Coding of 3D Videos based on Visual Discomfort

Dogancan Temel and Ghassan AlRegib

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA, 30332-0250 USA
{cantemel, alregib}@gatech.edu

***Abstract*—We propose a rate-distortion optimization method for 3D videos based on visual discomfort estimation. We calculate visual discomfort in the encoded depth maps using two indexes: temporal outliers (TO) and spatial outliers (SO). These two indexes are used to measure the difference between the processed depth map and the ground truth depth map. These indexes implicitly depend on the amount of edge information within a frame and on the amount of motion between frames. Moreover, we fuse these indexes considering the temporal and spatial complexities of the content. We test the proposed method on a number of videos and compare the results with the default rate-distortion algorithms in the H.264/AVC codec. We evaluate rate-distortion algorithms by comparing achieved bit-rates, visual degradations in the depth sequences and the fidelity of the depth videos measured by SSIM and PSNR.**

***Index Terms*—Rate distortion optimization, depth map, 3D video, perceptual quality, image fidelity, visual discomfort**

I. Introduction

Free Viewpoint Video *(FVV)* enables navigation inside a scene whereas Three-Dimensional Television *(3DTV)* provides the depth perception to the end user. 3D view synthesis methods are either based on 2D depth maps or sparse 3D scene structures. In terms of video coding, depth map-based methods are more feasible and applicable because they do not require large bandwidth and the acquisition system can be as simple as a stereo camera. When 3D view synthesis methods are considered, Depth-Image-Based Rendering (*DIBR*) is the most commonly preferred approach in the literature [1]. In DIBR-based 3D view synthesis, a depth map is required for each color frame. Depth map is basically a grayscale image where each pixel value corresponds to the relative distance between the camera reference frame and world reference frame. By using the relative depth information in the depth map and camera setup parameters such as baseline, focal length, convergence distance and relative position, color pixels can be mapped to the world frame in 3D and then they can be projected to a new 2D camera frame (virtual view) at the receiver side. Users can have the 3D experience by feeding the stereo image pair that consists of reference and virtual view to the 3D display system.

We can encode the depth sequences using the Advanced Video Coding *(AVC)* standard. A detailed overview of the *H.264/MPEG-4 AVC* is provided in [2]. As it is explained in [2], rate control approaches consist of three main steps: target bit allocation, virtual buffer based bit-rate control and adaptive quantization. In order to understand the structure of AVC and how it performs these main steps, authors in [3] explain the rate control in three levels as: GOP Level rate control, Picture Level rate control and Basic Unit Level rate control (optional).

In the GOP Level rate control, Quantization Parameter (*QP*) is initialized based on the available channel bandwidth and QP for the rest of the pictures in the GOP are calculated according to the formulas described in the technical notes by Joint Video Team *(JVT)* [3]. In the Picture Level rate control, control system consists of pre-encoding and post-encoding steps. We are interested in the pre-encoding stage of the stored pictures where controller performs Rate Distortion Optimization *(RDO)* by setting QPs of each picture. QP value assignment depends on the objective quality of the pictures and Mean Absolute Difference (*MAD*) is used in the MPEG-4 AVC standards. The last level of rate control is the Basic Unit Level rate control where rate control is performed for a group of continuous macro-blocks in addition to GOP and Picture Levels.

In this paper, we use Picture Level rate control for coding *3D* videos. Since depth sequences are not directly presented to the end user, we need to use quality metrics that consider the depth perception of the Human Visual System *(HVS)*. Authors in paper [4] investigate the relationship between quality of synthesized view and the quality of the depth maps. They illustrate that PSNR and SSIM values of depth maps do not correlate well with the rendered view quality and they conclude that these metrics are not suitable for measuring degradation based on compression. In this paper, we compute the factors that can lead to visual discomfort instead of assessing the fidelity in the depth sequences. Objective quality assessment of 3-D videos based on visual discomfort was proposed in [5] as *3VQM*. Authors derive an ideal depth and define the error metric as the absolute value of the difference between the estimated depth and the ideal depth. Using the error metric, temporal and spatial characteristics of the DIBR-based 3-D videos are calculated in the form of temporal outliers, temporal inconsistencies and spatial outliers. Effectiveness of *3VQM* in capturing errors and inconsistencies is evaluated in [6]. Authors validate *3VQM* by showing that it is more accurate, coherent and consistent compared to PSNR and SSIM that are calculated over synthesized views.

