Jin *et al. EURASIP Journal on Image and Video Processing* 2014, **2014**:6
http://jivp.eurasipjournals.com/content/2014/1/6

EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3D video

Lina Jin[*], Atanas Boev, Karen Egiazarian and Atanas Gotchev

Abstract

3D video is expected to provide an enhanced user experience by using the impression of depth to bring greater realism to the user. Quality assessment plays an important role in the design and optimization of 3D video processing systems. In this paper, a new 3D image quality model that is specifically tailored for mobile 3D video is proposed. The model adopts three quality components, called the cyclopean view, binocular rivalry, and the scene geometry, in which the quality must be quantified. The cyclopean view formation process is simulated and its quality is evaluated using the three proposed approaches. Binocular rivalry is quantified over the distorted stereo pairs, and the scene quality is quantified over the disparity map. Based on the model, the 3D image quality can then be assessed using state-of-the-art 2D quality measures selected appropriately through a machine learning approach. To make the metric simple, fast, and efficient, final selection of the quality features is accomplished by also considering the computational complexity and the CPU running time. The metric is compared with several currently available 2D and 3D metrics. Experimental results show that the compound metric gives a significantly high correlation with the mean opinion scores that were collected through large-scale subjective tests run on mobile 3D video content.

Keywords: 3D quality assessment; Cyclopean view; Binocular rivalry; Linear regression

1 Introduction

Recently, with the rapid advances being made in 3D video technologies, mobile 3D video has become a subject of interest for both the entertainment and consumer electronics industries. Mobile 3D video offers a number of challenges, because it is expected to deliver a high-quality experience to the mobile users while using limited resources, including lower bandwidths and error-prone wireless channels. One of the greatest challenges is the evaluation of 3D video quality in a perceptual manner. Normally, a 3D video system includes several signal processing stages, e.g., scene capture and content creation, video format conversion, encoding, transmission, possible post-processing at the receiver side, and rendering and display of the image. Each stage may contribute to the degradation of the 3D visual quality, and the errors that occur at certain steps may propagate through the chain. Therefore, quality assessment (QA) plays an important role in the design and optimization of the system in relation to the prospective users, systems, and services.

QA of any multimedia content is best performed subjectively, i.e., by asking test participants to give their opinions on different aspects of the quality of the content that they experienced. While it is highly informative in that it directly reflects human perception, subjective evaluation has many limitations. It is a time-consuming and expensive process and is not suitable for real-time quality monitoring and adjustment of the systems. Therefore, research on objective QA usually follows the subjective studies to design algorithms that can automatically assess multimedia quality in a perceptually consistent manner. Consider, for example, a wireless multimedia network system: a server can be dedicated to the evaluation of the delivered content quality using objective QA measures, and the results can be used to control and allocate the streaming resources. At the encoding and decoding stages, objective QA can also be used to optimize the

[*] Correspondence: lina.jin@tut.fi
Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 10, Tampere 33720, Finland

encoding and rendering algorithms. Objective QA of conventional (i.e., 2D) images and video have been an active research topic for several decades, but the research work on QA for 3D images and video is relatively young and less mature.

A 3D video can be defined as time-varying imagery that supports the binocular visual cue, which, in combination with other 3D visual cues, delivers a realistic perception of depth. In its simplest form, 3D video is formed using two separate video channels (i.e., left and right) in which the time-synchronized frames form stereo pairs. Early attempts to objectively quantify 3D video images have applied 2D metrics to each frame of the stereo pair. Each frame is viewed as a single image for which the quality is measured separately, and then the overall 3D quality is calculated by averaging over time and space (i.e., the mean of the left and right channel quality values). This approach, however, hardly corresponds to the actual binocular mechanisms of the human visual system (HVS) and, thus, hardly correlates with the subjective quality scores. Recently, the inclusion of some 3D factors as part of the quality evaluation process has been attempted [1]. In [2], a 3D discrete cosine transform (DCT)-based stereo QA method was proposed for mobile 3D video. The method attempts to model the mechanisms of binocular correspondence formation, using the information in the neighboring blocks and contrast masking by grouping similarly sized 4×4 blocks of pixels in the left and right channels for joint analysis in the 3D DCT domain. In [3], the local depth variance for each reference block is used to weigh the quality metric proposed in [2] appropriately. In [4], a monoscopic quality component and a stereoscopic quality component for measurement of stereoscopic image quality have been combined. The former component assesses the monoscopically perceived distortions caused by phenomena such as blurring, noise, and contrast change, while the latter assesses the perceived degradation of the binocular depth cues only. In [5], an overall stereo quality metric was proposed through the combination of image quality with disparity quality using a nonlinear function. In [6], the 3D video quality was analyzed on the basis of being composed of two parts: the stereoscopic 2D video quality and the depth map quality. In [7], a quality metric for color stereo images was proposed based on the use of the binocular energy contained in the left and right retinal images, which was calculated using the complex wavelet transform (CWT) and the bandelet transform. The authors of [8] proposed two approaches based on depth of image-based rendering to compare synthesized views and occlusions. Authors in [9] proposed an objective model for evaluation of the depth quality using subjective results. In [10], the performances of several state-of-the-art 2D quality metrics were compared for quantification of the quality of stereo pairs formed from two synthesized views. In [11] the authors studied the perception of stereoscopic crosstalk and performed a set of subjective tests to obtain mean opinion scores (MOS) of stereoscopic videos. They attempted to predict the MOS by combination of a structural similarity index (SSIM) map and pre-filtered dense disparity map. The quality metric proposed in [12] attempts to predict the perceived quality of color stereo video by a combination of contrast sensitivity function (CSF) filters with rational thresholds.

In [1], an analysis of the factors that influence the 3D quality of experience has been conducted. According to that analysis, the following HVS properties should be taken into account in the design of 3D quality metrics [13]. First, the HVS perceives '2D' types of degradation after they are combined in the cyclopean view and not individually in the left and right channels. Therefore, it is meaningful to measure 2D artifacts on the cyclopean view. The forms of degradation related to the 3D geometry and perceived through disparity are characterized as '3D' artifacts. Thus, the cyclopean image of both the degraded and the reference video streams should be extracted and compared, along with the binocular disparity that is presented in the degraded stream. Second, while the 2D and 3D artifacts can be assessed separately, the content in one visual path may influence the other. The binocular perception of depth is influenced by pictorial depth cues. It is possible that there may be masking or facilitation between the depth cues that come from the two visual paths. Consequently, the 3D quality is influenced by the 2D content. The perception of the asymmetric quality depends on the scene depth. Artifacts in the cyclopean view may be masked by the convergence process. Consequently, the 2D quality is then influenced by the 3D content. The overall quality of a 3D scene is therefore a combination of the 'cyclopean' and 'binocular' perceptual qualities.

Based on the above analysis, a new model for the assessment of 3D image quality is investigated in this paper. The model considers three components: the cyclopean view, binocular rivalry, and the depth presence. This general model aims to reflect the peculiarities of 3D scene perception. These peculiarities include the fusion of the left and right (stereo) images into a single (cyclopean) image and its 2D quality, the possible influence of binocular rivalry on visual comfort, and the influence of the depth presence on correct perception of the 3D scene geometry. The investigation aims to find suitable features to quantify the qualities of these three components in a 3D image to enable their combination, leading to an objective metric that is in accordance with the objective opinion. An abundant set of features that are used in the state-of-the-art 2D QA metrics is adopted, and a machine learning approach is applied to find the best combinations of these features. With regard to the formation of the cyclopean view, three different quality models are investigated that depend on whether the image fusion process is simulated at pixel level or at block level. The binocular disparity, i.e., the differences between the images seen in each eye, is an important cue that the HVS uses to perceive 3D scenes. However, artifacts in a stereoscopic pair may introduce unnatural stereoscopic correspondences that cannot be interpreted by the binocular HVS. These effects are perceived as a binocular rivalry, and this binocular rivalry must be quantified. The binocular suppression theory states that masking and facilitation effects exist between the images that are perceived by each eye [14]. It is anticipated that the masking between the eyes works in a similar manner to the masking effects between the different spatial orientations. In this paper, a local method for binocular rivalry evaluation is proposed that quantifies the quality of the binocular rivalry between the viewed left channel and right channel. Also, the depth presence is quantified using the disparity map, which gives the apparent motion between corresponding pixels in the left and right images.

To fuse the three proposed components in a perceptually driven manner, two mobile 3D video databases and related subjective tests are used [15,16]. Earlier subjective studies aimed to set more precise limits for acceptance of the quality experienced when both the compression artifacts using different 3D video coding methods and varying amounts of depth are presented. They have also taken a more systematic approach to the examination of depth versus compression artifacts by varying a dense set of parameters that influence quality. In the first mobile 3D video database, the number of compression artifacts has been varied by selecting five quantization parameters (QPs) and the strength of the depth effect was varied by selecting two camera baseline ranges. The video sequences in the second 3D video database have been encoded using four different coding methods, including H.264/AVC Simulcast, H.264/AVC multiview video coding (MVC), mixed resolution stereo coding (MRSC), and video plus depth (V + D). The encoding parameters have been chosen in accordance with the settings of the prospective system for mobile 3D video delivery [15] to evaluate the perceived quality provided by each type of content. The combinations of the quality features according to our model, leading to the quality metric, are tested on both databases. The results show that this metric outperforms the current popular metrics over different 3D video formats and compression methods.

2 Image processing channel in stereo vision

A simplified model of the stereoscopic HVS is presented in Figure 1. The model follows the main functional stages of binocular vision, as discussed in [1]. In the first stage, the light captured by the eyes is processed separately in each eye. A set of perceptual HVS properties are produced by this processing, including light adaptation, contrast sensitivity, and low chromatic resolution. These properties can be modeled by luminance masking, conversion to a perceptual color space, and CSF-based masking, as shown in Figure 2. In the next stage, the visual information passes through the lateral geniculate nucleus (LGN), where the inputs from both eyes are processed together. It is assumed that the LGN decorrelates the stereoscopic signal and then forms the so-called cyclopean view [17]. The visual information is then fed to the V1 brain center, which is sensitive to patches with different spatial frequencies and orientations. The processes in the LGN and the V1 center can be modeled as multichannel decomposition, followed by binocular, spatial and temporal masking, as shown in Figure 2 [17].

