Quality as Event*, but does not explicitly consider the context in its definition as *Quality of Excellence*. Due to this ambivalence QoP is therefore also not a suitable definition of video quality.

2.4 Quality of Experience

Quality of Experience (QoE) complements the signal fidelity focused QoS and purely perceptual QoP by aiming to capture the *quality* as truly subjectively experienced by considering video quality as *Quality as Event* according to Definition 2.4. One popular definition of QoE is provided in the ITU recommendation ITU-T P.10/G.100 [124]:

Definition 2.7 (*Quality of Experience—QoE*) The overall acceptability of an application or service, as perceived subjectively by the end-user.

NOTE 1—Quality of experience includes the complete end-to-end system effects.

NOTE 2—Overall acceptability may be influenced by user expectations and context.

This definition considers the subjectively perceived quality with respect to the human observer's expectations and context, thus it seems to provide a reasonable interpretation of *Quality as Event*. It does, however, define quality purely in the terms of acceptability, but following the argument by Möller [214], acceptability itself is based at least partly on the QoE. Therefore, this can not be a suitable definition of QoE.

In its Whitepaper on the Definitions of QoE [176], *Qualinet*, the *European Network on Quality of Experience in Multimedia Systems and Services*, provides a more holistic definition of QoE in the sense of *Quality as Event* by extending the definition of QoE beyond the pure acceptability as in the ITU-T definition [176]:

Definition 2.8 (*Qualinet: Quality of Experience—QoE*) Is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state.

The quality formation process resulting in a QoE rating according to Definition 2.8 is illustrated in Fig. 2.2. Input parameters of this process are the context, representing the nature of the event both externally and internally, in our case with the viewing environment and the human observers' constitution, and the physical signal itself, represented in our case by the video. The quality formation part consists of two overall parts, a quality perception path, describing the sensation and perception of the visual signal, and the reference path that translates the context of the quality event into the expectation of a certain desired quality. The experienced quality is then gained by comparing the desired quality with the perceived quality, finally describing the quality of the complete event as QoE. Thus Definition 2.8 of the QoE provides a suitable interpretation of *Quality as Event*.

Moreover, we can observe that QoS and QoP can be considered as contributing aspects to the overall QoE in this definition: firstly, the objective properties of the input signal described by the *Quality as Qualitas* with QoS influence the sensing of the stimuli at the beginning of the perception path. Secondly, the perception path resulting in the perceived quality represents the *Quality as Excellence* as described by the QoP. The definition triple of QoS, QoP and QoE resulting from the quality formation process is also similar to the triple model for QoE with a sensorial, perceptual and

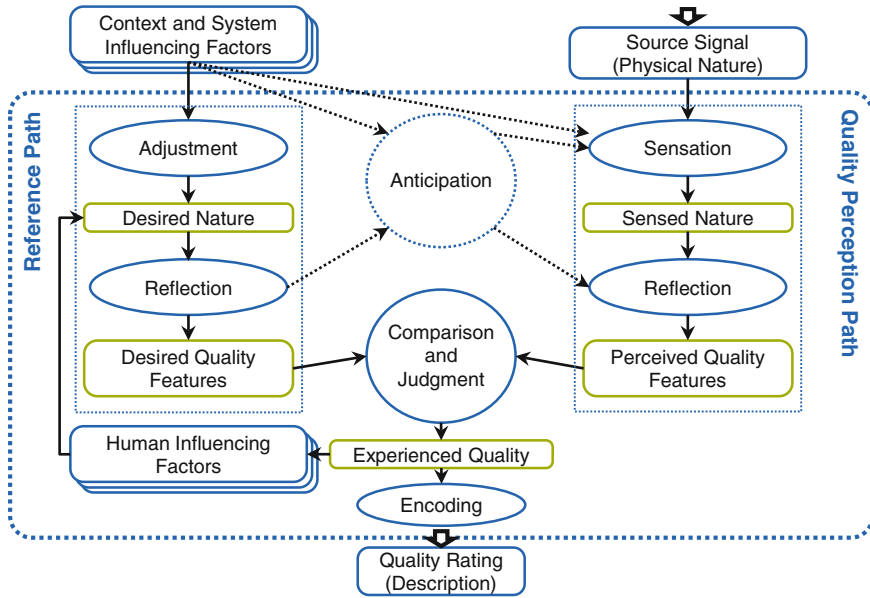


Fig. 2.2 Quality formation process resulting in a QoE rating: reference path on the *left* and quality perception path on the *right* (adapted from [176])

emotional step as suggested by Pereira [249, 250]: the sensorial part is represented roughly by the QoS, the perceptual part by the QoP and the emotional part by the reference path, resulting in the overall QoE. Additionally, this process supports the argument in Sect. 2.3 that QoP is not a suitable definition of video quality, as we have no access to the result of the perception path itself and thus the QoP: we can only observe the result of the quality formation process in the form of the QoE.

