VIDEO RETARGETING BASED FRAME-COMPATIBLE STEREO VIDEO CODING

Siao-Wei Chen, Ming-Feng Tsai, Jui-Chiu Chiang

Department of Electrical Engineering, National Chung Cheng University, Taiwan

ABSTRACT

Due to more and more request of visual entertainment with improved perceptual realism, stereo video offering enhanced realism and interactivity is highly desired. Usually, the stereo video is packed into a single video and each view preserves only half resolution to achieve an efficient delivery. In this paper, a novel frame-compatible stereo video coding technique based on content-aware video retargeting is presented. Despite of combing the stereo video into a single channel by uniform downsampling, we propose to sub-sample the stereo video in an unequal way, by taking the saliency map of the video into considering. During the reconstruction of the stereo video with original resolution, the regions with higher saliency attention are less distorted, relying on a smaller downsampling factor, instead of 2. The experimental results reveal that the proposed technique achieves up to 43% bitrate saving as compared to the most popular frame-compatible stereo formats.

Index Terms— Frame compatible, stereo video, video retargeting, content-aware

1. INTRODUCTION

Recently, the advance in 3D display technology, enabling the depth perception, provides a better visual experience in our daily life. Usually stereo videos are captured by two cameras deployed in parallel for the same scene. The data amount of two-channel video is two times of that of traditional 2D video, and a wider bandwidth is needed accordingly. To reduce the cost in terms of bandwidth occupation and hardware, it is desired to deliver the stereo content through the existing channel specified for conventional 2D video. Consequently, frame-compatible stereo video formats were proposed [1] and received considerable attention currently. In addition to the reuse of the existing infrastructure and hardware, one noticeable advantage of such formats is that there is no need to design additional codec and the 2D codec is able to encode and then decode the stereo video with some supplemental information embedded in the bitstream, such as SEI message in H.264/AVC [2], which mentions the packing type used in the frame-compatible stereo video. Then the stereo video can be recovered by extracting each view from the decoding video. Usually, an upsampling procedure needs to be employed to reconstruct the stereo video with the original resolution. Typically, frame packing types include side-by-side (SbS), top-bottom (TB), column interleaved, row interleaved, and checkerboard. Besides, the temporal multiplexing format is generated by downsampling each video in the temporal direction and combining the downsampled two videos into a single video in an interleaving manner. Basically, the SbS and TB formats are more popular than the others due to a higher coding efficiency guaranteed, as well as a better quality in the reconstructed images and hence are the mandatory formats supported by HDMI 1.4 [3].

In addition to those formats supported by H.264/AVC, some literatures present different strategies to pack the stereo video [4-5]. In [4], the stereo video with resolution 1280×720 in each view is merged into a video with resolution 1920×1080, where the left view image is put on the left side and the right view image is split into three parts and put on the remaining space. Although the delivery of the stereo video with resolution 1280×720 is achieved in this way, the right view suffers from an artifact presenting discontinuities between boundaries of the three parts. To enhance the image quality of the stereo video delivered based on frame-compatible formats, several extensions based on scalable video coding and multi-view coding maintaining the backward compatibility were proposed [6-7].

In this paper, we propose a new frame-compatible stereo format based on a HVS (Human Visual System) mechanism where the regions the people pay a higher attention will be less distorted during the downsampling and upsampling processes by assigning a downsampling factor smaller than 2. On the other hand, the regions with minor interest will be assigned a higher downsampling factor. And a much better R-D (rate-distortion) performance is provided, compared to the formats currently used in the 2D coding standards.

2. PROPOSED SCHEME

The frame-compatible formats supported in H.264/AVC combine the two video with half data by downsampling each 2D video in a uniform manner. Consequently, the distortion after upsampling to the original resolution could be uniformly distributed within the entire image. Due to the HVS property, not all the content inside the image draw the same attention for viewers, and the ROI (region of interest) should be much better reconstructed to ensure an enhanced

image perception. Thus, the idea behind the proposed frame-compatible format is to reduce the distortion of the ROI region during the frame-compatible stereo video coding and reconstruction.

