applied sciences

MDPI

Article

No-Reference Quality Assessment of Transmitted Stereoscopic Videos Based on Human Visual System

Md Mehedi Hasan [1,*], Md. Ariful Islam [1], Sejuti Rahman [1], Michael R. Frater [2] and John F. Arnold [2]

[1] Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[2] School of Engineering and Information Technology, University of New South Wales, Canberra 2600, Australia
* Correspondence: mmhasan@du.ac.bd

Abstract: Provisioning the stereoscopic 3D (S3D) video transmission services of admissible quality in a wireless environment is an immense challenge for video service providers. Unlike for 2D videos, a widely accepted No-reference objective model for assessing transmitted 3D videos that explores the Human Visual System (HVS) appropriately has not been developed yet. Distortions perceived in 2D and 3D videos are significantly different due to the sophisticated manner in which the HVS handles the dissimilarities between the two different views. In real-time video transmission, viewers only have the distorted or receiver end content of the original video acquired through the communication medium. In this paper, we propose a No-reference quality assessment method that can estimate the quality of a stereoscopic 3D video based on HVS. By evaluating perceptual aspects and correlations of visual binocular impacts in a stereoscopic movie, the approach creates a way for the objective quality measure to assess impairments similarly to a human observer who would experience the similar material. Firstly, the disparity is measured and quantified by the region-based similarity matching algorithm, and then, the magnitude of the edge difference is calculated to delimit the visually perceptible areas of an image. Finally, an objective metric is approximated by extracting these significant perceptual image features. Experimental analysis with standard S3D video datasets demonstrates the lower computational complexity for the video decoder and comparison with the state-of-the-art algorithms shows the efficiency of the proposed approach for 3D video transmission at different quantization (QP 26 and QP 32) and loss rate (1% and 3% packet loss) parameters along with the perceptual distortion features.

Keywords: quality assessment; stereoscopic video; disparity index; human visual system; no-reference

Citation: Hasan, M.M.; Islam, M.A.; Rahman, S.; Frater, M.R.; Arnold, J.F. No-Reference Quality Assessment of Transmitted Stereoscopic Videos Based on Human Visual System. *Appl. Sci.* **2022**, *12*, 10090. https://doi.org/10.3390/app121910090

Academic Editor: Andrea Prati

Received: 2 September 2022
Accepted: 26 September 2022
Published: 7 October 2022

1. Introduction

Quality Assessment is an imperative aspect of video services aimed at human observers in applications such as television, Blu-ray, DVD, mobile TV, web TV, gaming, and video streaming. An objective 3D video QA is a statistical mathematical model that approximates the results for video perception that would be obtained from typical human viewers [1]. No-reference (NR) models can objectively estimate a video's quality based on the received frames that have been subjected to distortions from coding and transmission losses [2]. Since it does not require any information from the source video, it generates less precise scores when evaluating the quality of a video compared to full- and reduced-reference approaches, but can be applied in many real-time applications for which source information is unavailable. NR models are widely used for continuous quality monitoring at the receiver end in video playback and streaming systems [1], with a basic system shown in Figure 1. The figure shows a conventional quality measure system for video transmission where stereoscopic videos are differently encoded, transmitted, and then decoded in the receiving system. For any kind of transmission losses or distortion, the end devices perform some checking of the videos, perform real-time error concealment, and then show it to the viewer.

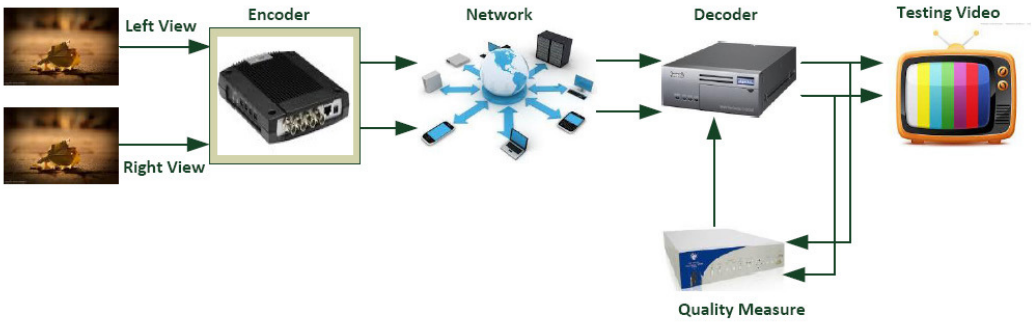


Figure 1. Typical model of no-reference video quality measure.

3D video transmission has received significant research attention driven by commercial applications in recent years [1,3,4]. More attention has been paid to analyzing and mitigating the effects by applying various techniques in the 3D video transmission chain and display to ensure better Quality of Experience (QoE) [5]. However, the impacts of artifacts introduced into a 3D video by its transmission system have not received as much interest as those in a 2D video [6] although they influence the overall image quality similarly. Because stereoscopic 3D video has two separate channels, each of which may experience unrelated attenuation, transmission errors over unreliable wireless communication channels are higher than in 2D video. For example, lag in a view can lead to a time-lapse out-of-sync process, reducing the comfort of 3D [7] viewing. In addition, the methods used to reduce these spurious structures (e.g., error masking) do not work as well for 3D videos as they do for 2D [8] videos. To create a good 3D depth perception [9,10] and avoid the competition of two binoculars [11,12], ideally the two 3D channels should be maintained and synchronized.

Numerous factors, including spatial or temporal frequencies, binocular depth cues, and transmission media, can significantly affect the video quality experienced by users during 3D viewing. Previous studies of binocular vision and its sensitivity showed that the behaviors of viewers during 2D and 3D viewing are different and the observations are tightly linked to the way in which 3D videos are perceived [13–15]. Existing 2D image or video quality models are not adequate for measuring 3D visual experiences since they do not incorporate various 3D perceptual characteristics, including at least the most positive (binocular depth) and negative (cross-talk and visual discomfort) factors of S3D [7,12].

As shown in Figure 2, the brain employs binocular disparity to extract depth information from two-dimensional (2D) retinal pictures. The difference in the coordinates of similar features in two stereo views is referred to as "binocular disparity". The horopter is a curving line that connects all sites that have a zero retinal disparity (same relative coordinate), with the points positioned at it having the same perceived distances from a human subject's fixation point. Panum's fusional area is located around the horopter and is a zone within which views with non-zero retinal disparities can be merged binocularly; however, things located outside of this area result in double pictures. Panum's area size varies across the retina and is regulated by the spatial and temporal characteristics of the fixation object [16]. When a person focuses on an item, a picture of that object is formed on the retina, and objects that are closer to or farther away from the accommodation distance are perceived as fuzzy pictures. Objects that lay within a limited zone around the accommodation point, however, might be regarded as having a high resolution (i.e., not blurred), with the extent of this region known as the depth of field (DOF) [16].

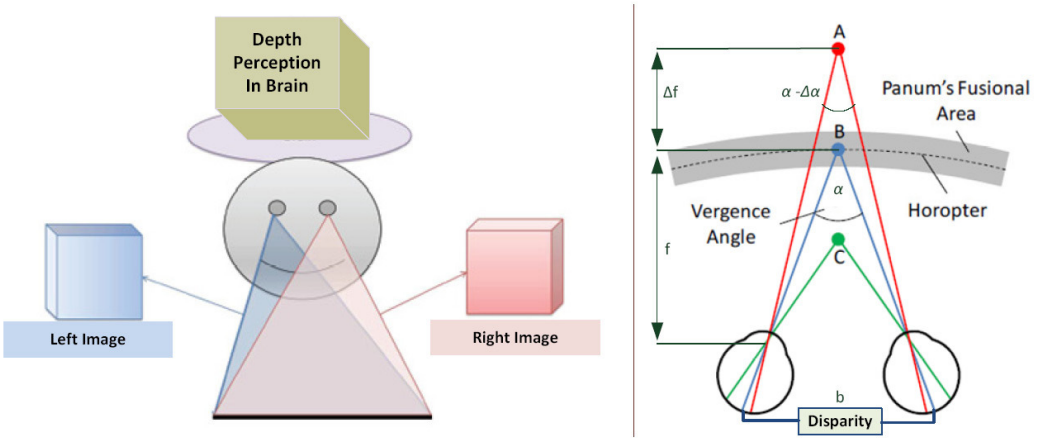


Figure 2. Binocular disparity and corresponding depth perception in the brain.