In practical use, however, the context is often not explicitly considered in the assessment of QoE, as QoE is usually assessed in formal subjective testing according to standardised methodologies and in standardised test environments. But it can be argued that this provides at least a standardised external context, even if the internal context of the subjects participating in a test may vary from subject to subject. Hence, if we consider the formal subjective testing as a sufficiently well-defined context and thus *event*, we can consider the results of these tests as a representation of the QoE for this specific context. Besides the definitions of QoE discussed above, alternative definitions have been proposed that explicitly include the business aspect e.g. Geerts et al. [71], Laghari and Connelly [169], Perkis et al. [251] and/or the influence of different demographics e.g. Geerts et al. [71], Laghari and Connelly [169] into the definition of QoE. For an in-depth discussion of QoE, I refer to Möller and Raake [215].

Based on these considerations and the *Qualinet* QoE definition in Definition 2.8, *video quality* in this book is therefore defined as a derivation of QoE:

Definition 2.9 (*Video Quality*) Is the degree of delight or annoyance of a subject in formal subjective testing, expressed by the resulting ratings on a specific scale provided to the subject. It results from the fulfilment of the subject's expectations with respect to the utility of the presented video sequences in the light of the subject's personality and current state.

On the one hand, this definition is less general than Definition 2.8, reflecting the focus on video and subjective testing, on the other hand, the enjoyment as one of the criteria for the subject's expectations is omitted, as in subjective testing the utility of the processed or distorted video is in the focus. Similarly, it is assumed that the subject's delight or annoyance is recorded on a specific scale provided to the subject and *formal subjective testing* in this context refers to the use of standardised testing methodologies and environments in the subjective testing.

Due to the fact that video in the context of video quality metrics is considered to only contain visual stimuli, *visual quality* is often used synonymously with *video quality* in order to emphasise that the quality evaluation is limited to visual stimuli. In this book, I therefore use *visual quality* when referring to the subjective ratings and/or prediction results of the metrics in order to highlight the visual nature of the assessment and prediction task, and *video quality* when referring to the overall quality evaluation or metrics.

2.5 Video Quality Assessment

Based on my definition of video quality, this section describes the two aspects that define *formal subjective testing*, the viewing environment and the used subjective testing methodologies. *Formal* in this context means that both the used methodologies and the other test parameters are not designed especially for each test, but are rather based on standards or well-established best practices. The overall goal behind using a formalised testing setup is the elimination or at least significant reduction of possible biases introduced due to the test environment or methodology. A comprehensive list of possible biases in subjective quality assessment is given in the review by Zieliński et al. [380]. Although only focusing on audio, the assumptions behind many methods are similar to video and therefore the review also provides an insight into the biases that are likely to be encountered in video quality assessment. The secondary, but often equally important goal of a formalised testing setup is to ensure the reproducibility of the results across different testing sites.

In the planning of a subjective test, the processing or distortions to be studied and the media or applications to be targeted must be defined e.g. the evaluation of new coding technologies for certain resolutions or the assessment of the influence of wireless channel errors on video communication. Depending on the overall goal of a test, different testing methods are chosen, but also the testing environment may be adapted if needed.



Fig. 2.3 ITU-R BT.500 compliant subjective testing laboratory at TUM: controlled lighting, colour-neutral mid-grey background, adaptability to different displays and test setups

2.5.1 Environment

The testing environment consists of the evaluation room and displays used to present the videos in the subjective testing. For both room and display properties, usually ITU-R recommendation BT.500 [115] is used as a guideline and depending on the overall test goal, modifications can be made on basis of this recommendation.

