

Regionally Adaptive Filtering for Asymmetric Stereoscopic Video Coding

Ying Chen¹, Ye-Kui Wang², Moncef Gabbouj¹, and Miska M. Hannuksela²

¹Department of Signal Processing

Tampere University of Technology, Tampere, Finland
{ying.chen, moncef.gabbouj}@tut.fi

²Nokia Research Center

Tampere, Finland
{ye-kui.wang, miska.hannuksela}@nokia.com

Abstract—In asymmetric stereoscopic video coding, one view can be coded in a lower resolution of the other. In this scenario, stereoscopic video can be compressed with only moderately increased bandwidth and complexity compared to 2D mono-view video coding. The subjective quality degradation of this scenario can be negligible compared to coding two views with original resolution. The low-resolution view can be predicted from the high-resolution view to achieve higher coding efficiency. In this paper, a regionally adaptive filtering algorithm is proposed to generate a predictor of a macroblock (MB) or MB partition of the low-resolution view from the high-resolution view. Different filters are applied for different picture regions. Disparity motion matching and clustering are applied in the encoder for generation of regionally adaptive filters. Simulation results show that the proposed algorithm results in up to 27% bit-rate saving compared with methods without adaptive filtering.

I. INTRODUCTION

Multiview video technologies have gained significant interest recently. As views are highly correlated, efforts have been undertaken by the Joint video team (JVT) to reduce redundancy between coded views in the Multiview Video Coding (MVC) standard [1], which is an extension to the Advanced Video Coding Standard (H.264/AVC). Many display arrangements for multiview video are based on rendering a different image to the viewer's left and right eyes. For example, when data glasses or autostereoscopic displays are used, only two views are observed at a time in typical multiview applications, such as 3D TV [2], although the scene can often be viewed from different positions or angles. Based on the concept of asymmetric coding, one view in a stereoscopic pair can be coded with a lower fidelity, while the perceptual quality degradation can be negligible [3]. Thus, stereoscopic video applications may be feasible with moderately increased complexity and bandwidth requirement compared to mono-view applications, even in mobile applications domain [4].

It is desirable to have an Asymmetric Stereoscopic Video (ASV) coding system based on MVC, where one view is compliant to the existing mono-view standard, i.e., H.264/AVC, and the other view can be coded with techniques that provide high efficiency by exploiting redundancies

between views. Approaches have been proposed in the literature to enable inter-view prediction in an ASV codec. In [4], a downsampling process is invoked before inter-view prediction [4]. In [5], direct motion compensation (MC) scheme has been proposed to substantially reduce the complexity without compression efficiency loss.

In the context of mono-view video coding, 2D non-separable adaptive filters have been proposed for the interpolation of values of non-integer sample positions a motion vector points to [6]. In this paper, regionally adaptive filtering algorithms focusing on the integer sample positions are applied to reduce the correlation between the high-resolution picture and the low-resolution picture further and address the potential focus mismatch problem. The proposed techniques target stereoscopic video applications with minor bandwidth increase compared to mono-view video communications but with subjective quality comparable to coding two views with roughly double the bandwidth. The proposed techniques provide about 8% bit-rate saving on average, and 27% bit-rate saving at most, which is equivalent to more than 0.7 dB luma peak signal-to-noise (PSNR) gain for the low-resolution view.

The rest of this paper is organized as follows. In Section II, ASV and the disparity motion compensation methods are described. In Section III, the proposed regionally adaptive filtering algorithm is presented. Implementation details and simulation results are provided in Section IV. Discussions are given in Section V and Section VI concludes the paper.

II. ASYMMETRIC STEREOSCOPIC VIDEO

A typical prediction structure of stereoscopic video is shown in Fig. 1. Pictures in each view form a hierarchical bi-predictive (B) temporal prediction structure. The base view (view 0, denoted as S0) is independently coded and the other view (view 1, denoted as S1) is dependent on view 0. Note that the MVC standard support more views predicted from each other in the view dimension in a hierarchical manner [1]. In ASV, view 0 is in the original resolution (e.g. VGA) and view 1 is coded in a quarter resolution (e.g. QVGA) of view 0. ASV approaches are motivated by the suppression theory of binocular vision [3], which indicates that the perceived sharpness and depth effect of a mixed-resolution stereoscopic

pair is dominated by the higher-quality component. It is foreseen that a 2D mono-view mobile system can be enhanced to a stereoscopic mobile system with about 25% transmission bandwidth and decoder complexity increase.

