COMPRESSION IMPROVEMENT VIA REFERENCE ORGANIZATION FOR 2D-MULTIVIEW CONTENT

Pavel Nikitin[⋆†] *Marco Cagnazzo*[⋆] *Joel Jung*[†]

[⋆] Telecom ParisTech, LTCI, 46 rue Barrault, Paris
[†]Orange Labs, 4 avenue du 8 mai 1945, Guyancourt

ABSTRACT

One of the most challenging goals of future immersive services is to enable the observation of a scene from any viewpoint, thus making free-navigation possible under certain constraints. In order to provide such kind of services with smooth navigation, a huge amount of views should be available on the client's device. In particular, it is important for the case of 2D-multiview content, where cameras are positioned on a 2D grid in order to provide both horizontal and vertical parallax. This kind of content requires a large coding rate; therefore improving the compression performance of video encoders is especially relevant in this case. This paper studies how the encoder configuration affects the compression, by taking into account the spatial position of each camera. Four parameters are addressed in this work: coding order of the views, the number of reference lists, the number of reference pictures, and the ordering of pictures in the reference lists. An average of 12.0% bitrate saving is achieved for medium bitrate and 11.1% for low bitrate compared to the state of the art techniques.

Index Terms— Compression, free-navigation, 2D-multiview, reference pictures lists, coding order

1. INTRODUCTION

Immersive video formats for free navigation, such as 2D-multiview content (i.e. multiview with both horizontal and vertical parallax) are extremely demanding in terms of coding resource, therefore efficient compression is of crucial importance. In general, compression efficiency can be improved by modifying the tools (normative modifications) and by optimzing the encoder (non-normative modifications). In this study, we consider a non-normative modification of MV-HEVC [1] related to the coding order, to the reference list construction, and to the number of reference pictures in reference lists. The reference pictures are those pictures that have already been decoded and can thus be used to predict a block of pixels in the current picture. HEVC allows to use up to fifteen reference pictures, organized in one or two lists. Reference pictures are signaled by indexes in the reference lists. The inter-view prediction process is identical to the usual temporal prediction, the only difference is that the reference pictures come from different views and are taken at the same time instant; in opposition, when temporal prediction is performed, the reference pictures come from the same view, but at different time instants. The optimization of the reference list for temporal prediction has been widely studied for "classical" (i.e., single view) video content [2]; likewise, there have been some works on the selection of reference view for optimizing inter-view prediction in the case of horizontal-parallax multiview video: for example, Maugey et al. [3] have investigated in detail the case of 1D-multiview content and the free navigation scenario along the 1D path. However, in the case of 2D-multiview, the optimization of the reference list structure has been overlooked in the literature. In one of the few works dealing with this topic [4], it is claimed that serpentine coding order among views, depicted in Fig. 1, gives the best RD performance, but no proof is given. Moreover, it is not clear how to generalize the prediction structures that works well in the temporal case to the interview case: for example, hierarchical bi-directional prediction is very effective for temporal prediction while is expected to be less performing for inter-view prediction since there is less correlation between the views and the motion vectors are much larger. In order to find the best encoding configuration one can perform an exhaustive search, which requires an encoding of the content with tested configuration and a measurement of the encoder performance, but this approach would require such a huge amount of computational power to be practically unfeasible. For example if we encode sixteen views with all possible configurations and coding orders using the current implementation of multiview codec (two reference lists and up to fifteen reference pictures in each list), the total amount of encoder's configurations will be larger than 10^{77}. Our goal is to provide clues about how to choose an effective encoder configuration based on geometrical relationships between cameras without an explicit testing of all combinations.

2. ANCHOR AND CONTEXT

In this paper we consider the compression of one access unit (AU) of both texture and depth, i.e., the textures and the depth

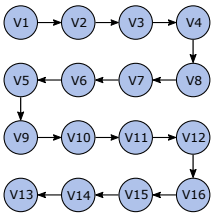


Fig. 1: Coding order of the anchor configuration.