The perceptual properties of the binocular vision suggest that the visual information is simultaneously processed in two different pathways, as shown in Figure 3. One pathway performs a fusion process using the binocular information to form a cyclopean view, which is a 2D representation of the scene as if it was observed from a virtual point that appears between the eyes [1]. During fusion, the HVS attempts to reconstruct details that are available to one eye only, which allows the observer to reconstruct any partially occluded details of the scene. The other pathway compares the images that have been projected onto each retina and extracts the distance information (also known as binocular depth cues [17]). Larger

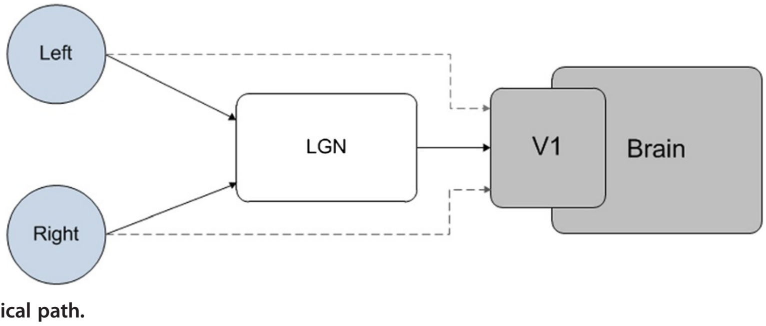

Figure 1 Model of the optical path.

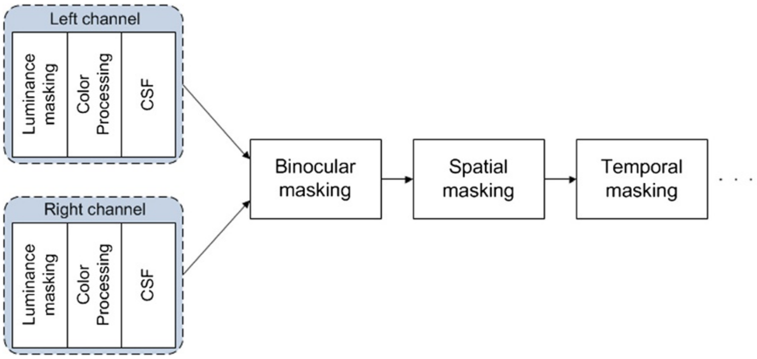

Figure 2 Image processing channel for realization of the model.

differences between the retinal images result in a more pronounced binocular depth. However, if these differences are too large, the images from the two eyes cannot be fused, and instead of the cyclopean view, the HVS perceives binocular rivalry [18]. Binocular rivalry is one of the major sources of visual discomfort in 3D video. This phenomenon can be caused by several factors, including physical misalignments, luminance, color, reflection, hyperconvergence, hyperdivergence, and ghosting [19].

Based on this model, we assume that the quality of a 3D image is perceived as a combination of two components: the quality of the cyclopean view, and the quality of the binocular image. The subjective experiments in [15] show that the presence of depth influences the perceived quality, and this influence can be either positive or negative, depending on the content. As described in [1], the same amount of blockiness is graded differently in scenes with differently pronounced depths. The presence of stereoscopic depth also affects the perceived overall quality. Larger binocular differences will increase the perceived binocular depth but may also reduce the quality of the cyclopean view. This effect is not monotonic, which indicates that there might be an ‘optimal’ global depth for a 3D scene on portable autostereoscopic displays, at which the HVS has the lowest sensitivity to any cyclopean image degradation.

3 Feature-based quality estimation

In this section, we propose a new 3D QA model composed of three components: the quality of the cyclopean view, the prominence of the binocular depth, and the presence of binocular rivalry. The block diagram of our model is shown in Figure 4. We select a set of features that (potentially) quantify each quality component. Combinations of these features are then matched against the MOS that were obtained from subjective quality tests.

3.1 Cyclopean view assessment

The quality of the cyclopean view can be measured in a full-reference setting. The first step is to create the cyclopean views of the reference and the distorted stereo pairs. When both cyclopean views are available, we can compare the structural differences between the two cyclopean views using an ordinary full-reference 2D quality metric.

One way to create the cyclopean view is to generate a dense disparity map of the stereo pair and reconstruct the view from an intermediate observation position. In case the corresponding pixels in the two observations have different colors or intensities, the mean values of both properties are taken. This roughly corresponds to the way that the cyclopean view is fused by the HVS in the absence of stereoscopic rivalry [1]. However,

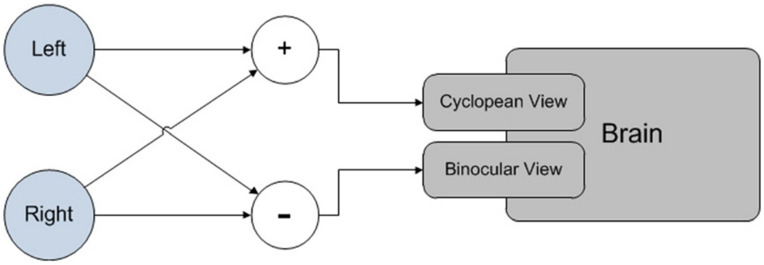

Figure 3 Model of the binocular fusion and depth extraction process.

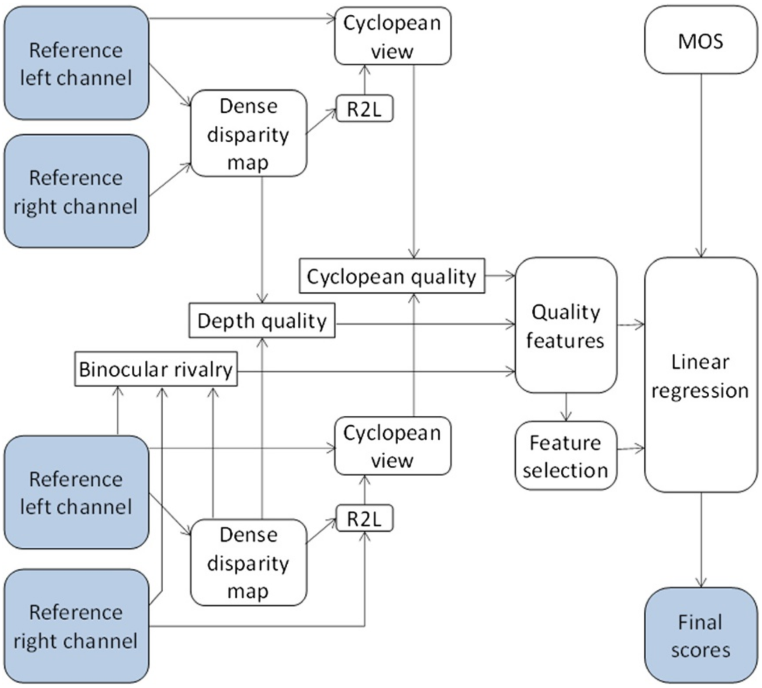

Figure 4 Block diagram of proposed 3D metric.

rendering of the intermediate camera involves the interpolation of pixels from both views. To reduce the influence of any interpolation errors, we can fuse the two views and reconstruct an observation that matches one of the views. This can be done by warping one of the views onto the other - for example, by rendering the right view using the left view pixels and the disparity map - and then fusing the two views. Because we aim to assess the structural differences between the two cyclopean views, we assume that this transformation would still allow any distortions in either view to be quantified. Wherever occlusions occur, the available pixels from the opposite view are used. In our approach, we calculate a dense disparity map and an occlusion map between the left and right images using a color-weighted local search method [20].

Using the disparity map, the pixels in the right channel are then mapped to their positions in the left channel, which is denoted here as a ‘right to left’ mapping, i.e., R2L:

$$\mathrm{R2L}(x,y) = I^{\mathrm{R}}(x+\Delta(x,y),y), x = 1\ldots N, y = 1\ldots M, \qquad (1)$$

where (x, y) indicates the pixel location, M, N indicates the image size of one channel, I^{R} is the image from the right channel, and $\Delta(x, y)$ is the pixel shift for the pixel at position (x, y). Occluded pixels are handled by replacing them with corresponding pixels from the left image:

$$R\tilde{2}L(x,y) = \begin{cases} I^{\mathrm{L}}(x,y), & \text{if} \quad \Omega(x,y) = 1 \\ \mathrm{R2L}(x,y), & \text{otherwise}, \end{cases} \qquad (2)$$

where I^{L} is the left image, Ω is the binary occlusion map, and $\Omega(x, y) = 1$ marks the occluded pixels. The final cyclopean view, I^{cyc}, is generated as the mean of the left image and the mapped image from the right image:

$$I^{\mathrm{cyc}} = \frac{I^{\mathrm{L}} + \mathrm{R\tilde{2}L}}{2}. \qquad (3)$$

The cyclopean view formation process is shown in Figure 5, and an example of the cyclopean view obtained is given in Figure 6.

When the cyclopean view is obtained, we then apply three quality evaluation models. Hereafter, we use the notation *QA* to denote *any* quality assessment measure, which compare the similarity (dissimilarity) between images or image blocks. Specific QAs are discussed in Section 3.4 where they are indexed (e.g., QA_1, QA_2...) to denote the particular assessment measure.

The first model assumes QA on a global basis:

$$Q_1^{\mathrm{CV}} = \mathrm{QA}\left(I_{\mathrm{ref}}^{\mathrm{cyc}}, I_{\mathrm{dis}}^{\mathrm{cyc}}\right), \qquad (4)$$

where $I_{\mathrm{ref}}^{\mathrm{cyc}}$ and $I_{\mathrm{dis}}^{\mathrm{cyc}}$ are the cyclopean images that were obtained from the reference and distorted stereo pairs, respectively.

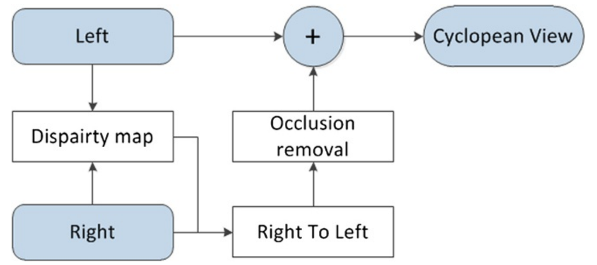

Figure 5 Flow chart for generation of the cyclopean view.

