A RESOLUTION ADAPTIVE INTERPOLATION TECHNIQUE FOR ENHANCED DECODING OF SCALABLE CODED VIDEO

Marta Mrak, Nikola Sprljan, Ebroul Izquierdo

Multimedia and Vision Lab, Dept. of Electronic Engineering
Queen Mary, University of London
Mile End Road, E1 4NS, London, UK
{marta.mrak, nikola.sprljan, ebroul.izquierdo}@elec.qmul.ac.uk

ABSTRACT

Subpixel accurate motion compensated temporal filtering introduces a significant coding gain in scalable 3D wavelet video codecs. The influence of the chosen subpixel interpolation technique has not yet been fully analysed in the context of resolution scalability. That problem is addressed in this paper. It is shown that support for increased accuracy and resolution adaptive spatial interpolation needs to be featured in a scalable video decoder, when low resolution sequences are targeted. Using the proposed resolution adaptive filters based on sinc kernels leads to improved decoding performance at low resolution in the sense of achieving higher quality while reducing the complexity of the system.

1. INTRODUCTION

With the expansion of video applications, the need for video coding enabling seamless delivery for various displaying platforms is becoming acute. Intensive research activities on diverse algorithms for scalable video coding have been undertaken in the past. Currently this technology is becoming mature and reaching the phase of wide commercial exploitation. Besides popular hybrid DCT-like based codecs [1], efficient 3D wavelet codecs that produce embedded bitstreams have been developed [2-5]. These codecs use motion compensated temporal filtering (MCTF) to remove temporal redundancies [6].

Following the traditional coding approach, in which the temporal transform precedes the spatial transform, spatial domain MCTF (SD MCTF) 3D wavelet video codecs produce embedded video streams featuring spatial, temporal and quality scalability. Unfortunately, these codecs suffer from artefacts known as "MCTF drift", when resolution scalability is selected [7]. This drawback appears when high spatial subbands are discarded by the extractor, because low spatial subbands do not carry the same information as low resolution sequence obtained directly from the original video. Due to the order in which the temporal and spatial transforms are applied and due to the inaccuracies of the motion estimation, estimation errors and artefacts are transferred from low resolution to higher temporal frame subbands and vice-versa. While significant R-D performance saturation is caused by this phenomenon, e.g., convergence towards PSNR below 40 dB, it is not visually disturbing [8]. This holds for a case when the reference sequence used for distortion estimation is the one directly extracted from the original sequence by downsampling.

So far, methods for improved low-bitrate, low-resolution performance have been proposed [9], [10]. In this paper we consider low-resolution, yet high quality sequence improvement that is achieved by higher motion precision decoding and improved spatial interpolation for MCTF synthesis.

As a significant contribution to temporal decorrelation comes from subpixel accurate motion compensation, interpolation is an important tool in video coding. However, the designs of interpolation filters used in scalable video coding [9], [11], anticipate application of the same interpolation filters for each decoding resolution.

The prediction and update steps for MCTF synthesis on lower resolutions, as formulated in the following section, determine how the motion parameters have to be modelled for lower resolution decoding. Further analysis of spatial dependencies of temporal frame pixels over various resolutions defines the interpolation filter design for lower decoding resolution. Finally, the test results show possible improvements of both visual quality and PSNR when the proposed interpolation scheme is integrated in a scalable video decoder.

2. TEMPORAL FILTERING AND LOWER RESOLUTION DECODING

In a SD MCTF scenario, MCTF is performed on the highest resolution frames. For temporal filtering a wavelet transform is used and therefore it can be factored into lifting steps. Invertible sub-pixel accurate lifting has been proposed in [11]. Following their interpretation, using Haar wavelets, temporal high (H) and low (L) pass frames are obtained from original or lower temporal frames A and B according to (1) and (2):

$$H(m_0,n_0)=\frac{1}{\sqrt{2}}\cdot\left(B(m_0,n_0)-\tilde{A}\left(m_0-d_{m_0},n_0-d_{n_0}\right)\right) \tag{1}$$

