Adaptive Filtering in View Synthesis Prediction for Multiview Video Coding

Shinya Shimizu*, Hideaki Kimata*, and Yoshimitsu Ohtani*
* NTT Cyber Space Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847 Japan
E-mail: shimizu.shinya, kimata.hideaki, ohtani.yoshimitsu@lab.ntt.co.jp Tel: +81-46-859-2703

***Abstract*—View synthesis prediction has been studied to achieve efficient inter-view prediction. Existing view synthesis prediction methods generate the predicted pictures by using pictures decoded at the other views and geometric information of the scene. However, it is difficult to obtain such geometric information correctly. In addition, these conventional methods have no ability to compensate the inter-view difference in image signals caused by individual camera characteristics and the non-Lambert reflection of objects.**

The method proposed herein can compensate both the inter-view signal mismatch and incorrect depth information by using an asymmetrical adaptive filter and the weighted average of wiener filter and the median filter. The proposed compensation process is applied to a geometrically compensated picture to minimize the effect of warping-based view synthesis. Experiments show that the proposed method reduces the bitrate by up to 7% relative to view synthesis prediction based on the general adaptive filtering method.

I. Introduction

Multiview video is attracting a lot of interest for realizing advanced visual media services like Free-viewpoint Television (FTV) and three dimensional video (3D Video) [1], [2]. Recent progress on technologies for multiview video processing will make such services possible in the near future.

Multiview video coding is one of the most important technologies because multiview video generates much larger data sets than general mono-view video. The recent international standard MPEG-4 AVC/H.264 Annex.H Multiview Video Coding (MVC) achieves efficient encoding for multiview video [3]. However, the total bitrate yielded by MVC is proportional to the number of views. This means that MVC is not so suitable when a multiview video that consists of a large number of views is to be transmitted. Since a large number of views are required for FTV, multiview auto-stereoscopic displays, and baseline adjustable advanced stereo displays, it is desirable to be able to generate videos at arbitrary viewpoints from just a limited number of viewpoints.

According to the well-known theory of "Plenoptic sampling", depth information of the scene is necessary to generate arbitrary views when multiview video is captured by sparsely arranged multiple cameras [4]. There are two approaches to obtaining the scene depth information at the display side, 1) conducting online depth estimation in the process of view generation at the display side [5], and 2) transmitting depth information estimated at the provider side [6]. The first approach requires not only high computation power at the display side but also relatively dense multiview video material. Therefore, a lot of interest is being paid to implementing the second approach by the representation that combines multiview video with multiview depth maps [7]. One disadvantage of this representation is that it increases the bitrate although MVC can help somewhat.

View synthesis prediction (VSP) is one of the promising technologies with which to efficiently code multiview video by using multiview depth information [8]. A predicted picture is synthesized by warping the image signals of the reference pictures into the coding target view. Compared to disparity compensated prediction, VSP can finely compensate scene geometry. The degree of the preciseness is highly dependent on the correctness of the depth information. However, it is difficult to estimate scene depth correctly and the encoded depth information might have some coding noise. In addition, it is impossible to compensate the inter-view mismatch caused by individual camera characteristics and the non-Lambert reflection of objects. As a result of these problems, existing VSP techniques fail to reduce the bitrate drastically.

In this paper, we propose an adaptive filtering method in VSP that can compensate both geometrical miss-warping and the inter-view mismatch of image signals. In Section II, we describe related works on view synthesis prediction and inter-view signal compensation. The proposed method is presented in Section III. Section IV introduces the experiment conducted and its results, and Section V concludes this paper.