Visual Discomfort Metric *(VDM)* is introduced in this paper. It is basically a modified version of *3VQM* to estimate perceived compression errors instead of depth map estimation errors. 3D videos are encoded with *VDM* by analyzing the spatial and temporal characteristics of the depth sequences to make the video coding content-adaptive. The rest of this paper is organized as follows. Section II describes visual discomfort estimation. In Section III, the interaction between encoder and the visual discomfort estimation is explained. Distortion assessment is provided in Section IV and rate-distortion analysis is discussed in Section V. Finally, the concluding remarks are stated in Section VI.

II. VISUAL DISCOMFORT ESTIMATION

Objective quality metrics are commonly used to asses video quality in streaming applications. These metrics are required to be real time and highly correlated with the subjective assessment. Pixel-based quality metrics such as MAD, MSE and PSNR are used because of the simplicity of implementation. However, these pixel-based methods do not correlate well with subjective tests. In the proposed work, we focus on visual discomfort instead of pixel-wise degradations.

Visual Discomfort Metric (*VDM*) is an adaptive version of *3VQM* [5]. *3VQM* was used to calculate depth map estimation errors. In contrast, in this paper, we assume that ground truth depth maps are error free. We estimate compression-based quality degradation by comparing the ground truth and processed depth maps that are used for rendering virtual views. The depth map error definition in *3VQM* is given as:

$$\Delta Z = |Z_{ideal} - Z_{GT}|, \tag{1}$$

where Z_{ideal} is the ideal depth map that is defined in *3VQM* as the depth map that results in distortion-free video and Z_{GT} is the Ground Truth depth map. In *VDM*, the error definition is modified by replacing ideal depth with ground truth depth and ground truth depth with compressed depth as follows:

$$\Delta \hat{Z} = |Z_{GT} - Z_{processed}|, \tag{2}$$

where $\Delta \hat{Z}$ is the absolute value of the difference between the ground truth depth map and the processed depth map and $Z_{processed}$ is the depth map after compression. In the following parts, we define the visual discomfort indexes used in *3VQM* and *VDM*.

Compressing the depth map leads to spatial inconsistencies in the depth values. These discontinuities result in relocated pixels in the synthesized views, which cause visual discomfort. We measure spatial discomfort using the standard deviation of $\Delta \hat{Z}$ and call the quantity as Spatial Outlier *(SO)*:

$$SO = STD(\Delta \hat{Z}). \tag{3}$$

Depth map error patterns can also temporally vary because of the compression artifacts. These artifacts result in impulsive intensity changes around textured regions and flickering around flat regions. Temporal variation of the depth map errors can be modeled by the standard deviation of the difference between the depth map errors in consecutive frames. We define these errors as Temporal Outliers *(TO)* as follows:

$$TO = STD(\Delta \hat{Z}_{t+1} - \Delta \hat{Z}_t), \tag{4}$$

where STD is the standard deviation, $\Delta \hat{Z}_t$ and $\Delta \hat{Z}_{t+1}$ are the predicted depth map error in frame t and $t+1$, respectively. Temporal depth consistency is a significant factor in visual comfort. Temporal depth inconsistency can be measured by quantifying the excessive and fast changing disparities using the standard deviation of the difference of consecutive depth frames. We call this quantity as Temporal Inconsistencies *(TI)* and is given as follows:

$$TI = STD(\hat{Z}_{t+1} - \hat{Z}_t). \tag{5}$$

In *3VQM*, the pooling of these indexes is based on computing the complements of the indexes so that *3VQM* decreases as the image gets distorted. The *SO* index is masked with the logical intersection of *SO* and *TO* indexes to avoid considering visual discomfort sources more than once. Finally, the powers of the discomfort indexes in *3VQM* are determined according to an offline training process and the combination is scaled with a coefficient. *3VQM* formulation is given in [5] as follows:

$$3VQM = K(1 - SO(SO \bigcap TO))^a (1 - TO)^b (1 - TI)^c, \tag{6}$$

where $K = 5.0$, $a = 8.0$, $b = 8.0$, and $c = 6.0$.

