RATE ALLOCATION BETWEEN VIEWS IN SCALABLE STEREO VIDEO CODING USING AN OBJECTIVE STEREO VIDEO QUALITY MEASURE

Nukhet Ozbek[1], A. Murat Tekalp[2], and E. Turhan Tunali[1]

[1]Ege University, International Computer Institute
Bornova, Izmir, Turkey, 35100
[2]Koc University, College of Engineering
Sariyer, Istanbul, Turkey, 34450

ABSTRACT

It is well-known that in stereoscopic 3D video systems humans perceive good quality 3D video as long as one of the eyes sees a high quality view. Hence, in stereo video encoding/streaming, best rate allocation between views can be addressed by reduction of the spatial resolution, frame rate, and/or quantization parameter of the second view with respect to the first view. In this paper, we address selection of the rate allocation strategy between views for our recently developed scalable multi-view video codec (SMVC) [1] to obtain the best rate-distortion performance. Since 3D video quality perception does not correlate well with the overall PSNR of the two views, we propose a new quantitative measure for stereo video quality as weighted combination of two PSNR values and a jerkiness measure. The weights are determined by means of correlating subjective quality test results and the objective measure scores on a set of test videos. DSCQS test methodology is used for subjective evaluation of stereo videos. Experimental results are presented to demonstrate how the objective and subjective 3D video quality varies for different choices of rate allocation between the views.

Index Terms − stereo video coding, DSCQS, jerkiness

1. INTRODUCTION

3D video and free viewpoint video are new types of media for next generation broadcast television and streaming applications. MPEG Ad-Hoc Group for 3D Audio and Video (3DAV) is now working on a new standard [2], where new prediction structures as well as processing tools are being investigated for efficient multi-view video coding (MVC). Multi-view video transport over the Internet requires effective rate allocation in which the available bandwidth should be allocated among the views.

The objective of rate allocation between views is to maximize the quality of the final 3D presentation while satisfying various constraints. The most challenging part of quantitative analysis of video adaptation is to define adequate measures or methods for estimating quality. Conventional signal level measures like PSNR need to be modified when video quality is compared at different spatio-temporal resolutions and/or in 3D [3].

In stereoscopic video, an objective quality metric is not commonly used but instead, some subjective quality evaluation methods are utilized. In [4] and [5], the double-stimulus continuous-quality (DSCQ) scale method, which described in ITU-R Recommendation 500, is used to explore the response of the human visual system to mixed-resolution stereo video sequences.

It is well-known from the studies that for appropriate 3D perception from stereo video, the right and left views need not be encoded with full temporal, spatial, and SNR resolutions. This can be used to benefit in effective transport of multiple view video, where one of the views is sent with full resolution, whereas the spatial, temporal and/or SNR resolution of other view(s) can be dynamically adapted according to video content and network conditions [6]. With scalable coding of multi-view video, the encoding can be done once and off-line. In a point-to-point transmission scenario, bitstreams at various spatial, temporal and SNR resolutions can be extracted dynamically on demand. Alternatively, transport of interactive (free-view) 3DTV over IP can be achieved by receiver-driven multicast, where the receiver can subscribe to receive each view at some desired temporal, spatial and/or SNR resolution.

The scalable extension of H.264/AVC is selected as the starting point of the Scalable Video Coding (SVC) work [7]. It specifies temporal scalability by means of a lifting framework on motion-compensated temporal filtering (MCTF). For spatial scalability, a combination of motion-compensated prediction and over-sampled pyramid decomposition is proposed [8]. SNR scalability is achieved by residual quantization with little modification to H.264/AVC. In [1], we introduced the SMVC codec which based on the scalable extension of H.264/AVC (JSVM) and presented coding results that are superior to simulcast scalable coding of multiple views.

This paper presents subjective experimental test results for different scalability options of stereoscopic video coding. A new metric for objective quality measure is proposed so that video adaptation can be performed by optimizing rate-visual distortion. The original and modified PSNR values are demonstrated. Section 2 gives a brief summary of our scalable stereoscopic video codec (SMVC). Section 3 explains the test methodology used in visual quality tests and the new PSNR metric which proposed. Section 4 describes how the jerkiness is measured as a motion content indicator. Section 5 provides experimental results. Conclusions are drawn in Section 6.

