DUJASE Vol. 4(1) 11-18, 2017 (January)

Feature Based No-Reference Perceptual Depth Assessment Model for Mobile 3D Video Applications

Iffat Alam and Z. M. Parvez Sazzad

Dept. of Electrical and Electronic EngineeringUniversity of Dhaka, Dhaka-1000, Bangladesh.

iffat.alam28@gmail.com, sazzad@du.ac.bd

Received on 20.12.15, Accepted for publication on 18.06.2017

ABSTRACT

Depth perception is one of the most important characteristics which separate 3D videos from traditional 2D videos. In this work, a feature based no-reference perceptual depth assessment model has been proposed for symmetric and asymmetric coded stereoscopic videos. This model extracts disparity and temporal features to evaluate the perceived depth of mobile 3D videos. The disparity feature is estimated by using block based structural similarity index between the corresponding blocks of left and right view and for temporal feature the jerkiness is estimated between the consecutive frames for both left and right view. The estimated features are then combined to give a single predicted score. The performance of the model is verified by subjective experiment data. The result indicates that the prediction performance of the proposed model is satisfactory.

Keywords: No-reference, stereoscopic 3D video, Deptl, Symmetric, Asymmetric.

1. Introduction

With the development of stereoscopic image and video technology, there is no doubt that all conventional 2D media are soon going to be replaced by 3D media to improve the quality of experience for all media applications from broadcasting [1] to more specialized applications such as remote education [2], robotic navigation [3], medical applications [4] and many more. 3D video is gaining a world-wide popularity both in cinema and broadcasting industries as it is a technology that will extensively enhance the user's visual experience. One of the major concerns of such technology is to provide sufficient visual quality, especially if 3D video is to be transmitted over a limited bandwidth. In case of 3D videos, depth perception is one of the most important characteristics which separate them from 2D videos as depth perception enables us to detect distance between different objects/structural contents in the video. Therefore, the necessity to define appropriate methods to assess the perceived depth of the processed stereoscopic images and videos is becoming more evident. Assessing 3D video is a challenging issue because it is affected by video quality, depth perception and visual comfort. It is particularly challenging to evaluate the depth when the stereoscopic image consists of two views with different quality. In [5], a depth perception assessment index is proposed for stereoscopic images using a phase-shift model. They used Gabor filter to compute the responses of left and right images respectively, and proposed a phase-shift model for computing disparity maps based on phase gradient and phase difference information. In [6], a source of information for absolute depth estimation was proposed based on the whole scene structure that does not rely on specific objects. It was demonstrated that by recognizing the properties of the structures present in the image, the scale of the scene can be estimated and therefore its absolute mean depth. Hwang et al. [7] proposed a visual attention and depth assisted stereo image quality assessment model which consists of stereo attention predictor, depth variation and stereo distortion predictor. Faria et al. [8] proposed a stereoscopic depth perception approach inspired by the primary visual cortex using the stimulus response of the receptive field profiles of binocular cells for disparity computation. Lebreton et al. [9] characterized depth information provided by the source sequences as an important factor because it validates whether the content is suitable for 3D video services. Boev et al. [10] combined monoscopic and stereoscopic quality components from the 'Cyclopean' image and disparity map respectively for stereo-video evaluation. In [11], a subjective experiment is presented to study the relation between blur/sharpness and depth. It extended the concept of just noticeable blur (JNB) at different depths for 2D videos to 3D videos. In [12], a 3D visual attention model is proposed for stereoscopic image quality assessment based on 2D saliency model, center bias, depth cue. The perceptual depth can be assessed either through subjective tests or through objective metrics. Subjective assessment refers to the process of collecting the opinions of a large number of viewers in the form of opinion scores that rate the visual perception of a video. These scores are then averaged to get mean-opinion-score (MOS). Even though it is the most accurate way to assess a video, it is not suitable for real time applications. Therefore, objective evaluation is becoming an ever increasing requirement to monitor the visual perception in real time. Consequently, no-reference evaluation method is highly desired at end user terminals as the pristine reference video will not be always available. In this work, a feature based no-reference perceptual depth assessment model is proposed for symmetric and asymmetric coded stereoscopic videos. This model extracts disparity and temporal features to evaluate the perceptual depth. The disparity feature is extracted to measure the perceived depth of the stereo video. Finally, video jerkiness is estimated as the temporal feature. The outline of this paper is as follows: Section 2 describes the details of the proposed model. The experimental results and performance evaluation with subjective experiment data are given in Section 3. Finally, conclusions are drawn in Section 4.

