

A NEW SCALABLE MULTI-VIEW VIDEO CODING CONFIGURATION FOR MOBILE APPLICATIONS

Hoda Roodaki¹, Mahmoud Reza Hashemi¹, Shervin Shirmohammadi²

¹Multimedia Processing Laboratory, School of Electrical and Computer Engineering, College of Engineering, University of Tehran

²Distributed and Collaborative Virtual Environments Research Laboratory (DISCOVER Lab) School of Information Technology and Engineering (SITE), University of Ottawa
h.roodaki@ut.ac.ir, rhashemi@ut.ac.ir, shervin@site.uottawa.ca

ABSTRACT

Transmission of multi-view video content is not practical in most mobile environments due to the limited bandwidth and processing power of mobile devices. To support such environments, one can limit the number of views that are being transmitted, known as Scalable Multi-view Video Coding (SMVC). In this paper, we propose a new view selection method for view scalability in multi-view video coding in mobile environments, which uses inter and intra-view dissimilarities to determine the most suitable views for the base layer corresponding to the prediction structure and user selected limited number of views. By selecting more correlated views for the base layer, the proposed method provides an improved performance, as confirmed by simulation results, even when all the enhancement layers are dropped due to network limitations.

Index Terms — Scalable multi-view video coding, mobile video, dissimilarity estimation.

1. INTRODUCTION

In recent years, multi-view video coding has gained popularity in a variety of applications such as immersive teleconference, 3DTV, and free view point video. 3D Video is a key application for next-generation mobile devices [1]. Multi-view video coding (MVC) has been developed as an extension of the H.264 standard [2] to support 3D video and to improve compression and quality. However, even with this improvement, 3D video is currently beyond the capabilities of most mobile devices and environments.

View scalability, which reduces the number of views that are being transmitted, is a viable alternative to address this concern [3]. The views that are being selected for the base layer have a significant impact on the coding performance. Furthermore, in a mobile environment, it is likely that the enhancement layers are discarded altogether. In that case, because of the inherent inter-view dependencies of an MVC bitstream, nothing can be decoded at the receiver. In this paper, we propose a new base layer view

selection scheme for scalable MVC which addresses the above concerns. In our technique, we restrict the prediction structure of base layer views, such that they are encoded, and consequently decoded, independent of views in the enhancement layers. In order to minimize performance degradation, these views should be selected such that they are more correlated to each other and have less correlation to remaining views. Our method uses the correlation among different views, and also among frames within a view as a metric to choose best candidate views for the base layer. Experimental results indicate that adjacent views are not necessarily the most correlated ones, and that our proposed method indeed leads to a higher efficiency in terms of compression rate and overhead.

The rest of this paper is organized as follows. Related multi-view and scalable multi-view coding methods are reviewed in the next section. The proposed method is presented in section 3. Simulation results are provided in section 4, followed by concluding remarks in section 5.

2. RELATED WORK

While 3D video has become the subject of extensive research recently, mobile 3D video is still understudied. Nokia's Mobile MVC Prototype is a rare example of mobile 3D video, where MVC has been used to encode, transmit, and display 3D video on Nokia's N800 Internet Tablet [4], although it is not scalable in the sense that all views are sent to the receiver irrespective of whether or not the receiver can handle them. To be scalable and to allow access to selected views with minimum decoding effort, MVC should support a scalable bitstream structure, where view scalability is defined as a functionality that enables the decoder to decide the number of views to be decoded based on its processing overhead [3].

The current MVC coding prediction structure is shown in Figure 1. First V0 is encoded using single view video coding. Its prediction is limited to reference pictures in temporal domain. Next, V2 is encoded using the reconstructed frames of V1 as an additional reference. Similarly, V1 is encoded using reconstructed frames of V0

and V2 in addition to temporal references. Considering the inter-view dependencies in this scheme, it is clear that all views must be received at an end point in order to decode the video correctly. For view scalability; however, the goal is to reduce the number of views sent to some of the resource-constrained receivers. The general approach to achieve this goal is to select some views as base and the remaining views as enhancement layers. Depending on the capability of the mobile receiver, the base view (as a minimum) and one or more enhancement views might be transmitted to the receiver. Since the base layer might still have dependencies on the enhancement layers, as described above, several synthesis methods have been proposed in the literature to regenerate lost or untransmitted enhancement layers at the decoder side. These methods, some of which are described next, would then allow the base view to be decoded without the aid of the enhancement layers.