In *VDM*, *TI* contributes to the discomfort metric as in *3VQM*. However, power assignment of *TO* and *SO* are modified. Instead of calculating the powers of complement of the metrics, we calculate the power of discomfort metrics and then take the complement. Moreover, power assignment of *TO* and *SO* are content-adaptive in *VDM*. Spatial and temporal information indexes are calculated as defined in [7]. To calculate spatial information index, luminance channel of each frame is filtered with a Sobel operator and then the standard deviation is computed over pixels. This procedure is performed for all frames and the maximum value represents the video sequence. Temporal information index is calculated by taking the difference between consecutive frames, calculating the standard deviation over the pixels for each frame and then selecting the maximum index over time. S_{Inf} stands for the cube root of Spatial Information index and T_{Inf} stands for the cube root of Temporal Information index as shown in Equation (7) and Equation (8), respectively. We use S_{Inf} as the power of *SO* index and T_{Inf} as the power of *TO* index as in Equation (9).

$$S_{Inf} = \sqrt[3]{max_{time}\{std_{space}[Sobel(F_n)]\}}, \tag{7}$$

$$T_{Inf} = \sqrt[3]{max_{time}\{std_{space}[F_n(i,j) - F_{n-1}(i,j)]\}}, \tag{8}$$

where F_n: current frames, F_{n-1}: previous frame, std: standard deviation, max_{time}: max operator that selects the maximum index over time (over all the frames in the video).

The *TI* index can capture the depth map estimation errors as described in *3VQM* [5]. However, when we asses the change in the perceived quality with respect to varying compression errors, *TI* is not highly correlated with the subjective results. As depth videos are quantized more coarsely, depth maps become smoother. Difference of consecutive frames becomes less significant when depth maps are smoothed, which leads to a lower *TI* index. Formulation of *VDM* is given in Equation (9). When we assign a negative value to parameter *c*, *TI* will decrease the value of *VDM* as we quantize the depth videos more coarsely. However, negative powers of the complement of *TI* linearly decreases with the increasing quantization parameter which is correlated with PSNR more than SSIM or perceived quality. Therefore, we exclude *TI* from *VDM* by assigning 0.0 to parameter *c* in Equation (9). Finally, we take the logical intersection of *SO* and *TO* indexes out of the equation since visual discomfort becomes more severe when we have both type of distortions at the same pixel locations. The resulting measure is given as follows:

$$VDM = K(1 - SO^a)(1 - TO^b)(1 - TI)^c, \qquad (9)$$

where $K = 1.0$, $a = S_{Inf}$, $b = T_{Inf}$ and $c = 0.0$.

III. RATE-DISTORTION OPTIMIZATION

The H.264/AVC pipeline is shown in Figure 1 [8]. At the AVC encoder, we have an access to both original and processed (compressed) videos. Therefore, full reference metrics can be used to measure distortion. In the default rate control mechanism of H.264, Mean Absolute Difference (*MAD*) is used as the distortion metric. However, we need to use distortion metrics that consider the structure of the content and perception instead of basic pixel-wise comparisons. Especially when we encode depth sequences instead of color sequences, distortion metrics should correlate with the errors in rendered 3D views. As described in Section II, we use *VDM* to estimate the depth map compression errors. Frames are compressed with the maximum quantization parameter (*QP*) that satisfies minimum *VDM* requirements which is calculated as the mean *VDM* of all the frames when they are quantized with constant QP in the range of 30 to 49. At first we encode depth sequences with constant QPs and obtain a lookup table for minimum *VDM* values for each sequence. Then, we initialize the AVC encoder and compress the depth frames. *VDM* is calculated by using the ground truth frame and the compressed frame, if *VDM* is higher than the threshold we increase the QP for the next frame, otherwise we decrease the QP. We set minimum QP as 30 and maximum QP as 50.