The geometry of stereopsis depicted in Figure 2 corresponds to the experimental setting of the two-needle test [17] with a fixation point B. The theoretical depth discrimination Δf, according to it, can be determined from the definition of the convergence angle α by applying $d\alpha/df$, which provides

$$\Delta f = -\frac{f^2}{b}\left(1 + \frac{b^2}{4f^2}\right) \cdot \Delta\alpha \tag{1}$$

with b the interpupillary distance and f the mean object distance. With $4f^2 \gg b^2$, which is even valid for a near point distance $f_{near} = 250$ mm, Equation (1) simplifies to the common form $\Delta f \approx -f^2 \frac{\Delta\alpha}{b}$. In binocular vision, $\Delta\alpha$ is the stereoscopic acuity (the smallest detectable depth difference) which is required to get a proper binocular or 3D depth perception. Scientifically, it is observed that $\Delta\alpha = 10\,arcsec$, (a tolerable value under photopic lighting conditions) which is added to move the view from fixation point A $(\alpha - \Delta\alpha)$ to (α) and for more depth perception B (α) to C $(\alpha + \Delta\alpha)$. The minimum detectable depth difference being on the order of 0.3 mm for a fixation distance of 650 mm. Among other factors, stereoscopic acuity is affected mainly by parameters such as an object's luminance and spatial frequency, and angular distances from fixation and object motion. Therefore, a multidimensional 3D visual experience model needs to be defined to incorporate the aforementioned acuity factors based on their perceptual importance.

Effective 3D video quality evaluation schemes should be designed based on the observations which will allow us to (i) synchronize between the two views of stereoscopic video, (ii) adjust parameters to maximize overall quality and comfort, (iii) control visual quality during transmission, and (iv) define the levels of quality for specific video services. The main goal of this research work is designing an NR quality metric for evaluating 3D videos considering different aspects of human binocular perception and real-time transmission. In Section 2, we include a summarized discussion of existing relevant research on 3D quality assessment and Section 3 introduces the proposed NR-based approach with a detailed algorithm presented in the subsections. The experiments and final results are discussed in Section 4. Finally, Section 5 concludes this research with the directions on possible future works.

2. Previous Relevant Research

S3D QA algorithms could be categorized based on the perspective binocular perception. The first consists of 2D-based 3D QA models which do not utilize depth information from stereo pairs. In [18], some well-known 2D image quality metrics (PSNR, SSIM, MSSIM, VSNR, VQM, Weighted SNR, and JND) are introduced and their capabilities for the stereoscopic image were investigated. Structural Similarity (SSIM) is a Full-Reference (FR) objective video quality metric proposed by Wang et al. [19] which is based on the

assumption that human visual perception is highly adapted for extracting structures from a scene. It compares local patterns of pixel intensities normalized for luminance and contrast. In [20], several 2D objective video quality metrics (SSIM, Universal Quality Index (UQI), C4, and RR Image QA (RRIQA)) for left and right images are combined using an average image distortion approach and visual acuity technique based on disparity distortion for the QA of a stereo image. The approaches in [21] apply 2D QA algorithms independently on the left and right perspectives, and then aggregate the two scores (by different means) to forecast 3D quality. In addition, Ryu et al. [22] presented a 3D quality score as the weighted sum of the left and right perspectives' quality scores. According to Meegan et al. [23], the binocular sense of the quality of asymmetric MPEG-2 distorted stereo pictures is approximately the average of the two views, but the perception of asymmetric blur in a distorted stereo image is mostly influenced by the higher-quality view. Since most approaches accumulate standard 2D image or video features rather than considering the psychovisual aspects of 3D, the outcome directed the investigation of a binocular rivalry or acuity measure by obtaining the disparity between the left and right views.

The second category of models considers the binocular (depth) information and estimates the disparity map in the overall process. Shen et al. [1] presented an NR QA approach imitating the HVS perception route, and from the fused and single view, they derived the features throughout the global feature fusion sub-network. Liu et al. [14] developed a new NR stereoscopic image quality perception model that incorporated monocular and binocular features by the relationship between visual features and stereoscopic perception. Blenoit et al. [24] developed an FR 3D QA algorithm that computes the differences between the quality scores of the left and right videos calculated from reference and distorted views, as well as a distorted disparity map. C4 [25] and SSIM [26] are used to compute these three quality scores, and different combinations of them are employed to obtain the final results. Their findings suggest that discrepancy information can improve the SSIM-based 3D QA algorithm (called 3D-SSIM). You et al. [27] expanded the concept of forecasting the 3D quality of stereo pairs by using SSIM and the mean absolute difference (MAD) of their predicted disparity maps. Bensalma et al. [12] presented a 3D QA technique based on assessing the difference in binocular energy between reference and tested stereo pairs, which takes into account the potential influence of binocular effects on perceived 3D quality. In the RR 3D QA proposed by Hewage et al. [28], just the depth map's edge information is sent, and the PSNR of the reference frame is used to estimate the overall score. Akhter et al. [29] proposed an NR 3D QA algorithm which extracts features and estimates the disparity map, with a logistic regression model then used to predict the 3D quality scores. Wang et al. [30] introduced a 3D QA model based on the suppression theory of binocular fusion, in which 2D Image Quality Metric (IQM) distortion maps are generated for both the left and right views. Then, a binocular spatial sensitivity module is incorporated with these maps to generate the final quality metric.

Some recent feature-based approaches also contributed significantly in the NR-based video quality assessment research. Varga et al. [31] provided a potent feature map for NR-VQA that draws inspiration from Benford's law. It is shown that the first-digit patterns recovered from the video volume data's various transform domains are quality-aware attributes that can be efficiently projected towards sensory quality reporting. Based on the research findings, the authors in [32] suggested a support vector regression (SVR)-based supervised learning strategy to tackle the no-reference video quality assessments (NRVQA) issue. Authors claimed that the suggested method provides satisfactory accuracy on real aberrations and competitive intensity on conventional (fake) aberrations. Ebenezer et al. [33] suggested a fresh working prototype for no-reference video quality assessment (VQA) that is grounded on the inherent characteristics of video's space-time chips. The phrase "space-time chips" (ST-chips) refers to a brand-new, quality-aware feature set that we characterize as confined spatial slices of multimedia data that follow the motion flow of the local region. They demonstrated that the parameters from such frameworks are distorted-sensitive and can thus be utilized to forecast the grade of movies using generalized

dispersion fitting to the band-pass histograms of space-time chips. Saad et al. proposed a blind video evaluation model (no reference or NR) that is not specific to the [34] distortion. This method is based on the spatial-temporal model of video scenes in the signal processing domain (discrete cosine transform) and a computational model that classifies the motion occurring in the scenes to predict the quality video.

Deep learning has received attention in recent years due to the availability of benchmark video QA databases. Varga et al. [2] described a novel, deep learning-based strategy for NR-VQA that used several pre-trained convolutional neural networks (CNN) to classify the probable image and video distortions in parallel. To extract the spatial and temporal features from the stereoscopic videos, H. Imami et al. [3] proposed a quality assessment method based on a 3D convolution neural network with capturing the disparity information. Zhang et al. [35] proposed a synthesized video denoising algorithm based on CNN for the elimination of temporal flicker distortion and enhanced 3D synthetic video perceptual quality. Feng et al. [36] presented a multi-scale feature-directed 3D CNN for QA that used 3D convolution to catch the spatiotemporal features as well as a novel multi-scale unit to accumulate multi-scale information. Jin et al. [4] proposed a no-reference image quality assessment method for measuring 3D composite images based on visual entropy-oriented multi-layer feature analysis. However, all of the above approaches combine high computational complexity, creating longer latency, and thus lack real-time capability for end-user electronics after transmission.

In general, it is difficult to judge the quality of the perceived depth because the depth of ground truth is generally not available during transmission. These models can only evaluate depth quality using an estimated disparity map (calculated from an empty stereo pair or a distorted stereo pair) which is significantly affected by the accuracy of the sound. In addition, most existing 3D video quality models are validated using symmetrically encoded videos. In addition, conventional quality assurance methods depend on FR or RR criteria, making it difficult to judge from damaged transmitted or broadcast movies, especially when the original video is not available to the user or receiver. Due to the asymmetric quality of their encryption and the degradation caused by packet loss and network delays, the application of these quality measures to many real-world situations is limited. Our proposed approach overcomes these challenges by accumulating perceptual 3D features in an NR-based manner applicable to video streaming and streaming platforms along with a reduced computational complexity to meet the real-time capacity of the decoder and measure the optimal 3D video quality.