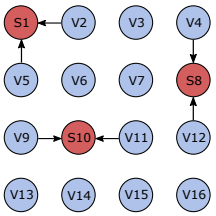


Fig. 2: Examples of synthesis configurations.

from all views at one given time instant. The reason is that we want to explore the best practices for inter-view prediction, without for the moment considering the influence of temporal prediction. This aspect will be studied in future works. We study the impact of the coding order and of the reference lists construction on the RD performance. We set as the anchor an MV-HEVC encoder with direct-P inter-view structure, one reference list, one reference image in the list and serpentine coding order, as shown in Fig. 1. Our anchor fits to the MPEG-I's common test conditions [1]. In particular, the texture and depth frames are compressed separately, and after decoding texture and depth information, virtual views on the positions of captured views are synthesized in order to test how the compression impacts the quality of the synthesis. The process of selection input views for synthesis is described in [1], some of the synthesis configurations are shown in Fig. 2. The RD performance of the proposed schemes are compared to the anchor using the Bjontegaard Delta Rate [5]. Following the CTCs [1], we compute the BD-rate on the compressed views (video BD-rate for short) and on the synthesized views (synth BD-rate). We consider five values of the quantization parameter (QP), that is 25, 30, 35, 40, 45, the first four accounting for a medium bitrate range, the last for four a low bitrate range. For this study we have used five 2D-multiview sequences in a 4-by-4 2D configuration (i.e., 16 views in 4 rows and 4 columns): Technicolor Painter [6], ULB Unicorn [7], Orange Shaman, Orange Dancing, and Orange Kitchen [8]. Technicolor Painter and ULB Unicorn are captured scenes and Orange Shaman, Orange Dancing, and Orange Kitchen are computer generated sequences. The depth maps for captured sequences are estimated as described in [9] for Technicolor Painter and in [10] for ULB Unicorn sequence.

3. REFERENCE LIST OPTIMIZATION

3.1. Number of reference lists

We start by assessing how much we can gain by using two reference lists and bi-prediction. The experiments have been conducted for five sequences for the first AU. On average we have gained 0.7% video BD-rate for medium bitrate and 0.3% for low bitrate. For synthesized view we have gained 2.8% and 2.6% for medium and low bitrate respectively.

3.2. Number of reference picture

In this experiment our goal is to check how much the RD performance depends on the number of reference views. We performed fifteen separate encoding of our content, increasing the number of reference pictures at each time using the serpentine scanning order and two reference lists. The results are shown in Fig. 3. The usage of seven references gives the best performance for compression while for synthesis BD-rate the best number of reference pictures is eight. For compression it is the best configuration for all the sequences both for medium and low bitrate. The maximum compression performance is achieved for Technicolor Painter sequence with -15.7% and -13.7% BD-rate gain for medium and low bitrate respectively. The lowest compression performance is achieved for Orange Shaman with -3.0% and -1.8% BD-rate gain. For synthesis the best configuration for medium bitrate is one with nine references with -10.4% BD-rate gain, and for low bitrate with eight references that provides -9.7% BD-rate gain. We deduce from these results that using lists with larger number of references increases the signaling cost without being compensated for by better prediction.