IV. DISTORTION ASSESSMENT

Lossy compression methods lead to artifacts, which result in visual distortions. These distortions distract the users and degrade the quality of experience. We perform compression using ver. 18.5 of H.264/AVC reference software and CABAC is used as the entropy coding method. The video sequences used in this work are obtained from *3DMobile* project video database and they can be sorted as follows: `Balloons`, `Champagne Tower`, `Kendo`, `Lovebirds` and `Pantomime` [9]. In this section, we show how *VDM*, SSIM and PSNR perform under varying compression ratios. *QP* is set to *30*, *35*, *40*, *45* and *49*. Quality metrics are calculated using ground truth and processed depth sequences. In order to represent the metrics at the same plot, PSNR is normalized as shown in Figure 2.

It is possible to recognize the visual degradations in the depth sequences especially when *QP* is set to 45 or 49 as it can be observed for `Kendo` and `Lovebirds` sequences, see Figure 3. Depth maps are not directly viewed by subjects and the quality of these sequences may not highly correlate with the perceived 3D quality. However, structural deformations in the depth maps can still be a good indication of the perceived quality as explained in the rest of the section. For low QP values, visual degradations in the depth maps are not obvious, especially at low resolution. On the contrary, PSNR always shows linear decrease with the increase in QP as it is plotted in Figure 2. Thus, we conclude that PSNR is not highly sensitive to the content and visual degradations in the depth sequences, which is understood in the research community. In contrast, *VDM* and SSIM usually have a slightly steeper decrease in the last two stages. *VDM* and SSIM have similar curves for `Balloons`, `Champagne Tower` and `Pantomime` sequences and they correlate with the visual degradations in the depth sequences. For `Kendo` and `Lovebirds` sequences, we can look at the compressed depth frames in Figure 3 to analyze the behavior of these metrics.

In order to understand the behavior of *VDM* for `Kendo`, we need to consider the temporal and spatial information indexes that are given in Table I. `Kendo` has the second highest spatial information index and highest temporal information index. Visual discomfort metrics vary between 0.0 to 1.0 and when we take higher powers of these metrics, we make *VDM* more sensitive to visual discomforts. Therefore, *VDM* has a steeper decrease for `Kendo` sequence. When QP is increased up to 45 and then to 49, we can observe that `kendo` stick loses its uniformity as it is shown in Figure 3. When the pixels of the same object have different depth values, they will also have different disparity values. Thus, quantization errors such as the ones around kendo stick will cause perceivable visual discomfort that will degrade the quality of experience for the end user. The slope of the SSIM curve slightly decreases when QP is higher than 40 which means SSIM is not highly sensitive to degradations around foreground objects that lead to visual discomfort.

`Lovebirds` depth sequence has the lowest spatial and temporal complexity as in Table I. Therefore, *VDM* is expected to be less sensitive to the visual discomfort. As QP is increased, we can see the blurring artifacts around foreground subjects in Figure 3. If we consider the background of the depth frames more carefully, we can observe blockiness artifacts especially when $QP = 49$. These kind of blockiness artifacts degrade the quality of experience for the end user. Objective metrics are supposed to slightly decrease until QP is set to 45 and they should significantly decrease at $QP = 49$. When PSNR curve is considered, we can see that it linearly decreases as QP is increased and the slope decreases between 45 and 49 which contradicts with the visual degradations. SSIM significantly decreases for most of the QP values whereas it increases when QP is changed from 45 to 49 which negatively correlates

