ADAPTIVE M-BAND HIERARCHICAL FILTERBANK FOR COMPLIANT TEMPORAL SCALABILITY IN H.264 STANDARD

C. Bergeron, C. Lamy-Bergot[*]

THALES Land & Joint Systems
EDS/SPM Department
F-92704 Colombes

B. Pesquet-Popescu[†]

ENST
TSI Department
F-75013 Paris

ABSTRACT

This paper presents a solution of temporal scalability for video encoded H.264/MPEG-4 AVC bitstreams. Achieved through the concept of adaptive M-band hierarchical filterbanks, the temporal scalability is performed thanks to the application of a frame shuffling operation which allows to keep backward compatibility with the standard. Simulation results show that this scalability is obtained with no degradation in terms of subjective and objective quality.

1. INTRODUCTION

Following the ever increasing demand for efficient, simple and easily applicable video coding standard that could be applied to settings as different as wired and wireless communications, ITU-T and ISO have established a common specification, denoted H.264 or MPEG-4 AVC [1], which provides a significant compression gain when compared to former standards and is easily adaptable to networked applications. Targeting applications as diverse as visiophony over wired or wireless links, high quality video services for streaming over satellite or lower quality streaming for video services over the Internet, H.264 presents one major drawback when channel varying applications are concerned : it does not include scalability. Solutions are currently being proposed in the literature or within SVC standardisation group to remedy to this problem, which generally plan to modify the H.264 syntax to integrate PFGS (Progressive fine granular scalability) coding or subband decompositions [2, 3].
In the meantime, motion-compensated (MC) spatio-temporal subband decompositions have gained a lot of interest due to their fine granular spatial/temporal/SNR scalability features combined with state-of-the-art compression performance[4]. In particular, the temporal scalability in these codecs is achieved through multi-resolution dyadic (and even triadic [5]) filterbanks. However, these structures are open-loop and some of the powerful tools in H.264/AVC like the in-loop deblocking filter are not easy to apply.

In this paper, we present some temporal scalable solutions fully compliant with H.264/AVC and show that they can be easily interpreted and generalized in the framework of adaptive M-band hierarchical filterbanks. They combine a hierarchical representation with a closed-loop structure and preserve (or even improve) the coding performance of the original non scalable scheme.

The paper is organised as follows. Section 2 introduces proposed hierarchical filterbank structures and discusses their interest for video coding and scalability. An application of such filterbanks is proposed and discussed in Section 3. Section 4 describes a practical setup for easily applying filtering in a compliant way to an H.264 codec, through the application of an interleaver. Finally, experimental results are presented in Section 5 and conclusions are drawn in Section 6.

2. M-BAND HIERARCHICAL FILTERBANKS

We propose a generic filterbank structure that provides a hierarchy of output subbands, containing one "intra" (equivalent to a low-pass) subband and several levels of detail subbands, ordered according to their importance. Each detail subband is obtained through a closed prediction loop, which is different from existing temporal wavelet schemes. Fig. 1 illustrates the proposed concept for a scalability of factor 2 and two detail levels, but it is easy to generalize this construction to other sub-sampling factors. The Group of Pictures (GOP) size is $M = 2^L - 1$, where L is the number of temporal resolution levels (note again the difference with a wavelet filterbank GOP size and structure).

In Fig. 1, D_1, resp. D_2 denote delays that can be chosen such as to design different GOP encoding orders. For example, for $D_1 = Z^{-2^{L-1}}$, $D_2 = Z^{2^{L-1}}$, we get a symmetrical encoding structure, with an Intra frame in the middle of the GOP (see also Fig. 2, while for $D_1 = Z^{2^{L-1}}$, $D_2 = Z^{2(2^{L-1})}$ the Intra frame is encoded at the beginning of

[*]This work was partially supported by the European Community through project IST-FP6-001812 PHOENIX.

[†]This work was partially supported by the European Community through the project IST-FP6-1-507113 DANAE

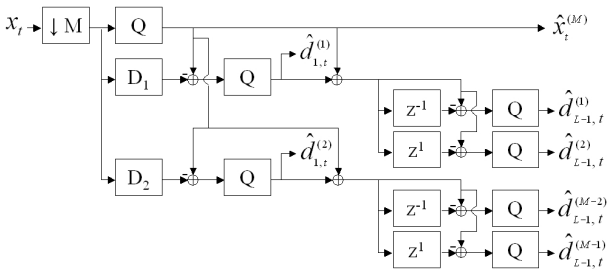


Fig. 1. Example of hierarchical filterbanks decomposition.

the GOP (see also Fig. 3). By inactivating one of the two detail branches starting with a certain level, one can obtain a different hierarchy of frames (*e.g.*, as illustrated in Fig. 6). Note that alternating at different stages of the scalable structure filterbanks with different sub-sampling factors (see Fig. 4) provides a flexible manner to achieve variable frame rates.

