1-1-2007

# Video coding with 3D wavelet transforms

Joseph Bradley Boettcher

# VIDEO CODING WITH 3D WAVELET TRANSFORMS

By

Joseph B. Boettcher

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Engineering
in the Department of Electrical & Computer Engineering

Mississippi State, Mississippi

December 2007

VIDEO CODING WITH 3D WAVELET TRANSFORMS

By

Joseph B. Boettcher

Approved:

| | |
|---|---|
| Nicholas H. Younan<br>Professor of Electrical<br>& Computer Engineering<br>(Graduate Program Director) | Roger L. King<br>Professor of Electrical<br>& Computer Engineering<br>(Associate Dean) |
| James E. Fowler<br>Professor of Electrical<br>& Computer Engineering<br>(Director of Thesis) | Lori M. Bruce<br>Professor of Electrical<br>& Computer Engineering<br>(Committee Member) |
| Susan M. Bridges<br>Professor of Computer Science & Engineering<br>(Committee Member) | |

Name: Joseph B. Boettcher

Date of Degree: December 14, 2007

Institution: Mississippi State University

Major Field: Engineering (Computer Engineering)

Major Professor: Dr. James E. Fowler

Title of Study: VIDEO CODING WITH 3D WAVELET TRANSFORMS

Pages in Study: 40

Candidate for Degree of Master of Science

Video coding systems based on 3D wavelet transforms offer several advantages over traditional hybrid video coders. This thesis proposes two 3D wavelet-based video-coding approaches. In the first approach, motion compensation with redundant-wavelet multihypothesis, in which multiple predictions that are diverse in transform phase contribute to a single motion estimate, is deployed into the fully scalable MC-EZBC video coder. The bidirectional motion-compensated temporal-filtering process of MC-EZBC is adapted to the redundant-wavelet domain, wherein transform redundancy is exploited to generate a phase-diverse multihypothesis prediction of the true temporal filtering. In the second approach, a video coder is proposed that does not perform motion compensation explicitly, instead relying on the motion-selective characteristics of the 3D dual-tree discrete wavelet transform to isolate moving features. The transform coefficients are coded with binary set-partitioning using $k$-d trees in an algorithm that exploits within-subband spatiotemporal coherency as well as cross-subband correlation to achieve efficient coding.

# ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. James E. Fowler, whose enthusiastic and dedicated approach towards scientific research inspired and motivated me throughout my graduate studies. His guidance and support over the course of my work were of invaluable help. I would also like to thank my colleagues Justin Rucker and Kristen Parker for all their help over the years. Finally, I would like to thank my family and friends for their constant support and encouragement.

TABLE OF CONTENTS

# LIST OF TABLES

LIST OF FIGURES

FIGURE Page

# CHAPTER I

# INTRODUCTION

As the role of digital video in modern technology continues to grow, there is an increasing demand for more efficient video-compression techniques. The traditional hybrid video-coding architecture, as exhibited in the MPEG standards, involves a two-stage approach in which a video signal is first temporally decorrelated via block-based motion compensation, then spatially decorrelated by a 2D spatial transform. This architecture requires a feedback loop in which frames that are reconstructed from a motion estimate are used as reference frames in the motion prediction of subsequent frames. Although this approach has been effective, it has several drawbacks. The feedback loop creates drift, allowing noise in the reconstructed frames to propagate. In addition, the presence of the feedback loop impedes scalability—the ability to produce video sequences at varying bit rates, frame rates, or spatial resolutions from a single encoded bitstream.

As an alternative to the traditional motion-compensation feedback loop, temporal redundancy can be removed from a video signal by applying a wavelet transform in the temporal direction. Thus, the use of 3D wavelet transforms, in which a 2D spatial wavelet transform is combined with a wavelet transform in the temporal direction, has become a popular approach to achieving highly scalable video compression. However, for the temporal wavelet decomposition to produce subbands of beneficial quality, motion compensation is still needed to prevent filtering across dissimilar regions in the

temporal signal. Motion-compensated temporal filtering (MCTF) solves this problem by using a motion estimate to guide the temporal transform in the direction of predicted motion. Accurate motion estimation is important to the success of MCTF, since filtering across poorly matched regions can result in low-quality temporal subbands with "ghosting" artifacts [1].

