MACROBLOCK-LEVEL MODE BASED ADAPTIVE IN-BAND MOTION COMPENSATED TEMPORAL FILTERING

Anyu Gao, Nishan Canagarajah and David Bull

Image Communications Group, Centre for Communications Research, University of Bristol,
Merchant Ventures Building, Woodland Road, Bristol BS8 1UB, United Kingdom,
E-mail: {anyu.gao, nishan.canagarajah, dave.bull}@bristol.ac.uk

ABSTRACT

This paper presents an adaptive in-band motion compensated temporal filtering (MCTF) scheme for 3-D wavelet based scalable video coding. The proposed scheme solves the motion mismatch problem when motion vectors from the LL subband are inaccurately applied to the highpass subbands in decoding high spatial resolution video. Specifically, we compare the macroblock residue energy in the highpass frames obtained by using motion vectors from both the LL and highpass subbands, and then adaptively transmit different sets of motion vectors based on whether mismatch has occurred in the highpass subbands. Macroblocks in the higher temporal levels favour the selection of highpass subbands' motion vectors because the motion estimation process becomes less accurate as temporal level increases. The modes information, which specifies whether the LL subband motion vectors or the highpass subbands' motion vectors are used by the current macroblock, is coded by run-length coding. Experimental results show that the proposed scheme improves both the visual quality and PSNR for high resolution decoding with comparison to other in-band MCTF schemes. Furthermore, our scheme requires only modifications when performing MCTF in the highpass subbands, thus, the original strength of in-band MCTF for decoding low spatial resolution video is well preserved.

Keywords: Wavelet transform, in-band MCTF, motion mismatch

1. INTRODUCTION

The open-loop 3-D wavelet scalable video coding [1] [2] based on motion compensated temporal filtering (MCTF) has attracted great attention in recent years. This class of video coding schemes eliminates the "drift" problem suffered by predictive coding schemes like [3], and is also able to provide combined temporal, spatial and SNR scalabilities with high compression efficiency.

Traditional 3-D wavelet coding schemes exploit temporal redundancy by performing MCTF in the spatial domain, i.e. on the original frames. This process has the potential of introducing motion mismatch when decoding video at low resolution due to motion vector (MV) down-scaling. Therefore, in-band MCTF based schemes [4] [5] [6] have been proposed. In in-band schemes, the original frames first undergo typically one or two levels of spatial discrete wavelet transform (DWT), called pre-temporal spatial DWT, and the prediction and update steps are subsequently performed in each of the lowpass and highpass spatial subbands. Since each subband (resolution) now has its own motion field, the above mentioned problem is naturally solved.

Conventional in-band schemes perform motion estimation (ME) on all pre-temporal spatial subbands [5] (denoted multi-scheme) and transmit all the resulted MVs. This is uneconomical for low bit rates, since there are certain correlations between these MV sets. Thanks to the interleaving algorithm [6], ME can be performed on only the LL subband [7] (denoted single scheme); the highpass subbands can use the same set of MVs for prediction and update. However, if the MVs from the LL subband do not capture the underlying motion in the highpass subbands, mismatch artefacts will appear in the decoded video. In [8], a subband-based adaptive approach has been proposed. It removes the mismatch by additionally transmitting the highpass subband MVs. However, the cross-band motion information correlation is not well exploited since some macroblocks (MBs) in the highpass subbands do not need their own MVs to be transmitted to the decoder.

We extend the idea in [8], and propose a MB-level adaptive in-band MCTF scheme that transmits only the necessary highpass subbands' MVs so that a better motion-texture trade-off can be achieved. The MV selection decision is made by detecting motion mismatch on the MB-level in the highpass spatial subbands. The rest of the paper is organised as follows: In section 2, we give some background information on MCTF; Section 3 analyses the motion mismatch problem in the highpass spatial subbands caused by the single scheme; the proposed adaptive scheme is detailed in section 4; section 5 presents the experimental results in both PSNR and visual quality with comparison to other in-band schemes. Conclusions and future work are given in section 6.

