A BACKWARD COMPATIBLE 3D SCENE CODING USING RESIDUAL PREDICTION

Shinya Shimizu, Hideaki Kimata, Kazuto Kamikura, and Yoshiyuki Yashima

NTT CyberSpace Laboratories, NTT Corporation

ABSTRACT

In order to represent 3D space, we have proposed to use the representation that consists of multi-view video plus a single view depth map. This format is backward compatible with the MPEG-C Part 3 (a.k.a. ISO/IEC 23002-3). This paper proposes a coding scheme on this 3D space representation that is efficient even if low delay decoding functionality is required. We apply the residual prediction framework to the view synthesis prediction errors. Experiments show that the proposed scheme achieves up to about 7.7% bitrate reduction compared to multi-view video coding with disparity compensation, even if the depth map video is added. Furthermore, the proposed scheme doesn't require any syntax changes to the conventional video coding standard and it needs few modifications on the circuits of conventional video codecs; it might be possible to reuse almost all of the components. The backward compatibility and reusability achieved by the proposed scheme are quite important to reduce manufacturing costs and time to market.

Index Terms— multi-view video, depth map, MVC, MPEG-C Part 3, backward compatibility

1. INTRODUCTION

Free viewpoint television (FTV) and three-dimensional (3D) video are attracting a lot of interest from not only the research area but also the marketplace because they can provide a highly realistic sensation, which is important in many fields, e.g. entertainment and education [1, 2]. Many kinds of formats have been proposed to represent these advanced visual media.

Recently, the ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) released MPEG-C Part 3 (a.k.a. ISO/IEC 23002-3), which is single video plus per sample depth. This format can realize efficient and backward compatible 3D video because the video stream of the video map is attached to the conventional single video stream. This backward compatibility is very important in reducing manufacturing costs and time to market. Unfortunately, it is impossible to offer a very strong depth feeling and large parallax with this format because of the large occlusion areas created in such cases. Furthermore, it is very difficult to support free viewpoint navigation.

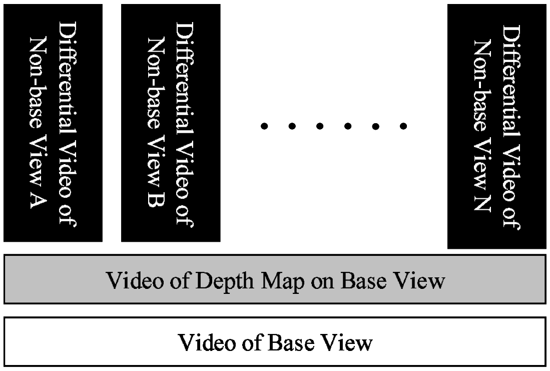
Differential Video of Non-base View A
Differential Video of Non-base View B
. . . . . .
Differential Video of Non-base View N
Video of Depth Map on Base View
Video of Base View

Fig. 1. Proposed hierarchical bitstream structure.

Multi-view video (MVV) is another type of representation for 3D scenes. MVV is simple but has the ability to support 3D videos that have large parallax and wide free viewpoint navigation. MVV is very useful in storing and displaying 3D scenes as it is; Multi-view Video Coding (MVC) is currently under development in the Joint Video Team (JVT) of MPEG and ITU-T VCEG to realize future FTV and 3D video applications [3].

However, a lot of operations are required if the same MVC bitstream is to be displayed on a multitude of different terminals, which have different requirements in terms of the number of views and/or the distance among views. Furthermore, this format has backward compatibility only with the conventional single view video, not with MPEG-C Part 3. This is not preferable to the market because MPEG-C Part 3 has already been released as 3D scene representation.

We have proposed another a format that consists of a single view video on a certain viewpoint, depth map on the same viewpoint of the single view video, and some differential videos [4]. We call the viewpoint of the first component the base view. The first two components are completely the same as the components of the MPEG-C Part 3. This format can be encoded into the hierarchical structure illustrated in Fig.1. As can be seen, this format has backward compatibility with the MPEG-C Part 3. Moreover this format can be converted into MVV, because one differential video stands for the differences between a single view video on one viewpoint and a synthesized video, which is generated from the base view video and the depth map. This makes it possible to offer strong depth feeling, large parallax, and wide range free viewpoint navigation.

1-4244-1484-9/08/$25.00 ©2008 IEEE

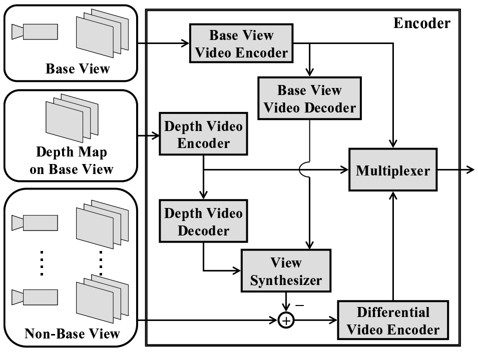

Fig. 2. Hierarchical encoder of MVV plus single depth video.

