A Simplified Scalable Wavelet Video Codec with MCTF Structure

Bin Wang[1], K.K. Loo[1], P.Y. Yip[2], M.F. Siyau[3]
[1]*School of Engineering and Design, Brunel University, UK*
[2]*School of Computer Science, University of Hertfordshire, UK*
[3]*Dept. Elec., Comp. & Eng., London South Bank University, UK*
Email: Bin.Wang@brunel.ac.uk, Jonathan.Loo@brunel.ac.uk

Abstract

In recent years, wavelet-based image and video coding systems that utilize a wide range of spatial-temporal-SNR scalability with state-of-the-art coding performance has been proposed in the literature. The strong market acceptance of the new scalable technology stems from advances in network technology as well as various requirements from terminal users since computer network is becoming the main media transmission. This paper presents a simplified scalable wavelet video coding structure using MCTF with 5/3 filter. This structure does not include motion estimation, hence entropy coding for motion vectors is omitted. The proposed codec generates four temporal resolution layers, with up to four spatial resolution levels and six variable quantization levels for certain sequence. Simulation results show the codec has great flexibility feature and the reconstructed video quality is thought to be of acceptable quality.

1. Introduction

In the last decade the telecommunication industry has been experiencing a digital revolution. The need for accessing multimedia content over heterogeneous networks has brought an extensive interest within the research and standardization communities to try the possibility of applying scalable wavelet feature into multimedia processing and transmission in particular within the area of wavelet-based scalable video coding with motion-compensated temporal filtering (MCTF) [1] where if used in an optimal manner will achieve full-scalability in terms of video size, resolution and quality.

In this paper, based on some existing technologies, a simplified scalable wavelet video coding structure using MCTF with 5/3 filter is presented. The remarkable feature in the proposed coding structure is that there is no motion estimation during MCTF; hence it is not necessary to deal with motion vectors in encoding process. Moreover, a flexible function can achieve spatial-temporal changes in the proposed codec for which can generate up-to four levels of spatial resolution, four layers of temporal resolution and six variable quantization levels maximum. The experimental results demonstrated a good PSNR as well as acceptable reconstructed video quality.

The paper is organized as follows. In section 2, background introduction on discrete wavelet transform (DWT) [2], MCTF and scalability is presented briefly. Subsequently, in section 3, the proposed scalable wavelet video codec architecture and the bitstream representation are presented. Section 4 presents some experimental results and section 5 concludes the paper.

2. Background

The following subsections provide a brief background on DWT, MCTF and Scalability.

2.1 Discrete wavelet transform

DWT, based on sub-band coding, is found to yield a fast computation of wavelet transform. It has both low-pass and high-pass filters with rescaling to decompose input video into several levels. The popular decomposition called Mallat-tree algorithm [3] has been employed in our design. For its reversibility, Mallat-tree algorithm is able to decompose video frame into levels at encoder and reconstructed at the decoder.

2.2 Motion compensated temporal filtering

In the scalable video context, MCTF has been known to provide an open-loop encoding structure [4]. This lifting scheme consists of three steps i.e. polyphase operation, prediction and update. In the analysis filter bank, the input signal Si is divided into two parts (odd S2i+1 and even S2i) after polyphase decomposition; the even part is motion compensated

0-7695-2650-0/06/$20.00 (c) IEEE

(MC) using motion prediction while the odd part in the prediction stage is known as the prediction residual:

$$Hi = S2i + 1 - P(S2i) \quad (1)$$

The Hi is then again compensated in the updated stage. Depending on the two MC operators in P (.) and U (.), the expression can be decided. For example, if they are liner and invertible i.e. U (P(s)) = 1/2, then

$$Li = S2i + \mathbf{U}(S2i+1 - \mathbf{P}(S2i)) = S2i + \mathbf{U}(S2i+1) \quad (2)$$

And in the synthesis filter bank, it is the reverse process compared with the analysis filter bank.

The spatial-domain (t+2D) and in-band (2D+t) approaches with new feature of open-loop video coding schemes according to the way how motion-aligned temporal transform is performed [4]. In 2D+t, the video frame is first spatial decomposed and MCTF is carried out in wavelet domain, possibly with subsequent further spatial decompositions while in t+2D, MCTF is applied to video frame directly in spatial domain before the spatial decomposition. Some finding from previous testing [5] have shown that the architecture of in-band provides the additional feature of independent operation of the MCTF at each spatial level, leading to more efficient resolution scalability and additional degrees of freedom for coder design. In this paper, we utilize the modified in-band MCTF (IBMCTF) in our design structure.