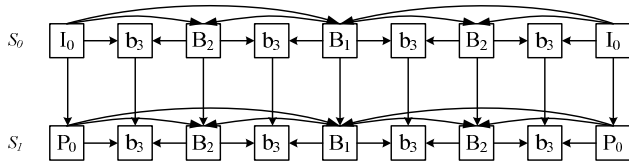


Figure 1. Typical prediction structure for stereoscopic video.

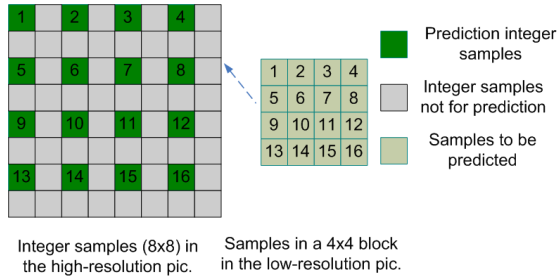


Figure 2. DMC when a motion vector points to integer samples in the picture in view 0.

To decrease the bandwidth of view 1 further, inter-view prediction from view 0 to view 1 can be applied. One solution to enable inter-view prediction has been proposed in [4]. Since inter-view prediction, as the case in MVC, utilizes the motion compensation (MC) in H.264/AVC to realize so-called disparity compensation, view 0 must be downsampled for inter-view prediction [4].

To eliminate the potential extra buffer requirement or complexity increase caused by downsampling, another solution, based on a Direct MC (DMC) from the high-resolution (view 0) pictures, was proposed in [5]. In DMC, disparity compensation is done directly in the high-resolution picture of view 0. So, if a disparity motion vector of a low-resolution (view 0) picture points to integer or half sample positions in the virtually downsampled view 0 picture, it points to even or odd sample positions in the high-resolution (view 0) picture and a MC prediction block is formed from those integer samples. This is illustrated in Fig. 2, where the samples in a 4x4 block can be predicted from 16 pixels in a view 0 picture consisting of either all odd or all even integer samples (each sample is directly predicted from the sample marked with the same number in the figure). If the motion vector virtually points to quarter-sample positions, it points to half-sample positions in the high-resolution picture in view 0 and the DMC averages two neighbouring integer samples [5]. As reported in [5], the performance of the DMC is similar to downsampling based solutions proposed in [4]. However, since the inter-view picture is of high resolution, more information can be potentially utilized for inter-view prediction algorithms, based on DMC.

III. REGIONALLY ADAPTIVE FILTERING

For multiview content, phenomena, such as imperfect calibration, different camera parameters, and focus changes

across views, may lead to less efficiency in the inter-view prediction based on H.264/AVC MC or DMC. During DMC, only $\frac{1}{4}$ of the integer pixels in the high-resolution picture are considered for compensation and $\frac{3}{4}$ of the other pixels in the prediction area are not used. Those pixels can be potentially beneficial in coding of the low-resolution view. Moreover, multiview sequences can have regions with depth level difference. Regions of different depth levels are affected, e.g., blurred at different extents, because of focus mismatch. This makes the picture-level global adaptive filter [7] less efficient, since the resulting filter may not be optimal for some regions. To exploit the information in the high-resolution view more elegantly, a regionally adaptive filtering algorithm is proposed. Multiple optimal filters are applied to different regions of an inter-view picture. The proposed algorithm is realized as follows.

A. Regionally Adaptive Filter Generation

Assume that we have K depth levels in a picture, the following optimization problem is to be solved.

$$H^* = \arg \min_{\mathbf{H}} (e^2) = \arg \min_{\mathbf{H}} \left(\sum_{i=1 \dots K} \sum_{s_p^i \in \mathcal{S}^i} (b_p^i - U_p^i \cdot H^i)^2 \right) \quad (1)$$

wherein $H^* = \{H^1, \dots, H^K\}$ are the K filters and \mathcal{S}^i is the set of the pixels belonging to the i -th depth level. b_p^i are the sample values of pixels in set \mathcal{S}^i and U_p^i are the vectors containing sample values of the group of pixels corresponding to b_p^i . The filter applied in inter-view prediction targeting on the i -th depth level is $H^i = [h_1^i \ h_2^i \ \dots \ h_N^i]^T$, wherein N is the length of the filter and also the number of pixels in a group of corresponding pixels, which are used to obtain one prediction value. To solve this problem, the first step is to identify different sample sets \mathcal{S}^i and the next step is to solve each optimization problem for each depth level. Each problem can be solved by Least Mean Square (LMS) algorithm, the same algorithm used in [7].

