WAVELET-BASED VIDEO CODING: OPTIMAL USE OF MOTION INFORMATION FOR THE DECODING OF SPATIALLY SCALED VIDEO SEQUENCES

Davide Maestroni , Augusto Sarti, Marco Tagliasacchi, Stefano Tubaro

Dipartimento di Elettronica e Informazione, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milan, Italy (Europe)
phone: +39 02 2399 3647, fax: +39 02 2399 3413, email: maestron/sarti/tagliasa/tubaro@elet.polimi.it

ABSTRACT

In this paper we discuss the how to best handle motion vectors in spatially scalable wavelet-based video decoders. Motion vectors with full resolution are normally included in the bit-steams relative to spatially scaled version of a video sequence. When a low-resolution version of the original sequence is received, the decoder must scale the motion vectors accordingly. We will show that the motion vector scaling (truncation) is not the best solution and that better results can be obtained by interpolating the subsampled sequence to full resolution using of the wavelet synthesis low-pass filter. We illustrate the results of experiments carried out with an in-band wavelet-based fully scalable coder that performs spatial analysis, followed by temporal filtering. Emphasis is given to the computation of the Overcomplete DWT in the spatially scalable scenario.

1. INTRODUCTION

Nowadays video streaming is, in fact, ubiquitous, as more and more devices are able to render image sequences. As a consequence, it is no longer possible to produce multiple encoded representations of the signal in order to adapt to the decoder's characteristics. In fact, there is an ever increasing requirement of sending an encoded representation that is adapted to the device and network characteristics, in such a way that the coding is performed only once while decoding may take place several times at a different resolution, frame rate and quality. We refer to these requirements in terms of spatial, temporal and rate scalability, respectively. Wavelet-based video coders-decoders are able to fulfill such scalability requirements while achieving a good level of performance when they work at full spatio-temporal resolution. We can identify two families of wavelet-based video coders: SD-MCTF (Spatial-Domain Motion-Compensated Temporal Filtering) [1] and IB-MCTF (In-band Motion-Compensated Temporal Filtering) [2]. The former applies temporal filtering first along motion trajectories in order to reduce temporal redundancy. A simple Haar filter is often employed in this phase, although longer filters such as 5/3 filters have been recently shown to enable better compression. The output of temporal analysis is then 2D filtered in the spatial domain to reduce spatial redundancy. Wavelet coefficients are then entropy-coded using any of the wavelet-based still image compression algorithms (JPEG2000, SPIHT, EZBC). In literature, MCTF-EZBC is the state-of-the-art scalable video coder that implements a SD-MCTF scheme with EZBC coding of the wavelet coefficients. This is the reference implementation within the MPEG Ad-Hoc group on Scalable Video Coding.
In-Band MCTF swaps temporal and spatial analysis in such a way that the motion estimation/compensation phase is carried out in the wavelet domain. Because of the shift-variance of the critically sampled DWT, the motion-compensated temporal filtering takes place in the Overcomplete DWT (ODWT) domain. The ODWT is a redundant subband decomposition of the input signal that removes the subsampling operations, thus achieving shift invariance. On the positive side, the IB-MCTF approach gets rid of blocking artifacts even when a block-based motion compensation algorithm is employed. On the other hand, this approach is computationally quite demanding and memory consuming due to the ODWT computation.
Both SD-MCTF and IB-MCTF coders decompose each group of pictures (GOP) in a plurality of spatio-temporal subbands. Scalability is achieved by pulling from the encoded bitstream only those subbands that represent the sequence at the desired frame rate and resolution up to a given quality (i.e. quantization) level.
The rest of this paper is organized as follows: Section 2 illustrates the motion vector handling and addresses spatial scalability. Section 3 extends what shown in Section 2 to the IB-MCTF scenario. Section 4 shows some experimental results.