2. Proposed No-Reference Model

The block diagram of the proposed no-reference (NR) perceptual depth assessment model is shown in Figure 1. The model extracts the following features:

- Disparity Feature
- Temporal Feature

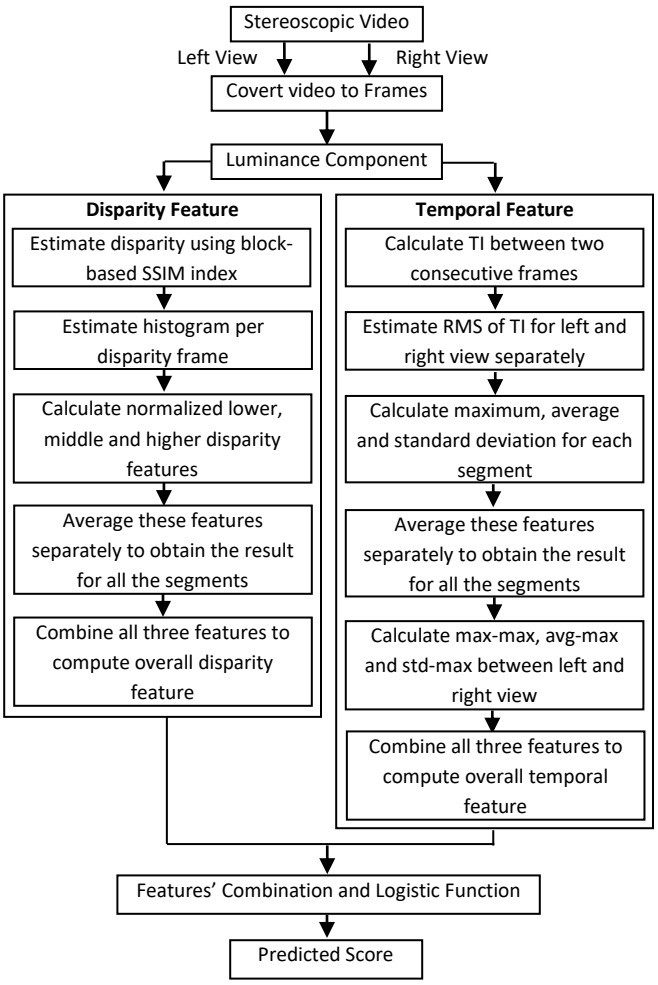
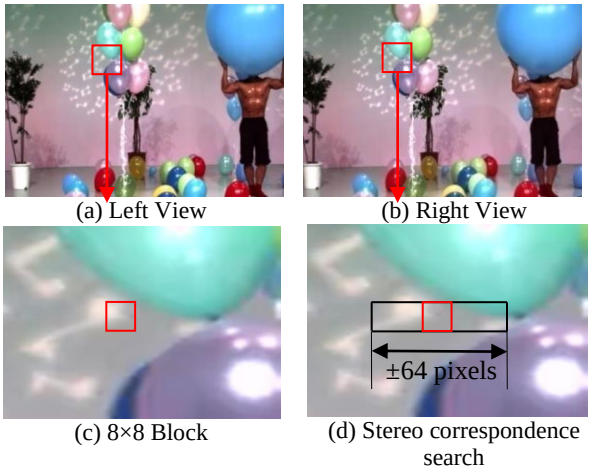


Fig. 1: Proposed depth assessment model

2.1 Disparity Feature

In this work, disparity is estimated by considering the structural similarity [13] between the corresponding blocks of left and right view [14]. In case of stereoscopic videos, a frame consists of two views i.e. left view and right view. During disparity estimation, each pixel in left image is matched with their corresponding pixels in the right image so that the corresponding pixels are the projections of the same 3D position. In case of standard stereo setup, we can consider that the camera movement is only along the horizontal direction. As a result, the displacement between the left and right view is considered to be in the horizontal direction only. Therefore, the corresponding pixels are constrained to lie on the same row.



(a) Left View (b) Right View (c) 8×8 Block (d) Stereo correspondence search

Fig. 2: (a) Left view; (b) Right view; (c)-(d) Expanded view of (a) and (b); For a 8×8 block in the left image (c) a stereo correspondence search is conducted in the right image along ±64 pixels in the horizontal direction (d).

