MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Video Coding Without Motion Compensation Using a 3-D Dual-Tree Wavelet Transform

Beibei Wang, Yao Wang, Ivan Selesnick and Anthony Vetro

TR2004-138 December 2004

Abstract

This paper explores the use of a recently introduced 3-D dual-tree discrete wavelet transform (DDWT) for video coding. The 3-D DDWT is an attractive video representation because it isolates motion along different directions in separate subbands. However, it is an overcomplete transform with 8:1 or 4:1 redundancy. Based on the effectiveness of the iterative projection-based noise shaping scheme proposed by Kingsbury on reducing the number of coefficients, and the investigation about the correlation between subbands at the same spatial/temporal location, both in the significance map and in actual coefficient values, a new video coding scheme using 3D DDWT is proposed. The proposed video codec does not require motion compensation and provides better performance than 3D SPIHT, both objectively and subjectively, despite the fact that the raw number of coefficients resulting from the 3-D DDWT is much more than that of the conventional 3-D DWT. The proposed coder allows full scalability in spatial, temporal and quality dimensions.

Picture Coding Symposium

Video Coding Without Motion Compensation Using a 3-D Dual-tree Wavelet Transform

Beibei Wang[1], Yao Wang[1], Ivan Selesnick[1] and Anthony Vetro[2]

[1] Polytechnic University, Electrical and Computer Engineering Dept, Brooklyn, NY
[2] Mitsubishi Electric Research Laboratories, Cambridge, MA
Email: (bb_w, yao)@vision.poly.edu, selesi@duke.poly.edu, avetro@merl.com

ABSTRACT

This paper explores the use of a recently introduced 3-D dual-tree discrete wavelet transform (DDWT) for video coding. The 3-D DDWT is an attractive video representation because it isolates motion along different directions in separate subbands. However, it is an overcomplete transform with 8:1or 4:1redundancy. Based on the effectiveness of the iterative projection-based noise shaping scheme proposed by Kingsbury on reducing the number of coefficients, and the investigation about the correlation between subbands at the same spatial/temporal location, both in the significance map and in actual coefficient values, a new video coding scheme using 3D DDWT is proposed. The proposed video codec does not require motion compensation and provides better performance than 3D SPIHT, both objectively and subjectively, despite the fact that the raw number of coefficients resulting from the 3-D DDWT is much more than that of the conventional 3-D DWT. The proposed coder allows full scalability in spatial, temporal and quality dimensions.

1. INTRODUCTION

The standard separable discrete wavelet transform (DWT) provides a multi-resolution representation of a signal and has established an impressive reputation for video compression. Several recently proposed DWT-based video coders have achieved coding efficiency similar to or slightly better than block-based hybrid video coders [1]. In addition, such coders provide a scalable representation of the video in spatial resolution, temporal resolution and quality. But the poor directional selectivity of the multidimensional DWT can lead to checkerboard artifacts at the low bit rate range.

This work was supported in part by the National Science Foundation under grant CCF- 0431051.

An important recent development in wavelet-related research is the design and implementation of 2-D multiscale transforms that represent edges more efficiently than does the DWT. Kingsbury's complex dual-tree wavelet transform (DT-CWT) is an outstanding example [2]. The DT-CWT is an overcomplete transform with limited redundancy (2^m :1 for m-dimensional signals). This transform has good directional selectivity and its subband responses are approximately shift-invariant. The 2-D DT-CWT has given superior results for image processing applications compared to the DWT [2, 3]. Recently, Selesnick and Li introduced a 3-D version of the dual-tree wavelet transform and showed that it has superior motion selectivity [4].

The major challenge to apply the 3-D complex DDWT for video coding is it is an overcomplete transform with 8:1 redundancy. In our current study, we chose to retain only the real parts of the wavelet coefficients, which can still lead to perfect reconstruction, while retaining the motion selectivity. This reduces the redundancy to 4:1 [4].