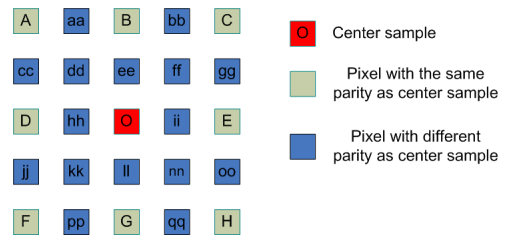


Figure 3. Corresponding group of pixel with 25 samples.

B. Locating the Corresponding Group of Pixels

The best matching sample in the inter-view picture (named center sample in this paper) for a pixel can be located by adding the disparity motion vector to the sample position. A corresponding group of pixels that contain the center sample and the nearest samples (with the same or different parity as the center sample) are shown in Fig. 3. In this paper, all those N (25 even and odd) integer samples are used for filtering a predicted value for one sample in view 1.

C. Disparity Clustering for Depth Level Segmentation

In this sub-section, a method is proposed to segment a picture into different depth regions by the disparity motion vectors (DMVs). To get the DMVs, the corresponding block in the picture of view 0 is found by block matching. We utilize a simple 16x16 block matching algorithm, which reduces the complexity. It is applied for each MB of the view 1 picture and only the integer positions in view 0 are searched.

Given the number of desired depth levels K for a picture, a K -means algorithm is utilized to cluster the DMVs into K classes, by minimizing the following squared error function:

$$E = \sum_{i=1}^K \sum_{v_j \in V_i} \|v_j - \mu_i\|^2 \quad (2)$$

where the K clusters of DMVs are $V_i, i=1 \dots K$ and $\mu_i, i=1 \dots K$ is the centroid (or the mean) of all the DMVs $v_j \in V_i$. $\|\cdot\|$ is the Euclidean norm.

The K -means problem is solved by Lloyd's algorithm [8], in which the centroids are initialized and then updated based on the following repeatedly executed steps:

1. Classify all DMVs into different clusters based on the current centroids. A DMV is classified to the cluster with the nearest centroid to this DMV.
2. Recalculate the centroids: $\hat{\mu}_i, i=1 \dots K$.
3. If the centroids are not changed, that is $\|\mu_i - \hat{\mu}_i\| < \varepsilon, \forall i=1 \dots K$, terminate the iteration; else, set $\mu_i = \hat{\mu}_i, i=1 \dots K$ and return to Step 1.

In this paper, ε is set to 1.

To address multiple depth levels, there are K sets of samples $S^i, i=1 \dots K$ in a picture. After the clustering of the DMVs, segmentation is done to divide the picture into different depth regions, each of which contains a DMV being classified into the corresponding DMV cluster.

After the filters are obtained, they are applied to the inter-view picture to get multiple filtered reference pictures.

D. Relevant Regions Selection for the Adaptive Filters

The optimization problem, described in equation (1), seeks the least squared error solution for a specific sample set in a picture. As a matter of fact, view 1 is coded in a hybrid way, which enables not only inter-view prediction (that utilizes adaptive filters) but also conventional H.264/AVC (intra-view) modes: inter prediction and intra prediction. The MBs or MB partitions for which intra-view modes are selected cannot benefit from the adaptive filter. So allowing those MBs to be considered for the adaptive filter generation can lead to less optimal filters, which are less sensitive to the prediction errors of those samples finally predicted by inter-view prediction. The regions that incline to use the generated filters are the relevant regions for the adaptive filtering generation.