2. BACKGROUND ON MCTF

A breakthrough in the implementation of MCTF is the lifting scheme [1] [2] that guarantees perfect reconstruction. Lifting based MCTF performs wavelet transform in two sequential steps, the prediction and the update steps. In our experiments, the bi-directional 5/3 wavelet is used due to its better complexity-efficiency trade-off comparing to other wavelet transforms [9]. The prediction and update steps for 5/3 lifting are:

$$h_k = f_{2k+1} - \frac{1}{2}[W_{2k\to 2k+1}(f_{2k}) + W_{2k+2\to 2k+1}(f_{2k+2})] \quad (1)$$

$$l_k = f_{2k} + \frac{1}{4}[W_{2k-1\to 2k}(h_{k-1}) + W_{2k+1\to 2k}(h_k)] \quad (2)$$

where f_k denotes the original input frames, and $W_{k_1 \to k_2}(f_{k_1})$ denotes a motion compensated mapping operation that maps frame k_1 onto the coordinate system of frame k_2.

The prediction step in equation (1) forms the temporal highpass frame h_k, which is the motion compensated residue. The update step in equation (2) forms the corresponding temporal lowpass frames l_k. The update step serves to ensure efficient lowpass filtering of the input frames along the motion trajectories. This predicting/updating operation continues on the lowpass frames in each temporal level until the highest temporal level where in general only one lowpass frame will be left. Perfect reconstruction comes naturally by reversing the order of the lifting steps and replacing additions with subtractions as follows:

$$f_{2k} = l_k - \frac{1}{4}[W_{2k-1 \to 2k}(h_{k-1}) + W_{2k+1 \to 2k}(h_k)] \quad (3)$$

$$f_{2k+1} = h_k + \frac{1}{2}[W_{2k \to 2k+1}(f_{2k}) + W_{2k+2 \to 2k+1}(f_{2k+2}) \quad (4)$$

The MCTF process in the spatial domain can be extended to the subband/wavelet domain by performing prediction and update steps on the spatially transformed highpass and lowpass wavelet coefficients. In order to eliminate the shift-variant problem in the critically-sample DWT domain, in-band MCTF is always performed in the overcomplete DWT (ODWT) domain [4] [5] [6].

Suppose that a 1-level pre-temporal DWT is applied to the original frames. This will result in each frame being transformed into 4 spatial subbands, namely LL, HL, LH and HH subbands in their critically-sampled DWT representations. It should be noted that the highpass subbands (HL, LH and HH) are necessary in forming their ODWT representations at the encoder. However, these highpass subbands are not present in decoding low resolution video. In this case, the decoder will use interpolation to produce a set of "low-quality references" [4] [7].

3. PROBLEM ANALYSIS

Traditionally, ME is performed on all pre-temporal subbands [5], each subband then uses its own MVs to perform prediction and update. For a 1-level pre-temporal DWT, this process will result in 4 sets of MVs, which is uneconomical (see Table 1) in terms of motion-texture trade-off since there are certain correlations between these MV sets.

As mentioned previously, ME can be performed only on the LL subband [7], the highpass subbands can then use the same set of MV to perform prediction and update. Generally, this scheme works reasonably well for the first few temporal levels. However, as temporal level increases, the ME process of the LL subband generally becomes less accurate due to larger motion displacement between any two lowpass frames and their lower-quality (due to inaccurate update) comparing to the higher temporal levels. If the less accurate motion information is applied to the corresponding highpass frames, mismatch will appear in the highpass subbands which then translate into annoying visual artefacts in the reconstructed high-resolution video.

Figure 1 (left) shows an example of the inaccurately predicted/updated highpass subbands from the highest level temporal by encoding the *foreman* sequence. Note the illuminated mismatch areas in the HL and LH subbands and the lines around face and neck in the HH subband. The corresponding reconstructed frame with visual artefacts around foreman's face, neck and his helmet is shown in Figure 1 (right).

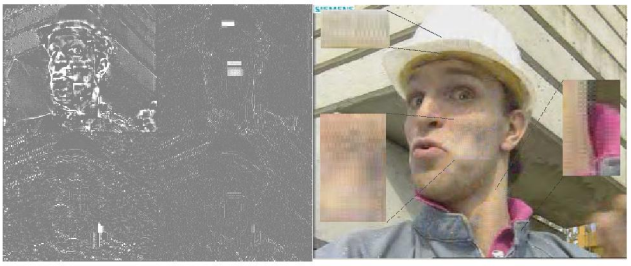

Figure 1: Wavelet-domain highpass subbands motion mismatch (left) and visual artefacts in the reconstructed video (right) of frame 89 (highest temporal level of a 4-level MCTF) for *foreman* using the single scheme, bit-rate: 256kbps

From equations (3) and (4), it can be seen that if there are significant errors in the spatial highpass subbands in the highest temporal level highpass frame, these errors would not only deteriorate the current temporal level but also propagate to subsequent lower temporal levels due to the recursion property of inverse MCTF, and hence the quality of all the reconstructed frames in the current GOP will be degraded.