3. Materials and Methods

3.1. Applied Datasets

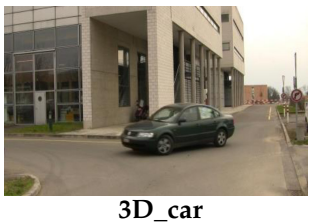
For our experiment, we used three different stereoscopic sequences (3D_02, 3D_car and 3D_03) from the RMIT3DV [37] and EPFL [38] video datasets summarized in Table 1. All had a 3-s playback duration, full HD resolution of 1920 × 1080 and 25 frames per second, and consisted of different pictorial contents, such as low or high contrast and object movements, and textures. The two sequences, 3D_02 and 3D_car, which were from the EPFL dataset, showed a person bicycle-riding on a road and a car taking a turn while moving, respectively. The other sequence, 3D_03 which was from the RMIT3DV dataset, was called Flag Waving in the state library which consists of a very fast motion. These videos consisted of different low to high level of 3D depth perception and various motions.

3.2. Proposed Method

The assessment of the quality of a distorted 3D video is an important part of building and organizing advanced immersible media delivery platforms. As shown in Figure 3, our proposed QA method is implemented by extracting the binocular and perceptual features of a stereoscopic video, which are combined afterward to generate a QA Objective Metric (QAOM). At first, perceptual attributes were taken into account to estimate the binocular features that influence the quality of a stereoscopic video. A new video quality index

based on two image/video features has been devised to evaluate the perceived quality of transmitted stereoscopic videos, with the extracted features accumulated according to the tube suppression theory. The disparity index was developed by taking into account similar aspects of stereoscopic film, as well as edge detection, which was used to assess binocular vision impairment and distortions due to packet loss in the network.

Table 1. Experimental video datasets.

Experimental S3D Videos	Name	Disparity	Video Characteristics
 **3D_02**	Bicycle Riding	Moderate	Low contrast, high object motion
 **3D_car**	Car Moving	High	Low contrast, low object motion
 **3D_03**	Flag Waving	High	High Contrast, random object motion

3.2.1. Dissimilarity Measure Based on Disparity Index

Firstly, stereo-matching with a block-based technique [39] is constructed to create a matrix of error energy for extracting disparity information from a pair of stereoscopic frames, based on which an intermediary disparity map [40] is generated from the two views. An averaging filter for eliminating unreliable disparity estimates is applied for enhancing the disparity map's reliability.

We denote $L_v(i,j,c)$ as the left-view RGB format, $R_v(i,j,c)$ the right-view RGB format and error energy in an RGB format. The error energy $E_{eng}(i,j,d)$ for a block size of $p \times q$, can be expressed as

$$E_{eng}(i,j,d) = \frac{1}{3pq} \sum_{x=i}^{i+p} \sum_{y=j}^{j+q} \sum_{c=1}^{3} (L_v(x,y+d,c) - R_v(x,y,c))^2 \quad (2)$$

Here, d is the disparity and c has value of $1,2,3$ which correspond to the RGB color space's red, green, and blue components, respectively. The error energy matrix $E_{eng}(i,j,d)$ is smoothed for the predefined difference search range S $(d = 1\ to\ S)$ by applying averaging filters multiple times and eliminating very abrupt energy changes that may correspond to incorrect matching, with its recurring use uncovering global trends in error energy. The

whole process makes this algorithm a region-based approach. The averaging filtering of the $E_{eng}(i,j,d)$ is described as follows for a $p \times q$ window size.

$$E_{eng}(i,j,d) = \frac{1}{3pq} \sum_{x=i}^{i+p} \sum_{y=j}^{j+q} E_{eng}(x,y,d) \tag{3}$$

Applying the averaging filter iteratively to the error energy matrix for each disparity, the disparity (D) with the minimum error energy of $E_{eng}(i,j,d)$ is selected as the best disparity estimation of pixel (i,j), of the disparity map as shown in Figure 4. The process can be summarized as follows: firstly, estimate the error energy matrix for every disparity D in the search range.

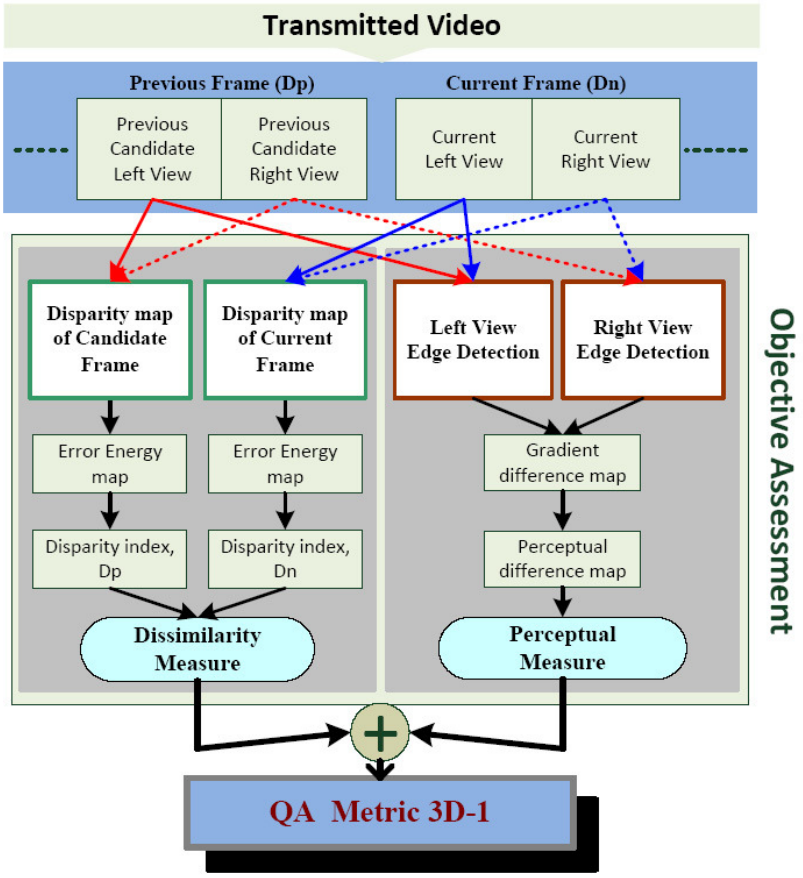


Figure 3. Block diagram of a QA metric for 3D video.

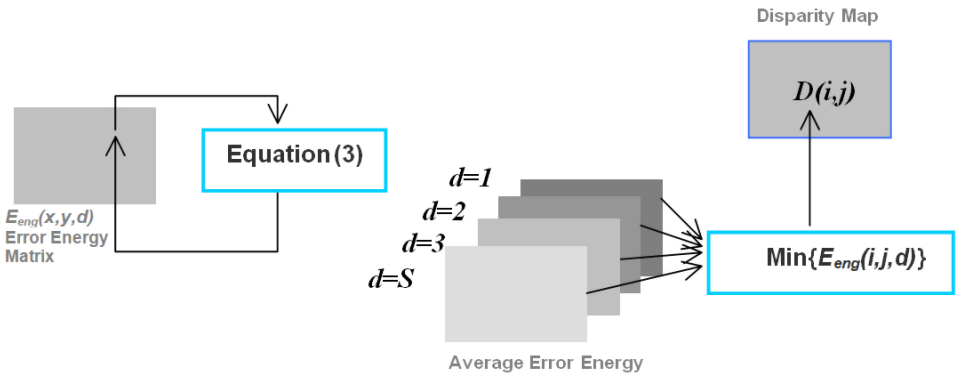


Figure 4. Estimation of the disparity map from the minimum of smoothed error energy.

The average error energy obtained from the stereoscopic images is utilized to remove unreliable disparity estimates from the disparity map $D(i,j)$, which is also calculated by the similar block-matching approach as shown in Equation (2). Apply the averaging filter iteratively to every error matrix for every disparity value, and finally, for each pixel (i,j), assign the minimum error energy $Min[E_{eng}(i,j,d)]$ to its disparity $D(i,j)$ in the disparity map.

