JOINT ADAPTATION OF SPATIAL RESOLUTION AND SAMPLE VALUE QUANTIZATION FOR ASYMMETRIC STEREOSCOPIC VIDEO COMPRESSION: A SUBJECTIVE STUDY

Payman Aflaki[a], Miska M. Hannuksela[b], Jussi Hakala[c], Jukka Häkkinen[b,c], Moncef Gabbouj[a]

[a]Department of Signal Processing, Tampere University of Technology, Tampere, Finland;
[b]Nokia Research Center, Tampere, Finland;
[c]Dept. of Media Technology, Aalto University, School of Science and Technology, Espoo, Finland

ABSTRACT

A novel asymmetric stereoscopic video coding method is presented in this paper. The proposed coding method is based on uneven sample domain quantization for different views and is typically applied together with a reduction of spatial resolution for one of the views. Any transform-based video compression, such as the Advanced Video Coding (H.264/AVC) standard, can be used with the proposed method. We investigate whether the binocular vision masks the coded views of different types of degradations caused by the proposed method. The paper presents a subjective viewing study, where the proposed compression method is compared with two other coding techniques: full-resolution symmetric and mixed-resolution stereoscopic video coding. We show that the average subjective viewing experience ratings of the proposed method are higher than those of the other tested methods in six out of eight test cases.

Index Terms— Low bit-rate video coding, quantization, downsampling, asymmetric stereoscopic video, subjective assessment.

1. INTRODUCTION

Asymmetric stereoscopic video is one division of ongoing research for compression improvement in stereoscopic video, where one of the views is sent with high quality, whereas the other view is degraded and hence the bitrate is reduced accordingly. This technique is based on the psycho-visual studies of stereoscopic vision in human visual system (HVS) which demonstrated that the lower quality in a degraded view presented to one eye is masked by the higher quality view presented to the other eye, without affecting the visual perceived quality (binocular suppression theory [1]). The quality difference between the views of a stereoscopic video is commonly achieved by removing spatial, frequency, and temporal redundancies in one view more that in the other. Different types of prediction and quantization of transform-domain prediction residuals are jointly used in many video coding standards. In addition, as coding schemes have a practical limit in the redundancy that can be removed, spatial and temporal sampling frequency as well as the bit depth of samples can be selected in such a manner that the subjective quality is degraded as little as possible.

In [2], a set of subjective tests on a 24" polarized stereoscopic display comparing symmetric full-resolution, quality-asymmetric full-resolution, and mixed-resolution stereoscopic video coding were presented. The performance of symmetric and quality-asymmetric full-resolution bitstreams was approximately equal. The results showed that in most cases, resolution-asymmetric stereo video with a downsampling ratio of 1/2 along both coordinate axes provided similar quality as symmetric and quality-asymmetric full-resolution stereo video. These results were achieved under the same bitrate constraint.

Objective quality metrics are often able to provide a close approximation of the perceived quality for single-view video. However, in the case of asymmetric stereoscopic video, there are two views with different qualities, and it has been found that objective quality assessment metrics face some ambiguity on how to approximate the perceived quality of asymmetric stereoscopic video [3].

In this paper, we propose a novel compression method for one view of stereoscopic video coding, while the other view is coded conventionally. Our aim is to study the proposed method for asymmetric stereoscopic video due to the fact that it introduces different compression artifacts than those of conventional coding methods and hence the human visual system might mask the coding errors of one view by the other view. Consequently, this paper verifies the assumption that binocular suppression is capable of masking the proposed uneven sample-domain quantization with a systematic subjective comparison of the proposed method with two other compression techniques, namely symmetric and mixed-resolution stereoscopic video coding.

This paper is organized as follows. Section 2 presents the proposed compression method. The test setup and test material are described in Section 3, while Section 4 provides the results. Finally, the paper concludes in Section 5.

2. PROPOSED COMPRESSION METHOD

2.1 Overview

The proposed encoding approach is depicted in Fig. 1. While the proposed method is applied to the right view in Fig. 1, it can equally be applied to the left view. The proposed coding method consists of the transform-based encoding step for the left view and three steps for the right view: downsampling, quantization of the sample values, and transform-based coding. First, the spatial resolution of the image is reduced by downsampling. The lower spatial resolution makes it possible to use a smaller quantization step in transform coding and hence improves the subjective quality compared to a coding scheme without downsampling. Moreover, downsampling also reduces the computational and memory

resource demands in the subsequent steps. Second, the number of quantization levels for the sample values is reduced using a tone mapping function. Third, transform-based coding, such as H.264/AVC encoding, is applied.

