MULTIPLE DESCRIPTIONS SCALABLE VIDEO CODING

Christophe Tillier, Béatrice Pesquet-Popescu

Télécom Paris
Signal and Image Proc. Dept.
46, rue Barrault, 75634 Paris, FRANCE
e-mail : {tillier, pesquet}@tsi.enst.fr

Mihaela van der Schaar

Univ. of California Davis
Dept. of Elect. and Computer Eng.
One Shields Avenue, 3129 Kemper Hall
Davis, CA 95616-5294

ABSTRACT

Multiple description coding (MDC) is a joint source-channel coding technique specifically designed for real-time multimedia applications over best effort switched packet networks (such as Internet), in order to cope with packet losses due to transmission errors or network congestion. Error resilience of transmitted bitstreams is thus significantly increased, but this does not solve problems like bitstream adaptation to bandwidth variations or receiver characteristics, which are in turn addressed by scalable coding techniques.

In this paper, we present a new method of multiple description coding of scalable video, combining the scalability features with MDC. It is based on a motion-compensated spatio-temporal subband decomposition, where the redundancy is tunable in a frame-like non-linear temporal representation.

1. INTRODUCTION

Video communication over Internet and wireless networks is becoming increasingly popular. However, reliable transmission of the video over such networks poses many challenges. This is not just due to the inherently lower bandwidth provided by these networks as compared with traditional delivery networks, but also due to the associated problems such as congestion, competing traffic, fading, interference, mobility, all of which lead to losses. Multiple description coding (MDC) includes a set of techniques that can improve the robustness of video to such losses [1], [2], [3], [4]. MDC involves creating correlated coded representations of the video and transmitting them on separate channels for improved error resilience. This is done in a way that acceptable video quality can be obtained using a subset of the descriptions, with the quality improving as the number of subsets received increases. The motivation for using MDC is to introduce redundancy at the source coder to combat transmission failures. In this sense, MDC is a way of accomplishing joint source and channel coding. This trade-off between resilience and redundancy needs to be properly exploited for successful video delivery. Moreover, studies have shown [5] that MDC has advantages over other error resilient coding techniques (such as layered coding with unequal error protection) when the network is very lossy (e.g. packet loss rate higher than 25%). However, most MDC coding techniques proposed so far are built on top of a non-scalable motion compensated prediction framework. A key disadvantage of this approach is that non-scalable MDC can only improve the error resilience of video transmitted over unreliable wired and wireless networks, but it is not able to address two other important challenges associated with the robust transmission of video over unreliable networks: adaptation to bandwidth variations and receiving device characteristics. In other words, a shortcoming of several existing MDC techniques is that the achievable redundancy and one-description distortion is fixed. To accommodate varying network environments and QoS requirements, it is desirable to have a coder that can realize variable trade-off between redundancy and one-description distortion, or essentially total rate and average distortion. In a preliminary work [6], we have developed a new scheme for MDC, using wavelet based interframe coding schemes [7], that results in highly efficient and error resilient bitstreams, which are also spatio-temporal-SNR scalable to allow easy adaptation to network and device variations. We use the inherent ability of lifting based motion compensated temporal filtering schemes [8], [9], [10] to recover missing information, and partition the resulting bitstreams so that the video may be transmitted over multiple channels. We vary the amount of redundancy introduced through this partitioning to increase the robustness to packet-losses, and present video quality results under different loss scenarios and bandwidth conditions.

In this paper we propose an improved scheme for multiple description scalable video coding, in which the redundancy is achieved by temporal oversampling of a motion compensated 3-band lifting structure [11]. The advantage of these schemes are that the 3-band representation allows for a reduced redundancy compared with an oversampled 2-band scheme. The redundancy factor can be tuned according to the number of decomposition levels, resulting in a very flexible and robust scheme.

The paper continues in Section 2 by reviewing the motion-compensated 3-band lifting scheme. In Section 3 we introduce the new MDC temporal scalable video codec. Section 4 presents some simulation results and we conclude in Section 5.

