Feature Based No-Reference Continuous Video Quality Prediction Model for Coded Stereo Video

Z. M. Parvez Sazzad, Rafik Bensalma, and Mohamed Chaker Larabi
XLIM-SIC Lab, Dept. of Signal and Communication, University of Poitiers, France
E-mail: sazzad@univdhaka.edu, rafik.bensalma@univ-poitiers.fr, chaker.larabi@univ-poitiers.fr

Abstract

In this paper, we propose a continuous no-reference video quality evaluation model for MPEG-2 MP@ML coded stereoscopic video based on spatial, temporal, and disparity features with the incorporation of human visual system characteristics. We believe edge distortion is a major concern to perceive spatial distortion throughout any image frame which is strongly dependent on smooth and non-smooth areas of the frame. We also claim that perceived depth of any image/ video is mainly dependent on central objects/ structures of the image/ video contents. Thus, visibility of depth is firmly dependent on the objects' distance such as near, far, and very far. Subsequently, temporal perception is mostly based on jerkiness of video and it is dependent on motion as well as scene content of the video. Therefore, segmented local features such as smooth and non-smooth area based edge distortion, and the objects' distance based depth measures are evaluated in this method. Subsequently, video jerkiness is estimated based on segmented temporal information. Different weighting factors are then applied for the different edge distortion and depth features to measure the overall features of a temporal segment. All features are calculated separately for each temporal segment in this method. Subjective stereo video database, which considered both symmetric and asymmetric coded videos, is used to verify the performance of the model. The result indicates that our proposed model has sufficient prediction performance.

Introduction

There is no doubt that all conventional 2D media are going to be replaced by immersive 3D media in near future to improve the quality of experience for all media applications from broadcasting [1] to more specialized applications such as robotic navigation [2], and medical treatments [3]. There are many alternative technologies for 3D video display and communication including holographic, volumetric and stereoscopic; stereoscopic video seems to be the most developed technology at the present [4]. Stereoscopic video consists of two videos (left and right views) captured by closely located (approximately the distance between two eyes) two video cameras. These views constitute a stereo pair and can be perceived as a virtual view (i.e., not an actual camera view) in 3D by human observers with the rendering of corresponding view points. Therefore, the codec used in 2D video material can still be applied independently on the left and right views of a stereo video pair to save valuable bandwidth and storage capacity, though MPEG Ad-Hoc group for 3D audio and video is working on a new standard for efficient multi-view video coding [5]. Although the technologies required for 3D video are emerging rapidly, the effect of these technologies as well as video compression on the perceptual quality of 3D viewing has not been thoroughly studied. Therefore, perceived quality of 3D video is always an important issue to evaluate the performance of all 3D imaging applications and subjective quality assessment is the most accurate method for it. However, it is time consuming and expensive. In addition, this kind of assessment is not suitable for real time monitoring applications. Therefore, objective evaluation is an ever increasing requirement to monitor perceptual video quality in real time. Consequently, no-reference (NR) quality evaluation is more important to monitor video quality at end user terminals where reference videos are not available.

Although, nowadays 3D media quality evaluation is getting more attention in video quality expert group (VQEG) community, a very few efforts has concentrated till now to develop 3D quality metric specifically for video. Most of these works start with 2D metrics and try to incorporate information about 3D. In [6], a compound full-reference (FR) stereo-video quality metric is proposed composed of two elements; stereoscopic quality, and monoscopic quality. Monoscopic quality evaluates perceived distortions caused by blur, noise, contrast change etc., and the measure is a comparison between initial stereo-frames assumed to have perfect quality, and stereo-frames resulting from some distorting processing. Stereoscopic quality assesses the perceived degradation of binocular depth cues only, and it measures the amount of binocular cues preserved between image pairs. The features of each frame are averaged to get the prediction value for a sequence. In [7], the selection of the rate allocation strategy between views is addressed for scalable multi-view video codec to obtain the best rate-distortion performance by using objective stereo video quality measure. The work is inspired by the hypothesis that humans perceive good quality stereoscopic / 3D video as long as one of the eyes sees a high quality view. It seems as 2D video quality prediction because they do not take into account any depth information to the metric which is one of the most important factors for 3D perception. In [8], a FR stereoscopic video quality assessment method is proposed based on a well known 2D video quality model, VQM. The evaluation considers color video and depth information (H.264 codec). The method uses both the objective color video quality measured using VQM, and the objective quality of the average of the rendered left (color image) and right (depth image) views measured using VQM. A similar analysis is performed on color plus depth map-based stereo video sequences using VQM metric in [9]. The applicability of conventional 2D video metrics such as PSNR, SSIM, and VQM to 3D video with different packet loss conditions is investigated on a small dataset both for the case of stereoscopic video and monoscopic video with depth information [10]. A depth map based 3D video quality metric is proposed in [11] for perceptual depth and visual fatigue. In order to estimate depth map, the algorithm considers three features, namely depth range, vertical misalignment, and temporal consistency. To obtain the depth map that consider the three features, feature based disparity vectors is estimated by using Scale invariant feature transform. The method is then integrated into a single value which indicates visual fatigue. In [12] the authors investigate the variation of the subjective quality of depth perception with the quantization level of the depth maps. In particular the paper focused on the color plus depth representation of 3D video. The Just Noticeable Difference in Depth values at various depth levels is experimentally derived. In [13], a continuous NR objective quality assessment model is proposed for MPEG-2 MP@ML coded stereoscopic videos based on spatio-temporal segmentation that use the perceptual differences of local features such as edge and non-edge. Spatial distortions and disparity measures of a stereoscopic pair frame are calculated based on aforementioned features. In [14], [15], two FR stereoscopic quality metrics are proposed based on 3D-DCT with take into account HVS properties, such as contrast sensitivity function (CSF) and luminance masking [16], [17]. In the metrics, 3D-DCT is used to analyse the perceptual similarity of blocks in stereo frames grouped using disparity correspondence and block-matching. The methods are inspired by the properties of binocular vision such as binocular fusion and binocular difference. Properties of the binocular vision suggest that the visual information is simultaneously processed in two different ways. In [18], a perceptual model for stereo video quality evaluation is proposed, which mainly consists of three steps: wavelet-based perceptual decomposition, contrast conversion and masking, pooling and quality mapping. The work tries to incorporate some human visual system properties such as the Contrast Sensitivity Function, Multi-channel and Masking to the algorithm. In [19], a NR metric is proposed to predict quality of experience for 3D videos/ images. The algorithm mainly tries to assess visual comfort associated with viewing stereo images and videos. The algorithmic measure is to extract statistical features from disparity and disparity gradient maps as well as indicators of spatial activity from images. In [20], [21], authors propose respectively both FR and NR objective video quality measure for DIBR-based stereoscopic 3D videos. The 3VQM metrics attempt to evaluate the elements of visual discomfort which is calculated by the approach of ideal depth estimation. The metrics have also been used to derive the ideal depth estimate in a no-reference scenario. The ideal depth estimate is then evaluated by the three distortion measures: Temporal error outliers, Spatial error outliers and Temporal inconsistency. The three indexes are combined to form the proposed video quality measures which are verified against subjective rating.