2.3 Scalability

Scalability is attractive due to the capability to reconstructing low resolution or low quality video signal from partial bitstream; this makes a simple solution in adaptation to network and terminal capability.

2.3.1 Spatial Scalability

Spatial scalability means that the video at multimedia resolutions (e.g. CIF, QCIF, SDTV and HDTV) with a factor of 2 in horizontal and vertical resolution. During the experiment, we use DWT module to implement these changes. There are four different resolutions i.e. HDTV, SDTV, CIF and QCIF, the relation between them are HDTV≈4SDTV, SDTV≈4CIF, CIF=4QCIF. Hence, for the four standard resolutions, we can set maximum four spatial levels for HDTV, three spatial levels for SDTV, two spatial levels for CIF and one spatial level for QCIF.

2.3.2 Temporal Scalability

Temporal scalability generates frame layers where the lower one is encoded to provide the basic temporal rate and the enhancement layer is coded with temporal prediction with respect to the lower layer. A hybrid temporal structure around four layers is used in our module. In each layer, it is an independent MCTF scheme, including prediction and update stage. It has been proved that it is very efficient after using hybrid structure in video coder [6].

3. The Proposed Scalable Codec

The main encoding structure using wavelet scalable video coding is illustrated in the Figure 1. The difference between the 2D+t and the simplified 2D+t is that there is no MV in the latter one.

The input video frames are divided into two parts: low frequency band and high frequency band after performing the DWT function, then utilizing complete–to-overcomplete DWT (CODWT) [7] to reconstruct the high-frequency band. The number of spatial level depends on the sample resolution. In our structure, we set QCIF as the lowest level for DWT. For example, HDTV can have four levels, namely, HDTV, SDTV, CIF and QCIF and low-frequency band QCIF with high-frequency band CIF can compose reconstructed CIF and the same way for SDTV and HDTV. We recognize the lowest level QCIF as the low frequency band and the others as high frequency band.

For low frequency band, it deals with MCTF function directly. In Figure 2, the multi-layer MCTF structure with 5/3 filter used in our design codec is shown. With the bidirectional 5/3 filter used in this scheme, each frame needs two reference frames. There are four layers based on a GOP size of 8, each layer should perform MCTF separately except layer 0 and the procedure works in a top to bottom approach at the encoder. At layer 0, the eight frames, which are the original low frequency band from the lowest level after DWT, will serve as an input sequence into layer 1. At layer 1, the nearby two frames are reference frames for the corresponding one in prediction stage for the bidirectional scheme of the 5/3 filter. At this point the prediction residual H is immediately created. Similarly, the reconstructed frame L is produced in the update stage in the same manner. Then all the L frames from layer 1 are received by layer 2 as input. The same process is repeated at layer 2 and layer 3 until there is only one reconstructed frame L in the lowest layer. At last, for each GOP, the necessary bitstream is composed for transmission only covering the reconstructed frame L at layer 3 and the whole error frames H from layer 1, 2 and 3. Some coefficients have been attached in every layer, e.g. 1/2 at the prediction

stage while 1/4 at the update stage. These coefficients can be changed depending on the accuracy. In MCTF, we can optimize the content of each GOP and reduce maximum energy it contained.

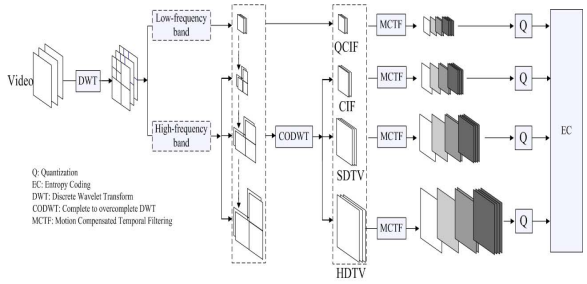

Figure 1. The encoding structure that performs open-loop framework in simplified codec for HDTV.

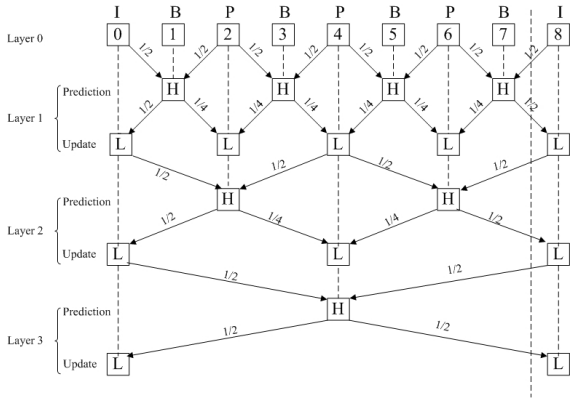

Figure 2. The four-layer MCTF scheme with GOP size 8.