To reduce the number of coefficients, Kingsbury proposed an iterative projection-based noise shaping (NS) scheme [3], which modifies previously chosen large coefficients to compensate for the loss of small coefficients. In our previous work [5], we have found that noise shaping applied to 3-D DDWT can yield a more compact set of coefficients than from the 3-D DWT. The fact that noise shaping can reduce the number of coefficients to below that required by DWT (for the same video quality) is very encouraging.

In [5], the vector entropy study validates our hypothesis that only a few bases have significant energy for an object feature. The relatively low entropy of the significance vector across subbands suggests that the whereabouts of significant coefficients may be coded efficiently by coding the significance bits across subbands jointly. The fact that coefficient values do not have strong correlation among the subbands, on the other hand, indicates that the benefit from vector coding the magnitude bits across the subbands may be limited. Based

on the above investigation, we proposed a video codec in this paper that doesn't require motion compensation and allows full spatial, temporal and quality salability.

This paper is organized as follows. Section 2 briefly introduces the 3-D DDWT and its properties for video representation. Section 3 describes the proposed codec in detail. Section 4 presents the coding results of the video codec using 3-D DDWT. The final section summarizes our work and discusses future work for video coding using 3-D DDWT.

2. 3-D DUAL-TREE WAVELET TRANSFORM

The design and the motion-selectivity of dual-tree filters are described in [6] and [2]. A Daubechies-like algorithm for the construction of Hilbert pairs of short orthonormal (and biorthogonal) wavelet bases yields pairs of bases, which can be used to efficiently implement the motion-selective wavelet transform [6]. The dual-tree wavelet transform is implemented by first applying separable transforms and then combining subband signals with simple linear operations. So even though it is non-separable (and therefore free of some of the limitations of separable transforms), it inherits the computational efficiency of separable transforms.

Figure 1 illustrates the difference between the standard 3-D DWT and the 3-D DDWT. The figure depicts the wavelets (i.e. the basis functions) associated with the 3-D DWT and the 3-D DDWT respectively. As illustrated, the 3-D DWT mixes different orientations in one wavelet basis, but the 3-D DDWT is free of this effect. The 3-D DDWT has many more subbands (a subband refers to the coefficients associated with one wavelet basis) than the 3-D DWT (28 high subbands instead of 7, 4 low subbands instead of 1). Only a subset of the high subbands is drawn in Fig.1 for DDWT. The 28 high subbands isolate 2-D edges with different orientations that are moving in different directions. Because of this motion selectivity, the 3-D DDWT is likely to be substantially more effective for the representation of video than the 3-D DWT.

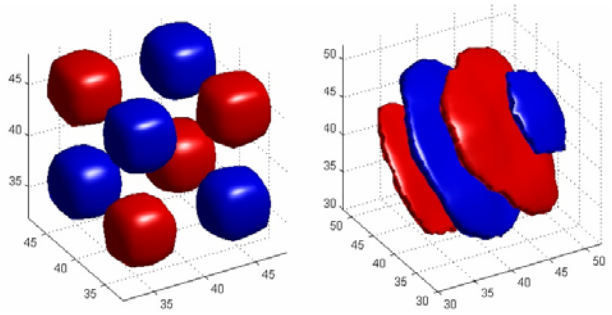


Fig.1 Isosurfaces of a typical 3-D DWT (left) and a typical 3-D DDWT (right). For the 3-D DDWT, each subband corresponds to motion in a specific direction.

A core element common to all state-of-the-art video coders is motion-compensated temporal prediction, which is the main contributor to the complexity as well error-sensitivity of a video encoder. Because the subband coefficients associated with the 3-D DDWT directly capture moving edges in different directions, it may not be necessary to perform motion estimation explicitly. This is our primary motivation for exploring the use of 3-D DDWT for video coding.

One major obstacle for applying the 3-D DDWT for video coding is that it is an overcomplete transform by a factor of eight or four (if only the real parts of the coefficients are retained). However, our previous investigation [5] has shown that after noise shaping, 3-D DDWT needs fewer coefficients than 3-D DWT to achieve the same video quality for all sequences. This result is very encouraging and has prompted us to explore the use of this new transform for video coding.