Human Visual System (HVS) modelling is very important to evaluate perceptual media quality objectively whatever the media in 2D or 3D. There are two types of HVS models: neurobiological model [22] and psychophysical vision model. Models based on neurobiology aim to estimate the actual low-level process in the eye and optical nerve. However, these are not suitable in real-world application, because of their complexity [23]. The psychophysical models are used to predict aspects of the human vision, which are relevant to picture quality, such as color perception, contrast sensitivity, temporal and pattern masking etc. In this work, we propose a features based no-reference stereo video quality assessment model both for symmetric and asymmetric coded video which is inspired by different aspects of HVS characteristics. The model consists of three features: spatial, temporal, and disparity. The metric uses perceptual difference of smooth and non-smooth areas to measure edge distortion as spatial feature, and consider importance of central objects/ structure and its distance to measure perceive depth as disparity feature. Finally, video jerkiness is estimated as temporal features with the incorporation of video motion and scene contents. Here, we limit our study to MPEG-2 MP@ML codec video with different bit rates. The subjective experiment results on stereo videos dataset are used to train and test the model. 3D video quality assessment is required to incorporate multidimensional perceptual factors: depth, 3D video impairments (i.e., mainly crosstalk), and visual comfort etc., the combine effect of these factors reflects overall 3D perceptual quality. The rest of the paper is organized as follows: A section describes details of our proposed model. Results with the subjective experiments are discussed in the next section and finally, the paper is concluded in the last section.

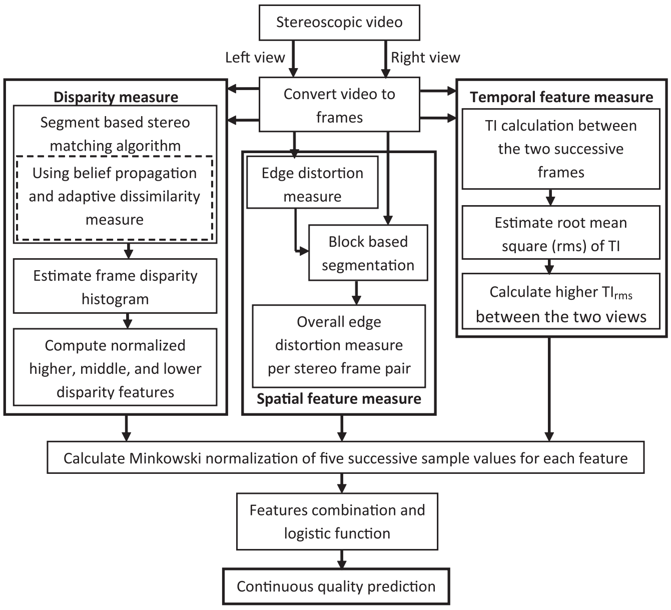
Stereoscopic video
Left view
Right view
Convert video to frames
Disparity measure
Segment based stereo matching algorithm
Using belief propagation and adaptive dissimilarity measure
Estimate frame disparity histogram
Compute normalized higher, middle, and lower disparity features
Edge distortion measure
Block based segmentation
Overall edge distortion measure per stereo frame pair
Spatial feature measure
Temporal feature measure
TI calculation between the two successive frames
Estimate root mean square (rms) of TI
Calculate higher TI_{rms} between the two views
Calculate Minkowski normalization of five successive sample values for each feature
Features combination and logistic function
Continuous quality prediction

Figure 1. Proposed NR quality evaluation model.

Proposed No-Reference Model

In this section, we provide a brief discussion about our proposed features based objective model. The computational model consists of three feature measures:

- Spatial feature: Edge distortion measure
- Disparity feature: Depth measure
- Temporal feature: Jerkiness measure

All features are calculated with the incorporation of different aspects of human visual system (HVS) characteristics. Subsequently, each feature is distinctly calculated for each temporal segment of length fifteen successive frames. Thus, we get two sets of mathematical features per second. Because, all reference video clips are in 30 fps as well as subjective sample was taken 2/sec. The block diagram of the proposed model is shown in Figure 1.

Spatial Feature: Edge Distortion Measure

Human visual system (HVS) is very sensitive to edge/ structural information in viewing field. Therefore, Human eye can easily perceive any degradation of edge information which reflects perceptual quality in spatial domain. Consequently, perceived distortions in spatial domain should be strongly dependent on smooth and non-smooth areas of the viewing field. For example, in theory, the visual distortions of an image increase with an increased rate of compression. However, the relationship between the distortions and the level of compressions is not always straight forward. It strongly depends on the texture contents of an image as well. Therefore, smooth and non-smooth areas based edge distortion measure is used for spatial feature estimation. Here, zero-crossing technique is employed for edge detection. In this section, we estimate edge distortion in spatial domain by using zero-crossing edge detector. The both views of stereo video sequence are converted into frame and consider selected frames only with luminance component to reduce computational cost. The frames were selected at regular intervals of time with two frames skipping. For example, the consecutive selected frames are 1, 4, 7,..... Firstly, we apply the block (8×8) based segmentation algorithm to the left and right frames individually to classify smooth, and non-smooth blocks in the frames [27]. Secondly, we calculate zero-crossing of each 8×8 block of the stereo frame pair separately for left and right frames. Thirdly, we average each value of zero-crossing independently for smooth, and non-smooth blocks of each frame of the stereo pair. And finally, total zero crossing for each stereo frame pair is estimated based on minimum zero-crossing value between the left and right frames distinctly for smooth, and non-smooth blocks. Here minimum zero-crossing is considered to take into account highest edge degradation between the two views. The measure of zero-crossing within each block of the frames are calculated horizontally and then vertically.