In order to measure the disparity feature, the left image is segmented into non-overlapping 8×8 blocks. For each 8×8 block of the left image, the corresponding block search in the right image is conducted up to ±64 pixels using the Structural Similarity index (SSIM) measure which is shown in Figure 2. The SSIM index is defined as

$$Q = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

where, $x = \{x_i \mid i = 1,2,\ldots,64\}$ and $y = \{y_i \mid i = 1,2,\ldots,64\}$ be the left and right image blocks respectively with

$$\mu_x = \frac{1}{64}\sum_{i=1}^{64} x_i \quad (2)$$

$$\mu_y = \frac{1}{64}\sum_{i=1}^{64} y_i \quad (3)$$

$$\sigma_x^2 = \frac{1}{63}\sum_{i=1}^{64} (x_i - \overline{x})^2 \quad (4)$$

$$\sigma_y^2 = \frac{1}{63}\sum_{i=1}^{64} (y_i - \overline{y})^2 \quad (5)$$

$$\sigma_{xy}^2 = \frac{1}{63}\sum_{i=1}^{64} (x_i - \overline{x})(y_i - \overline{y}) \quad (6)$$

And, $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$ are two constants to stabilize the division with weak denominator where L is the dynamic range of the pixel-values, and $k_1 = 0.01$, $k_2 = 0.03$ by default.

The resultant SSIM index is a decimal value between -1 and 1. The value 1 is achieved if and only if $x_i = y_i$ for all $i = 1, 2,\ldots,64$. For each 8×8 block of the left image, disparity index (DI) is found by searching the position of maximum quality index up to ±64 pixels of the right image.

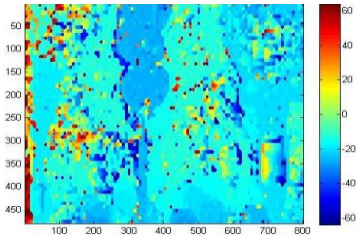
$$DI = \max\begin{Bmatrix} Q_{-64}, Q_{-63}, Q_{-62}, \ldots\ldots, \\ Q_{-1}, Q_0, Q_1, \ldots\ldots, Q_{63} Q_{64} \end{Bmatrix} \quad (7)$$

In this way, the disparity map can be found for each pixel in a stereo view. Eventually, this disparity estimation process is conducted for each frame of the stereoscopic video. Subsequently, a depth map of the stereo frame of Balloon sequence is shown in Figure 3.

Finally, after obtaining the depth map, the histogram of the disparity frames is estimated. The lower, middle and higher parts of the histogram are considered and then these values are normalized considering the highest disparity value. Subsequently, these three normalized disparity features are considered to measure depth in this method.

(a) Balloon Left View

(b) Balloon Right View



(c) Depth Map

Fig. 3: Depth map of the Balloon sequence.

We consider,

- Lower disparity: $h(0), h(1), h(2)$

where $h(0), h(1), h(2)$ indicate number of disparity pixels with pixel's displacement 0, 1 and 2 respectively.

- Middle disparity: $h\left(\frac{d}{2}-1\right), h\left(\frac{d}{2}\right), h\left(\frac{d}{2}+1\right)$
- Higher disparity: $h(d-2), h(d-1), h(d)$

where d is the maximum pixel disparity.

- V is the variance of the disparity frames.

For normalized disparity,

$$NDl(f) = V \times \left\{ \frac{h(0) \cdot (0+1) + h(1) \cdot (1+1)}{M \times N \times (d+1)} + \frac{h(2).(2+1)}{M \times N \times (d+1)} \right\} \quad (8)$$

$$NDm(f) = V \times \left\{ \frac{h\left(\frac{d}{2}-1\right) \cdot \left(\left(\frac{d}{2}-1\right)+1\right)}{M \times N \times (d+1)} + \frac{h\left(\frac{d}{2}\right) \cdot \left(\frac{d}{2}+1\right) + h\left(\frac{d}{2}+1\right) \cdot \left(\left(\frac{d}{2}+1\right)+1\right)}{M \times N \times (d+1)} \right\} \quad (9)$$

$$NDh(f) = V \times \left\{ \frac{h(d-2) \cdot \left((d-2)+1\right)}{M \times N \times (d+1)} + \frac{h(d-1) \cdot \left((d-1)+1\right) + h(d) \cdot (d+1)}{M \times N \times (d+1)} \right\} \quad (10)$$

where $NDl(f), NDm(f)$ and $NDh(f)$ are respectively lower, middle and higher disparity features of a stereo frame pair.

We consider 8 frames per temporal segment. The lower disparity feature for a temporal segment, s is calculated as:

$$NDl(s) = \frac{1}{8}\sum_{f=1}^{8} NDl(f) \quad (11)$$

Similarly, the middle and higher disparity features for a temporal segment can be calculated as:

$$NDm(s) = \frac{1}{8}\sum_{f=1}^{8} NDm(f) \quad (12)$$

$$NDh(s) = \frac{1}{8}\sum_{f=1}^{8} NDh(f) \quad (13)$$

Finally, the total lower, middle and higher disparity features are calculated by taking the average of $NDl(s), NDm(s)$ and $NDl(s)$ for all the 15 segments as the sequences were 8 seconds long with 15 fps (i.e. 120 frames in total). Lastly, all three disparity features are combined by some weighting factors to estimate the overall disparity feature.

$$ND = NDl(s)^{\gamma_1} \cdot NDm(s)^{\gamma_2} \cdot NDh(s)^{\gamma_3} \quad (14)$$

where γ_1, γ_2 and γ_3 are the weighting factors.

