THE EFFECT OF MOTION VECTOR TRUNCATION ON SPATIALLY SCALABLE WAVELET BASED VIDEO CODERS

Marco Tagliasacchi, Stefano Tubaro, Augusto Sarti, Davide Maestroni

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: tagliasa/tubaro/sarti@elet.polimi.it, davide.mae@tiscali.it

ABSTRACT

In this paper we describe how to properly handle motion vectors in spatially scalable wavelet based video coders. When a lower resolution version of the original sequence is received, the decoder has to scale the motion vectors accordingly. We demonstrate that motion vector truncation is not the right choice and that better results can be obtained by interpolating the subsampled sequence to full resolution by means 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 ubiquitous as more and more devices are able to reproduce image sequences. There exists the compelling requirement of sending an encoded representation of a sequence that is adapted to the device and network characteristics in such a way that coding is performed only once while decoding takes place several times at a different resolution, frame rate and quality. We refer to these requirements as spatial, temporal and rate scalability respectively. Wavelet based video coders are able to fulfill scalability yet retaining good performances when decoding at full spatio-temporal resolution. We 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, even though more recently longer filters such as 5/3 filters has been shown to provide 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 wavelet coefficients. It 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 motion estimation-compensation is carried out in the wavelet domain. Because of the shift variance of the critically sampled DWT, 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. As a positive aspect, the IB-MCTF approach get rid of blocking artefacts even when a block based motion compensation algorithm is employed. On the other hand, it is computationally 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 details motion vectors handling to address spatial scalability. Section 3 extends what stated in Section 2 to the IB-MCTF scenario. Section 4 illustrates experimental results. Section 5 concludes this paper.

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 $l=1,...,L$, $k=1,...,K$. 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
we scale at half the original resolution, the decoder receives spatial scalability, only a subset of their subbands are sent. If an approximation of the original signals:

Temporal synthesis reconstructs a low resolution trajectory. 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 in adopts from a theoretical point of view the following approaches:

- The first equivalence holds as far as we use an orthogonal transform to reconstruct a full resolution approximation of the signals. 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:

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

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

$$\sum_{n=0}^{N-1} e_{rec}^2(n) = \int_{-\pi}^{\pi} |X_{rec}(\omega)|^2 \, d\omega - \int_{-\pi}^{\pi} [Y(\omega) - \hat{Y}(\omega)]^2 \, d\omega$$

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

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

We obtain:

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

If we constrain the displacement to be integer, the previous expression turns out to be zero if $d$ is even. Lower resolution version of the original signal. Conversely, if $d$ is odd, the expression of the error becomes:

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

Figure 3 depicts the power spectrum of $G(\omega + \pi)$ and $H(\omega)$ together with their product.
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. If the components are odd they are also rounded to the nearest integer. The reconstruction error turns out to be:

$$\sum_{n=0}^{N-1} e_2^2(n) = \sum_{n=0}^{N-1} \left| y_{rec}(n) - y_1(n) \right|^2$$

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:

$$\sum_{n=0}^{N-1} e_2^2(n) = \sum_{n=0}^{N-1} e_{rec}^2(n) = \int |x(\omega)|^2 H(\omega)^2 d\omega = \int [1 - \rho^2] |x(\omega)|^2 H(\omega)^2 d\omega$$

The equivalence holds because the interpolating filter \( H(\omega) \) can be seen as an inverse Haar DWT (that is an orthogonal transform) where the observation frequency subband is the subsampled signal and the high frequency subband coefficients are set to zero. Having fixed \( |H(\omega)|^2 \), we are not able to state that the following inequality holds for any signal $x(n)$:

$$\sum_{n=0}^{N-1} e_{rec}^2(n) \geq \sum_{n=0}^{N-1} e_2^2(n)$$

Since:

$$\|h_1(\omega)\|^2 \leq |\hat{g}(\omega)|^2$$

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 $x(n)$.

$$\text{Err}_{r_1} = E\left[ \sum_{n=0}^{N-1} e_{rec}^2(n) \right] = \int 2 |1 - \rho^2| S_x(\omega) |H(\omega)|^2 d\omega$$

$$\text{Err}_{r_3} = E\left[ \sum_{n=0}^{N-1} e_2^2(n) \right] = \int 2 |1 - \rho^2| S_x(\omega) |H(\omega)|^2 d\omega$$

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

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

As illustrated in Figure 4 for any $\rho$ in [0,1] $\text{Err}_{r_1} > \text{Err}_{r_3}$ and their ratio is higher for $\rho$ 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 artefacts 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 5 shows a block diagram of a system implementing the ODWT using the algorithm à trous [3]. Due to the lack 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 output of the $h_1(n)$ filter 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 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 computation is not performed at full resolution.
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 $H(\omega)$ 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. It represents a reasonable reference indeed, since MCTF attenuates temporal aliasing, hence producing a better sequence from a subjective standpoint. 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 setting to zero only those ODWT coefficients corresponding to critically sampled locations of the missing subbands - scenario (b)

Figure 7 shows an example taken from the reconstructed sequence. IB-MCTF-3 turns out to yield the best objective 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.

5. CONCLUSIONS

In this paper we proved from a theoretical point of view 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 get better performances in the spatial scalability scenario.

REFERENCES