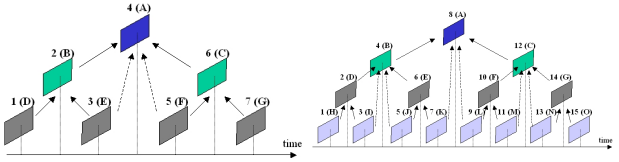
A common characteristics of these schemes is that the frames situated at the top of the hierarchy are very distant temporally and this distance decreases as one goes to further refinement details, making easier the prediction at these stages. This is somehow different from the temporal wavelet framework, where the "temporal distance" between frames is smaller for first prediction levels than for the last levels. Another interesting feature of these structures is the self decodability of each GOP, which is very useful in error-prone environments to avoid error propagation.

3. APPLICATION TO THE CASE OF H.264 VIDEO STANDARD

It is proposed here to rely on subband like decomposition presented in Section 2 with taking advantage of existing H.264 properties to avoid requiring any modification of or addition to the standardised codec, and remain consequently compliant with each of its profiles, and do not require the possibility to use bi-directionnal frames.

3.1. Filtering scheme for GOPs of size $2^L - 1$

Considering again GOPs of $2^L - 1$ frames, denoted by their original time reference $\{1, 2, \ldots, 2^L - 1\}$, the aim of the operation is to intelligently distribute the frames so that the encoding process that will follow is perform efficiently. A random GOP decomposition pattern will not be the most efficient one, first because one wants to obtain a regular frame rate when using the temporal scaling, and second because a better compression and coding efficiency can be obtained when placing the reference frames close to the predicted ones. The regular repartition pattern corresponding to the M-band decomposition, which ensures the temporal scalability feature is obtained as follows. The first reference image is coded in Intra, and placed for instance at the middle of the GOP. This is followed by using in each remaining sub-frames the middle frame for the second level of reference and so on. This imposes to have an encoding and decoding order different from the visual one, as represented on the Figures by the coding references {A, B, ..., F, ...}. Each frame uses then as reference a close one of upper reference level, as illustrated by the arrows in Fig. 2-(a) and (b). It is to be noted that for the most important frames (first levels of importance), the coding efficiency is not optimal, since the separation between prediction and reference frames in original order of GOP can be greater than one. Still, this may be compensated by the fact that the latest frames should offer better compression rate, for they are closer to the Intra one (decreasing separation between the reference and predicted frames). In the case where one wants to use an Intra frame as first encoded image, the GOP decomposition can easily be adapted, as illustrated in Fig. 3.



(a) GOP size 7 (b) GOP size 15

Fig. 2. Filtering applied to GOP of $2^L - 1$ frames.

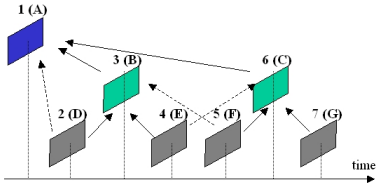


Fig. 3. Filtering applied to 7 frames GOP, with I frame first.

3.2. Generalisation to other GOP sizes

The GOP decomposition principle presented before can be generalised to any GOP sizes, and easily adapted when one wants sub-frame rates of target ratio R greater than 2, albeit with some possible efficiency loss. One will :

- select $R - 1$ reference frames at each level (with for instance the Intra frame the first of them) as the medium ones, where medium values are defined as $m_i = \lfloor i(GOP_{size} + 1)/R \rfloor$ for $i = 1, ..., R - 1$, and each part between mediums are defined as sub-GOPs ;
- repeat for each sub-GOP : take as reference frames the medium ones and define accordingly R remaining sub-GOPs.

This is illustrated in Fig. 4 for a GOP of 17 frames, with sub-frame rate $R = 3$ for the two first levels and sub-frame rate $R = 2$ for the last decomposition level.

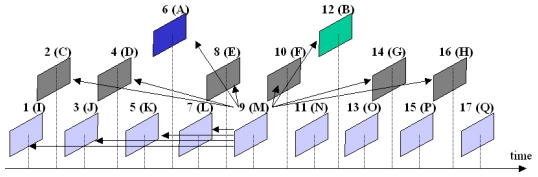


Fig. 4. Filtering applied to 17 frames GOP.