2.2. Temporal Feature

To compute the temporal feature, maximum jerkiness between the consecutive frames is estimated for both left and right view frames. Jerkiness of any stereoscopic video depends on the motion and scene contents of the video. In temporal domain jerkiness is really annoying for human eye. To measure video jerkiness as a temporal feature, the luminance intensity variation of pixels between the consecutive frames is used. The temporal feature extraction is shown for left view frames in Figure 4.

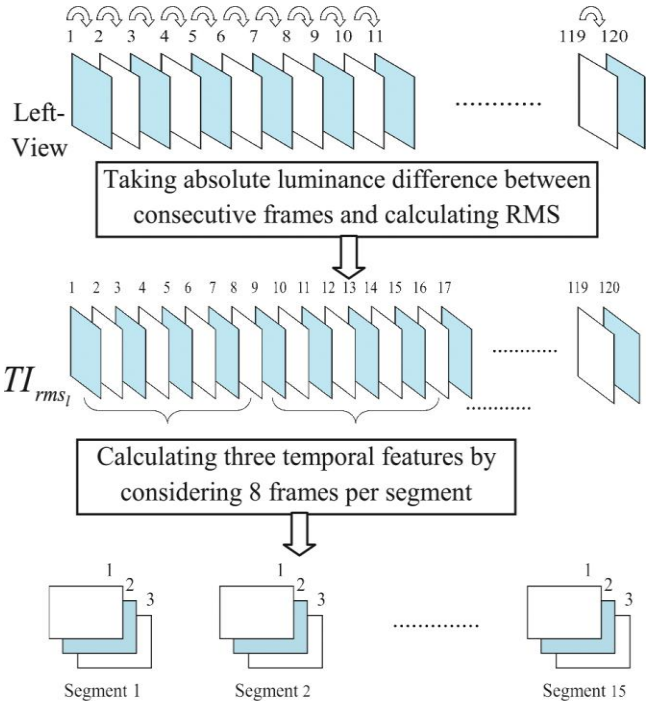


Figure 4: Temporal feature extraction is shown for left view frames.

Firstly, the temporal information, TI i.e. the absolute luminance difference between consecutive frames is estimated separately for left and right views.

For left view:

$$TI_l(m,n,t) = |x_l(m,n,t+k) - x_l(m,n,t)| \quad (15)$$

where $k = 1$ and $t = 1,2,3,\ldots$ are the selected frame numbers.

Secondly, the deviation of the temporal information is calculated:

$$TI_{d_l}(t) = \sqrt{\overline{TI_l^2(m,n,t)} - \overline{TI_l(m,n,t)}^2} \quad (16)$$

Thirdly, the root mean square is estimated:

$$TI_{rms_l}(t) = \sqrt{TI_{d_l}^2(t) + \overline{TI_l(m,n,t)}^2} \quad (17)$$

Similarly, $TI_{rms_r}(t)$ is estimated for right view frames.

Now considering 8 frames per temporal segment, three temporal features are calculated for each temporal segment from $TI_{rms_l}(t)$ and $TI_{rms_r}(t)$ separately.

For left view:

First temporal feature is computed by:

$$TI_{max_l}(s) = max\begin{Bmatrix} TI_{rms_l}(t), TI_{rms_l}(t+1), \\ TI_{rms_l}(t+2), \ldots\ldots, \\ TI_{rms_l}(t+7) \end{Bmatrix} \quad (18)$$

where t = 1, 9, 17, …

Second temporal feature is computed by:

$$TI_{avg_l}(s) = \frac{1}{8}\sum_{t=1}^{8} TI_{rms_l}(t) \quad (19)$$

Third temporal feature is computed by:

$$TI_{std_l}(s) = \sqrt{\frac{1}{7}\sum_{t=1}^{8}\left(TI_{rms_l}(t) - \overline{TI_{rms_l}(t)}\right)^2} \quad (20)$$

Similarly, these three temporal features are calculated for right view as, $TI_{max_r}(s), TI_{avg_r}(s)$,and $TI_{std_r}(s)$ respectively.