Room In order to avoid any unnecessary distractions from the visual assessment task for the subjects, the walls of the room are mid-grey as a colour-neutral compromise between a too dark or too bright background that could possibly conflict visually with the video shown on the display. Similarly, the test room should be sufficiently sound-proof to minimise distractions to the test subjects. Flicker-free, uniform lighting with a colour temperature of 6500 K equivalent to daylight provides the room's illumination and the lighting should be adjustable, so that a ratio of 0.15 between the peak illumination of the used display and the background illumination behind the display can be achieved. The seating of the test subject should allow for a variable distance between display and test subjects, depending on the used display's screen size. A typical example for a ITU-R BT.500 compliant video quality evaluation laboratory is shown in Fig. 2.3. Pinson et al. [254], however, recently suggested in their comparative international study that uncontrolled environments may also be suitable for video quality assessment, as many factors kept constant in the controlled environment, e.g. lighting or wall colour, do not seem to influence the results significantly.

Displays According to ITU-R BT.500 a reference monitor should be used for presenting the videos. Depending on the media or application under test, the display is calibrated to a specific colour gamut, gamma, white point and luminance. For the colour gamut, gamma and white point, ITU-R BT.601 [110] is used for standard definition television (SDTV) and ITU-R BT.709 [105] is used for high definition television (HDTV). The luminance should be 100–200 cd/m² in compliance with ITU-R BT.500 and a reasonable choice is 120 cd/m² as recommended in SMPTE

RP166 [305]. The viewing distance is depending on the display's screen size and expressed in terms of the screen height H . Common viewing distances are $3H$ for HDTV [104] and $6H$ for SDTV [103]. Instead of using an expensive reference monitor, a calibrated high-quality display can be a valid alternative as indicated by Keimel and Diepold [139].

Although ITU-R BT.500 still requires a CRT display for the video quality evaluation, both a study by VQEG [334, Appendix VII] and Pinson and Wolf [253] have shown that the results gained with LCD displays are statistically equivalent to the results gained with CRTs. Therefore LCDs are a valid contemporary choice to replace the increasingly rare and outdated CRTs. ITU-R BT.2022 [114] describes some additional considerations that should be taken when using non-CRT displays for subjective testing and the ITU-R recently published a recommendation for the viewing environment when using LCDs [117]. Pinson et al. [254] suggest additionally that unless the aim of test is in the assessment of different equipment, display calibration, display type and viewing distance may not have a significant influence on the results. Similar results were achieved in a smaller comparison of the results gained by performing a test with a reference display, a consumer display and a home cinema projector by Redl et al. [270].

Crowdtesting Crowdtesting describes the use of crowdsourcing to perform subjective testing. Instead of a dedicated video quality assessment laboratory, the subjective testing is performed distributed in the Internet using web-based applications. One obvious advantage is a more demographically and geographically diverse group of subjects, more representative of the general population. From a more economically point of view, the costs associated with subjective testing can be lowered significantly using crowdtesting, as on the one hand the reimbursement of the test subjects can be lower and on the other hand less investment in equipment has to be made.

In moving the subjective assessment into the Internet, however, many parameters in the subjective testing are no longer fixed and may not even be controllable at all. First studies by Keimel et al. [149, 150] have shown that crowdtesting can indeed provide results similar to the results from formal subjective testing in a laboratory environment. But as also discussed by Keimel et al. [148], many challenges still remain before crowdtesting is a universally acceptable replacement for formal subjective testing. For an in-depth discussion of and best practices for crowdtesting, I refer to Hoßfeld and Keimel [89] and Hossfeld et al. [92].

2.5.2 Testing Methodologies

Testing methods describe a set of certain aspects that define the test setup and process in detail. Often methods from the recommendations ITU-R BT.500 [115] and ITU-T P.910 [125] are chosen. Both suggest similar methods, but the former is focused on broadcasting, whereas the latter is focused on telecommunications applications. Instead of limiting this section only on the discussion of standardised methods, the

aim of this section is rather to provide an understanding of the assumptions on which these standardised testing methods rely.

One important aim of these methods is to avoid the introduction of additional biases into the subjects' rating due to the assessment task. Considering the quality formation process in Sect. 2.4, one source for introducing biases is the encoding or mapping of the subjects' quality rating to the quality scale provided by the used methodology. This is not a problem unique to video quality assessment and can be observed whenever judgements need to be quantified. Poulton [264] provides a comprehensive general overview and discussion about the different types of bias that can be encountered in quantifying judgements, and Zieliński et al. [380] elaborates the biases discussed in [264] in the context of subjective assessment for auditory stimuli.

