6 Quality evaluation of 3D video

Your goals for this “Quality evaluation of 3D video” chapter are to learn about:

- Perceptual quality evaluation of 3D video.
- Subjective 3D video quality evaluation methods.
- Objective 3D video quality evaluation methods.
- Real-time 3D video quality evaluation methods.

Even though the initial developments for 3D video services are in place, the acceptance of these services is dependent on the user satisfaction of the reconstructed 3D video quality. Therefore, extensive quality evaluation studies are necessary to study the effect of camera arrangement, data representation, coding, transmission and display techniques on the perceived quality of 3D video. Some of the stereoscopic image impairments introduced by the 3D video system are keystone distortion, depth-plane curvature, crosstalk, size distortions, cardboard effect, picket fence effect, image flipping and shear distortion. Moreover, depending on the coding approaches (e.g. DCT) being used, conventional coding impairments like blockiness, blur will be introduced to the reconstructed 3D video. These impairments in stereoscopic video will influence multi-dimensional perceptual attributes such as image quality, depth perception, presence, naturalness, etc. A detailed analysis is necessary to study how these 3D percepts influence the overall perceived quality in general. For instance, the study presented in [100] concludes that excessive disparities can cause eye strain and therefore degrade the perceived image quality. Mostly psychophysical experiments are conducted to measure and quantify 3D perceptual attributes. In addition to that, explorative studies can be utilized to get unbiased attitudes and views for emerging technologies like 3D video. For instance, focus groups can be formed to evaluate the impact of new stereoscopic image systems through group discussions [101]. This method also can be employed to evaluate the added value of depth. Moreover, explorative studies will help better understanding the attributes of a multi-dimensional construct like image quality, depth perception, viewing experience, etc.

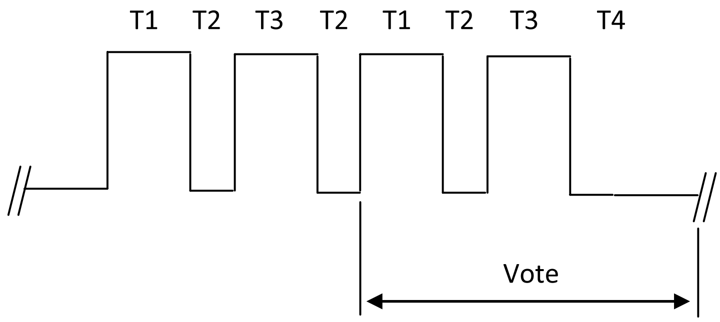
Psychophysical scaling paradigms can be classified into two main categories [102], namely;

- Performance-oriented methods
- Appreciation oriented methods.

The performance oriented assessment methods are utilized to measure the effectiveness of a specific task whereas appreciation oriented methods measure and quantify the perceptual attributes of new media types and decide whether the content is pleasing or not. Appreciation oriented quality evaluation methodology for stereoscopic TV pictures is described in ITU-R BT.1438 recommendation [103]. Most of the subjective evaluation procedures in this recommendation are based on the ITU quality evaluation recommendation for TV pictures (i.e. ITU-R BT.500.11) [104]. In addition to the measurement of image quality, other 3D perceptual attributes like presence, naturalness and eye strain can be measured using the same experimental paradigms. The main evaluation strategies mentioned in [103] are;

- Single-Stimulus-Continuous-Quality-Scale (SSCQS) method: The quality is assessed individually for each stereoscopic image sequence in the stimulus set.
- Stimulus comparison method: Series of stereoscopic image sequences are presented sequentially in time and observers are asked to assign a relation between two consecutive stereoscopic video sequences
- Double-Stimulus-Continuous-Quality-Scale (DSCQS) method: Alternately, an unimpaired stereo image sequence (reference) and an impaired stereo image sequence (test) are shown. The reference and test image sequences are presented twice. For both stereo image sequences (reference and test) observers assess the overall picture quality separately.

The presentation method of DSCQS is illustrated in Figure 6.1. The final quality rating (i.e. opinion score) of this method is the difference of individual scores for the reference and impaired image sequences. Subsequently the individual opinion scores are averaged across all the subjects in order to obtain the Mean Opinion Score (MOS). The confidence intervals can also be specified to indicate the individual differences.