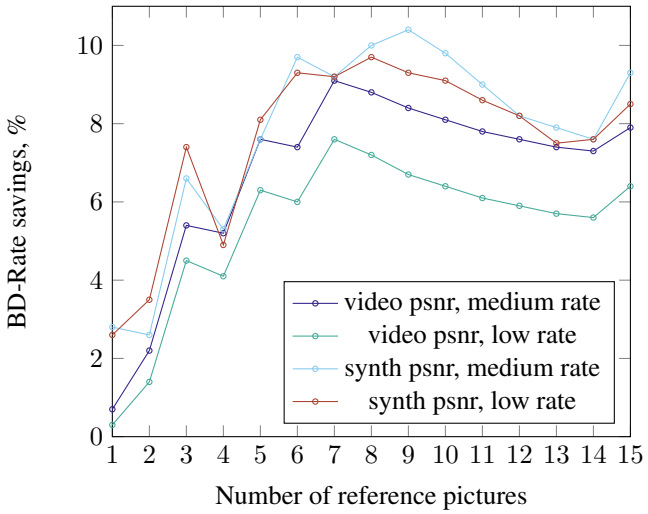


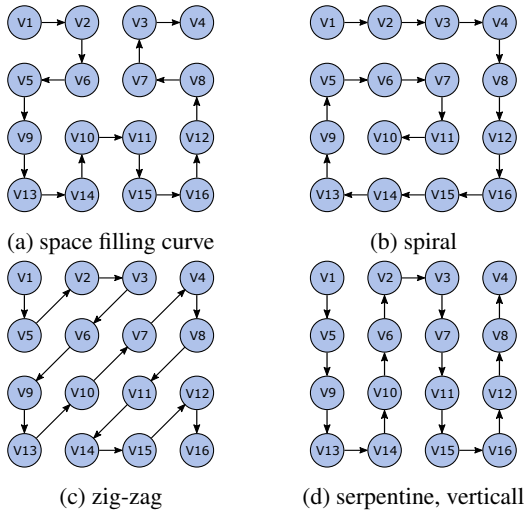
Fig. 3: Dependency of BD-Rate gain on the number of reference pictures.

3.3. Reference list ordering

In order to reduce the signaling of reference picture's index, we propose to modify the construction of reference lists. We calculate the distance between the camera that corresponds to the current view and cameras that correspond to all already decoded views, and the references views are put the lists in increasing distance order. The calculation of the distance between cameras is performed based on the camera parameters. The tie-break rules are the following: if the equidistant views are horizontal and vertical neighbors of the current view (e.g.

Configuration	video BD-rate		synth BD-rate	
	medium	low	medium	low
Desc Desc	-10.6%	-9.3%	-13.0%	-11.8%
Asc Asc	-9.1%	-7.6%	-9.2%	-9.2%
Asc Desc	-7.1%	-5.3%	-8.6%	-7.7%
Desc Asc	-10.3%	-9.1%	-11.7%	-10.1%
Proposed	**-11.0%**	**-9.7%**	**-12.6%**	**-12.6%**

Table 1: Dependency of BD-rate on the different ways of reference list ordering.



(a) space filling curve (b) spiral

(c) zig-zag (d) serpentine, verticall

Fig. 4: Coding orders

V6 and V9 references for V10, see Fig. 1), the order is H-V in list 0 and V-H in list 1. Otherwise they are ordered by putting the last coded view first (e.g., V5 and V7 for V10, see Fig. 1). The proposed configuration is compared with some basic reference list organizations, where we have put the pictures in reference list either in ascending (Asc) or descending (Desc) coding order. The results of the experiments are represented in Table 1. By applying the proposed method of reference lists construction we have obtained 11.0% BD-rate gain for medium and for low bitrates, and 12.6% for synthesized views. Thus, putting reference pictures into the list starting from the closest to the current view provides the best BD-rate gains.

4. IMPACT OF CODING ORDER ON PERFORMANCE OF COMPRESSION

4.1. Experiments for different coding orders

In this study the goal is to find out how the compression efficiency depends on the coding order of the views. As is shown in the literature [11][12][13] different coding orders are beneficial for the different type of content, but often one particular coding order is selected without giving an explanation of the

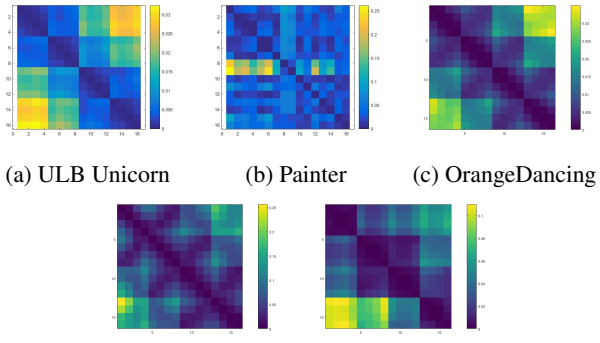
Coding order	video BD-rate		synth BD-rate	
	medium	low	medium	low
SH	-11.0%	-9.7%	-12.6%	-12.6%
SV	-10.8%	-9.4%	-4.3%	-3.7%
SFC	-10.8%	-9.5%	-10.4%	-10.4%
ZZ	-10.2%	-8.5%	-13.2%	-12.8%
SpE	-8.7%	-7.4%	-6.9%	-8.4%
SpC	**-12.0%**	**-11.1%**	**-8.7%**	**-11.0%**

Table 2: Impact on performance of different coding orders: serpentine horizontal(SH), serpentine vertical (SV), space filling curve (SFC), zig-zag (ZZ), spiral from the edge (SpE), and spiral from the center (SpC).

choice. In Fig. 4 four different coding orders are depicted: space filling curve (4a), spiral from the edge (4b), zig-zag (4c), and serpentine vertical (4d). For spiral we test two variants: starting from the edge V1 to the center V10, and in other direction from V10 to V1.