The decoding end consists of the transform-based decoding step for the right view and three respective steps for the left view: transform-based decoding, inverse quantization of sample values, and upsampling. In the first step, the bitstream including coded transform-domain coefficients is decoded to a sample-domain picture. Then, the sample values are rescaled to the original value range. Finally, the image is upsampled to the original resolution i.e. the same resolution as of the left view or to the resolution used for displaying.

In the following sub-section, the key novel parts of the proposed coding scheme, namely the quantization of the sample values in the encoder and their inverse quantization in the decoder are described in details.

2.2 Quantization and inverse quantization of sample values

This step of the proposed compression method reduces the number of quantization levels for luma samples. In addition, the original luma sample values are remapped to a compressed range. Hence, the contrast of the input images for transform-based coding and the output images from transform-based decoding is smaller compared to the contrast of the respective original images. The remapping to a compressed value range is typically done towards the zero level, and hence the brightness of the processed images is reduced too.

The proposed method includes the following key steps:

1) Before transform-based encoding: reduction of the number of luma quantization levels in the sample domain and scaling of luma sample values to a compressed value range.
2) After transform-based decoding: Re-scaling of the decoded sample values in such a way that the original sample value range of the luma sample values is restored.

When the same quantization step size is used for transform coefficients in transform-based encoding, the bitrate of the video where sample values are quantized becomes smaller than that of the same video without sample value quantization. This reduction in bitrate depends on the ratio of the number of luma quantization levels divided by the original number of luma quantization levels, which typically depends on the bit depth. Ratios closer to 0 have very good compression outcome but the quality drop is severe. On the other hand, applying a ratio close to 1 keeps the quality close to the original quality with a smaller relative bitrate reduction. We found ratios greater than or equal to 0.5 to be practical.

The presented sample value quantization operation is lossy, i.e., it cannot be perfectly inverted, when integer pixel values are in use. Hence, the original pixel values can be only approximately restored by the inverse quantization of sample values.

Based on informal subjective results, the sample value quantization is proposed to be applied only to the luma component. This is because the bitrate saving achieved by quantization of the two chroma components caused a more severe subjective quality reduction than the same bitrate saving achieved by quantizing the luma component more coarsely.

The quantization of sample values can be done in various ways. For example, tone mapping techniques can be exploited [4]. In this paper, linear luma value quantization with rounding is used as expressed as:

$$q = round\left(\frac{i*w}{2^d}\right) = (i * w + 2^{d-1}) \gg d \quad (1)$$

where:

q is the quantized sample value
round is a function returning the closest integer
i is the input value of the luma sample
w is the explicit integer weight ranging from 1 to 127
d is the base 2 logarithm of the denominator for weighting

Since Eq. (1) is implemented using integer multiplication, addition, and bit shifting, it is computationally fast. As the sample value range is reduced, the value of w is required to be smaller than 2^d. With this limitation, Eq. (1) is identical to the formula used for H.264/AVC weighted prediction. The ratio ($w / 2^d$) is referred to as the luma value quantization ratio.

Inverse quantization of sample values to their original value range is achieved by:

$$r = round(q' * \frac{2^d}{w}) \quad (2)$$

where:

r is the inverse-quantized output value
q' is the scaled value of the luma sample as output by the transform-based decoder
Other parameters are the same values as used in the sample value quantization.

Eq. (2) requires one floating or fixed point multiplication and a conversion of the floating or fixed point result to integer by rounding. If it is preferred to use integer arithmetic in the decoder rather than in the encoder, it is possible to apply Eq. (2) in the encoder and Eq. (1) in the decoder with the condition that w is greater than 2^d.

