VERSATILE LAYERED DEPTH VIDEO CODING BASED ON DISTRIBUTED VIDEO CODING

G. Petrazzuoli, C. Macovei, I.-E. Nicolae, M. Cagnazzo, F. Dufaux, B. Pesquet-Popescu

Institut Mines-Télécom, Télécom-ParisTech / CNRS LTCI / UMR 5141

ABSTRACT

Video content constitutes today a large part of the data traffic on the Internet. This is allowed by the capillary spreading of video codec technologies: nowadays, every computer, tablet and smart phone is equipped with video encoding and decoding technologies. As a matter of facts, the video content often exists in different formats, that, even though be incompatible among each other, still have a significant mutual redundancy. The incompatibility prevents an efficient exploitation of the scalability, that on the other hand is a very important characteristic when it comes to efficient network use.

An interesting alternative to classical scalable video, is using distributed video coding (DVC) for the enhancement layers. In the envisaged scenario, clients have different decoders for the base layer, adapted to the characteristic of their device. However they can share the same enhancement layer, since DVC allows encoding frames independently from the reference that will be employed at the decoder.

This approach has been considered in the past in order to improve temporal and spatial scalability. In this work we review the existing approach, we improve them using more recent DVC techniques and we perform a new analysis for the emerging multi-view applications.

1. INTRODUCTION

The Internet is an heterogeneous collection of network, where users can have different resources in terms of memory and computational complexity. Today the largest part of the Internet traffic is related to video applications such as video conference, video streaming, downloading and sharing. A trivial way to take into account the different requests of the users is to encode the different versions of a video at different quality and store all the versions on a video server. Then, only one of these versions is sent to each user. Obviously, among the different versions of the same video there will be a huge redundancy. Scalable video coding (SVC) [1] has been developed as an extension of H.264/AVC for encoding the different versions of the video by eliminating redundancies as much as possible. SVC enables to encode the video once, but the users can chose the parameters of the video by selecting only a subset of the bit stream used for encoding the video. Then, the bit stream is divided in base layer (that consists in the layer at lowest quality) and other enhanced layers, that are sent to the user only if requested. There are three main types of scalability: temporal, spatial and quality. The temporal scalability enables the user to decode the video at lowest frame rate. This is possible using hierarchical B-frames such as in H.264/AVC. Spatial scalability enables the user to decode the video at different spatial resolution. Quality scalability means that for each enhanced layer that is sent, the PSNR of the decoded image w.r.t. the real one increases. However, besides these "classical" forms of scalabilty, today new ones appear, associated to the emerging formats such as multi-view video (MVV) and multi-view video-plus-depth (MVD): we have view scalability when a subset of the total views is decodable without having to decode all the views, and component scalability when the the access to one component (texure or depth) does not rely on the decoding of the other.

One of disadvantages of classical scalable approaches is that each enhanced layer is strictly dependent from the previous ones. Moreover, an enhanced layer cannot be decoded, if the previous one should be not correctly received and decoded. In order to make each layer independent of the other ones, [2], [3], [4] and [5] propose to apply Distributed Video Coding (DVC) for encoding the video. DVC is based on distributed source coding [6, 7]. In this paradigm, dependent sources are independently encoded but jointly decoded. Under some constraints on the statistical characteristics of the sources, the loss in terms of rate-distortion performance are negligible w.r.t. classical joint source coding. Concerning scalability, this means that with DVC we can encode the different layers independently. Then, the decoding is independent from which information is available at the decoder side. In this way, we can have different base layers sharing the same enhancement layer encoded in DVC. This can allow a remarkable bandwidth saving, above all when many different codecs are considered. Due to the different video coding techniques present nowadays on a network, (for example H.264/AVC with its different profiles, HEVC, MPEG-2, MPEG-4), it would be necessary to encode the video in all these formats. On the contrary, if DVC is used, one only version of the enhanced layers is sufficient for all the users independently of the technique used for the base layer. We have analysed also the RD performance when scalable DVC is ap-

plied on view domain in the context of multiview distributed video coding. Moreover, several solutions are possible that allow view scalability: of course, a trivial solution is using the same single view encoder on each view (Simulcast); a more effective approach is based on the use of the multiview extension of H.264/AVC, called H.264/MVC. In this work, we compare the performance of multiview scalable DVC w.r.t. these classical approaches for view scalability.

The paper is structured as follows. In Section 2, we provide a state-of-the-art about scalable video coding, distributed video coding and scalable DVC. In Section 3, we describe in details our analysis and comparison, and we conclude the paper in Section 4.