Figure 1 depicts the block diagram of the proposed scheme. First, for each image in each view, a saliency map is built to identify the image energy distribution in a per-pixel basis. Then a global salience map is built considering the salience maps within one GOP (group of picture). After that, each image will be divided into several strips depending on similarity among the global salience value of each column/row. The salience value is cumulated in column for the SbS based frame-compatible format, and is cumulated in row for TB based. Once the strips are determined, each strip is assigned an appropriate scaling factor revealing its significance over the entire image and a sub-sampled image is obtained. Following that, the sub-sampled images from two views are combined into a single image in term of SbS or TB formats, and a frame-compatible format is ready to be encoded by the 2D codec. In the decoder, each strip is upsampled according to the scaling factor assigned with the supplementary information embedded in the bitstream indicating the partitions and scaling factors of strips in each frame. The details of each step will be described in the followings.

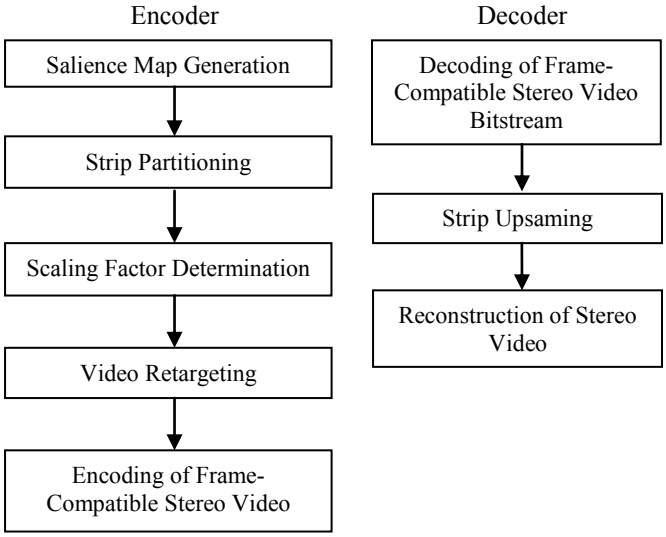


Figure 1. The block diagram of the proposed scheme.

2.1. Salience Map Generation

The saliency map (SM) is generated by the method proposed in [8], where the information derived from the luminance contrast, the color-double-opponent and the gradient are linearly combined. To ensure temporal consistency for the retargeting based frame-compatible stereo video, and consequently the coding results in a promising R-D performance, the strip partitions are determined by a global salience map generated from the saliency in the temporal trajectory within the same GOP, instead of the saliency map of the current image. Note that, if each image is performed retargeting independently, the same object in different frames could be assigned different scaling factor and becomes alike, which leads to unsatisfied coding performance due to reduced inter-frame prediction exploited. The global saliency map for each GOP is obtained by selecting the maximum value at the same coordinate covered in the time window, expressed as below:

$$SM_G(x,y,m) = \max(SM(x,y,m(N-1)+n)), for \ \ 0 \le n \le N-1 \tag{1}$$

where (x,y), N and m denote, respectively, the image coordinate, the GOP size and the GOP index. Then the saliency value in the global saliency map is cumulated for each line (column or row). However, only the top 25% of the saliency values of each line are summed up considering their representative property, and this map is called global importance map (IM_G). Note that, for the images within one GOP, there is only one global importance map and the strip partitions for these images will be exactly the same accordingly.

2.2 Strip Partitioning

The strip partition method is inspired by the algorithm presented in [9], but with some modifications here. In the beginning, each global importance map is uniformly partitioned into K strips. Let b_k and b_{k+1} denote the left partition boundary and the right partition boundary for the k_{th} strip, where $0 \le k \le K$-1. Assume x axis denote the dimension to be sub-sampled and its width is W. Then b_0=0 and b_K=W. For the given boundary values b_{k-1} and b_{k+1}, the boundary value b_k for the k_{th} strip will be updated accordingly to (2).

$$b_k = \arg\min_{b_{k-1} \le b < b_{k+1}} (D(b)) \tag{2}$$

$$\text{where } D(b) = \sum_{x=b_{k-1}}^{b-1} \left| IM_G(x) - f_{k-1} \right|^2 + \sum_{x=b}^{b_{k+1}-1} \left| IM_G(x) - f_k \right|^2$$

$$f_k = \frac{\sum_{x=b}^{b_{k+1}-1} IM_G(x)}{b_{k+1} - b} \quad \text{and} \quad f_{k-1} = \frac{\sum_{x=b_{k-1}}^{b-1} IM_G(x)}{b - b_{k-1}}, \tag{3}$$

The criterion to determine the boundary is based on the assumption that the line with similar significant level should be grouped into the same strip. After b_k is determined, the boundary value for the strip nearby can be updated and the algorithm performs iteratively until the convergence is achieved. Besides, after all the boundary values are computed, two strips will be merged if the difference between their significant values is below a pre-defined threshold. In this way, the data of the supplementary

information sent to the decoder can be reduced if global importance image is divided with fewer strips.