2. THREE-BAND MOTION-COMPENSATED LIFTING SCHEME

The motion-compensated 3-band scheme on which the proposed MDC system is based is presented in Fig. 1.

We use a "Haar-like" motion-compensated (MC) 3-band (3B) scheme, as presented in [11], where the two predict operators are the simplest forward and backward temporal predictors only based

on one past resp. future frame, and the update operators are also very simple, involving only one detail frame. The three output subbands are thus computed as:

$$h_t^+(\boldsymbol{n}) = x_{3t+1}(\boldsymbol{n}) - x_{3t}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^+),\ t \in \mathbb{N}$$
$$h_t^-(\boldsymbol{m}) = x_{3t-1}(\boldsymbol{m}) - x_{3t}(\boldsymbol{m} - \boldsymbol{v}_{3t-1}^-),\ t \in \mathbb{N}^*$$
$$l_t(\boldsymbol{p}) = \frac{1}{2}x_{3t}(\boldsymbol{p}) + \frac{1}{4}\left[h_t^+(\boldsymbol{p} + \boldsymbol{v}_{3t+1}^+) + h_t^-(\boldsymbol{p} + \boldsymbol{v}_{3t-1}^-)\right],$$

where we denoted by $\boldsymbol{v}_t^+$ the forward motion vector used to predict frame t and by $\boldsymbol{v}_t^-$ the backward motion vector corresponding to the same frame. The spatial positions $\boldsymbol{m}$, $\boldsymbol{n}$, $\boldsymbol{p}$ are on the same motion trajectory respectively in the frames $3t-1$, $3t+1$ and $3t$.

This scheme enters the classical lifting framework, where the perfect reconstruction is easily achieved by inverting the order of operations and the sign of the operators.

3. TEMPORAL SCALABLE MDC OF VIDEO

First, we present how the two descriptions are built, then we analyse the redundancy of the scheme. Further, the reconstruction is discussed when one description is lost and when both of them are received and finally building the MDC scheme on a temporal decomposition with several levels is presented.

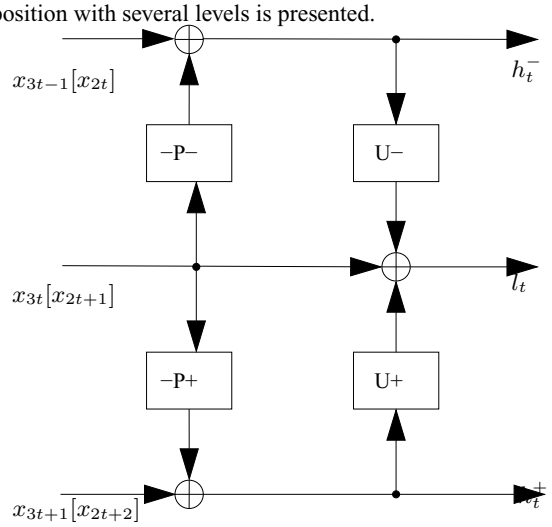

Fig. 1. Three-band temporal lifting scheme (in brackets, the input polyphase components for the oversampled structure used in the MDC system).

3.1. Building the two descriptions

An MDC scalable video system has been proposed in [6], where the two descriptions had in common the low-pass band and the two types of detail subbands belong to different descriptions.

The multiple description scheme we propose here is built on the temporal 3B MC structure in Fig.1, by subsampling by a factor 2, instead of a factor 3. This leads actually to an overcomplete temporal 3B structure, with non-linear operators involving ME/MC. The two descriptions correspond to the output subband frames of this scheme for even, resp. odd t.

The temporal synchronisation of the frames in the two descriptions thus obtained with the original frames and with the subbands resulting from a non redundant 3B encoder is illustrated in Fig. 2. We keep the lower index notation for the temporal moment, while the upper index refers to the description number.