3.3. Influence of the H.264 multi-reference capability

Contrarily to previous video coding standards that were using simple reference mode, that is to say for which the prediction of inter frames could only be done with respect to a given preceding picture, H.264 allows to use up to 16 different frames as reference for each P-slice. In practice, this capability allows that all previous pictures in encoding order can be used as reference for the current frame while keeping the scalability feature, as each level of refinement is encoded one after the other. This is illustrated in Fig. 4 for the 9^{th} frame (M in encoding order), where all its possible references are shown by arrows. In particular, it is to be noted that the frame can be predicted from macroblocks of previous frames at the same refinement level. When the multi-reference option is activated, the proposed structure actually becomes time-variant.

To better illustrate this, Table 1 provides an example done with 'Foreman' sequence frames (QCIF format, 15Hz, 64 kbps, 126 frames) of respective percentages of MB being predicted from previous frames in the case of decomposition given by Fig. 2-(a). Naturally, those figures depend of the considered sequence, the quantification parameters and the size of the GOP. Nevertheless it was observed in simulations that at least about 45 % of the MB were predicted from the previous frame in display order, and that the Intra frame is used about twice more as direct reference than in the classical encoding scheme.

	Reference MB in previous frames (%)					
current	#-1	#-2	#-3	#-4	#-5	#-6
B	100					
C	21.8	78.2				
D	8.9	74.1	17			
E	17.4	6.6	43.9	32.1		
F	10.8	2.8	54.3	1	31.1	
G	29.1	3.3	6.9	46.4	1.6	12.8

Table 1. Influence of multi-reference capability : repartition of reference MB in previous frames for the scalable scheme.

4. SIMULATION CHAIN

The purpose of the presented scheme is to introduce temporal scalability within an a priori non-scalable codestream by shuffling the frames in a GOP to distribute them as regularly as possible. This will be easily implemented based on the consideration that two different frame numbering solutions do exist in the H264/MPEG-4 AVC standard. The first one (*frame_num*), correspond to the decoding order of access units, but does not necessarily indicate the final display order that the decoder will use. The second one (POC or *Picture Order Count*) corresponds to the display order of the decoded frames (or fields) that will be used by the decoder for display order.

As presented in Section 3, the most important frames, corresponding to those decoded from the lowest frame rates, are regularly distributed and spaces between them are filled with the less important frames, that will be decoded only at higher frame rates. Re-arranging the frames according to their encoding order, it can be seen that the re-arranged sequence can then be encoded as classically done by any H.264/MPEG-4 AVC encoder, and decoded accordingly. A complete backward compatibility at the decoder side can then be obtained by forcing the encoder to use the original decoding order as POC values.

The simulation chain used to obtain the scalability feature is presented in Fig. 5. The shuffling operation, illustrated in the figure for the case of a 7 frames GOP, is applied directly on the video sequence to be encoded by means of an interleaver, before the actual H.264 encoding process which is only modified to the extent of knowing the used shuffling table, to allow for the insertion of the correct decoding order in the POC fields. The H.264 codestream transmitted is then fully compliant with the standard, and can be directly decoded by any standard compliant H.264/MPEG-4 AVC decoder. The only drawback of this scheme is that the shuffling operation introduces a delay and the necessity of buffering frames, both at the encoder and decoder sides.

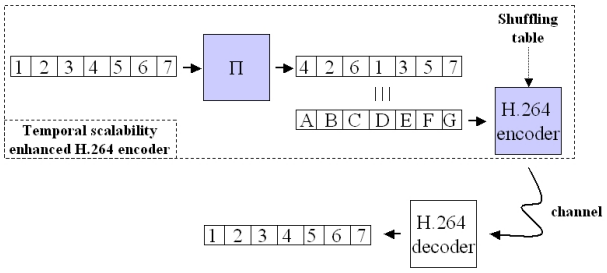


Fig. 5. Simulation chain.

5. NUMERICAL RESULTS

The simulations were done with the Joint verification Model (JM) version 8.4 [6]. Unfortunately, this verification model limits the number of frames that can be used as reference for inter prediction. As such, results presented hereafter were drawn only for GOP of size 7 to allow for decomposition to be fully applied.

Furthermore, to better illustrate the interest of the scalable filtering method introduced in this paper, we present the results that would be obtained with a fully pyramidal reorganisation of the GOP. Such a decomposition, illustrated by Fig. 6, can be seen as a simple variation of the shuffling operation performed to decompose the GOP. Practically, it will provide better results in terms of compression as each frame is at minimal distance of its main reference ones in this scheme, contrarily to what happens with the scalable method, where the regular repartition leads to increasing the distance from their reference for the frames of the first refinement levels. On the other hand, the pyramidal decomposition does not provide an efficient scalability feature. Indeed, if the last refinement levels are lost, the reconstructed sequence will present long gaps of inactivity.