$$\begin{aligned}L\left(m_0-\bar{d}_{m_0},n_0-\bar{d}_{n_0}\right)=\tilde{H}\left(m_0-\bar{d}_{m_0}+d_{m_0},n_0-\bar{d}_{n_0}+d_{n_0}\right)\\ +\sqrt{2}\cdot A\left(m_0-\bar{d}_{m_0},n_0-\bar{d}_{n_0}\right)\end{aligned} \tag{2}$$

where (d_{m_0}, d_{n_0}) and ($\bar{d}_{m_0}$, $\bar{d}_{n_0}$) denote full resolution motion vector components and their integer parts, respectively. $\tilde{X}$ stands for interpolated frame X and index 0 denotes full resolution $r=0$.

This research was partially supported by the European Commission under contract FP6-001765 aceMedia.

The original sequence synthesis can then be realised using the following equations:

$$A\left(m_0-\bar{d}_{m_0},n_0-\bar{d}_{n_0}\right)=\frac{1}{\sqrt{2}}\cdot\left(L\left(m_0-\bar{d}_{m_0},n_0-\bar{d}_{n_0}\right)-\tilde{H}\left(m_0-\bar{d}_{m_0}+d_{m_0},n_0-\bar{d}_{n_0}+d_{n_0}\right)\right) \quad (3)$$

$$B\left(m_0,n_0\right)=\sqrt{2}\cdot H\left(m_0,n_0\right)+\tilde{A}\left(m_0-d_{m_0},n_0-d_{n_0}\right) \quad (4)$$

Using an invertible subpixel accurate approach, coding efficiency depends on the spatial interpolation. It has been shown that the optimal interpolation can be achieved using separable sinc interpolation [11].

An important property of 3D wavelet codecs is that the bitstream offers embedded resolution levels. If the encoder performs R 2D decomposition levels, spatial subbands (LL_R, (LH_R, HL_R, HH_R),..., (LH_0, HL_0, HH_0)) are embedded in the bitstream. From a scalable encoded bitstream, lower resolution sequences can be extracted directly avoiding complex and in many cases unfeasible transcoding operations.

When an application requires a sequence of resolution $r > 0$, the scalable video extractor discards high-pass subbands of temporal frames. The reconstruction will be carried out from L_r and H_r frames whose dimensions are 2^r times smaller than the original frames. From these observations, a general form of the temporal synthesis equations can be derived that is useful for lower resolution sequence decoding:

$$A_r\left(m_r-\overline{\left(\frac{\bar{d}_{m_0}}{2^r}\right)},n_r-\overline{\left(\frac{\bar{d}_{n_0}}{2^r}\right)}\right)=\frac{1}{\sqrt{2}}\cdot\left(L_r\left(m_r-\overline{\left(\frac{\bar{d}_{m_0}}{2^r}\right)},n_r-\overline{\left(\frac{\bar{d}_{n_0}}{2^r}\right)}\right)-\tilde{H}_r\left(m_r-\overline{\left(\frac{\bar{d}_{m_0}}{2^r}\right)}+\frac{d_{m_0}}{2^r},n_r-\overline{\left(\frac{\bar{d}_{n_0}}{2^r}\right)}+\frac{d_{n_0}}{2^r}\right)\right) \quad (5)$$

$$B_r\left(m_r,n_r\right)=\sqrt{2}\cdot H_r\left(m_r,n_r\right)+\tilde{A}_r\left(m_r-\frac{d_{m_0}}{2^r},n_r-\frac{d_{n_0}}{2^r}\right) \quad (6)$$

From (5) and (6) it can be seen that if the original motion information has to be preserved, the original subpixel accuracy $1/M$ increases for $1/2^r$. However, scalable video decoders often neglect this fact, as they do not support subpixel interpolation under $1/M$. If R is the lowest resolution that can be extracted in an SD MCTF codec, the decoding should be supported to an accuracy of up to $1/(M\cdot 2^R)$.

