Scene-Cut Processing in Motion-Compensated Temporal Filtering

Maria Trocan and Béatrice Pesquet-Popescu

ENST, Signal and Image Processing Department
46, rue Barrault, 75634 Paris, France
{trocan, pesquet}@tsi.enst.fr

Abstract. Motion-compensated temporal filtering (MCTF) is a powerful technique entering scalable video coding schemes. However, its performance decreases significantly if the video signal correlation is poor and, in particular, when scene-cuts occur. In this paper we propose an improved structure for MCTF by detecting and processing the scene-cuts that may appear in video sequences. It significantly reduces the ghosting artefacts in the temporal approximation subband frames, providing a higher quality temporal scalability, and dramatically improves the global coding efficiency when such abrupt transitions happen.

1 Introduction

The 3-D subband schemes ($t + 2D$) exploit the temporal interframe redundancy by applying an open-loop temporal wavelet transform over the frames of a video sequence. Temporally filtered subband frames are further spatially decomposed and can be encoded by different algorithms such as 3D-SPIHT [1] or MC-EZBC [2].

A weakness of the existing $t + 2D$ video codecs is related to the way the temporal filtering behaves near scene changes. Usually, the input video signal is partitioned into GOPs and temporally filtered without checking the correlation between the GOP frames. Moreover, the sliding window implementation of the temporal filtering is done using frames from adjacent GOPs in the processing of the current GOP. When the input signal involves complex motion transitions, and especially scene-cuts, this can translate into inefficient prediction/update operations, leading to poor quality results and also to reduced temporal scalability capabilities.

Several attempts to avoid the artefacts related to these abrupt changes have already been proposed for *hybrid* coding, such as the scene-cut detection and content-based sampling of video sequences [3] or video segmentation using encoding cost data [4], alleviating but not completely solving this problem.

In this paper we present a motion-compensated temporal transform coding scheme, specifically adapted to the detection and processing of the uncorrelated shots of the input video sequence. After the scene-cuts are detected, we encode each set of frames between two consecutive scene-cuts separately, by adapting the temporal filtering to cope with arbitrary number of frames in a shot. An advantage of our scheme is that scene-cuts once eliminated, MCTF efficiency is maximal, as for highly-correlated video signals. The problem is related to border effects, and therefore is much easier to cope with in case of Haar MCTF. However, it has been shown [5–7] that the use of longer bidirectional filters, like the 5/3 filter bank, can take better advantage of the temporal redundancy between frames. Existing methods for adaptive GOP structure in the MCTF framework [8, 9] basically detect changes and limit the number of temporal decomposition levels based on a measure of unconnected pixel percentage. However, compared with our approach, this technique does not make a strict correspondence between the scene cut and the GOP boundary. Our proposed approach varies the GOP size only on the frames previous to the transition, and these frames are encoded in several GOPs of power of two sizes. In this way, the scene cut does not span any GOP. We present therefore our scene-cut processing method in the framework of 5/3 MCTF, but our proposal can be adapted to other temporal filters.

The paper is organized as follows: in the next section, we recall the classical motion-compensated 5/3 temporal transform and present the method proposed in this framework for scene-cuts detection and processing. Section 3 illustrates by experimental results the coding performance of the proposed scheme. We conclude in Section 4.

2 Scene-cut Detection and Processing

The MCTF approach consists in a hierarchical open-loopsubband motion- compensated decomposition. Let us denote by x_t the original frames, t being the time index, and by h_t and l_t the high-frequency (detail) and low-frequency (approximation) subband frames, respectively. For the 5/3 filterbank implemented in lifting form, the operators allowing to compute these subbands are bidirectional, and the equations have the following form (see also Fig. 1):

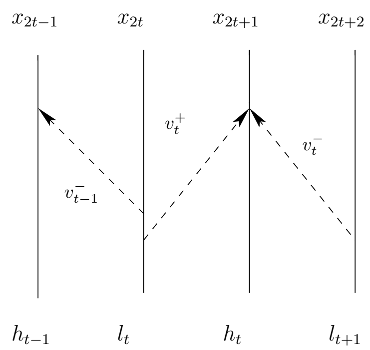


Fig. 1. MCTF with bidirectional predict and update lifting steps.

