Sampling in Transform Domain for Improved QoE of 3D Frame-Compatible Video Coding

Jin Li, Jan De Cock, Peter Lambert, Rik Van de Walle
Ghent University-iMinds, Department of Electronics and Information Systems - Multimedia Lab,
Gaston Crommenlaan 8 bus 201, 9050 Ledeberg-Ghent, Belgium

Abstract - **3D frame-compatible formats are considered a promising solution to 3D distribution due to their compatibility to existing systems. However, since sub-sampling is applied in order to pack the left and right views into a single frame, the Quality of Experience (QoE) has been a major concern. Existing methods in the related literature make use of low-pass filtering to halve the original samples, leaving room for further QoE improvement. In this paper, the sub-sampling is performed in the transform domain and designed to coincide with the H.264/AVC coding structure. The methodology compensates for the fact that sub-sampling tends to truncate the same information as the encoder quantization. In this way, the QoE degradation additionally caused by the sub-sampling is minimized. Experimental results show a significant improvement over competing methods, roughly 2.5dB higher and 75% BD-rate reduction on average. In addition, the evaluation criterion is also discussed by taking into account the impacts of sampling and coding process.**

I. INTRODUCTION

Frame-compatible formats offer a solution to introduce 3D services in existing environments [1], [2]. This format can represent stereo video in such a way that is compatible with existing codecs and delivery infrastructure.

As a result, the video can be compressed with existing encoders, transmitted through existing channels and decoded by existing receivers and players. Due to these minimal changes, stereo services can be quickly deployed to capable displays. So far, HDMI v1.4 has announced its support [3], and H.264/AVC also adopts it in Supplement Enhancement Information (SEI) [4].

The drawback of representing the stereo signal in this way is that spatial or temporal resolution would be lost. In order for the packed format to have the same resolution as the original views, sub-sampling is introduced to halve the samples. At the receiver, the decoded video is up-sampled to retrieve the left and right views. However, the QoE is degraded by the sampling process and the coding process.

So far, sampling for 3D frame-compatible formats has not been intensively researched. However, it is of importance to investigate different sampling approaches for this kind of application, because it is not only a pure sampling process but also has a great impact on the subsequent coding process. In the related literature, sampling is usually performed using a low-pass filter followed by decimation [5]. Although a good trade-off between the video quality and compression ratio was claimed, based on the results we simulated, the video quality needs to be further improved.

In this paper, sampling in the transform domain is examined for the application of 3D frame-compatible formats. Since the high frequency coefficients tend to be removed in the encoding process, the proposed sub-sampling is performed in the transform domain using the same transform structure as H.264/AVC. Therefore, the information removed by the sub-sampling has a very high probability to be also dropped by the quantization process. In this way, the information lost by sub-sampling is minimized. In addition, the evaluation criterion is also discussed by taking into account the impacts of both the sampling and the coding process.

The rest of this paper is organized as follows. In Section 2, the coding framework of 3D frame-compatible formats and two sampling techniques are briefly reviewed. Section 3 describes the proposed sampling technique and the proposed evaluation method. The simulation results are presented in Section 4. Finally, Section 5 concludes the paper.

II. CODING FRAMEWORK OF 3D FRAME-COMPATIBLE FORMATS

With a frame-compatible format, the two views are decimated and then packed into a single video. The packing schemes such as interleave and checkerboard are specified by H.264/AVC. However, experiments show that SbS and TaB formats tend to have better results in terms of PSNR versus bitrate [2]. Practically, some source material is already interlaced, thus SbS format is usually preferred over other formats.

A coding framework for 3D frame-compatible formats is shown in Fig.1. The left and right views are first sub-sampled before being packed into a single view. Then, the 3D frame-compatible format is passed to the encoder. The bit-stream is transmitted over channels to the decoder. At the decoding side, the 3D frame-compatible format is

