Overview of View Synthesis Prediction for Multi-view Video Coding

Yo-Sung Ho, Cheon Lee, and Kwan-Jung Oh

Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
{hoyo, leecheon, kjoh81}@gist.ac.kr

Abstract: With a wide range of viewing angles, the multi-view video can provide more realistic feeling at arbitrary viewpoints. However, because of a huge amount of data from multiple cameras, we need to develop an efficient coding method. One of the promising approaches for multi-view video coding is view synthesis prediction which generates an additional reference frame. In this paper, we explain the principal idea of view synthesis prediction and propose a view interpolation and coding scheme. For view interpolation, we employ initial disparity estimation, variable block matching, and disparity error correction methods. After generating an intermediate view image, we apply a 'VIP P-picture' coding method with the additional reference frame, where we have five additional motion estimation modes and a modified motion vector prediction scheme. By computer simulations, we have improved the coding gain by 0.66 dB on average for the well synthesized sequences.

1. Introduction

Improvements of signal processing and network transmission techniques have enabled various multimedia services. Recently, demands for interactive and realistic contents are growing rapidly. Multi-view video is a good candidate to satisfy such demands. However, as multiple cameras are used to acquire the multi-view video, the amount of data to be processed increases tremendously. Since it is a serious limitation for applying distributive services, it is required to develop an efficient coding scheme for the multi-view video without significant sacrifice of visual quality [1-2].

The multi-view video can be applied to various areas including FVV (free viewpoint video), FTV (free viewpoint TV), 3DTV, surveillance, and home entertainment. Currently, the JVT (joint video team) of MPEG (moving picture experts group) and VCEG (video coding experts group) is working on the standardization of MVC (multi-view video coding).

Unlike the single view video, since the multi-view video has high correlation among neighboring image frames, we can use reconstructed images at the adjacent viewpoints to encode images at the current viewpoint. In this effort, several algorithms have been studied and the view interpolation prediction is one of the promising approaches for efficient multi-view video coding. We can generate a synthesized image for the current view by using depth/disparity information and the adjacent reconstructed images. In order to encode the current viewpoint image, we can exploit the interpolated image as an additional reference frame.

In this paper, we describe two methods for view synthesis: 3D image warping and view interpolation. Then, we propose an efficient view interpolation scheme for multi-view video coding, and a 'VIP P-picture' coding method that exploits the synthesized image as an additional reference frame.

2. View Synthesis for Multi-view Video Coding

The view synthesis prediction is a reliable approach to improve the coding efficiency of MVC. In order to generate an additional reference image, we can use a 3D warping or view interpolation method. While the 3D warping method performs a projection process using the depth information and camera parameters, the view interpolation method exploits disparity information between the adjacent viewpoint images. Since the depth and disparity information are not supported in most cases, we need to obtain such information from the input images.

2.1 3-D Warping

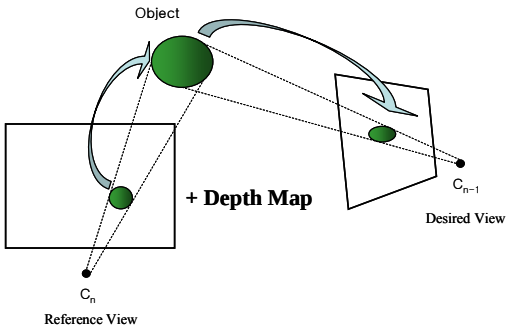


Fig. 1. 3-D Warping

If the depth information is available at every viewpoint, a virtual viewpoint image can be rendered from the nearby viewpoint images by projecting the pixels of the original image to their proper 3D locations and re-projecting them onto the new picture [3]. If the multi-view video supports camera parameters, we can apply the 3D warping method.

Let A_n, R_n, and t_n denote the intrinsic matrix, the rotation matrix, and the translation vector of the camera C_n, respectively. Using a point $\mathbf{x}_n = [x_n, y_n]$ on the image I_n captured by C_n and the depth of the point $D(\mathrm{x}_n)$ can be mapped onto other camera I, where $i \in \{1 \ldots N\}$. N is the number of cameras. At first, the point is projected into the 3D space from the 2D image plane by

$$X = R_n \cdot A_n^{-1} \cdot [\mathbf{x}_n \ 1]^T \cdot D(\mathbf{x}_n) + t_n \tag{1}$$

where X denotes the 3D point. Then, a point X in the 3D space can be projected onto the desired viewpoint, for instant I_{n-1}, by

$$\mathbf{x}_{n-1} = A_{n-1} \cdot R_{n-1}^{-1} \cdot (X - t_{n-1}) \tag{2}$$

After mapping x_{n-1}, we can obtain an image re-projected into the desired viewpoint.