II. Related Works

A. View Synthesis Prediction (VSP)

VSP offers efficient inter-view prediction for MVC. VSP applies pixel-wise warping by using scene depth and camera parameters; its fine compensation ability offers better prediction accuracy. VSP generates the predicted picture by using pixel correspondence. The corresponding pixels are identified by the inverse-projection of pixels and the re-projection of the reconstructed three-dimensional points. Eq.1 defines the inverse-projection and the re-projection is given by Eq.2.

$$g = R_a^{-1} A_a^{-1} \left(u_a, v_a, 1\right)^T d - t_a \tag{1}$$

$$k \left(u_b, v_b, 1\right)^T = A_b R_b \left(g + t_b\right) \tag{2}$$

, where A, R, and t denote the intrinsic matrix, rotation matrix, and translation vector of the camera, respectively. k is a scalar value and d denotes the camera-object distance. (u, v) denotes the pixel coordinates. Subscripts, a and b, denote views. Some

previous works take a as the reference view[9] and others take b as the reference view[8], but there is no inherent difference in the quality of synthesized views. Our VSP takes b as the reference view.

The performance of VSP highly depends on the accuracy of depth information and camera parameters. However, it is almost impossible to obtain them precisely in practical situations. In addition, when depth information used in VSP is either digitalized or encoded, they always contain some distortion. Even though these problems yield large and undesirable effects, no existing VSP method takes these errors into account. In this paper, we propose an adaptive filter model that can compensate some error in the depth information.

B. Inter-view Signal Compensation

One of the problems of simple VSP is its inability to handle inter-view inconsistencies in the image signals because simple VSP uses image signals on the other views as predicted signals. Many studies have tried to overcome this problem. For example, Yamamoto *et al.* proposed color compensation via look-up-tables [10]. This method can handle inter-view illumination changes, but not inter-view focus mismatch.

Adaptive reference filtering (ARF) was proposed to compensate the inter-view focus mismatch [11]. ARF utilizes 2D filters to generate additional inter-view reference pictures whose focus is similar to that of the coding target picture. The coefficients of these filters are estimated to minimize the residual energy of disparity compensated prediction. Because the main purpose of ARF is compensating inter-view focus mismatch, ARF assumes that the filter can be symmetrical with respect to the x- and y-axes.

This paper proposes an adaptive filtering method to increase the performance of VSP. Although ARF can compensate inter-view focus mismatch efficiently, VSP has another big problem which should be considered. It is the difficultly of depth estimation and coding noise on depth information. Our proposal can compensate not only inter-view mismatch but also error in the depth information.

III. Adaptive Filtering For VSP

A. Adaptive Filtering after View Synthesis

Multiview video exhibits inter-view focus and illumination mismatch which are caused by the different camera-object distances and camera heterogeneity. In order to improve the performance of VSP by compensating these while considering the error possible in the depth information, we propose to conduct adaptive filtering as the post process of view synthesis.

The proposed scheme proceeds as follows. First, the geometry-compensated picture is generated by warping the image signals of the decoded inter-view reference pictures, as in the existing VSP. Next, an adaptive filter is constructed by minimizing the difference between the coding target picture and a filtered geometry-compensated picture. The filter model used here is described later. Finally, the constructed filter is applied to the geometry-compensated picture to generate the view synthesized picture.

Compared to the straight-forward combination of VSP and ARF, where view synthesis is performed by using adaptively filtered inter-view reference pictures, our proposal offers two benefits; 1) lower computational complexity and 2) higher prediction accuracy. In terms of computational complexity, the proposed method computes only one filter and applies it once while the conventional approach requires as many filters as inter-view reference pictures used in the view synthesis process. As for prediction accuracy, the proposed scheme can outperform the existing approach because artificial noise in the frequency domain, which is caused by the pixel-wised warping process, is reduced by the filter in the final stage.

B. Adaptive Asymmetric Filter with Median Offset

In order to reduce the influence of depth error by adaptive filtering, we propose an asymmetric filter with median offset for adaptive filtering. The proposed filter model is expressed by Eq. 3.

$$P_{x,y} = \left(\sum_{i=-m}^{m} \sum_{j=-n}^{n} H_{i,j} S_{x+i,y+j} \right) + w \underset{\substack{-n \le i \le n \\ -m \le j \le m}}{Median} \left(S_{x+i,y+j} \right) \tag{3}$$

, where P is the view synthesized picture as the prediction picture, S is the geometry-compensated picture, the subscript (x, y) denotes the pixel position within the coding target picture, and $Median$ is a function that returns the median of the values input. No symmetry is assumed for the filter coefficients $\{H_{i,j}\}$.