The second model evaluates the cyclopean view in a block-by-block fashion, as shown in Figure 7. In the left channel, an 8 × 8-sized reference block A starting at coordinates (i,j) is selected. The corresponding block in the disparity map is denoted by Δ_{ij}. In the right channel, the block with the same coordinates (i,j) is marked B'. Using the disparity map, the corresponding block B is then found by taking the median of the disparity values in the disparity patch Δ_{ij}:

$$\hat{d} = \left\lceil \mathrm{median}\{\Delta_{ij}\}_{8\ \times\ 8} \right\rceil, \tag{5}$$

where Δ_{ij} is the disparity mapping with coordinates (i,j), $\{\cdot\}_{8\ \times\ 8}$ indicates an 8 × 8 block, and $\hat{d}$ can be a positive value, zero, or a negative value. The model assumes that the quality of the block is represented by the quality of the better channel of the two,

$$Q_2^{\mathrm{CV}} = \frac{\sum_{i=1}^{N_{\mathrm{blk}}} \max\left(q_i^{\mathrm{L}}, q_i^{\mathrm{R}}\right)}{N_{\mathrm{blk}}}, \tag{6}$$

where N_{blk} is the number of blocks, and q_i^{L} and q_i^{R} are the quality scores of the left and right channels, respectively,

$$q_i^{\mathrm{L}} = \mathrm{QA}(A_{\mathrm{ref}}, A_{\mathrm{dis}}) \tag{7}$$

$$q_i^{\mathrm{R}} = \mathrm{QA}(B_{\mathrm{ref}}, B_{\mathrm{dis}}), \tag{8}$$

where A_{ref} is the reference block in the original left image $I_{\mathrm{ref}}^{\mathrm{L}}$, I_{dis} is the corresponding block in the distorted left image $I_{\mathrm{dis}}^{\mathrm{L}}$, and B_{ref} and B_{dis} are the corresponding blocks in $I_{\mathrm{ref}}^{\mathrm{R}}$ and $I_{\mathrm{dis}}^{\mathrm{R}}$, respectively.

The third model closely follows the second model but assumes that the block quality is represented by the average of the quality values measured in the left and right channels:

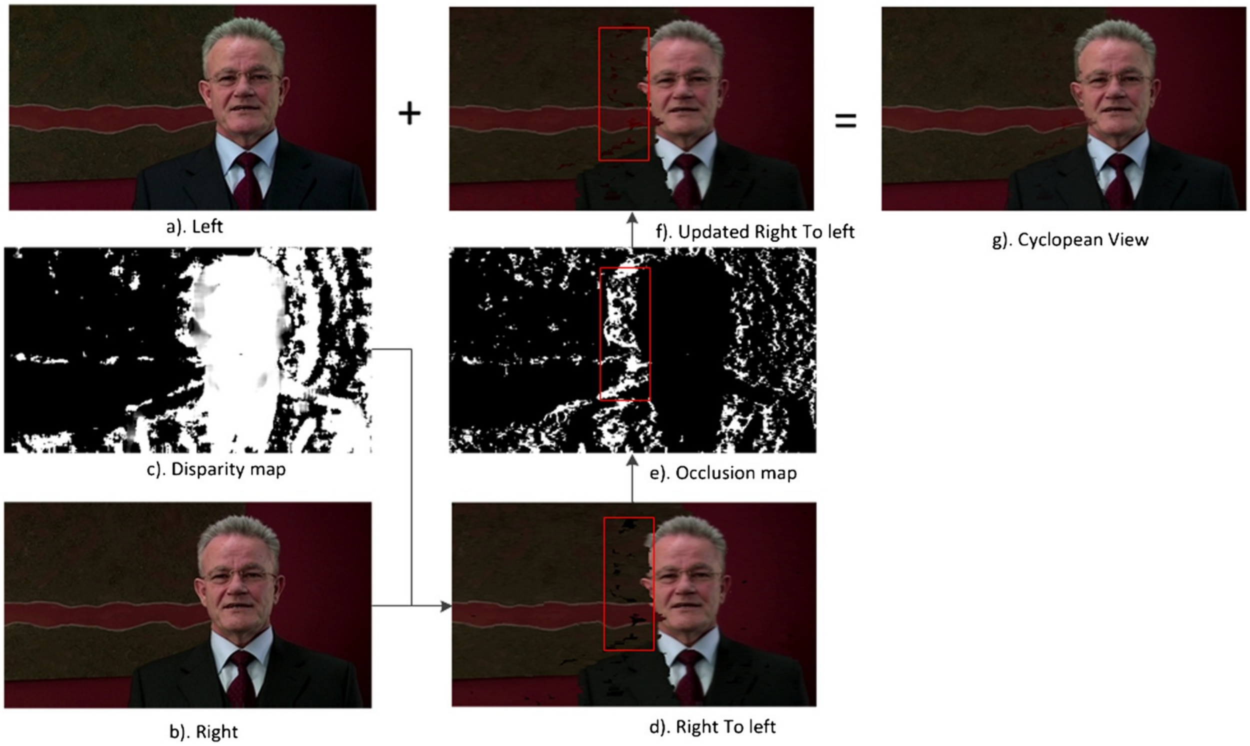

Figure 6 An illustration example of forming a cyclopean view. (a) Left. **(b)** Right. **(c)** Disparity map. **(d)** Right to left. **(e)** Occlusion map. **(f)** Updated right to left. **(g)** Cyclopean view.

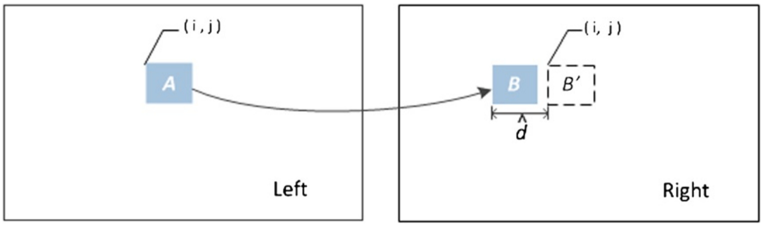

Figure 7 **Location of similar blocks for the second binocular rivalry quality model.**

$$Q_3^{\mathrm{CV}} = \frac{\sum_{i=1}^{N_{\mathrm{blk}}} \left(q_i^{\mathrm{L}} + q_i^{\mathrm{R}}\right)/2}{N_{\mathrm{blk}}}. \tag{9}$$

3.2 Binocular rivalry assessment

Binocular rivalry occurs when the eyes attempt to converge on a single point in a scene, but the images seen by the two eyes are not sufficiently similar. Binocular rivalry can occur naturally in a complex 3D scene with numerous occlusions. However, the presence of severe artifacts in only one of the channels can cause unnatural binocular rivalry, which is perceived as a severe stereoscopic artifact. Binocular rivalry can be measured in a non-reference setting, i.e., by analyzing the distorted pair only. We assume that regardless of whether or not the rivalry is present in the original pair, its presence in the distorted pair would be equally disturbing. In our approach, we use the dense depth map to find the corresponding blocks in the two channels and measure the cumulative difference between the corresponding blocks, as follows:

$$Q^{BR} = \frac{\sum_{i=1}^{N_{\mathrm{blk}}} QA(A_{\mathrm{dis}}, B_{\mathrm{dis}})}{N_{\mathrm{blk}}}. \tag{10}$$

3.3 Binocular depth assessment

In this paper, we evaluate the presence of the binocular depth by estimation of the dense depth map for the stereo pair. We calculate a dense disparity map using the color-weighted local-window method described in [20]. The quality of depth Q^{DQ} is then studied as follows:

$$Q_1^{\mathrm{DQ}} = \mathrm{QA}(\Delta_{\mathrm{ref}}, \Delta_{\mathrm{dis}}), \tag{11}$$

where Δ_{ref} is the disparity map from the original stereoscopic image, and Δ_{dis} is the disparity map from the distorted stereoscopic image. Here, QA denotes a QA function that uses one of the candidate features, as described in the Section 3.4.

3.4 Candidate features

Each of the three quality components described above relies on a comparison function denoted by QA. However, the data that are compared are not in the same modality in each case; in one case, we measure the similarity between the images, while in another we compare disparity maps. These cases are interpreted in different ways by the HVS, and the optimum similarity measure would be different for each case. To determine the most suitable measure in each case, we have selected and tested ten state-of-the-art QA methods.

We denote the original input image (block) by u and the distorted image (block) by v. The first quality feature is calculated based on the mean squared error (MSE), which is the most popular difference metric used in image and video processing:

$$\mathrm{QA}_1(u, v) = \frac{1}{MN} \sum_i \sum_j \left(u_{ij} - v_{ij}\right)^2. \tag{12}$$

The MSE is chosen because it is simple to calculate, has clear physical meaning, and is mathematically convenient in the context of optimization.

The second quality feature is the gradient-normalized sum-of-squared-difference (SSD) [21]. The result is normalized with reference to the gradient map and is calculated as the mean of the SSD. Any local intensity variations in the textured areas between u and v are thus penalized:

$$\mathrm{QA}_2(u, v) = \frac{1}{MN} \sum_i \sum_j \frac{\left(u_{ij} - v_{ij}\right)^2}{\left\|\nabla u_{ij}\right\|^2 + 1} \tag{13}$$

where ∇u_{ij} is the gradient value of input signal u.

Many studies have confirmed that the HVS is more sensitive to low-frequency distortions rather than to those at high frequencies. It is also very sensitive to contrast changes and noise. Therefore, the third measure aims to remove the mean shifting and contrast stretching in the manner shown in [22]. The measure is calculated in 8 × 8 blocks and uses the decorrelation properties of the block DCT and the effects of the individual DCT coefficients on the overall perception:

$$\mathrm{QA}_3 = \frac{1}{N_{\mathrm{blk}}} \sum_{i=1}^{M-7} \sum_{j=1}^{N-7} E_{\mathrm{w}}(u - v) \tag{14}$$

$$E_{\mathrm{w}}(u) = \frac{1}{64}\sum\nolimits_{i=1}^{8}\sum\nolimits_{j=1}^{8}\mathrm{DCT}(u)_{ij}^{2}Tc_{ij}, \tag{15}$$

where Tc is the matrix of correction factors for each of the 8 × 8 DCT coefficients, which was normalized based on the JPEG quantization table in [22].

The fourth quality measure is inspired by [23], which was designed based on [22] by taking the CSF and the between-coefficient contrast masking of the DCT basis functions into account. In the same manner shown in [22], the measure operates with the values of the DCT coefficients of the 8 × 8 pixel block. The model allows each DCT coefficient to calculate its own maximum distortion value that is not visible because of the between-coefficient masking. It is assumed that the masking degree of each coefficient $\mathrm{DCT}(u)_{ij}$ depends upon its square value (power) and on the human eye sensitivity to this DCT basis function as determined using the CSF. Several basis functions can jointly mask one or several other basis functions. Then their masking effect value depends upon the sum of their weighted powers [23]. The final formula is expressed as follows:

$$\mathrm{QA}_4 = \frac{1}{N_{\mathrm{blk}}}\sum_{i=1}^{M-7}\sum_{j=1}^{N-7}E_{\mathrm{w}}(u-v)\cdot\mathrm{MaskEff}, \tag{16}$$

where MaskEff is the reduction of the masking and contrast operation given in [23].