In the next section, the interpolation required for the temporal synthesis as defined by (5) and (6) is analysed.

3. SPATIAL INTERPOLATION AND INVERSE MCTF

Subpixel accurate motion compensation in scalable video codec implies the application of spatial interpolation. Even if the coder performs pixel accurate motion compensation, lower resolution decoding will need interpolation since originally integer motion vectors may point to subpixel positions.

Among various interpolation methods, sinc interpolation enables optimal MCTF decomposition. Theoretically, it also introduces the smallest distortion error into the signal defined by its samples.

Sinc interpolation can be performed as a separable filtering process, so that frame X can first be interpolated in a horizontal direction, and then the resulting frame $\tilde{X}^{\eta}$ in vertical direction. 1-D interpolation takes into account contributions of the nearest samples and these are proportional to $\sin(t)/t$, where t is the distance between the interpolated sample and the original sample.

Because of computational cost and annoying visual artefacts (ringing effects), in image/video coding it is desirable to use windowed interpolation. Following filter design from [11], we use the Hamming window. Our interpolation filters $\mathbf{h}_p^T=\left(h_p^T\left(-T/2+1\right),\ldots,h_p^T\left(T/2\right)\right)$ can be of any number of taps T and mainly differ from the previous approaches in the way that support wide range of fractional position p of the interpolation samples.

We interpolate pixels of A_r and H_r frames as:

$$\tilde{X}_r^{\eta}\left(i,\lfloor j\rfloor\right)=\sum_{t=-T_r/2+1}^{T_r/2}\left(h_{i-\lfloor i\rfloor}^{T_r}\left(t\right)\cdot X_r\left(\lfloor i\rfloor+t,j\right)\right) \quad (7)$$

$$\tilde{X}_r\left(i,j\right)=\sum_{t=-T_r/2+1}^{T_r/2}\left(h_{j-\lfloor j\rfloor}^{T_r}\left(t\right)\cdot\tilde{X}_r^{\eta}\left(i,\lfloor j\rfloor+t\right)\right) \quad (8)$$

where i and j are subpixel coordinates whose range depends on the targeting resolution. At the encoder side $r = 0$, and at the decoder side $r \le R$. Such interpolation can be distinguished from earlier approaches since in our proposal the filtering process is resolution adaptive.

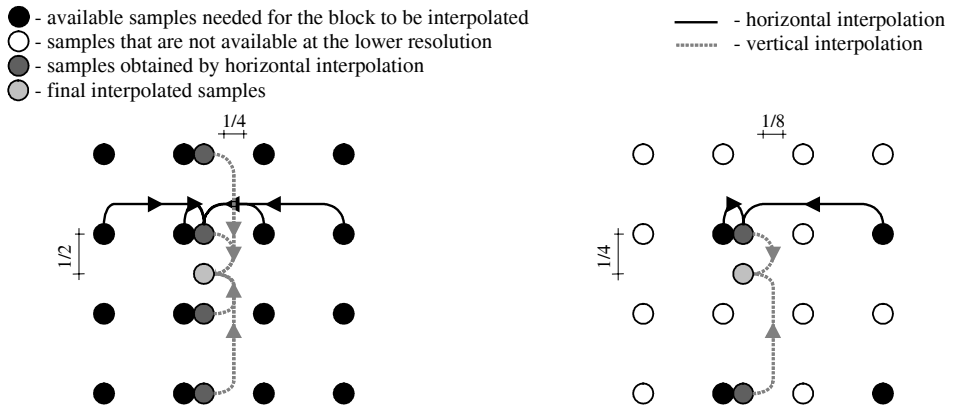
At the decoder, the interpolation has to capture the same features around the subpixel to be interpolated. Therefore, it may take fewer pixels in order to cover the same area. That is, for $r > 0$, the value of T_r corresponds to $T_0 / 2^r$ since only the pixels that correspond to the same spatial position should be used in the calculation of the interpolated pixels. This is illustrated in Figure 1 for $T_0 = 4$. As at low resolutions (Figure 1.b) the number of available pixels that cover the corresponding area of the frame is also low, the interpolation takes fewer samples into account. At the same time, as shown in (5) and (6), the subpixel accuracy becomes higher. Consequently, for each resolution and subpixel position used by the decoder, $\mathbf{h}_p^{T_r}$ must be estimated such that