2. BACKGROUND IN SCALABILITY AND DVC

2.1. Temporal Scalability in H.264/AVC (SVC)

The scalable extension of H.264/AVC [1] has been proposed in order to take into account the different resources in terms of memory and complexity of the user, for temporal, spatial and quality scalability. Let us consider a video stream divided into a base layer (BL) and in n enhanced layers. The base layer consists of only I frames or P-frames, whose reference frame is in the BL. The n enhanced layers can be obtained by introducing hierarchical B-frames. The B-frames of the k-th enhanced layer can be obtained by using as reference the frames of the previous enhanced layers (from 1 to $k-1$). With a simple dyadical structure, if the original video is at f frames per second (fps), the BL layer is at f/N fps and the k-th enhanced layer will be at $2^k \frac{f}{N}$ fps. The H.264/SVC standard also allows a flexible (i.e. non-dyadic) definition of temporal dependencies between frames.

2.2. DVC and DISCOVER interpolation algorithm

In this section we describe one of the most popular framework for DVC, the Stanford codec [8]. In this codec, the video stream is split into Key Frames (KFs) and Wyner-Ziv Frames (WZFs). Borrowing the terminology from the predictive video coding context, a KF and all the following WZFs before the next KF are said to form a group of pictures (GOP). Hence the distance between two successive KFs is called GOP size. The KFs are INTRA coded (i.e. without motion estimation and compensation). The Wyner-Ziv Frames are fed into a systematic channel coder. The systematic part is discarded and the parity bits are sent to the decoder. At the decoder side, an estimation of the Wyner-Ziv Frame is needed. It can be obtained by interpolation of the already decoded frames. This estimation is called Side Information (SI) and it can be considered as a noisy version of the true WZF. The channel decoder must correct these estimation errors by using the parity bits. Then, the encoding of the WZFs is completely independent from how the KFs have been encoded and decoded. The European project DISCOVER [9] implemented the Stanford architecture and defined effective tools for coding the KFs and the WZFs. It has become the reference technique for distributed monoview and multiview video coding. In DISCOVER the SI is generated by a linear motion interpolation algorithm of the closest frames available at the decoder side. In a previous work [10], we have proposed an high order motion interpolation (HOMI) based on 4 images. This algorithm improves the RD performance of classical interpolation techniques.

2.3. Scalable DVC

One of disadvantages of SVC is that each layer depends strictly from the previous one. With DVC, the different layers can be encoded and decoded independently. This means that the base layer can be encoded with any technique without affecting the decoding of the WZFs. In particular, the temporal scalability is intrinsic in DVC. Indeed, the procedure of encoding and decoding for GOP size larger than two is very similar to the structure of hierarchical B-frames of H.264. Let us consider a GOP size equal to 4. Then, let I_{k-2} and I_{k+2} be two consecutive KFs. These frames are used for the estimation of the WZF at instant k. Once that this frame has been decoded, the frame I_k is available at the decoder side. It can be used along with the KFs for obtaining the estimation of the WZFs at instant $k-1$ and $k+1$. Tagliasacchi et al. [5] proposed a temporal scalable DVC for the PRISM codec. In this scheme the base layer has been obtained by H.263+/INTRA. The enhanced layer had been obtained by using algorithms for linear motion interpolation. In [2] and [3] a comparison of temporal scalable DVC w.r.t H.264/AVC has been performed. Moreover, for DVC coding they used an overlapped block motion compensation based side information generation module and an adaptive virtual channel noise model module. They obtained that the RD performance of scalable DVC improves the performance of H.264/INTRA but does not surpass the RD of SVC. Then, they suggest to use DVC only if there are some constraints in terms of complexity and memory at the encoder side. The independence of the enhanced layer w.r.t. the base layer for DVC has been emphasized by [2] and [4]. Indeed, also if we change the anchor frames, the enhanced layers does not change for DVC, On the contrary, another enhanced layer is needed each times that the INTRA Frames of H.264/AVC are coded in a different manner. The quality scalability is also automatically obtained with DVC: the parity bits generated by the encoder are used for improving the quality of the side information. Then, the more parity bits are sent by the decoder, the better the quality of the decoded frames is. Each set of parity bits progressively improves the PSNR of the decoded WZFs. Solutions for spatial scalability have been proposed by [2] and [11].

In [12] and [13] the temporal scalability is extended for multiview video coding. Ozbek et al. [12] suppose to have

two cameras : the right view is temporally predicted and the left view is predicted from the right one. They extend this structure for multiview by supposing that only one view camera depends from itself and the other ones are predicted by this reference view. Drose et al. [13] suppose that only a central camera is coded independently of the other one. The temporal stream is coded with a certain GOP. In the position of the I frames, the frames of the other cameras are P-frames depending on the frames of the central camera, as the view progressive architecture of H.264/MVC [14]. The other frames are coded only by exploiting temporal correlation.