These estimates are for points around object boundaries, resulting from object occlusion in images which may be recognized by examining the high error energy $D_{ne}(i,j)$ in $E_{eng}(i,j)$. The d is omitted from the rest of the equations since the average is taken from a disparity search range $S\,(d = 1\ to\ S)$. To increase the confidence of $D(i,j)$, a simple threshold is applied to filter out unreliable difference estimates and obtain a new disparity map.

$$\widetilde{D}(i,j) = \begin{cases} D(i,j) & , E_{eng}(i,j) \leq Th_{eng}(i,j) \\ D_{ne}(i,j) = 0 & , E_{eng}(i,j) > Th_{eng}(i,j) \end{cases} \tag{4}$$

where $D_{ne}(i,j)$ represents the unestimated high-energy value of block mismatch due to errors in comparing pairs of views or unreliable parallax estimation. Th_{eng} also determines whether the disparity estimate is reliable, which can be determined for the averaged filter tolerance coefficient (α), as shown in Figure 4, where the error energy threshold is given as

$$Th_{eng}(i,j) = \alpha \cdot E_{eng}(i,j) \tag{5}$$

where α is a tolerance factor to adjust the reliability of the filtering process; smaller values make $\widetilde{D}(i,j)$ more reliable. On the other hand, decreasing α erodes the disparity map by removing more disparity points in the map. Then, the disparity index can be calculated as the average of $\widetilde{D}(i,j)$ as

$$D_n = Mean[\widetilde{D}(i,j)] \tag{6}$$

A stereoscopic video maintains a nearly constant D_n ratio between left and right views. Considering the assumption that there are no scene changes in the video, this statement will be explained and analyzed in the experimental part. If there is any packet loss or distortion in either or both views, D_n will deteriorate, thereby demonstrating dissimilarity between the views which is measured as

$$S_m = \left| (\sum_{i=n-p}^{n-1} D_i) - (D_n \cdot P) \right| \times \frac{D_n}{10} \tag{7}$$

where D_n is the disparity index of the current frame and P the number of disparity indices of the previous candidate frames being compared with it to measure the dissimilarity (S_m). The D_n value is divided by 10 to get the final S_m score within the range from 0 to 1 and produce an equal ratio (weight factor) of S_m with the perceptual edge-based measure which equally contributes to the final QA score of $QAOM_3D$, as discussed in a later section.

3.2.2. Edge-Based Perceptual Difference Measure

Considering the human visual system, edges are the most essential perceptual elements. They record key events and changes in the features of an image, such as discontinuities in depth and surface orientations and fluctuations in scene illumination. The losses of edge magnitudes in visually relevant parts of an image are calculated in this subsection.

Because of its minimal computing complexity and excellent efficiency for recognizing edges, we employ a Sobel filter [41] to find edge regions of a given video frame A. As indicated in Equation (8), the operator constructs derivative approximations using two 3×3 kernels convoluted with the original image—one for horizontal and one for vertical changes. If we consider A to be the source picture, and Gx and Gy to be two images that comprise the horizontal and vertical derivative estimates at each position, the computations are as follows:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * A \ , \ G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \tag{8}$$

Here, $*$ denotes the convolution operation for the two-dimensional signal processing. We can combine the resulting gradient approximations to get the gradient magnitude for the current frame (G_c) as follows:

$$G_c = \sqrt{{G_x}^2 + {G_y}^2} \tag{9}$$

The resulting edge size map for the previous frame (G_P) of the video was determined similarly. After that, the edge strength difference (E_{diff}) between the current frame and the previous frame for the perceptual relevant region [42] in the image is computed, and the pixel position where the edge strength lies is defined as greater than the current frame's edge strength map average ($\overline{G_c}$). Furthermore, the threshold at which edge size differences are perceptible is taken as half the standard deviation of the actual edge size (σG_c) [43]. The mean and standard deviation of the current frame, both left and right, are calculated, and their mean is taken as the threshold. So, the distinct difference from the edge (E_c) at pixel location (i, j) is calculated as:

$$\Delta E_{diff}(i,j) = \begin{Bmatrix} \frac{\sum_{i=n-p}^{n-1}[G_c(i,j) - G_P(i,j)]}{P} & , G_c(i,j) > \bar{G}_c \\ 0 & , G_c(i,j) \leq \bar{G}_c \end{Bmatrix} \tag{10}$$

$$E_c(i,j) = \begin{Bmatrix} \Delta E_{diff}(i,j) & , \Delta E_{diff}(i,j) > \frac{\sigma G_c}{2} \\ 0 & , \Delta E_{diff}(i,j) \leq \frac{\sigma G_c}{2} \end{Bmatrix} \tag{11}$$

where P is the number of previous candidate frames relative the current frame to measure the difference. For a current video frame (c), the sum of the $E_c(i,j)$ is defined as the total difference (E_cD) for that frame, that is, the maximum perceived difference between the two views which is computed as

$$D_E = MAX(Left_{E_cD}, Right_{E_cD}) \tag{12}$$

3.2.3. Final Objective Quality Measure

The final QA is achieved by combining the two stereoscopic perceptual measures, dissimilarity and perceptual difference, to assess the property of the present stereoscopic 3D video due to the unavailability of the original sending video.

We assume that both the edge and disparity features have equal weights with respect to perceptual characteristics and dominance. Therefore, the distortion measures developed in Equations (7) and (12) are averaged and this average distortion is subtracted from 1, to obtain the overall quality of the stereoscopic video impairment metric $QAOM_{3D}$:

$$QAOM_{3D} = (1 - \frac{S_m + D_E}{2}) \tag{13}$$

4. Results

According to the proposed method, the dissimilarity between the left and right frames is measured by considering the inter-view similarity disparity information, which is a block-based stereoscopic matching approach. Based on a 3×3 window size for block matching and a predetermined disparity search range of $d = 1$ to 40, the reliable disparity, depth map, and error energy were calculated. Equations (2)–(5) show details of the steps for calculating the intermediate stages of the dissimilarity measure and Figure 5 shows some results for the Bicycle Riding dataset.