The fifth measure is based on the feature similarity index (FSIM) method proposed in [24]. FSIM was designed to compare the low-level feature sets of the reference image and the distorted image. Phase congruency (PC) and the gradient magnitude (GM) are used in FSIM and play complementary roles in the characterization of the local image quality. The measure is defined as

$$\mathrm{QA}_5 = \frac{\sum_i\sum_j S_{\mathrm{L}}\left(u_{ij}, v_{ij}\right)\mathrm{PC}_m\left(u_{ij}, v_{ij}\right)}{\sum_i\sum_j \mathrm{PC}_m\left(u_{ij}, v_{ij}\right)} \tag{17}$$

$$\mathrm{PC}_m(x, y) = \mathrm{MAX}\{\mathrm{PC}(x), \mathrm{PC}(y)\}, \tag{18}$$

where PC is the phase congruency operation of [25], and $S_{\mathrm{L}}(u, v)$ is the similarity map formed by combination of the similarities of PC and GM as $S_{\mathrm{L}} = S_{\mathrm{PC}} \times S_{\mathrm{GM}}$. S_{PC} and S_{GM} are calculated as

$$S_{\mathrm{PC}}(u, v) = \frac{2\mathrm{PC}(u)\cdot\mathrm{PC}(v) + T_1}{\mathrm{PC}^2(u) + \mathrm{PC}^2(v) + T_1} \tag{19}$$

$$S_{\mathrm{GM}}(u, v) = \frac{2\mathrm{GM}(u)\cdot\mathrm{GM}(v) + T_2}{\mathrm{GM}^2(u) + \mathrm{GM}^2(v) + T_2}, \tag{20}$$

where T_1 and T_2 are positive constants. In our work, in addition to the compound measure QA_5, we also consider the individual components, i.e., the PC and the GM, separately, and thus form the sixth and seventh measures, respectively:

$$\mathrm{QA}_6 = \frac{\sum_i\sum_j S_{\mathrm{PC}}\left(u_{ij}, v_{ij}\right)\cdot\mathrm{PC}_m\left(u_{ij}, v_{ij}\right)}{\sum_i\sum_j \mathrm{PC}_m\left(u_{ij}, v_{ij}\right)} \tag{21}$$

$$\mathrm{QA}_7 = S_{\mathrm{GM}}(u, v). \tag{22}$$

The SSIM metric proposed in [26] is considered in the formation of the eighth candidate quality measure. The measure is composed using the luminance comparison $l(u,v)$, the contrast comparison $c(u, v)$ and the structure comparison $s(u,v)$, as follows:

$$\mathrm{QA}_8 = l(u, v)\cdot c(u, v)\cdot s(u, v), \tag{23}$$

$$l(u, v) = \frac{2\mu_u\mu_v + c_1}{\mu_u^2 + \mu_v^2 + c_1}, \tag{24}$$

$$c(u, v) = \frac{2\,\mathrm{cov}_{uv} + c_2}{\sigma_u^2 + \sigma_v^2 + c_2}, \tag{25}$$

$$s(u, v) = \frac{\sigma_{uv} + C_3}{\sigma_u\sigma_v + C_3}, \tag{26}$$

where μ_u and μ_v are the means of u and v, respectively, σ_u^2 and σ_v^2 are the variances of u and v, respectively, cov_{uv} is the covariance of v, c_1 and c_2 are the two variables used to stabilize the division with a weak denominator, and $c_3 = c_2 / 2$. In this paper, QA_9 is defined as the luminance comparison and $\mathrm{QA}_{10} = \frac{2\sigma_{uv}+c_2}{\sigma_u^2+\sigma_v^2+c_2}$, which is a simplified formula for $c(u, v) \times s(u, v)$, as shown in [26].

3.5 Machine learning methods for feature fusion

As described in the previous sections, the proposed quality approach aims to combine three different measures, by separately measuring the quality of the cyclopean view, the binocular rivalry, and the presence of depth. The limited knowledge of the subjective quality perception of 3D images means that it is not possible to predict which of the QA models will produce the best correlation with the subjective scores. Therefore, to find the best combination of quality measures and image features, we adopt a machine learning approach.

We assume that the best combination of features can be found by linear regression. Given a set of quality measures $\varphi_{(k,l)}$, the MOS over a set of test videos Θ_k are predicted using linear combinations where

$$\Theta_k = \hat{\theta}_0 + \sum\nolimits_{l=1}^{\mathrm{L}}\varphi_{(k,l)}\hat{\theta}_l, \tag{27}$$

or

$$\Theta = \begin{bmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_K \end{bmatrix} = \begin{bmatrix} 1, \varphi_{(1,1)}, \varphi_{(1,2)}, \ldots, \varphi_{(1,L)} \\ 1, \varphi_{(2,1)}, \varphi_{(2,2)}, \ldots, \varphi_{(2,L)} \\ \vdots \\ 1, \varphi_{(K,1)}, \varphi_{(K,2)}, \ldots, \varphi_{(K,L)} \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_L \end{bmatrix}, \tag{28}$$

where the vector Θ represents the subjective scores, L is the number of quality measures, K is the number of test stimuli (videos), and $\hat{\theta}_{0,1,2,\ldots,L}$ are the parameters of the model. The above linear model in vector form can also be rewritten as an inner product:

$$\Theta = \varphi^T \hat{\theta}. \tag{29}$$

To fit the linear model to a set of training data, $\hat{\theta}$ is normally determined using the least squares method [27]:

$$f(\theta) = \sum_{i=1}^{K} \left(\overrightarrow{\Theta}_i - \varphi_i^T \theta \right)^2 = \left(\overrightarrow{\Theta} - \varphi^T \theta \right)^T \left(\overrightarrow{\Theta} - \varphi^T \theta \right), \tag{30}$$

where $f(\theta)$ is the cost function, and θ can be chosen to minimize $f(\theta)$ using its derivatives, where

$$\begin{aligned} \nabla_\theta f(\theta) &= \nabla_\theta \left(\overrightarrow{\Theta} - \varphi^T \theta \right)^T \left(\overrightarrow{\Theta} - \varphi^T \theta \right) \\ &= \varphi^T \overrightarrow{\Theta} - \varphi^T \varphi \theta = 0. \end{aligned} \tag{31}$$

Then,

$$\theta = \left(\varphi^T \varphi \right)^{-1} \varphi^T \overrightarrow{\Theta}, \tag{32}$$

where the array of quality measures φ is formed by the proposed 3D quality models, where $\varphi = [Q^{CV}, Q^{BR}, Q^{DQ}]$.

Efficient solution of Equation 28. Using Equation 32 requires a simple, reasonable, and efficient quality measure array and the use of subjective scores from properly conducted subjective experiments. The subjective experiments are described in the Section 4.

4 Mobile 3D video test content and related subjective tests

Two mobile 3D video databases and their corresponding subjective tests have been used for this study. The first database, denoted by '3D database I', contains four stereoscopic videos, called Akko&Kayo, Champagne Tower, Pantomime, and LoveBirds1, with varying levels of compression artifacts and depth presence [16]. Thumbnails of the videos in this database are shown in Figure 8. The database has 60 videos and consists of four scenes; each scene is captured in stereo using three different baselines, and each captured video is compressed by an H.264 encoder using five different values for the QP.

The original videos are high-resolution multiview videos. They have been converted into stereo videos with lower resolution by suitable rescaling. To maintain the variable depth levels, each video sequence has been retargeted by selecting different camera pairs from the available multiview video sequences. For all sequences, the left camera has been retained, while the right camera was selected at two different depth levels called the *short* baseline and the *wide* baseline. In addition, a monoscopic video sequence was generated by repeating the left channel sequence in the place of the right channel sequence. This would effectively present a 2D view with no depth effects on the 3D display. The short baseline produces a 3D scene within a limited disparity range but with visible 3D effects. The wide baseline provides an optimal

Figure 8 Contents of 3D video database I.

disparity range for the mobile stereoscopic display by setting the right camera position. All sequences were then downscaled to lower resolutions for the target display device. After that, each video was encoded using the H.264/AVC Simulcast method in intra-frame mode. The QP was selected in the [25, 30, 35, 40, 45] range and compression was independently applied to the left and right channels.

Thirty-two observers were involved in the subjective tests and were equally distributed in terms of gender with an age range between 18 and 37. The test materials were presented one by one in a pseudo-random order. The display device used was an autostereoscopic screen with a resolution of 428 × 240 pixels per view. After each clip, the test participants were asked to provide overall quality scores on a scale from 0 to 10 and indicate the acceptability of the quality for viewing the mobile 3D video on a yes/no scale. At the beginning of each session, a training set of seven clips was shown. Each test stimulus was shown twice during the test. A set of dummy videos was also shown at the beginning and in the middle of each test session. A total of 164 video clips were shown to each observer [15]. The overall ratings of the stereoscopic videos were finally ranked in terms of their MOS.

The second database contains six different videos spanning different genres of mobile 3DTV and video: these videos are Bullinger, Butterfly, Car, Horse, Mountain, and Soccer2, as shown in Figure 9. This set of videos is intended to represent a range of stereoscopic videos with different content properties, including varying spatial details, temporal changes, and depth complexity. Each video sequence lasts 10 s.

The sequences were encoded using four different methods: H.264/AVC Simulcast, H.264/AVC MVC, MRSC, and V + D. The encoding parameters were chosen as shown in Table 1 [15]. Coding was carried out using two codec profiles: the baseline profile and the high profile. The simple baseline profile uses an IPPP prediction structure and context-adaptive variable-length code (CAVLC) [28] prediction. The group of picture (GOP) size was set at 1. This refers to the low-complexity encoder for mobile devices. The more complex high profile enables hierarchical B-frames with GOP sizes of 8 and context-based adaptive binary arithmetic coding (CABAC) quantization. Because of the variable compressibilities of the different sequences, individual bit rate points were determined for each sequence [15]. The QP of the codec was set at 30 for high quality and 37 for low quality. In total, the database has six reference sequences and 96 distorted 3D video sequences.

Subjective tests were carried out with 87 test participants that were evenly divided in terms of gender and with ages ranging between 16 and 37 years. The visualization process was performed by following the same test procedure and using the same autostereoscopic display as that used in the tests with 3D database I. The MOS for both tests are of the same scale.