Content selection The used video sequences should cover a sufficiently large variety of content for the targeted media or applications and the processing or distortion under assessment should be able to produce noticeably degradation in the selected content. Even though the content should be sensitive to degradations, it should still be realistic content, conceivably be part of real-life media and applications. This requirement is often expressed as *critical, but not unduly so* [115]. Additionally, the media and applications targeted in the test are usually also influencing the selection of the content. In the context of video quality assessment, an unprocessed video sequence containing a specific content is also often called *source*. For the majority of testing conditions, the video sequences are 10 s long to allow for a sufficiently large number of different test conditions within the test. Besides these economical considerations, the length of 10 s can also be justified by the fact that even if the video quality is assessed continuously, at most the last 9–15 s of the video are considered by the subjects in the quality assessment as demonstrated by Pinson and Wolf [252] and also indicated by Aldridge et al. [1]. Some general criteria for the selection of adequate content are suggested in ITU-R BT.1210 [111] and by Pinson et al. [255].

General structure Each test consists of two major parts: a training and the test itself as illustrated in Fig. 2.4. Both consist of multiple basic test cells (BTC), each representing one specific test condition and including a separate block for recording the subjects' quality rating. The separation of assessment and voting allows the subjects to exclusively focus on the assessment task while the video is shown. The content and test condition of two successive BTCs should be different in order to avoid the transfer of the bias from the preceding to the current BTC caused by the similarity in content and test condition [264]. In order to avoid fatigue in the test subjects, the test itself should last no longer than 30 min and it may therefore be necessary to split larger test into different test sessions, where each session has the same overall structure as outlined above. Regarding the influence of the assessment task on the viewing behaviour of the subjects, a study by Le Meur et al. [177] indicates that the subjects' eye movements do not change significantly between the free viewing of the video without assigned task and the viewing of the video with the task to assess its quality.

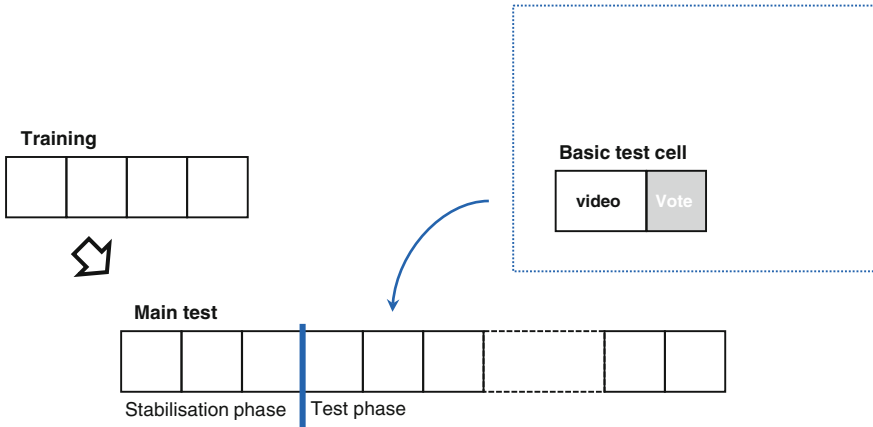


Fig. 2.4 General structure of a subjective test: training phase and main part, consisting of a stabilisation phase and the test phase itself

Before the test starts, the *training* provides the test subjects with an introduction to the test setup and methodology. Also it offers the subjects an opportunity to practice the assessment task and, if necessary, ask for assistance from the test supervisor. The instructions are usually given by the test supervisor and it is sensible to use a prepared script or presentation for these instructions, especially if the supervisor changes from subject to subject or the test is part of a joint campaign with other laboratories. In order to avoid influencing the subjects, the content in this training should be different from the test itself, but should still exhibit similar distortions as in the subsequent test. Should subjects be unable to perform the required assessment task they are consequently excluded.

After a successful training of the subjects, the test itself commences in two phases: first a stabilisation phase, followed by the test phase.

The *stabilisation phase* provides the subjects with an indication of the range of visual quality they will encounter during the test. The aim of providing quality anchors is to reduce or at least control the absolute contraction, centring and range equalising bias. Contraction bias describes the avoidance of the test subjects to use the extremes of the scales and thus a contraction of the subjects' ratings in the direction of the scale centre. The centring bias describes the tendency of subjects to shift their quality range towards the centre of the assessment scale, so that their ratings are symmetric towards the scale's centre. Lastly, the range equalising bias describes the effect that even if the subjects only use a small range of their inner quality scale, they map it to the complete range of the assessment scale [264].