After all the signals (low frequency band and high frequency band) is processed using MCTF, they are quantized and coded by the entropy coder for transmission. Figure 3 shows bitstream representation of the proposed scalable video codec for the case of input HDTV sequence, and example of bitstream composition for different spatial and temporal resolution.

The decoder can automatically acquire necessary bitstream from the lowest level QCIF to HDTV depending on the network environment and the user terminal's spatial-temporal display requirement as demonstrated in Figure 4. For spatial changes, if input video frame is of an SD resolution (704×576) and the client wants to display the video in CIF (352×288) at its terminal, the combination bitstream between QCIF and CIF will be necessary. It is known that better video quality can be obtained from the higher the resolution hence there is need to use more video information from upper resolution levels to reconstructed the picture. For temporal variation, we may have to drop some frames in order to suit the network transmission condition. In this scenario, certain bitstream will be selected from the different resolution to compose into output. Figure 3 (b) and (c) demonstrates the bitstream composition for different spatial and temporal resolution.

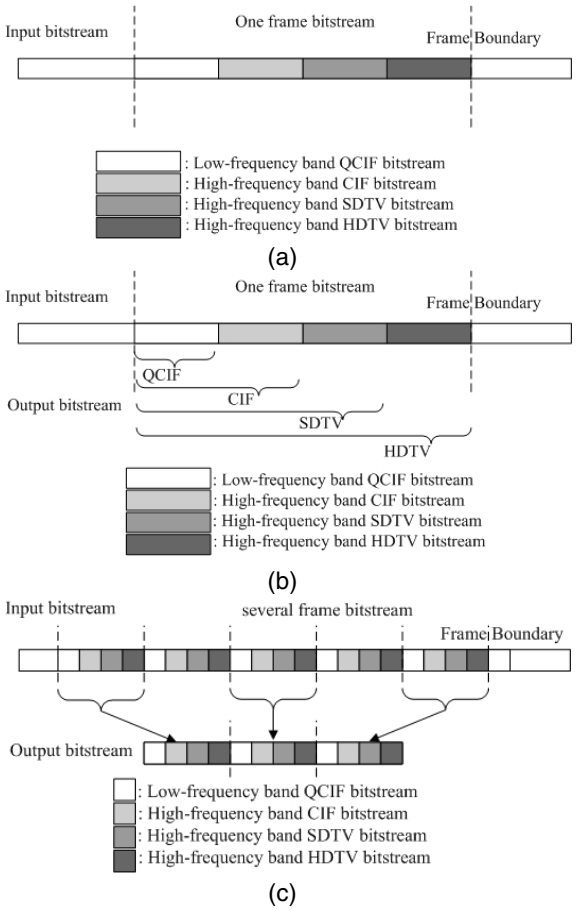

Figure 3. (a) Bitstream representation of the proposed scalable video codec, (b) Bitstream composition for different spatial resolution. (c) Bitstream composition for different temporal resolution.

The feature of the proposed codec can be summarized as follows:

Multi-level decomposition: DWT is used to decompose a video sequence into various spatial resolutions i.e. QCIF, CIF, SDTV and HDTV. In each spatial level, open-loop MCTF encoding is used to decompose up to four temporal layers provide temporal scalability from layer 0, layer 1, layer 2 and layer 3.

"Zero value" complementation: After DWT, it is critical to construct a complete picture from the high frequency subbands so that higher spatial level and its temporal resolution can be obtained. Instead of the conventional way of using CODWT, we complete the high frequency subband with "zero value" complementation. It has been demonstrated that it has

certain advantages in comparison to the conventional CODWT.

Flexible multi-level quantization: There are six various quantization levels: $2^n \in \{4, 8, 16, 32, 64, 128\}$. For low and high frequency subbands, the quantization level can be different. Generally, high level quantization level is set to low frequency band while lower for high frequency subbands. The quality of picture improved with the increase of quantization level. From our simulation,, it has been found that the difference of PSNR between level {32} and {64} and {128} is quite small, therefore, we can make conclude that level {32} is the best choice considering both compression ratio and picture quality.