5 Feature selection

Both subjective experiments were performed while following the same protocol and using the same device and the same quality evaluation scale. Therefore, we were able to combine the entries from the two databases into a single group of opinion scores within the same scale. We picked 70% of the entries by random selection for forming a training set. The rest of the entries were included in a test set. We measured the prediction performances of the different feature groups using the Spearman rank-order correlation coefficient (SROCC). The SROCC output is in the [−1, 1] range, where a higher absolute value

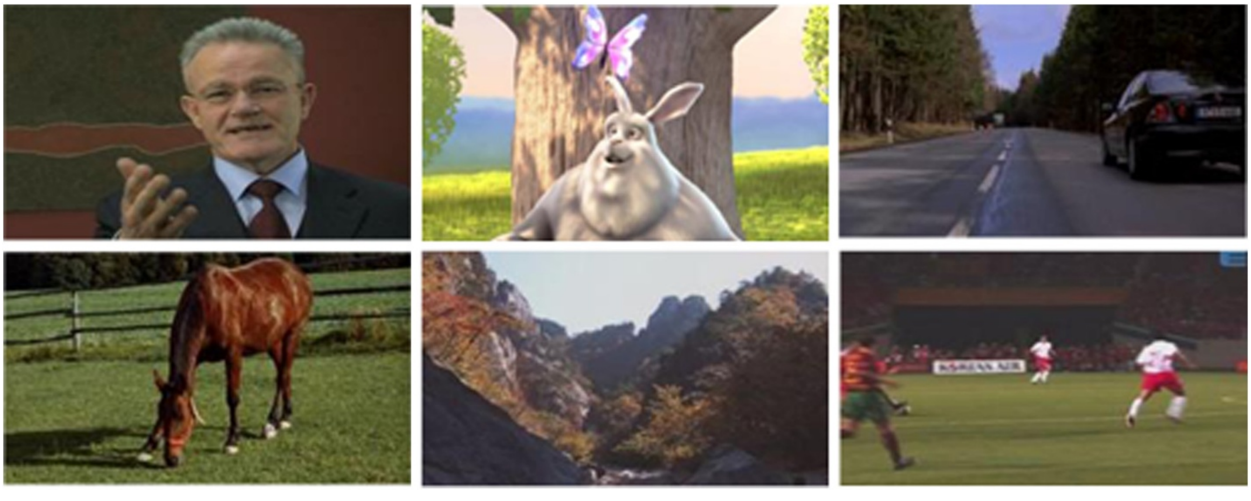

Figure 9 Contents of 3D video database II.

Table 1 Codec settings of the two profiles

Profile	Baseline	High
GOP size	1 (IPPP)	8 (hierarchical B-frames)
Symbol mode	CAVLC	CABAC
Search range	48	48
Intra-period	16	16
Quality level	QP (30, 37)	QP (30, 37)

or SROCC indicates a stronger monotonic relationship between the MOS and the values that were predicted using the metric.

The set of feature candidates consists of 50 items, numbered between $\mathcal{F}_1$ and $\mathcal{F}_{50}$. There are three quality components: the cyclopean view (denoted by Q^{CV}), the binocular rivalry (Q^{BR}), and the depth quality (Q^{DQ}). The quality of the cyclopean view is assessed using three alternative approaches: global comparison $\left(Q_1^{\mathrm{CV}}\right)$, block-wise selection of the better channel $\left(Q_2^{\mathrm{CV}}0029\right.$, and the block-wise average $\left(Q_2^{\mathrm{CV}}\right)$. A set of ten measures was applied to each quality component. The feature candidates are listed in the first row of Table 2. The measures are listed in the first column of the same table. For example, $\mathcal{F}_1$ indicates the quality assessment QA$_1$ under cyclopean view model 1, i.e., $\left\{Q_1^{\mathrm{CV}}, \mathrm{QA}_1\right\}$; $\mathcal{F}_{33}$ indicates the quality assessment QA$_3$ under the binocular rivalry model, i.e., $\{Q^{\mathrm{BR}}, \mathrm{QA}_3\}$. The quality measures that are not relevant to the comparison of the depth maps are excluded from the experiments. These combinations are marked with a dash in Table 2.

We use a regression fitting to measure the performances of the individual features. First, the output of each candidate feature listed in Section 3.4 was normalized to the range [−10, 10], using logistic fitting as follows:

$$f(x) = \beta_1\left(1 + \frac{\beta_2 - \beta_3}{\beta_3 + e^{-\frac{x}{\beta_4}}}\right). \tag{33}$$

The parameters β_1, β_2, β_3, and β_4 have been selected in each individual case so that the output of each feature fits into the desired range. Then we evaluate the performance of each i feature in terms of Spearman correlation. The results of this evaluation are given in Table 2 in columns 2 to 6. The combined performances of all quality measurements applied to a given component are shown on the bottom row of the table, and this measure is denoted by SROCC$_1$. The results in this row indicate the applicability of a single component for use in subjective quality prediction. The combined performance values for the single quality measure when applied to all components are given in the last column of the table, which is labeled SROCC$_2$. These results indicate the applicability of a given quality measure.

From these results, we can see that the use of a single quality component is insufficient because the quality values predicted by a single component do not correlate well with the subjective scores. The best correlation is achieved when using feature $\mathcal{F}_{15}$, i.e., $\left\{Q_2^{\mathrm{CV}}, \mathrm{QA}_5\right\}$. Using the cyclopean view components (e.g., $\left\{Q_{i\in[1,2,3]}^{\mathrm{CR}},\right.$ $\left.\mathrm{QA}_{i\in[1,\ldots,10]}\right\}$), we can achieve SROCC values of more than 0.9. This result can be interpreted as evidence that the 2D quality of the cyclopean view is a major component of the overall perceived quality.

In the next experiment, we attempted to find a combination of features and quality measures that produced a good trade-off between prediction accuracy and

Table 2 Spearman correlations of each quality feature and each quality component

Metrics	$Q_1{}^{\mathrm{CV}}$	$Q_2{}^{\mathrm{CV}}$	$Q_3{}^{\mathrm{C}}$	Q^{BR}	Q^{DQ}	SROCC$_2$
	$\mathcal{F}_1 \sim \mathcal{F}_{10}$	$\mathcal{F}_{11} \sim \mathcal{F}_{20}$	$\mathcal{F}_{21} \sim \mathcal{F}_{30}$	$\mathcal{F}_{31} \sim \mathcal{F}_{40}$	$\mathcal{F}_{41} \sim \mathcal{F}_{50}$	
QA$_1$	0.5591	0.6022	0.4784	−0.0109	0.4660	0.8240
QA$_2$	0.6703	0.7058	0.5849	0.2204	0.5124	0.7830
QA$_3$	0.5816	0.6511	0.5379	0.0345	-	0.6739
QA$_4$	0.5850	0.6839	0.5529	0.0656	-	0.7029
QA$_5$	0.7664	*0.8101*	0.7964	0.1085	0.4769	0.8598
QA$_6$	0.5588	0.5840	0.5240	−0.0675	0.4199	0.7543
QA$_7$	0.7493	0.7347	0.7389	0.3531	0.5870	0.7969
QA$_8$	0.5218	0.5234	0.4452	−0.0119	0.5146	0.7433
QA$_9$	0.4210	0.4680	0.3919	0.1617	0.4881	0.5375
QA$_{10}$	0.5247	0.5192	0.4460	−0.0327	0.3914	0.7407
SROCC$_1$	0.9025	0.9144	0.9164	0.6357	0.6806	0.9711

The highest correlation value is marked in italic.

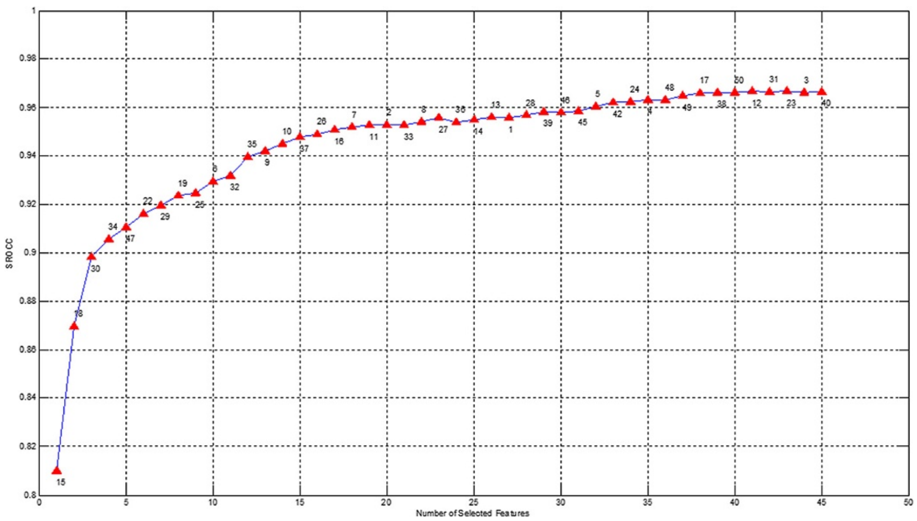

Figure 10 Quality performances with sequential feature selection.

computational complexity. We performed a sequential feature search, looking for the best combination of $n + 1$ features using the best combination of n features and adding one feature at a time. In this manner, we were able to extract 45 features until we reached the SROCC value of 0.97, as shown in Figure 10. By studying the performance improvements introduced by each feature selection (as shown in Figure 11), we see that a combination of four or five features will result in a good accuracy vs. complexity trade-off. The difference in performance for each two consecutive number of features is given in Figure 11, and the difference between the performance for four and five features is marked with a red circle. The first five features in the sequential search are $\{\mathscr{F}_{\{15,18,30,34,47\}}\}$, where $\mathscr{F}_{\{15,18,30\}}$ evaluates the cyclopean view, $\mathscr{F}_{34}$ evaluates the binocular rivalry, and $\mathscr{F}_{47}$ evaluates the depth quality.

The computational complexities of the best performing combinations of four or five features are shown in Table 3 and in Figure 12. The Big O complexity, the McCabe complexity, and the CPU running time for each combination are shown in Table 3. The Big O notation specifically describes the worst-case scenario. The McCabe complexity was proposed in [29] and was also called the cyclomatic complexity or the conditional complexity. McCabe describes the independent paths through the source code as a directed graph. The McCabe complexity is calculated from the cyclomatic number of its graph [29]. The CPU time listed in Table 3 is the time taken to run ten images in each QA$_i$ using MATLAB 2012b on the Win64 OS with the Intel Core Duo E8400 CPU. For comparison, the last row of Table 3 contains the complexity of dense depth estimation and the time it needs to calculate the disparity map of the ten images using search window of 50 pixels on the same computer. Disparity estimation is a step which is required for the calculation of all considered features (see Figure 4), and its computational complexity is in the same range as the complexity of the features. The McCabe complexity and the CPU time of all candidates are compared with the complexity of disparity estimation in Figure 12.