2.3 Scaling Factor Determination

Once the strip partitions are found, we need to decide an appropriate scaling factor for each strip. Since f_k in (3) represents the averaged importance value for the strip, a parameter g_k is defined here to present the relative importance for the k_{th} strip over all the strips in the global important map, and it is expressed as below:

$$g_k = \frac{f_k}{\sum_k f_k} \tag{4}$$

Then the initial scaling factor s_k for the k_{th} strip is computed after taking into account the final strip number K_f and the ratio between the desired output size W_{out} and the original dimension, as shown in (5).

$$s_k = g_k \times \frac{W_{\text{out}}}{W} \times K_f \tag{5}$$

Note that, if the scaling factor s_k obtained in (5) is larger than 1(it happens when the strip has a definitely high significant value), it will be changed to one to prevent the unreasonable situation for downsampling case. However, with such an adjustment, it is possible that the output size does not satisfy the desired one, and an additional procedure is carried out here. Let the initial output size be W'_{out}. We discuss two cases here, including that W'_{out} is larger than W_{out} and W'_{out} is smaller than W_{out}. For the case that W'_{out} is larger than W_{out}, it means we should do some scaling furthermore and a smaller scaling factor will be assigned to the strip with the lowest scaling factor, according to (6), where $s^{\min}$, $l^{s\min}$ denote the lowest scaling factor and the corresponding width of this trip, respectively. On the other hand, if W'_{out} is samller than W_{out}, a higher scaling factor will be assigned to the strip with the highest scaling factor. Although it happens seldom, this additional procedure can be performed on the other strips if the desired size is not achieved after adjusting the lowest/highest scaling factor, where the scaling factor after adjustment should always be between 1 and 0.

$$s = \begin{cases} s^{\min} - \dfrac{(w'_{out} - w_{out})}{l^{s\min}} & , \quad w'_{out} > w_{out} \\ s^{\max} + \dfrac{(w_{out} - w'_{out})}{l^{s\max}} & , \quad w'_{out} < w_{out} \end{cases} \tag{6}$$

Then, the retargeting is carried out on the input image according to the results of strip partitioning and scaling factor. Then two retargeted videos are combined into a single video, followed by encoding. The downsampling filters used in the scalable video coding (SVC) standard [10] are adopted here to individually perform downsampling for each strip according to the assigned scaling factor. The bitstream sent to the decoder will contain the strip partitions and scaling factors as supplementary information. This information will be encoded by entropy coding and the overhead can be ignored due to the few data amounts.

2.4 Upsamping in the Decoder

The filters used for upsampling in SVC are adopted here, for both the luminance and the chrominance. Since strips may have different scaling factors, the upsampling for each strip is performed independently where the upsampling factor is the inverse of the scaling factor and the appropriate filter can be determined accordingly. After the combined stereo video is reconstructed, the left view and right view can be extracted and displayed on the 3D television. If the device does not support the stereo display, it would be feasible to display one view with original resolution.

3. EXPERIMENTAL RESULTS

Two stereo videos are used here to evaluate the proposed scheme, and the encoding setting is summarized in Table 1. The frame-compatible stereo video is encoded as the base view in JMVC platform due to the consideration of a possible extension for full resolution case. The initial strip number used is 10.