Phases of presentation:

T1 = 10 s	Test sequence A
T2 = 3 s	Mid-gray
T3 = 10 s	Test sequence B
T4 = 5-11 s	Mid-gray

Figure 6.1: DSCQS presentation structure

The DSCQS and SSCQS methods are utilized in most the experiments described in this book as main subjective quality evaluation methods as these methods are recommended by standardization bodies for stereoscopic video quality measurements and are in wider usage in 3D video research [103], [105-107]. Furthermore, all subjects are screened for their visual acuity (using the Snellen chart), good stereo vision (using the TNO stereo test), and good colour vision (the Ishihara test). Moreover, the 3D displays will be calibrated using the GretagMacbeth Eye-One Display 2 calibration device and test environments (e.g. home viewing conditions) will be set according to the specifications set by ITU-R BT.500.11 recommendation.

The perceptual quality of asymmetrically coded colour and depth map sequences are measured and evaluated in [130]. Moreover, the effect of packet losses on the perceptual quality is also studied. The quality is measured across two perceptual attributes namely, image quality and depth perception. Furthermore, the relationships are derived among these measured perceptual attributes.

Subjective tests for each 3D video system parameter (e.g. camera angle, coding) change is not an efficient method to evaluate the quality due to several reasons. The most prominent reasons are the time consumption, enormous effort necessary, and the requirements for special test environments (e.g. standard test laboratories). Therefore, candidate objective quality measures of 3D video have become a compromise way of measuring the quality.

Therefore, candidate objective quality measures (i.e. PSNR) of colour image sequence and depth image sequence are utilized to represent the effectiveness of proposed algorithms in this book. PSNR is derived by setting the Mean Squared Error (MSE) in relation to the maximum possible value of the luminance (see Equations 6.1 and 6.2).

For n-bit value this is as follows,

$$MSE = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[g(i,j) - G(i,j)\right]^2}{M \cdot N}$$

Equation 6.1

$$PSNR = 20 \cdot \log_{10}\left(\frac{2^n - 1}{\sqrt{MSE}}\right)$$

Equation 6.2

Where g (i,j) is the original signal at pixel (i,j), G (i,j) is the processed signal and M × N is the picture size. The resultant is a single number in decibels (dB).

Even though PSNR scores of depth image are indicative, it may not represent the depth as perceived by the human observers. Therefore, the objective quality measures of rendered left and right views using the DIBR method are also used to quantify the depth perception. In order to obtain PSNR ratings the left and right video rendered using the impaired colour and depth image sequences are compared against the left and right video rendered using the original/reference colour and depth map sequences.

These objective measures may or may not strongly correlate with the quality attributes of 3D video as measured by subjective tests. Studies have found out that there is a high correlation between subjective ratings and individual objective quality ratings of 3D video components (e.g., average PSNR and SSIM of left and right video or colour and depth video) [123]. For instance, depth perception is highly correlated to the average PSNR of the rendered left and right image sequences [123]. This could be due to the loss of correspondence between left and right objects and reduction of monocular depth cues as a result of compression and transmission errors. This means that we could use individual objective quality measures of different 3D video components to predict the true user perception in place of subjective quality evaluation, through a suitable approximation derived based on correlation analysis. However, with some 3D source representations such as the colour and depth map 3D image format, it may be difficult to derive a direct relationship between objective measures and subjective quality ratings. For instance, the objective quality of the depth map may have a very weak correlation on its own with the overall subjective quality, because the depth map is used for projecting the corresponding colour image into 3D coordinates and it is not directly viewed by the end users. Individual quality ratings of left and right views may not always account for depth reproduction of the scene. Therefore, the next phase of 3D objective quality metrics includes a methodology to quantify the effect of binocular disparity of 3D scenes in addition to a conventional image/video quality assessment methodology. For instance in [122], in addition to image quality artifacts, disparity distortion measures were also incorporated to evaluate the overall 3D video quality. The article showed improved performance over the method which does not account for the correspondence information of stereoscopic views. The latest 3D image/video quality metrics evaluate depth reproduction in addition to usual image artifacts (such as blockiness) using specific image features (e.g., edge, disparity and structural information of stereoscopic images) which are important for the HVS in both 2D and 3D viewing. For instance the method proposed in [124] shows high correlation values with subjective quality results (Mean Opinion Score, MOS): the correlation coefficient with subjective quality ratings is as high as 0.95; this outperforms the method based on 2D image quality + disparity [122] and other conventional 2D quality metrics separately applied to left and right views (see Table 6.1). The reported performance figures in Table 6.1 are obtained using the same 3D dataset. These observations confirm that accurate 3D image quality metrics should be designed to also consider binocular disparity distortions. All the methods described above are Full-Reference (FR) methods and need the original 3D image sequence to measure the quality by comparison, hence they are not suitable for the evaluation of the quality "on the fly" in real-time transmission applications such as interactive 3D video streaming. In this case the solution is to use Reduced-Reference (RR) or No-Reference (NR) metrics which do not require the original image for quality assessment, but either no information (NR) or just some side-information about it (RR) requiring few bits for its transmission. Most of the NR metrics are designed specifically for a known set of artifacts (e.g., JPEG compression) and cannot be deployed in a more general scenario. In case of RR metrics, side-information is generated from features extracted from the original 3D image sequence and sent to the receiver-side to measure 3D video quality. Since the reference side-information has to be transmitted over the channel, either in-band or on a dedicated connection, the overhead should be kept at a minimum level. The next section describes how we could measure 3D video quality "on the fly" using RR and NR methods and provides a brief description of the existing methods.