To get the relevant regions, consisting of chosen MBs, the following function is proposed to select the MBs.

$$f(MB_t) = \begin{cases} 1 & \text{Distortion}(MB_t) \leq T \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

wherein $f(\cdot)$ equals to 1 indicates that the t -th MB is selected as relevant MB and $Distortion(\cdot)$ returns the MSE (Mean Square Error) distortion between the original signal and the predicted signal of an MB. So we have:

$$S^i = \{s_t^j \mid s_t^j \in MB_t, f(MB_t) = 1, v_t \in V_i\}, i=1 \dots K \quad (4)$$

For simplicity, MB_t denotes the t -th MB in the picture in view 1. The threshold T is content dependent and it is decided as follows:

$$Rate = |D| / NumMB, D = \{MB_t \mid f(MB_t) = 1\} \quad (5)$$

wherein $|\cdot|$ stands for the cardinality of a set and $NumMB$ is the number of MBs in a picture of view 1. When the $Rate$, the percentage of MBs that are chosen for inter-view prediction, is set, the threshold T can be fixed by ordering the distortion values of all the MBs in a picture. Next section will give more detail on the $Rate$ values.

IV. IMPLEMENTATION AND SIMULATION

The proposed algorithm was implemented into the MVC reference software, JMVM (Joint Multiview Video Model) version 5 [9]. The tested sequences were *Exit*, *Ballroom (BR)*, *Rena*, *Race1*, *Akko&Kayo (AK)*, *Breakdancers (BD)* and *Flamenco2 (FL)*. For each test sequence, the first two views were selected to be coded as view 0 and view 1, respectively, in our simulation. The low-resolution input views were generated by utilizing the MPEG-4 downsampling filter, which is [2, 0, -4, -3, 5, 19, 26, 19, 5, -3, -4, 0, 2]/64. After decoding, the low-resolution pictures were upsampled by the H.264/AVC interpolation filter ([1, -5, 20, 20, -5, 1]/32) for PSNR calculation. Other parameters, e.g., the temporal prediction structure and the motion estimation search range followed the MVC common test conditions specified in [10].

In this paper, a fixed $Rate$ value was used for all pictures in a sequence. The $Rate$ values used are shown in Table I.

The rate distortion (RD) performances for the proposed regionally adaptive filtering (RAF) method, DMC [6] as well as the picture-level adaptive filter (PAF) proposed in [7] were compared and the results are listed in Table II. Note that, in the table, a positive bit-rate saving or a positive Δ PSNR value indicates that the algorithm on the left is better than the right one. The results were generated using the Bjontegaard measurements [11] based on the bit-rates and average PSNR values of the four test points corresponding to different QP values.

TABLE I. RATES FOR DIFFERENT TEST SEQUENCES

Sequence	Exit	BR	Rena	Race1	AK	BD	FL
Rate (%)	75	55	85	60	70	90	80

TABLE II. COMPARISON BETWEEN RAF, DMC AND SIMULCAST (FOR VIEW 1)

Sequence	RAF vs DMC		RAF vs GAF	
	Bit-rate saving	Δ PSNR	Bit-rate saving	Δ PSNR
Akko&Kayo	-2.56%	-0.092	3.61%	0.111
Ballroom	13.41%	0.307	4.66%	0.113
Exit	2.32%	0.042	2.02%	0.039
Race1	-0.47%	-0.008	6.02%	0.114
Rena	27.43%	0.776	0.68%	0.019
Breakdancers	14.08%	0.312	7.86%	0.177
Flamenco2	2.16%	0.081	1.07%	0.039
Average	8.05%	0.203	3.70%	0.087

The proposed method gives more than 10% bit-rate saving for three of the test sequences. The average bit-rate saving is 8% for the low-resolution view. As shown in Table II, on average RAF approximately doubled the bit-rate saving of PAF. RAF improved the compression efficiency for the sequences for which PAF performed poorly, i.e., *Race1*, *Akko&Kayo* and *Exit*. Meanwhile, it further increases the gain for the other sequences significantly, such as *Ballroom* and *Breakdancers*.

The RD curves of the four methods are shown in Fig. 3 for the *Breakdancers* sequence. As shown in Fig. 3, RAF is much better than the simulcast coding, this observation confirms that the proposed RAF based on DMC can achieve substantial bandwidth decrease compared to simulcast coding and make the whole bandwidth much closer to that of 2D mono-view H.264/AVC video. On top of the DMC, adaptive filtering algorithms provided extra gains. The bit-rate saving of RAF is more significant than that of PAF. Note that for PAF, we selected the best configuration in [7]. DMC outperforms simulcast with more than 1.3 dB [5], and the proposed RAF algorithm further increases the coding efficiency. So RAF on top of DMC can greatly reduce the bandwidth of the ASV.