Unlike MVV, this format requires an encoded depth map which raises the possibility of increasing the total bit amount. Our solution is an efficient coding framework, where all components are encoded by the conventional single view video encoder as if they were regular video signals [4]. Fig.2 overviews the previously proposed hierarchical encoder. Actually, this encoder can achieve totally efficient compression, but there is a problem with encoding differential videos without any temporal prediction because differential videos have less spatial correlation than regular videos.

To achieve FTV functionality, the images to be used for view generation should be decoded as fast as possible and user has free choice in requesting the view. Given this background, when low delay decoding across both time and view is required, it is necessary to encode differential videos without any temporal prediction. It is due to this requirement that the previously proposed encoder fails to achieve efficient compression. In this paper, we propose a new coding scheme that achieves efficient compression while keeping as many of the benefit as possible.

This paper is organized as follows. Section 2 provides a brief review of the previous framework; its problems are also discussed. We propose the new coding scheme in Section 3. The experiment and its results are presented in Section 4. Finally, we conclude in Section 5.

2. HIERARCHICAL CODING OF MVV AND SINGLE DEPTH WITH USING CONVENTIONAL CODER

Already mentioned above, our earlier proposal encodes MVV after converting it into the representation that consists of base view video, depth map on base view, and differential videos on the other views [4]. A differential video is generated by subtracting a synthesized video, which is generated from the base view video and its depth map by using computer vision techniques, from the normal single view video on another viewpoint. In order to achieve totally efficient compression, we also proposed to encode all the components individually by the conventional video encoder as if they were regular video signals, and compose a hierarchical bitstream as illustrated in Fig. 1.

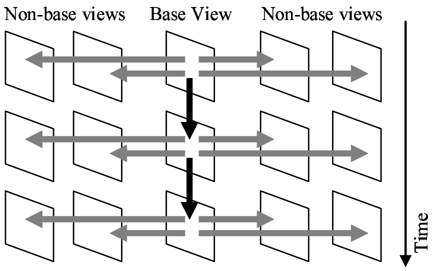

Fig. 3. Prediction structure for low delay decoding.

This hierarchical structure is one of the features of this framework. It realizes backward compatibility with MPEG-C Part 3. This is highly desired by the market. Another benefit of the hierarchical structure is that extremely flexible view scalability is achieved. Thus it is possible to transmit/decode only those views that are unnecessary. This flexible view scalability is very useful in enhancing interoperability because it becomes possible for the same bitstream to be displayed on a multitude of different terminals and over networks with various performance attributes. Another feature of this scheme is that conventional video encoders can be used. This feature significantly reduces development costs and time to market.

It is obvious that there are spatial and temporal correlations on depth, which expresses camera-object distance. Most of the signals in a differential video are caused by incorrect depth, inaccurate camera parameters, mismatch of projection model, local illumination changes, and occlusions. Generally these factors are spatially and/or temporally correlated. Therefore, this framework can achieve totally efficient compression by using a conventional predictive video coding standard like H.264/AVC.

Unfortunately, there is almost no spatial correlation between the signals of occluded areas and unoccluded areas. This is because most of the original video signals are removed in the unoccluded areas but not in the occluded areas. Accordingly, it is impossible to predict the signals spatially beyond the boundary of an occlusion. In other words, compression performance around the boundaries of occlusions is quite low if intra prediction is applied to encode the signals on such areas in differential videos.

The occluded areas are, however, temporally correlated, so it is possible to predict the signals if temporal prediction can be applied. This is why the previously proposed scheme can achieve totally efficient compression without being affected by these spatial discontinuities.

Thinking about the FTV application, it is necessary to decode the images which are used for generating the image on user's selected viewpoint as fast as possible. There is no idea which viewpoint is selected in advance. This means that it should support a low delay decoding across both time and view. It is possible to fulfill this requirement by using the special prediction structure showed in Fig.3.