3.2. Redundancy analysis for one level

Let us now compute the redundancy factor of the above scheme. We denote by L the sequence length. In the non-redundant encoder the frames are processed by three: $3n$, $3n+1$, $3n+2$, with $n \in \{0, \ldots, N\}$. The sequence length is therefore $L = 3(N+1)$. In the redundant encoder the frames processed in a GOF are $2n'$, $2n'+1$, $2n'+2$, where $n' \in \{0, \ldots, N'\}$. The number of output frames in the two descriptions is therefore $L' = 3(N'+1)$ and the redundancy factor, denoted by ρ will be: $\rho = \frac{L'}{L}$. In order to compute ρ, we need an entire number of GOFs in each description (and we consider the same number of GOFs in the two descriptions) and also in the non-redundant encoder. This means that the temporal moment $3n+2$ for $n = N$ needs to be the same as $2n'+2$ for $n' = N'$. We get then $3N = 2N'$. By introducing $P = \frac{N}{2} = \frac{N'}{3}$, we obtain:

$$\rho = \frac{L'}{L} = \frac{3(N'+1)}{3(N+1)} = \frac{3(3P+1)}{3(2P+1)} = \frac{3P+1}{2P+1}$$

For $P \to \infty$, the redundancy factor $\rho \to 3/2$, as expected, but it is slightly higher for small P.

3.3. Reconstruction of one redundant level

If one of the description is lost, then only three over four of the original frames can be directly reconstructed. The missing frames can be interpolated from their neighbors, by averaging these frames, after motion compensation. In order to be able to perform the motion compensation, the available motion vector fields are extended in the opposite direction to obtain the missing fields.

If both descriptions are received, all the original frames can be decoded, so perfect reconstruction is achieved. Moreover, for each even frame $2n$, $n \in \{1, \ldots, N-1\}$ two reconstruction options are possible at the decoder side, from one description or from the other. This redundancy can be exploited to improve the quality of the reconstructed frame. For example, it can be obtained as the mean of the two lower quality reconstructed frames obtained independently from the two descriptions.

3.4. Multiple levels

The compression performance of the subband scheme being dependent on the number of temporal decomposition levels, the above redundant scheme can be extended to several levels. A possible extension consists in interlacing the frames of the two approximation subbands at the first temporal level and then iterating the scheme on this new "sequence". At the second level we have therefore $1/2 * L$ detail frames and $1/2 * 1/2 * L$ approximation frames. The redundancy factor is here $1/(1 + 1/2 + 1/4) = 7/4$ (asymptotically). For a number of decomposition levels $j_{max} \to \infty$, the redundancy factor tends to 2.

Original sequence	0	1	2	3	4	5	6	7	8
Non redudant encoder	h_1^-	l_1	h_1^+	h_2^-	l_2	h_2^+	h_3^-	l_3	h_3^+
1st description	h_1^{1-}	l_1^1	h_1^{1+}		h_3^{1-}	l_3^1	h_3^{1+}		
2nd description			h_2^{2-}	l_2^2	h_2^{2+}		h_4^{2-}	l_4^2	h_4^{2+}

Fig. 2. Temporal synchronisation of the frames in the two descriptions (one decomposition level).

Note that in our redundancy analysis we only considered the number of frames. However, approximation and detail frames are not allocated the same number of bits, which also depends on the temporal level. The bit allocation procedure can also be used to tune the quality of the reconstructed sequence from both descriptions, taking into account that one over two frames in this case are obtained by averaging frames from the two descriptions.

In order to reduce the redundancy of the scheme, we propose to iterate in a different way (see Fig. 3). For the first levels, groups of three frames are alternatively processed in one description or the other, the resulting approximation frames interlaced and only at the last level (in Fig. 3, the second one) the redundancy is achieved by overlapping the GOFs. The reconstruction from one or both descriptions follows the same principles as for the scheme with one decomposition level.

The redundancy in this case is introduced at the coarsest temporal decomposition level. We have for j_{max} levels an asymptotic redundancy factor $\rho = 1 + 1/3^{j_{max}} * 3/2$, much reduced compared with the first scheme.