The filter coefficients $\{H_{i,j}\}$ and the weight w are derived frame by frame to minimize the difference with respect to the coding target picture:

$$\underset{H,w}{min} \sum_{x,y} \left(O_{x,y} - S_{x,y} \right)^2 \tag{4}$$

, where O is the current frame to be encoded. The optimal parameters can be determined by taking the derivative with respect to each parameter and solving the simultaneous equation where all derivatives are equal to zero.

According to the physical model of video capture, it may be possible to compensate focus mismatches by some kind of symmetric filtering. However, it is inevitable that the geometry-compensated picture will have some small geometrical mismatch against the coding target picture because the quantized depth information can 't express the disparities precisely. These small displacements can be removed by applying a kind of spatial interpolation filter. As a result of combining these filter models, the proposed filter model uses asymmetrical filter coefficients. Since we make no assumption about a priori knowledge of the level of focus mismatch and depth error, all coefficients must be adaptively optimized for each sequence.

The proposed filter model uses median offset in order to increase its robustness toward coding noise in the depth information. When VSP uses depth maps encoded by the conventional video codec, mosquito noise appears around the strong depth edges. This noise results in the salt-and-pepper

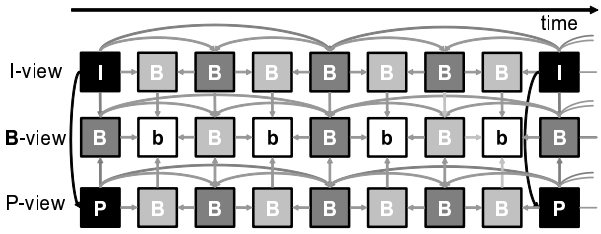

Fig. 1. Hierarchical B structure for MVC(GOP Size 8)

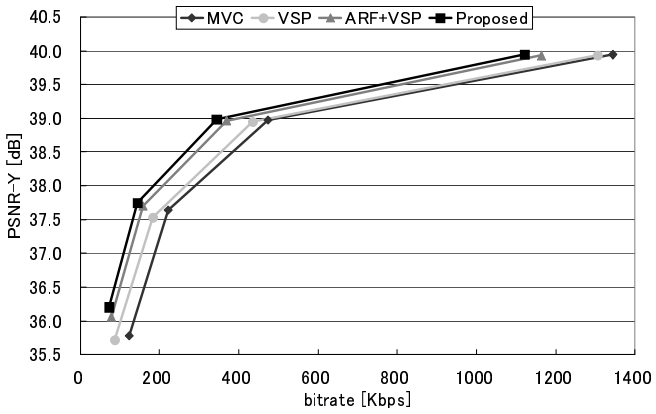

Fig. 2. Overall RD performance

artifacts common in geometry-compensated pictures. It is well-known that the salt-and-pepper artifacts can be reduced by median filtering. However, it depends on the quality of depth information as whether the best results are yielded by using the median filtered value, the adaptively filtered value, or their weighted average. Therefore, the proposed filter model is designed to include a median offset term with a variable weighting factor.

IV. Experiments

A. Conditions

We implemented the proposed method on JMVM version 8.0[12], which is the test model of MVC, and conducted an experiment to assess the efficiency of our method. VSP was implemented as the Inter16x16 mode. This means only 16x16 pixel units were allowed. To distinguish VSP mode from normal Inter16x16 mode, a one bit flag was encoded. The VSP macroblock did not encode the reference index, motion/disparity vector, transform flag, or prediction residuals, i.e. a skip macroblock. We used a 5x5 filter ($m = n = 2$), so the total number of coefficients was 26. Therefore, twenty six 32-bit floating point numbers were encoded in the slice header.