Table 1. Encoding environment for frame-compatible stereo video

Platform	JVMC 8.5 [11]
GOP	16
QP	22, 27, 32, 37
Prediction	Hierarchical B
Test sequences	Ballons, Kendo
Resolution	1024×768 (each view)
Frame rate	30Hz
Total frame number	96 frames

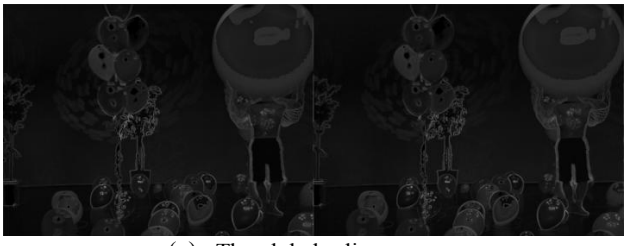
Figure 2 illustrates the global saliency map for both views, the associated column-based distributions, and the corresponding strip partitions on the test image “Ballons”. Figure 2(a) demonstrates that the motion trajectory is well preserved in the global salience map, and the same object will be scaled with similar scaling factor in consecutive frames within one GOP. The red lines appeared in Figure 2(c) denotes the strip boundaries. Although the initial strip number is 10, the final strip number is 4 for both the stereo images. Hence, the overhead to transmit the additional information about strips is really can be ignored. Furthermore, Figure 2(b) & (c) verify that the columns with similar importance value are divided into the same strip and the scaling factor below the Figure 2(c) is able to reveal the importance of the strip over the entire image based on human visual system.

3.1 Comparison with Previous Works

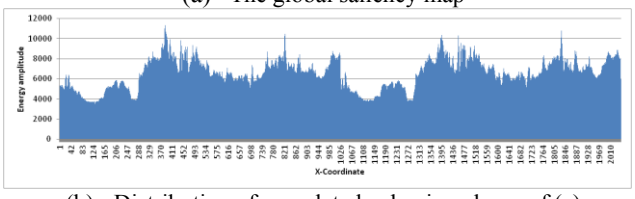
To verify the efficiency and effectiveness of the proposed technique, comparisons with previous works are discussed here. In addition to the conventional frame-compatible formats supported in the current 2D video standard, the

work presented in [5] is also compared. The work in [5] realized the downsampling of each view based on the checkerboard format and the packing is adaptively determined between SbS or TB formats to preserve the edge energy as much as possible. If the edge energy in the vertical direction is higher than that in the horizontal direction, the TB format is used for packing; otherwise, the SbS format is selected. Although the work in [5] can reach better coding efficiency, compared to the conventional packing formats, only the global energy maximization is considered. The work presented here explores the way to perform a better downsampling for each view, instead of the way for packing the stereo video, and how to maximize the human visual perception for the reconstructed stereo video is focused.

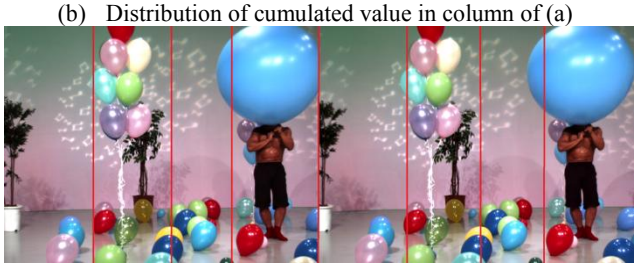
Figures 3 to 4 present the R-D performance comparisons for two test videos, where "single layer" denotes the coding of left view only, and the PSNR is for the left view and the right view has similar value. Our downsampling technique can be employed on the vertical or horizontal direction, where "Proposed_SbS"/"Proposed_TB" denotes that the downsampling for each view is performed on the horizontal/vertical direction and the packing is then carried out as side-by-side/top-bottom format. Those curves show that the proposed scheme has the best performance. Tables 2 to 3 summarize the results in terms of Bjøntegaard Delta (BD) [12]. It shows that the bitrate saving of the proposed scheme is up to 43% as compared to the conventional TB format, and 10% as compared to [5].



(a) The global saliency map



(b) Distribution of cumulated value in column of (a)



0.40	0.56	0.45	0.60	0.35	0.57	0.45	0.58

(c) The strip partitions on the input color image

Figure 2. Illustration of saliency map based strip partitions.

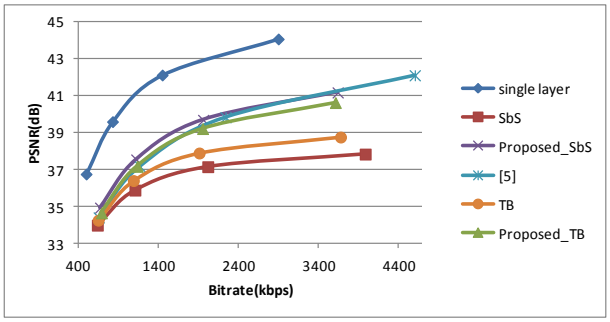


Figure 3. R-D performance comparison for the sequence "Ballons".