$$
\begin{cases} h_t = x_{2t+1} - \frac{1}{2}\big(\mathcal{F}(x_{2t}, \boldsymbol{v}_t^{+}) + \mathcal{F}(x_{2t+2}, \boldsymbol{v}_t^{-})\big) \\ l_t = x_{2t} + \frac{1}{4}\big(\mathcal{F}^{-1}(h_{t-1}, \boldsymbol{v}_{t-1}^{-}) + \mathcal{F}^{-1}(h_t, \boldsymbol{v}_t^{+})\big) \end{cases} \tag{1}
$$

where $\mathcal{F}(x_t, \boldsymbol{v}_t)$ is the motion prediction operator, compensating the frame x_t by projection in the direction of the motion vector field $\boldsymbol{v}_t$, and $\boldsymbol{v}_t^{+}$, $\boldsymbol{v}_t^{-}$ are the forward and backward motion vectors predicting x_{2t+1}, respectively . The notation $\mathcal{F}^{-1}(h_t, \boldsymbol{v}_t)$ corresponds to the compensation of the h_t frame in the opposite direction of the motion vector field $\boldsymbol{v}_t$. Indeed, in general the motion prediction is not an invertible operator. Unconnected and multiple connected pixels are processed as detailed in [10].

When the input sequence involves complex motion transitions, this can translate to inefficient prediction/update operations, leading to poor quality results and temporal scalability capabilities, as illustrated in Fig.2. One can remark in particular the energy of the detail frames, which need to be encoded, and also the poor visual quality of the approximation frame, very penalizing for temporal scalability.

In the following, we suppose the scene-cuts have been detected, and we present the algorithm used for change detection at the end of this section.

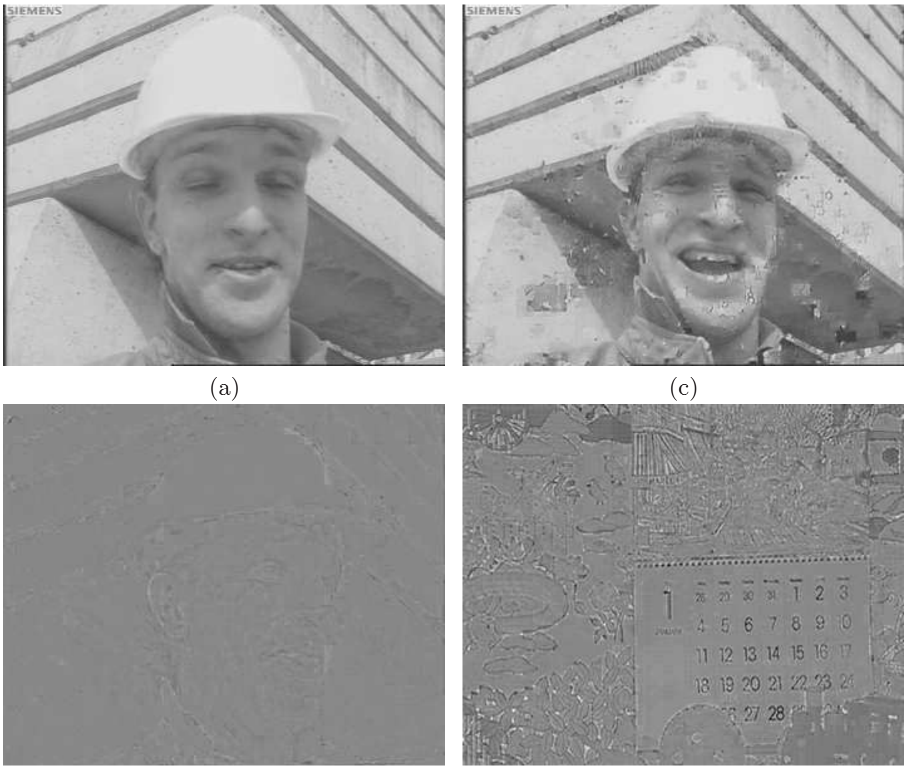
First, the temporal filtering needs to be changed in order not to filter over a scene-cut. The second modification is related to the encoding of the last group of frames (GOF) before the scene-cut.