Now these three temporal features are calculated for all the segments of a video sequence for left view by:

$$TI_{max_l} = \frac{1}{F}\sum_{s=1}^{F} TI_{max_l}(s) \quad (21)$$

$$TI_{avg_l} = \frac{1}{F}\sum_{s=1}^{F} TI_{avg_l}(s) \quad (22)$$

$$TI_{std_l} = \frac{1}{F}\sum_{s=1}^{F} TI_{std_l}(s) \quad (23)$$

where F = 15 which is the total number of temporal segments.

Similarly, for right view the temporal features for all the segments are calculated as, TI_{max_r}, TI_{avg_r} and TI_{std_r}.

Finally, three temporal features are calculated by taking the maximum temporal feature between the two views using the following equations.

$$TI_{max-max} = max\left(TI_{max_l}, TI_{max_r}\right) \quad (24)$$

$$TI_{avg-max} = max\left(TI_{avg_l}, TI_{avg_r}\right) \quad (25)$$

$$TI_{std-max} = max\left(TI_{std_l}, TI_{std_r}\right) \quad (26)$$

Lastly, all three temporal features are combined by some weighting factors to estimate the overall temporal feature.

$$TI = TI_{max-max}^{\theta_1} \cdot TI_{avg-max}^{\theta_2} \cdot TI_{std-max}^{\theta_3} \quad (27)$$

where θ_1, θ_2 and θ_3 are the weighting factors.

2.3 Features Combination

To constitute a stereoscopic depth prediction model, following features combination equation is considered to integrate the disparity and temporal features.

$$S = \alpha + \beta(TI) \cdot (ND) \quad (28)$$

where α and β are the model parameters and TI, ND represent the overall temporal and disparity features.

A logistic function is used as the non-linearity property between the human perception and the physical features [15]. Finally, the *MOS* prediction score *(MOS_p)* is derived by:

$$MOS_p = \frac{4}{1 + exp[-1.0217(S-3)]} + 1 \quad (29)$$

The model parameters and weighting factors are estimated by an optimization algorithm with the subjective test data. Here, Particle Swarm Optimization (PSO) algorithm is used for optimization [16].

3. Performance Evaluation

In this work, subjective experiment data are used to evaluate the performance of the proposed model. All subjects participating in the subjective experiment were students of University of Dhaka. The experiments were performed to estimate the mean opinion score (MOS) for perceptual depth of the stereo video sequences. These perceptual depth scores are used to train and test the proposed model.

Table 1: Subjective test conditions and parameters

Method	DSQS
Coder	H.264
Bit Rates	4 kinds (100, 150, 250 and 400 Kbps)
Stereo Video Clips	6
Video Resolution	480×800 pixels (24 bit/pixel, RGB)
Each clip length, and frame rate	8 sec, and 15 fps
Subjects	31 (Non expert, Students)
Display	4.3-inch, LCD 3D Auto-stereoscopic
Display Resolution	480×800 pixels
Viewing Distance	Adjustable viewing distance
Room Illumination	Dark

The subjective experiment was conducted by using Double Stimulus Quality Scale (DSQS) method. Sixteen test sequences were created from each of the reference sequences [17] by symmetric and asymmetric combinations of four bit rates - 100, 150, 250 and 400 Kbps for left view and right view. The resolution of each video sequence was 480×800 pixels. The duration of each sequence was 8 seconds with 15 frames per second. The subjective test conditions and parameters are summarized in Table 1. Each subject was shown 96 test videos in a random order. In each sequence, two versions of the same video clip were shown in succession. First one was the reference and the second one was the test sequence which was rated on a discrete five point scale. The five point 'perceptual depth' scales are Not perceptible at all = 1, Slightly perceptible = 2, Fairly perceptible = 3, Easily perceptible = 4, and Strongly perceptible = 5. Note that the numerical values attached to each category were only used for data analysis and were not shown to the subjects. Mean opinion scores (MOSs) were then computed for each stereo sequence after post-experimental screening according to ITU-R BT 500-11 recommendation [18]. The effect of bit rates on depth perception was examined by analyzing MOS at different bit rate combinations for the six sequences used in our experiment. Out of six sequences, four sequences are selected in this work based on the content of the videos which comprised of indoor and outdoor scenes that ranges from low to medium motion, where the scenes were filmed at both close and faraway distances. The sequences are divided into two parts for training and testing. The training dataset consists of two sequences Balloon and Newspaper and the testing dataset consists of two other sequences Poznan Street and Lovebird.

3.1 Training Result

The model parameters and weighting factors estimated by the Particle Swarm Optimization (PSO) algorithm with the training dataset are shown in Table 2. In order to provide quantitative measures on the performance of our proposed NR depth prediction model, we followed the standard performance evaluation procedures employed in the video quality experts group (VQEG) FR-TV Phase II test [19], where mainly four evaluation metrics, correlation coefficient (CC), Spearman rank order correlation coefficient (SROC), average error (AVE), and root mean square error (RMSE) between MOS and MOS prediction (MOS_p) were used. The CC, SROC, AVE (calculated on the scale of 5), RMSE of the training dataset are shown in Table 3.