2.2 View Interpolation

The view interpolation method, proposed by Chen and Williams [4], can be used to reconstruct arbitrary viewpoints using the optical flow between two input images. Based on this method, Droese [5] proposed a new view interpolation method in the disparity domain for the multi-view video. Here, the disparity can be defined as a pixel distance in the horizontal coordinate of the corresponding pixels in the two images, which is described as

$$I_L(x, y) = I_R(x + d, y) \tag{3}$$

where d is the disparity of a pixel at an intermediate viewpoint between the left and right images. It can also be represented by two disparities with a parameter α ($0 \le \alpha \le 1$) that is related to the position of the pixel at the intermediate viewpoint between two anchor images.

$$\begin{aligned} I_\alpha(x,y) &= I_L(x+\lfloor \alpha d \rfloor, y) \\ &= I_R(x+\lfloor (1-\alpha) d \rfloor, y) \end{aligned} \quad (4)$$

where $\lfloor \rfloor$ is a rounding operation to integer values and $I_\alpha(x,y)$ indicates the intensity value at the intermediate viewpoint.

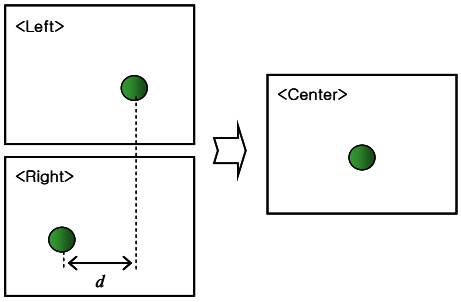


Fig. 2. View Interpolation

3. Depth Estimation for View Synthesis

Ince *et al.* [6] has developed a depth map estimation algorithm for view synthesis prediction in multi-view video coding. For depth map estimation, we can match the corresponding epipolar lines by employing the epipolar geometry.

Combining Eq. (1) and Eq. (2), we can write x_{n-1} as a function of x_n and $D(x_n)$ within a scaling factor by

$$\mathbf{x}_{n-1}(\mathbf{x}_n, D(\mathbf{x}_n)) = A_{n-1} \cdot R_{n-1}^{-1} \cdot (R_n A_n^{-1} [\mathbf{x}_n \ 1]^T D(\mathbf{x}_n) + t_n - t_{n-1}) \quad (5)$$

Using Eq. (5), we try to minimize the prediction error $P(x)$ among all possible depth values.

$$P(\mathbf{x}) = \Psi\left(I_n[\mathbf{x}_n] - I_{n-1}[x_{n-1}(\mathbf{x}_n, D(\mathbf{x}_n))]\right) \quad (6)$$

where Ψ is an error function. A depth value is determined by minimizing

$$D(\mathbf{x}) = \arg\min_{D_i(\mathbf{x})} P(\mathbf{x}) \quad (7)$$

where $D_i(x) = D_{min} + iD_{step}$, $i = \{0 \ldots (D_{max} - D_{min})/K\}$, and K is the number of possible depth values.

Based on the block matching operation, Ince *et al.* proposed several methods to improve the depth map: hierarchical estimation, regularization, and median filtering. In Fig. 3, we have compared those depth maps visually.

4. Proposed View Interpolation Scheme

4.1 Initial disparity estimation

Most stereo matching algorithms set the maximum disparity value. However, the accurate maximum value is very difficult to predict before actual disparity observation. To overcome this problem, we estimate the initial disparity. Using the coarse initial value, we can find more accurate disparities in the next step.

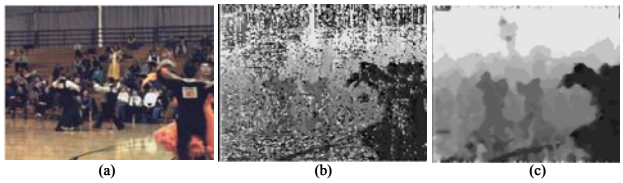


Fig. 3. Estimated Depth Maps. (a) Original Image of Ballroom sequence (b) Result of original block-based depth estimation. (c) Final result of improved depth estimation.

Since stereoscopic images have the ordering constraint property, we can estimate disparities hierarchically using region subdivision. After we examine the most outstanding block of steep luminance changes with the search range of [0, width], we divide the image into two regions according to the position of the outstanding block. The second outstanding block is also estimated with a reduced search range at the divided region repeatedly. During this process, each block has one disparity value.