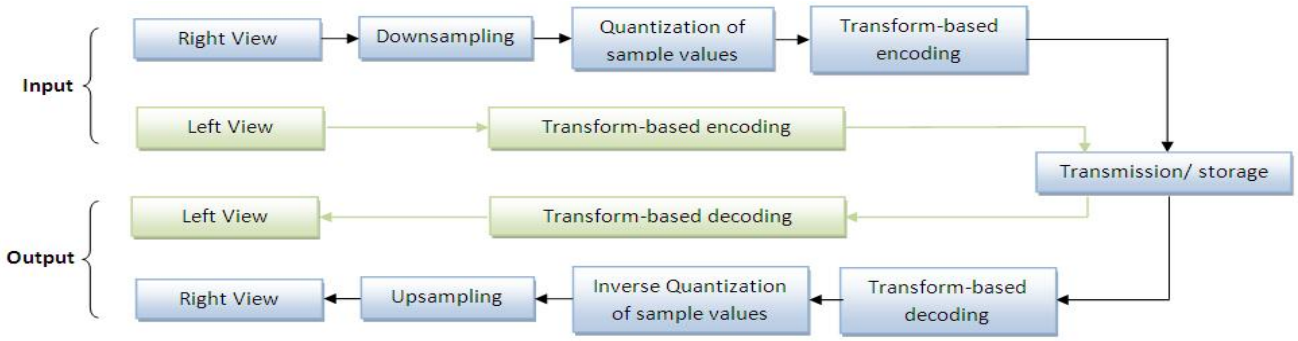

Fig. 1. Diagram of proposed compression method

3. TEST SETUP

3.1 Preparation of test stimuli

The subjective assessments were performed with four sequences: Undo dancer, Kendo, Newspaper, and Pantomime. Undo dancer, exemplified in Fig. 2, is a synthetically created photorealistic multiview sequence including a dancing person, reproduced from a motion capture. The other three sequences are common test sequences in the 3D Video (3DV) ad-hoc group of the Moving Picture Expert Group (MPEG). The sequences were downsampled from their original resolutions to the resolutions mentioned in Table 1 in order to be displayed on the used screen without scaling (see Section 3.2). The filters included in the JSVM reference software of the Scalable Video Coding standard were used in this and other subsequent downsampling and upsampling operations [5].

For each sequence, we had the possibility to choose between several camera separations or view selections. This was studied first in a pilot test of 9 subjects. The test procedure of the pilot test was similar to that of the actual test presented in Section 3.2. Several camera views were available for each sequence in the pilot test, and based on the subjective scores achieved, the 4 cm and 5 cm camera separations were chosen for Undo dancer and the rest of test sequences, respectively.

Several bitstreams were coded for each sequence with the following coding methods:

1. Full resolution symmetric stereoscopic video by coding to both views with H.264/AVC. No downsampling or quantization of luma sample values.
2. Mixed resolution stereoscopic video by downsampling the right view and subsequently applying H.264/AVC coding to it while coding the full-resolution left view with H.264/AVC.
3. The proposed coding scheme including downsampling, quantization of luma sample values, and H.264/AVC coding to the right view and coding the left view with H.264/AVC.

The coded left view for each sequence was identical regardless of the coding method. The left view was kept unchanged, because we wanted to assess the perception and acceptability of the left and right eyes presented with different types of quality degradations as caused by transform-domain quantization, spatial downsampling, and sample-domain quantization and to reduce the number of factors which could affect the subjective rating. Joint optimization over both views for the quantization step size for sample values and transform coefficients as well as for the spatial resolution was left to another subjective experiment. As the bitrate of the right view for each bitstream of the same sequence was kept the same regardless of the coding method used, there was a fair comparison between the coding methods.

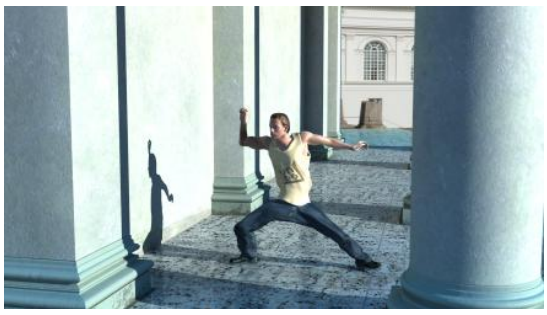

Fig. 2. A frame of Undo dancer sequence

Table 1. Spatial resolution of the right view

	Full	5/6	3/4	1/2
Undo dancer	960x576	800x480	720x432	480x288
Others	768x576	640x480	576x432	384x288

In order to have a representative set of options for the second coding method (with downsampling and transform-based coding), three bitstreams per sequence were generated, each processed with a different downsampling ratio for the right view. The subjective results achieved for stereoscopic video in [2] motivated us to use downsampling ratios equal to or greater than 1/2. Hence, downsampling was applied to obtain a spatial resolution of 1/2, 3/4, and 5/6 relative to the original resolution along both coordinate axes. Table 1 presents the spatial resolution of the right view used for different sequences.