For zero-crossing in horizontal direction: Let the test frame signal is *x(m, n)* for m ∈ [1, M] and n ∈ [1, N], a differencing signal along each horizontal line is calculated by

$$d_h(m,n) = x(m,n+1) - x(m,n), \tag{1}$$

n ∈ [1, N-1] and m ∈ [1, M]

Here, zero-crossing is calculated by second order derivative with sign identification and multiplication

$$d_{h-sign}(m,n) = \begin{cases} 1 & \text{if } d_h(m,n) > 0 \\ -1 & \text{if } d_h(m,n) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$d_{h-mul}(m,n) = d_{h-sign}(m,n) \times d_{h-sign}(m,n+1) \tag{3}$$

We define for n ∈ [1, N-2]:

$$z_h(m,n) = \begin{cases} 1 & \text{if } d_{h-mul}(m,n) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where "1" and "0" are respectively indicated zero-crossing occurring and not, and also the size of $z_h(m,n)$ is $M \times (N-2)$. The horizontal zero-crossing of a block (8 × 8), ZC_{bh}, is estimated as follows:

$$ZC_{bh} = \sum_{i=1}^{8} \sum_{j=1}^{8} z_h(i,j) \tag{5}$$

Thus, we can calculate zero-crossing of each available block of the left and right frames.

Similarly, the vertical feature of zero-crossing (ZC_{bv}) of the block is calculated. Therefore, the overall zero-crossing feature, ZC_b, per block is given by:

$$ZC_b = \frac{ZC_{bh} + ZC_{bv}}{2} \tag{6}$$

Consequently, the average zero-crossing values for smooth, and non-smooth areas of the left frame are calculated by:

$$ZCl_p = \frac{1}{N_p} \sum_{b=1}^{N_p} ZC_{bp} \tag{7}$$

$$ZCl_n = \frac{1}{N_n} \sum_{b=1}^{N_n} ZC_{bn} \tag{8}$$

where N_p, and N_n are respectively the number of smooth, and non-smooth blocks of the frame. Similarly, the average zero-crossing values of ZCr_p, and ZCr_n for the right frame are calculated. In this text, the subscripts 'p', and 'n' are indicated smooth and non-smooth areas respectively.

We then calculate the total zero-crossing features of smooth, and non-smooth areas of the stereo frame. For the total zero-crossing features (ZC_p, and ZC_n), we estimate minimum values between the left and right frames by the following algorithm:

$$ZC_p\ (ZCl_p, ZCr_p) = min(ZCl_p, ZCr_p) \tag{9}$$

$$ZC_n\ (ZCl_n, ZCr_n) = min(ZCl_n, ZCr_n) \tag{10}$$

Let, $ZC_p(f)$ be the total zero-crossing for smooth areas of a stereo frame pair, f. Then we calculate the feature for each temporal segment by the following equation:

$$ZC_p(s) = \frac{1}{F} \sum_{f=1}^{F} ZC_p(f) \tag{11}$$

where $ZC_p(s)$ represents zero-crossing for a temporal segment, s for smooth areas. And "F" denotes number of frames in a temporal segment, s. Here, the value of F = 5. Subsequently, we determine the zero-crossing ($ZC_p(s)$) of the temporal segment for smooth areas by using available past four temporal segment values. Here, zero-crossing value of every temporal segment is estimated by Minkowski normalization with the consideration of the values of past four temporal segment and the current temporal segment to take into account of human recency effect. The equation is given by:

$$Z_p(s) = \Big(\frac{1}{5} \sum_{i=1}^{5} ZC_p(s-i)^{\gamma}\Big)^{\frac{1}{\gamma}} \tag{12}$$

where s = 1,2,3,...... represent successive temporal segment in temporal domain. The value of ZC is in the range [0,64]. Similarly, the zero-crossing feature, $Z_n(s)$, for non-smooth areas is computed. Here, the value of γ is considered as 0.5. Lastly, the overall zero-crossing, Z_s, for each temporal segment is calculated by

$$Z(s) \quad = \quad Z_p(s)^{w_1} \cdot Z_n(s)^{w_2} \tag{13}$$

where w_1, and w_2 are the weighting factors for the zero-crossing of smooth, and non-smooth areas.

Disparity Feature: Depth Measure

We claim that human eye is mostly influenced by central objects/ structural information of the image/ video contents for depth perception. Accordingly, visibility of the depth is highly dependent on the objects distances such as near, far, and very far. Pixel displacement between the left and right views is called disparity. A pixel's disparity is inversely to its depth in the stereo

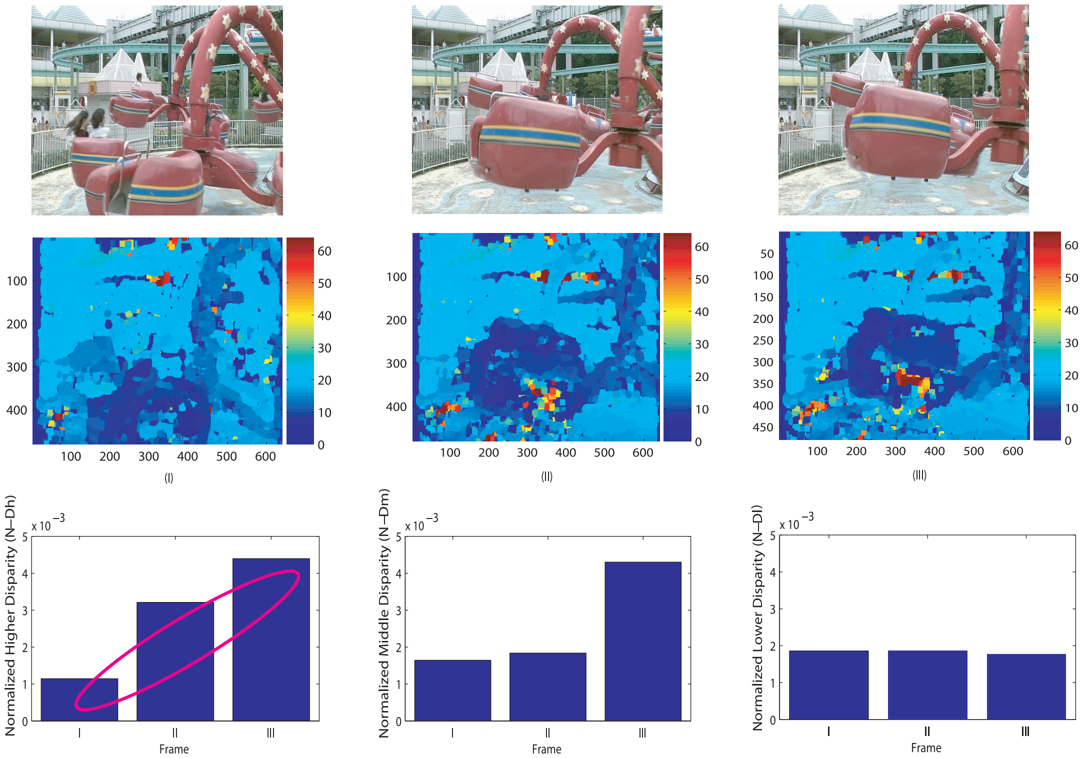

Figure 2. Normalized histogram of higher, middle, and lower disparity of the three different frames with depth maps.