In [8], we proposed a subband-level adaptive in-band MCTF scheme (denoted the subband-adaptive scheme) that removes the motion mismatch by selectively transmitting the MVs of the entire related highpass spatial subbands. This approach assumes that when mismatch occurs in 1 MB in 1 highpass subband, it is also likely to occur in other MBs in the current and the rest highpass subbands. However, for sequences with large areas of smooth motions, transmitting the highpass subbands' MVs of the entire subbands may not be efficient in terms of utilising the total bit-budget. Furthermore, MBs in areas with smooth motions may in fact be better predicted in terms of reducing the prediction error energy using the MVs of the collocated lowpass subbands' MBs [7]. Table 1 compares the number of motion bits generated by performing ME on the second 64 frames of the *foreman* sequence using the approaches from [5] [7] [8]. As can be seen, although the subband-adaptive scheme reduces the overall motion bits significantly comparing with the multi-scheme, some highpass subbands' MVs are in fact unnecessarily transmitted to the decoder. The objective of the proposed approach is therefore to find a more efficient way in the MV selection process to eliminate motion mismatch as well as suppressing the MCTF prediction error, so that both the visual quality and PSNR performances can be improved.

T-level	Multi	Single	Subband-adaptive
1	52248	23976	24024
2	36672	18800	20248
3	24200	13976	17960
4	15456	9800	14272
Total	128576	66552	76507

Table 1: Number of motion bits generated by a 4-level MCTF of the second 64 frames (2nd GOP for bitstream truncation [11]) for CIF *foreman* using the multi- [5], single [7] and subband-adaptive [8] in-band schemes, 1-level pre-temporal 9/7 DWT is used

4. MB-LEVEL ADAPTIVE IN-BAND MCTF

From the discussions in the previous section, it is intuitive that the MCTF may be performed more efficiently if the MV selection process occurs at the macroblock level.

In equation (1), it is shown that the highpass frame h_k is the residue left after motion compensation. In regions where the motion model captures the actual motion, the energy in the highpass frames will be close to zero. On the other hand, when the motion model fails, this energy will increase, as shown in Figure 1 (left). We use this criterion to determine whether to perform single in-band or multi in-band MCTF for a certain MB. The energy in a MB is defined as:

$$E_{MB} = MSE = \sum_{y=0}^{Y-1}\sum_{x=0}^{X-1}[c^2(x,y)/(Y*X)] \quad (5)$$

where $c(x,y)$ is the wavelet coefficient at coordinate (x,y) within the macroblock; Y and X are the height and width of the macroblock.

We then define the macroblock energy ratio between the motion compensated macroblock obtained by single and multi- in-band schemes as:

$$\alpha = \frac{E_{MB\_Single}}{E_{MB\_Multi}} \quad (6)$$

If α exceeds a pre-defined threshold value α_0, a mismatch is expected to occur in the highpass subbands, and the highpass MVs are used to prevent the mismatch; on the other hand, if α is below the threshold value, which means mismatch is unlikely to occur, therefore, the MVs of the collocated MB from LL subband is used to perform MCTF. Adjusting the value of α_0 allows us to trade coding efficiency for visual quality (i.e. reduction of artifacts). We use smaller α_0 for lower temporal levels and larger α_0 for higher levels, since the motion accuracy generally decreases as temporal level increases as previously mentioned. We also observed from our experiments that if, for example, a mismatch is detected in the HL subband, the collocated MBs in other highpass subbands are also likely to contain mismatch errors (see Figure 1 left). Therefore, if α exceeds α_0 for one MB in one subband, then the collocated MBs from other highpass subbands are also expected to have mismatch and hence will have their own MVs transmitted. A simplified block diagram of the proposed adaptive scheme is shown in Figure 2.