Table 2: Model parameters and weighting factors

$\alpha = 75.99013$	$\beta = -77.2595$	
$\theta_1 = -0.02525$	$\theta_2 = -0.01684$	$\theta_3 = -0.00311$
$\gamma_1 = 0.024923$	$\gamma_2 = 0.020574$	$\gamma_3 = -0.01452$

Table 3: Evaluation results for training dataset

CC	SROC	AVE	RMSE
0.87	0.85	0.23	0.295

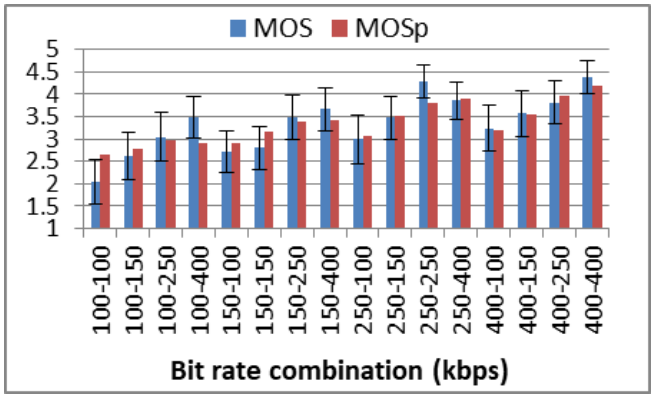


Fig. 5: MOS and MOS_p scores of Balloon sequence

Figure 5 shows the comparison between MOS and MOS_p scores for the Balloon sequence. The MOS scores are shown with 95% confidence interval. The figure indicates

that the MOS_p scores are within 95% confidence interval for almost all the bit rate combinations except at (100-400) Kbps and (250-250) Kbps where the deviations are noticeable. Both MOS and MOS_p scores exhibit lowest depth perception scores at (100-100) Kbps, i.e. 2.04 and 2.64 respectively. The highest depth perception scores are achieved at (400-400) Kbps for both subjective and objective assessments. At (400-400) Kbps, the MOS and MOS_p scores are 4.38 and 4.19 respectively. The asymmetric bit rate combinations (250-400) Kbps and (400-250) Kbps exhibit consistency between the subjective and the objective scores. At (250-400) Kbps, the MOS score is 3.86 and MOS_p score is 3.91 and at (400-250) Kbps, the MOS score is 3.81 and MOS_p score is 3.98. Moreover, we see that there is a noticeable variation between the MOS and MOS_p scores at the bit rate combination (250-250) Kbps. The MOS score is 4.28 while the MOS_p score is 3.82. In addition, the bit rate combinations (100-400) Kbps and (400-100) Kbps show poor MOS and MOS_p scores when compared with (250-250) Kbps even though the total bit rate of (250-250) Kbps, (100-400) Kbps and (400-100) Kbps are the same. So it can be said that if the difference between left view and right view bit rate is greater, then the symmetric combination gives somewhat better prediction result than the asymmetric combinations.

The analysis shown in Figure 6 represents the comparison between MOS and MOS_p scores for the Newspaper sequence.

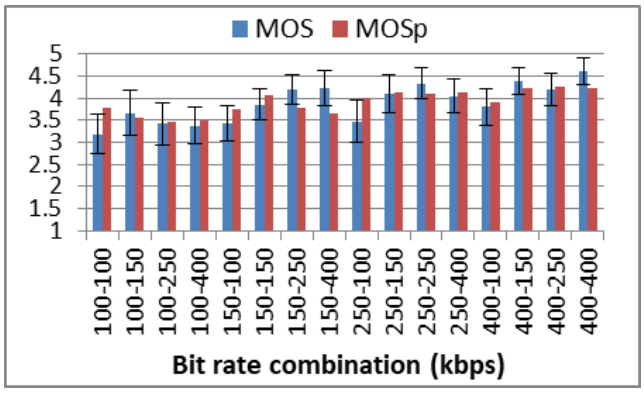


Fig. 6: MOS and MOS_p scores of Newspaper sequence

Here, we can see that almost all the MOS_p scores are within 95% confidence interval except at (100-100) Kbps and (150-400) Kbps where the deviations are noticeable. Both the MOS and MOS_p scores are above 3 for all bit rate combinations. The bit rate combination (400-400) Kbps exhibits highest depth perception scores for both subjective and objective scores. At (400-400) Kbps, MOS and MOS_p scores are 4.62 and 4.23, respectively. In addition, same as the Balloon sequence, the bit rate combinations (250-400) Kbps and (400-250) Kbps show consistency between the subjective and objective scores. Moreover, the scores of (250-400) Kbps and (400-250) Kbps are comparable to (400-400) Kbps. Moderate results are obtained for (250-250) Kbps bit rate combination for both subjective and objective scores. At (250-250) Kbps, MOS score is 4.3 and MOS_p score is 4.1. The bit rate combinations (100-400) Kbps and (400-100) Kbps show consistency between MOS and MOS_p scores but the scores are not as good as (250-250) Kbps. Therefore, the same assessment as Balloon sequence holds true for the Newspaper sequence.

