FREQUENCY BASED ADAPTIVE SPATIAL RESOLUTION SELECTION FOR 3D VIDEO CODING

Payman Aflaki[a], *Dmytro Rusanovskyy*[b], *Miska M. Hannuksela*[b], *Moncef Gabbouj*[a]

[a]Department of Signal Processing, Tampere University of Technology, Tampere, Finland;
[b]Nokia Research Center, Tampere, Finland;

ABSTRACT

Downsampling applied to texture views prior to encoding can increase the subjective quality of decoded video. In our study, we show that spatial resolution selection based on traditional pixel-based distortion metrics, such as Mean Square Error (MSE) is weakly correlated with the resolution selection based on subjective quality of coded video. To overcome this problem, we propose a novel frequency-based distortion metric which is shown to resemble subjective quality of coded video more accurately compared to conventionally used MSE-based metric.

Index Terms— MVC, resolution adjustment, objective quality metrics, subjective assessment, frequency power spectrum

1. INTRODUCTION

3D video coding standardization is a recent activity targeting at enabling a variety of display types, including autostereoscopic multiview displays and stereoscopic displays, as well as user-adjustable depth perception. To enable this functionality, multiple high-quality views shall be available in decoder/display side. Due to the natural limitations of content production and content distribution technologies, capturing and delivery of a large number of high-quality views to user side is a very challenging task under the current video coding technologies. To assess available solutions for this challenge, the Moving Picture Experts Group (MPEG) issued a Call for Proposals for 3D video coding technologies (hereafter referred to as the 3DV CfP) [1] which enables rendering of a selectable number of views within a certain viewing range without increasing the required bitrate compared to conventional bandwidth. More than 20 solutions were submitted to the CfP and they were evaluated through a rigorous formal subjective quality assessment performed by the MPEG and its partners.

A significant number of responses to the CfP utilized the Multiview Video plus Depth (MVD) data format and were based on the H.264/MVC video coding standard [2]. The MVD data format consists of natural texture image and its associated depth map data image. The use of MVD data format and Depth-Image-Based Rendering (DIBR) algorithms [3] at the decoder side allows rendering required number of intermediate views from limited input views. However these views (both texture and depth) should be encoded and transmitted to the decoder.

The H.264/MVC is the state-of-the-art coding standard in the field of multiview video coding (MVC) which utilizes inter-view and temporal redundancies in multiview data. However, the resulting bitrate of MVC coded MVD data (texture and depth views) exceeds the bandwidth reserved for conventional 2D video services. As a result, significant research was done to further decrease the bitrate while preserving subjective quality of decoded views and preserving the compatibility with existing H.264/MVC video coding technology.

Adaptive spatial resolution adjustment for coded video data is a potential approach to decrease the bitrate. If the same encoding parameters are utilized, downsampling of video data prior to encoding leads to bitrate reduction. In this design, the overall system distortion is a combination of conventional coding distortion and reduction of high frequency components due to low pass filtering introduced by downsampling. The video coding system should be designed properly to balance these distortions in order to achieve a subjectively superior visual quality of decoded video data.

The spatial downsampling proposed in [4, 5] improves compression at low bitrates. An adaptive decision is made for appropriate downsampling and quantization mode according to local visual significance. The downsampling ratio is automatically adjusted from 1/4 to 1 according to local image contents. Authors in [6] proposed an adaptive downscaling ratio decision approach for better compression of multiview video. The proposed method is based on a trade-off between the distortion introduced by downsampling and distortion introduced by quantization. The results indicated that using bit-rate adaptive mixed spatial resolution coding for both views and depth maps can achieve savings in bit-rate, compared to Full Resolution (FR) multiview coding when the quality of synthesized views is considered. In [7] authors utilized adaptive downsampling to improve performance of H.246/AVC video coding. In this work, it is proposed to optimize the spatial resolution through a rate distortion optimization, where distortion of downsampling and coding processes were averaged.