To this end, both the predict and update steps have to be modified near the end of the first scene, as illustrated in Fig. 3. For sequences processed homogeneously, the temporal subbands resulting from the MCTF are encoded by GOFs of 2^L frames, where L is the number of temporal decomposition levels that were performed. When a scene-cut occurs in a sequence, the GOF just before the change will have in general a different number of frames. If we denote its number of frames by A_n and write this number as:

$$
A_n = (a_0 a_1 \dots a_{L-1})_2 = \sum_{l=0}^{L-1} a_l 2^l,
$$

then we shall decompose the GOF in smaller GOFs, in decreasing order of their size: $a_l\, 2^l,\ l \in \{0, \dots, L-1\},\ a_l \in \{0, 1\}$, which will be filtered and encoded separately. This also corresponds to changing the number of temporal decomposition levels and filtering operations for these sub-GOFs. Indeed, we can do only l temporal decomposition levels for a sub-GOF of size 2^l, $l < L$. Moreover, the prediction across the scene-cut is inhibited, as well as the usage of the reverse motion vector field over the same transition, during the update step. After the scene-cut, the normal filtering with “sliding window” is started, the effect of the scene cut being only a slight modification of the filters to take into account the induced border effects.

Now that we have explained the modifications in filtering and coding in order to take into account scene changes, we turn to the detection of such transitions.



(a) (c)

(b) (d)

Fig. 2. Approximation (a) and detail (b) frames in a GOF without scene-cut. Approximation (c) and detail (d) frames when the GOF contains a scene-cut (first part: "foreman" sequence, second part: "mobile" sequence).

Several criteria for scene-cut detection have been proposed in the literature, like: the variation of the relative energy of the displaced frame difference (DFD) along the sequence [11], the energy and angle distribution of the motion vector fields in consecutive frames [12], by keeping track of the percentage of the unconnected pixels, estimated after motion estimation [13] or using unsupervised segmentation and object tracking [14].

For our simulation we have used, as detection criteria, the variation of the relative energy of the DFD along the sequence. If the displaced frame difference between two successive frames is computed as:

$$d_t = DFD(x_t, x_{t+1}) = x_{t+1} - \mathcal{F}(x_t, \boldsymbol{v}_t) \tag{2}$$

then the variation of the relative energy of the DFD is computed as:

$$\mathit{\Delta}_{2t} = \frac{d_{2t}^2}{d_{2t-1}^2} \tag{3}$$

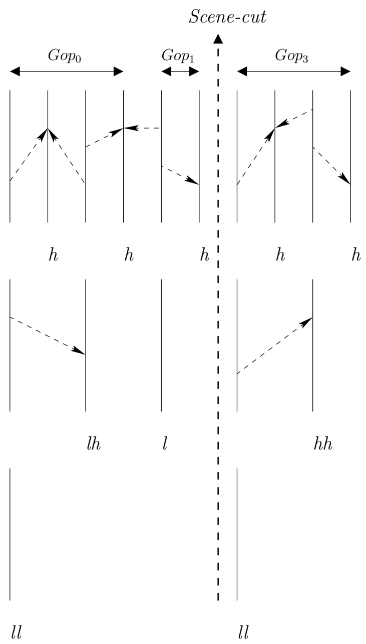


Fig. 3. Scene-cut processing over two temporal levels for a 10-frames video shot.

When the input signal is highly-correlated, the variation of the relative energy of the DFD along the sequence is almost constant (i.e. $\Delta \approx 1$). We say a scene-cut is detected when the variation of relative energy has a rapid change. For appropriately chosen parameters τ_1 and τ_2, we say that the scene-cut occurs after the frame x_{2t+1} when:

$$\begin{cases} |\Delta_{2t} - 1| < \tau_1 \\ |\Delta_{2t+1} - 1| > \tau_2 \end{cases} \tag{4}$$