Method	CC	SSE	RMSE
SSIM (Structural SIMilarity)	0.837	0.965	0.159
VQM (Video Quality Metric)	0.932	0.423	0.106
Proposed in [8]: 2D image quality + Disparity	0.901	0.608	0.126
Proposed in [10]	0.947	0.341	0.095

Table 6.1: Correlation between objective 3D image/video measures and subjective quality

6.1 Real-time 3D video quality evaluation strategies

The measured image quality at the receiver-side can be used as feedback information to update system parameters "on the fly" in a "QoE-aware" system design approach [117][125]. However, measuring 3D video quality in real time is a challenge mainly due to the complex nature of 3D video quality and also the fact that the amount of side-information to be sent to measure the quality with RR methods is larger compared to 2D image/video applications. The emerging RR and NR quality evaluation methods are based on image features associated to the characteristics of the HVS. Some of these features are related to image perception (e.g., luminance, contrast) and some are related to depth perception (e.g., disparity, structural correlations). An appropriate selection of these features is crucial to design an effective 3D image/video quality assessment method. The selected features should be able to quantify image and depth perception related artifacts with a minimum overhead. If the overhead is significant, the feasibility of deploying the designed RR method is reduced. Figure 6.2 shows how the extracted edge information is employed to measure 3D video quality in the RR method proposed in [126]. In this method, luminance and contrast details of the original and distorted images are utilized to count for conventional image artifacts, whereas edge information based structural correlation is employed to measure the structural/disparity degradation of the 3D scene, which is directly affecting rendering using colour plus depth map based 3D video. In order to reduce the overhead for side-information (i.e., extracted features of the reference image) lossless compression mechanisms can be deployed for its compression. An extra effort should be also made to send the side-information without corruption using a dedicated channel or highly protected forward channel. Visual attention models could also be utilized to find 3D image/video features which attract significant attention during 3D viewing. However, a direct relationship between visual attention and image perception for 3D images and video is yet to be found. NR methods are the most suitable for real-time 3D video applications since these do not consume any bandwidth for the transmission of side information. However, their performance and application domain is limited since they rely solely on the received 3D image/video sequence and other contextual information (e.g., Hybrid-NR methods: packet loss rate, bit-error rate). It may be impossible to count for all the artifacts imposed along the end to end 3D video chain without referring to the original image sequence. This is why most of the proposed NR metrics are limited to a specific set of artifacts [127].