In this paper, we perform a set of subjective tests showing that MSE-based resolution selection cannot

estimate the subjective results accurately. Hence, a novel algorithm for adaptive spatial resolution selection based on frequency-based distortion metric is presented. Results prove that this method is capable of better estimating the subjective quality comparing MSE-based approach.

The rest of paper is organized as follows. Section 2 describes proposed methods. The test material, setup, and results are presented in sections 3 and 4 for objective and subjective experiments, respectively. Section 5 discusses the results. Finally, conclusions are given in Section 6.

2. PROPOSED SPATIAL RESOLUTION SELECTION METHODS

The level of distortions introduced by lossy video coding systems is typically controlled by a Quantization Parameter (QP) where higher QP corresponds with low bitrates but higher coding distortions. In the case of multi-resolution encoding and under constrained bitrate, different QP values are associated with selected resolutions. This association between different resolutions and QPs, providing a same target bitrate, can be estimated in advance and specified to the encoder through a properly designed lookup table. In such design, video data at lower spatial resolution can be coded at a lower QP under the same bitrate constrain and less coding distortions are introduced to coded data. However, process of resolution rescaling introduces its own distortion through a low pass filtering of input data. To solve this rate-distortion optimization problem, encoder should take both of these distortions in consideration. In this paper we present two methods for encoder to make decision on the spatial resolution of texture data under constrained bitrate:

1) Mean Square Error based method
2) Frequency Power Spectrum based method

The two proposed methods are described in detail in following sub-sections.

2.1 Pixel-based distortion metric

In this method the MSE over FR encoded image is calculated against the original image. For downsampled schemes, the encoded image with different downsampling ratios is upsampled to FR and then the MSE is calculated against the original. Considering that under the same bitrate constrain, different resolutions provide different MSE values, therefore, the resolution providing the least MSE value will be selected as the candidate which should be utilized for encoding. In this step, we consider a fine interval for MSE values in which a lower resolution is preferred. In other words, if the MSE value of the lower resolution is in a predefined and fixed interval of MSE values of a higher resolution, the lower resolution will be selected. Selecting a lower resolution favors a lower computational complexity.

2.2 Frequency-based distortion metric

Our approach is based on the assumption that image quality degradation caused by downsampling and coding can be better evaluated in frequency domain, rather than in the pixel domain. Since downsampling and block-based coding with scalar quantization are both suppressing high frequencies, we can evaluate introduced degradation throw analysis of high frequency components of 2D Discrete Cosine Transform (DCT) spectrum.

Let us introduce the following notation: $F\{\ \}$ as a separable 2D forward DCT and $w=(w1,w2)$ as coordinates of DCT coefficients. 2D DCT being performed over the whole image s size of M×N results in 2D DCT spectrum of the same size M×N:

$$S(w1,w2)=F\{s(x,y)\}, \qquad (1)$$

where F is the function performing the 2D DCT transfer while $x=0,...M-1$ and $y=0...N-1$ are spatial coordinates of the image s , and $w1=0,...M-1$ and $w2=0,..N-1$ are coordinates in the 2D DCT spectrum S.

Transform coefficient which are located in the right-bottom section of spectral image are associated with high frequency components (**HFC**) of image I and we select these information for further analysis as follows:

$$\begin{aligned}&\mathbf{HFC}(s)=S(w1,w2)\\ &w1=T1..M-1\\ &w2=T2..N-1\end{aligned}, \qquad (2)$$

where terms T1 and T2 are boundaries that specify **HFC** in horizontal and vertical directions of 2D DCT spectrum, respectively.

In our method we compare **HFC** of 2D DCT coefficients computed for the following image:

- *UF*: Uncompressed image at the Full Resolution
- *CF*: Compressed image at the Full Resolution
- *CL*: Compressed image at Low Resolution

Note that original image is downsampled prior to encoding and upsampled to FR after decoding to produce *CL*.