3. 3-D DUAL-TREE DISCRETE WAVELET TRANSFORM FOR VIDEO CODING (DDWTVC)

The proposed coder (DDWTVC) first applies the noise-shaping method to determine the coefficients to be retained, and then applies a bit plane coder to code the retained coefficients. The low subbands and high subbands are coded separately, each with three parts: significance map coding, sign coding and magnitude refinement.

3.1. Coding of the significance map

Although 3-D DDWT has 28 high subbands, only a few subbands have significant energy for an object feature and the typical combination of significant subbands at the same spatio-temporal location is quite predictable. This has been verified by measuring the entropy of the vector containing the significance bits at the same location across the 28 subbands [5]. The significance bit is either 0 or 1 depending on whether or not the coefficient becomes significant in the current bit-plane. The relatively low vector entropy (much lower than 28) prompted us to apply adaptive arithmetic coding for the significance vector. Though the vector dimension is 28, for each bit plane, only a few patterns appear with high probabilities. To utilize the different statistic information of the high subbands in each bit plane, individual adaptive arithmetic codec is applied for each bit plane separately.

For the four low subbands, vector coding is used to exploit the correlation among the spatial neighbors (2 x 2 regions) and four low subbands. If the coefficient is already significant in a previous bit plane, the corresponding component of the vector is deleted in the current bit plane. After the first several bit planes, the largest dimension is reduced below 10. Although in

theory there are $2^{\wedge 16}$ possible symbols to code, only a few patterns are likely. These significance symbols are coded using adaptive arithmetic coding for each bit plane and different vector sizes individually.

The proposed video coder codes 3-D DDWT coefficients in each scale separately. Our experiments show that 3-D DDWT doesn't have strict parent-children relationship as 3-D DWT does [7]. Noise shaping destroys such a relationship further. So the spatial-temporal orientation trees used in 3-D SPIHT [7] are only applied in the finest stage, which has a lot of zero coefficients.

The vector coding efficiently codes the significance maps across 28 high subbands jointly. But the spatial dependence of significance bits in each subband has not been explored in the current video codec. Recognizing that context-based coding is an effective means to explore such dependence, we have explored the use of context-based arithmetic vector coding. A main difficulty in applying context models is that the complexity grows exponentially with the number of pixels included in the context. Because some subbands have stronger correlation, we don't need to include all 28 high subbands in the context model. The context models using neighboring coefficients in the uncorrelated subbands should be more helpful and informative. We have tested the efficiency of various contexts, which differ in the chosen subbands and spatial neighbors. The study so far however has not yielded significant gain over direct vector coding.

3.2. Coding of the sign information

This part is used to code the sign of the coefficients if they become significant. We have found that the sign bits, in each individual subband and at the same position across different subbands, exhibit substantial statistical dependencies and can be predicted well. We apply arithmetic coding to code the prediction error.

Our experiments show that four low subbands have very predicable signs. This sign property is due to the particular way the 3-D DDWT coefficients are generated. The orthonormal combination matrix of the four separable 3-D DDWT is given in [4] as follows:

$$\psi_a(x,y,z) = 0.5(\psi_1(x,y,z) - \psi_2(x,y,z) - \psi_3(x,y,z) - \psi_4(x,y,z))$$
$$\psi_b(x,y,z) = 0.5(\psi_1(x,y,z) - \psi_2(x,y,z) + \psi_3(x,y,z) + \psi_4(x,y,z))$$
$$\psi_c(x,y,z) = 0.5(\psi_1(x,y,z) + \psi_2(x,y,z) - \psi_3(x,y,z) + \psi_4(x,y,z))$$
$$\psi_d(x,y,z) = 0.5(\psi_1(x,y,z) + \psi_2(x,y,z) + \psi_3(x,y,z) - \psi_4(x,y,z))$$

where $\psi_1, \psi_2, \psi_3, \psi_4$ are real 3-D wavelets defined in [4]. By applying this combination matrix to each subband, the 3-D oriented dual-tree wavelet transform is obtained. Because the low subbands $\psi_1, \psi_2, \psi_3, \psi_4$ are always positive (because they are low-pass filtered values of the original image pixels), ψ_a is almost always negative and other three low subbands are almost all positive. This property of low subbands is used to code low subbands sign information very efficiently.