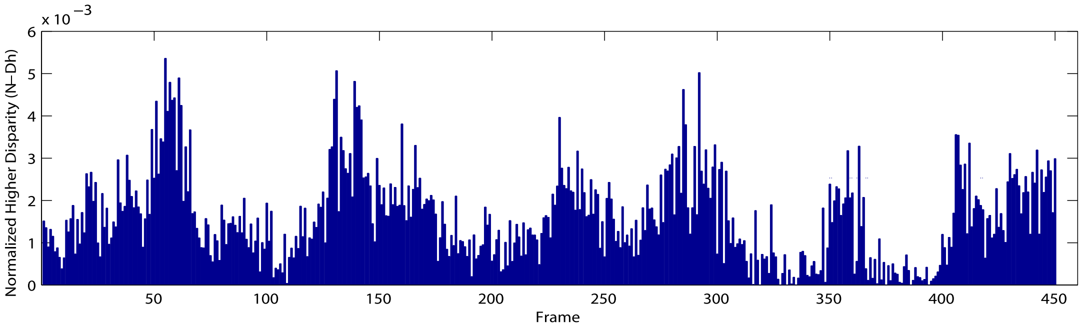

Figure 3. Normalized histogram of higher disparity for Chip-5 (Amusement park).

view. Therefore, pixels displacement are calculated by a segment based stereo matching algorithm [26] and evaluate disparity histogram. Subsequently, we estimate three normalized histogram values from lower, middle, and upper part of the histogram in order to take into account the objects relative distances. In order to verify the evidence, we analyse a stereo video clip of Amusement park (Clip-5), see Figure 5. In the video clip, a rotation merry-go-round from a steady camera perspective can be seen and depth of the central objects is changing with the rotation. To explain it more clearly, we consider three frames of the clip where central object is relatively very far, far, and near in frames 1st, 2nd, and 3rd respectively (see Figure 2). The figure also shows normalized higher, middle, and lower disparity values for the three frames. The bar graph in Figure 2 for normalized higher disparity indicates that 1st frame's disparity is lower compared to 2nd frame and also 2nd frame is lower than 3rd frame. Consequently, the analysis confirms that the object depth perception is higher in 3rd frame and eventually in 2nd, and 1st frame. Similarly, the normalized higher disparity histogram for all frames of the clip is shown in Figure 3. The bars in the figure follows a changing nature that also confirms the variation of depth perception of the clip. Thus, higher, middle, and lower normalized disparity measures are used in this model as disparity feature. Although, higher disparity feature is mainly approved the depth perception we also want to consider middle, and lower disparity features to take into consideration of relative depth of other objects of the scene.

Disparity histogram based depth measure is presented in this section. In order to measure disparity feature, we use a segment based stereo matching algorithm. The approach is conducted by using belief propagation and a self adapting dissimilarity measure. Details of the algorithm are discussed in [26]. Firstly, the stereo matching algorithm is applied to the consecutive selective frame pairs in a stereo sequence. Eventually, we get each pixel's disparity of the frame pairs. Secondly, we calculate histogram of the disparity frames. Thirdly, upper, middle, and lower parts of the histogram are considered and then normalized these values. Subsequently, these three normalized disparity features are considered to measure depth in our method. The depth maps of three sample stereo frame

pairs are shown in Figure 2. Colors in the depth maps that are indicated by vertical color bars in right are estimated depths of the frames pairs.
We consider

- Lower disparity: *h(0), h(1), and h(2)*
  where *h(0), h(1), and h(2)* indicate number of disparity pixels with pixel's displacement 0, 1, and 2 receptively.

- Middle disparity: *h*($\frac{d}{2}-1$)*, h*($\frac{d}{2}$)*, and h*($\frac{d}{2}$*+1)*

- Higher disparity: *h(d-2), h(d-1), and h(d)*
  where *d* is the maximum pixel disparity/ displacement.

For normalized disparity:

$$NDl(f) = \frac{h(0)\cdot(0+1)+h(1)\cdot(1+1)+h(2)\cdot(2+1)}{M\times N\times(d+1)} \tag{14}$$

$$\begin{aligned} NDm(f) \quad = \quad & \frac{h(\frac{d}{2}-1)\cdot((\frac{d}{2}-1)+1)+h(\frac{d}{2})\cdot((\frac{d}{2})+1)}{M\times N\times(d+1)} \\ & + \frac{h(\frac{d}{2}+1)\cdot((\frac{d}{2}+1)+1)}{M\times N\times(d+1)} \end{aligned} \tag{15}$$

$$\begin{aligned} NDh(f) = & \frac{h(d-2)\cdot((d-2)+1)+h(d-1)\cdot((d-1)+1)}{M\times N\times(d+1)} \\ & + \frac{h(d)\cdot(d+1)}{M\times N\times(d+1)} \end{aligned} \tag{16}$$

where *NDl(f), NDm(f), and NDh(f)* are respectively lower, middle, and higher disparity features of a stereo frame pair. Subsequently, the lower disparity feature, *NDl(s)* for a temporal segment, s is calculated by:

$$NDl(s) = \frac{1}{F}\sum_{f=1}^{F} NDl(f) \tag{17}$$

Similarly, *NDm(s), and NDh(s)* are estimated. Finally, the total disparity features of a temporal segment are calculated using Minkowski normalization with considering the past four disparity samples by the Equation 12. Let, *NDL(s), NDM(s), and NDH(s)* be the total Minkowski features of a temporal segment for lower, middle, and higher disparity respectively. Lastly, all three disparity features are combined by some weighting factors to estimate overall disparity feature of a temporal segment by the following equation.

$$ND(s) \quad = \quad NDL(s)^{w_3}\cdot NDM(s)^{w_4}\cdot NDH(s)^{w_5} \tag{18}$$

where w_3, w_4, and w_5 are the weighting factors.