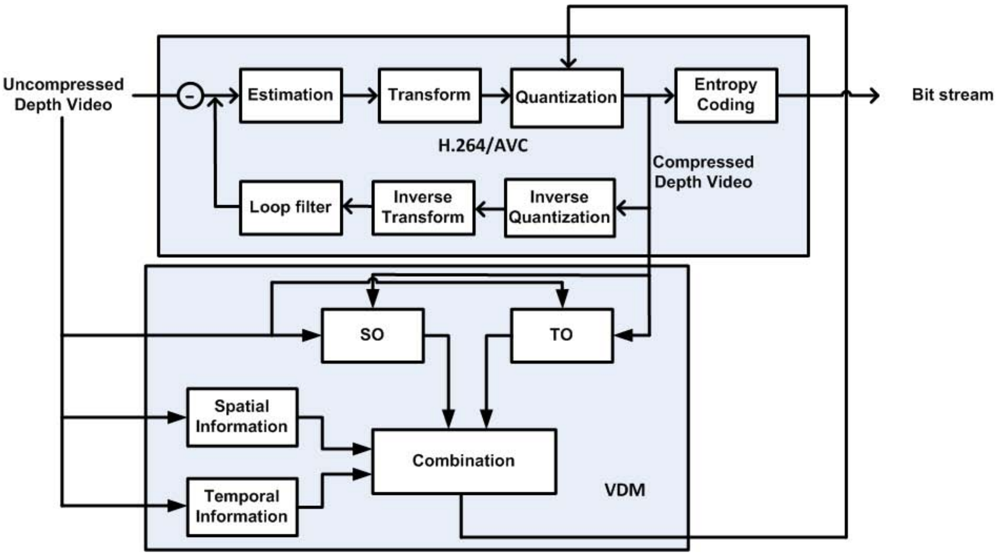


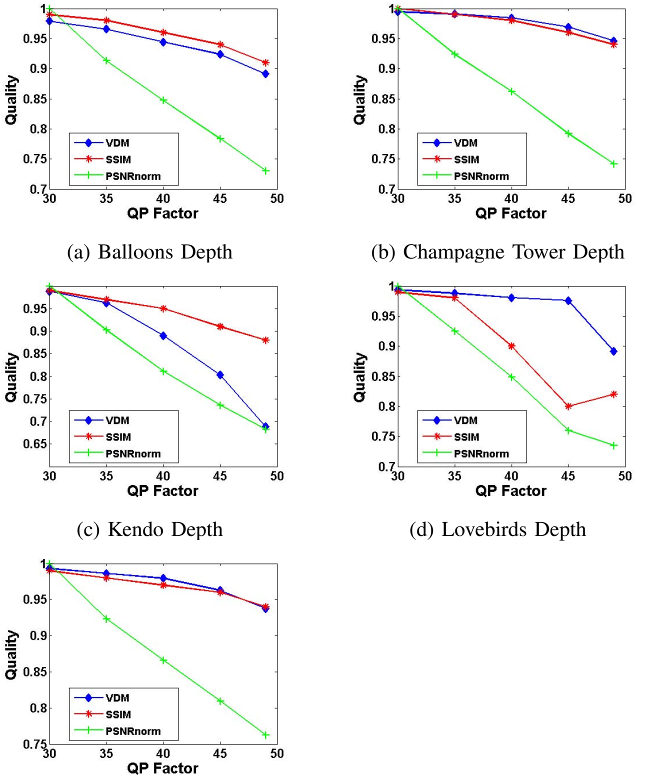
Fig. 1. Rate Distortion Optimization Pipeline

with the visual degradations. In the case of *VDM*, it highly correlates with visual degradations by slightly decreasing until $QP = 45$ and significantly decreasing when QP reaches 49. *VDM* is less sensitive to the visual discomfort compared to other sequences. However *VDM* is still capable of estimating the degradations in the quality of experiences, especially for high QPs in Lovebirds sequence.

VDM is able to estimate the visual discomfort in all of the sequences. However is can be oversensitive if the content is both spatially and temporally complex as we observe in the `Kendo` sequence, especially for the QP values between 35 and 45. SSIM correlates with the expected quality of experience for Balloons, Champagne Tower and Pantomime sequences. However, it is not highly sensitive to degradations in `Kendo` sequence for high QPs. In lovebirds, SSIM is oversensitive to the degradations for low QPs and it is insensitive to the expected degradations for high QPs. PSNR decreases linearly for all of the sequences and it does not highly correlate with the expected quality of experience.