978-3-901882-50-0

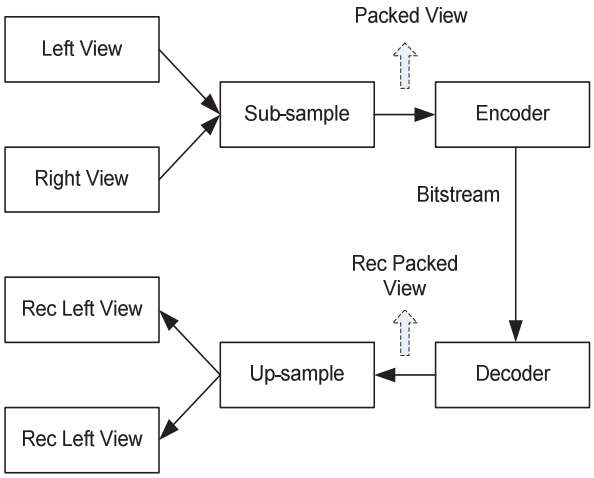

Fig.1 Coding framework of 3D frame-compatible formats

reconstructed and then de-multiplexed into two video contents corresponding to the source inputs.

As a lossy process, it is no doubt that sub-sampling has a great impact on QoE. In addition, because the sub-sampling process will determine what information to be preserved for coding, it significantly affects the following compression process, i.e., the coding efficiency. Dolby [5] specifies a sub-sampling method in spatial domain using a low-pass filter followed by decimation. The low-pass filter for luma component is [-5 19 29 -68 -47 305 558 305 -47 -68 28 19 -5] and that for chorma components is [-20 -13 84 154 84 -13 -20]. In the specification, Dolby claims that this filter could achieve a good trade-off of sharpness and compression ratio.

Although it turns out that sampling greatly affects the QoE as well as the coding efficiency for this application, intensive research has not been performed. However, it is of importance to develop such sampling methods that coincide with the coding structure. In theory, if the sub-sampling tends to drop the same information that the quantization will do, the additional QoE degradation caused by sampling could be minimized. Besides, the implementation should be friendly for existing systems and infrastructure.

III. SAMPLING IN TRANSFORM DOMAIN

In this paper, sub-sampling in transform domain is proposed in order to coincide with the structure of H.264/AVC which has been widely adopted by the industry. In this section, the methodology will be described in detail, which is followed by a discussion concerning the performance evaluation on this topic.

As a major evolution to other coding standards, H.264/AVC deploys the 4×4 transform block. Moreover, in order to facilitate the implementation, the transform matrix $\boldsymbol{A}$ is decomposed into a new transform matrix $\boldsymbol{C}$ and a scaling matrix $\boldsymbol{E}$ which is integrated into the later quantization process.

$$\boldsymbol{Y} = (\boldsymbol{CXC^T}) \otimes \boldsymbol{E} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \boldsymbol{X} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 1 & -1 & -2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \right) \otimes \begin{bmatrix} a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4} \\ a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4} \end{bmatrix} \quad (1)$$

where $\boldsymbol{Y}$ is the transform block and $\boldsymbol{X}$ represents the 4×4 samples in spatial domain.

Following the transform process, quantization is performed to remove a lot of high frequency coefficients in order to achieve a good compression ratio. In many video coding standards, quantization has been the major lossy coding technique if it is not the only one, which leads to the video quality degradation.

As shown in Fig.1, sub-sampling is introduced to encode the 3D frame-compatible formats. Thus, there exist two major processes that cause information loss. In the sub-sampling process, some information will be lost, and during the following quantization process, high frequency information will be further truncated.

On the other hand, the video quality perceived at the decoder side highly depends on the amount of received original information. Therefore, theoretically it is always beneficial to maximize the overlap between the information removed by sub-sampling and by quantization. That is, the errors could be minimized if both the sub-sampling process and the quantization tend to truncate the same information, i.e., high frequency information.

Based on above analysis, a sub-sampling method in transform domain is proposed here. For an input vector $\boldsymbol{X_N} = \{x_0, x_1, \dots, x_{N-1}\}$, the DCT output vector $\boldsymbol{Y_N} = \{y_0, y_1, \dots, y_{N-1}\}$ is given by the relation

$$\boldsymbol{Y_N} = \boldsymbol{T_N X_N} \quad (2)$$

where the $N \times N$ transform matrix for N-point DCT is written as

$$T_N(i,j) = \sqrt{\frac{2}{N}} \begin{cases} 1/\sqrt{2}, & i = 0 \\ cos\dfrac{(2j+1)i\pi}{2N}, & otherwise \end{cases} \quad (3)$$

Then, the high frequency coefficients in $\boldsymbol{Y_N}$ are truncated and a new vector $\boldsymbol{Y_{N/2}}$ is therefore obtained as $\boldsymbol{Y_{N/2}} = \{y_0, y_1, \dots, y_{N/2-1}\}$.