We used the *breakdancers* sequence, which is one of the MVC test sequence provided by Microsoft Research. Multi-view depth maps were also provided [6], In the experiments, the multiview depth maps were encoded by MVC with the Basis QP equal to 36.

We encoded three views, views 3, 4, and 5, all of which contain inter-view focus mismatch, with the reference structure shown in Fig.1. A hierarchical B structure (GOP 8) was applied in the temporal direction. We used one I-view, an H.264/AVC compatible view, one P-view, where inter-view prediction is allowed in one direction, and one B-view, where inter-view prediction is allowed in both directions. The other important coding conditions are listed in Table I.

In order to verify the effects of filtering on the geometry-compensated picture, not reference pictures, we also implemented ARF. Note that only one filter was chosen for one reference picture because it is impossible to use multiple reference pictures for one view in the view synthesis process. We also evaluated the effects of the asymmetrical filter and the use of the weighted median value. In other words, we tested a symmetrical filter with respect to x- and y-axes, a circular symmetric filter, and the compensation model without the median term. The 5x5 symmetrical filter with respect to x- and y-axes can be described as

$$H = \begin{pmatrix} a & b & c & b & a \\ d & e & f & e & d \\ g & h & i & h & g \\ d & e & f & e & d \\ a & b & c & b & a \end{pmatrix}, \tag{5}$$

and the circular symmetric filter is the case of 5 where $h = f$, $c = e = g$, and $b = d$.

B. Experimental Results

Fig.2 plots the rate-distortion curves. The "MVC" curve plots the coding results for H.264/AVC MVC, without VSP, and the "VSP" curve shows the coding performance for the existing VSP, with no compensation. The "ARF+VSP" curve is for the VSP with ARF reference pictures. The "Proposed" curve expresses the performance of the proposed method.

As can be seen, the proposed method achieves the best performance at all rate points examined. At low bitrates, the proposed method not only reduces the bitrates, but also improves the PSNR values. This fact shows that the proposed method succeeds in increasing prediction accuracy. The proposed method is also very effective at high bitrates unlike the existing VSP. The lack of improvement in the PSNR values shows that VSP picture quality is almost the same as the target

TABLE I
Coding parameters

Parameter	Value
GOP size	8
Anchor period	8
Number of reference frames	2
Motion estimation scheme	FME
Entropy coding method	CABAC
Hadamard transform	used
RD-optimized mode decision	used
Layer0 QP (Basis QP)	22, 27, 32, 37
Layer1 QP	Layer0 QP + 3
Layer2 QP	Layer1 QP + 1
Layer3 QP	Layer2 QP + 1

Fig. 3. Example of original image(top), before filtering(bottom-left), and after filtering(bottom-right)

quality, so it may be necessary to encode prediction residuals if the target quality is raised.

Table II shows the bitrate reductions and PSNR gains achieved by the different compensation models. These values are calculated relative to MVC using the Bjøntegaard measure[13]. For the "VSP+AF" condition, no median filtering is applied, which is the same with ARF. The difference between "VSP+AF" and "ARF+VSP" is the order of view synthesis and adaptive filtering. "VSP+AF" applies adaptive filtering after view synthesis while "ARF+VSP" applies adaptive filtering first. From the difference between ARF and axes-symmetric AF, the proposed compensation scheme, which applies adaptive filtering to the geometry-compensated picture, introduces about 1% bitrate reduction. The introduction of the median term and asymmetry constraint bring further bitrate reductions of the order of 3-5% and 1-3%, respectively. The total improvement relative to VSP with the existing ARF method is about 7%. In addition, the proposed method offers improvements in computational complexity because our method requires only one filtering operation on one coding picture while the ARF scheme applies filtering for each reference picture. Fig.3 shows examples of coding target picture, geometry-compensated picture, and VSP picture.