Depth Sequences	TInf Index	SInf Index
Balloons	1.44	2.23
Champagne Tower	1.38	2.11
Kendo	1.88	2.20
Lovebirds	1.23	1.99
Pantomime	1.64	1.99

TABLE I
SPATIAL AND TEMPORAL INFORMATION INDEXES OF DEPTH SEQUENCES



(a) Balloons Depth

(b) Champagne Tower Depth

(c) Kendo Depth

(d) Lovebirds Depth

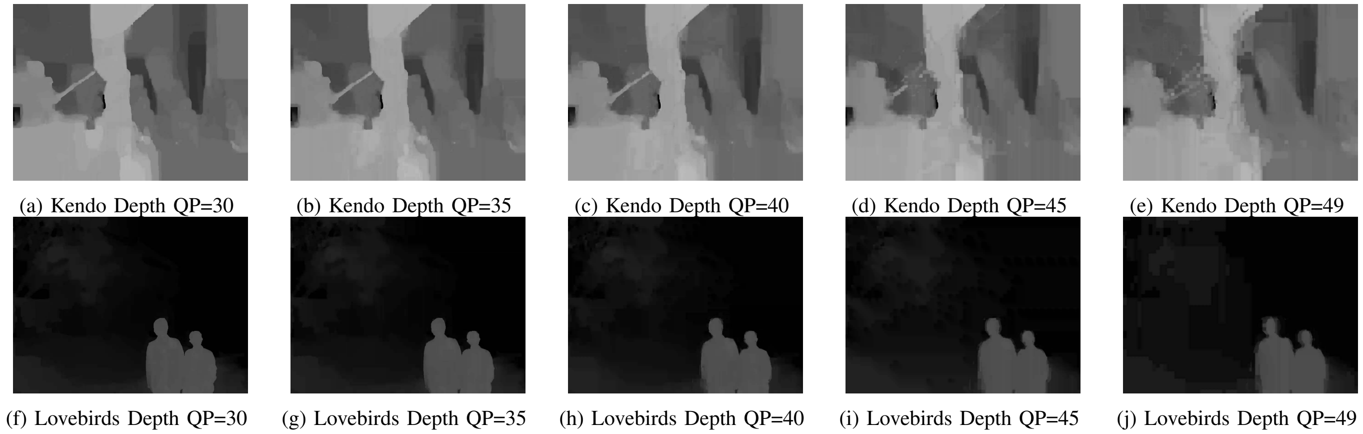
(e) Pantomime Depth

Fig. 2. Objective quality results for compressed (lossy) depth video sequences

V. RATE-DISTORTION EVALUATION

In this paper, we compare three configurations for rate-distortion optimization. At first, we encode the depth sequences without enabling rate-distortion optimization (*WRDO*) and setting quantization parameter to 40. Secondly, we enable standard rate-distortion optimization (*SRDO*) with initial $QP = 40$, minimum $QP = 30$ and maximum $QP = 50$. The distortion metric used in default *RDO* is Mean Absolute Difference (*MAD*), which is a fidelity metric based on pixel-wise differences. In the final case, we perform *VDM* based rate-distortion optimization (*VDM-RDO*), which is explained in Section III. Bit-rate is given in terms of *kbits/sec* and fidelity is calculated with PSNR(dB) and SSIM. Rate-distortion optimization results are summarized in Table II.

The standard rate-distortion optimization (SRDO) enables bit-rate savings for all sequences except `Lovebirds`.


(a) Kendo Depth QP=30 (b) Kendo Depth QP=35 (c) Kendo Depth QP=40 (d) Kendo Depth QP=45 (e) Kendo Depth QP=49

(f) Lovebirds Depth QP=30 (g) Lovebirds Depth QP=35 (h) Lovebirds Depth QP=40 (i) Lovebirds Depth QP=45 (j) Lovebirds Depth QP=49