2. MOTION VECTOR HANDLING FOR SPATIALLY SCALABLE WAVELET-BASED VIDEO CODERS

Wavelet-based video coders address spatial scalability in a straightforward way. At the end of spatio-temporal analysis each frame k of a GOP of size K represents a temporal subband further decomposed into spatial subbands up to level L as illustrated in Figure 1. Each frame thus consists of the following subbands: $LL_k^{(L)}$, $LH_k^{(l)}$, $HL_k^{(l)}$, $HH_k^{(l)}$ with

Work developed within the FIRB-VICOM project (www.vicom-project.it), funded by the Italian Ministry of University and Scientific Research (MIUR); and within the VISNET project, a European Network of Excellence (www.visnet-noe.org)

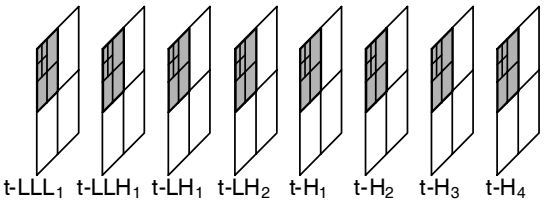


Figure 1 - When a sequence is scaled at half its resolution (s = 1) only the greyed spatial subbands are sent for each GOP

l=1…L, k = 1…K. Let us assume that we want to send and decode a sequence whose resolution is 2^s times lower than the original one. For example, if *s* is equal to one a CIF resolution sequence would be decoded at QCIF resolution. We need to send only those subbands $LL_k^{(L)}$, $LH_k^{(l)}$, $HL_k^{(l)}$, $HH_k^{(l)}$ with *l=s+1…L* if *s* < *L*. Else, if *s* = *L* only $LL_k^{(L)}$ is sent. At the decoder side, spatial decomposition and motion-compensated temporal filtering is inverted in the synthesis phase. The problem that must be addressed is that the motion vector field, available at the decoder side, normally has a full resolution, while the received subbands represents a lower resolution version of the sequence. In the rest of this paper we assume for simplicity of exposition that the motion field is represented by motion vectors having integer components. We want to compare from a theoretical point of view the following approaches:

(a) the motion vectors are truncated and rounded in order to match the received sequence resolution;
(b) the original motion vectors are retained, while a full resolution sequence is interpolated starting from the received subbands.

In the implementation available to us MCTF-EZBC adopts the former approach. It is computationally simpler while not as efficient as the latter in terms of reconstruction quality.

In order to state things formally let us concentrate our attention on a one-dimensional discrete signal *x(n)* and its translated version by an integer displacement *d*, i.e. *y(n) = x(n – d)*. Their 2D counterpart are the reference and the current frame respectively. This way we are are neglecting motion compensation errors due to complex motion, reflections and illumination changes. Temporal analysis is carried out with a "lift" implementation of the Haar transform along the motion trajectory *d*:

$$H(n) = \frac{1}{\sqrt{2}}\left[y(n) - x(n-d)\right] = 0$$

$$L(n) = \frac{1}{\sqrt{2}}\left[x(n) - H(n+d)\right] = \frac{1}{\sqrt{2}}x(n)$$

L(n) and *H(n)* are wavelet transformed and, in the case of spatial scalability, only a subset of their subbands are sent. If we scale at half the original resolution, the decoder receives the following signals:

$$H_L(n) = 0; \quad L_L(n) = \frac{x_L(n)}{\sqrt{2}} = \frac{1}{\sqrt{2}}\left[x(k) * h(k)\right]_{k=2n}$$

Temporal synthesis reconstructs a low resolution approximation of the original signals:

$$\hat{x}_L(n) = \sqrt{2}L_L(n) - H_L\left(n + \frac{d}{2}\right) = x_L(n)$$

$$\hat{y}_L(n) = \sqrt{2}H_L(n) - \hat{x}_L\left(n - \frac{d}{2}\right) = x_L\left(n - \frac{d}{2}\right)$$