3 Experimental results

For simulations, we considered a high-definition video sequence (HD format: 1920×1280, 60 fps) from the "Erin Brockovich" movie, containing 180 frames and 3 scene-cuts: after the 44^{th}, the 80^{th} and respectively, the 161^{th} frame. Moreover, in order to work on a representative set of test sequences, we also built several test sequences obtained by concatenating parts of the standard CIF sequences at 30 fps: Foreman and Mobile (i.e.: MF_18×16 - video file containing the first 18 frames from Mobile and the next 16 frames from Foreman, FM_16×16 - with the first 16 frames from Foreman, followed by the first 16 frames from Mobile). The aim was to test all possible configurations for the number of frames in the GOF previous to the scene-cut. In order to detect the abrupt scene transitions, the values of τ_1 and τ_2 were empirically determined as being equal to 0.1 and

0.4, respectively. These parameters ensured that all the scene-cuts were detected and no false alarms appeared for the considered sequences. Sequences with fade or dissolve transitions can be processed with the described MCTF scheme, but the detection method should be replaced with an appropriate one, as described in [15].

The target number of decomposition levels for motion-compensated 5/3 temporal filtering is $L = 4$. The coding procedure is based on the MC-EZBC codec and the used motion estimation algorithm is a Hierarchical Variable Size Block Matching (HVSBM) one. The motion vectors have been estimated with $1/8^{th}$ pixel accuracy and the temporal subbands were spatially decomposed over 4 levels with the biorthogonal 9/7 wavelet. The encoding of the entire YUV sequence was performed, but the results are further expressed only in terms of average YSNR.

YSNR (dB)	6000 kbs	8000 kbs	12000 kbs
SC-MCTF	36.4227	36.8639	37.6387
MCTF	34.9281	35.7519	36.5217

Table 1. PSNR results of 5/3 MCTF with and without scene-cut processing for "Erin Bronckovich" - (HD, 60fps, 180 frames).

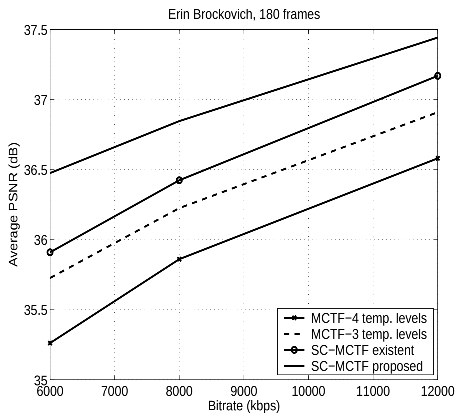


Fig. 4. Rate-distortion curves for 180-frames HD Erin Brockovich sequence.

The importance of correctly processing the scene-cuts is illustrated in Fig. 4, Fig. 5 , as well as in the Tables 1-2, where the rate-distortion performances for 5/3 MCTF with (denoted in these tables by SC-MCTF) and without (simply denoted by MCTF) scene-cut processing are compared. It can be easily noticed that in all the cases our scheme performs better, achieving a gain between 0.5

MF_18x16 sequence (30fps)				
YSNR (dB)	512 kbs	768 kbs	1024 kbs	1536 kbs
SC-MCTF	30.1185	32.3141	33.7612	35.6489
MCTF	23.9811	28.5192	30.4135	32.8334
FM_16x16 sequence (30fps)				
YSNR (dB)	512 kbs	768 kbs	1024 kbs	1536 kbs
SC-MCTF	30.3151	32.7043	34.1021	35.9510
MCTF	26.4706	30.3061	31.8275	33.8650

Table 2. PSNR results of 5/3 MCTF with and without scene-cut processing for the "MF_18x16" and "FM_16x16" sequences.

dB and 2.0 dB over a classical MCTF. Results in Fig. 4 indicate that reducing the GOF size (from 16 to 8 frames) can alleviate the problem of scene-cuts by decreasing their influence, but a correct processing of these zones allows us both to take advantage of the temporal correlation in homogeneous shots and to increase the coding efficiency. It can also be observed that our proposed technique outperforms the one described in [8, 9].