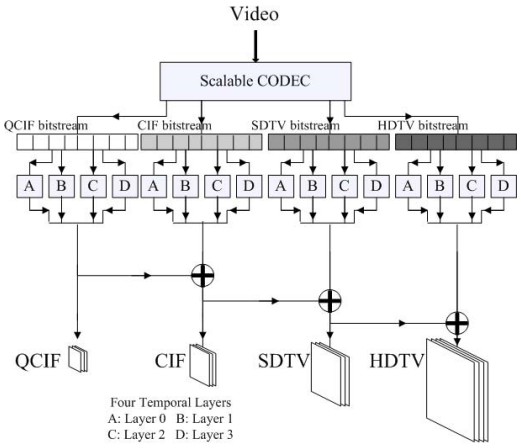

Figure 4. Bitstream composition for different spatial-temporal resolution displaying.

4. Experimental Results

Based on the analysis obtained in the previous section, we carried out our experiments using 3 video sequences i.e. "Foreman" (176x144), "News" (352x288) and "mobcal_ter" (720x576). The GOP size for all 3 video sequences is set to 8 with a frame rate of 15 frames/sec. Different quantization levels were applied, $2^n \in \{4, 8, 16, 32, 64, 128\}$, to obtain the six rate-distortion points. In this case, the spatial scalability is varied betweem QCIF to SDTV while temporal scalability is varied from layer 0 to layer 3. Figure 5 to 7 show PSNR versus bit per pixel (bpp) of the QCIF "foreman", CIF "News" and SDTV "mobcal_ter" respectively under same condition i.e. the same spatial resolution, four different layers in temporal and six different quantization levels. The trend of each line is quite similar and the increment after the quantization level of 32 is very small, which indicates this is an ideal point combined with factors of calculation efficiency and frame quality.

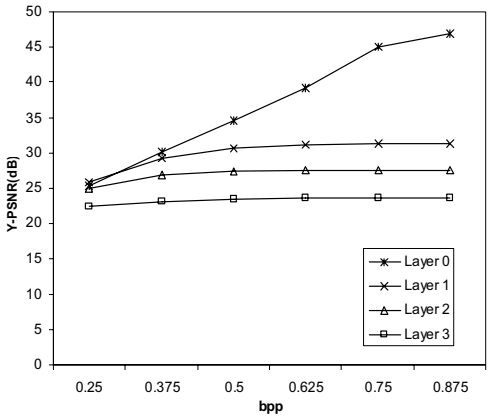

Figure 5. PSNR vs. bpp for "Foreman" sequence

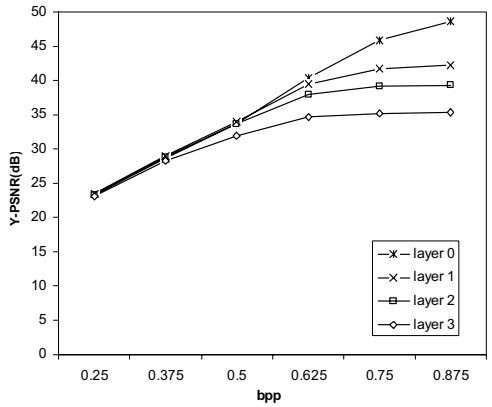

Figure 6. PSNR vs. bbp for "News" sequence

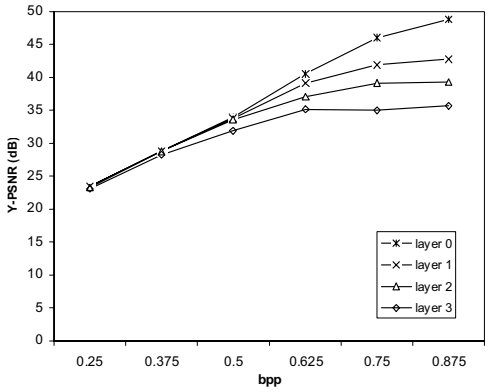

Figure 7. PSNR vs. bit per pixel for "mobcal_ter" sequence

Figure 8 shows only the Y component of the first frame of SDTV "mobcal_ter" sequence under the same spatial resolution SDTV and quantization level {32} but different temporal layers. From Figure 8(a) to (d), the corresponding layer is layer 0 to layer 3. The picture of every layer is very clear and the quality is good. Here, it is observed that the deterioration of the overall performance of the sequences is very minor as the layer increases.