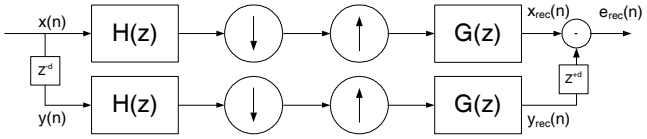


Figure 2 – Reconstruction error computation without motion vector truncation (Scenario (b))

For reasons better explained in Section 4, we compute the reconstruction error using the spatially low-pass filtered and subsampled version of the original frames as reference:

$$e_1(n) = \hat{x}_L(n) - x_L(n) = 0$$

$$e_2(n) = \hat{y}_L(n) - y_L(n) = x_L\left(n - \frac{d}{2}\right) - y_L(n)$$

We derive the solution for scenario (b) first, since we will see (a) as a particular case. The decoder reconstructs an interpolated version of the original sequence. This is accomplished by setting to zero the coefficients of the missing subbands before performing the wavelet synthesis. It is worth pointing out that this is equivalent to estimating the missing samples using the wavelet scaling function as interpolating kernel. The energy of the reconstruction error turns out to be:

$$\sum_{n=0}^{N-1} e_{rec}^2(n) = \sum_{n=0}^{N-1} \left| x_{rec}(n-d) - y_{rec}(n) \right|^2 \tag{1}$$

Figure 2 illustrates how $e_{rec}(n)$ is computed starting from *x(n)* and *y(n)*. *H(z)* represents the analysis wavelet low-pass filter, while *G(z)* is the synthesis low-pass filter. In the rest of this paper we assume that they are Daubechies 9/7 biorthogonal filters. As they are nearly orthogonal, the equation (1) is satisfied. The reconstructed signal $x_{rec}(n)$ is an approximation of *x(n)* having the same number of samples. Therefore motion compensation can use the original motion vector *d*. In the Fourier domain we write:

$$X_{rec}(\omega) = \frac{1}{2}G(\omega)\left[X(\omega)H(\omega) + X(\omega+\pi)H(\omega+\pi)\right] \tag{2}$$

Equivalent expressions can be written for $y_{rec}(n)$. By Parseval's theorem the prediction error in equation (1) becomes (scenario b is considered):

$$\sum_{n=0}^{N-1} e_{rec(b)}^2(n) = \int_{-\pi}^{+\pi} \left| X_{rec}(\omega)e^{-j\omega d} - Y_{rec}(\omega) \right|^2 d\omega \tag{3}$$

By substituting (2) in (3) and recalling that:

$$y(n) = x(n-d) \Rightarrow Y(\omega) = X(\omega)e^{-j\omega d}$$

We obtain:

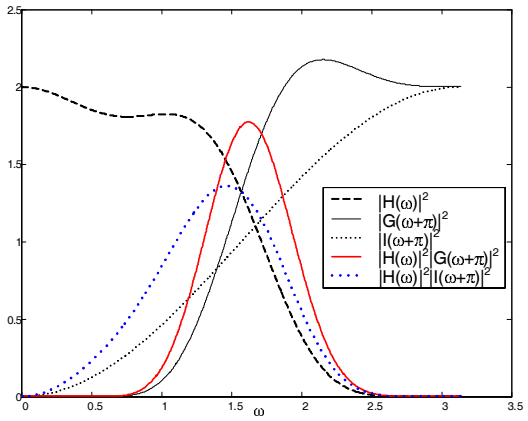
$$\sum_{n=0}^{N-1} e_{rec(b)}^2(n) = \frac{1}{4}\left|1 - e^{-j\pi d}\right|^2 \int_{-\pi}^{+\pi} |G(\omega)|^2 |X(\omega+\pi)|^2 |H(\omega+\pi)|^2 d\omega =$$

$$= \frac{1}{2}\left|1 - e^{-j\pi d}\right|^2 \int_{0}^{+\pi} |G(\omega+\pi)|^2 |X(\omega)|^2 |H(\omega)|^2 d\omega$$