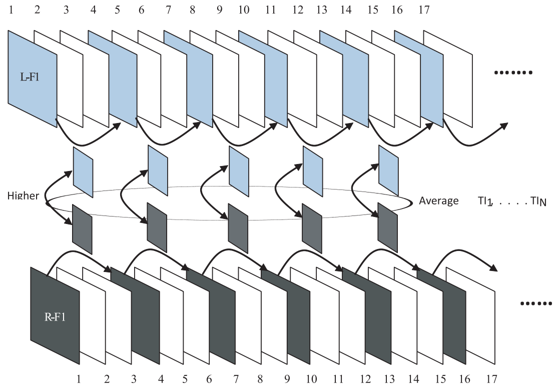

Figure 4. *Temporal feature extraction*

Temporal Feature: Jerkiness Measure

In order to compute temporal feature, we determine maximum jerkiness between the consecutive frames both for left and right views. Because jerkiness makes more annoying for human eye in temporal domain. Eventually, jerkiness of any stereo video is heavily dependent on motion and scene contents of the video sequence. Therefore, we estimate maximum jerkiness as temporal feature by considering highest motion and scene contents between the successive frames. To measure video jerkiness as a temporal feature, we use luminance intensity variation of pixels between the successive frames both in temporal and spatial domains. The frame selection criteria is the same which is used in others features extraction. The temporal feature extraction approach is shown in Figure 4. Firstly, absolute luminance difference (i.e., temporal information, TI) between the two successive selective frames are estimated separately for left and right views by the following: For left view:

$$TI_l(m,n,t) = |x_l(m,n,t+k) - x_l(m,n,t)| \tag{19}$$

where k = 3, and t = 1, 4, 7,..... are the selected frames numbers. Secondly, the deviation of the temporal information is calculated by:

$$TI_{d_l}(t) = \sqrt{\left(\overline{TI_l^2(m,n,t)} - \overline{TI_l(m,n,t)}^2\right)} \tag{20}$$

Thirdly, the root mean square, $TI_{rms_l}(m,n,t)$ is estimated by:

$$TI_{rms_l}(t) = \sqrt{\left(TI^2_{d_l}(t) + \overline{TI_l(m,n,t)}^2\right)} \tag{21}$$

Similarly, $TI_{rms_r}(t)$ is estimated for right view. Fourthly, the maximum temporal feature is computed between the two views using the following equation:

$$TI_{rms}(f) = max\Big(TI_{rms_l}(f), TI_{rms_r}(f)\Big) \tag{22}$$

Subsequently, the temporal feature, $TI_{rms}(s)$ for a temporal segment, s is calculated by:

$$TI_{rms}(s) = \frac{1}{F}\sum_{f=1}^{F} TI_{rms}(f) \tag{23}$$

Finally, the total temporal feature of a temporal segment are calculated using Minkowski normalization with considering the past four temporal samples by the Equation 12. Let, $MTI(s)$ be the Minkowski normalized temporal feature. Then, the temporal feature is updated by a weighting factor with the following equation:

$$TI(s) \quad = \quad MTI(s)^{w_6} \tag{24}$$

where w_6 is the weighting factor for adjusting temporal feature.

Features Combination

The following features' combination equation is considered to integrate the spatial, temporal, and disparity features in order to constitute a continuous stereo video quality prediction model.

$$S = \alpha(ND) + \beta Z + \gamma(TI) \tag{25}$$

where α, β, and γ are the method parameters. The proposed model performance is also studied without disparity by the following features combine equation:

$$S = \alpha + \beta Z + \gamma(TI) \tag{26}$$

A logistic function is used as the non-linearity property between the human perception and the physical features [28]. Finally, the obtained MOS prediction, MOS_p, per a temporal segment is derived by the following equation.

$$MOS_p = \frac{b_1}{1+exp[-b_2(S-b_3)]} + b_4 \qquad (27)$$

The model's parameters, weighting factors (w_1 to w_6), and the parameters of the logistic function are must be estimated by an optimization algorithm with the subjective test data. Here, Particle Swarm Optimization (PSO) algorithm is used for optimization [29].

Figure 5. *Left view of reference sequence (0 ∼ 75) sec*

Figure 6. *Left view of reference sequence (76 ∼ 150) sec*

Figure 7. *Left view of reference sequence (151 ∼ 225) sec*

Results

The proposed model's performance is evaluated by a subjective dataset. In the following sections, we provide details about the subjective experiments and the evaluation result of the proposed model.

Subjective Experiments

The Media Information and Communication Technology (MICT) lab conducted the subjective experiments on color stereo coded videos by using single stimulus continuous quality evaluation (SSCQE) method in which a processed video sequence was presented alone without being paired with its reference version [24]. The experiments considered fifteen stereo video clips of each 15 seconds length with 640 × 480 pixels, and 30 fps progressive format. All clips were combined together to create long sequences of 3 minutes 45 seconds. Left view (grey scale only) of a reference sequence is shown in Figures 5, 6, and 7. Video clips order was same in each sequence according to Figures 5, 6, and 7. Ten symmetric/asymmetric stereo video sequences were created by using MPEG-2 MP@ML encoder with four kinds of bit rates 2, 3, 5, and 8 Mbps. The selected bit rates combinations of left (L) and right (R) sequences are (L, R): (2, 2), (3, 2), (3, 3), (5, 2), (5, 3), (5, 5), (8, 2), (8, 3), (8, 5) and (8, 8) Mbps. Video clips order was same in every sequence. All reference clips were produced by NHK, Japan, and made available for research on stereo video. Each reference clip was 15 seconds length with 1920 × 1035 pixels, 30 fps of 24-bit/pixels RGB color space. An auto stereoscopic (SANYO) display was used in this experiment to display the stereoscopic video sequences and the subjects were instructed about the limited horizontal viewing angle to perceive 3D video correctly. Sixteen non-expert subjects (8 males and 8 females, with average age 23 years) with ages ranging from 20 to 32 participated in the experiment. Most of them were college/university student and also were non-experts in the area of video quality. All subjects were screened prior to participate the session for normal visual acuity with or without glasses, normal color vision, normal stereo depth perception and familiarity with the language. The subjective test conditions and parameters are summarized in Table 1. The subjects were asked to provide their overall perception of quality on a continuous quality scale marked with "Excellent", "Good", "Fair", "Poor", and "Bad". The subjective scores were quantized on a scale of [0...100], 0 being the worst quality and 100 being the best. The slider in the SSCQE test was not a stand-alone hardware device, but a graphical on-screen slider that was steered by moving the mouse up and down, i.e. vertical mouse movements were translated directly into slider shifts. Viewers familiarity with handling a computer mouse were an additional advantage. SSCQE judgements were given continuously at a sampling rate of 2/sec. In order to avoid any recency effects from the previous sequence pair, first clip (Clip-1, 15 sec) voting was rejected in each sequence. Therefore, total 420 samples were collected instead of 450 samples (3 minutes 45 seconds) for each sequence. Mean opinion scores (MOSs) were then computed for each stereo video sequence after post-experiment screening of the results according to ITU-R Rec. 500-10 [25]. Two outlier subjects were detected out of sixteen subjects. Discarding the outliers, the MOS had been computed for each sequence with the 95% confidential interval (CI). Here, the CI was estimated for each MOS value per 0.5 second. Out of 10 sequences, we consider six sequences, three symmetric ((5,5), (3,3), and (2,2) Mbps) and three asymmetric ((8,2), (5,3), and (3,2) Mbps) sequences in this work. Figures 8 and 9 show the continuous MOS scores of the six sequences.