In [3] the view-dependent geometry is encoded and transmitted in order to synthesize the required view at the decoder. In [5] an auxiliary bitstream is sent, in addition to the main bitstream that provides the information necessary for implementing view scalability. [6] and [7] have tried to encode the extra information more efficiently. They have proposed to combine scalability in the view and depth domain. In this approach in addition to view scalability, another form of scalability such as spatial scalability is also applied on depth information.

To summarize, when the available bandwidth forces the encoder to drop enhancement layers or limit the number of views in the base layer, most existing methods synthesize the missing views at the decoder using side information. This approach requires extra bandwidth, and processing power, which is not available in most mobile applications. In this paper, this concern has been addressed by more efficiently selecting the base layer views, and eliminating the need for synthesis by de-correlating the base and enhancement layers.

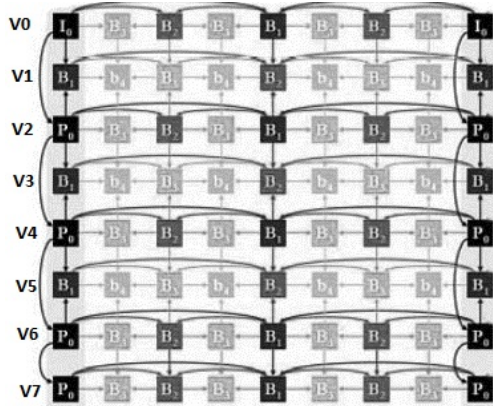


Fig 1. The current multi-view video coding prediction structure

As will be described in section 3, the contribution of our work is that unlike existing methods, in our method if

the enhancement layer is lost or dropped completely, no synthesis method is required at the receiver in order to regenerate enhancement layers views.

3. PROPOSED METHOD

In this paper, a new scalable multi-view video configuration has been proposed. In this new configuration, the base layer is encoded with a restricted prediction structure to de-correlate base and enhancement layer views as much as possible. In order to minimize the subsequent performance degradation, the most correlated views are being selected for the base layer.

In the first step, we extract a dissimilarity graph as depicted in Figure 2. Each node in the graph shows a frame and a weight is calculated for each edge of the graph illustrating the dissimilarity between two adjacent frames corresponding to the end nodes of this edge. In this graph, horizontal edges show the intra-view dissimilarity of frames within a specific view, and vertical edges show the inter-view dissimilarity of corresponding frames of two different views.

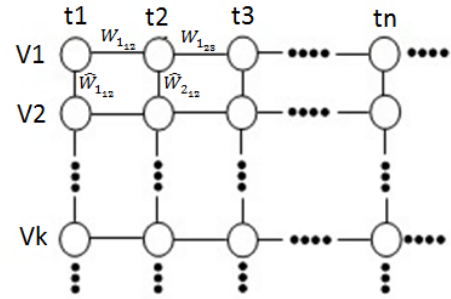


Fig 2. Dissimilarity Graph of Video Frames

For simplicity and without lack of generality, the similarity between frames in our method is measured by the correlation of their histograms. Let $H_{F_1,R}$, $H_{F_1,G}$, $H_{F_1,B}$ and $H_{F_2,R}$, $H_{F_2,G}$, $H_{F_2,B}$ be the single dimensional histograms of the RGB components of frames F_1 and F_2 respectively. The similarity of the two images can be defined as follows:

$$\begin{aligned} \text{similarity}(F_1, F_2) = & w_1 \cdot \rho(H_{F_1,R}, H_{F_2,R}) + \\ & w_2 \cdot \rho(H_{F_1,G}, H_{F_2,G}) + w_3 \cdot \rho(H_{F_1,B}, H_{F_2,B}) \end{aligned} \quad (1)$$

where w_1, w_2, w_3 are weights assigned to RGB components in the histogram, $w_1 + w_2 + w_3 = 1$ and ρ is the correlation of two single dimensional histograms. The dissimilarity of two frames (the weights of the graph in Figure 2) can be calculated as shown in the following equation:

$$\text{dissimilarity} = \frac{1 - \text{similarity}}{1 + \text{similarity}} \quad (2)$$

In this equation, the dissimilarity measure is inversely proportional to the similarity measure, and is equal to zero for two identical frames.