Since uncertainty is inherent in motion estimation, many video-coding systems use a combination of motion predictions, a concept known as multihypothesis motion compensation (MHMC) [2]. The MC-EZBC coder [3], which uses an MCTF approach and represents the state-of-the-art in fully scalable video coding, employs two forms of MHMC. First, the motion-estimation procedure operates with fractional-pixel accuracy, a form of spatial-diversity MHMC made possible by the lifting implementation of MCTF. In addition, MC-EZBC uses bidirectional MCTF, a form of temporal-diversity MHMC in which the multiple motion predictions come from different frames. However, it is possible to improve the MC-EZBC system with the addition of a third form of MHMC—multihypothesis prediction via transform-phase diversity. The first contribution of this thesis is the proposal of a system in which redundant-wavelet multihypothesis (RWMH) [4, 5] is embedded within the MC-EZBC framework. Taking place in the redundant wavelet domain, MCTF in the proposed system benefits from multiple motion predictions that are diverse in transform phase. Results from this work, first presented in [6], indicate that RWMH provides more accurate motion compensation for the MCTF process, leading to higher-quality temporal subbands and more efficient coding.

While most video-coding systems based on 3D wavelet transforms rely on a motion compensation procedure to guide the temporal transform, there is reason to consider avoiding motion compensation altogether. For example, block-based motion compensation can result in the appearance of visual artifacts in the reconstructed

video frames around block boundaries. Furthermore, motion-estimation procedures are usually the heaviest computational burden on the encoder. An alternative to MCTF has arisen recently in the form of the complex dual-tree discrete wavelet transform (DDWT) [7–9]. The DDWT is a redundant transform that, in the 3D case [9], produces four times as many subbands as the traditional critically-sampled discrete wavelet transform (DWT), with each subband oriented in a different spatiotemporal direction. When applied to a video signal, these orientations help isolate image features moving in different directions, providing inherent motion selectivity. The ability of the transform to describe motion without explicit motion estimation or compensation has motivated the use of the DDWT in video-coding systems [10–12] looking to avoid the computational complexity associated with motion estimation. However, since the 3D DDWT is four times redundant, efficient coding of the transform coefficients is a challenging task.

The coder proposed in [11, 12]—the DDWT video coder (DDWTVC)—exploits the fact that, although the DDWT is greatly redundant, there is a significant degree of correlation between coefficients residing at the same spatiotemporal locations in different subbands. The DDWTVC uses arithmetic coding of cross-subband vectors of coefficients to exploit this cross-subband correlation. However, large-magnitude DDWT coefficients typically occur rather sparsely in any given DDWT subband, with most of the coefficients being small or zero (i.e., insignificant coefficients). In fact, it has proven beneficial to apply a "noise-shaping" procedure [8] to deliberately increase the sparsity of the transform coefficients [10, 12]. Yet, DDWTVC does not explicitly exploit the fact that the insignificant coefficients tend to form spatiotemporally coherent regions within each subband. The second contribution of this thesis is the proposal of a video-coding system using the DDWT in conjunction with an embedded wavelet-based coder called binary set-splitting with $k$-d trees (BISK) [13–15]. This DDWT-BISK algorithm uses set-partitioning to exploit not only cross-subband correlation but also

spatiotemporal coherency within subbands to effectively represent the sparse coefficient volume. As experimental results have shown [16, 17], the resulting system outperforms other video coders, including DDWTVC, that do not perform explicit motion estimation or compensation. This work has inspired the extension of the DDWT into applications involving the coding of both still images [18] and hyperspectral imagery [19], although the focus in this thesis will be limited to video-coding applications.