By providing anchors for the complete quality range in the test including best and worst quality, the subjects are able to adjust their inner, intrinsic quality scale to the provided quality scale [264]. This reduces the contraction bias, as the subjects are now familiarised with the quality range they will encounter in the test, and controls the centring and range equalising bias as all subjects will now share the same intrinsic

quality scale. Usually the stabilisation phase consists of three to five video sequences [115]. Because the subjects are not aware of this implicit stabilisation phase, this provides an *indirect anchoring* [380]. The ratings gained from the subjects in the stabilisation phase are discarded before processing. If the test consists of multiple sessions, the study by Keimel et al. [151] indicates that the test conditions in the stabilisation phase of each session should be representative of the quality range in the complete test, not only of the quality range in the current session.

Following the stabilisation phase, the *test phase* itself commences. It consists of the BTCs representing the test conditions that should be assessed with respect to their video quality. For each BTC, a corresponding rating is recorded that provides the subjects' ratings.

Subjects Subjects participating in a test should be screened for normal (corrected) visual acuity and colour vision using e.g. Snellen and Ishihara charts for vision acuity and colour vision, respectively [115]. Although it is recommended to reject subjects failing these vision tests, Pinson et al. [254] recently suggested that slightly less than perfect visual acuity and colour vision do not seem to influence the quality assessment significantly.

Usually non-expert or naïve viewers are preferred. Non-expert in this context means that the subjects were not involved in defining the processing or distortions introduced in the test conditions and therefore have no preconceptions about the degradations of the video quality caused by specific distortions [115]. The assumption is that non-expert viewers are therefore more representative of the general population than experts. Experts, however, may be used in the design of the test in order to choose appropriate test conditions in a pilot test [125].

ITU-R BT.500 and ITU-T P.910 recommend to use at least 15 subjects in the subjective testing. Winkler [357] confirmed this minimum number of 15 subjects in simulations and based on the data from five different experiments, but Pinson et al. [254] recommend based on the results of their comprehensive study that at least 24 subjects should be used in a controlled environment, and at least 35 subjects should be used in an uncontrolled environment. Regarding the use of expert or non-expert viewers, Nezveda et al. [223] have suggested that when using expert viewers, fewer subjects are needed compared to using non-expert viewers in order to provide similar results. They caution, however, that using only comparably fewer expert viewers may only be suitable to identify general trends.

Rating scales Rating scales allow the test subjects to record their quality assessment for each of the BTCs. Depending on the aim of the test, discrete or continuous scales may either record the impairment or the absolute quality of the video sequences under test. In addition to indicating the range of the scale, corresponding labels for certain impairment or absolute quality categories are provided on the scale and depending on the method the scale's range is often also indicated with numerical labels. Usually the scales are divided in five-, nine- or eleven-point intervals, but still usually only five categorical labels are used. Examples of discrete and continuous scale are shown in Fig. 2.5.

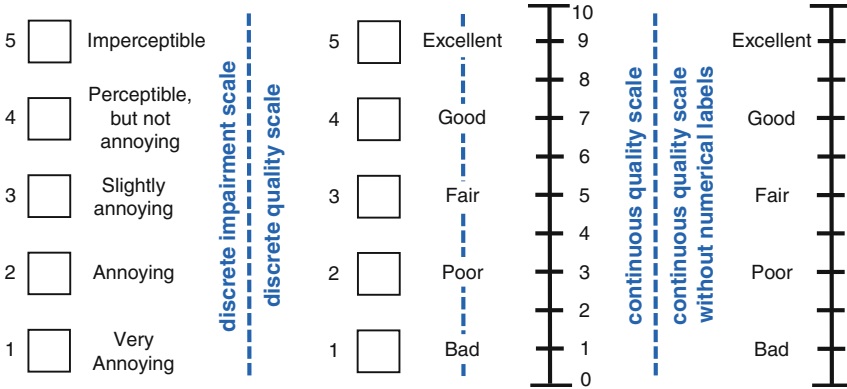


Fig. 2.5 Examples of rating scales used in subjective testing: discrete impairment scale, discrete and continuous quality scales with categorical and numerical labels as used in ACR [125], and continuous quality scale with only categorical labels as used in DSCQE [115]

Although the required scale is often explicitly defined in the corresponding standard for the used testing method, the used scales are often adapted depending on the test setup.

Continuous scales consist of a line with ticks in equally sized intervals and categorical labels indicating the position of certain impairment or absolute quality levels within the used scale. Often the ticks also provide additional numerical labels to indicate the available impairment or quality range. Even though a continuous scale suggests that more granular ratings could be achieved, quantisation effects usually occur, as test subjects tend to align their ratings with the labels and ticks, resulting in a quasi-discrete distribution of the ratings [380]. In the context of video quality, a study by Huynh-Thu et al. [97] and an eye-tracking experiment by Schleicher et al. [284] confirmed this behaviour.