The extracted inter-view weights of this graph for the “Exit” test sequence are summarized in Table 1. As indicated by these results, the most correlated views (with minimum dissimilarity) are not necessarily the adjacent views.

Table 1. Average inter-view dissimilarity between different views of the “Exit” sequence

	V0	V1	V2	V3	V4	V5	V6
V0							
V1	0.90						
V2	0.88	0.90					
V3	0.85	0.85	0.87				
V4	0.81	0.83	0.86	0.90			
V5	0.82	0.85	0.86	0.86	0.89		
V6	0.77	0.79	0.81	0.80	0.81	0.85	
V7	0.79	0.81	0.82	0.81	0.85	0.86	0.85

In order to find the most correlated views, first the total number of views is selected according to our bandwidth limitation. Let us assume that we want to send four views in our base layer. Hence, our base layer prediction structure should be restricted as shown in Figure 3.

In MVC the first view should be predicted temporally, so the view with minimum intra-view dissimilarity should be selected as V_0 in our restricted prediction structure. According to Figure 3, for better prediction of V_2 , it should be highly correlated to V_0 . Hence among all the remaining views, the view with minimum dissimilarity to V_0 is selected as V_2 . As well, the view with highest correlation to V_0 and V_2 is selected as V_1 and the view with minimum inter-view dissimilarity to V_2 is selected as V_3 .

With this scheme, the base layer can be decoded independent of enhancement layers and with relatively high quality. On the other hand, for enhancement layers we use the conventional prediction structure of Figure 1 to encode all views.

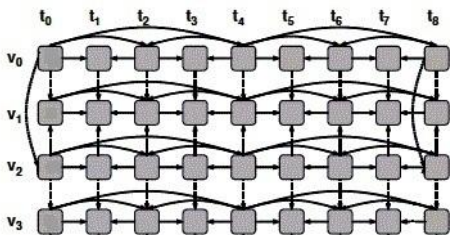


Fig 3. The restricted prediction structure for base layer

In our method, unlike the existing ones [3][5], in critical situations where the enhancement layer is dropped completely, no synthesis method is required at the receiver in order to regenerate enhancement layer views to solve the inter-view dependencies problem for decoding base layer. Hence, there is no need for side information, such as depth information or view geometry that is typically required by synthesis methods. Another advantage of the proposed

algorithm is that it can support free viewpoint as one of the main requirements of multi-view video coding. By knowing the user selected view direction of each terminal, the proposed method can determine the corresponding minimum number of proper views that meet the bandwidth constraints.

4. SIMULATION RESULT

The proposed method has been evaluated using three standard multi-view test sequences from [8] and [9]. Table 2 summarizes the properties of these sequences. Results have been obtained using the JMVC reference software version 8.2.

The implemented evaluation procedure is as follows. We have extracted the dissimilarity graph of eight views for each sequence. Table 1 and 3 show the average inter and intra-view dissimilarities for the “Exit” sequence. For this sequence, we chose view 0 as first view in our restricted prediction structure since it has the minimum mean intra-view dissimilarity among eight views according to Table 3. As we can see in Table 1, among the remaining views of this sequence, view 6 has the minimum dissimilarity to view 0. Hence, it should be selected as V_2 in our restricted prediction structure of Figure 3. View 1 has the highest correlation to 6, hence it is considered as V_1 , and view 3 with minimum dissimilarity to view 6 is chosen as V_3 .

To evaluate the performance of our view selection algorithm, we compare the overall quality of our selected views with the case where four adjacent views are selected as the base layer. We also tested our method for different number of total views and compared the results with the same case of adjacent views. The extracted PSNR and bitrate in our algorithm and for different number of selected views for “Exit” sequence are shown in Table 4. As we can see in most cases our selected views have better quality with lower bitrate. It should be noted that in these three standard sequences, the cameras are too close to each other and hence all views are highly correlated. Clearly, for sequences with larger camera spacing the proposed method will have a more significant improvement since in these cases the adjacent views are far less correlated.

All evaluations have been performed using four QP values (15, 20, 25 and 30). For each test the quality has been measured with Bjontegaard average BD-PSNRs and BD-Bitrates for each sequence [10]. The coding gain over adjacent views for each sequence is summarized in Table 5.