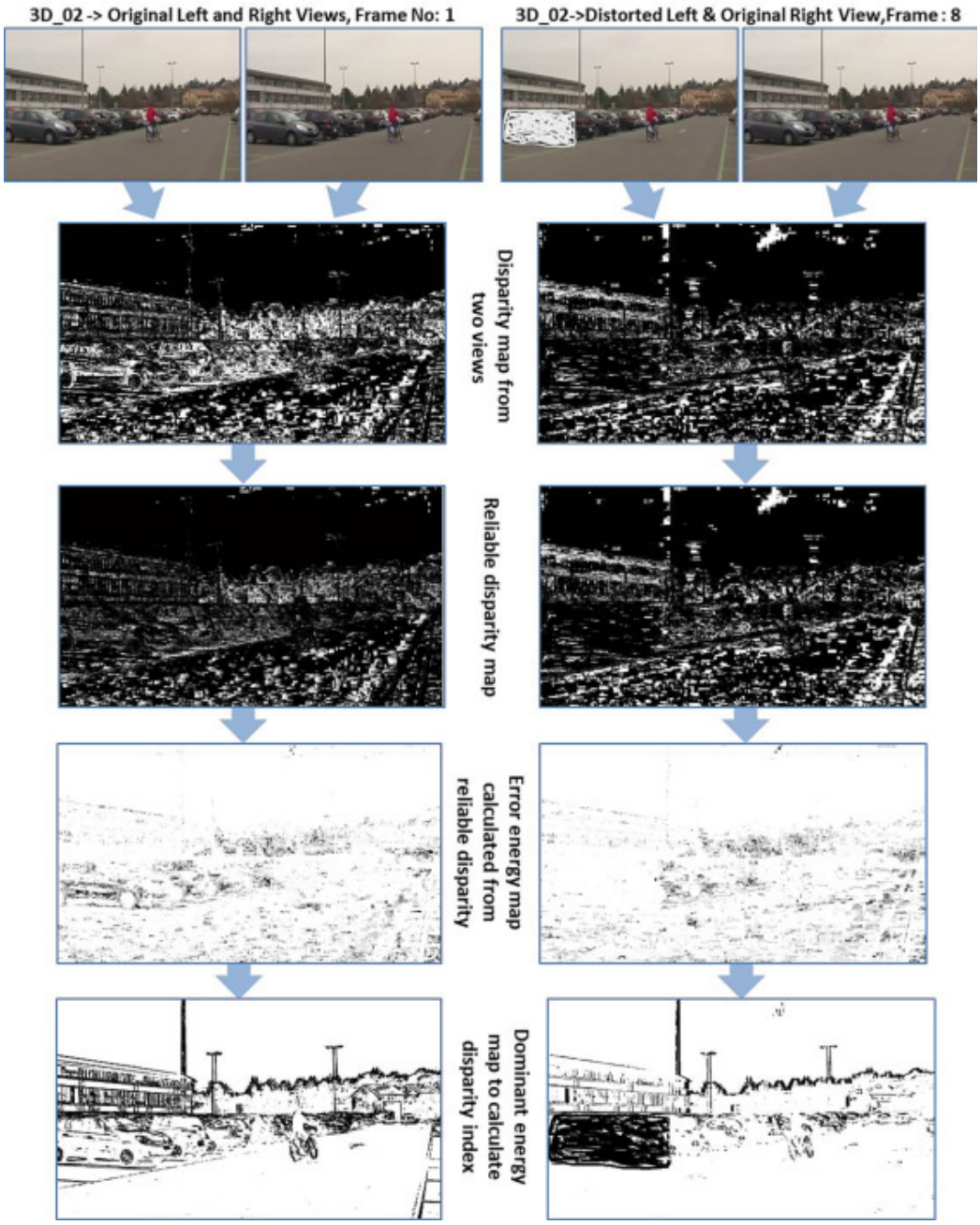


Figure 5. Steps for calculating disparity index from left and right frames of the Bicycle Riding video.

Using Equation (6), the disparity index (D_n) calculated from the reliable disparity measured between the left and right frames of the Bicycle Riding video was 4.155, a frame-by-frame comparison approach which continued for the entire video sequence. It was observed that the disparity indices remained relatively constant throughout a sequence assuming that there was no scene cut, with any error or distortion in either view reducing the value of D_n. To verify this observation, different kinds of errors, i.e., packet losses, distortions, and manual errors, were applied in one of the two views. Then, based on variations in the error energy, the disparity index was calculated from the left and right views. In Figure 5, the intermediate calculations and the corresponding maps are shown.

Using Equation (6), the disparity index (D_n) is calculated again from the reliable disparity between the distorted left and original right views of the Bicycle Riding video which was 3.497 and was less than the actual disparity index of 4.15 for two undistorted frames. Further examples of simulated packet losses and distortions are shown in Figures 6 and 7.

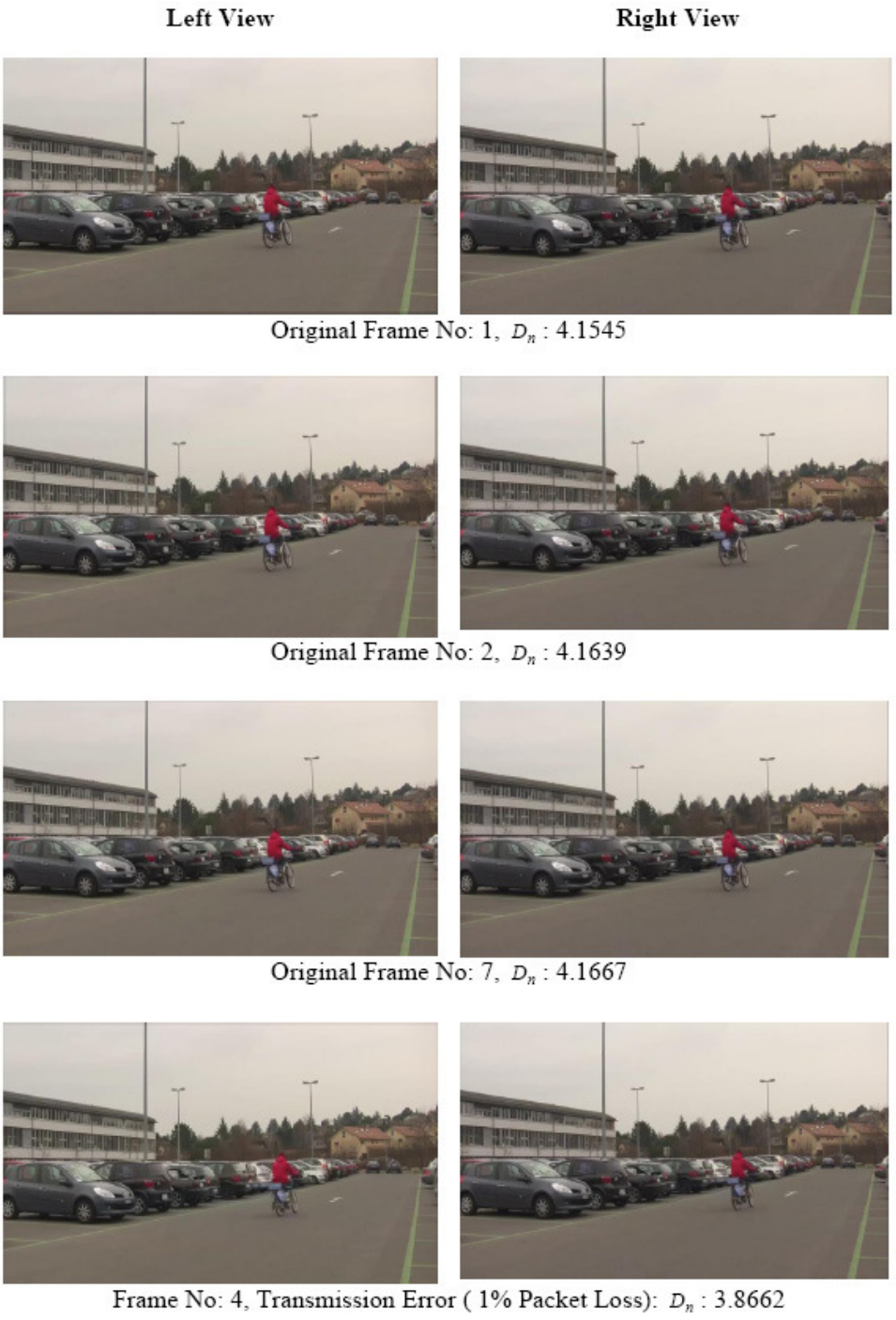


Figure 6. Different error scenarios induced in left views of 3D Video (part-1).

In Figures 6 and 7, different error mechanisms and artifacts induced in the frames can be seen, with the resultant disparity indices varying depending on the types of errors. In our experiment, packet losses were simulated using the JM H.264 reference software and, for various artifacts, the disparity indices were less than for those in undistorted frames. To verify that the left and right frames in a 3D video maintain a constant rate of disparity for proper viewing, we simulated different errors in 75 frames of the Bicycle Riding video sequence, with a graph of the disparity indices presented in Figure 8.