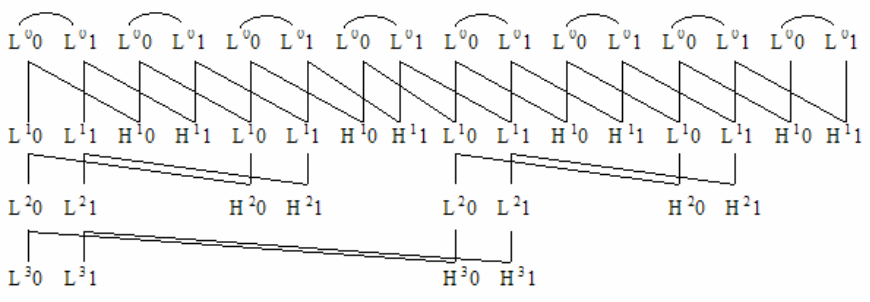

Fig. 1 SMVC prediction structure for N=2 and GOP=16.

2. SCALABLE STEREO VIDEO CODING

We used the SMVC [1] for scalable stereo video coding. The prediction structure of SMVC, which uses multiple reference frames to produce a difference (H) frame, is illustrated in Fig. 1 for the case GOP size is 16, where the first view is only temporally predicted [6]. We implemented SMVC as an extension of the JSVM reference software [9] by sequential interleaving of the first (right) and second (left) views in each GOP. The prediction structure supports adaptive temporal or disparity compensated prediction by using the present SVC MCTF structure without the update steps.

Every frame in left view uses past and future frames from its own view and the same frame from the right view for prediction. In every view, only first frame of the GOP (key frame) use just inter-view prediction so that subscribing to receive any view at some desired temporal resolution can be possible. Fixed QP-setting is used for each view and in between all temporal levels.

The bit stream extractor and decoder modules in the JSVM are modified accordingly in order to recover the last temporal layer as the left view. Since we have two views, the effective GOP size reduces to half the original GOP size shown in Fig. 1, where even and odd numbered Level 0 frames at the decomposition stage correspond to the first and the second view, respectively. Thus, number of temporal scalability levels is decreased by one. However, spatial and SNR scalability functionalities remain unchanged with the proposed structure.

In this study, we extracted right (R) and left (L) bitstreams for each video using the following options:

1) R: 30 Hz, 4SIF, Qp=28; L: 30 Hz, 4SIF, Qp=28
2) R: 30 Hz, 4SIF, Qp=28; L: 15 Hz, 4SIF, Qp=28
3) R: 30 Hz, 4SIF, Qp=28; L: 30 Hz, 4SIF, Qp=34
4) R: 30 Hz, 4SIF, Qp=28; L: 30 Hz, SIF, Qp=28
5) R: 30 Hz, 4SIF, Qp=28; L: 15 Hz, 4SIF, Qp=34
6) R: 30 Hz, 4SIF, Qp=28; L: 15 Hz, SIF, Qp=28
7) R: 30 Hz, 4SIF, Qp=28; L: 30 Hz, SIF, Qp=34

The quality of each scalability option is evaluated as discussed in the next section, where we propose an objective measure to select the best scalability option, hence rate allocation between views, for all videos.

3. STEREO/3D VIDEO QUALITY

In this section, we introduce a new objective measure for 3D/stereo video quality evaluation. This measure has two free parameters, which are determined to match the results of the objective measure with those of subjective evaluation tests described also in this section.

3.1 An Objective Measure for Stereo Video Quality

Since 3D video quality perception does not correlate well with the overall PSNR of the two views, we propose a new quantitative measure for stereo video quality as weighted combination of two PSNR values and a jerkiness measure. We assume that artifacts due to spatial and quality (SNR) are accounted by the PSNR metric. We also assume that the PSNR of the second view is deemed less important for 3D visual experience [4]. However, PSNR alone does not account for motion jitter artifacts sufficiently well. For temporal artifacts *Jerkiness* measure is employed.