4. SIMULATION RESULTS

Spatio-temporal coefficients and motion vectors (MV) are encoded within the MC-EZBC framework [7, 12], where MV fields are first represented as quad-tree maps and MV values are encoded with a 0-order arithmetic coder, in raster-scan order. The MC temporal filteirng is performed using Hierarchical Variable Size Block Matching (HVBSM) algorithm with block sizes varying from 64×64 to 4×4 and an 1/8th pel accuracy for MC.

We have tested the proposed algorithm on several CIF sequences at 30fps. In Figs. 4 and 5, we compare the rate-distortion performance of the non-robust 3-band scheme with that of the MDC central decoder on "foreman" and "mobile" sequences. Two decomposition levels were used, corresponding to 15% redundancy. One can remark that the loss in coding performance in noiseless environment of the redundant codec is around 1 dB. The available bitrate is equally distributed between the two descriptions. The lateral distortions are thus represented for the global bitrate, each one of them using however only half of it. Note that the second description has higher reconstruction error, which is related to a slight asymmetry in the reconstruction scheme. Indeed, when loosing the first description, the detail frames of the second description at the first level are missing, while when loosing the second description the details of the first one exist. In Fig. 5 note for example that at 500 kbs the reconstruction when only the first description is received has comparable distortion with that of the non-redundant

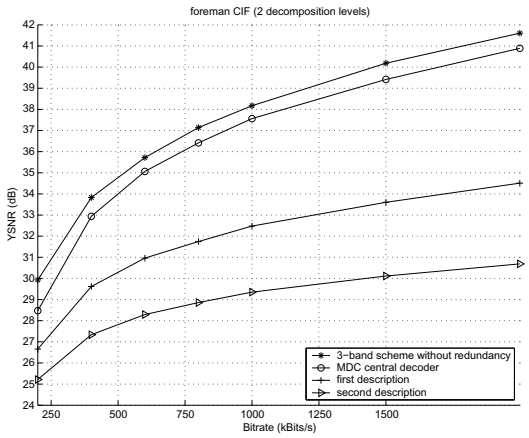

Fig. 4. Central and lateral distortions of the MDC scheme compared with the non robust 3-band codec for two decomposition levels ("foreman" CIF sequence, 30 fps). The bitrate corresponds to the global rate for the robust codec (both descriptions).

codec, which shows that the interpolation strategy is very effective at this bitrate.

5. CONCLUSION

In this paper we have presented a new framework for building multiple descriptions of scalable video coding, in which the robustness is achieved by temporal oversampling of a motion-compensated 3-band lifting structure. The redundancy factor is tunable over the number of temporal decomposition levels, and for typical values (15%), the loss in coding performance in noiseless environment is only around 1 dB compared with the non robust scheme. The reconstruction based on a single description has good performance, due to a motion-compensated temporal interpolation of missing frames.

6. REFERENCES

[1] V.A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Transactions on Information Theory*, vol. 39, pp. 821–834, 1993.