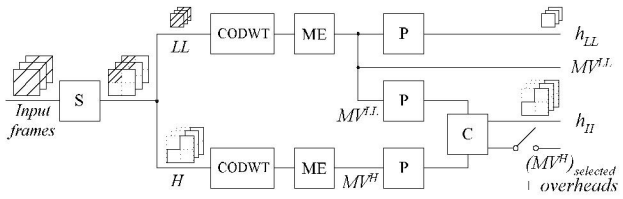

Figure 2: Block diagram of the proposed adaptive in-band scheme

In Figure 2, the blocks S, ME and P denote the pre-temporal spatial DWT, ME and MCTF prediction respectively; the highpass subbands are collectively denoted as H, hence h_{LL} and h_H are the highpass temporal subbands of the LL and other highpass spatial subbands respectively. C denotes the comparison operation that determines whether an MB from the highpass subbands should perform prediction using MV^{LL} or MV^H. Finally, all MVs from MV^{LL} and a selected set from MV^H, together with some overhead information are embedded into the bitstream.

The proposed scheme requires two types of additional overhead information to be included in the final bitstream. 1) A flag bit for each frame-level elementary ME process, indicating whether the coming MV bitstream contains highpass subbands' MVs or not, and 2) a 1-bit MV mode per MB for all MBs in the highpass subbands to specify whether this MB and the collocated highpass subbands' MBs should use their own MVs to perform inverse MCTF. Both overheads are essential for decoder synchronisation.

The first type of overhead is un-coded because it only takes a small amount of the bit-budget. For example, a CIF encoding with 1-level pre-temporal DWT and 4-level in-band MCTF would have 15 elementary ME processes, and hence only require 15 bits for flag information. The mode information on the other hand, consumes more bits than the flags. For the above example with MB size of 16x16, a total number of (176/16)*(144/16) = 99 bits are required for 1 elementary ME. Given an acceptably efficient motion estimator, most MBs in the highpass subbands can be predicted using the corresponding LL subband MVs, hence this type of MBs takes a much higher percentage than highpass MBs that should use their own MVs for MCTF prediction. Taking this property into consideration, we adopt the simple run-length coding (RLC) technique to code the mode information. We will show in the experimental results that the amount of additional overhead incurred by run-length coding is worthy because the proposed method singles out all the unnecessary highpass MVs that would have been transmitted by the subband-adaptive approach in [8]. As a result, the smooth region highpass MBs are better predicted by lowpass MVs, and hence more bits are saved for texture coding. It is also worth noting that the proposed scheme should be applied to sequences with considerable complex motions (e.g. foreman, football etc.). For less motive sequences (e.g. Akiyo), the added MV mode overhead, may instead worsen the motion-texture trade-off since there may be no significant mismatch in the highpass subbands.

5. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed adaptive scheme in comparisons with the multi- [5], single [7] and the subband-adaptive schemes [8]. These results were obtained by encoding the CIF sequences of *foreman* (300 frames@30frames/second) with 4-level 5/3 MCTF and 1-level pre-temporal 9/7 DWT. The ME and motion compensation operations use variable-sized blocks similar to H.264 [10].

We implemented the proposed in-band MCTF using MPEG's reference software [11] on 3-D wavelet video coding. In-band ME is always performed in the ODWT domain using the "high-quality reference" for both encoding and decoding.

Table 2 shows the mean PSNR[1] by decoding at a number of bit-rates. The values of α_0 are set to 60, 30, 15 and 4 for temporal levels 1, 2, 3 and 4 respectively, and these values are obtained through several experiments. As can be seen, the proposed MB-adaptive scheme outperforms the single scheme [7] and subband-adaptive scheme [8] for up to 0.1dB and 0.18dB respectively.

[1] $PSNR_{MEAN} = (4 \cdot PSNR_Y + PSNR_U + PSNR_V)/6$

bit-rate (kbps)	Multi	Single	Subband-adaptive	MB-adaptive
128	32.9837	33.6675	33.5893	33.7618
160	33.9062	34.5294	34.4541	34.5993
192	34.6436	35.1625	35.0889	35.2207
224	35.1924	35.591	35.5389	35.6484
256	35.6082	36.0091	35.9531	36.0589
384	36.9207	37.2562	37.2111	37.3107
512	37.8199	38.1409	38.0894	38.1998

Table 2: PSNR comparisons for multi, single, subband-adaptive and the proposed MB adaptive in-band MCTF

Table 3 below shows the number of motion bits generated by each of the four schemes. The proposed MB-adaptive approach further reduces the number motion bits required for MCTF by the subband-adaptive scheme. The bit savings and the removal of the mismatch, together with the efficient use of MV^{LL} on the highpass subbands' MBs contribute to the PSNR improvement in Table 2.