4.2 Variable block-based disparity estimation

Since the Droese's method uses block-based disparity estimation, it may detect a wrong disparity value when the given block is located at the boundary region. To avoid this problem, we propose a variable block-based disparity estimation algorithm, as shown in Fig. 4. Once we determine a disparity for a basic block, we compare the cost values of the large and small blocks. If the difference of two cost values is larger than the pre-determined threshold value T, we subdivide the block into smaller blocks for disparity estimation. The minimum block size is 2x2.

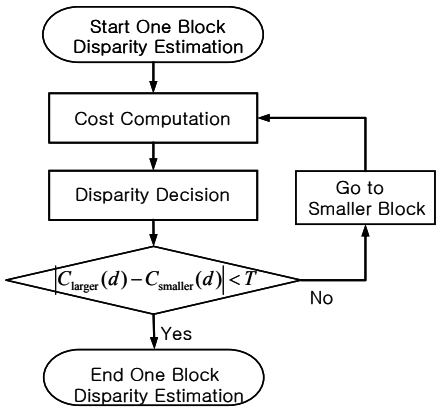


Fig. 4. Variable Block-based Disparity Estimation

4.3 Pixel-level disparity estimation

The final disparity estimation is performed at the pixel level. The search range is adjusted by Eq. (8), where $D(x,y)$ is the disparity obtained at the previous step. For pixel-level disparity estimation, a small search range is enough to estimate the accurate disparity. The procedure for pixel-level disparity estimation is similar to the previous step, but the cost function is modified by adding C_{StDev} term, as shown in Eq. (9). C_{StDev} means the difference of standard deviations for the matching and reference blocks.

$$\begin{aligned} \text{MinRange} &= D(x,y) - \text{SearchRange}/2 \\ \text{MaxRange} &= D(x,y) + \text{SearchRange}/2 \end{aligned} \quad (8)$$

$$\begin{aligned} C(x,y,d) = C_{sim}(x,y,d) &+ \lambda \cdot C_{reg}(x,y) \\ &+ \gamma \cdot C_{StDev}(x,y,d) \end{aligned} \quad (9)$$

4.4 Disparity error correction

In order to improve accuracy and consistency of disparities, we correct disparity errors. If several regions of different disparity values belong to one object, some regions may have erroneous disparity values. By checking the cost value using a large block, we can replace those disparities by one disparity value. Median filtering is very useful to reduce disparity estimation errors.

4.5 View synthesis using disparity map

The center view image can be synthesized by

$$I_\alpha(x,y) = (1-\alpha)\cdot \hat{I}_L(x+\alpha\cdot D(x,y),y) + \alpha\cdot \hat{I}_R(x+(\alpha-1)\cdot D(x,y),y) \quad (10)$$

where $\hat{I}$ is interpolated value, when $\alpha \cdot D(x,y)$ or $(\alpha-1)\cdot D(x,y)$ is not an integer. Here we assume that illumination is changed linearly.

5. 'VIP P-picture' Coding

Fig. 5 shows the basic coding structure of MVC that employs H.264/AVC [7]. We call the image frames at T0 and T8 as the 'anchor frames' that are coded using only inter-view prediction. View interpolated prediction is applied to the B-views, such as S1, S3, and S5. For images in the B-view, we synthesize the intermediate images using two neighboring images in the adjacent views at the same temporal reference.

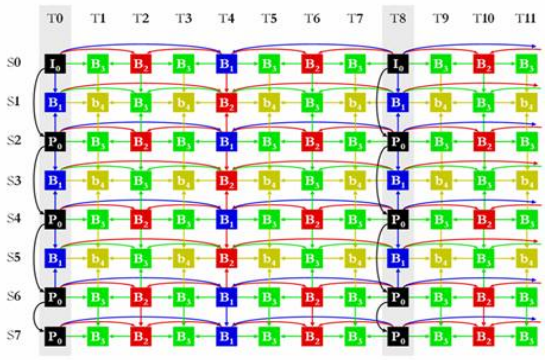


Fig. 5. Reference Prediction Structure of MVC

5.1 Additional coding modes

We propose a view interpolated prediction P-picture (VIP P-picture) coding method where we use the synthesized intermediate image as an additional reference frame for encoding the B-views. In 'VIP P-picture' coding, we have five additional macroblock modes: 'VIP_SKIP', 'VIP_16x16', 'VIP_8x16', 'VIP_16x8', and 'VIP_P8x8'. We call these modes as the VIP modes. Especially, the 'VIP_SKIP' mode refers to a co-located block of the synthesized intermediate image and it does not encode the motion information and residual data at all.