Following the truncation, $N/2$ –point inverse transform is applied as

$$\boldsymbol{X'_{N/2}} = \boldsymbol{T^{-1}_{N/2} Y_{N/2}} \quad (4)$$

where $\boldsymbol{X'_{N/2}}$ represents the pixel vector after the sub-sampling. In this way, high frequency information is removed and the resolution of the original image is halved.

At the receiver side, up-sampling is applied in order to de-multiplex the reconstructed packed video into the left

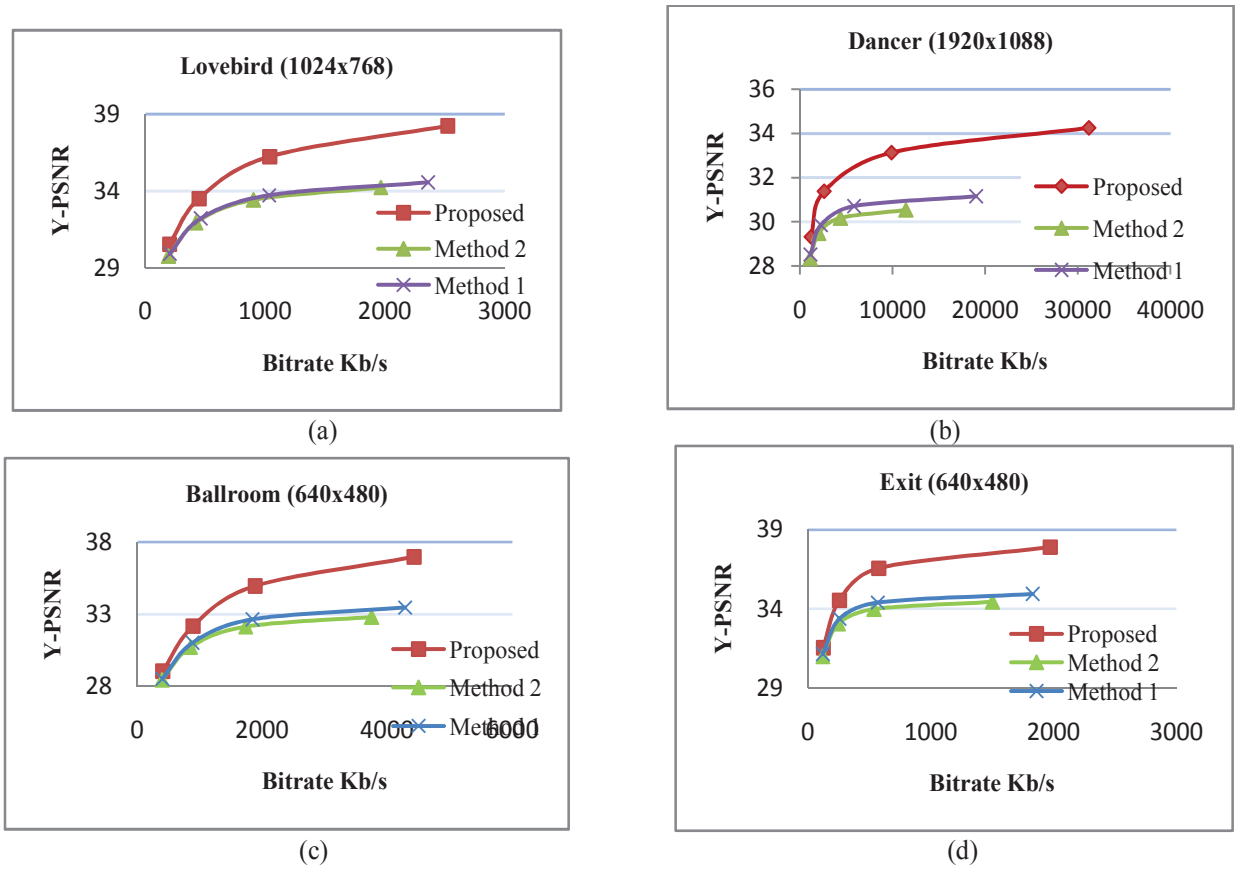

Fig.2. PSNR comparison among the proposed scheme, Method 1 specified by Dolby and Method 2 using MPEG-4 filtering, (a) Lovebird, (b Dancer), (c) Ballroom, and (d) Exit.