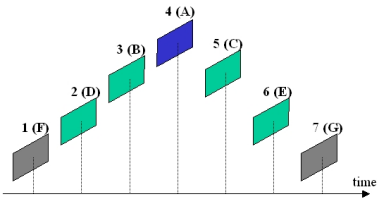


Fig. 6. Pyramidal decomposition for a GOP of 7 frames.

The mean MSE values for 'Foreman', 'Mobile' and 'Akiyo' reference sequences (QCIF, 15 Hz, M=7) are given in Table 2 for three different decompositions. The first one tally with the classical encoding solution (IPPPPPP), the second one with the pyramidal decomposition corresponding to Fig. 6 shuffling pattern and the third one with the scalable decomposition obtained with Fig. 2-(a) shuffling pattern. In each case, the quantification parameters have been adjusted to yield a 64 kbps target bitrate with a better resolution for the higher levels of the decompositions. Quite logically, it can be observed for the scalable and pyramidal methods curves that the three first frames in terms of importance have always better PSNR values than the four other frames, and it is interesting to note that this is obtained with no degradation of the mean MSE value.

		Normal	Pyramidal	Scalable
Foreman	bitrate	63.84	63.77	63.7
(196 fr.)	Av. MSE	32.19dB	32.36dB	32.28dB
Mobile	bitrate	63.93	63.59	63.94
(147 fr.)	Av. MSE	23.96dB	24.05dB	24.08dB
Akiyo	bitrate	63.99	63.76	63.61
(147 fr.)	Av. MSE	40.19dB	40.33dB	40.32dB

Table 2. Comparing the performance of shuffling and standard modes for different reference video sequences.

Those results are illustrated by Fig. 7 which gives the PSNR evolution for the three first GOPs of the 'Foreman' sequence. The scalable and pyramidal decompositions show the interest of placing the Intra frame in the middle of the GOP and their comparison shows that the scalability feature is obtained at minimal cost, as these 196 frames sequences yield average MSE of 32.19 dB and 32.36 dB respectively, while the classical ordering one has a mean MSE of 32.19 dB for same bitrate.

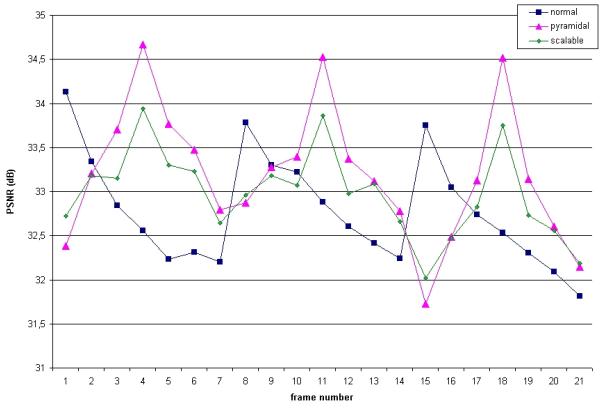


Fig. 7. 'Foreman' PSNR evolution for different modes.

6. CONCLUSIONS

A solution allowing to introduce temporal scalability in standard compliant video encoded H.264/AVC bitstreams is proposed, that rely on a decomposition of the GOP. This decomposition was placed in the framework of adaptive hierarchical filterbanks, and it was shown that in practice it can be applied via a simple frame shuffle operation thanks to H.264/AVC standard properties. Simulation results show that scalability is obtained with no or little performance degradation compared to the classical or other non-scalable decompositions.

7. REFERENCES

[1] *Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264, ISOIEC 14496-10 AVC), Doc JVT-G050r1*, Geneva, Switzerland, May 2003.

[2] L. Blaszak, M. Domanski, A. Luczak, and S. Mackowiak, "Avc video coders with spatial and temporal scalability," in *Proc. of PCS'03*, Saint-Malo, France, April 2003, pp. 41–47.

[3] H. Schwarz, D. Marpe, and T. Wiegand, *Subband extension for H.264/AVC, Doc JVT-K023*, Munich, Germany, March 2004.

[4] J.-R. Ohm, *Multimedia Communication Technology*, Springer, 2004.

[5] C. Tillier and B. Pesquet-Popescu, "3d, 3-band, 3-tap temporal lifting for scalable video coding," in *Proc. of IEEE ICIP2003*, 2003.

[6] *Joint verification model for H.264 (JM 8.4)*, http ://iphome.hhi.de/suehring/tml, July 2004.