$$p\in\left\{\frac{1}{M\cdot 2^r},\frac{2}{M\cdot 2^r},\ldots,\frac{M\cdot 2^r-1}{M\cdot 2^r}\right\}.$$

At the decoder side, only $\mathbf{h}_p^{T_0}$ for T_0 are needed.

Examples of 1-D $\mathbf{h}_p^{T_r}$ interpolation filters are given in Table 1 for accuracy up to 1/16, i.e., subpixel positions $p \in \{1/16, 2/16,...,15/16\}$ and a) $T = 8$ and b) $T = 4$ taps.

Considering issues of complexity, at the encoder side the computational cost of the interpolation depends on the used motion estimation process, which determines the number of subpixel values to be computed. At the decoder side only subpixels that contribute to the motion compensation process have to be computed. Consequently, the computational cost of interpolation decreases by a factor of 4 per pixel for each subsequent lower resolution level.



a) Interpolation of a pixel of fractional positions (1/4, 1/2) at the original resolution $r = 0$

b) Interpolation of a pixel of fractional positions (1/8, 1/4) at the original resolution $r = 1$

Figure 1. Resolution depended subpixel interpolation

Table 1. Coefficients $h_p^T(t)$ of interpolation filters (rows) for 1/16 subpixel accurate MCTF of taps a) 8 and b) 4.

a) $T = 8$

subpixel position (p)	neighbouring pixel position (t) -3	-2	-1	0	1	2	3	4
1/16	-0.0040	0.0156	-0.0497	0.9941	0.0584	-0.0181	0.0049	-0.0013
1/8	-0.0072	0.0284	-0.0902	0.9742	0.1249	-0.0381	0.0105	-0.0026
3/16	-0.0095	0.0383	-0.1215	0.9409	0.1985	-0.0594	0.0168	-0.0040
1/4	-0.0109	0.0452	-0.1437	0.8950	0.2777	-0.0812	0.0233	-0.0053
5/16	-0.0117	0.0492	-0.1571	0.8380	0.3609	-0.1026	0.0300	-0.0068
3/8	-0.0117	0.0505	-0.1624	0.7713	0.4465	-0.1223	0.0363	-0.0081
7/16	-0.0113	0.0495	-0.1605	0.6968	0.5323	-0.1393	0.0420	-0.0094
1/2	-0.0105	0.0465	-0.1525	0.6165	0.6165	-0.1525	0.0465	-0.0105
9/16	-0.0094	0.0420	-0.1393	0.5323	0.6968	-0.1605	0.0495	-0.0113
5/8	-0.0081	0.0363	-0.1223	0.4465	0.7713	-0.1624	0.0505	-0.0117
11/16	-0.0068	0.0300	-0.1026	0.3609	0.8380	-0.1571	0.0492	-0.0117
3/4	-0.0053	0.0233	-0.0812	0.2777	0.8950	-0.1437	0.0452	-0.0109
13/16	-0.0040	0.0168	-0.0594	0.1985	0.9409	-0.1215	0.0383	-0.0095
7/8	-0.0026	0.0105	-0.0381	0.1249	0.9742	-0.0902	0.0284	-0.0072
15/16	-0.0013	0.0049	-0.0181	0.0584	0.9941	-0.0497	0.0156	-0.0040