Table I BD-rate comparison (%)

	Lovebird	Dancer	Ballroom	Exit
Method 1	68.9	136.1	49.0	56.3
Method 2	70.6	145.2	54.5	64.9

view and the right view. The up-sampling is a reverse process of the sub-sampling. Firstly, two-point transform is performed as

$$\boldsymbol{Y}''_{N/2} = \boldsymbol{T}_{N/2}\boldsymbol{X}''_{N/2} \qquad (5)$$

where $\boldsymbol{X}''_{N/2}$ is a vector composed of $N/2$ reconstructed pixels, $\boldsymbol{Y}''_{N/2}$ represents the transform coefficients. Then, $\boldsymbol{Y}''_{N/2}$ is padded with $N/2$ zeros so that a new N-point vector $\boldsymbol{Y}''_{N}$ is formed. Finally, based on (4) an N -point inverse transform is applied on $\boldsymbol{Y}''_{N}$ to construct a new pixel block $\boldsymbol{X}''_{N}$ containing N pixels. Therefore, up-sampling is realized.

In this paper, N is defined as 4 in order to coincide with the coding structure of H.264/AVC and to preserve the original information as much as possible.

In order to compare the performance among different methods, the objective video quality is measured in terms of PSNR and SSIM versus bitrate. However, the PSNR should not be calculated between the packed video and the reconstructed packed video, since it does not take into account the impact that the sampling has on the video quality. Instead, in this paper the PSNR is defined by the distortion between the original left and right views and the reconstruct left and right views which are to be observed by the clients.

VI. EXPERIMENTAL RESULTS

The proposed model was tested using the JM 18.2 against the methods specified by Dolby and using a MPEG-4 filtering. The PSNR versus bitrates are plotted based on the experimental results. As mentioned in Section 3, the PSNR is calculated based on the original left and right views and the reconstructed left and right views. Experiments were carried out with the Baseline profile.

The method specified by Dolby is referred as "Method 1" where sub-sampling uses the low-pass filter defined in Section 2 and the up-sampling applies Lanczos filter as [3 0 -17 0 78 128 78 0 -17 0 3]. The other method which is referred as "Method 2" utilizes the MPEG-4 filter to down-sample the original views into lower-resolution as [2, 0, −4, −3, 5, 19, 26, 19, 5, -3, -4, 0, 2]. While the low-resolution pictures are up-sampled by the Lanczos filter.

Figure 2 shows the comparisons between the proposed method and the other two approaches. According to the results, the proposed method significantly outperforms the approaches that perform the sub-sampling in the spatial domain. It is also shown that Method 1 defined by Dolby is slightly better than Method 2 which uses MPEG-4 filtering.