Each of these images undergo 2D DCT and **HFC** coefficients for *CF* and *CL* spectral images are compared against the ***HFC*** of the UF image:

$$\begin{aligned}dCF(w1,w2)&=\mathbf{HFC}(UF(w1,w2))-\mathbf{HFC}(CF(w1,w2)) \qquad (3)\\ dCL(w1,w2)&=\mathbf{HFC}(UF(w1,w2))-\mathbf{HFC}(CL(w1,w2))\end{aligned}$$

The differential spectral images *dCF* and *dCL* are computed coefficient-wise for all transform coefficients that belong to the specified **HFC**. Since transform coefficients of *dCF* and *dCL* are computed over the entire image *s*, a large number of them would have magnitude close to zero. These coefficients would not reflect noticeable components of the image *s*, but their cumulative magnitude might affect the decision making. In order to avoid this, we filter *dCF* and

dCL with commonly used in transform-based filtering hard-thresholding [8]. This non-linear filtering operation $T\{\ \}$ is performed over each transform coefficient of *dCF* and *dCL* as following:

$$T\{Y(w)\} = \begin{cases} Y(w), & |Y(w)| \geq T3 \\ 0, & \text{else} \end{cases}, \quad (4)$$

where *Y(w)* is original transform coefficient, and *T(Y(w))* filtered transform coefficient and *T3* is a threshold specifying an expected level of the noise present in the current image.

Following the filtering, we compute arithmetic mean of transform coefficients presented in *T(dCF)* and *T(dCL)* and utilize this value as a distortion metric.

$$\text{cost}(CF) = \frac{1}{n} \cdot \sum_{w2=T2}^{N-1} \sum_{w1=T1}^{M-1} T(dCF(w1, w2)) \quad (5)$$

$$\text{cost}(CL) = \frac{1}{n} \cdot \sum_{w2=T2}^{N-1} \sum_{w1=T1}^{M-1} T(dCL(w1, w2)) \quad (6)$$

where term n determines number of samples within **HFC** and computed as: $n = (N - T2 - 1) \cdot (M - T1 - 1)$

Optimal resolution for coded image is selected as such that provide minimal cost of the metric presented in (5):

$$resolution = \arg\min_{\text{cost}} (\text{cost}(CF), \text{cost}(CL)) \quad (7)$$

3. OBJECTIVE EXPERIMENTS

3.1 Test material and setup

Test sequences and input views utilized in this study are the same as specified in 3DV CfP for case C2 [1]. Modified JM 17.2 reference software [10] with extended multiview profile was utilized for encoding multiview texture data. Four Rate Points (RP) specified in the CfP were utilized for the encoding procedure.

The content of the sequences remains relatively similar, hence, only the statistics of the first frame are utilized in this study. However, this method can be easily extended to be utilized at Group of Picture (GOP) levels or scene cuts. If the codec supports the change on spatial resolution of frames through the encoding process, it might be subjectively beneficial to utilize the proposed method in scene cuts. Utilization of first frame statistics is due to similar content of each sequence and controlling the increase of complexity.

Considering that constant QP settings were required in the CfP, a target bitrate was met by coding FR scheme with different QP values and choosing the QP value that produced the closest bitrate to the bitrate point given in the CfP. Under the same bitrate constraint, downsampling with lower resolutions enables encoding with lower QP values compared to the QP utilized in FR encoding. Based on our statistical results, the ratios between QP values for downsampling ratios of 3/4 and 1/2 are 0.88 and 0.74, respectively. QP values around this estimated value for lower resolutions were tested to achieve the closest bitrate to the given bitrate point in the CfP.