For high subbands, we have found that the current coefficient tends to have the same sign as its neighbors in the low pass direction, but have the opposite sign to its high pass neighbors. The prediction from the low pass neighbor is more accurate than that from the high pass neighbor.

The coded binary valued symbol is the product of the predicted and real sign bit. To exploit the statistical dependencies, we apply the similar sign context models of 3-D embedded wavelet video (EWV) [8].

3.3. Magnitude Refinement

This part is used to code the value of significant coefficients in the current bit plane. Because only a few subbands have strong correlation based on a prior study [5], the magnitude refinement is done in each subband individually. The context modeling is used to explore the dependence among the neighboring coefficients. Context models similar to the EWV method [8] are applied to 3-D DDWT here.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the coding performance of the proposed video codec using 3-D DDWT. The comparisons are made to 3-D SPIHT [7], which also does not use motion compensation. Only the comparisons of luminance component Y are presented here. All experimental results are obtained by actually running the codec software.

For both DDWTVC and 3-D SPIHT, 3-level wavelet decompositions are applied. The 3-D SPIHT uses the Daubechies (9, 7)-tap filters. For DDWTVC, the Daubechies (9, 7)-tap filters are used at the first level, and Qshift filters in [4] are used below level 1.

4.1. Video quality comparison

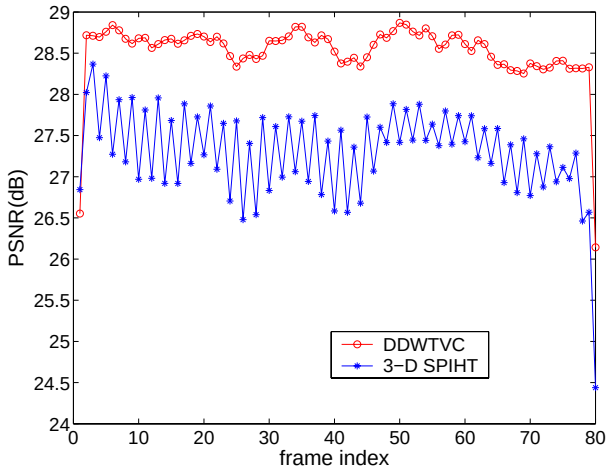
Two CIF video sequences "mobile-calendar" and "Stefan" are used for testing. Both sequences have 80 frames with a frame rate of 30 fps.

Table 1 lists the average PSNR of the two sequences at different bit rates. For a video sequence which has many edges and motions, such like "mobile-calendar", DDWTVC outperforms 3-D SPIHT more than 1 dB. For sequence "stefan", DDWTVC offers 0.2~0.4 dB better PSNR results. Figure 2 plots the PSNR vs. frame index for the two sequences. The figure shows that DDWTVC has more stable PSNR quality from frame to frame, which

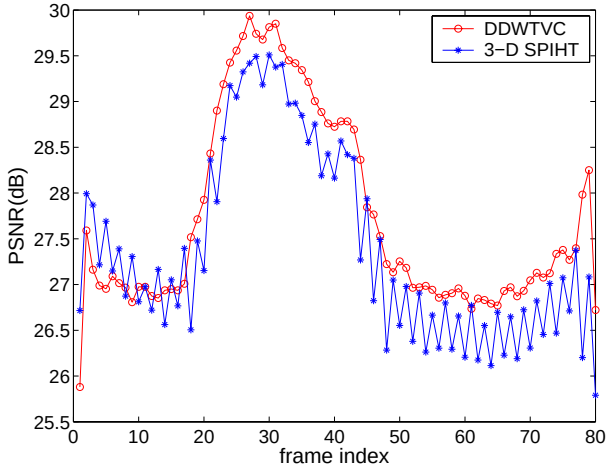
contributed to its better visual quality. Subjectively, DDWTVC has better performance than 3-D SPIHT for both sequences. Considering that 3-D DDWT has four times raw data than 3-D DWT and has the redundancy caused by symmetric extension, the coding results are very promising.