If we constrain the displacement to be integer, the previous expression turns out to be zero if *d* is even. Conversely, if *d* is odd, the expression of the error becomes:

$$\sum_{n=0}^{N-1} e_{rec(b)}^2(n) = 2\int_{0}^{+\pi} |G(\omega+\pi)|^2 |X(\omega)|^2 |H(\omega)|^2 d\omega \tag{4}$$

Figure 3 depicts the energy spectrum of $G(\omega+\pi)$ and $H(\omega)$ together with their product. We can conclude that the error energy depends on the frequency characteristics of the signal and it is close to zero if its energy is mostly



concentrated at low frequencies. In fact in this case the approximation we get interpolating with $G(\omega)$ is very much similar to the original. On the other hand equation (4) suggests that the error is zero also when the energy is concentrated at high frequencies.

Figure 3 – Frequency responses of the filters cited in the paper

This counterintuitive result can be explained as follows: when an high frequency signal passes through the system in Figure 2 it is almost cancelled by the low-pass analysis filter $H(\omega)$. The error turns out to be zero because both $x_{rec}(n)$ and $y_{rec}(n)$ have little residual energy.

The error in scenario (a) can be derived as a special case of scenario (b). Since the received signal has lower resolution than the motion field, vectors are truncated (scaled and rounded). The reconstruction error turns out to be:

$$\sum_{n=0}^{N/2-1} e_2^2(n) = \sum_{n=0}^{N/2-1} \left| x_L\left(n - round\left[\frac{d}{2}\right]\right) - y_L(n) \right|^2$$

In order to find a frequency domain expression for this scenario we can observe that the operation of truncating and rounding motion vectors is equivalent to interpolating the low resolution version received by the decoder with a sample&hold filter and then applying the full resolution motion field. As a matter of fact the error turns out to be:

$$\sum_{n=0}^{N-1} e_{rec(a)}^2(n) = \int_0^{+\pi} 2\left|\frac{e^{j\omega}}{\sqrt{2}} - \frac{1}{\sqrt{2}}\right|^2 |X(\omega)|^2 |H(\omega)|^2 d\omega =$$

$$= \int_0^{+\pi} 2|I(\omega+\pi)|^2 |X(\omega)|^2 |H(\omega)|^2 d\omega$$

Having fixed $|H(\omega)|^2$, we are not able to state that the following inequality holds for any signal:

$$\sum_{n=0}^{N-1} e_{rec(a)}^2(n) \geq \sum_{n=0}^{N-1} e_{rec(b)}^2(n) \qquad (5)$$

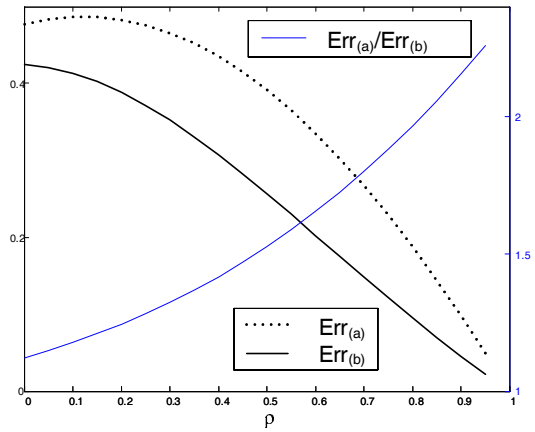
Since:

$$|I(\omega)|^2 \not\geq |G(\omega)|^2 \text{ for all } \omega$$

Nevertheless, if we assume that most of the energy is concentrated at low frequencies, inequality (5) holds. In order to enforce this intuition let us take the expectation on both sides of (5) with respect to $x(n)$.

$$Err_{(a)} = E\left\{\sum_{n=0}^{N-1} e_{rec(a)}^2(n)\right\} = \int_0^{+\pi} 2|I(\omega+\pi)|^2 |S_x(\omega)|^2 |H(\omega)|^2 d\omega$$

$$Err_{(b)} = E\left\{\sum_{n=0}^{N-1} e_{rec(b)}^2(n)\right\} = \int_0^{+\pi} 2|G(\omega+\pi)|^2 |S_x(\omega)|^2 |H(\omega)|^2 d\omega$$