The *jerkiness* measure should represent the user perception of the motion in a video, and can be defined as the product of the motion activity and the temporal distance between two consecutive frames [10]. The regional MA_n^r is defined as follows for n-th frame:

$$MA_n^r = \frac{\sum_{i=1}^{N_{NZ}} (|mvx_i| + |mvy_i|) + N_I.P_I}{N_I + N_{NZ}} \qquad (1)$$

where N_I and N_{NZ} are the number of Intra Macroblocks (MBs) and the number of non-zero motion vectors MBs in region r of the n-th frame respectively, P_I is a weight for Intra MBs that set to 128 (the maximum of motion vector for a 64x64 motion search window) and mvx_i and mvy_i are the horizontal and vertical components of the i-th motion vector. We divide a frame spatially into nine regions which include equal number of MBs in each. So, the modified jerkiness definition is as follows

$$J_n = K.Max(MA_n^1,..,MA_n^9).(t_n - t_{n-1}) \qquad (2)$$

where $(t_n - t_{n-1})$ is 1/15 sec. for a 15 fps video. $\boldsymbol{K}$ is a constant to be found experimentally in order to represent motion activity content of the video in an effective way.

Hence, we propose a new frame based quality measure

$$Q_n = \begin{cases} (1-\alpha).PSNR_n^{Right} + \alpha.PSNR_n^{Left} - J_n, & if\ temporal\ scaling \\ (1-\alpha).PSNR_n^{Right} + \alpha.PSNR_n^{Left}, & otherwise \end{cases} \quad (3)$$

It has two free variables, α and K which can be found by *Least Squares Fitting* the MOS (Mean Opinion Score) values of the subjective test results.

Finally, we define a sequence based measure Q by taking average of Q_n values, which can be expressed as

$$Q = \frac{1}{N}\sum_{n=0}^{N-1} Q_n \quad (4)$$

3.2 Subjective Quality Evaluation

For subjective evaluation of the encoded stereo videos, we have adopted DSCQS Test method, implemented under the 3DTV project [3], where non-experts and inexperienced assessors are used. The two videos are evaluated by the assessor on a continuous scale ranging from 0 to 100 with help of two sliders. Multiple assessors are shown two conditions, A and B (two stereoscopic videos), consecutively one of which is always the source and the other is the tested condition applied on the source. The identity of the videos, whether it is the source or the test condition, should be known by the experimenter but not by the assessors. The next pair of conditions is shown after the assessors establish an opinion.

For the analysis of the test results, each evaluation is graded between 0-100 and the difference between the scores of source image and the test condition is calculated to find the score of that test condition on that image by the assessor. After all these scores are calculated, the values are normalized to fit in 0-100. And as a final step, to find the scores of each scaling option (test condition) the average of all the scores over the assessors and images are taken. Scores of the options can be compared with their closeness to the number to which zero score is mapped during the normalization process.

4. EXPERIMENTAL RESULTS

In order to meet time requirements of assessment test, we use only 4 video sets, balloons, flowerpot, soccer2, and train, with 7 scaling options which stated in Table 1. The first three sequences are 720x480 in size and 30 fps. Train is 720x576 in size and 30 fps. All videos are encoded at 240 frames long.

After all the assessors (10 viewers) finish the test, the scores are evaluated and normalized. Average MOS scores and confidence intervals for each option are shown in Table 2. Due to the normalization, 0 (best quality) is mapped to 42, and the success of the options can be measured by closeness of their mean to 42. As it can be seen, the mean of the original video is not exactly 42, which is due to the expected misjudgment of the assessors.

Table 3 and Fig. 6 give the total bitrates (view 0 and view 1) in kbps and modified PSNR (Q) values in dB. The free variables are set as follows: α = 1/3, and K=1/4. It can be seen from the results that scaling in one dimension (options 2, 3, and 4) shows better performance than combined scalability options (options 5, 6, and 7). According to the bitrates and MOS scores, these seven operating points can be reduced to three without loosing visual quality. For instance, when scaling the second view of Balloons sequence, option 3 should be kept whereas 2 and 4 can be skipped. Similarly, among combined scalability options, option 7 should be kept but option 5 and 6 can be skipped. The optimum scaling option among [2,3,4] or [5,6,7] set may vary for different videos depending on motion content and spatial details. However, we note that the smallest bitrate in each set matches the highest Q in each set. Our proposed objective measure enables to select the rate-distortion optimized scaling option for rate allocation among the views.