Fig. 3. Compressed `Kendo` and `Lovebirds` sequences

Whereas, *VDM* based rate-distortion optimization (*VDM*-RDO) results in bit-rate savings for all of the sequences. In terms of bit-rate, *VDM*-RDO outperforms SRDO in all the sequences except `Kendo`. As mentioned in section IV, `Kendo` has the second highest spatial information index and highest temporal information index which means *VDM* is oversensitive for `Kendo` compared to other sequences. Thus, *VDM*-RDO allocates more bits to `Kendo` to avoid visual discomfort. On average, 82.69 $kbits/sec$ is required for WRDO, SRDO results in 54.10 $kbits/sec$ and VMD-RDO in 38.99 $kbits/sec$. As a trade off, rate-distortion optimization leads to lower fidelity metric values. PSNR decreases by 1.77 dB for SRDO and 4.03 dB for *VDM*-RDO. In terms of SSIM, it remains at 0.95 for SRDO and decreases to 0.91 for *VDM*-RDO. The main decrease in SSIM occurs at `Lovebirds` sequence for which SSIM does not highly correlate with visual degradations in depth sequences. PSNR decrease illustrates pixel-wise fidelity degradation and it does not represent perceived quality.

Depth Sequences	WRDO	SRDO	VDM-RDO
Bit-rate (kbits/sec)			
Balloons	93.58	48.79	42.53
Champagne Tower	56.58	50.45	26.45
Kendo	153.84	57.83	67.76
Lovebirds	37.66	49.55	19.47
Pantomime	71.81	48.86	38.75
AVERAGE	**82.69**	**54.10**	**38.99**
PSNR (dB)			
Balloons	36.08	33.51	32.49
Champagne Tower	39.73	38.78	35.45
Kendo	35.11	30.63	30.96
Lovebirds	39.56	41.40	34.92
Pantomime	39.37	36.70	35.90
AVERAGE	**37.97**	**36.20**	**33.94**
SSIM			
Balloons	0.96	0.94	0.93
Champagne Tower	0.98	0.98	0.95
Kendo	0.95	0.89	0.90
Lovebirds	0.90	0.97	0.81
Pantomime	0.97	0.95	0.95
AVERAGE	**0.95**	**0.95**	**0.91**

TABLE II
RATE DISTORTION METRICS CALCULATED OVER 200 FRAMES

VI. CONCLUSION

In this paper, we propose a rate-distortion optimization method for DIBR-based 3D videos. Instead of using fidelity metrics such as PSNR and SSIM, we use content adaptive visual discomfort measure *VDM*. Compared to standard rate-distortion optimization, on average, we can save 15.11 $kbits/sec$ on bit-rate. As a price of bit-rate savings, *VDM* results in 2.26 dB decrease in PSNR and 0.04 is SSIM in terms of image fidelity. The main contribution of the proposed approach is saving from the bit-rate while maintaining the quality of experience level by taking perception into consideration.

REFERENCES

[1] C. Fehn, "Depth-image-based Rendering (DIBR), Compression, And Transmission For A New Approach On 3DTV," *Proc. of SPIE*, vol. 5291, pp. 93–104, 2004.

[2] A. Puri, X. Chen, and A. Luthra, "Video coding using the h.264/mpeg-4 avc compression standard." in *Signal Processing: Image Communication*, vol. 2, 2003, pp. II–892–II–895 vol.2.

[3] J. V. Team, "Text description of joint model reference encoding methods and decoding concealment methods," 2005.

[4] D. V. S. X. De Silva, W. A. C. Fernando, S. Worrall, and A. Kondoz, "A novel depth map quality metric and its usage in depth map coding," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, 2011, pp. 1–4.

[5] M. Solh, G. AlRegib, and J. M. Bauza, "3VQM: A vision-based quality measure for DIBR-based 3D videos," in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, July 2011, pp. 1 –6.

[6] D. Temel and G. AlRegib, "Effectiveness of 3VQM in Capturing Depth Inconsistencies," in *IEEE IVMSP Workshop*, Seoul, Korea, June 10-12 2013.

[7] ITU-T, "P.910: Subjective video quality assessment methods for multimedia applications," in *Tech. Rep.*, 2008.

[8] J. V. Team, "H.264/14496-10 avc reference software manual," 2009.

[9] Mobile 3dtv - 3d video database, http://sp.cs.tut.fi/mobile3dtv/stereo-video/. [Online]. Available: http://sp.cs.tut.fi/mobile3dtv/stereo-video/