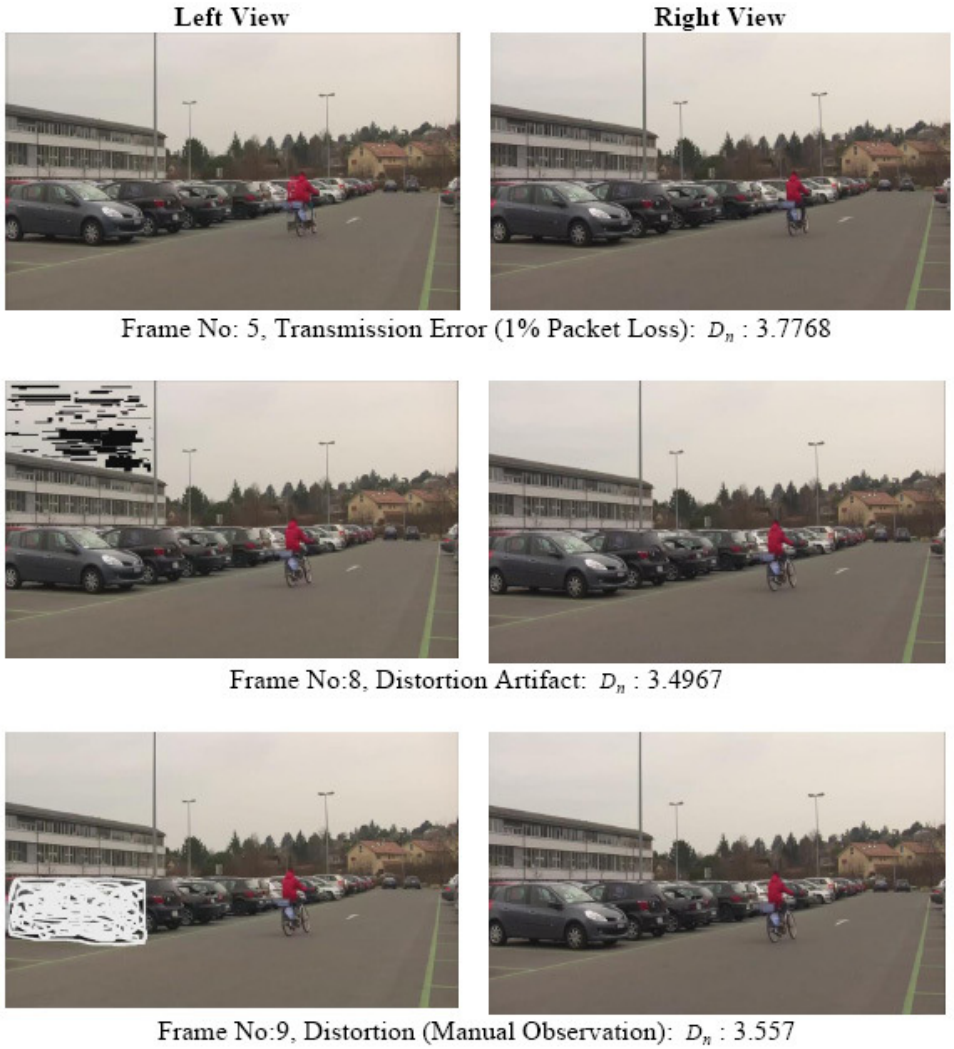


Figure 7. Different error scenarios induced in left views of 3D Video (part-2).

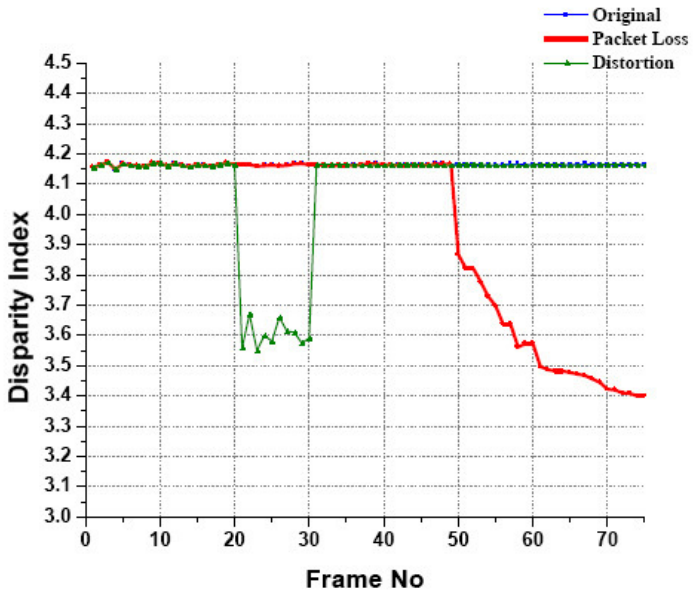
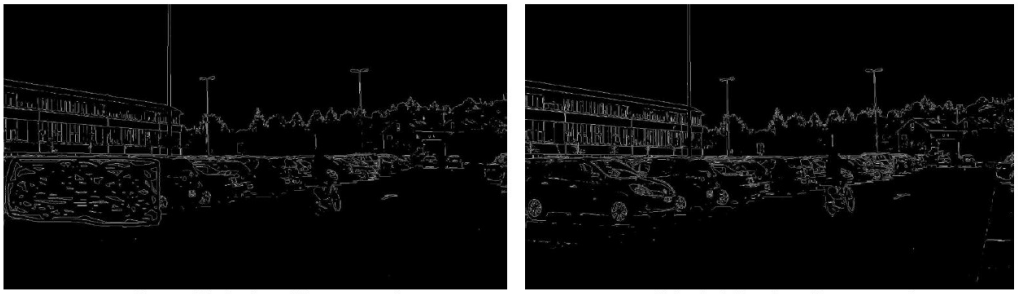


Figure 8. Disparity indices with respect to frames for the Bicycle Riding 3D video.

For a packet loss, due to the dissimilarity, the views result in a deteriorated disparity (in frames 50 to 75). Further, for each distorted artifact (in frames 20 to 30), the value of the index decreased and then increased back to its previous value for the original undistorted frame in both views. On that occasion, S_m of the current frame is measured by collating it with the previous frame indices of disparity using Equation (7). Different numbers of previous frames could be selected as reference ones by selecting different values of P.

For the second part of our experiment, an edge was selected as the most significant perceptual feature for consideration and Sobel edge detection as the most prominent and widely used mask for detecting edges in an image. In our proposed method, the edges of the current frame were compared with those of the previous one to determine the distortion between them. The edges of a frame detected using Equations (8) and (9) are shown in Figure 9, in which those of one distorted and one original frame from Figure 5 are presented.



Distorted left view (frame no. 8) Original right view (frame no. 8)

Figure 9. Sobel edge detection in a stereoscopic frame.

Each edge-detected frame was compared with the previous frame in the same view to observe edge distortions. From our experimental observations, it was noted that, for undistorted videos, the perceptual edge difference was small. However, if any prominent error-induced distortions were present in a frame, this measure demonstrated a significant difference compared with the previous one and may have surpassed a predefined threshold in terms of edge accumulation. According to Equations (10) and (11), the threshold at which we can observe the magnitude of the edge differences was half the standard deviation of the current edge magnitude (σG_c). Based on Equation (12), the maximum edge difference considered was that between the separate left and right frames.

It was obvious that the motions in a video affected the accumulation of edge differences, as shown by the differences between the Sobel edge-detected original right views in frames 4 and 5 in Figure 10 in which the edges indicate relative motions between the frames.

Figure 10. Difference in Sobel edge accumulation due to motion.

5. Discussion

According to our observations, for temporal information, this edge accumulation was low even for high-motion videos. The perceptual edge difference D_E was calculated using Equation (12), with the results obtained from the original and distorted frames of the Bicycle Riding stereoscopic video shown in Table 2. For low-motion videos, the D_E values were usually in the range from 0.05 to 0.10 and, for high-motion ones between 0.10 and 0.20, as can be observed for the undistorted frames 1, 2, and 3 while those for frames 6 and 7 increased due to packet losses occurring in the previous frame. However, for error or distorted frames, edge difference accumulations could be as high as 1, depending on their percentages of transmission losses and distortions.

Table 2. Experimental analysis of the Bicycle Riding video.

Org. Video Sequence	Disp. Index, D_n	Ed. Diff., D_E	Imp. Video Sequence	Disp. Index, D_n	Ed. Diff., D_E
Original 1	4.155	-	Original 1	4.155	-
Original 2	4.164	0.1097	Original 2	4.164	0.1097
Original 3	4.172	0.1099	Original 3	4.172	0.1099
Original 4	4.149	0.1120	**1% Packet loss 4**	**3.866**	**0.3680**
Original 5	4.166	0.1085	**1% Packet loss 5**	**3.776**	**0.3832**
Original 6	4.162	0.1125	Original 6	4.162	0.0894
Original 7	4.160	0.1195	Original 7	4.160	0.1340
Original 8	4.158	0.1212	**Distorted 8**	**3.497**	**0.4167**
Original 9	4.170	0.1238	**Distorted 9**	**3.557**	**0.5743**
Original 10	4.169	0.1195	**Distorted & 1% Packet loss 10**	**3.225**	**0.6608**

Org. = Original, Imp. = Impaired, Disp. = Disparity, Ed. = Edge, Diff. = Difference.

In Table 2, it can be seen that when distortion or packet loss (P. loss) is added, the disparity index ratio of left and right views (D_n) significantly decreased, a score also applied to measure the dissimilarity (S_m) of the video. Eventually, in Equation (13), the dissimilarity and perceptual variance metrics were merged to construct the QA metric of a 3D video. To demonstrate the effectiveness of our method, we examined two video datasets. To speed up the process, only the most recent received frame was utilized to calculate the edge difference perceptual measure. Different ratings were obtained for frames 2 to 10, as shown in Table 3, with distorted or degraded frames marked as **D** with the frame number. Finally, a video's $\widetilde{Q}\widetilde{A}OM_{3D}$ score can be represented as the overall mean of the $QAOM_{3D}$ scores obtained from each frame. In our approach, the concluding quality scores of the video trials contain 75 frames for different loss settings obtained from our proposed QAOM and two popular algorithms are shown in Table 4.