Since temporal scaling is only applied in option 2, 5, and 6, the jerkiness is accounted just for these scaling options. Fig. 2 to 5 shows the Q_n values of different scaling options so that the effect of jerkiness can be seen in frame basis.

Table 1: Scaling options for the left view.

OPT1	full spatial, full temporal, full SNR	4SIF, 30 Hz, QP=28
OPT2	full spatial, ½ temporal, full SNR	4SIF, 15 Hz, QP=28
OPT3	full spatial, full temporal, base SNR	4SIF, 30 Hz, QP=34
OPT4	base spatial, full temporal, full SNR	SIF, 30 Hz, QP=28
OPT5	full spatial, ½ temporal, base SNR	4SIF, 15 Hz, QP=34
OPT6	base spatial, ½ temporal, full SNR	SIF, 15 Hz, QP=28
OPT7	base spatial, full temporal, base SNR	SIF, 30 Hz, QP=34

Table 2: MOS scores and confidence intervals.

	BALN	FLOW	SOCCER2	TRAIN
ORIG	48.9 ± 0.9	47.8 ± 1.0	45.4 ± 0.9	49.4 ± 1.7
OPT1	48.6 ± 1.4	47.3 ± 2.3	54.0 ± 1.4	48.1 ± 1.3
OPT2	50.0 ± 1.4	58.5 ± 1.8	56.0 ± 1.8	54.8 ± 2.3
OPT3	53.9 ± 1.9	57.8 ± 1.7	65.6 ± 2.6	47.9 ± 2.3
OPT4	54.3 ± 1.7	52.6 ± 1.9	60.3 ± 1.7	47.1 ± 2.2
OPT5	58.8 ± 1.8	61.6 ± 2.3	66.1 ± 1.9	66.9 ± 1.9
OPT6	63.6 ± 1.8	43.8 ± 2.5	70.4 ± 2.3	55.3 ± 1.9
OPT7	60.6 ± 1.7	60.4 ± 1.7	60.5 ± 1.7	46.9 ± 2.0

Table 3: Overall bitrate and Q values for scaling options.

opt	BALN		FLOW		SOCCER2		TRAIN	
1	4232	37.31	3622	35.52	3168	37.35	4425	35.58
2	3515	35.16	3194	34.66	2670	35.4	3735	33.6
3	2668	35.27	2303	33.48	2161	35.4	2716	33.46
4	2878	35.19	2451	33.37	2282	34.96	2977	33.44
5	2497	33.15	2204	32.63	2043	33.53	2561	31.49
6	2628	33.08	2327	32.54	2126	33.1	2738	31.48
7	2428	34.65	2092	32.85	1990	34.47	2451	32.87

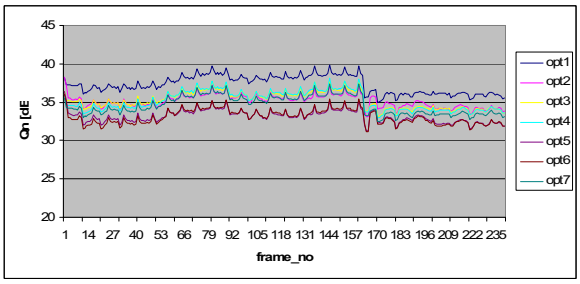

Fig. 2 Q_n values of scaling options for Balloons sequence.

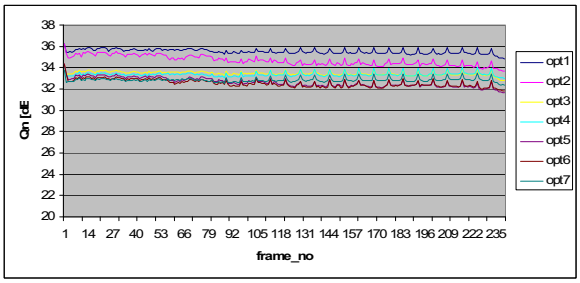

Fig. 3 Q_n values of scaling options for Flowerpot sequence.