To find an optimal group of features, we estimated the performances of all combinations of five and six features. Since the computational overhead for deriving dense disparity map is the same in each case, we did not take it into account in the feature selection process. We found that 5 groups of five features and 18 groups of six features had SROCC scores that were higher than 0.93. The best performing groups of five features are listed in Table 4, and the best performing groups of six features are listed in Table 5. The complexity levels of each group were calculated based on the McCabe complexities, and

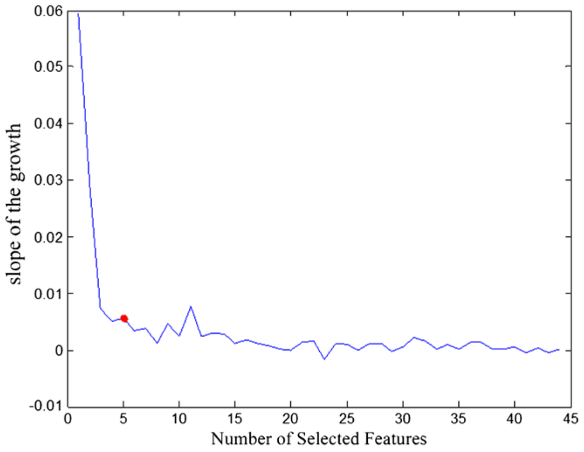

Figure 11 Performance improvements of different numbers of selected features.

Table 3 QA computational complexity

Metrics	Corresponding $\mathcal{F}_i$	Big O	McCabe	Time (s)
QA_1	$\mathcal{F}_1$, $\mathcal{F}_{11}$, $\mathcal{F}_{21}$, $\mathcal{F}_{31}$, $\mathcal{F}_{41}$	$O(N)$	2	0.156
QA_2	$\mathcal{F}_2$, $\mathcal{F}_{12}$, $\mathcal{F}_{22}$, $\mathcal{F}_{32}$, $\mathcal{F}_{42}$	$O(\delta N)$	9	0.328
QA_3	$\mathcal{F}_3$, $\mathcal{F}_{13}$, $\mathcal{F}_{23}$, $\mathcal{F}_{33}$, -	$O(N^2)$	15	7.535
QA_4	$\mathcal{F}_4$, $\mathcal{F}_{14}$, $\mathcal{F}_{24}$, $\mathcal{F}_{34}$, -	$O(N^2)$	22	22.32
QA_5	$\mathcal{F}_5$, $\mathcal{F}_{15}$, $\mathcal{F}_{25}$, $\mathcal{F}_{35}$, $\mathcal{F}_{45}$	$O(N^2)$	25	36.93
QA_6	$\mathcal{F}_6$, $\mathcal{F}_{16}$, $\mathcal{F}_{26}$, $\mathcal{F}_{36}$, $\mathcal{F}_{46}$	$O(N^2)$	22	37.02
QA_7	$\mathcal{F}_7$, $\mathcal{F}_{17}$, $\mathcal{F}_{27}$, $\mathcal{F}_{37}$, $\mathcal{F}_{47}$	$O(\delta N)$	3	0.484
QA_8	$\mathcal{F}_8$, $\mathcal{F}_{18}$, $\mathcal{F}_{28}$, $\mathcal{F}_{38}$, $\mathcal{F}_{48}$	$O(N^2)$	9	0.827
QA_9	$\mathcal{F}_9$, $\mathcal{F}_{19}$, $\mathcal{F}_{29}$, $\mathcal{F}_{39}$, $\mathcal{F}_{49}$	$O(N^2)$	9	0.343
QA_{10}	$\mathcal{F}_{10}$, $\mathcal{F}_{20}$, $\mathcal{F}_{30}$, $\mathcal{F}_{40}$, $\mathcal{F}_{50}$	$O(N^2)$	9	0.718
Dense disparity estimate	-	$O(\delta N)$	10	11.30

the CPU times are shown in Table 3 and Figure 12. From Table 4, we see that the previously found feature group, $\mathcal{F}_{\{15,18,30,34,47\}}$, is not the group with the lowest complexity, with a McCabe complexity of 108 and a running time for a single image set of 6.11 s. The fastest quality measure, $\mathcal{F}_{\{25,28,30,41,48\}}$, does not contain a component that is sensitive to binocular rivalry. Therefore, by considering the complexity, the correlation performance, and the sensitivity of the metric to different artifacts, we selected the second-fastest feature group $\mathcal{F}_{\{25,28,33,41,48\}}$ for the final quality metric. The output of each feature was normalized according to formula (33). The weighting and normalization coefficients used for each feature are given in Table 6.

This selection is also confirmed by the results of the full searches over six features. These combinations reach correlation performances of 0.93, but at considerably higher computational costs. However, we can see that the feature evaluation components from the first two groups (CV and BR) tend to dominate the best performing combinations. It should be noted that the performance is calculated based only on the training subset of test videos, and by selecting a three-component combination, we aim to provide a balanced combination for a wider, and possibly more diverse, set of videos.

6 Comparative results

The prediction performance of an objective quality metric can be evaluated in terms of accuracy, monotonicity, and association. We use the normalized root mean squared error (RMSE), the SROCC, and the Pearson linear correlation coefficient (PLCC) to quantify the corresponding performance properties of our metric. Before calculation of the correlation performance, we apply a logistic fitting function to all quality metrics under comparison.

The subjective experiments performed on the two sets of test sequences have been analyzed in [15,16]. Some findings relevant to our current work are summarized here. The results of subjective experiments, involving 3D database I were interpreted in [15] as both the artifact level and the presence of stereoscopic depth affect the user acceptance of and satisfaction with the 3D video sequences. Also, according to the subjective test results for database II [16], MVC and the V + D approach provide the best subjective quality for all compression levels. We believe that a well-performing 3D quality metric should be able to predict these subjective preferences.

We compared the feature group proposed in Section 5 (i.e., $\mathcal{F}_{\{25,28,33,41,48\}}$) with several state-of-the-art quality metrics. The results are as shown in Tables 7 and 8. The metrics that were intended for 2D image quality [i.e., peak signal-to-noise ratio (PSNR), SSIM, normalized root mean squared error (NRMSE), and PSNR-HVS] have been applied separately to the left and right channels and the final results have been averaged. In the PSNR case, the MSE derived in each channel was averaged in advance. Four metrics that predict the quality of the stereoscopic content have been included in the comparison: PHVS3D [2], PHSD [3], 3DBE [7], and the stereo metric, described

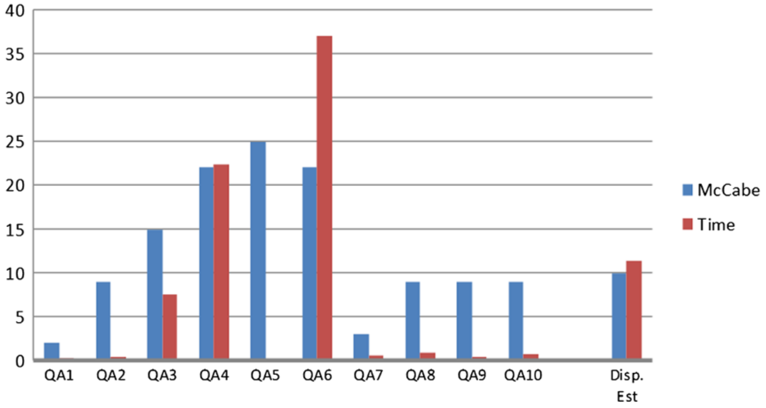

Figure 12 QA computational complexity comparison.

Table 4 Computational comparisons for five quality features (full search)

	Feature selection	Q^{CR}	Q^{BR}	Q^{DQ}	McCabe	Time (s)	SROCC
1	$\mathscr{F}_{\{25,28,30,41,48\}}$	+		+	54	3.95	0.9260
2	$\mathscr{F}_{\{15,16,17,25,29\}}$	+			84	11.16	0.9255
3	$\mathscr{F}_{\{9,15,16,32,35\}}$	+	+		90	11.15	0.9243
4	$\mathscr{F}_{\{25,28,34,41,48\}}$	+	+	+	67	6.11	0.9242
5	$\mathscr{F}_{\{25,28,33,41,48\}}$	+	+	+	*60*	*4.63*	*0.9236*
Sequential	$\mathscr{F}_{\{15,18,30,34,47\}}$	+	+	+	68	6.13	0.9106

The values for the selected set of features are marked in italic.

in [5]. The quality values for 3DBE were kindly provided by the authors of the metric in [7]. All metrics work on the luminance components of the images.

The SROCC, PLCC, and normalized RMSE values for each compared QA on 3D databases I and II can be seen in Tables 7 and 8, respectively. For visual comparison, prediction results for databases I and II are shown in Figures 13 and 14, respectively. To quantify the performance in terms of their different aspects, the videos from both databases are grouped into several subsets. Test sequences in 3D database I are classified into three subsets based on the depth levels in Table 7: 'mono' is used for monoscopic sequences, and short and wide are used for stereoscopic sequences. 3D database II is grouped into four subsets based on the encoding methods used, i.e., MRSC, MVC, SIM, and V + D. The two algorithms with the best performance levels are marked in bold.

Table 6 Linearization and weighting coefficients used in the final quality metric

	$\mathscr{F}_{25}$	$\mathscr{F}_{28}$	$\mathscr{F}_{33}$	$\mathscr{F}_{41}$	$\mathscr{F}_{48}$
Weighting coefficient	1.8627	−1.0692	0.1202	0.4880	−0.4443
β_1	39.08	9.896	0.309	0.088	9.896
β_2	−166.6	370.8	3.281	9.749	370.8
β_3	4,483	3,577	−0.809	−0.866	3,577
β_4	0.139	0.114	803.6	454.2	0.1142

The predominant distortions in this database are caused by DCT-based compression and are manifested as blocking and smearing artifact characteristic for harsh quantization levels. These distortions affect the cyclopean view quality and can be detected by quality metrics, sensitive to texture degradation. PSNR-HVS produces the third best performance on the mono set, where the SROCC, PLCC, and RMSE values are 0.921, 0.917, and 0.716, respectively. PHSD and PHVS3D also correlate well with the MOS in that database. PHSD is an improved version of PHVS3D, in which the disparity errors are considered. The SSIM metric, if used separately in each channel, does not correlate well with the subjective scores of 3D database I. The proposed combination of five quality features has the best correlation with the MOS, which were compared using either SROCC or PLCC. The overall correlations of SROCC, PLCC, and RMSE reach 0.935, 0.924, and 0.684 correspondingly. Most QA metrics fail on the wide baseline sets. The proposed metric shows higher correlations on *all* subsets,

Table 5 Computational comparisons for six quality features (full search)