TABLE 1. SPATIAL RESOLUTION SELECTION BASED ON MSE-BASED METHOD

	Rate Points			
	RP1	RP2	RP3	RP4
Poznan Hall2	FR	FR	FR	FR
Poznan Street	FR	FR	FR	FR
Undo Dancer	FR	FR	FR	FR
GT_Fly	FR	FR	FR	FR
Kendo	1/2	1/2	3/4	3/4
Balloons	1/2	1/2	3/4	3/4
Lovebird1	1/2	3/4	3/4	3/4
Newspaper	1/2	1/2	3/4	3/4

3.2 Results of MSE-based method

The MSE method resulted in Rate Distortion (RD) curves presenting the distortion by MSE. The lowest MSE per specific bitrate and encoding scheme is selected considering an interval equal to 5% as presented in sub-section 2.1. Resolution selection based on MSE RD curves is presented in Table 1 where 1/2, 3/4, and FR present schemes where the sequences have resolution of 1/2, 3/4, and FR, respectively.

Table 1 shows that the MSE-based method resulted in the selection of 1/2 or 3/4 resolution for the 1024×768 sequences, while FR was consistently selected for the 1920×1088 sequences. In a subjective assessment of expert viewers, a resolution lower than FR was generally preferred not only for the 1024×768 sequences but also for the 1920×1088 sequences, when the viewing conditions of the CfP were used. This finding was also supported by the results of the CfP [11] as follows. We submitted coded sequences using the resolutions in Table 1. The same codec was used to encode sequences of different resolutions; hence the compression performance should be approximately equivalent regardless of the resolution. We compared the subjective evaluation results of our submission to the H.264/MVC anchor bitstreams by linearly interpolating the bitrates where H.264/MVC anchor results gave the same subjective quality as our submission in RP1 and RP2. It was found that the average bitrate reduction of RP1 and RP2 yielding the same subjective quality was about 20 percent units higher for the 1024×768 sequences in the C2 coding conditions. Comparing this bitrate reduction to that for the 1920×1088 sequences gave indications that an appropriate spatial resolution selection played an essential role in the subjective quality of the 1024×768 sequences and that the subjective quality of coded 1920×1088 sequences could be improved by downsampling.

TABLE 2. SPATIAL RESOLUTION SELECTION BASED FREQUNCY-BASED APPROACH

	Rate Points			
	RP1	RP2	RP3	RP4
Poznan Hall2	1/2	1/2	1/2	1/2
Poznan Street	1/2	1/2	1/2	1/2
Undo Dancer	1/2	1/2	1/2	1/2
GT_Fly	1/2	1/2	1/2	1/2
Kendo	1/2	1/2	1/2	1/2
Balloons	1/2	1/2	1/2	1/2
Lovebird1	1/2	FR	FR	FR
Newspaper	1/2	1/2	1/2	1/2

3.3 Results of the frequency-based method

Resolution selection based on the distortion metric presented in sub-section 2.2 is reported in Table 2. The thresholds we used in our experiment are $T1 = 0.65 * width$, $T2 = 0.65 * height$, and $T3 = 0.2 * \mathbf{HFC}(UF)$ but the scheme is quite flexible to these thresholds. Note that these results differ from those achieved by MSE-based method (see Table 1).

Results in Table 2 show that the proposed metric favored selection of the 1/2 resolution consistently for the 1920x1088 sequences. As explained in the previous sub-section, such selection of resolutions was supported by expert viewing and also the subjective evaluation results of the CfP suggested that a lower resolution than FR could be appropriate for the 1920x1088 sequences. Nevertheless, we wanted to verify the resolutions provided by the proposed method through a systematic subjective test as explained in Section 4. Frequency based distortion metric-based method failed to select the resolution with the highest subjective quality for Lovebird1. It might be due to relatively higher (~2.5 times) cost(*CL*) value compared to the rest of 1024x768 sequences. The higher cost(*CL*) might be because of false edges due to the original sequence having JPEG-like blocking artifacts. This means downsampling eliminated more high frequency components for Lovebird1.

4. SUBJECTIVE EXPERIMENT

Subjective assessment was performed on three out of four 1920×1088 sequences. The input views and synthesized views utilized in our experiment are the same as specified in 3DV CfP for case C2 [1].