StSDLC [42] is a metric which calculates impairments in the range from 0 (low distortion) to 1 (high distortion). In our experiment, for our proposed method to be comparable with the others, we presented its values in reverse, i.e., as 1 to 0. We used the H.264 JM encoder to create several forms of impairments, using 26 and 32 quantization parameter settings to reduce the overall quality of a video and 1% and 3% packet losses simulated by the encoder and, to create distortions, different kinds of compression artifacts and manual degradation. We found that our approach performed equally well as the full-reference-based ones and, most importantly, did not require an original video or image. Therefore, it is very significant for quantifying transmitted or broadcast videos for which the original video is not available at the receiver end.

Table 3. $QAOM_{3D}$ scores for original and distorted videos from Bicycle Riding, Car Moving, and Flag Waving datasets.

Bicycle Riding				**Car Moving**				**Flag Waving**			
Original Video		***Distorted Video***		***Original Video***		***Distorted Video***		***Original Video***		***Distorted Video***	
Frm No	***QAOM* Score**	**Frm No**	***QAOM* Score**	**Frm No**	***QAOM* Score**	**Frm No**	***QAOM* Score**	**Frm No**	***QAOM* Score**	**Frm No**	***QAOM* Score**
2	0.9436	2	0.9436	2	0.9562	2	0.9562	2	0.9028	2	0.9028
3	0.9433	3	0.9433	3	0.9621	3	0.9621	3	0.9136	3	0.9136
4	0.9322	**Pl 4**	**0.7568**	4	0.9498	4	0.9498	4	0.8877	4	0.8877
5	0.9426	**Pl 5**	**0.7934**	5	0.9478	5	0.9478	5	0.8762	**Pl & Ep5**	0.7746
6	0.9439	6	0.8895	6	0.9512	6	0.9512	6	0.8821	**Pl & Ep6**	0.7438
7	0.9400	7	0.9326	7	0.9388	**D 7**	**0.7532**	7	0.8946	**Pl & Ep7**	0.7511
8	0.9389	**D 8**	**0.6679**	8	0.9522	**D 8**	**0.7725**	8	0.9033	**Pl & Ep8**	0.7647
9	0.9395	**D 9**	**0.7010**	9	0.9552	**D 9**	**0.7214**	9	0.9086	**Pl & Ep9**	0.7781
10	0.9388	**Pl & D10**	**0.6163**	10	0.9439	**D 10**	**0.6825**	10	0.9055	**Pl & Ep10**	0.7850

Frm No = Frame No, Pl = Packet Loss, D = Distorted, Ep = Error Propagation.

Table 4. Comparison of quality scores from the proposed and different prominent approaches.

Experimental Methods	**Reference Criterion**	**Impairment Parameter**	**Overall Video Quality Score**		
			Bicycle Riding	**Car Moving**	**Flag Waving**
SSIM [19]	**Full Reference**	**QP 26**	**0.9875**	**0.9763**	**0.9826**
		QP 32	0.9545	0.9586	0.9745
		Packet Loss (1%)	0.9437	0.9325	0.9556
		Packet Loss (3%)	0.9385	0.9086	0.9305
		Noise & Distortion	0.8877	0.8221	0.8936
StSDlc [42]	**Full Reference**	**QP 26**	0.9568	0.9482	0.9536
		QP 32	0.9397	0.9222	**0.9332**
		Packet Loss (1%)	**0.7950**	0.8045	0.8536
		Packet Loss (3%)	0.7859	0.7883	0.8319
		Noise & Distortion	0.7134	0.7725	0.7943
BLIIND [34]	**No Reference**	**QP 26**	0.9624	0.9611	0.9528
		QP 32	0.9555	0.9589	0.9423
		Packet Loss (1%)	0.8822	**0.8524**	**0.8779**
		Packet Loss (3%)	**0.7029**	0.7523	0.7884
		Noise & Distortion	0.7428	0.7325	0.7621
Proposed $\widetilde{Q}\widetilde{A}OM_{3D}$	**No Reference**	**QP 26**	0.9528	0.9598	0.9325
		QP 32	**0.9438**	**0.9385**	0.9004
		Packet Loss (1%)	0.7765	0.8026	0.7881
		Packet Loss (3%)	0.6782	**0.7245**	**0.7011**
		Noise & Distortion	**0.6164**	**0.6523**	**0.6286**

In addition, we compare our approach to the approach called BLIIND [34], where the author applied an NR-based method to find the video quality. The method works well for estimating the videos with nominal noises where QP is 26 and 32. It shows moderate results even for 1% packet losses. However, for higher distortions and losses, the method fails to track the level of distortions since it is a non-distortion method that deliberately avoids structural errors in the temporal video sequences. However, the proposed method performs really better than the frequency domain-based high computational BLIIND method. It performs significantly well for higher distortion and noises in the video sequences. In addition, the performance for QP32 depicts the performance of the method for light level noises in the video. Although it, it does not outperform all the approaches, the proposed

method works adequately well in terms of lower computation time. The method considers human visual sensitive features without the presence of a reference frame which is very much required for transmitted videos.

This research was conducted to implement a low-computational QA algorithm which could be used in real-time stereoscopic video transmission. To determine the influence of the binocular artifact in transmission, a series of experiments was performed, with the stereoscopic video streams individually encoded. To simulate real packet losses, the JM reference software was used. Finally, we have designed an objective no-reference quality measurement approach based on the HVS features of stereoscopic videos which combines the dissimilarity measure obtained based on the disparity index between the left and right views, and a perceptual difference measure calculated by considering the difference in the edge magnitude between temporal frames.

6. Conclusions

This study aims of incorporating binocular rivalry measure in QA which will lead us to design a robust error-resilient 3D video communication free from perceptual ambiguity and binocular rivalry with the focus on developing error concealment strategies that are applicable in real-time and various categories of 3D displays which do not deteriorate the user's visual perceptual comfort. To avoid the detrimental effect of binocular ambiguity and visual discomfort, the transmission system could be designed to take into account these challenging psychovisual issues; for example, based on feedback regarding the measure of 3D video quality at the receiver end, parameters of the transmission system could be changed "on the fly" to mitigate distortions and generate 3D views with minimal distortions [44]. To mitigate the loss of quality, the transmission system can allocate extra resources to that view or enhance the level of protection to reduce error for that 3D video channel. This type of distortion measure or impairment meter is useful for assuring error-resilient video communication and boosts the possibility of efficiently fusing 3D video content and improving overall user QoE.

Author Contributions: Conceptualization, M.M.H. and J.F.A.; methodology, M.M.H.; software, M.M.H.; validation, M.M.H., M.A.I. and S.R.; formal analysis, M.M.H.; investigation, M.M.H. and S.R.; resources, M.A.I.; data curation, M.M.H. and M.A.I.; writing—original draft preparation, M.M.H.; writing—review and editing, S.R., M.R.F. and J.F.A.; visualization, S.R.; supervision, M.R.F. and J.F.A.; project administration, J.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the University of Dhaka.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express their gratitude to the University of Dhaka for funding the project and Professor John F. Arnold for continuously providing research guidance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, L.; Chen, X.; Pan, Z.; Fan, K.; Li, F.; Lei, J. No-reference stereoscopic image quality assessment based on global and local content characteristics. *Neurocomputing* **2021**, *424*, 132–142. [CrossRef]
2. Varga, D. No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion. *Sensors* **2022**, *22*, 2209. [CrossRef] [PubMed]
3. Imani, H.; Zaim, S.; Islam, M.B.; Junayed, M.S. Stereoscopic Video Quality Assessment Using Modified Parallax Attention Module. In *Digitizing Production Systems*; Springer: Cham, Switzerland, 2022; pp. 39–50.
4. Jin, C.; Peng, Z.; Zou, W.; Chen, F.; Jiang, G.; Yu, M. No-Reference Quality Assessment for 3D Synthesized Images Based on Visual-Entropy-Guided Multi-Layer Features Analysis. *Entropy* **2021**, *23*, 770. [CrossRef]
5. Hewage, C.T.; Martini, M.G. Quality of experience for 3D video streaming. *IEEE Commun. Mag.* **2013**, *51*, 101–107. [CrossRef]