Following this prediction structure, all differential videos on non-base views should be encoded without temporal pre-

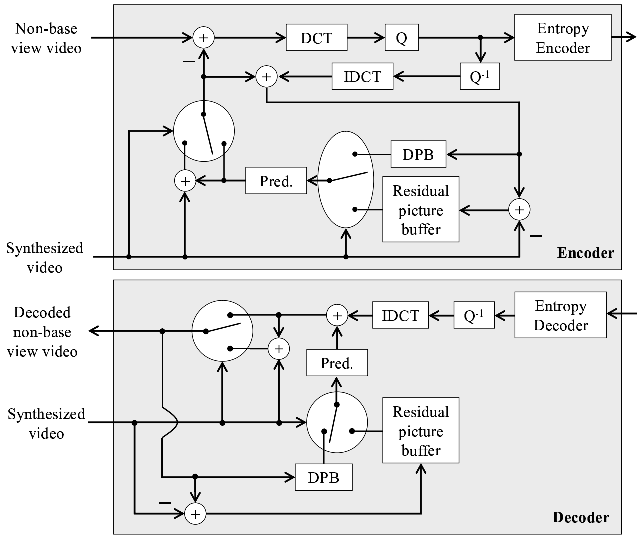

Fig. 4. Proposed non-base view video encoder/decoder.

dictions. Therefore, the previously proposed framework fails to achieve totally efficient compression. This paper proposes an adaptive encoding scheme that keeps the hierarchical structure of bitstream while minimizing the number of modifications that need to be made to the conventional predictive video coding scheme.

3. ADAPTIVE RESIDUAL PREDICTION FRAMEWORK

Fig.4 shows the proposed encoder/decoder for the non-base views. Both base view video and depth map are coded using the previously proposed framework, in other words, they are coded individually by the conventional single view video coder (see Fig.2).

Already mentioned in Section 2, the degradation in coding efficiency is mainly caused by the fact that there is almost no spatial correlation between the differential video signals in occluded and adjacent unoccluded areas. But there might be spatial correlation between its original video signals in the occluded area and those in the adjacent unoccluded area. So it might be possible to increase the coding efficiency by using original video signals as reference of prediction on occluded areas.

The decoder can know the areas where occlusion is happened by checking view synthesized video, which must be generated to decode non-base view videos, because the view synthesis fails at occluded areas. Fig.5 shows two examples of view synthesized pictures on different views and different times. View synthesis was failed at the black areas, which include all the uncovered areas, not only occluded areas. Therefore, in the proposed scheme, codec changes the reference signals adaptively by means of the failure of view synthesis. To be exact, if there is at least one pixel where view synthesis is failed in a current coding block, this block uses local decoded original video signals of its non-base view as reference. Otherwise, the block uses local decoded differential video signal as reference.

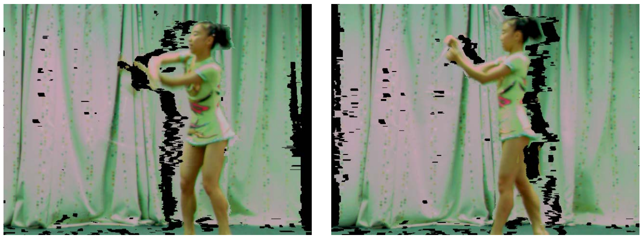

Fig. 5. Examples of view synthesized pictures.

It might be possible to realize the same adaptive selection by encoding additional information bits at each block. However, in contrast to this method, the proposed one can not only prevent bitrate increase but also avoid syntax changes. This is valuable in reducing manufacturing costs.

Introducing the adaptive reference raises inconsistency in the bit depth of coding signals. This inconsistency increases the circuit size, which raises manufacturing cost. Therefore, the proposed scheme uses residual prediction on the view synthesized error signals. These error signals are the same as those present in the differential video signals because both are the differences between the original video and the view synthesized video. Thus this residual prediction means the prediction on differential video. In the residual prediction framework, the predicted signals are added to the view synthesized signals to generate signals to predict the original video signals. As a result, the bit depth of coding signals becomes consistent in this residual prediction framework; it is possible to minimize the increase in circuit size.

Employing the residual prediction framework also offers consistency with conventional prediction methods such as disparity compensation prediction (DCP) in the bit depth of both coding target signals and prediction signals. This consistency makes it possible to use any of the conventional inter-view prediction methods. Offering this option can prevent compression efficiency from decreasing in regions where the quality of view synthesizing is poor.

4. EXPERIMENTAL RESULT AND DISCUSSION

In order to assess the efficiency of our coding scheme, we conducted an experiment. We used two MVC test sequences, ”Rena” and ”Akko&Kayo” for the experiment [5]. Rena sequence is composed of 16 views arranged in a horizontal line. Akko&Kayo sequence is composed of 15 views arranged in a 2D array. We set the center view, numbered 46 in Rena and 48 in Akko&Kayo, as the base view. We implemented the proposed scheme and the scheme described in [4] on JMVM, which is a reference encoder of MVC. We used fixed QPs. Depth map was generated by using the method described in [6], which is based on the multiple-baseline stereo algorithm. We used 3D warping for view synthesis. Depth map video

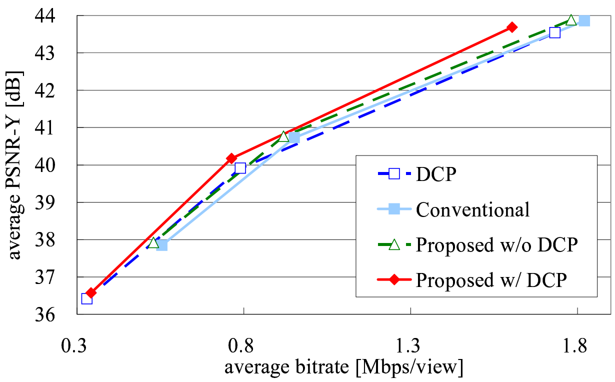

Fig. 6. Total RD performances on Rena.