V. Conclusion

We proposed an adaptive filtering method for view synthesis prediction. The proposed method can compensate not only the inter-view signal mismatch caused by the different camera-object distances and camera heterogeneity, but also errors in the depth information by using an asymmetrical adaptive filter with weighted average of the wiener filter and median filter. The proposed compensation process is applied to the view synthesized picture in order to minimize the effect of warping-based view synthesis. Experiments show that the proposed method reduces the bitrate by about 35% relative to MPEG-4 AVC/H.264 Annex.H Multiview Video Coding.

In this paper, multiview depth maps were encoded at the same bitrate even if the total target bitrate was changed. It is obvious that the quality of depth maps affects the coding performance even though the proposed method has the ability to compensate the noise in the depth maps. Therefore, one future work is studying the impact of depth map quality. Furthermore, we plan to expand our VSP scheme by encoding prediction residuals and will consider smaller block sizes.

TABLE II
Simulation results (Bjøntegaard delta)

Method	Symmetry	BD-Rate	BD-PSNR
ARF+VSP	-	28.74%	0.47dB
VSP+AF	No	30.88%	0.50dB
	Axes	29.76%	0.48dB
	Circular	27.49%	0.45dB
Proposed	No	35.55%	0.59dB
	Axes	32.95%	0.54dB
	Circular	32.46%	0.54dB

Acknowledgment

We would like to thank Interactive Visual Media Group, Microsoft Research, for providing the multiview videos and high quality depth maps.

References

[1] Masayuki Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, July 2006.

[2] Aljoscha Smolic, Karsten Müller, Philipp Merkle, Christoph Fehn, Peter Kauff, Peter Eisert, and Thomas Wiegand, "3d video and free viewpoint video - technologies, applications and mpeg standards," in *Proc. ICME2006*, July 2006, pp. 2161–2164.

[3] "H.264 : Advanced video coding for generic audiovisual services," Recommendation, ITU-T, March 2009, Pre-published.

[4] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum, "Plenoptic sampling," in *Proc. ACM SIGGRAPH 2000*, 2000, pp. 307–318.

[5] Yuichi Taguchi, Keito Takahashi, and Takashi Naemura, "Real-time all-in-focus video-based rendering using network camera array," in *Proc. 3DTV-Conference*, May 2008, pp. 241–244.

[6] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, 2004.

[7] Aljoscha Smolic, Karsten Müller, Kristina Dix, Philipp Merkle, Peter Kauff, and Thomas Wiegand, "Intermediate view interpolation based on multiview video plus depth for advanced 3d video system," in *Proc. ICIP2008*, 2008, pp. 2448–2451.

[8] Sehoon Yea and Anthony Vetro, "View synthesis prediction for rate-overhead reduction in ftv," in *Proc. 3DTV-Conference*, May 2008, pp. 145–148.

[9] Yuichi Taguchi and Takeshi Naemura, "View-dependent coding of light fields based on free-viewpoint image synthesis," in *Proc. ICIP2006*, October 2006, pp. 509–512.

[10] Kenji Yamamoto, Masaki Kitahara, Hideaki Kimata, Tomohiro Yendo, Toshiaki Fujii, Masayuki Tanimoto, Shinya Shimizu, Kazuto Kamikura, and Yoshiyuki Yashima, "Multiview video coding using view interpolation and color correction," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17, no. 11, pp. 1436–1449, 2007.

[11] P. Lai, Y. Su, P. Yin, C. Gomila, and A. Ortega, "Adaptive filtering for cross-view prediction in multi-view video coding," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, January 2007, vol. 6508.

[12] Pandit Pandit, Anthony Vetro, and Ying Chen, "Jmvm 8 software," JVT Doc. JVT-AA208, April 2008.

[13] Gisle Bjøntegaard, "Calculation of average psnr differences between rd-curves," VCEG Doc. VCEG-M33, April 2001.