Discrete scales are similar, but unlike a continuous scale, only discrete choices are available to the subjects: each discrete option is represented by a box with a categorical and numerical label indicating the corresponding impairment or quality level. Considering the quantisation effect for continuous scales, the difference in the ratings resulting from using discrete or continuous scales are often negligible as indicated in the two following studies: Huynh-Thu et al. [97] presented results for the single stimulus ACR method, indicating that for both discrete and continuous scales with absolute quality categorical labels there are no statistically significant differences in the subjects' ratings. Additionally, no difference for scales with five, nine or eleven ticks was found. For the double stimulus DSCQS and DSIS II methods, Corriveau et al. [47] also presented results that indicate equivalence not only between continuous and discrete scales, but also between impairment and quality scales. In addition, Svensson [320] suggests that discrete scales provide better stability with respect to the intra-rater agreement in different tests.

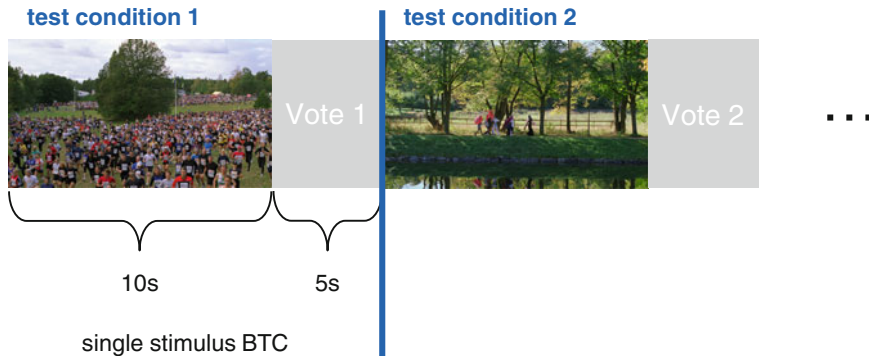


Fig. 2.6 Single stimulus BTC: presentation of the test condition, followed by a voting block

Especially in a larger, international test, a potential issue for both types of scales is the position of the categorical labels with respect to the range represented by the corresponding scale, as depending on the impairment or quality, the labels' position may not be a suitable representation of the labels interpretation in specific languages [347, 380]. This was confirmed for impairment and quality labels commonly used in video quality assessment in a study by the ITU-R [102]. Despite this, the recent study by Pinson et al. [254] suggest that for English, French, German and Polish differences in the interpretation of the labels do not influence the overall results significantly.

Single stimulus methods Single stimulus methods consist of BTCs with a straightforward structure: each BTC consists of one test condition, followed by a voting block indicating the number of the current BTC and presented on a neutral mid-grey background as illustrated in Fig. 2.6. This simple structure allows for short BTCs and with a video sequence length of 10s, followed by a voting block of 5s, each single stimulus BTC lasts 15s. Hence, we are able to achieve 4 BTC/min and with a maximum test duration of 30 min, we can therefore present 120 different test conditions per test.

One disadvantage of single stimulus methods, however, is their context dependency as they suffer from the sequential contraction bias: if the preceding test condition represented a high quality stimulus, the magnitude of the current stimulus' quality is overestimated, and if the preceding test condition represented a low quality stimulus, the magnitude of the current stimulus' quality is underestimated [264]. In order to avoid this, each test condition is shown multiple times during the test session in different context i.e. a different preceding test condition, and ITU-R BT.500 suggests three repetitions of each test condition in a test session. The rating for a test condition is then determined by averaging the results over all presentations, where ITU-R BT.500 also recommends to consider the first presentation as a stabilisation and therefore discard it in the averaging. In contrast, ITU-T P.910 suggest only two to four repetitions in total, but not for each test condition. Note, that with these repetitions the number of different test conditions per test is reduced accordingly. As a

side effect, the multiple ratings of the same test condition can be used to assess the consistency of a subject's rating as an indicator of their reliability [125].