5.2 Modified motion vector prediction

In H.264/AVC, we encode the difference between the real and predicted motion vectors. In this work, we modify the current motion vector prediction scheme to be more suitable for VIP. If a macroblock mode is not a VIP mode, its motion vector is predicted using the existing method. However, if a macroblock mode is one of the VIP modes, the motion predictor considers only the neighboring blocks that are encoded by referring to the synthesized intermediate image. The modified motion vector prediction scheme reduces coding bits for the motion vector.

6. Experimental Results

In order to evaluate the proposed view interpolation scheme and the 'VIP P-picture' coding method, we check PSNR values of the synthesized intermediate images and coding efficiency of the 'VIP P-picture' coding. We use the JMVM 1.0 (joint multi-view video model) provided by MPEG/JVT as the reference software. 'Akko&Kayo', 'Rena', and 'Ballroom' sequences are used for the computer simulations.

The size of the basic block for the disparity estimation is 16x16. The threshold value in Fig. 4 is chosen to be 13, and the search range for the pixel-level disparity estimation is 5~15. The window size of the median filter is 12x12. The reference coding structure is shown in Fig. 5 [7]. The QP values are 22, 27, 32, and 37 [8]. For the existing and new VIP picture coding methods, the search ranges for motion estimation are 96 and 48, respectively.

6.1 Determination of center view position

In general, we assume that the center view is located in the middle ($\alpha = 0.5$) of two adjacent views. However, sometimes the center view is not located at the center position. To derive the accurate position of the center view, we calculate α using Eq (9). For 'Akko&Kayo' and 'Rena' sequences, α is close to 0.5; however, α for 'Ballroom' is quite different from 0.5. The calculated α values for S1, S3, and S5 are 0.557, 0.471, and 0.533, respectively.

$$\alpha = \frac{-D_L}{|D_R - D_L|} \quad (9)$$

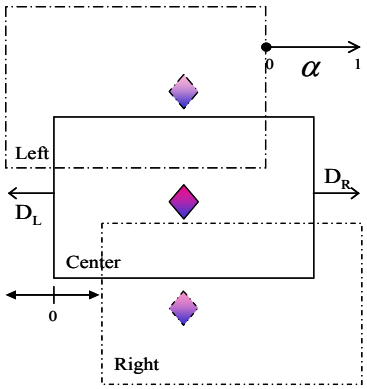


Fig. 6. Alpha Value Calculation

6.2 Results of view interpolation scheme

Table 1 shows the average PSNR values for the synthesized intermediate images. The proposed method improves the quality of interpolated images by 1~4 dB. 'Akko&Kayo' and 'Rena' show better results than 'Ballroom' in terms of the PSNR values because the former two sequences have less occlusion regions and those disparity values are smaller than the latter one.

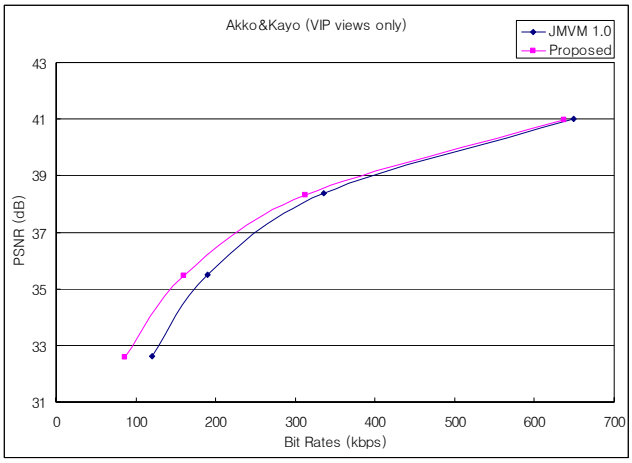
Table 1. View Interpolation Results: Average PSNR of 30 Frames

Test Sequences	Previous Method			Proposed Method		
	Max. Search Range			Search Range		
	30	40	50	5	10	15
Akko&Kayo	27.8	31.5	30.4	33.0	32.7	32.3
Rena	28.4	27.5	26.4	32.6	32.7	32.8
Ballroom	20.7	21.0	21.4	25.3	25.3	25.3