Table 1: Subjective test conditions and parameters

Method	SSCQE
Samples	2/sec
Coder	MPEG-2 MP@ML
Bit Rates	4 kinds (2, 3, 5 and 8 Mbps)
Stereo video clips	15
Video resolution	(640×480) 24-bit/pixel, RGB
Each clip length	15 sec
Stereo sequences	10 (Each length 3 min 45 sec)
Subjects	16 (Non expert, students)
Display	10-inch, LCD 3D Auto stereoscopic
Display resolution	640×480 pixels (LR: 320 × 480)
Viewing distance	4H (H = Picture height)
Room illumination	Dark

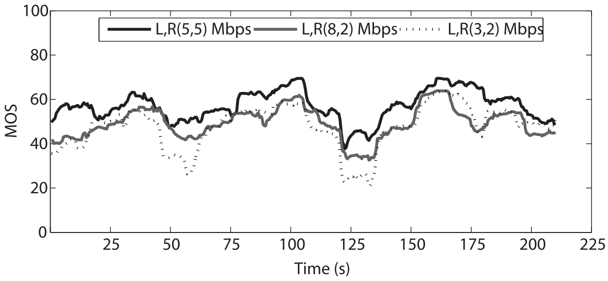

Figure 8. *Continuous MOS scores of three sequences (Seq-1, 2, and 3)*

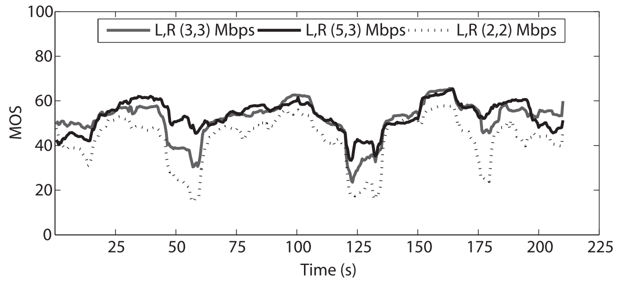

Figure 9. *Continuous MOS scores of three sequences (Seq-4, 5, and 6)*

Performance Evaluation

In order to verify the performance of our proposed model, we consider six stereo video sequences, three symmetric ((5,5), (3,3), and (2,2) Mbps) and three asymmetric ((8,2), (5,3), and (3,2) Mbps) and divide the sequences into two parts for training and testing. The training dataset consists of three Seq-1(5,5), Seq-2(8,2), and Seq-3(3,2) symmetric/asymmetric coded stereo video sequences. The testing dataset consists of the others three symmetric/asymmetric coded stereo sequences, Seq-4(3,3), Seq-5(5,3), and Seq-6(2,2), and also there is no overlapping between the training and testing. The parameters and weighting factors are obtained by the PSO algorithm with the training sequences are shown in Table 2.

Table 2: Model parameters and weighting factors

α = 6.865239	β = -37.1035	γ = 8.911039
w_1 = -0.007833	w_2 = 0.013315	w_3 = 0.006783
w_4 = 0.039649	w_5 = -0.018944	w_6 = 0.028219
b_1 = 58.04346	b_2 = -3.64833	b_3 = -21.3781
b_4 = 0.852088		

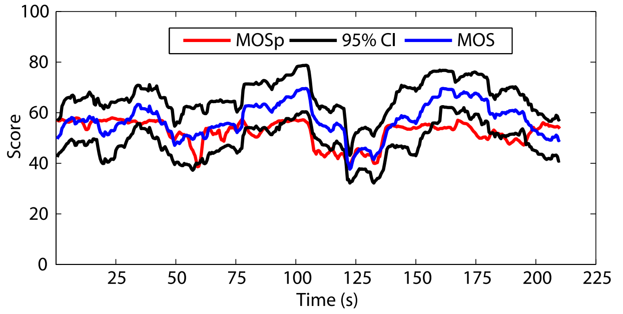

Figure 10. MOS_p *scores with 95%CI of Seq-1, symmetric: (L,R:5,5) Mbps*

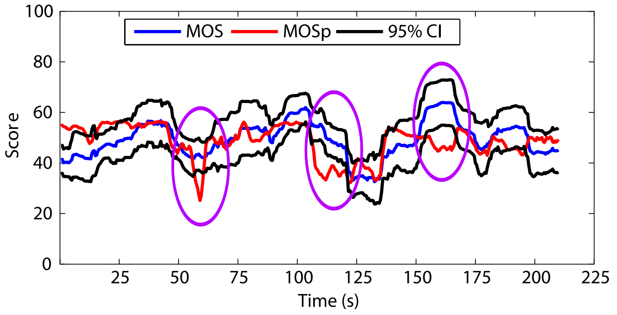

Figure 11. MOS_p *scores with 95%CI of Seq-2, asymmetric: (L,R:8,2) Mbps*

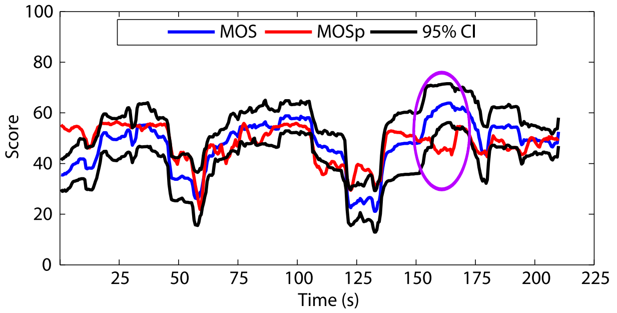

Figure 12. MOS_p *scores with 95%CI of Seq-3, asymmetric: (L,R:3,2) Mbps*

As our proposed model is designed for continuous quality prediction, the conventional image/ video quality evaluation criteria such as, Pearson linear correlation coefficient, Root mean square error, Spearman rank order correlation coefficient, and outlier ratio are not suitable for this evaluation. Because, the evaluation are calculated based on point to point samples between subjective and predicted scores. However, in SSCQE method, each human response time is individual. Therefore, it is quite hard to find any synchronization within subjects with respect to subjective scores and time. Moreover, our model's predicted samples are very particular to the temporal segments of a stereo video sequence. Therefore, we believe that the most important thing for continuous quality prediction is how continuously the prediction confined within the confidential interval. The continuous MOS prediction (MOSp) for every sequence with 95%CI and MOS are shown in Figures 10 to 15. Figures 10 to 15 indicate that the model's continuous prediction consistency is sufficient except some clips in sequences, seq-2(8,2), seq-3(3,2), and seq-6(2,2).