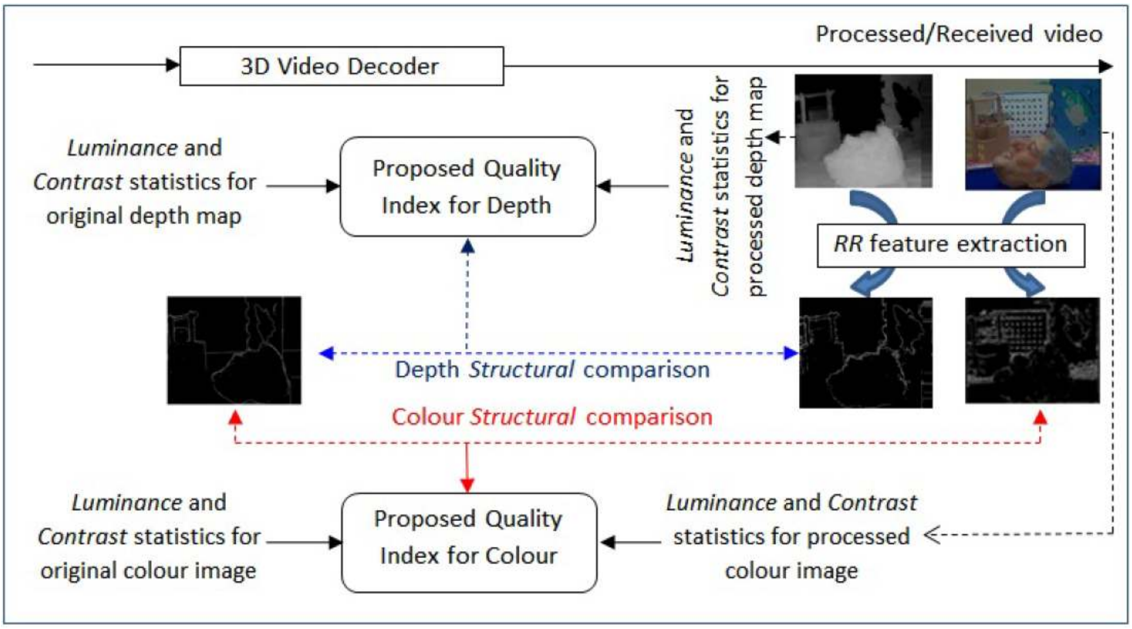


Figure 6.2: Reduced reference edge based 3D video quality metric [12].

Table 6.2 reports a few existing NR and RR quality metrics for 3D image/video. This table explains which image features are used to measure the overall perception and how much the different metrics are correlated with subjective quality scores (i.e., MOS) and with existing Full-Reference methods. It can be observed that most of these methods show a high degree of correlation with subjective MOS and Full-reference methods. However, these metrics are focused on one or two specific 3D perceptual attributes. The combined effect of these perceptual attributes which is directly related to user 3D QoE has not been addressed to date. The methods in [128] and [127] are evaluated using the same image database whereas others are evaluated using different data sets. Since some of these metrics, e.g., NR metrics ([127] and [129]) are designed for a particular types of image artifacts (e.g., JPEG compression), it is not always possible to compare the performance of a NR metric with another objective quality model in a common dataset. On the other hand, due to the overhead associated with RR metrics compared to zero overhead for NR metrics, the usage and advantages of these methods are significantly different. In addition, due to some practical reasons (intellectual property rights, different source 3D video formats, e.g., colour + depth vs. left and right images, unavailability of ground truth depth maps, etc.), it is not always feasible to compare the performance of two different 3D quality evaluation algorithms in a common dataset. The lack of reliable and comprehensive 3D image/video databases is another major challenge faced by researchers and developers, making difficult to effectively compare the performance of emerging objective and subjective quality evaluation methods with that of the existing methods.

Quality Metric	Method (NR or RR)	Artifacts	Features used to measure image artifacts (IA) and disparity (D)	CC	ROCC	OR	RMSE
Cyclop [14]	RR	JPEG symmetric and asymmetric coding artifacts	IA: Contrast sensitivity (spatial frequency and orientation); D: coherence of cyclopean images	0.981	0.950	0.050	-
Sazzad et al. [13]	NR	JPEG symmetric and asymmetric coding artifacts	IA: Blockiness and zero crossing of edge, flat and texture areas; D : average zero crossing of plane and non-plane areas	0.960	0.920	0.069	-
Solh et al. [15]	NR	Depth map and colored video compression, depth estimation (stereo matching), and depth from 2D to 3D conversion	IA D: Temporal outliers (TO), temporal inconsistencies (TI), and spatial outliers (SO) using ideal depth estimate for each pixel	0.916	0.1003	0.8	1.686
Hewage & Martini [12]	RR	H264 compression and random packet losses	IA: Luminance, structure and contrast D : edge based structural correlation	Colour: 0.9273 (vs. FR); Depth: 0.9795 (vs. FR)			Colour: 0.0110 (vs. FR); Depth: 0.0064 (vs. FR)

Table 6.2: No-Reference (NR) and Reduced-Reference (RR) methods for 3D image/video

6.2 Challenges for real-time 3D video quality evaluation

The possibility to measure 3D image/video quality in real time, as requested by 3D video applications, is hindered by several issues. The major challenge is how we could measure the effect of all perceptual attributes (e.g., depth, presence, naturalness, etc.) associated with 3D viewing. The lack of availability of common 3D image/video databases is also detrimental for the advance in this discipline. The following paragraphs briefly discuss these challenges and possible solutions foreseen.