In ITU-R BT.500 [115], two basic variants are suggested: variant I, the *single stimulus (SS)* method, and variant II, the *single stimulus with multiple repetition (SSMR)* method, where the later includes multiple repetitions of each test condition to minimise the context effect. Both variants are frequently used with a discrete five-point impairment or quality scale with categorical and numerical labels, unofficially called *single stimulus impairment scale (SSIS)* e.g. in [208, 230] and *single stimulus multimedia (SSMM)* e.g. in [232, 351], respectively. Although not specified in ITU-R BT.500, often discrete nine- or eleven point quality scales are used e.g. a eleven-point scale for the SSMM by Oelbaum et al. [232]. Less frequently used extensions of these types are the *single stimulus numerical categorical scale (SSNCS)* that uses a discrete eleven-point scale with only numerical labels, and the *non-categorical scale* that uses a continuous scale without any labels.

In ITU-T P.910 [125], the *absolute category rating (ACR)* method is described. It uses a discrete absolute quality scale with categorical and numerical labels. Depending on the required discriminability, five-, nine- or eleven-point scales may be used, but also an option to use a continuous scale is provided in ITU-T P.910. Similar to the SSMR method, multiple presentations of each test condition are advised for the ACR method in ITU-T P.910. A derivation of the ACR method is the *absolute category rating with hidden reference (ACR-HR)*, where an undistorted reference version of each content is included in the test, unknown to the subjects. For each subject, the rating of the hidden reference is then used to calculate a differential rating between each test condition and the corresponding undistorted reference with the same content. This can be considered as a rough correction of subjects' biases in the ratings at least with respect to the upper end of the provided quality scale.

Double stimulus methods Double stimulus methods extend the BTCs of the single stimulus methods with a *reference*, representing an undistorted version of the same content as in the test condition. Additionally, both the video representing the test condition and the reference video are usually repeated as illustrated in Fig. 2.7. This allows the subjects to gain an overall impression in the first presentation, followed by a detailed consideration of their rating in the second presentation. The video sequences representing the test condition and the reference are denoted with the letters *A* and *B*, respectively. Depending on the used method, the reference is either identified explicitly to the subjects or not. The order of the reference and the test condition is the same in each presentation, but each BTC may have a different order. Similar to the adaptation of the rating scales, both the order and the explicit identification of the reference are often adapted as needed.

One advantage of double stimulus methods is that the explicit reference in the BTC can be considered as *one-sided direct anchoring* [380]. Hence we can avoid on the one hand the sequential contraction bias that occurs in the single stimulus methods, on the other hand we provide a direct high quality anchor in each BTC comparable to the stabilisation phase, allowing the subjects to adjust their inner quality scale at least to the upper end of the provided quality scale in each BTC anew. Although multiple replications of each test condition are therefore not necessary, ITU-T P.910 suggests



Fig. 2.7 Double stimulus BTC: presentation of reference *A* and test condition *B*, repeated once and then followed by a voting block

also for double stimulus methods two to four repetitions in total for reliability testing of the subjects [125]. Corriveau et al. [47] have provided experimental evidence that for double stimulus methods, in particular the DSCQS method, the context dependency is indeed significantly lower compared to single stimulus methods.

One practical disadvantage, however, are larger BTCs: assuming again a video sequence length of 10 s, a voting block length of 5 s and a *A/B* label block of 2 s, each double stimulus BTC lasts 53 s. Hence, we are able to achieve approximately 1 BTC/min and with a maximum test duration of 30 min, we can therefore present only 30 different test conditions per test, compared to up to 120 test conditions per test for the single stimulus methods.

In ITU-R BT.500 [115], two double stimulus methods are described: the *double stimulus impairment scale (DSIS)* method and the *double stimulus continuous quality scale (DSCQS)* method. The DSIS method uses a discrete five-point impairment scale in two variants: variant I that consist only of one presentation of test condition and reference, and variant II that consists of two presentations. In a BTC of the DSIS method, the reference is always the first video i.e. *A* and the test condition is the second video i.e. *B*. Moreover, this fixed order is announced to the subjects. Variant I of the DSIS method is also defined as the *degradation category rating (DCR)* method in ITU-T P.910 [125]. The DSCQS method is aimed at assessing the difference in quality between two videos. It uses two continuous scales with categorical labels and ticks in equal intervals for each BTC. Unlike the DSIS method, the subjects are not aware which video is the reference and they have therefore to provide separate ratings for video *A* and *B*, resulting in a differential rating $A - B$ for each BTC.