If we model the signal as a wide-sense stationary noise with correlation coefficient ρ the signal power spectrum is:

$$S_x(\omega) = \frac{1-\rho^2}{\left|1-\rho e^{j\omega}\right|^2}$$

Figure 4 – Comparison between the (normalized) error in scenario (a) and (b) with WSS input and correlation coeff. ρ

As illustrated in Figure 4 for any ρ in [0,1] $Err_{(a)} > Err_{(b)}$ and their ratio is higher for ρ close to 1, meaning that the penalty due to motion vector truncation with respect to interpolating at full resolution with $G(\omega)$ is greater when the input signal has energy concentrated in the low frequency range.

3. SPATIALLY SCALABLE IN-BAND MCTF

In-Band Motion Compensated Temporal Filtering (IB-MCTF) represents a valid alternative to conventional SD-MCTF since the reconstructed sequence does not suffer from blocking artifacts at low bitrates even when a block-matching algorithm is employed for motion compensation. On the other hand IB-MCTF is computationally and memory demanding since motion estimation-compensation takes place in the ODWT domain. Figure shows a block diagram of a system implementing the ODWT using the *algorithm à trous* [3]. Due to the absence of decimators the ODWT is shift invariant. Note that $h_0(n)$ represents the low-pass analysis filter (it is the same as $h(n)$ in our previous discussion), while $h_1(n)$ the high-pass analysis filter. The concepts stated in the previous sections have been deduced in the SD-MCTF scenario. Nevertheless they holds true in the IB-MCTF case, the only significant difference being the computation of the Overcomplete DWT of the reference frame. The received subbands are used to reconstruct a full resolution version of the signal $x_{rec}(n)$. Since this is a low-pass approximation of $x(n)$ the output of the high pass filter $h_1(n)$ of the critically sampled DWT is zero. On the other hand, contrary to intuition, the correspondent sub-band of the ODWT is not zero. More specifically, it is always equal to zero only in those locations corresponding to the critically sampled coefficients (indeed, by critically sampling the ODWT we end up with the DWT). For this reason the computation, at the decoder side, of the ODWT cannot be stopped prematurely, i.e. by setting to zero all the coefficients of the missing subbands, otherwise a sensible drop in reconstruction quality would be observed.

Experimental results demonstrated a drop up to 1dB in the PSNR if the ODWT is not computed in a complete way.

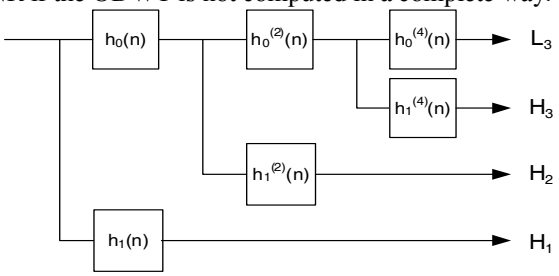


Figure 5 - Overcomplete DWT (ODWT) computed through the algorithm à trous. $h^{(k)}(n)$ is the dilated version of h(n) obtained inserting k-1 zeros between two consecutive samples.

4. EXPERIMENTAL RESULTS

We carried out several experiments in order to put in practice the principles stated in the previous sections. In order to assess the objective quality of the reconstructed sequence at reduced spatial resolution we used the low-pass filtered and subsampled version of the original sequence as a reference. We have chosen $H_0(\omega)$ as anti-aliasing filter. This approach differs from the one adopted when assessing temporal scalability [4], where the references used are the unquantized temporal low frequency frames output of the first level of the MCTF pyramid. On the other hand, as far as spatial scalability is concerned the output of MCTF is not a good reference, especially when working at full pixel motion accuracy. When using it as a reference objective results (PSNR) does not always match subjective evaluation criteria. For this reasons we argue that the low-pass subsampled frames can be deemed to be a better reference. Figure 6 shows the average Y PSNR of the *Mobile&Calendar* sequence spatially scaled from CIF to QCIF resolution. We set motion accuracy to full pixel and full reconstruction frame rate in our experiments. We compared three different coders:

- IB-MCTF-1: our implementation of an in-band coder with motion vector truncation - scenario (a)
- IB-MCTF-2: same as IB-MCTF-1 but with full resolution motion vectors and setting to zero all ODWT coefficients of the missing subbands - scenario (b)
- IB-MCTF-3: same as IB-MCTF-2 but computing ODWT coefficients in a complete way -scenario (b)

Figure 7 shows an example taken from the reconstructed sequence. IB-MCTF-3 turns out to yield the best objectice and subjective results for all test sequences. The reason why IB-MCTF-1 does not grow above 20dB even at high bitrates is that in this case the lossless reconstruction differs from the reference due to the non invertibility of the MCTF phase, which is caused by motion vector truncation. We also tested the implementation of MCTF-EZBC available to us. As expected it yielded results similar to the IB-MCTF-1 case.

Although scenario (b) turns out to be the best choice as far as reconstruction quality is concerned, it is more computationally demanding. In our experiments we observed that decoding is between 2.1 and 2.3 times slower than scenario (a) for CIF sequences scaled down to QCIF resolution. In addition to this it is more memory demanding, especially in the IB-MCTF case, since the decoder works on a full resolution version of the spatially scaled sequence.

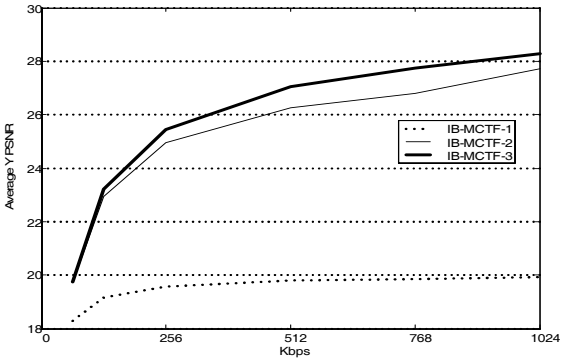


Figure 6 - Mobile&Calendar scaled at QCIF@30fps

Figure 7 - Mobile&Calendar (frame 33) scaled at QCIF resolution at 256Kbps. Left IB-MCTF-1. Right IB-MCTF-2

We have also carried out experiments using fractional resolution of the motion vectors. The results are very similar to those shown in Figure 6. The gap between IB-MCTF-2 and IB-MCTF-3 remains about always the same, while the gain of IB-MCTF-2 whit respect to IB-MCTF-1 reduces when the resolution of the motion vectors increases (about 2 dB at 512 kbps and a m.v. resolution of a quarter of a pixel).

5. CONCLUSIONS

In this paper we theoretically proved that truncating the motion vectors gives a poorly reconstructed sequence when it is decoded at lower spatial resolution. Better results can be obtained by setting to zero the missing subband and interpolating at full resolution before performing temporal synthesis. Future works will address an adaptive optimal interpolation of the low resolution subbands with filters other than $G(\omega)$ in order to achieve better performance in the spatial scalability scenario.

REFERENCES

[1] P. Chen, "Fully Scalable Subband/Wavelet Coding". Ph.D. thesis – Rensselaer Polytechnic Institute – Troy, NY, May 2003

[2] J.C. Ye, M. van der Schaar, "Fully Scalable Overcomplete Wavelet Video Coding using Adaptive Motion Compensated Temporal Filtering". In Proceedings of VCIP2003, July 2003, Lugano, Switzerland

[3] S. Mallat, "A Wavelet Tour of Digital Signal Processing", Academic Press, 1998

[4] J. W. Woods, P. Chen, "Cross-Check Method for Objective Comparison of Scalable Video Coders", ISO/IEC JTC1/SC29/WG11, MPEG2003/M9585, Pattaya, TH, March 2003