3. SYSTEM PERFORMANCE ANALYSIS

Let us consider now the works of [2] and [3]: we perform a performance analysis of DISCOVER w.r.t. some relevant video coding standards: H.264/AVC, H.264/AVC with a low complexity profile, the emerging HEVC. The low-complexity profie of H.264/AVC is obtained by switching off rate distortion optimization.

In our use-case, we have to send the different bit streams of the different standard. If a user has the BL of H.264/AVC cannot decode the B-frames coded by HEVC and viceversa. . For these reasons, it is necessary to send the bitstreams of H.264/AVC and HEVC. But if we suppose that all the users have a DVC decoder, the enhanced layers can be coded with a Wyner-Ziv codec, one bitstream is sufficient for all the users. We have extended the temporal scalable video coding along the view axis in multiview videos. We suppose that we have K cameras. One camera out of V is a Key camera. The other ones are Wyner-Ziv camera. The base layer consists of sending only the Key views. The other views are hierarchical coded, as in the temporal domain as depicted in Fig. 1. Let us suppose that one of four cameras is a Key cameras and let 0 and 4 be two of these cameras. Then, in the first enhanced layer, the view number 2 is send and for the second layer the cameras 1 and 3 are sent. This structure is used both for the DVC codec and for H.264/MVC. If the DVC DISCOVER is not used, in order to take into account that some users can not have the H.264/MVC decoder, we are forced to encode and store also a simulcast version of this video where each camera is independently coded. For this reasons, the performance of scalable multiview distributed video coding are compared w.r.t. H.264/MVC and H.264/Simulcast.

In order to perform a complete analysis of the different methods for scalable video coding, we suppose that we have two scenarios. One scenario where users have different decoders: we suppose that each video stored on the video server is coded in H.264/AVC, H.264/AVC low-complexity (with no RD optimization) and HEVC. Even if the base layers 1 and 2 are compatible, the corresponding enhancement layers will be not, since they are predicted against possibly different images. In this context we are then obliged to send all the base layers and all the enhanced layers. Another scenario is that only

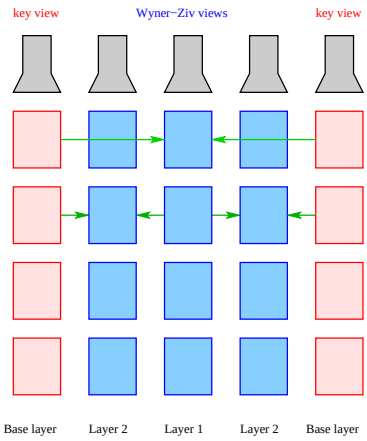

Fig. 1. Example of multiview scalable video coding with $V = 4$

the base layer is INTRA coded with H.264/AVC or HEVC, and the enhanced layers are encoded with the DISCOVER DVC scheme. These means that the enhances layers are independent from which is the base layer available for each user. For the scalable monoview we have considered the MPEG sequences *party scene* and *BQSquare*, respectively at spatial resolution of 832×480 and 416×240. Their frame rates are of 60 and 50 respectively. We have considered a GOP size of 4, and then we can suppose that we have a base layer and two enhanced layers. The frame rate for the base layer is respectively 12.5 and 15. We have then consider DISCOVER with the base layer (that means the KFs) encoded with H.264/AVC, with HEVC and H.264/(low complexity). We have performed a rate-distortion analysis of DVC w.r.t. the scenario where we are obliged to send H.264/AVC, HEVC and H.264/low complexity (see Tab. 1) and we have obtained up to 23.58% of bit reduction and up to 3.54 dB of PSNR improvement. Indeed, when standard video techniques are used for the enhanced layer, we are obliged to encoded these layers with all the considered standard. With DVC, since the enhanced layers are independent of the BL, we can use the same set of parity bits independently of which BL is available to the user. In the context of multiview video coding, we have considered the Xmas sequence at 480×640 resolution and we have compared the RD performance of DVC w.r.t. Simulcast+H.264/MVC (see Fig. 2 and 3). Indeed, if some user have not the codec of H.264/MVC, we are forced to send on the net also the Simulcast version, where all the view are independently encoded.