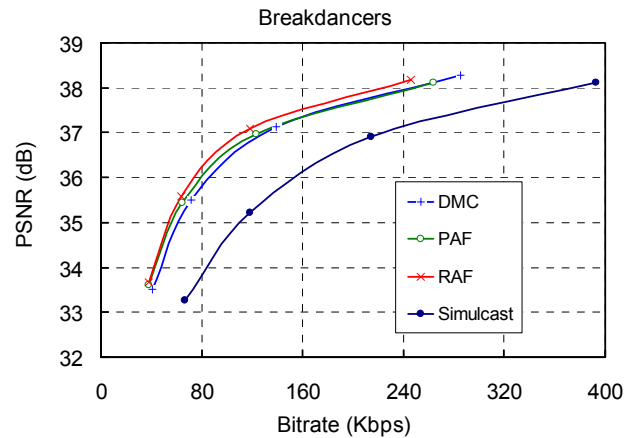
V. DISCUSSION

RAF is suitable for those sequences that have multiple depth levels and provides better coding efficiency than PAF because it better adapts the distribution of the disparity motion vectors. However, it requires more reference pictures, which actually increases the decoder memory requirement. There is always a tradeoff between the resource consumption in the decoder and the bandwidth efficiency; this is especially the case here, as more depth levels require a larger memory size. This is the reason we select 2 or 3 depth levels.

VI. CONCLUSIONS

Asymmetric stereoscopic video coding has the second view coded in quarter resolution compared to the resolution of the H.264/AVC compliant base view, with unnoticeable subjective quality decreases. To further decrease the

bandwidth of the low-resolution view, regionally adaptive filtering method was proposed to generate better predicted signal in inter-view prediction. The main motivation is based on the fact that the objects in a scene can have different depth levels thus different optimizations should be applied for different regions. By searching and classifying the disparity motion vectors, different regions can be segmented and get different filters. Compared with the direct motion compensation method without filtering, on average 8% and up to 27% bit-rate savings can be achieved. With the proposed method, stereoscopic video applications can be realized with minor bandwidth increase to the H.264/AVC based mono-view applications.

Figure 4. Rate distortion curves for the *Breakdancers* Sequence.

REFERENCES

- [1] "Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding," ISO/IEC JTC1/SC29/WG11, Doc. W9978, Hannover, Germany, 2008.
- [2] A. Vetro, W. Matusik, H. Pfister, J. Xin, "Coding Approaches for End-to-End 3D TV Systems," Picture Coding Symposium, 2004.
- [3] Julesz B., Foundations of Cyclopean Perception, University of Chicago Press, Chicago, IL, USA, 1971.
- [4] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, J. Kim, "Asymmetric coding of stereoscopic video for transmission over T-DMB," Proc. 3DTV-CON 2007, Kos Island, Greece, May 2007.
- [5] Y. Chen, S. Liu, Y.-K. Wang, M. M. Hannuksela, H. Li and M. Gabbouj, "Low complexity asymmetric multiview video coding," IEEE International Conference on Multimedia & Expo, 2008.
- [6] Y Vatis, B. Edler, D. T. Nguyen, J. Ostermann, "Motion and Aliasing-Compensated Prediction Using a Two-dimensional Non-separable Adaptive Wiener Interpolation Filter," IEEE International Conference on Image Processing, 2005.
- [7] Y. Chen, Y.-K. Wang, M. M. Hannuksela, and M. Gabbouj, "Picture-level Adaptive Filter for Asymmetric Stereoscopic Video," IEEE International Conference on Image Processing, 2008.
- [8] S. P. Lloyd, "Least Squares Quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.
- [9] P. Pandit, A. Vetro, Y. Chen, "JMVM 5 software," JVT-X208, Geneva, Switzerland, Jun.-Jul. 2007.
- [10] Y. Su, A. Vetro, A. Smolic, "Common Test Conditions for Multiview Video Coding," JVT-T207, Klagenfurt, Austria, Jul. 2006.
- [11] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, Mar. 2001.