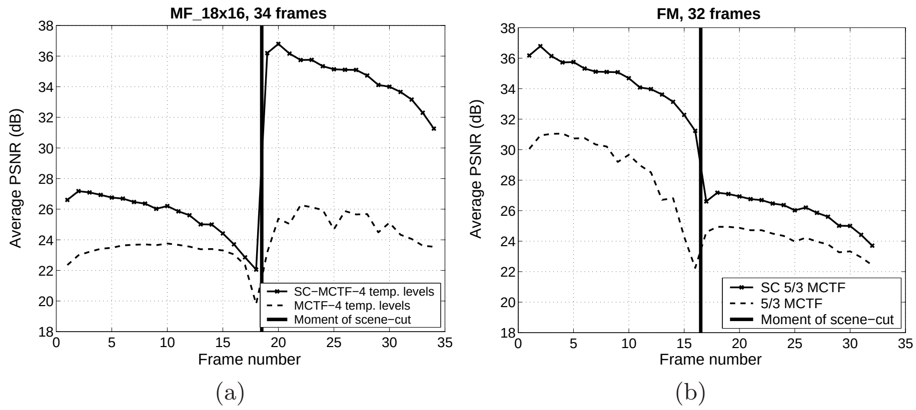


Fig. 5. PSNR for the "MF_18x16" (a) and "FM_16x16" (b) sequences, with and without scene-cut processing. Scene-change after the 18^{th} and respectively, 16^{th} frame.

4 Conclusion and Future Work

In this paper, we proposed an improved version of the 5/3 MCTF coding scheme, able to detect and process the scene-cuts appearing in video sequences. The lifting structure of the filters has been modified such that the filtering does not encompass the scene-cut. Moreover, the coding units were reduced to accommodate this change. As can be observed from the experimental results, our method gives an average YSNR gain of about 1.5 dB on the tested video sequences and higher for frames close to the scene-cut.

The presented method supposes the scene-change detection algorithm to be applied before starting the encoding process. In future work, we will focus on improvements allowing to process in one pass the video sequence, by performing the scene-cut detection during the encoding process.

References

1. B.-J. Kim, Z. Xiong, and W.A. Pearlman, "Very low bit-rate embedded video coding with 3-D set partitioning in hierarchical trees (3D-SPIHT)," *IEEE Trans on Circ. and Syst. for Video Tech.*, vol. 8, pp. 1365–1374, 2000.
2. S. Hsiang and J. Woods, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *ISCAS*, Geneva, Switzerland, 2000, pp. 589–595.
3. B. Shahraray, "Scene change detection and content-based sampling of video sequences," *SPIE*, vol. 2419, 1995.
4. R. L. De Queiroz, G. Bozdagi, and T. Sencar, "Fast video segmentation using encoding cost data," Tech. Rep., Xerox Corporation, 1999.
5. Y. Zhan, M. Picard, B. Pesquet-Popescu, and H. Heijmans, "Long temporal filters in lifting schemes for scalable video coding," doc. m8680, Klagenfurt MPEG meeting, July 2002.
6. J.-R. Ohm, "Complexity and delay analysis of MCTF interframe wavelet structures," doc. m8520, Klagenfurt MPEG meeting, July 2002.
7. D. Turaga and M. van der Schaar, "Unconstrained temporal scalability with multiple reference and bi-directional motion compensated temporal filtering," doc. m8388, Fairfax MPEG meeting, 2002.
8. Y. Wu and J.Woods, "MC-EZBC Video Proposal from RPI," Tech. Rep. MPEG04/M10569/S15, ISO/IEC JTC1/SC29/WG11, 2004.
9. P. Chen and J.Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans.on CSVT*, vol. 14, pp. 1183–1194, 2004.
10. C. Tillier, B. Pesquet-Popescu, and M. Van der Schaar, "Weighted average spatio-temporal update operator for subband video coding," *ICIP*, Singapore,Oct. 2004.
11. Y. Tsaig, *Automatic Segmentation of Moving Objects in Video Sequences*, Ph.D. thesis, Department of Computer Science, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel.
12. F. Porikli and Y. Wang, "Automatic video object segmentation using volume growing and hierarchical clustering," Technical Report TR2004-012, MERL-Mitsubishi Electric Research Laboratory, 2004.
13. J. Konrad and M. Ristivojevic, "Video segmentation and occlusion detection over multiple frames," *SPIE VCIP*, San Jose, 2003.
14. S.-C. Chen and M.-L. Shyu, "Video scene change detection method using unsupervised segmentation and object tracking," *ICME*, Tokyo, 2001.
15. B.T. Truong, C. Dorai, and S Venkatesh, "Improved fade and dissolve detection for reliable video segmentation," *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, pp. 961–964, 2000.