Combination number	Feature selection	Q^{CR}	Q^{BR}	Q^{DQ}	McCabe	Time (s)	SROCC
1	$\mathscr{F}_{\{9,15,16,22,32,35\}}$	+	+		99	11.1868	0.9358
2	$\mathscr{F}_{\{15,16,22,29,32,35\}}$	+	+		99	11.1868	0.9352
3	$\mathscr{F}_{\{15,16,19,22,32,35\}}$	+	+		99	11.1868	0.9331
4	$\mathscr{F}_{\{9,12,15,16,32,35\}}$	+	+		99	11.1868	0.9329
5	$\mathscr{F}_{\{15,16,24,29,32,35\}}$	+	+		112	13.3864	0.9325
6	$\mathscr{F}_{\{9,15,16,24,32,35\}}$	+	+		112	13.3864	0.9321
7	$\mathscr{F}_{\{12,15,16,29,32,35\}}$	+	+		99	11.1868	0.9317
8	$\mathscr{F}_{\{15,16,23,29,32,35\}}$	+	+		105	11.9075	0.9317
9	$\mathscr{F}_{\{9,15,16,23,32,35\}}$	+	+		105	11.9075	0.9315
10	$\mathscr{F}_{\{9,15,16,21,32,35\}}$	+	+		92	11.1696	0.9313
11	$\mathscr{F}_{\{15,16,21,29,32,35\}}$	+	+		92	11.1696	0.931
12	$\mathscr{F}_{\{5,9,15,16,17,35\}}$	+	+		109	14.8621	0.9309
13	$\mathscr{F}_{\{9,15,16,17,32,35\}}$	+	+		93	11.2024	0.9309
14	$\mathscr{F}_{\{5,15,16,17,29,35\}}$	+	+		109	14.8621	0.9309
15	$\mathscr{F}_{\{12,25,28,30,41,48\}}$	+		+	63	3.9781	0.9306
16	$\mathscr{F}_{\{12,15,16,19,32,35\}}$	+	+		99	11.1868	0.9305
17	$\mathscr{F}_{\{9,12,25,28,41,48\}}$	+		+	63	3.9406	0.9304
Sequential	$\mathscr{F}_{\{15,18,30,34,47,22\}}$	+	+	+	77	6.1606	0.9162

Table 7 Spearman and Pearson correlations of compared metrics on 3D video database I

		PSNR	SSIM	PSNR-HVS	NRMSE	3DBE	Ref [5]	PHVS3D	PHSD	Proposed
SROCC	Mono	0.875	0.704	0.921	0.857	0.782	0.703	*0.929*	0.918	**0.939**
	Short	0.883	0.683	0.907	0.874	0.602	0.702	0.910	*0.935*	**0.956**
	Wide	0.850	0.599	0.877	0.833	0.549	0.609	0.896	*0.934*	**0.952**
	All	0.864	0.623	0.886	0.841	0.649	0.631	*0.917*	0.865	**0.935**
PLCC	Mono	0.874	0.768	0.917	0.903	0.801	0.755	*0.927*	0.915	**0.949**
	Short	0.877	0.756	0.915	0.920	0.577	0.739	0.914	*0.928*	**0.942**
	Wide	0.820	0.681	0.865	0.857	0.530	0.671	0.887	*0.920*	**0.941**
	All	0.843	0.707	0.885	0.879	0.613	0.694	*0.906*	0.844	**0.924**
RMSE	Mono	0.864	1.237	*0.716*	0.747	1.201	1.307	**0.695**	1.026	0.858
	Short	0.823	1.206	0.698	0.707	1.497	1.312	*0.693*	0.750	**0.556**
	Wide	0.921	1.346	0.781	0.830	1.587	1.449	*0.774*	0.821	**0.635**
	All	0.874	1.293	0.737	0.783	1.438	1.362	*0.715*	0.873	**0.684**

Values in bold indicate the best score; values in italic indicate the second best.

and the SROCC values of the short and wide baseline subsets are quite consistent at 0.956 and 0.952, respectively.

3D database II contains a wider range of video distortions, most notably some cases of severe binocular rivalry. Such distortions do not affect large areas of the image but are immediately visible to the observer. As a result, quality metrics assessing texture quality tend to grade such cases as being of high quality, while observers grade them as having annoying artifacts. Most of the metrics included in our comparison fail on the 'V + D' set, particularly PSNR, PSNR-HVS, PHVS3D, and SSIM, for which the PLCC values are less than 0.1. This can be attributed to the presence of binocular rivalry artifacts which are caused by view rendering based on the estimated depths. For most videos exhibiting stereoscopic distortions, 2D metrics fail to predict the subjective scores. The overall SROCC values of PSNR and PSNR-HVS are only 0.254 and 0.227, respectively. Although the results for SSIM and NRMSE are slightly improved, their overall SROCC values are still very low. Among the 3D quality metrics, the PHVS3D metric does not perform well, but the improved PHSD version has the second best correlation with all the MOS in the database. Finally, in Table 8, we see that the metric proposed in this paper shows better performance because it is sensitive to a wider range of stereoscopic distortions.

Table 8 Spearman and Pearson correlations of compared metrics on 3D video database II

		PSNR	SSIM	PSNR-HVS	NRMSE	Ref [5]	PHVS3D	PHSD	Proposed
SROCC	MRSC	0.076	0.398	0.399	0.452	0.562	0.306	*0.649*	**0.910**
	MVC	0.328	0.587	0.423	0.608	0.803	0.264	*0.810*	**0.973**
	SIM	0.368	0.543	0.418	0.561	0.699	0.292	*0.778*	**0.932**
	V + D	0.050	0.205	0.094	0.316	0.522	0.221	*0.737*	**0.877**
	All	0.254	0.443	0.227	0.413	0.646	0.323	*0.799*	**0.942**
PLCC	MRSC	0.196	0.361	0.369	0.487	0.519	0.271	*0.536*	**0.906**
	MVC	0.293	0.432	0.380	0.617	0.649	0.202	*0.726*	**0.963**
	SIM	0.301	0.420	0.411	0.561	0.564	0.317	*0.778*	**0.940**
	V + D	−0.170	0.099	−0.057	0.211	0.308	0.002	*0.653*	**0.907**
	All	0.236	0.379	0.223	0.425	0.541	0.294	*0.730*	**0.942**
RMSE	MRSC	1.581	1.475	1.588	1.488	1.315	1.468	*1.242*	**0.667**
	MVC	1.785	1.529	1.685	1.437	1.404	1.812	*1.258*	**0.544**
	SIM	1.883	1.699	1.854	1.691	1.579	1.769	*1.240*	**0.664**
	V + D	1.674	1.771	1.869	1.877	1.641	1.611	*1.143*	**0.754**
	All	1.735	1.623	1.753	1.633	1.491	1.671	*1.222*	**0.661**

Values in bold indicate the best score; values in italic indicate the second best.

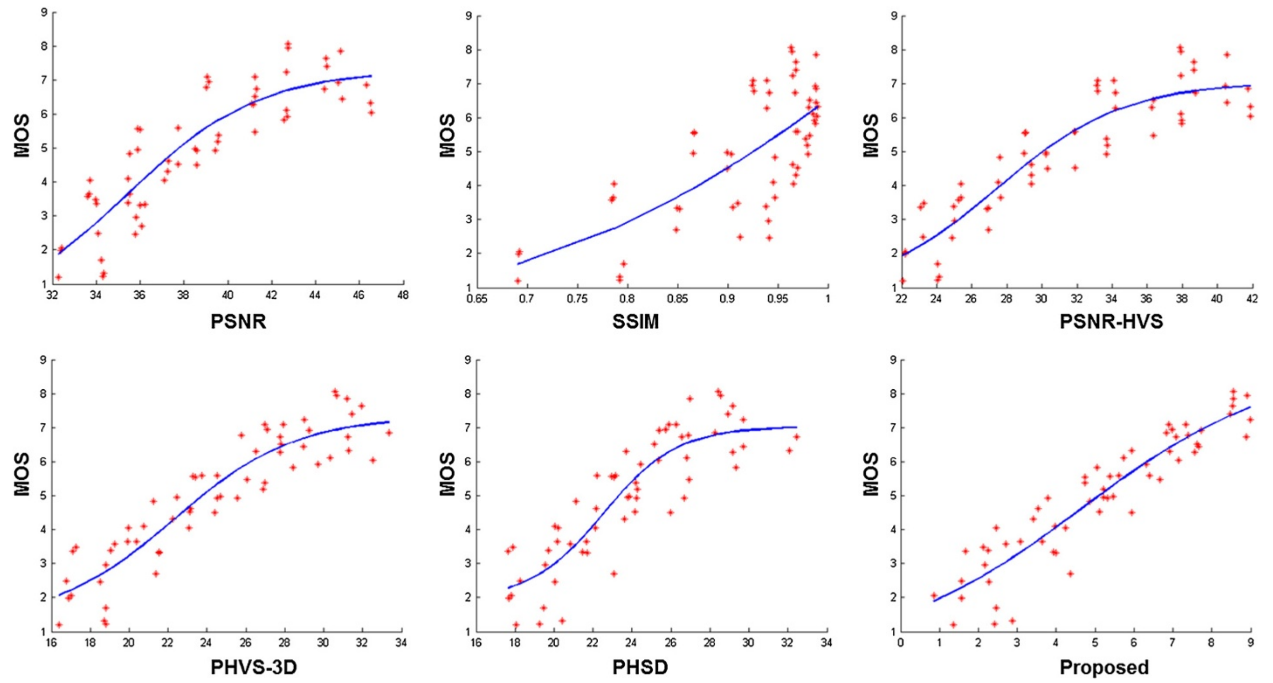

Figure 13 Logistic fitting figures of the different and proposed metrics on 3D video database I.