3.2 Testing Result

The model parameters and weighting factors, that are estimated using the training dataset, are applied on the testing dataset to evaluate the performance of our proposed model. The evaluation result for the testing dataset is shown in Table 4.

Table 4: Evaluation results for testing dataset

CC	SROC	AVE	RMSE
0.77	0.74	0.39	0.51

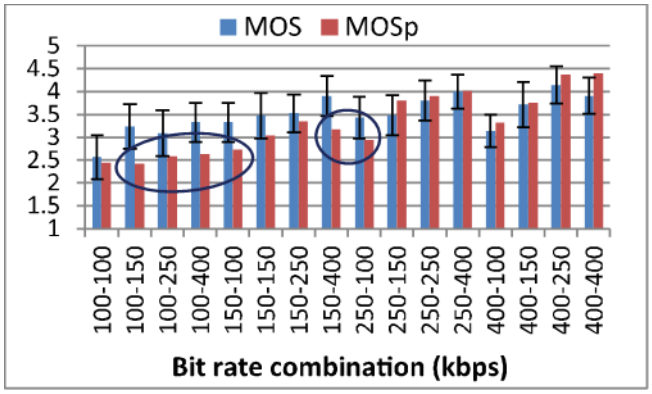


Fig. 7: MOS and MOS_p scores of Poznan Street sequence

Figure 7 shows the comparison between MOS and MOS_p scores for the Poznan Street sequence. Here, when the MOS_p scores are compared with the MOS scores, noticeable deviations are observed at the low bit rate combinations i.e. the MOS_p scores are lower than the MOS scores. The Poznan Street sequence has high motion i.e. the video content change between adjacent frames is high. Even though at low bit rate combinations where the quality is low, the subject can easily identify the depth of the video. Whereas the model tries to quantify the highest degradation between the two views and therefore at low bit rate combinations it cannot predict the depth of the video and gives the prediction scores lower than the subjective scores. However, in case of high bit rate combinations, the MOS_p scores show consistency with the MOS scores. The highest MOS score is obtained at (400-250) Kbps which is 4.14 and its corresponding MOS_p score is 4.3 which shows consistency. In addition, (250-400) Kbps bit rate combination shows consistency between the MOS and MOS_p scores which are 4 and 4.01 respectively. The bit rate combination (250-250) Kbps shows comparable results. However, the bit rate combinations (100-400) Kbps and (400-100) Kbps show poor and inconsistent MOS and MOS_p scores when compared with (250-250) Kbps even though the total bit rate of (250-250) Kbps, (100-400) Kbps and (400-100) Kbps are the same.