As the number of potentially suitable combinations for the downsampling ratio and the luma value quantization ratio is large, their joint impact on the subjective quality was studied first to select particular values for the downsampling ratio and the luma value quantization ratio for the subsequent comparisons between the different coding methods. To reveal potential dependencies at different quantization step sizes for transform coefficients, the bitstreams were generated with several quantization parameter values. Subjective assessment revealed that downsampling ratio 3/4 along with luma value quantization ratio 5/8 tended to provide the best relative subjective results. Thus, these values were consistently used in the subsequent comparisons.

In order to prevent fatigue of test subjects from affecting the test results, only two sets of bitstreams at different bitrates were included in the test. Table 2 presents the selected Quantization Parameter (QP) values for the full-resolution symmetric coding, the resulting bitrates, and the respective average luma PSNR values for the right view of each sequence coded using different coding methods. The PSNR values were derived from the decoded sequences after inverse quantization of sample values and upsampling to the full resolution.

Table 2. Tested bitrates per view, respective QP values per sequence for both higher quality (HQ) and lower quality (LQ), and the respective PSNR values for different coding techniques

		Pantomime	Dancer	Kendo	Newspaper
QP	HQ	41	42	43	42
	LQ	44	45	45	45
Bitrate (Kbps)		445.8 343.9	301.5 224.6	280.3 238.5	148.0 115.4
Proposed (PSNR-dB)		31.9 30.6	29.1 28.3	34.1 33.1	30.7 29.5
FR (PSNR)		31.9 30.0	29.2 27.7	33.3 32.0	30.0 28.3
1/2 (PSNR)		31.7 30.9	29.1 28.3	35.5 34.7	31.7 30.7
3/4 (PSNR)		32.5 31.0	29.5 28.5	34.7 33.5	31.3 29.8
5/6 (PSNR)		32.3 31.0	29.8 28.3	34.1 32.9	29.9 29.2

3.2 Test Procedure

12 subjects participated in this experiment of which 7 were women and 5 men. Their age differed from 19 to 32 years with an average of 23.6 years. The candidates were subject to thorough vision screening. Candidates who did not pass the criterion of 20/40 visual acuity with each eye were rejected. All participants had a stereoscopic acuity of 60 arc sec or better. Test clips were displayed on a 24" polarizing stereoscopic screen having the total resolution of 1920×1200 pixels and the resolution of 1920×600 per view when used in the stereoscopic mode. The viewing conditions were kept constant throughout the experiment and in accordance with the sRGB standard [6] ambient white point of D50 and illuminance level of about 200 lux. Viewing distance was set to 93cm which is 3 times the height of the image, as used in some subjective test standards [7].

The subjective test started with a combination of anchoring and training. The extremes of the quality range of the stimuli were shown to familiarize the participants with the test task, the test sequences, and the variation in quality they could expect in the actual tests that followed. The test clips were presented one at a time in a random order and appeared twice in the test session. Each clip was rated independently after its presentation. A scale from 0 to 5 with a step size of 0.5 was used for the rating. The viewers were instructed that 0 means "very bad" or "not natural" and 5 stands for "very good" or "very natural".

4. RESULTS AND DISCUSSIONS

Fig. 3 shows the viewing experience subjective results for all sequences in two different bitrates. Based on the average subjective ratings, it can be seen that the proposed coding method outperformed the other tested coding methods in all cases for the higher bitrate. Furthermore, except for the Dancer sequence, it had similar performance than the best mixed-resolution test case in the lower bitrate. The mixed-resolution coding with 5/6 spatial resolution in the lower quality view outperformed the proposed method for the Dancer sequence at the lower bitrate, while the performance of the proposed method was better than or similar to the performance of the other methods. Moreover, the symmetric full-resolution coding method was clearly inferior to the other tested methods at the lower bitrate.