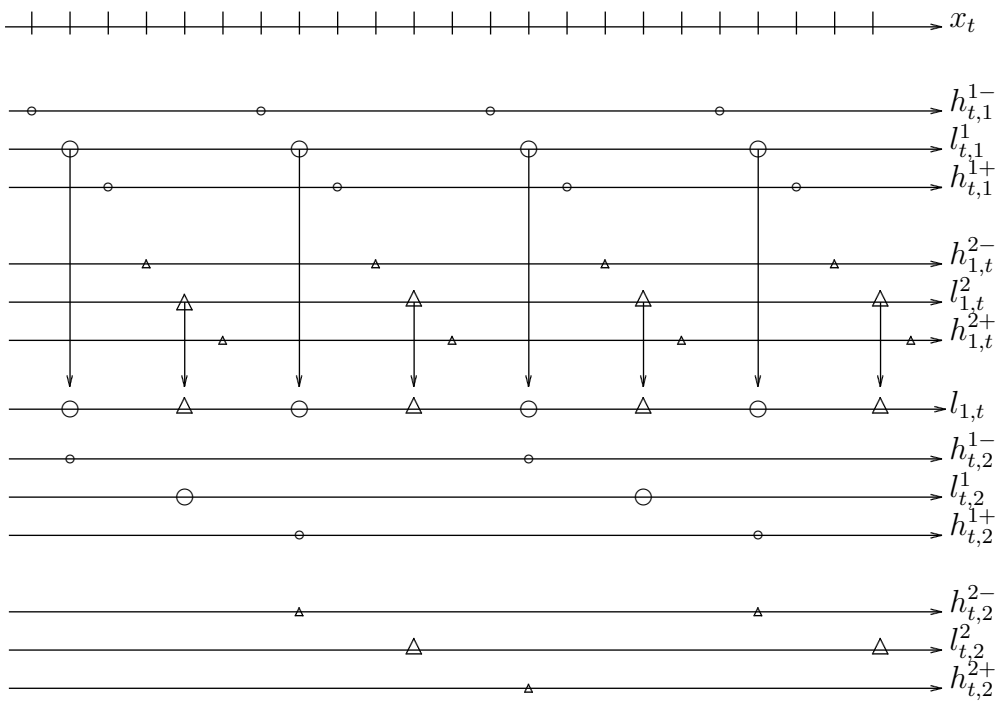

Fig. 3. MDC scheme with reduced redundancy for 2 temporal decomposition levels (first level is non-redundant, second level introduces redundancy). Circles and triangles represent the frames in the first, resp. second description.

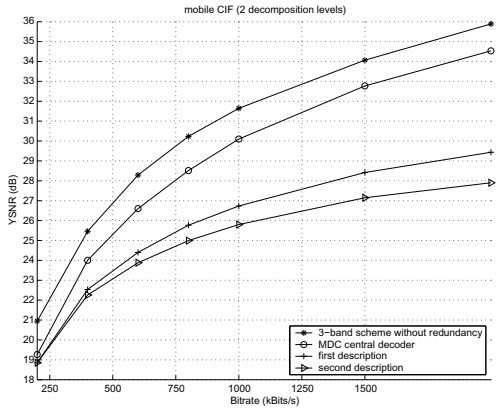

Fig. 5. Central and lateral distortions of the MDC scheme compared with the non robust 3-band codec for two decomposition levels ("mobile" CIF sequence, 30 fps). The bitrate corresponds to the global rate for the robust codec (both descriptions).

[2] V.K. Goyal and J. Kovacevic, "Optimal multiple description transform coding of gaussian vectors," in *Proceedings of the DCC'98*, Snowbird, UT, Mar. 1998, pp. 388 – 397.

[3] J. Apostolopoulos, T. Wong, W. Tan, and S. Wee, "On multiple description streaming with content delivery networks," in *Proceedings of the IEEE INFOCOM Conf.*, 2002, pp. 1736–1745.

[4] I.V. Bajic and J.W. Woods, "Domain-based multiple description coding of images and video," *IEEE Trans. on Image Proc.*, vol. 12, pp. 1211–1225, 2003.

[5] A.R. Reibman, H. Jafarkhani, Y. Wang, M.T. Orchard, and R. Puri, "Multiple-description video coding using motion-compensated temporal prediction," *IEEE Trans on Circ. and Syst. for Video Tech.*, vol. 12, pp. 193–204, 2002.

[6] M. van der Schaar and D.S. Turaga, "Multiple description scalable coding using wavelet-based motion compensated temporal filtering," in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Sept. 2003.

[7] S.J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Proc.*, vol. 8, pp. 155–167, 1999.

[8] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Salt Lake City, UT, May 2001.

[9] C. Tillier, B. Pesquet-Popescu, Y. Zhan, and H. Heijmans, "Scalable video compression with temporal lifting using 5/3 filters," in *Picture Coding Symposium, PCS-2003*, St. Malo, France, Apr. 2003.

[10] G. Pau, C. Tillier, B. Pesquet-Popescu, and H. Heijmans, "Motion compensation and scalability in lifting-based video coding," submitted to Image Communication, June, 2003.

[11] C. Tillier and B. Pesquet-Popescu, "3D, 3-band, 3-tap temporal lifting for scalable video coding," in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[12] "3D MC-EZBC software package," available on the MPEG CVS repository.