6.2.1 Measurement of different 3D perceptual attributes

Even though emerging 3D quality evaluation methods accurately predict a given quality attribute, the relationship among these perception attributes has not be thoroughly studied. The combined effect directly affects user experience and can be measured using emerging QoE indices. Therefore the current need is to understand how 3D audio/image processing and transmission artifacts affect the overall experience of the user, then identify audio, image and contextual features which can be used to quantify the overall effect on user experience. On the other hand, it is necessary to understand how the HVS perceives these 3D artifacts. For instance, there could be conflicts based on whether binocular suppression or binocular rivalry is taking place based on the artifacts in question. These aspects need extended attention in order to measure the overall experience of 3D viewing. In order to enable a unified approach to 3D objective quality subjective quality evaluation studies, standardization of these procedures are necessary. Several standardization activities are being carried out by VQEG, ITU (Recommendations: ITU-T P- and J-series), European Broadcasting Union EBU (3D-TV Group) and other Standards Developing Organizations (SDOs) in relation to 3D video subjective and objective quality evaluations. Currently, the Video Quality Expert Group (VQEG) is working (3DTV project) on creating a ground truth 3D video dataset (GroTruQoE dataset) using the pair-comparison method. This ground truth database will then be used to evaluate other time-efficient 3D subjective quality evaluation methodologies and objective quality models. In addition, the project also addresses the objective quality assessment of 3D video, with the plan to evaluate 3D quality of experience in relation to the visual quality, depth quality and visual comfort dimensions. Most of these findings are reported to objective and subjective 3D video quality studies in ITU-T Study Groups (SG) 9 and 12. EBU is also working on 3D video production, formats and sequence properties for 3DTV Broadcasting applications (e.g., EBU Recommendation R 135).

B. Lack of 3D image/video databases

There are several image/video quality databases for conventional 2D image/video artifacts, although only a few have been reported for 3D image/video artifacts. This prevents developers from using a common dataset to evaluate the performance of their metrics. Table 6.3 shows some of the reported 3D image/video databases in the literature. The amount of artifacts considered in these databases is limited. Most of them do not consider artifacts which could be introduced during transmission. Therefore it is a responsibility of the research community to produce comprehensive 3D video datasets covering a range of image and transmission artifacts and make available the developed 3D image/video dataset publicly.

3D image/video database	Creator	Artifacts
Mobile 3D video database	University of Tampere and Nokia	Crosstalk, blocking, colour mismatch and bleeding, packet losses for low-resolution video (only impaired sequences, no MOS values provided).
IRCCyN 3D image database	University of Nantes	JPEG, J2K, upsample/downsample, etc.
EPFL databases for images/videos	EPFL	Different camera distances
Kingston University video database	Kingston University-London	Packet losses
NAMA3DS1-COSPAD1	University of Nantes	H.264 and JPEG2000 compression artifacts
RMIT3DV	RMIT University	Uncompressed HD 3D video

Table 6.3: Available 3D image/video database

C. Visual attention models to develop RR and NR quality metrics

The attention of users during 3D viewing can be influenced by several factors including spatial/temporal frequencies, depth cues, conflicting depth cues, etc. The studies on visual attention in 2D/3D images found out that the behaviour of viewers during 2D viewing and 3D viewing is not always identical (e.g., centre bias vs. depth bias). These observations are tightly linked with the way we perceive 3D video. Therefore, effective 3D video quality evaluation and 3D QoE enhancement schemes could be designed based on these observations. There are still unanswered questions such as whether quality assessment is analogous to attentional quality assessment and also how attention mechanisms could be properly integrated into design of QoE assessment methodologies. A thorough study has not been conducted to date in order to identify the relationship between 3D image/video attention models and 3D image/video quality evaluation. Similar to the integrated model described above, attentive areas identified by visual attention studies can be utilized to extract image features which can be used to design No-Reference (NR) and Reduced-Reference (RR) quality metrics for real-time 3D video application. Furthermore, since visual attention models can predict the highly attentive areas of an image or video, these can be integrated into source and channel coding at the sender side. Emerging 3D saliency models incorporate 2D image, depth and motion information which can be applied to 3D video sequences. Most of the reported 3D saliency models are extensions of 2D visual saliency models by incorporating depth information. Table IV summarises a few 3D saliency models reported in the literature. There are two main types of depth integrated saliency models, namely: Depth weighted 3D saliency model and Depth saliency model based methods. The depth weighted saliency models weight the 2D saliency map based on depth information. In depth saliency models, the predicted 3D saliency map is derived based on the chosen weights for 2D and depth saliency maps.