The next chapter presents a review of both MCTF and complex wavelet transforms, additionally providing a more in-depth look at the MC-EZBC and DDWTVC video coders. Chap. III includes a discussion of RWMH, afterwhich the RWMH-EZBC video coder is described in detail and experimental results are presented. In Chap. IV, the proposed DDWT-BISK video coder is presented, along with experimental results. Finally, Chap. V offers conclusions drawn from the results of the experiments.

# CHAPTER II

# BACKGROUND

Over the past few decades, the use of wavelet transforms in image and video compression has become more prevalent. The ability of the wavelet transform to preserve the spatial structure of a signal while partitioning the frequency information into lowpass and highpass subbands makes it well-suited for the processing of signals with visual information. Since the majority of image-signal energy can be represented by low-frequency coefficients, wavelet transforms facilitate the embedded coding of images and video, where the most important information is coded first, and the detail information is added to the bitstream in successive passes. Embedded coding provides the means for scalability in video, since an embedded bitstream can be partially decoded to reconstruct a video sequence at varying quality levels, spatial resolutions, or temporal resolutions.

In this chapter, we will review two approaches to scalable video coding with 3D wavelet transforms. First, we will discuss motion-compensated temporal filtering, followed by an overview of the MC-EZBC video coder which uses the MCTF approach. Secondly, we will look at complex wavelet transforms, which will lead into a discussion of the DDWT video coder.

## 2.1 Motion-Compensated Temporal Filtering (MCTF)

While 3D transforms can be used in the coding of a variety of 3D data types, including volumetric medical imagery and hyperspectral remote-sensing imagery, their

use in video coding presents unique challenges. Video sequences often contain objects that move from frame to frame, creating high-frequency "edges" in the temporal signal which result in large coefficients in the temporal transform. To combat this problem, conventional motion-compensation techniques have been used to determine motion fields between frames so that the temporal transform can be applied along the path of motion rather than simply filtering between co-located pixels in each frame.

The basic approach to motion-compensated temporal filtering (MCTF), proposed by Ohm [20, 21], relies on block-based motion estimation, in which blocks of pixels in the current video frame are matched against blocks within a search window in the previous frame. The spatial displacement between blocks in the current frame and their closest match in the reference frame (based on some criteria, such as mean absolute error) is mapped by a set of motion vectors. Once the motion field is determined, a 2-tap wavelet filter applied along the motion trajectories will transform the current/reference frame pair into a lowpass/highpass frame pair. If the motion estimation is accurate, the resulting lowpass frame will contain most of the signal energy, while the highpass frame will contain mostly small-valued coefficients.

Because there is often overlap between blocks in the current frame and their best matches in the previous frame, particularly in cases where occluded objects in one frame are revealed in the next, there will not be a one-to-one connection for every pixel in the frame pair. In this case, special care must be taken when processing the "unconnected" pixels, that is, those that do not have a one-to-one motion-field connection. Two types of unconnected pixels exist: pixels in the reference frame that are not matched to any pixels in the current frame, and pixels in the current frame that are matched to pixels already used as a reference. Most commonly, filtering is performed on only connected pixels, while the other pixels are processed separately. For the unconnected pixels in the reference frame, their scaled values are placed in the lowpass frame after MCTF. For

the pixels in the current frame that are matched to pixels already used as a reference, the scaled difference between the current value and the reference value is placed in the highpass frame. Fig. 2.1 illustrates the MCTF process between two frames with unconnected pixels.

## 2.2 MC-EZBC

In [22], Hsiang and Woods introduced the MC-EZBC video coder, a fully scalable video coder employing block-based, spatial-domain MCTF. While the originally proposed coder was limited to unidirectional motion compensation, Chen and Woods later extended the MC-EZBC coder with bidirectional motion compensation in [3]. That is, when determining the motion field for a frame of video, both the preceeding and subsequent frames can be used as a reference. With the addition of bidirectional motion compensation, MC-EZBC yields state-of-the-art performance in scalable video coding.