Table 2 shows results of performance for different coding orders. On average the coding order that provides the best results for compression is spiral from the center, since the reference picture are as close as possible to the current one, and they are signaled using smaller index (i.e., less bits) in the reference list. Moreover in this configuration the intra picture which is encoded at higher quality, is in the center, so its average distance from the current encoded view is minimized. If we look at the results sequence-wise, we can observe that spiral from the center configuration is the best in terms of compression performance for all of the sequences except Technicolor Painter. For captured sequences the worst configuration is serpentine vertical, and for computer generated is spiral from the edge. The results for compression and synthesis are consistent for captured sequences.

These results clearly show that the content has a huge impact on the most efficient coding order. In particular we argue that the pictures that are more often used as reference (which necessarily are encoded early) must be very good predictors for the other views. This means not only that starting from the center is in general a good choice, but also that, if some view is different from the other (e.g. because of different color calibration), this can disrupt the prediction structure. Since the SpC coding order does not work well on the Technicolor Painter content, we wondered if some of the views are not well calibrated and affect the prediction process. Therefore we computed some metric to assess the similarity among views, in particular we choose the Kullback-Leibler divergence (KLD), or relative entropy. As we can observe from these results the KLD between views 8 and 9 for Technicolor Painter sequence, shown in Fig. 5b, is larger compared to the other views. It can be an indicator for a misalignment in capturing process of these views. It results in inconsistent predictors being used by the encoder.



(a) ULB Unicorn (b) Painter (c) OrangeDancing

(d) OrangeKitchen (e) OrangeShaman

Fig. 5: Graphical representation of the KLD between views i and j for different sequences: $A(i,j) = D(p_i||p_j)$, where p_k is the relative frequency of luminance values of view k, used as an estimator of the corresponding PDF, and $D(\cdot||\cdot)$ is the KLD.

4.2. Refinement based on different reference list allocations.

We have observed from previous tests that the intra-picture is often used as the reference even if its reference index is higher than of the others. One of the reason that this picture has higher quality is the lower QP, as it is described in CTC [1]. Taking into account this fact we We modify the reference list by anticipating the position of the Intra picture. However, we did not observe gains in compression efficiency. One of the main reasons for this behavior is that the intra picture is already located very close to the other views in this coding order.

4.3. Dependency on the number of reference picture.

In Section 3.2 we show that, using the anchor configuration, the best number of reference pictures is seven. However, by the time being, we know that the anchor configuration can be improved by using the SpC coding order and the distance-based list construction. Therefore we want to know what the best reference picture number is in this configuration. The results of the experiments are depicted in Fig. 6. As we can see from the figure using more than five references does not improve the compression performance. Most of the gains come only from the fact that we are using one horizontal and one vertical neighbor, which is why the configuration with two references is just slightly worse than with five. If we compare results obtained in Section 3.2 shown in Fig. 3 with the ones in Fig. 6, the curve for compression is smoother in Fig. 6 and saturates after configuration with the usage of five references. However, when 5 references are used, the encoding time increases in 3.2 times compared to the case where only one reference picture is used. Since the complexity of decoder does not depend on the number of used references, the providers of free-navigation services can make a decision regarding bitrate savings for the streaming of such content based on their available resources.

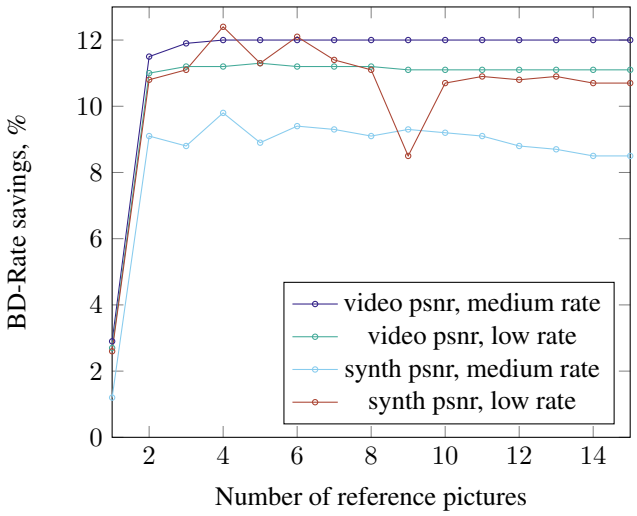


Fig. 6: Dependency of BD-Rate gain on the number of reference pictures for spiral coding order.