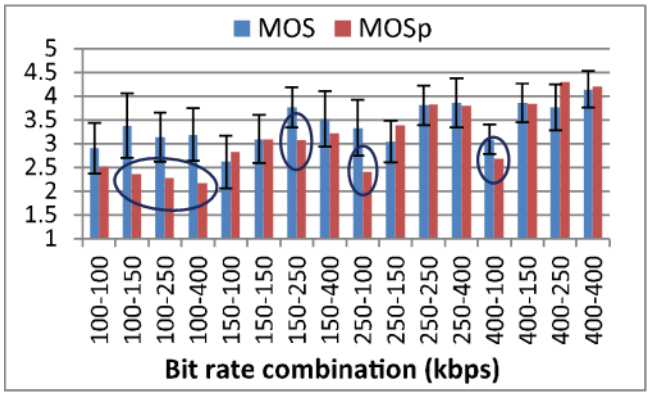


Fig. 8: MOS and MOSp scores of Lovebird sequence

Figure 8 shows the comparative analysis between MOS and MOS_p score for the Lovebird sequence. Same as the Poznan Street sequence when the MOS_p scores are compared with the MOS scores, noticeable deviations are observed at the low bit rate combinations i.e. the MOS_p scores are lower than the MOS scores. Lovebird sequence has medium motion content. The two objects in the video clip are the central objects and they slowly walks towards the camera. Just like the Poznan Street sequence, even at low bit rate combinations the subject could easily detect the depth of the video but the model could not determine the depth of the video accurately and gives the score lower than the subjective scores. On the other hand, in case of high bit rate combinations, the MOS_p scores show consistency with the MOS scores. The highest MOS score is obtained at (400-400) Kbps which is 4.14 and its corresponding MOS_p score is 4.2 which shows consistency. In case of the bit rate combination (400-250) Kbps there is a variation between the MOS and MOS_p scores. The MOS score is 3.8 whereas the MOS_p score is 4.2. However, the bit rate combination (250-400) Kbps shows consistency between the MOS and MOS_p scores. The bit rate combination (250-250) Kbps shows comparable and consistent results. However, same as before the asymmetric bit rate combinations (100-400) Kbps and (400-100) Kbps show poor and inconsistent results when compared with (250-250) Kbps. As far as we know there is no NR model for depth evaluation for mobile stereoscopic 3D videos. Therefore, it is clear from Figures 5, 6, 7, and 8 and also from Tables 3 and 4, that our proposed model performances are sufficient.

4. Conclusion

In this work, a no-reference perceptual depth assessment model is proposed based on disparity and temporal features for stereoscopic 3D videos for mobile applications. The performance of the proposed model has been evaluated by using the subjective experiments data. The high value of correlation coefficient and low value of average error, and root mean square error between MOS and MOS_p indicate sufficient accuracy of our proposed model. It is observed from the experimental results that the proposed model can achieve much higher accuracy and consistency with subjective experiment data if we incorporate human visual system (HVS) characteristics properly. In future, the research can be extended to incorporate properly the HVS system characteristics to the model to extract the disparity and temporal features.

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. A. M. William and D. L. Bailey, “Stereoscopic visualization of scientific and medical content,” in Proc. of the 33rd International Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH’06), pp. 26, ACM, Boston, Mass, USA, July-August 2006.
3. J. Baltes, S. McCann, and J. Anderson, “Humanoid Robots: Abarenbou and DaoDan,” RoboCup 2006-Humanoid League Team Description Paper, 2006.
4. C. F. Westin, “Extracting brain connectivity from diffusion MRI,” IEEE Signal Processing Magazine, vol. 24, no. 6, pp. 124–152, 2007.
5. Q. Jiang, S. Wang, K. Li and F. Shao, “Measuring depth perception for stereoscopic images using a phase-shift model,” Journal of Software, vol. 9, no. 10, pp. 2665-2671, Oct. 2014.
6. A. Torralba and A. Oliva, “Depth estimation from image structure,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, Sept. 2002.
7. J. J. Hwang and H. R. Wu, “Stereo image quality assessment using visual attention and distortion predictors,” KSII Transactions on Internet and Information Systems, vol. 5, no. 9, pp. 1613-1631, 2011.
8. F. Faria, J. batista, H. Araújo, “Stereoscopic depth perception using a model based on the primary visual cortex,” PloS One, vol. 8, no.12, pp. e80745, 2013.
9. P. Lebreton, A. Raake, M. Barkowsky, P. Le Callet, “Evaluating depth perception of 3D stereoscopic videos,” IEEE Journal of Selected Topic in Signal Processing, vol. 6, no. 6, pp. 710-720, Oct. 2012.
10. A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, G. B. Akar, “Towards compound stereo-video quality metric: a specific encoder-based framework,” in Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation, Denver, Colorado, pp. 218-222, 2006.
11. M. M. Subedar, L. J. Karam, “Increased depth perception with sharpness enhancement for stereo video,” in Proc. of SPIE-IS&T Electronic Imaging, vol. 7524, id. 75241B, Feb. 2010.
12. Q. Jiang, F. Duan, F. Shao, “3D Visual Attention for Stereoscopic Image Quality Assessment,” Journal of Software, vol. 9, no. 7, pp. 1841-1847, July 2014.
13. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004.
14. M. M. Akram, H. Siddiqa and Z. M. Parvez Sazzad, "Structural similarity based disparity estimation for stereoscopic images," 16th International Conference on Computer and Information Technology (ICCIT), 2013, pp.76-80, 8-10 March 2014.

15. 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.

16. J. Kennedy and R. Eberhart, “Particle Swarm Optimization,” in Proc. IEEE ICNN, Perth, Australia, vol. 4, pp. 1942-1948, 1995.

17. Mobile 3DTV content delivery optimization over DVB-H system. Retrieved from http://sp.cs.tut.fi/mobile3dtv/stereo-video/

18. ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Geneva, Switzerland, 2002.

19. VQEG: Final Report from the video quality experts group on the validation of objective models of video quality assessment, FR-TV Phase II, August 2003. Retrieved from http://www.vqeg.org/.