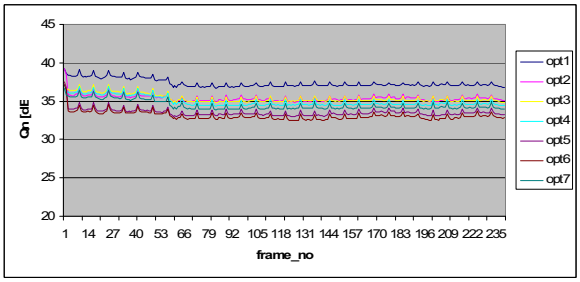

Fig. 4 Q_n values of scaling options for Soccer2 sequence.

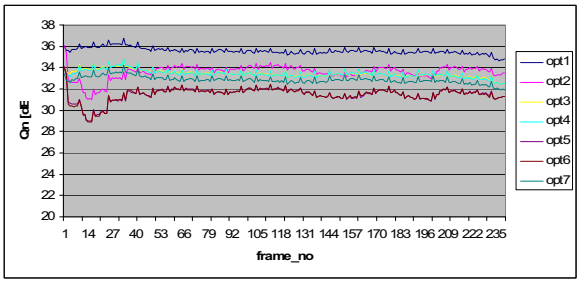

Fig. 5 Q_n values of scaling options for Train sequence.

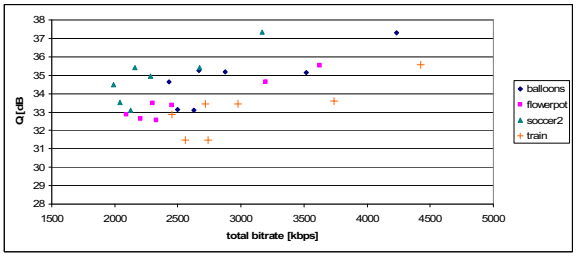

Fig. 6 Overall Rate-Distortion (Q) points for stereo videos.

6. CONCLUSIONS

Experimental results indicate that rate allocation between views by reducing the spatial resolution or increasing Qp value provides a better rate-distortion performance than reducing temporal resolution or any of the mixed scalability options.

By finding appropriate weights for the visual quality components, we accomplish to correlate the subjective test results to the objective quality measures. With the proposed measure, rate adaptation for stereoscopic videos can be performed to optimize rate-visual distortion.

7. ACKNOWLEDGEMENTS

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

The authors would like to thank Burak Gorkemli for his help while conducting subjective tests.

8. REFERENCES

[1] N. Ozbek, and A. M. Tekalp, "Scalable Multi-View Video Coding for Interactive 3DTV", *IEEE Int. Conf. on Multimedia & Expo (ICME)*, Toronto, Canada, July 2006.

[2] Aljoscha Smolic and Peter Kauff, "Interactive 3-D Video Representation and Coding Technologies," *Proceedings of the IEEE*, vol. 93, No. 1, January 2005.

[3] S. F. Chang, and A. Vetro, "Video Adaptation: Concepts, Technologies, and Open Issues", *Proceedings of the IEEE,* vol. 93, no.1, pp. 148-158, January 2005.

[4] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo Image Quality: Effects of Mixed Spatio-Temporal Resolution", *IEEE Transactions on Circuits and Systems for Video Technology*, vol 10, no. 2, pp. 188-193, 2000.

[5] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. B. Akar, M. R. Civanlar, A. M. Tekalp, "Temporal and Spatial Scaling for Stereoscopic Video Compression", *EUSIPCO 2006*, Florence, Italy, September 2006.

[6] N. Ozbek, and A. M. Tekalp, "Content-Aware Bit Allocation in Scalable Multi-View Video Coding", International Workshop on Multimedia Content Representation, Classification and Security, *LNCS 4105*, pp. 691-698, 2006.

[7] J. Reichel, H. Schwarz, M. Wien (eds.), "Scalable Video Coding – Working Draft 1," Joint Video Team (JVT), *Doc. JVT-N020*, Hong-Kong, China, Jan. 2005.

[8] H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "Combined Scalability Support for the Scalable Extension of H.264/AVC", *IEEE International Conference on Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, July 2005.

[9] J. Reichel, H. Schwarz, M. Wien, "Joint Scalable Video Model JSVM-4", *Doc. JVT-Q202*, Oct. 2005.

[10] G. Iacovoni, S. Morsa, and R. Felice, "Quality-Temporal Transcoder Driven by the Jerkiness", *IEEE Int. Conf. on Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, July 2005.