sequence	Mobile-Calendar		
Bitrate (kbps)	730	1000	1424
DDWTVC	25.73	26.87	28.51
3-D SPIHT	24.26	25.47	27.26
sequence	Stefan		
Bitrate (kbps)	730	1000	1424
DDWTVC	26.25	27.62	29.55
3-D SPIHT	25.82	27.29	29.33

Table 1: Average PSNR comparison of DDWTVC and 3-D SPIHT at the same bit rate.



(a) Mobile-Calendar at bit rate 1424 kbps



(b) Stefan at bit rate 1000 kbps

Figure 2: PNSR vs. frame index of two sequences at a given bit rate.

4.2. Scalability of DDWTVC

The proposed video codec using 3-D DDWT allows fully spatial, temporal and quality scalability, although how to design a rate-distortion optimized scalable video coder with noise shaping is an open research topic. Results in table 1 are obtained by choosing the best noise-shaping threshold among a chosen set, for each target bit rates. Specifically, the candidate thresholds are 128, 64, 32, and so on. Our experiments demonstrate that at low bit rate (less than 1Mbps for CIF), the coefficients set retained by noise shaping threshold 128 offers best results, and threshold 64 works best when bit rate is between 1 and 2 Mbps. If bit rate is above 2 Mbps, the codec uses coefficients obtained by threshold 32.

5. CONCLUSION

In this paper, a new video codec using the novel 3-D Dual-tree wavelet transform is proposed and tested on standard video sequences. The 3-D DDWT video codec (DDWTVC) applies adaptive vector arithmetic coding across subbands to efficiently code the significance bits jointly. The use of context models to explore the spatial dependence in each subband is also considered and is still under investigation. Both the sign and magnitude information for significant coefficients is coded using context-based arithmetic coding within each subband, with the sign bits coded predicatively, based on the properties of the dual-tree wavelet transform.

In terms of future work, the spatial dependence in each subband needs to be further explored to improve the coding efficiency. Another challenging open research problem is to design a rate-distortion optimized scalable video coder, so that each additional coefficient offers a maximum reduction in distortion without modifying the previous coefficients. For 3-D DDWT, the lifting steps [9] which map integers to integers may further reduce the coded data and improve the computation speed.

6. REFERENCE

[1] S-T Hsiang and J. W. Woods, "Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank", *Signal Processing: Image Communications,* vol. 16, pp. 705-724, May 2001

[2] N.G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", *Applied Computational Harmonic Anal,* vol. 10, no. 3, pp. 234-253, May 2001

[3] T. H. Reeves and N. G. Kingsbury, "Overcomplete image coding using iterative projection-based noise shaping", *ICIP 02,* Rochester, NY, Sept 2002.

[4] I. W. Selesnick, and K. Y. Li, "Video denoising using 2D and 3D dual-tree complex wavelet transforms", *Wavelet Appl Signal Image Proc. X (Proc. SPIE 5207)*, Aug 2003.

[5] B. Wang, Y. Wang, I. W. Selesnick and A. Vetro, "An investigation of 3D Dual-Tree wavelet transform for video coding", to appear in ICIP 04.

[6] I. W. Selesnick, "The design of approximate Hilbert transform pairs of wavelet bases", *IEEE Trans. on Signal Processing,* 50(5): 1144-1152, May 2002.

[7] B-J Kim and W.A. Pearlman, "An embedded wavelet video coder using three-dimensional set partitioning in hierarchical trees (SPIHT)", *Data Compression Conference, 1997. DCC '97. Proceedings*

[8] J. Hua; Z. Xiong; X. Wu, "High-performance 3-D embedded wavelet video (EWV) coding", *Multimedia Signal Processing, 2001 IEEE Fourth Workshop* on, 3-5 Oct. 2001

[9] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps", *J. Fourier Anal. Appl.*, 4 (no. 3), pp. 247-269, 1998.