When comparing the PSNR values presented in Table 2 with the subjective viewing experience results, one can see that PSNR was not representative of the subjective quality in this test.

5. CONCLUSIONS AND FUTURE WORK

A novel asymmetric stereoscopic video coding technique was introduced in this paper. The method is based on uneven quantization step size for luma sample values of different views, and it is typically jointly applied with downsampling. The proposed compression method was subjectively compared to full-resolution symmetric stereoscopic video coding and mixed-resolution stereoscopic video coding at different downsampling ratios. The average subjective viewing experience ratings of the proposed method were found to be higher than those of the other tested methods in six out of eight test cases. The results suggest that the human visual system is able to fuse views with different types of quality degradations caused by the proposed method. The provided results should be verified with a greater number of test sequences and more subjective tests to verify these conclusions.

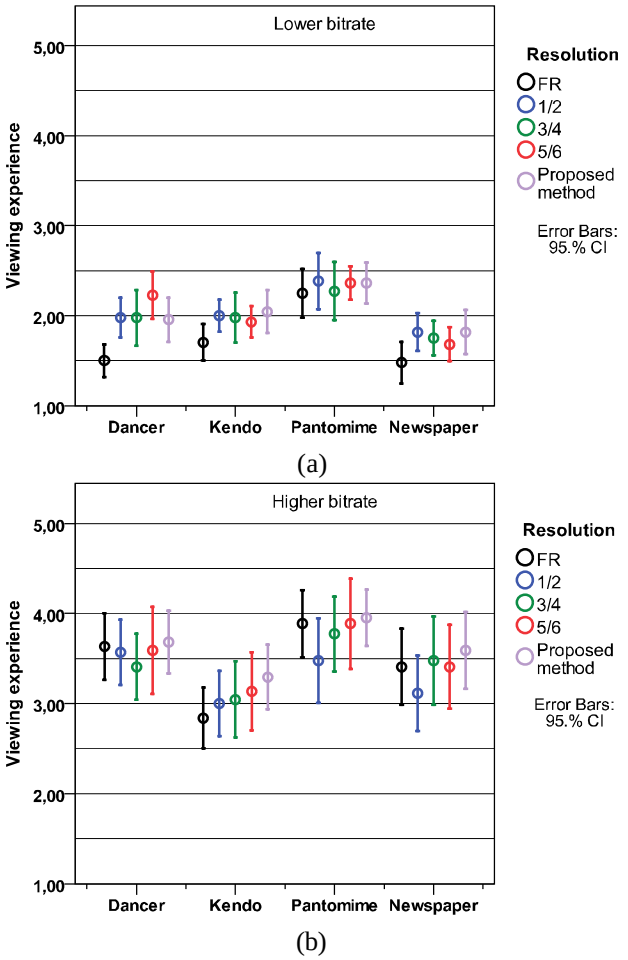

Fig. 3. Results of subjective tests for all sequences using two bitrates (a) lower bitrate (b) higher bitrate

6. REFERENCES

[1] R. Blake, "Threshold conditions for binocular rivalry," Journal of Experimental Psychology: Human Perception and Performance, vol. 3(2), pp. 251-257, 2001.

[2] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, "Subjective Study on Compressed Asymmetric Stereoscopic Video," Proc. of IEEE Int. Conf. on Image Processing (ICIP), Sep. 2010.

[3] P.W. Gorley, N.S. Holliman; "Stereoscopic image quality metrics and compression", Stereoscopic Displays and Virtual Reality Systems XIX, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol.6803, January 2008

[4] A. Segall, L. Kerofsky, S. Lei, "New Results with the Tone Mapping SEI Message," Joint Video Team, Doc. JVT-U041, Hangzhou, China, October 2006.

[5] JSVM Software http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm

[6] B. Girod and N. Färber, "Wireless video," Chapter 12 in the book "Compressed video over networks," Marcel Dekker, 2000.

[7] ITU-T, Subjective assessment methods for image quality in high-definition television, ITU-R BT.710-4