5. CONCLUSIONS

In this paper we have extensively studied the impact on the compression performance based on different parameters of encoder configuration. By using two reference lists we benefit from bi-prediction and can achieve **0.5%** BD-rate gain. By applying the proposed method of reference lists construction we have obtained **11.0%** BD-rate gain for medium bitrate and **9.7%** for low bitrate and **12.6%** for synthesized views. Thus, putting reference pictures in the list starting from the closest to the current view provides the best BD-rate gains. Moreover for most of the sequences spiral coding order provides the best results for video BD-rate, since the picture coded in Intra mode is located close to the most of coded views and is often used as a reference. However misalignment or inaccurate calibration of the cameras may disrupt the prediction efficiency, therefore coding structure that are effective in the general case can be less effective for content having this kind of problems, such as Technicolor Painter. The first five reference pictures in reference list contribute the most to the overall performance. We have gained **12.0%** in terms of BD-rate. Using lists with larger number of references increases the signaling cost without being compensated for by better prediction. Future work will be devoted to study the impact of reference list characteristics in the case when both temporal and interview predictions are used.

6. REFERENCES

[1] J. Jung, B. Kroon, R. Dore, G. Lafruit, and J. Boyce, "CTC on 3DoF+ and Windowed 6DoF," *ISO/IEC JTC1/SC29/WG11 MPEG2018, Doc. N17726*, July 2018.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Hierarchical b pictures," *Joint Video Team (JVT), Doc. JVT-P014*, July 2005.

[3] T. Maugey, G. Petrazzuoli, P. Frossard, M. Cagnazzo, and B. Pesquet-Popescu, "Reference view selection in dibr-based multiview coding," *IEEE Transactions on Image Processing*, vol. 21, pp. 1808–1819, 2016.

[4] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. Kovacs, and V. Kiran Adhikarla, "Subjective evaluation of super multi-view compressed content on high end light field 3D display," *Elsevier Signal Processing: Image Communication*, vol. 39, pp. 369–385, November 2015.

[5] G. Bjontegaard, "Calculation of average psnr differences between rd curves," *ITU-T SG16/Q6 VCEG 13th meeting, Doc. M33*, April 2001.

[6] D. Doyen, T. Langlois, B. Vandame, F. Babon, G. Boisson, N. Sabater, R. Gendrot, and A. Schubert, "Light field content from 16-camera rig," *ISO/IEC JTC1/SC29/WG11 MPEG2017, Doc. M40010*, January 2017.

[7] D. Bonatto, A. Schenkel, T. Lenertz, Y. Li, and G. Lafruit, "ULB high density 2D/3D camera array data set, version 2," *ISO/IEC JTC1/SC29/WG11 MPEG2018, Doc. M41083*, July 2017.

[8] P. Boissonade and J. Jung, "Proposition of new sequences for windowed-6DoF experiments on compression, synthesis, and depth estimation," *ISO/IEC JTC1/SC29/WG11 MPEG2018, Doc. M43318*, July 2018.

[9] D. Doyen and G. Boisson, "EE-Depth: Comparison of different DERS versions on TechnicolorPainter sequence," *ISO/IEC JTC1/SC29/WG11 MPEG2018, Doc. M42558*, April 2018.

[10] A. Schenkel, S. Fachada, and G. Lafruit, "ULB unicorn v2 EE-Depth results," *ISO/IEC JTC1/SC29/WG11 MPEG2018, Doc. M42336*, April 2018.

[11] S. Shi, P. Gioia, and G. Madec, "Efficient compression method for integral images using multi-view video coding," *IEEE 18th International Conference on Image Processing*, September 2011.

[12] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1461–1473, 2007.

[13] T. Chung, K. Song, and C.-S. Kim, "Compression of 2-d wide multi-view video sequences using view interpolationg," *IEEE 15th International Conference on Image Processing*, October 2008.