6. Biswas, M.; Frater, M.R.; Arnold, J.F.; Pickering, M.R. Improved resilience for video over packet loss networks with MDC and optimized packetization. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *19*, 1556–1560. [CrossRef]
7. Lambooij, M.; Fortuin, M.; Heynderickx, I.; IJsselsteijn, W. Visual discomfort and visual fatigue of stereoscopic displays: A review. *J. Imaging Sci. Technol.* **2009**, *53*, 30201-1. [CrossRef]
8. Wang, K.; Barkowsky, M.; Brunnström, K.; Sjöström, M.; Cousseau, R.; Le Callet, P. Perceived 3D TV transmission quality assessment: Multi-laboratory results using absolute category rating on quality of experience scale. *IEEE Trans. Broadcast.* **2012**, *58*, 544–557. [CrossRef]
9. Carreira, J.; Pinto, L.; Rodrigues, N.; Faria, S.; Assuncao, P. Subjective assessment of frame loss concealment methods in 3D video. In Proceedings of the Picture Coding Symposium (PCS), Nagoya, Japan, 8–10 December 2010; pp. 182–185.
10. Barkowsky, M.; Wang, K.; Cousseau, R.; Brunnström, K.; Olsson, R.; Le Callet, P. Subjective quality assessment of error concealment strategies for 3DTV in the presence of asymmetric transmission errors. In Proceedings of the 2010 18th International Packet Video Workshop (PV), Hong Kong, China, 13–14 December 2010; pp. 193–200.
11. Zhao, Y.; Zhang, Y.; Yu, L. Subjective Study of Binocular Rivalry in Stereoscopic Images with Transmission and Compression Artifacts. In Proceedings of the 2013 IEEE International Conference on Image Processing (ICIP), Melbourne, VIC, Australia, 15–18 September 2013; pp. 132–135.
12. Bensalma, R.; Larabi, M.C. A perceptual metric for stereoscopic image quality assessment based on the binocular energy. *Multidimens. Syst. Signal Process.* **2013**, *24*, 281–316. [CrossRef]
13. Alais, D.; Blake, R.; Blake, R. *Binocular Rivalry*; Bradford book; MIT Press: Cambridge, MA, USA, 2005.
14. Liu, Y.; Huang, B.; Yu, H.; Zheng, Z. No-reference stereoscopic image quality evaluator based on human visual characteristics and relative gradient orientation. *J. Vis. Commun. Image Represent.* **2021**, *81*, 103354. [CrossRef]
15. Zhang, P.; Jamison, K.; Engel, S.; He, B.; He, S. Binocular rivalry requires visual attention. *Neuron* **2011**, *71*, 362–369. [CrossRef]
16. Howard, I.P.; Rogers, B.J. *Binocular Vision and Stereopsis*; Oxford University Press: Oxford, UK, 1995.
17. Ogle, K.N. Some aspects of stereoscopic depth perception. *JOSA* **1967**, *57*, 1073–1081. [CrossRef] [PubMed]
18. You, J.; Jiang, G.; Xing, L.; Perkis, A. Quality of visual experience for 3D presentation-stereoscopic image. In *High-Quality Visual Experience*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 51–77.
19. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
20. Campisi, P.; Le Callet, P.; Marini, E. Stereoscopic images quality assessment. In Proceedings of the 2007 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007; pp. 2110–2114.
21. Yasakethu, S.; Hewage, C.T.; Fernando, W.A.C.; Kondoz, A.M. Quality analysis for 3D video using 2D video quality models. *IEEE Trans. Consum. Electron.* **2008**, *54*, 1969–1976. [CrossRef]
22. Ryu, S.; Kim, D.H.; Sohn, K. Stereoscopic image quality metric based on binocular perception model. In Proceedings of the 2012 19th IEEE International Conference on Image Processing (ICIP), Orlando, FL, USA, 30 September–3 October 2012; pp. 609–612.
23. Meegan, D.V.; Stelmach, L.B.; Tam, W.J. Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery. *J. Exp. Psychol. Appl.* **2001**, *7*, 143. [CrossRef]
24. Benoit, A.; Le Callet, P.; Campisi, P.; Cousseau, R. Quality assessment of stereoscopic images. *EURASIP J. Image Video Process.* **2008**, *2008*, 659024. [CrossRef]
25. Su, C.C.; Bovik, A.C.; Cormack, L.K. Natural scene statistics of color and range. In Proceedings of the 2011 18th IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, 11–14 September 2011; pp. 257–260.
26. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
27. You, J.; Xing, L.; Perkis, A.; Wang, X. Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis. In Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA, 13–15 January 2010.
28. Hewage, C.T.; Martini, M.G. Reduced-reference quality metric for 3D depth map transmission. In Proceedings of the 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), Tampere, Finland, 7–9 June 2010; pp. 1–4.
29. Akhter, R.; Sazzad, Z.P.; Horita, Y.; Baltes, J. No-reference stereoscopic image quality assessment. In Proceedings of the IS&T/SPIE Electronic Imaging, San Jose, CA, USA, 17–21 January 2010; International Society for Optics and Photonics: Bellingham, WA, USA, 2010; p. 75240T.
30. Wang, X.; Kwong, S.; Zhang, Y. Considering binocular spatial sensitivity in stereoscopic image quality assessment. In Proceedings of the 2011 IEEE Visual Communications and Image Processing (VCIP), Tainan, Taiwan, 6–9 November 2011; pp. 1–4.
31. Varga, D. No-Reference Video Quality Assessment Based on Benford's Law and Perceptual Features. *Electronics* **2021**, *10*, 2768. [CrossRef]
32. Dendi, S.V.R.; Channappayya, S.S. No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Trans. Image Process.* **2020**, *29*, 5612–5624. [CrossRef]

33. Ebenezer, J.P.; Shang, Z.; Wu, Y.; Wei, H.; Bovik, A.C. No-reference video quality assessment using space-time chips. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–6.
34. Saad, M.A.; Bovik, A.C.; Charrier, C. Blind Prediction of Natural Video Quality. *IEEE Trans. Image Process.* **2014**, *23*, 1352–1365. [CrossRef]
35. Zhang, H.; Zhang, Y.; Zhu, L.; Lin, W. Deep Learning-based Perceptual Video Quality Enhancement for 3D Synthesized View. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5080–5094. [CrossRef]
36. Feng, Y.; Li, S.; Chang, Y. Multi-scale feature-guided stereoscopic video quality assessment based on 3D convolutional neural network. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2095–2099.
37. Cheng, E.; Burton, P.; Burton, J.; Joseski, A.; Burnett, I. RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database. In Proceedings of the 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX), Melbourne, VIC, Australia, 5–7 July 2012; pp. 212–217.
38. Goldmann, L.; De Simone, F.; Ebrahimi, T. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In Proceedings of the IS&T/SPIE Electronic Imaging, San Jose, CA, USA, 17–21 January 2010; International Society for Optics and Photonics: Bellingham, WA, USA, 2010; p. 75260S.
39. Alagoz, B.B. Obtaining depth maps from color images by region based stereo matching algorithms. *arXiv* **2008**, arXiv:0812.1340.
40. Hasan, M.M.; Arnold, J.F.; Frater, M.R. No-reference quality assessment of 3D videos based on human visual perception. In Proceedings of the 2014 International Conference on 3D Imaging (IC3D), Liege, Belgium, 9–10 December 2014; pp. 1–6. [CrossRef]
41. Hasan, M.M.; Ahn, K.; Haque, M.S.; Chae, O. Blocking artifact detection by analyzing the distortions of local properties in images. In Proceedings of the 2011 14th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2011; pp. 475–480.
42. De Silva, V.; Arachchi, H.K.; Ekmekcioglu, E.; Kondoz, A. Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video. *IEEE Trans. Image Process.* **2013**, *22*, 3392–3404. [CrossRef] [PubMed]
43. Seo, J.; Liu, X.; Kim, D.; Sohn, K. An objective video quality metric for compressed stereoscopic video. *Circuits Syst. Signal Process.* **2012**, *31*, 1089–1107. [CrossRef]
44. Zhang, L.; Peng, Q.; Wang, Q.H.; Wu, X. Stereoscopic perceptual video coding based on just-noticeable-distortion profile. *IEEE Trans. Broadcast.* **2011**, *57*, 572–581. [CrossRef]