b) $T = 4$

subpixel position (p)	neighbouring pixel position (t) -1	0	1	2
1/16	-0.0290	0.9928	0.0388	-0.0026
1/8	-0.0488	0.9668	0.0878	-0.0058
3/16	-0.0605	0.9237	0.1465	-0.0097
1/4	-0.0653	0.8659	0.2142	-0.0147
5/16	-0.0648	0.7961	0.2896	-0.0209
3/8	-0.0603	0.7174	0.3712	-0.0283
7/16	-0.0534	0.6327	0.4571	-0.0365
1/2	-0.0451	0.5451	0.5451	-0.0451
9/16	-0.0365	0.4571	0.6327	-0.0534
5/8	-0.0283	0.3712	0.7174	-0.0603
11/16	-0.0209	0.2896	0.7961	-0.0648
3/4	-0.0147	0.2142	0.8659	-0.0653
13/16	-0.0097	0.1465	0.9237	-0.0605
7/8	-0.0058	0.0878	0.9668	-0.0488
15/16	-0.0026	0.0388	0.9928	-0.0290

4. RESULTS

The influence of interpolation filters on the decoding quality has been tested in SD MCTF environment. The test settings are listed in Table 2.

The decoding has been performed for the highest quality sequence (without quantization). Three test sequences are used - "City", "Bus" and artificial "Crew". The latest sequence is synthesized from the first frame of a high-resolution test sequence "Crew" by sliding window representing one part of this frame, moving in the bottom-right direction. In this way n-th frame of the artificial sequence corresponds to the sliding window displacement of $n/4$ pixels. This decreases the influence of MCTF drift on the interpolation analysis. For the sub-pixel window displacement a cubic spline interpolation of the frame pixels was used.

The results presented in Table 3 justify the application of resolution adaptive interpolation as in most cases PSNR results for sequences decoded using an interpolation filter with $T_r = T_0 / 2^r$ give the best results. However, the gain is relatively low which partially comes from the fact that the effect of the drift on PSNR masks the actual influence of chosen interpolation. In Figure 2 we show details from actual QCIF frames from the "City" sequence in which the worst PSNR results are observed.

Table 2. Test settings

	Encoder	Decoder
Subpixel accuracy	1/4	1/16
Interpolation filter tap	16	4, 8, 16
Resolution	CIF (352 × 288)	QCIF (176 × 144), QQCIF (88 × 72)
Frame rate	30 fps	30 fps

Table 3. PSNR test results for three sequences

Decoding tap	artificial "Crew"		"City"		"Bus"	
	CIF → QCIF	CIF → QQCIF	CIF → QCIF	CIF → QQCIF	CIF → QCIF	CIF → QQCIF
16	49.59 dB	55.50 dB	45.91 dB	46.68 dB	44.88 dB	44.92 dB
8	**49.91** dB	55.84 dB	46.25 dB	46.97 dB	**45.15** dB	45.12 dB
4	49.85 dB	**55.87** dB	**46.27** dB	**47.12** dB	45.00 dB	**45.14** dB

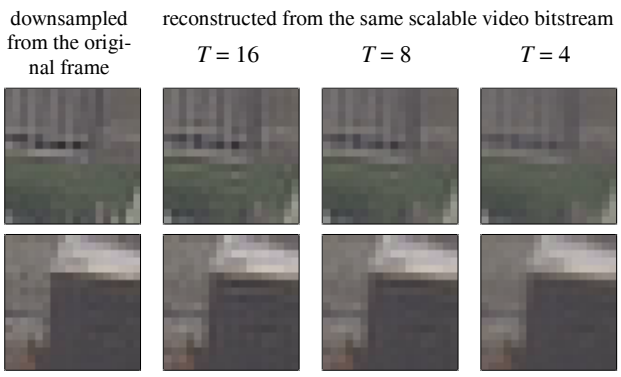


Figure 2. Details from the sequences decoded using different interpolation filters

From the enlarged blocks, a ringing effect caused by the exceedingly long interpolation filters $T = 16$ can be seen. On the other hand, the short filter $T = 4$ blurs the frames. Even if the PSNR results for these frames are not the best for $T = 8$, the corresponding visual quality is high, compared to the details taken from downsampled frame only. However, the application of very short filter of $T = 4$ does not introduce disturbing artefacts and gives good visual quality.