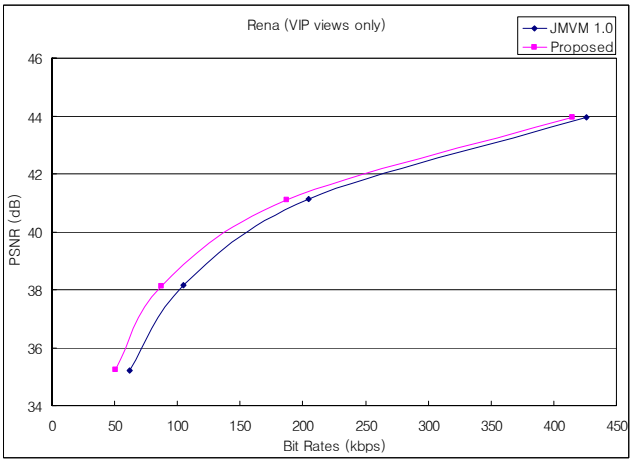
Unit: dB

6.3 Results of 'VIP P-picture' coding

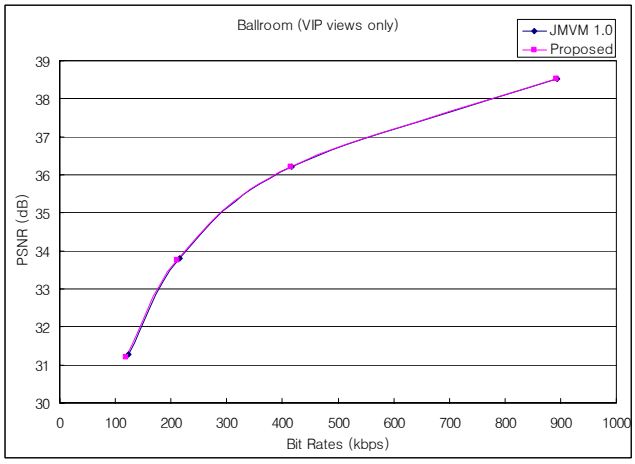
The new 'VIP P-picture' coding method is applied to the B-views. The number of B-views for 'Akko&Kayo', 'Rena', and 'Ballroom', are 6, 7, and 3, respectively. Figure 7 shows the experimental results for the 'VIP P-picture' coding. The proposed scheme shows similar or better results, compared to JMVM 1.0.



(a) Rate-distortion Curve for 'Akko&Kayo'



(b) Rate-distortion Curve for 'Rena'



(c) Rate-distortion Curve for 'Ballroom'

Fig. 7. Results of 'VIP P-picture' Coding

As shown in Fig. 7, the PSNR values of 'Akko&Kayo' and 'Rena' are improved by 0.74 dB and 0.57 dB, respectively. The result of 'Ballroom' is quite similar to the reference model since synthesized intermediate images of 'Ballroom' are not good enough to contribute to coding efficiency. These results indicate that improvement of the coding efficiency depends on the quality of the synthesized intermediate image.

7. Conclusion

In this paper, we have explained the main concept of view synthesis prediction and proposed a new view interpolation scheme for multi-view video coding. The proposed view interpolation scheme consists of initial disparity estimation, variable block-based disparity estimation, and disparity error correction methods. We have also proposed an efficient 'VIP P-picture' coding method that exploits the synthesized intermediate image as an additional reference frame for multi-view video coding. Experimental results demonstrate that the proposed view interpolation scheme improves image quality of the synthesized intermediate image and the proposed 'VIP P-picture' coding method achieves the PSNR gain by 0.66 dB on average over JMVM 1.0.

Acknowledgements

This work was supported in part by MIC through RBRC at GIST, and in part by MOE through the BK21 project.

References

[1] A. Smolic and P. Kauff, "Interactive 3D Video Representation and Coding Technologies," *Proceedings of the IEEE*, Spatial Issue on Advances in Video Coding and Delivery, Vol. 93, pp. 99-110, 2005.

[2] A. Smolic, K. Mueller, T. Rein, P. Eisert, and T. Wiegand, "Free Viewpoint Video Extraction, Representation, Coding, and Rendering," *Proc. of IEEE International Conference on Image Processing*, Vol. 5, pp. 3287-3290, 2004.

[3] H. Y. Shum, and S. B. Kang, "A Review of Image-based Rendering Techniques," *In IEEE/SPIE VCIP 2000*, pp. 2–13, 2000.

[4] S. Chen and L. Williams, "View Interpolation for Image Synthesis," *Computer Graphics (SIGGRAPH '93)*, pp. 279-288, 1993.

[5] M. Droese, T. Fujii, and M. Tanimoto, "Ray-Space interpolation based on Filtering in Disparity Domain," *Proc. 3D Conference*, pp. 213-216, 2004.

[6] S. Ince, E. Martinian, S. Yea and A. Vetro, "Depth estimation for view synthesis in multiview video coding," *Proc. of IEEE 3D TV Conference*, 2007.

[7] ISO/IEC JTC1/SC29/WG11, "Description of Core Experiments in MVC," N8019, 2006.

[8] JVT of ISO/IEC MPEG & ITU-T VCEG, "Common Test Condition for Multiview Video Coding," U211, 2006.