The same encoder as introduced in sub-section 3.1 was utilized for encoding multiview texture data and the following coding scenarios were evaluated:

- Full Resolution Scheme (FRS): 3DV coding on full resolution input
- Downsampled Scheme 1 (DS1): 3DV coding on downsampled texture with downsampling ratio 3/4 applied to both directions
- Downsampled Scheme 2 (DS2): 3DV coding on downsampled texture with downsampling ratio 1/2 applied to both directions

Each of these schemes produced a bit stream associated with rate points RP3 and RP1 given in 3DV CfP.

4.1 Test Procedure and Participants

Subjective viewing was conducted according to the 3DV CfP specification [1]. The 46'' Hyundai stereo display with passive glasses was utilized for displaying of the test material. The viewing distance was equal to 4 times the displayed image height (2.29m for HD sequences).

Subjective quality assessment was done according to Double Stimulus Impairment Scale (DSIS) method [12] with discrete unlabeled quality scale from 0 to 10 was used for quality assessment. Prior to the actual test, subjects were familiarized with test task, test sequences and with the variation in quality they could expect in the actual tests. The viewers were instructed that 0 stands for the lowest quality and 10 for the highest.

Prior to the participation in subjective viewing experiment, candidates were subject to a thorough vision screening. Candidates who did not pass the criteria (near and far vision, Landolt chart) of 20/40 visual acuity with each eye or color vision (Ishihara) were rejected. All participants had a stereoscopic acuity of at least 60 arc sec.

Subjective viewing was conducted with 30 subjects, (19 female, 11 male), aged between 18-29 years (mean: 23.7). The majority of the candidates (90%) were considered naïve as they did not work or study in fields related to information technology or video processing.

4.2 Subjective results

Subjective test results are depicted in Figure 1. It can be judged from the mean scores and confidence intervals presented in Figure 1 that subjective quality of DS2, associated to the lowest resolution, tends to be higher

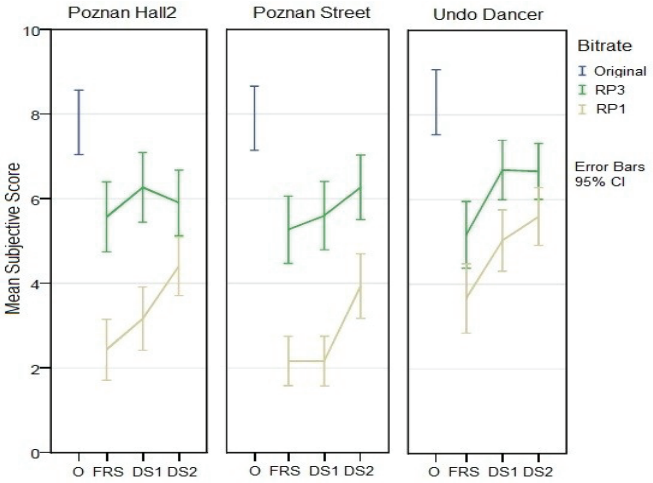

Figure 1. Subjective results for different encoding schemes

TABLE 3. SPATIAL RESOLUTION SELECTION BASED ON STATISTICAL SIGNIFICANCE ANALYSIS ON SUBJECTIVE RESULTS

	Rate Points	
	RP1	RP3
Poznan Hall2	1/2	1/2, 3/4, FR
Poznan Street	1/2	1/2
Undo Dancer	1/2	1/2, 3/4

compared to other schemes. The observation on significant differences between the encoding schemes was further analyzed using statistical analysis as presented in the paragraphs below.

Non-parametric statistical analysis methods, Friedman's and Wilcoxon's tests, were used as the data did not reach normal distribution (Kolmogorov-Smirnov: $p<.05$). Friedman's test is applicable to measure differences between several and Wilcoxon's test between two related and ordinal data sets [13]. A significance level of $p < .05$ was used.