Table 1. Bitrate reduction and PSNR gain

	vs DCP		vs Conventional	
	Bitrate	PSNR	Bitrate	PSNR
Rena				
Conventional	-1.81%	-0.05dB	-	-
Proposed w/o DCP	2.04%	0.11dB	3.70%	0.16dB
Proposed w/ DCP	7.70%	0.35dB	10.06%	0.45dB
Akko&Kayo				
Conventional	-17.99%	-0.80dB	-	-
Proposed w/o DCP	-5.38%	-0.28dB	11.03%	0.53dB
Proposed w/ DCP	6.17%	0.18dB	21.18%	1.04dB

was encoded losslessly with down sampling of the resolution. We used the low delay prediction structure illustrated in Fig.3.

Fig.6 plots the rate-distortion curves on Rena. The curve labeled ”DCP” means all non-base views were encoded by using DCP, which is the most popular method to encode MVV. We used JMVM to create this curve. ”Conventional” shows the coding performance achieved by the scheme described in [4]. ”Proposed w/o DCP” stands for the performance by the proposed scheme without using DCP: all pictures of non-base views were encoded as intra pictures. ”Proposed w/ DCP” shows the performance achieved by the proposed scheme using DCP: all pictures of non-base views were encoded as inter-view pictures. Except for the bitstream labeled ”DCP”, all the bitstreams included the depth map. Table 1 shows the bitrate reductions and PSNR gains as indicated by the Bjontegaard measure [7].

As it can be seen, the conventional scheme failed to achieve efficient compression. This is because the signals of the differential videos become less spatially correlated around the boundary of occlusion areas. The proposed scheme, on the other hand, offered higher encoding efficiency. Compared to the conventional scheme, a bitrate reduction of up to about 21.2% was achieved even for the prediction structure with low delay. Compared to MVC with DCP, the proposed scheme achieved up to about 7.7% bitrate reduction even if depth map video is included in only the proposed bitstream. Note that the big gap between ”Proposed w/ DCP” and ”Proposed w/o DCP” may be caused by the poor quality of the view synthesized videos; this means that there is a possibility of achieving more efficient compression by using a different view synthesize algorithm and a good depth map.

5. CONCLUSIONS

We have proposed an efficient coding scheme for multi-view video with single view depth map. The proposed scheme has a hierarchical coding structure in which the base view, depth maps, and non-base views are included one by one. As the result, bitstream generated by the proposed scheme is backward compatible with MPEG-C Part 3, which has already been released as an international standard format for FTV and 3D video. Since the proposed scheme employs the residual prediction framework, it can achieve more efficient compression even if low delay decoding is required. The proposed scheme doesn’t require any syntax changes and few modifications to the circuits of the conventional video codec are required; it might be possible to reuse almost all existing components. The backward compatibility and reusability offered by the proposed scheme are quite important in reducing manufacturing costs and time to market.

In this paper, we used 3D warping as the view synthesis algorithm. This is one of the simplest algorithms, so exploring new view synthesis algorithms suitable for the proposed coding scheme is one of our future works. We also note that the performance of encoding depth map is poor. Another future work is to develop a depth encoding algorithm. The camera setting on both Rena and Akko&Kayo are dense, so we plan to check our proposal’s performance on sparse and/or complicated MVV in order to fully determine its advantages and disadvantages.

6. REFERENCES

[1] M. Tanimoto, “Overview of Free Viewpoint Television,” *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, July 2006.

[2] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, “3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards,” in *Proc. ICME2006*, July 2006, pp. 2161–2164.

[3] A. Vetro, P. Pandit, H. Kimata, and A. Smolic, “Joint Draft 4.0 on Multiview Video Coding,” JVT Doc. JVT-X209, July 2007.

[4] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, “View Scalable Multiview Video Coding Using 3-D Warping With Depth Map,” *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17, no. 11, pp. 1485–1495, 2007.

[5] Y. Su, A. Vetro, and A. Smolic, “Common Test Conditions for Multiview Video Coding,” JVT Doc. JVT-T207, July 2006.

[6] S. Shimizu and H. Kimata, “CE3: MVC global depth map estimation,” JVT Doc. JVT-X076, July 2007.

[7] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” VCEG Doc. VCEG-M33, April 2001.