Baroncini [12] argues that this repetitive dual rating task in the DSCQS method leads to increased fatigue in the test subjects and therefore suggests an alternative to the DSCQS method based on a modification of the DSIS, variant II method, the *double stimulus unknown reference (DSUR)*. Unlike the DSIS method, the subjects are unaware if the reference is the first or second video. In the first presentation, the

subjects are therefore asked to identify if A or B is the reference, and only in the second presentation they should then rate the non-reference video. Thus no repetitive task as in the DSCQS method is required of the subjects, but rather two different, though still related tasks must be performed by the subjects. In [12] a five-point discrete impairment scale with categorical and numerical labels is proposed, but it is often used with an eleven-point discrete quality scale e.g. in [232].

Other Methods *Continuous quality evaluation* methods aim to assess temporal quality changes or fluctuations by presenting the subjects longer video sequence of up to five minutes and allows them to assess the quality continuously using a slider. Both single stimulus and double stimulus methods are defined in ITU-R BT.500 [115]: the *single stimulus continuous quality evaluation (SSCQE)* method and the *simultaneous double stimulus for continuous quality evaluation (SDSCE)* method. For the SSCQE, only the distorted video sequence is presented, whereas for the SDSCE both the distorted and the reference video sequence are presented simultaneously on a shared display or two different displays.

The interactive *Subjective Assessment of Multimedia Video Quality (SAMVIQ)* method defined in ITU-R BT.1788 [107] allows the test subjects to assess different test conditions of each content as often as needed in an interactive interface, including the undistorted reference for comparison. The rating itself is performed on a continuous quality scale with categorical and numerical labels from 0–100. Péchard et al. [243] have shown that SAMVIQ provides comparable results to ACR, but with fewer subjects. Due to its interactive approach, however, a SAMVIQ based test takes more time compared to an ACR based test.

Pair comparison methods are an alternative to the methodologies discussed so far. Instead of quantifying judgements, the subjects compare in each step two test conditions and decide which one is the preferred or better one. As the subjects do not need to quantify their judgements on a scale, we avoid the biases introduced by the mapping of the internal quality to the provided scale. By estimating the probability of choosing a test condition from the conditional probability that a certain test condition is preferred against another test condition with maximum-likelihood estimation, we can then use the Bradley-Terry-Luce model to obtain the rating of a test condition based on its probability [20]. One general disadvantage of the pair comparison, however, is that each test condition needs to be compared with every other test condition and assuming N different test conditions, this leads to $N(N - 1)$ pair comparisons, which leads to significantly more BTCs than for the other methodologies. The Pair comparison variant as described above is recommended as *pair comparison (PC)* method in ITU-T P.910 [125]. ITU-R BT.500 [115] additionally suggests variants that also include a scale to quantify the magnitude of the difference, but this may reintroduce biases into the ratings.

Processing of the results After completing the test, the ratings of all subjects are averaged for each test condition, resulting in a test condition's *mean opinion score (MOS)*, representative of its video quality. If only difference ratings are available e.g. for the ACR-HR method, the average is called *differential mean opinion score (DMOS)*. As an indication of the MOS' uncertainty, usually the 95 % confidence interval is additionally provided for each MOS.

In order to identify unreliable subjects, it is suggested in ITU-R BT.500 and ITU-T P.910 to perform a screening of the subjects according to a statistical criterion and reject those subjects' ratings that fail the screening. Pinson et al. [254], however, argue that only subjects that didn't understand the assessment task should be eliminated as it is often unclear why different subjects respond differently to the test conditions and therefore a strict statistical outlier removal may not be suitable. Still, outlier screening is usually done in subjective testing and two often used statistical criteria are discussed briefly in the following paragraphs.

ITU-R BT.500 [115] proposes to screen the subjects by comparing the rating of each subject for a test condition to the MOS of this test condition. If a subject's rating for a test condition deviates more than the sample standard deviation for this test condition multiplied by a factor that is dependent on the ratings distribution, counters are incremented, one if the subject's rating is too low, another if a subject's rating is too high. These counters therefore provide not only the number of rejected ratings, but also indicate the reason for the rejection. Should both the number of rejected ratings for a subject be above a certain threshold and the ratings are not exhibiting any recognisable offset, the subject is completely rejected and all its ratings discarded.

VQEG [334, Appendix V] suggests a simpler method by calculating first the Pearson correlation coefficient r_p between a subject's ratings and the MOS for all test conditions, followed by a comparison with a threshold. If $r_p < 0.75$, the subject is rejected and all its ratings are discarded.

Design of Video Quality Metrics with Multi-Way Data
Analysis

A data driven approach

Keimel, C.

2016, XV, 240 p. 52 illus., 2 illus. in color., Hardcover

ISBN: 978-981-10-0268-7