5. CONCLUSION

State-of-the-art SD MCTF codecs use the same interpolation filters at both encoder and decoder sides. That is, if the maximal coding accuracy is $1/M$, the codec uses M - 1 different interpolation filters of the same size. By extending the temporal synthesis equations, we have shown that in the case of spatial scalability, higher accuracy is needed and that the interpolation of lower resolution temporal frames needs reconsideration.

The analysis of spatial pixel dependences has shown that on lower resolution shorter filters can be used. Therefore we propose the application of an extended set of interpolation filters such that for each tap $T_r = T_0 / 2^r$, $r = 0,...,R$, filters for fractional subpixel positions of $\{1/(M\cdot 2^r),...., (M\cdot 2^r-1)/(M\cdot 2^r)\}$ are known.

Experimental results show that the application of our resolution adaptive scheme gives better results than the conventional approach. The observed improvements are in visual quality since ringing artefacts introduced by long interpolation filters are reduced, and in the PSNR gain is up to 0.5 dB. The most important feature of the proposed approach is reduced decoding complexity since for lower resolutions the underlying filters are shorter than non-adaptive ones.

6. REFERENCES

[1] *Scalable Extension of H.264/AVC*, ISO/IEC JTC1/SC29/ WG11, M10569/ S03, 68th MPEG Meeting, Munich, Germany, March 2004.

[2] S.-T. Hsiang and J. W. Woods, "Embedded Video Coding Using Invertible Motion Compensated 3-D Subband / Wavelet Filter Bank," *Signal Processing: Image Communication*, vol. 16, pp. 705-724, May 2001.

[3] A. Secker and D. Taubman, "Lifting-Based Invertible Motion Adaptive Transform (LIMAT) Framework for Highly Scalable Video Compression," *IEEE Trans. Image Processing*, vol. 12, no. 12, pp. 1530-1542, Dec. 2003.

[4] Y. Andreopoulos, M. Van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, J. Cornelis, "Complete-to-Overcomplete Discrete Wavelet Trans-forms for Scalable Video Coding with MCTF," *Proc. Visual Communications and Image Processing* (VCIP 2003), pp. 719-731, Lugano, Switzerland, 2003.

[5] V. Bottreau, M. Benetiere, B. Felts, and B. Pesquet-Popescu, "A fully scalable 3d subband video codec," *Proc. IEEE Int. Conf. on Image Processing* (ICIP 2001), vol. 2, pp. 1017-1020, Oct. 2001.

[6] J.-R. Ohm, "Three-dimensional Subband Coding with Motion Compensation," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 559-571, Sept. 1994.

[7] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and Jan Cornelis, "Fully-Scalable Wavelet Video Coding Using In-Band Motion Compensated Temporal Filtering", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing 2003* (ICASSP 2003), Hong Kong, China, vol. 3, pp. 417-420, March 2003.

[8] *Subjective test results for the CfP on Scalable Video Coding Technology*, ISO/IEC JTC1/ SC29/ WG11, M10737, 68th MPEG Meeting, Munich, Germany, March 2004.

[9] A. Secker and D. Taubman, "Highly Scalable Video Compression With Scalable Motion Coding," *IEEE Trans. Image Processing*, vol. 13, no. 8, pp. 1029-1041, Aug. 2004.

[10] M. Mrak, N. Sprljan, G.C.K. Abhayaratne, and E. Izquierdo, "Scalable Generation and Coding of Motion Vectors for Highly Scalable Video Coding," *Proc. 24th Picture Coding Symposium* (PCS 2004), San Francisco, USA, Dec. 2004.

[11] P. Chen and J.W. Woods, "Bidirectional MC-EZBC with lifting Implementation," *IEEE Trans. Circ. and Systems for Video Technology*, vol. 14, no. 10, pp. 1183-1194, Oct. 2004.