MCTF level	Multi	Single	Subband-adaptive	MB-adaptive
1	242352	115136	117816	115312
2	164464	87568	90096	87808
3	106424	62624	74840	65024
4	68920	44512	62752	49152
Total	582160	309840	345504	317296

Table 3: Motion bits usage comparisons for multi, single, subband-adaptive and the proposed MB adaptive in-band MCTF

Figure 3 below shows the same frame as in Figure 1 but reconstructed by the proposed MB-adaptive scheme. It is clear that the mismatch errors in the highpass subbands are eliminated. As a result, the reconstructed frame shown in Figure 3 (right) is now free of highpass mismatch artifacts.

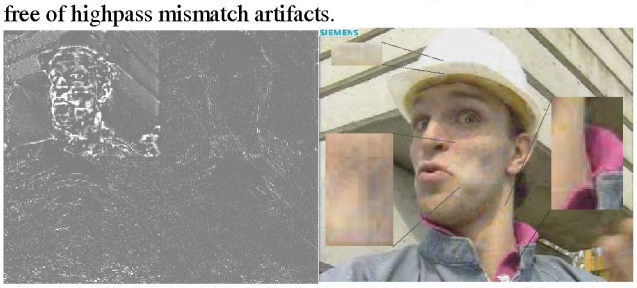

Figure 3: Highpass subbands (left) and reconstructed video (right) of frame 89 for *foreman* using the proposed MB-level adaptive scheme at 256kbps, refer to Figure 1 for comparison

6. CONCLUSIONS AND FUTURE WORK

We proposed a macroblock-level adaptive in-band motion compensated temporal filtering scheme based on motion mismatch detection in the highpass subbands. The proposed scheme solves the highpass-subband motion mismatch problem by adaptively transmitting different sets of motion vectors based on mismatch detection in the highpass subbands. Experimental results show that the proposed scheme improves both the visual quality and PSNR for high resolution decoding with comparison to other latest in-band MCTF schemes. Furthermore, our scheme only requires modifications when performing MCTF in the highpass subbands, hence the original strength of in-band MCTF for decoding low spatial resolution video is well preserved.

In the current scheme, we use empirical values to predict whether mismatch would occur if the LL subbands' MVs are applied to the highpass spatial subbands, and these values are determined after several experiments. For future work, we plan to embed the mismatch detection into the motion estimation process so that a more accurate set of α_0 values maybe obtained.

7. REFERENCES

[1] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Trans. Image Proc.*, vol. 12, pp. 1530- 1542, Dec. 2003.

[2] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, pp. 1183-1194, Oct. 2004.

[3] W. Li, "Overview of fine granularity scalability in MPEG4 video standard," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, pp. 301-317, Mar. 2001.

[4] A. M. Y. Andreopoulos, J. Barbarien, M. van der Schaar, J. Cornelis and P. Schelkens, "In-band motion compensated temporal filtering," *Signal Processing: Image Communication*, vol. 19, pp. 653-673, 2004.

[5] H. S. Kim and H. W. Park, "Wavelet-based moving-picture coding using shift-invariant motion estimation in wavelet domain," *Signal Processing: Image Communication*, vol. 16, pp. 669-679, 2001.

[6] J. C. Ye and M. van der Schaar, "Fully Scalable 3-D Overcomplete Wavelet Video Coding using Adaptive Motion Compensated Temporal Filtering," *Proc. SPIE Video Communications and Image Processing*, Jan. 2003.

[7] D. Zhang, J. Xu, F. Wu, W. Zhang, and H. Xiong, "Mode-Based Temporal Filtering for In-Band Wavelet Video Coding with Spatial Scalability," *Proc. SPIE Visual Communication Image Processing*, Jul. 2005.

[8] A. Gao, N. Canagarajah and D. Bull, " Adaptive in-band motion compensated temporal filtering based on motion mismatch detection in the highpass subbands," *Proc. SPIE Visual Communication Image Processing*, Jan. 2006.

[9] N. Mehrseresht, and D. Taubman, "An efficient content-adaptive motion compensated 3D-DWT with enhanced spatial and temporal scalability," *Proc. IEEE ICIP*, vol.2, pp. 1329- 1332, Oct. 2004.

[10] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Tranc. Circuits Syst. Video Technol.*, vol. 13, pp. 560- 576, Jul. 2003.

[11] R. Xiong, J. Xu, B. Feng, G. Sullivan, M-C. Lee, F. Wu and S. Li, "3D Sub-band Video Coding using Barbell Lifting," *ISO/IEC JTC/WG11* M10569, S05, Mar. 2004.