4. CONCLUSIONS AND FUTURE WORKS

In this paper we have performed an analysis in terms of RD performance for temporal scalable DVC w.r.t. classical scalable techniques. In contrast with the classical case, enhanced layers in DVC are independent from the BL. Then, if differ-

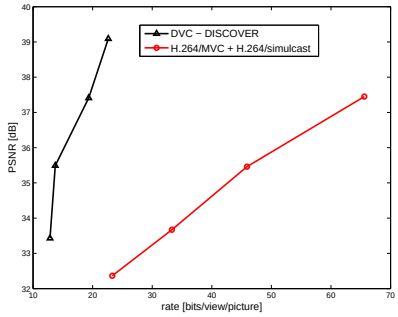

Fig. 2. RD curves for sequence Xmas - Layer 1

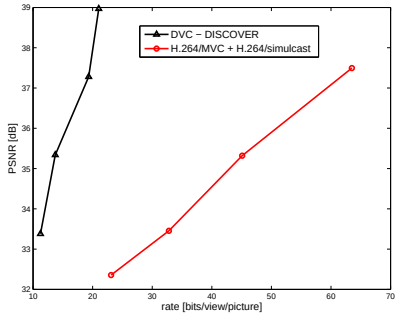

Fig. 3. RD curves for sequence Xmas - Layer 2

ent users have different decoders, using DVC, the same parity bitswill be sufficient to decode the enhanced layers independently of the codec used for the base layer, achieving a noticeable bandwidth saving. We have extended our analysis also to multiview video coding, ion order to take into account that some users can have the H.264/MVC codec and other ones cannot have it. Then, we should send also the Simulcast version of this video. If DVC is used, we can avoid to send these two versions.

method	Δ_R [%]	Δ_{PSNR} [dB]
BQSquare - layer 1		
DVC (KF coded with H.264/AVC)	-4.70	0.86
DVC (KF coded with HEVC)	-23.58	0.40
DVC (KF coded with H.264/AVC l.c.)	-4.73	0.20
BQSquare - layer 2		
DVC (KF coded with H.274/AVC)	19.17	3.54
DVC (KF coded with HEVC)	4.24	0.83
DVC (KF coded with H.264/AVC l.c.)	20.20	0.84
Party Scene - layer 1		
DVC (KF coded with H.264/AVC)	-12.79	0.80
DVC (KF coded with HEVC)	-16.36	1.07
DVC (KF coded with H.264/AVC l.c.)	-11.56	0.78
Party Scene - layer 2		
DVC (KF coded with H.264/AVC)	-13.71	1.06
DVC (KF coded with HEVC)	-18.22	1.08
DVC (KF coded with H.264/AVC l.c.)	-12.89	1.02

Table 1. RD performance by Bjontegaard metric w.r.t. H.264+HEVC+H.264(low-complexity)

5. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[2] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Codec-independent scalable distributed video coding," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 16 2007-oct. 19 2007, vol. 3, pp. III –9 –III –12.

[3] X. Huang, A. Ukhanova, E. Belyaev, and S. Forchhammer, "Temporal scalability comparison of the h. 264/svc and distributed video codec," in *Ultra Modern Telecommunications & Workshops, 2009. ICUMT'09. International Conference on*. IEEE, 2009, pp. 1–6.

[4] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Error-resilient scalable compression based on distributed video coding," *Signal Processing: Image Communication*, vol. 24, no. 6, pp. 437–451, 2009.

[5] M. Tagliasacchi, A. Majumdar, and K. Ramchandran, "A distributed-source-coding based robust spatio-temporal scalable video codec," in *Proc. Picture Coding Symposium*, 2004.

[6] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, July 1973.

[7] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–11, Jan. 1976.

[8] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Asilomar Conference on Signals and Systems*, Pacific Grove, California, Nov. 2002.

[9] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The Discover codec: Architecture, techniques and evaluation," in *Proceedings of Picture Coding Symposium*, Lisbon, Portugal, Nov. 2007.

[10] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "High order motion interpolation for side information improvement in DVC," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010.

[11] B. Macchiavello, F. Brandi, R.L. Queiroz, and D. Mukherjee, "Super-resolution applied to distributed video coding with spatial scalability," *Anais do Simpsio Brasileiro de Telecomunicaes*, 2008.

[12] N. Ozbek and A. Tekalp, "Scalable multi-view video coding for interactive 3dtv," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 213–216.

[13] M. Drose, C. Clemens, and T. Sikora, "Extending single-view scalable video coding to multi-view based on h. 264/avc," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2977–2980.

[14] A. Vetro, T. Wiegand, and G.J. Sullivan, "Overview of the stereo and multiview video coding extensions of the h.264/mpeg-4 avc standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626 –642, april 2011.