5. CONCLUSION

This paper proposes a new view selection method based on inter-view dependency for view scalability in multi-view video coding in mobile environments. The method alleviates restriction of mobile environments, such as low computing power and low bandwidth, by eliminating the need for synthesis at the receiver and by de-correlating base and

enhancement layers. It finds the solution by extracting inter and intra-view dissimilarity of frames; then, based on it and a prediction structure, the best views are allocated to base layer. Performance evaluations demonstrate that the proposed method achieves better compression rate compared to conventional methods with much less overhead.

6. ACKNOWLEDGEMENT

The authors acknowledge the research support and collaboration of Nokia Canada.

Table 2. Properties of test sequences

Sequence	Frame size	Frame rate (fps)	Camera Number	Camera Spacing	GOP length	Number of frames
Ballet	1024 x 768	15	8	20 cm	12	100
Break-dancer	1024 x 768	15	8	20 cm	12	100
Exit	640 x 480	25	8	20 cm	12	241

Table 3. Average Intra-view dissimilarity between different frames within each view of the “Exit” sequence

	V0	V1	V2	V3	V4	V5	V6	V7
Average Intra-view dissimilarity	0.97	0.9802	0.98	0.9799	0.9798	0.9802	0.9798	0.9785

Table 4. The comparison of quality and bitrate in kbit/s of our method and the adjacent view selection for the “Exit” Sequence

QP	Two Views				Three Views				Four Views			
	Our Selected Views		Adjacent Views		Our Selected Views		Adjacent Views		Our Selected Views		Adjacent Views	
	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate
30	36.96	674.5	36.68	719.3	37.13	865.86	36.78	981.88	37.26	1139.16	36.63	1365.38
25	38.58	1381.14	38.45	1481.4	38.74	1807.82	38.54	2084.74	38.89	2338.04	38.46	2928.8
20	40.29	3313.52	40.23	4397.9	40.36	4497.6	40.31	5844.18	40.47	5711.98	40.30	7701.14
15	43.05	9160.54	43.03	9455.2	43.04	13030	43.05	13681.52	43.08	16592.1	43.04	18406.8

Table 5. BD-PSNR and BD-Bitrate gain of our method with respect to adjacent view selection (positive numbers indicate improvement)

Sequences	Two View Selection		Three View Selection		Four View Selection	
	BD-PSNR	BD-Bitrate	BD-PSNR	BD-Bitrate	BD-PSNR	BD-Bitrate
Ballet	0.225	2.4035	0.1029	0.1878	0.0145	0.2070
Break-dancer	0.0261	-0.0326	0.263	0.9639	0.2391	0.8020
Exit	0.4506	0.9605	0.3801	1.0388	0.8216	1.5303

6. REFERENCES

- [1] M. Shafique, B. Zatt, S. Bampi, J. Henkel, “Power-Aware Complexity-Scalable Multiview Video Coding for Mobile Devices”, 28th Picture Coding Symposium, PCS 2010, Nagoya, Japan, Dec 2010.
- [2] ITU-T Rec. H.264 / ISO/IEC, Annex H, “Multiview Video Coding”, March 2010.
- [3] Sh. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, “View Scalable Multiview Video Coding Using 3-D Warping With Depth Map”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, Number 11, November 2007.
- [4] K. Willner, K. Ugur, M. Salmimaa, A. Hallapuro, J. Lainema, “Mobile 3D Video Using MVC and N800 Internet Tablet,” 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, Istanbul, Turkey, May 2008.
- [5] J. Lim, K. Ngan, W. Yang and K. Sohn, “A Multiview Sequence CODEC with View Scalability”, Signal Processing: Image Communication, Vol. 19, March 2004.
- [6] L.S. Karlsson, M. Sjostrom, “Multiview Plus Depth Scalable Coding in the Depth Domain”, 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, Potsdam, May 2009.
- [7] J. Cho, S. Cho, N. Hur, H. Lee, J. Jeong, “Effective Multiview Video Coding Using a Scalable Depth Map”, International Conference on Computational Intelligence for Modeling Control & Automation, Vienna, Austria, December 2008.
- [8] <http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload>, last access on March 10, 2011.
- [9] <ftp://ftp.merl.com/pub/avetro/mvc-testseq/orig-yuv/>, last access on March 10, 2011.
- [10] G. Bjontegaard, “Calculation of Average PSNR Differences between RD Curves,” ITU-T VCEG, VCEG-M33, Austin, USA, April 2001.