The three major miss prediction areas are marked by circles in the seq-2(8,2). The circles corresponding clips are Amusement park (clip-5), Festival with chromakey (clip-9), and Flower pot (clip-12). The first two clips (clip-5, and clip-9) are in high motion (i.e., video content changes of adjacent frames in the clips are very high) and camera work of these two videos are also high. Therefore, noise increases and decreases rapidly within a very short time in right view because of low encoding bit rate (2 Mbps). However, there is no significant variation of noise in Left view due to its high encoding bit rate (8 Mbps). Therefore, subject can not identify the low quality frames. Although, the low quality view suppresses perceptual quality the high quality view significantly restrains the perceptual quality. Whereas, our proposed model try to quantify the highest degradation between the two views and model can not follow the perceptual compromise significantly. However, the third miss prediction clip (clip-12) is very low motion and low content video with only two central objects (woman with flower pot). Therefore, noise variation is not high even in low bit rate right view. Moreover, the two objects in the clip are central objects and close to camera. Consequently, the low bit rate view could not suppress significantly the overall perceptual quality of the clip. On the other hand, our proposed model try to quantify the highest degradation between the two views with respect to spatial, temporal, and disparity features irrespective of any motion classification algorithm to classify the video clips into different motion group such as high, medium, and low, the model can not follow the variation of motion significantly to predict the quality. The same video clip of low motion (clip-12: woman with flower pot) is also miss predicted in the

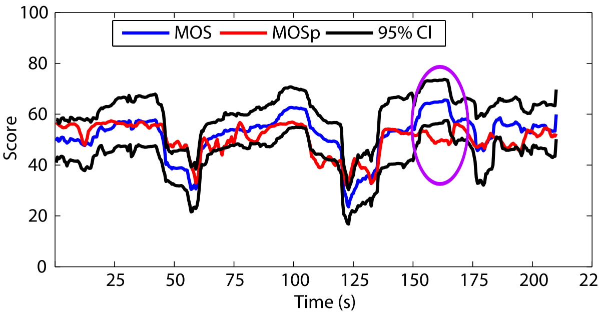

Figure 13. *MOS_p scores with 95%CI of Seq-4, symmetric: (L,R:3,3) Mbps*

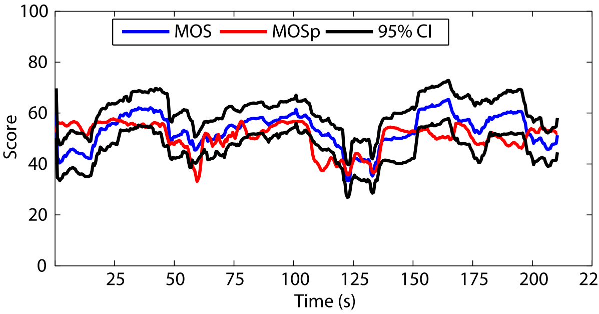

Figure 14. *MOS_p scores with 95%CI of Seq-5, asymmetric: (L,R:5,3) Mbps*

sequence, seq-3(3,2). Because, noise variation in the video clip is localized and not high even though the two views are in almost low bit rate. Moreover, the two objects in the clip are central objects and close to camera. Consequently, the subject could not identify the low quality frames of the clip. However, proposed model try to quantify the highest degradation between the two views during the features extraction. Therefore, the model can easily recognize the low quality frame and gives the prediction significantly low. In the sequence, seq-6(2, 2) the three major miss prediction areas are marked by circles. The circles corresponding clips are Amusement park (clip-5), Football (clip-10), and woman with flower pot (clip-12). Out of the three video clips, two (clip-5 and clip-10) are in very high motion and one (clip-12) in low motion. It has already been discussed regarding miss prediction of the low motion clip in low bit rate scenario. Here, the first two clips are in high motion (i.e., video content changing between adjacent frames in the clips are very high) and camera work of these two videos are also high. Consequently, noise increases and decreases too rapidly within a very short time in both views because of low encoding bit rate (2 Mbps). Therefore, subject can not identify those frames which are high quality. However, proposed model can easily recognize both the high and low quality frames. Therefore, the predictions are higher than the subjective scores.

Since, point to point evaluations are not an appropriate measure in continuous quality prediction we can consider only outlier ratio (OR) as a quantitative measure between the objective (MOSp) and subjective (MOS) scores that can closely indicate the prediction consistency of the model. The evaluation results for training and testing sequences are summarized in Table 3. It has been observed from Table 3 that the evaluation metric, OR is sufficient. Specifically, proposed model provides sufficient prediction consistency (lower OR). It has also been observed from Table 3 that the performance of the proposed model with disparity is better than without disparity for each sequence. The above results confirm that the proposed model's features extractions can be a significant measures for continuous stereo video quality prediction.

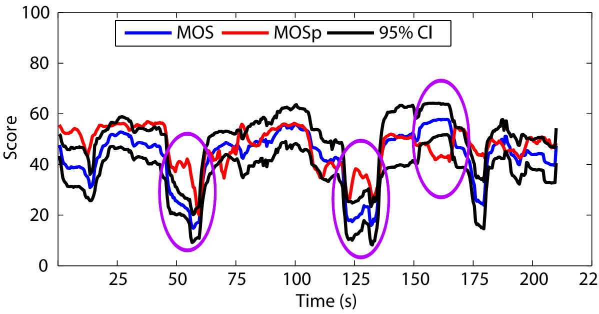

Figure 15. *MOS_p scores with 95%CI of Seq-6, symmetric: (L,R:2,2) Mbps*

Table 3: Evaluation results for training and testing

Seqs	*B.Rate(Mbps)*	Ave. 95%CI	*Training*	
			W. disp.	WO. disp.
			OR	OR
Seq-1	(L, R:5, 5)	± 8.875	0.0357	0.0952
Seq-2	(L, R:8, 2)	± 7.933	0.0786	0.1095
Seq-3	(L, R:3, 2)	± 7.879	0.0857	0.0976
			Testing	
Seq-4	(L, R:3, 3)	± 8.705	0.0048	0.0571
Seq-5	(L, R:5, 3)	± 7.404	0.0452	0.0690
Seq-6	(L, R:2, 2)	± 7.126	0.1310	0.1452

Conclusion

In this work, we presents a feature based NR computational quality evaluation model that can continuously predict perceptual video quality for MPEG-2 MP@ML coded stereoscopic videos. The model uses different HVS aspects to estimated the features. The three measures, such as edge distortion, depth, and jerkiness are determined in the approach. We verify the performance of the proposed model on a stereo database. The result show that the model performs quite well over wide range of video content. Future research can include to classify the video clips into different motion groups such as high, medium, and low with different emphasis so that the model can follow the perceptual variation of motion significantly to predict the quality. In order to incorporate depth perception reliably, the model can also be extended to identify central objects as well as the relation of the central objects to other objects. The aspect of jitter in the observer's reaction time to change in quality would deserve to be investigated as well in future work.