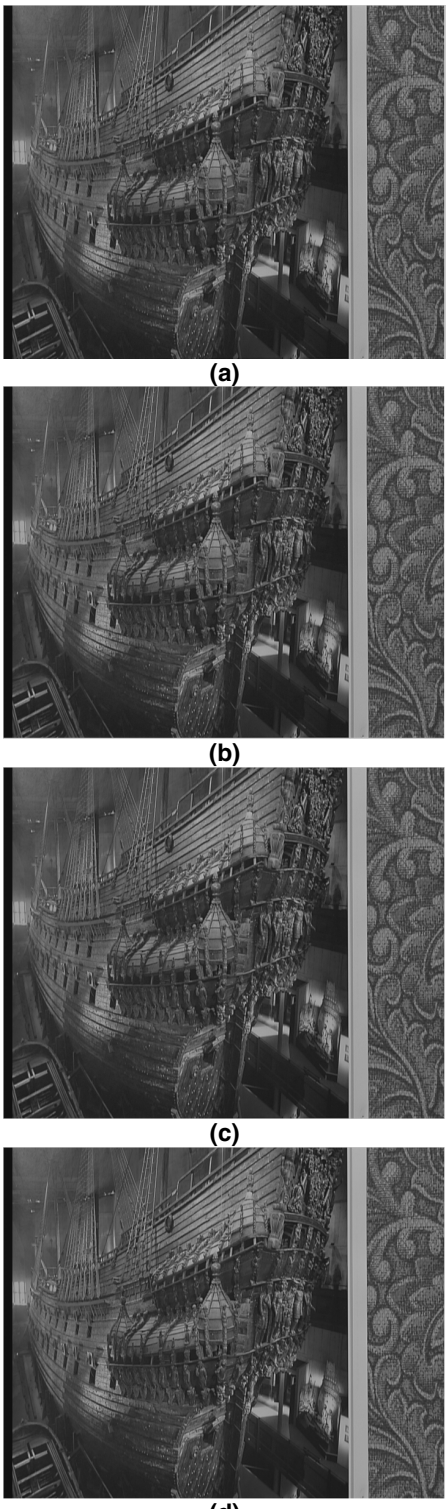

Figure 8. Y component of the first frame from "mobcal_ter" sequence at SDTV and quantization level {32}, (a) Layer 0, (b) Layer 1, (c) Layer 2, (d) Layer 3.

5. Conclusion

This paper demonstrates a simplified scalable wavelet video coding scheme. The codec produces a flexible spatial-temporal scalability with good picture quality at every spatial and temporal resolution. The encoding process uses DWT to decompose a video sequence into various spatial levels. For low frequency subband that is the lowest spatial level, open-loop MCTF encoding is used to produce the temporal scalability. CODWT is used to construct full picture from the high frequency subband, open-loop encoding is carried out to produce temporal scalability for higher spatial levels. The encoding process does not require motion vectors, hence omitted the need for entropy coding in this regard. The spatial-temporal scalability of the proposed codec makes the bitstream representation of a compressed video more adaptable and flexible to network condition and terminal user requirements.

6. References

[1] H. Schwarz, D. Marpe, T. Wiegand, "MCTF and Scalability Extension of H.264/AVC", Symposium Picture Coding , San Francisco, CA, USA, Dec. 2004

[2] A. Al Muhit, Md. S. Islam, M. Othman, "VLSI Implementation of Discrete Wavelet Transform (DWT) for Image Compression", 2nd Int. Conf. on Autonomous Robots and Agents, Palmerston North, New Zealand, Dec. 2004

[3] S. Deepika, "Efficient Implementations of Discrete Wavelet Transforms Using FPGAs", http://etd.lib.fsu.edu/theses/available/etd-11242003-185039/, Florida State University, Nov. 2003.

[4] R. Schaefer, H. Schwarz, D. Marpe, T. Schierl, T. Wiegand, "MCTF and scalability extension of H.264/AVC and its application to video transmission, storage, and surveillance". Proc. SPIE Visual Communications and Image Processing, vol. 5960, pp. 343-354, Jul. 2005.

[5] Y. Andreopoulos, M. Van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, J. Cornelis, "Open-loop, in-band MCTF for objective full-scalability in wavelet video coding," ISO/IECJTCI/SC29/WG11, m9026, Shanghai, China, Oct. 2002.

[6] M. Domański, S. Maćkowiak, L. Błaszak. " Efficient Hybrid Video Coders With Spatial and Temporal Scalability", Proc of IEEE Int. Conf. on Multimedia and Expo, Lausanne, Switzerland, Aug. 2002.

[7] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, J. Cornelis and P. Schelkens, "Complete-to-overcomplete discrete wavelet transforms: theory and applications", IEEE Trans. on Signal Processing, vol. 53, no. 4, pp. 1398-1412, Apr. 2005.