The following conclusions were obtained with statistical significance analysis presented above. In lower bitrates, DS2 has always better subjective results. In higher bitrates, all schemes have a similar performance for Poznan Hall2. In Poznan Street, DS2 has significantly a better subjective quality while DS1 and DS2 have a similar subjective quality for Undo Dancer and both are performing better than FRS. These results are reported in Table 3.

5. DISCUSSION

In this section, the objective results of the proposed methods are compared with subjective results available on a sub-set of test material. The subjective results are used as a reference and the performance of MSE- and DCT-based methods is evaluated based on similarity of their results with subjective results i.e. the more accurately estimating the subjective results, the better performing the method.

First, we compared the objective results achieved by MSE method with subjective results on available subset of test material. We noticed that MSE is not an appropriate metric to predict the subjective quality since only one of the resolution selections made by this method were aligned with subjective results. MSE results in all cases for HD sequences were favored to select the encoding schemes with FR while subjective results showed otherwise.

Second, resolution selection achieved by proposed method was compared with selection based on subjective test. In all cases the proposed method succeeded to estimate the subjective results correctly.

6. CONCLUSIONS

This paper tackled the problem of adaptive spatial resolution selection by comparing two methods. First, MSE value was calculated for FR and lower resolutions. The resolution with the smallest average MSE value was selected as the candidate to have the best subjective quality. This selection was compared then with subjective results on a subset of test material, revealing that the MSE-based method is not able to estimate the subjective quality accurately (one out of the six cases were estimated correctly). To solve this problem an objective metric based on FPS was described. The results confirmed that utilization of this algorithm succeeded to select the resolution with the best subjective quality whenever the subjective quality assessment results were available (all cases were estimated correctly). Hence, the proposed method is a potential candidate metric to select the resolution of the texture view prior to encoding by which the best perceived quality is assured.

ACKNOWLEDGMENT

The authors would like to thank Timo Utriainen, Emilia Pesonen, and Satu Jumisko-Pyykkö from the laboratory of the Human-Centered Technology of Tampere University of Technology for performing and providing the subjective results. Moreover, the authors thank Prof. M. Domañski, et al. for providing Poznan sequences and their camera parameters [9].

REFERENCES

[1] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Geneva, Switzerland, March 2011.

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.

[3] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in Proc. SPIE Conf. 5291, CA, U.S.A., Jan. 2004, pp. 93–104.

[4] W. Lin and D. Li, "Adaptive downsampling to improve image compression at low bit rates," IEEE Trans. Image Process., vol. 15, no. 9, pp. 2513–2521, Sep. 2006.

[5] V.A. Nguyen, Y.P. Tan, W.S. Lin, "Adaptive downsampling/upsampling for better video compression at low bit rate," IEEE ISCAS, pp.1624-1647, May 2008.

[6] E. Ekmekcioglu, S. T. Worrall, and A. M. Kondoz, "Bit-rate adaptive downsampling for the coding of multiview video with depth information," in Proc. 3DTV Conf., Istanbul, Turkey, May 2008, pp. 137–140.

[7] Ren-Jie Wang, Ming-Chen Chiena, and Pao-Chi Chang, "Adaptive downsampling video coding," Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7542, 2010

[8] R. Oktem, L. Yaroslavsky and K. Egiazarian, "Signal and Image Denoising in Transform Domain and Wavelet Shrinkage: A Comparative Study", in Proc. of EUSIPCO'98, Sept. 1998.

[9] M. Domañski, et al, "Poznan Multiview Video Test Sequences and Camera Parameters", MPEG 2009/M17050, October, 2009.

[10] JM reference software: http://iphome.hhi.de/suehring/tml/download

[11] http://mpeg.chiariglione.org/working_documents/explorations/3dav/3d-test-report.zip

[12] ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002.

[13] H. Cooligan "Research methods and statistics in psychology," (4th ed.). London: Arrowsmith., 2004.