References

[1] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, “Multiview Imaging and 3DTV,” IEEE Signal Processing Magazine, vol. 24, no. 6, pp. 10-21, (2007).

[2] J. Baltes, S. McCann, and J. Anderson “Humanoid Robots: Abarenbou and DaoDan,” RoboCup 2006-Humanoid League Team Description Paper, (2006).

[3] C. F. Westin, “Extraction brain connectivity from diffusion MRI,” IEEE Signal processing magazine, vol. 24, no. 6, pp. 124-152, (2007).

[4] N. Dodgson, “Auto stereoscopic 3-D displays,” IEEE Computer, vol. 38, no. 8, pp. 31-36, (2005).

[5] A. Smolic, P. Kauff, ”Interactive 3-D video representation and coding technology,” in Proc IEEE, Special Issue on Advances in Video Coding and Delivery, vol. 93, no. 1, pp. 98-110, (2005).

[6] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, G. Bozdagi Akar, “Towards compound stereo-video quality metric: a specific encoder-based framework,” IEEE SSIAI, Colorado, USA, March 2006.

[7] N. Ozbek, A. M. Tekalp, and E. T. Tunali, “Rate allocation between views in scalable stereo video coding using objective stereo video quality measure,” in Proc. IEEE ICASSP, Hawai, USA, Apr. 15-20, 2007.

[8] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, “Prediction of stereoscopic video quality using objective quality models of 2-D video,” Electronics Letter, vol. 44, no. 6, pp. 963-965, July 2008.

[9] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S. Villette, and A.M. Kondoz, “Quality evaluation of color plus depth map based stereoscopic video,” IEEE. J. Sel. Topics Signal Process., vol. 3, no. 2, pp 304318, 2009.

[10] S.L.P. Yasakethu, C.T.E.R. Hewage, W.A.C. Fernando, and A.M. Kondoz, “Quality analysis for 3D video using 2D video quality models,” IEEE Trans. Consum. Electron., vol. 54, no. 4, pp. 19691976, Nov. 2008.

[11] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn, “Depth map quality metric for three-dimensional video,” in Proc. SPIE-IS and T Electronic Imaging, Vol. 7237, Jan. 18-22, San Jose, USA, 2009.

[12] D. V. S. X. De Silva, W. A. C. Fernando, G. Nur, E. Ekmecioglu, and S.T. Worrall, “3D video assessment with just noticeable difference in depth evaluation,” in Proc. IEEE ICIP, Hong Kong, Sept. 26-29, 2010.

[13] Z. M. Parvez Sazzad, S. Yamanaka, and Y. Horita, “Spatio-temporal Segmentation Based Continuous No-reference Stereoscopic Video Quality Prediction,” in Proc. IEEE QoMEX, June 21-23, Trondheim, Norway, 2010.

[14] Jin L., Boev A., Gotchev A., Egiazarian K., “3D-DCT Base Perceptual Quality Assessment of Stereo Video,” in Proc. IEEE ICIP, Brussels, September, 11-14, 2011.

[15] Jin L., Boev A., Gotchev A., Egiazarian K.,“Validation of A New Full Reference Metric for Quality Assessment of Mobile 3DTV Content,” in Proc. EUSIPCO, Barcelona, Spain, August 29-September 2, 2011.

[16] Egiazarian K., Astola J., Ponomarenko N., Lukin V., Battisti F., Carli M., “New full-reference quality metrics based on HVS,” in Proc. of the Second International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2006.

[17] Ponomarenko N., Silvestri F., Egiazarian K., Carli M., Astola J., Lukin V., “On between-coefficient contrast masking of DCT basis functions,” in Proc. of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07, Scottsdale, Arizona, USA, 25-26 January, 2007.

[18] Z. Zhu, and Y. Wang, “Perceptual Distortion Metric for Stereo Video Quality Evaluation,” Wseas Trans. on Signal Processing, vol. 5, no. 7, pp. 241-250, 2009.

[19] A. Mittal, A. K. Moorthy, J. Ghosh, and A. C. Bovik, “Algorithm assessment of 3D quality of experience for images and videos,” in Proc. IEEE ICIP, Hong Kong, Sept. 26-29, 2010.

[20] M. Solh, G. AlRegib, and J. M. Bauza, “3VQM: A Vision-based Quality Measure For DIBR-based 3D Videos,” in Proc. IEEE ICME, Barcelona, Spain, July 2011.

[21] M. Solh, and G. AlRegib, “A no-reference quality measure for DIBR based 3D videos,” in Proc. IEEE ICIP Brussels, September, 11-14, 2011.

[22] R. Bensalma, and M. Chaker Larabi “Towards a perceptual quality metric for color stereo images,” in Proc. IEEE ICIP, Hong Kong, Sept. 26-29, 2010.

[23] S. Winkler, “Perceptual Video Quality Metrics A review,” in H. Wu and K. Rao, eds., “Digital video image quality and coding,” chap. 5, CRC press, 2006.

[24] S. Arata, Y. Horita, K. Honda, and T. Murai, “Continuous video quality evaluation of coded stereoscopic video,” Technical report of IEICE, IE2004-150(2005-1), 2005.

[25] ITU-R BT.500-10. Methodology for the Subjective Assessment of the Quality of Television Pictures.

[26] A. Klaus, M. Sormann, and K. Karner,“Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure,” in Proc. IEEE ICPR, Hong Kong, Aug. 20-24, (2006).

[27] Z. M. Parvez Sazzad, S. Yamanaka, Y. Kawayoke, and Y. Horita,“Stereoscopic image quality prediction,” in Proc. IEEE QoMEX, San Diego, CA, USA, July 29-31, (2009).

[28] Y. Horita, M. Miyahara, and T. Murai, “Estimation improvement in picture quality scale of monochrome still picture,” IEICE Trans. j80 (B-I), pp. 505514 (1997).

[29] J. Kennedy and R. Eberhart, “Particle Swarm Optimization,” in Proc. IEEE ICNN, Perth, Australia, pp. 1942-1948, (1995).