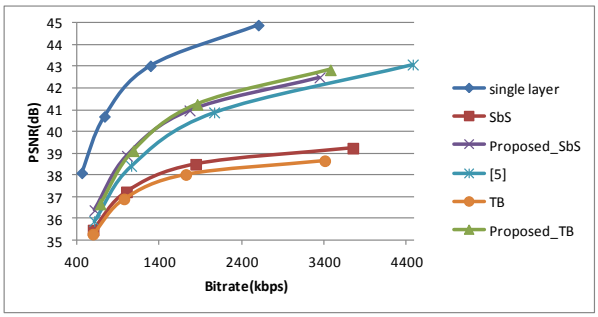


Figure 4. R-D performance comparison for the sequence "Kendo".

Table 2. BDBR and BDPSNR of proposed scheme with respect to the previous works for the sequence "Ballons".

	BDBR (%)	BDBR(dB)
Proposed_SbS vs. SbS	-39.32	2.16
Proposed_TB vs. TB	-24.88	1.32
Proposed_ SbS vs. [5]	-7.94	0.32

Table 3. BDBR and BDPSNR of proposed scheme with respect to the previous works for the sequence "Kendo".

	BDBR (%)	BDBR(dB)
Proposed_SbS vs. SbS	-39.52	1.99
Proposed_TB vs. TB	-43.04	2.47
Proposed_ TB vs. [5]	-10.55	0.42

4. CONCLUSION

In this paper, a new downsampling method for frame-compatible stereo video is developed. By considering the human visual property, the regions attract higher attention will have less distortion during the process of downsampling, followed by upsampling due to a larger scaling factor assigned. The experimental results show that the proposed scheme outperforms the existing techniques not only in the R-D performance, but also in the image quality of the reconstructed stereo video. In the future, this work will be extended to the case of high quality full resolution stereo, where how to maximize the inter-layer prediction between different image layers will be one of the research works.

REFERENCES

[1] A. Vetro, "Frame compatible Formats for 3D Video Distribution," in proc. of *IEEE International Conference on Image Processing*, pp.2405-2408, 2010.

[2] G. J. Sullivan, A. M. Tourapis, T. Yamakage, and C. S. Lim, Draft AVC amendment text to specify constrained baseline profile, stereo high profile, and frame packing SEI message, London, U.K., Joint Video Team (JVT) Doc. JVT-AE204, July, 2009.

[3] High-Definition Media Interface (HDMI) Specification Version 1.4, June, 2009.

[4] G. Ballocca, P. D'Amato, M. Grangetto, and M. Lucenteforte, "Tile format: A Novel Frame Compatible Approach for 3D Video Broadcasting," in proc. of *IEEE International Conference on Multimedia and Expo*, 2011.

[5] A.T. Chiang, H.M. Wang, J.F. Yang, and J.F. Wang, "A New Stereo Packing Format based on Checkerboard Sub-sampling for Efficient Stereo Video Coding," in proc. of *IEEE International Symposium on Circuits and Systems*, pp.385-388, 2012.

[6] G. Sullivan, W. Husak and A. Luthra, "Problem Statement for Scalable Resolution Enhancement of Frame-compatible Stereoscopic 3D video," ISO/IEC JTC1/SC29/WG11 N11526, Jul. 2010, Geneva.

[7] Y. Chen, R. Zhang and M. Karczewicz, "MVC Based Scalable Codec Enhancing Frame-Compatible Stereoscopic Video," in proc. of *IEEE International Conference on Multimedia and Expo*, 2011.

[8] J.-C. Chiang, C.-S. Hsieh, G. Chang, F.-D. Jou, and W.-N.Lie, "Region-of-interest based Rate Control Scheme with Flexible Quality on Demand," in proc. of *IEEE International Conference on Multimedia and Expo*, pp.238 - 242, 2010.

[9] J. S. Kim, J. H. Kim and C. S. Kim, "Adaptive image and video retargeting technique based on fourier analysis," in proc. of *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1730-1737, 2009.

[10]H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no.9, pp.1103-1120, Sep., 2007.

[11 JMVC 8.5, garcon.ient.rwthaachen.de, Sep. 2011.

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