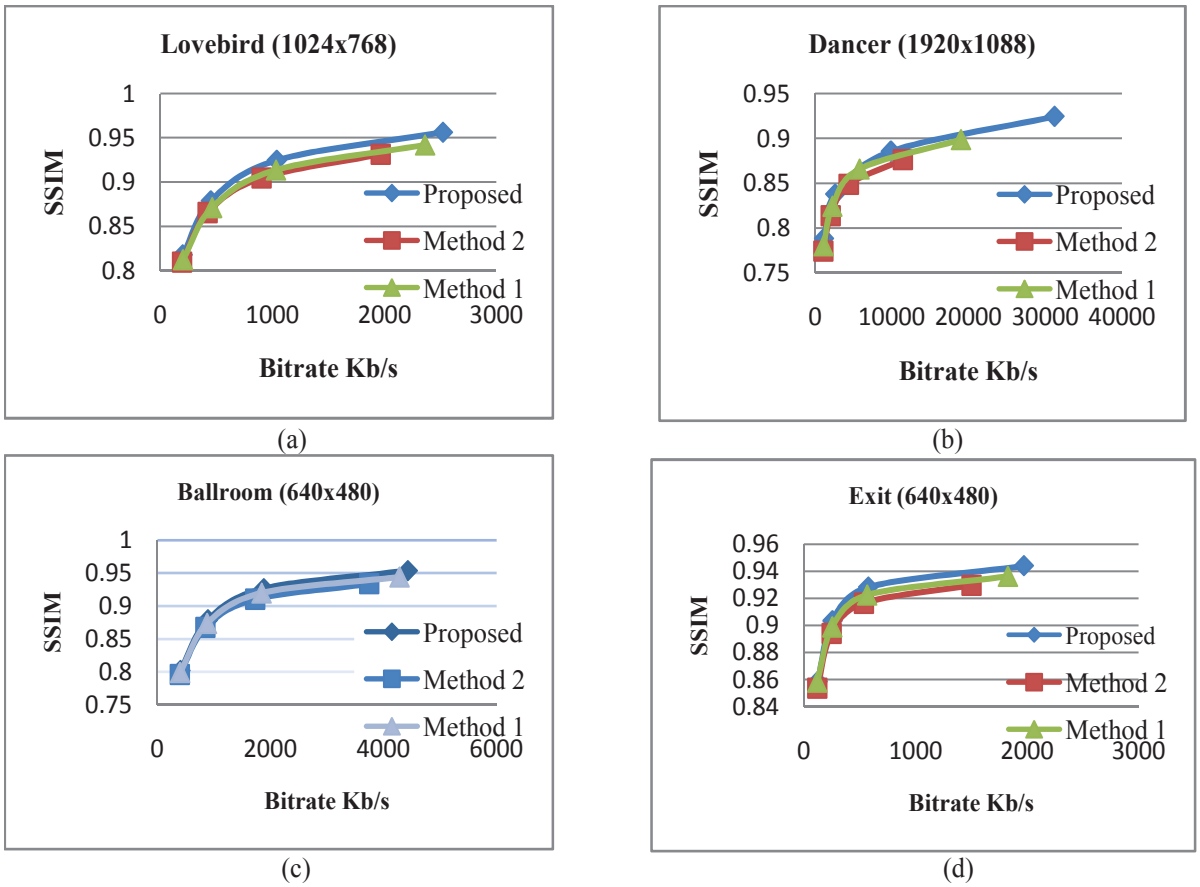

Fig.3. SSIM comparison among the proposed scheme, Method 1 specified by Dolby and Method 2 using MPEG-4 filtering, (a) Lovebird, (b) Dancer, (c) Ballroom, and (d) Exit.

Based on the figure, the proposed method achieves about 2.5dB higher on average over the other methods. Since the proposed sub-sampling is performed in the transform domain and applies the four-point transform which coincides with H.264/AVC, the sub-sampling and the quantization processes have a high probability to truncate the same high frequency information. Thus, the additional information loss caused by the sampling process is reduced.

Table I shows the comparisons in terms of BD-rate [6]. The BD-rate is calculated based on the proposed method. For instance, Method 2 requires 70.6% more bitrate to achieve the same PSNR as the proposed approach for Lovebird. It is obvious that the proposed sampling has the best performance over the methods in spatial domain.

Besides the traditional measurement such as PSNR, the method is also evaluated in terms of the structural similarity (SSIM). Figure 3 demonstrates the comparisons between the proposed scheme and competing methods. According to the reported results, the proposed method performs best in all cases in terms of both PSNR and SSIM.

V. CONCLUSIONS

In this paper, a sampling method is proposed for 3D frame-compatible formats. The sub-sampling is performed in the transform domain with a four-point DCT to coincide with the coding structure of H.264/AVC. In this way, the sampling process and the quantization process have a high probability to truncate the same high frequency information. Therefore, the additional information lost by sampling is minimized. The results show very attractive QoE. The evaluation criterion is also discussed by taking into account both the sampling and the coding processes.

ACKNOWLEDGEMENT

This work is part of 3DTV 2.0 project and financially supported by iMinds.

REFERENCES

[1] A. Vetro, "Frame Compatible Formats for 3D Video Distribution," *IEEE ICIP*, pp. 2405-2408, China, 2010.

[2] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard," *IEEE Special Issue on 3D Media and Displays*, vol. 99, issue 4, pp. 626-642,2011.

[3] HDMI Licensing, "HDMI Specification 1.4," 2009.

[4] G. J. Sullivan, "Draft AVC amendment text to specify Constrained Baseline profile, Stereo High profile, and frame packing SEI message," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, 2009.

[5] Dolby, "Dolby Open Specification for Frame-Compatible 3D Systems," 2010.

[6] G. Bjontegaard, "Calculation of Average PSNR Differences between RD curves,"VCEG-M33, 2001.