7 Conclusions

One of the biggest challenges in 3D QA is the calculation of the QA metric in a perceptual manner. In this paper, a novel full-reference stereoscopic quality metric that is applicable to mobile 3D video has been proposed. First, we built two 3D quality databases that were annotated with subjective test results in terms of their MOS. The databases include not only compression distortions but also differently pronounced depth and 3D format conversion distortions. According to the results of subjective

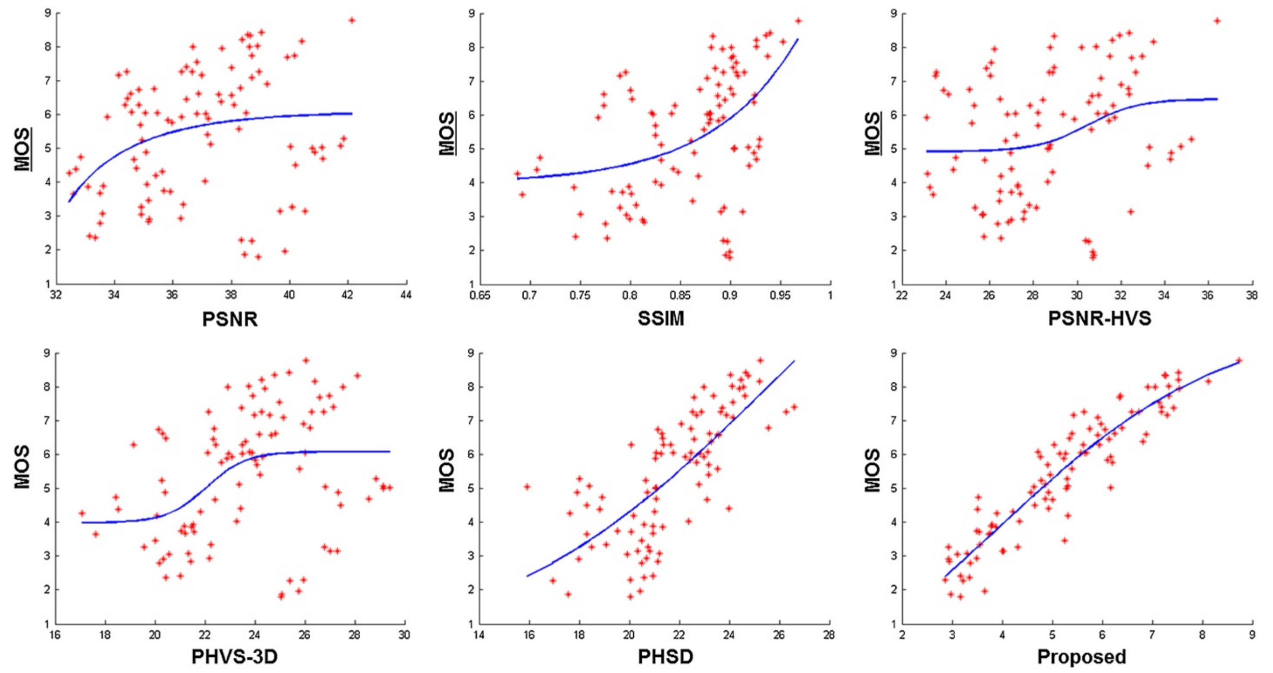

Figure 14 Logistic fitting figures of the different metrics and proposed metric on 3D video database II.

tests and interviews with the test participants [15], the number of compression artifacts is dominant in the evaluation of the content quality, whereas the presence of depth enhances the user experience. The viewers were very critical of the spatial quality and accepted only low numbers of artifacts in the content. The 3D effect enhances the user satisfaction and acceptance of the content; however, if the content is not presented with high spatial quality, then the content was declared to be less acceptable or completely unacceptable, regardless of the 3D effect. Motivated by these results, we modeled the 3D quality using three components: the cyclopean view, binocular rivalry, and the depth quality. The cyclopean view is simulated using three models. The first model generates a single cyclopean image by globally fusing the left and right views of a scene based on the properties of human stereo vision. The second and third models are based on local fusion methods, which calculate the quality on the block level between the left and right channels using a disparity map. Dissimilar visual stimuli between the two eyes bring binocular rivalry. In our approach, the amount of binocular rivalry is quantified by comparison of only the corresponding blocks in the distorted stereoscopic pair, using the disparity map that is provided by the reference pair. The differences between the images of a scene as seen by each eye are also used to form the perceived depth. The geometrical distortions are measured directly on the disparity map (and are called the depth quality).

Several QA methods are used to assess each quality component, with tests conducted using a training set that was extracted from the two available databases. To make the quality metric simple, fast, and efficient, the feature selection for all considered QAs is processed by studying their computational complexity and the CPU run times. Finally, six features are selected for the three components. The cyclopean view is measured by two quality assessment methods, i.e., QA_5 and QA_8, which are both under the third (local) cyclopean view model; binocular rivalry is evaluated using QA_3; and the depth quality is measured using the disparity map with QA_1 and QA_8. The experimental results have shown that the proposed metric significantly outperforms the current state-of-the-art quality metrics. We must note that our implementation does not take masking effects created by motion into account. This is will be studied in our future investigations. However, our experiments to date have shown that this masking plays a minor role in estimation of the quality. This observation has been confirmed by subjective tests on still images and videos with the same content, which resulted in very similar MOS.

Abbreviations

CSF: contrast sensitivity function; CWT: complex wavelet transform; DCT: discrete cosine transform; FSIM: feature similarity index; GM: gradient magnitude; HVS: human visual system; IQA: image quality assessment; LGN: lateral geniculate nucleus; MOS: mean opinion scores; MRSC: mixed resolution stereo coding; MSE: mean squared error; MVC: multiview video coding; PC: phase congruency; PLCC: Pearson rank-order correlation QA, quality assessment; QP: quantization parameters; SSD: sum-of-squared-difference; SSIM: structural similarity index; SROCC: Spearman rank-order correlation; V + D: video plus depth.

Competing interests

The authors declare that they have no competing interests.

Received: 10 June 2013 Accepted: 23 December 2013
Published: 21 January 2014

References

1. A Boev, M Poikela, A Gotchev, A Aksay, Modeling of the stereoscopic HVS, Mobile3DTV. Technical Report D5.3. (2010). Available at http://sp.cs.tut.fi/mobile3dtv/results/tech/D5.3_Mobile3DTV_v2.0.pdf. Accessed on 13 January 2014
2. L Jin, A Boev, A Gotchev, K Egiazarian, 3D-DCT base perceptual quality assessment of stereo video, in *IEEE 18th International Conference on Image Processing (IEEE ICIP2011)*. Brussels, 11–14 September 2011
3. L Jin, A Boev, A Gotchev, K Egiazarian, Validation of a new full reference metric for quality assessment of mobile 3DTV content, in *The 19th European Signal Processing Conference (EUSIPCO-2011)*. Barcelona, 29 August to 2 September 2011
4. A Boev, A Gotchev, K Egiazarian, A Aksay, GB Akar, Towards compound stereo-video quality metric: a specific encoder-based framework, in *IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 218–222. Denver, June 2006
5. J You, L Xing, A Perkis, X Wang, Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis, in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM*. Scottsdale, 13–15 January 2010
6. K Wang, K Brunnström, M Barkowsky, M Urvoy, M Sjöström, P Le Callet, S Touranchean, B Andrén, Stereoscopic 3D video coding quality evaluation with 2D objective metrics. Proc. SPIE Electronic Imaging , 8648 (2013). Stereoscopic Displays and Applications XXIV, 86481L, San Francisco, March 12, 2013. doi: 10.1117/12.2003664
7. R Bensalma, MC Larabi, Towards a perceptual quality metric for color stereo images, in *IEEE 17th International Conferences on Image Processing*. Hong Kong, 26–29 September 2010
8. E Bosc, R Pepion, P Le Callet, M Koppel, P Ndjiki-Nya, M Pressigout, L Morin, Towards a new quality metric for 3-D synthesized view assessment. IEEE J. Sel. Top. Sign. Proces. **5**(7), 1332–1343 (2011)
9. PR Lebreton, A Raake, M Barkowsky, P Le Callet, Evaluating depth perception of 3D stereoscopic videos. IEEE Sel. Top. Sign. Proces. **6**(6), 710–720 (2012)
10. P Hanhart, T Ebrahimi, Quality assessment of a stereo pair formed from two synthesized views using objective metrics, in *Proceedings of Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2013)*. Scottsdale, 30 January to 1 February 2013
11. L Xing, J You, T Ebrahimi, A Perkis, Assessment of stereoscopic crosstalk perception. Multimedia, IEEE Trans. **14**(2), 326–337 (2012)
12. A Maalouf, MC Larabi, CYCLOP: a stereo color image quality assessment metric, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1161–1164. Prague, 22–27 May 2011
13. P Seuntiëns, Visual Experience of 3D TV, Thesis. (2006)
14. MTM Lambooij, WA Ijsselsteijn, I Heynderickx, Visual discomfort in stereoscopic displays: a review, in *Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems XIV*. San Jose, 1–13 January 2007
15. D Strohmeier, S Jumisko-Pyykkö, K Kunze, G Tech, D Bugdayci, M Oguz Bici, Results of quality attributes of coding, transmission and their combinations, Mobile 3DTV Technical Report D4.3. , (2010). Available at http://sp.cs.tut.fi/mobile3dtv/results/tech/D4.3_Mobile3DTV_v1.0.pdf. Accessed on 13 January 2014
16. L Jin, A Boev, S Jumisko-Pyykkö, T Haustola, A Gotchev, Novel stereo quality metrics, MOBILIE 3DTV Technical Report D5.5, (2011). Available at http://sp.cs.tut.fi/mobile3dtv/results/tech/D5.5_Mobile3DTV_v1.0.pdf. Accessed on 14 January 2014
17. A Wandell Brian, *Foundations of Vision* (Sinauer Associates, Sunderland, 1995)

18. R Blake, A primer on binocular rivalry, including current controversies. Brain Mind **2**(1), 5–38 (2011)
19. S Knorr, K Ide, M Kunter, T Sikora, Basic rules for good 3D and avoidance of visual discomfort, in *International Broadcasting Convention (IBC)*. Amsterdam, 8–13 September 2011
20. S Smirnov, A Gotchev, M Hannuksela, Comparative analysis of local binocular and trinocular depth estimation approaches. Proc. of SPIE **7724**(2010) doi:10.1117/12.854765
21. S Baker, D Scharstein, J Lewis, S Roth, M Black, R Szeliski, A database and evaluation methodology for optical flow, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 243–246. Crete, Greece, 14–21 October 2007
22. K Egiazarian, J Astola, N Ponomarenko, V Lukin, F Battisti, M Carli, New full-reference quality metrics based on HVS, in *International Workshop on Video Processing and Quality Metrics*, p. 4. Scottsdale, January 2006
23. N Ponomarenko, F Silvestri, K Egiazarian, M Carli, J Astola, V Lukin, On between-coefficient contrast masking of DCT basis functions, in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 25–26. Scottsdale, January 2007
24. L Zhang, L Zhang, X Mou, D Zhang, FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
25. P Kovesi, Image features from phase congruency. Videre: J Comput Vision Res (1999). MIT Press. Volume 1, Number 3
26. Z Wang, A Bovik, H Sheikh, E Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
27. T Hastie, R Tibshirani, J Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer, Heidelberg, 2009)
28. T Wiegand, GJ Sullivan, G Bjøntegaard, A Luthra, Overview of the H.264/AVC video coding standard. IEEE Trans. Circ. Syst. Video Tech. **13**, 560 (2003)
29. TJ McCabe, A complexity measure. IEEE Trans. Soft. Eng. **SE-2**(4), 308 (1976)

doi:10.1186/1687-5281-2014-6
Cite this article as: Jin *et al.*: **Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3D video.** *EURASIP Journal on Image and Video Processing* 2014 **2014**:6.