In the MC-EZBC system, motion compensation is carried out via hierarchical, variable-size block matching (HVSBM) in which motion vectors are determined for large blocks of pixels in low-resolution frames, then refined for smaller sub-blocks at successively higher levels of detail. After the motion field is determined, the unconnected pixels are located and processed separately. To avoid inefficient temporal filtering, MCTF is performed only if fewer than half the pixels between a pair of consecutive frames are unconnected. If this criterion is met, then the frames are temporally filtered, resulting in highpass and lowpass temporal subbands. This process takes place for each pair of frames in a group of pictures (GOP), after which it is performed recursively on the lowpass temporal subbands, as depicted in Fig. 2.2. After MCTF is complete, the resulting temporal subbands go through spatial wavelet analysis, completing the 3D decomposition. Finally, the 3D subbands are encoded with the EZBC coder [23].

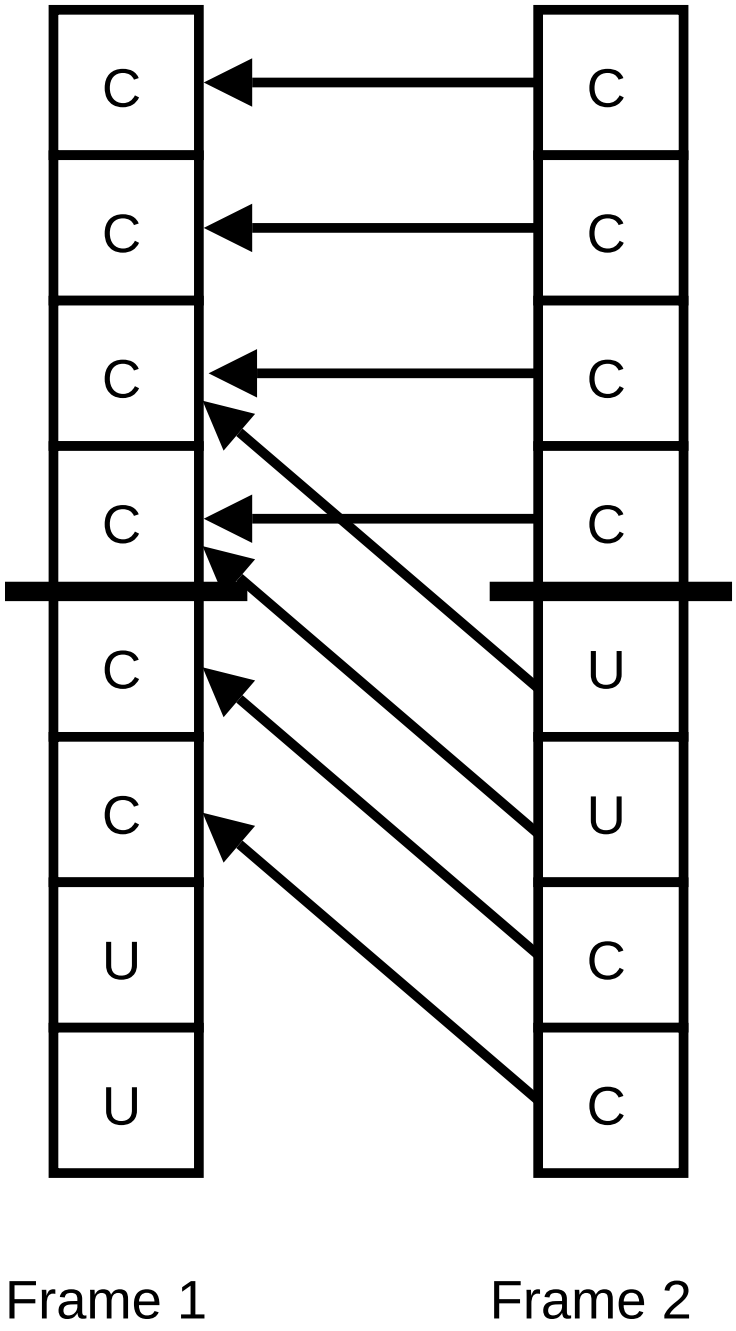



Figure 2.1: Designation of unconnected pixels (labeled ‘u’) after backward motion estimation in MCTF.

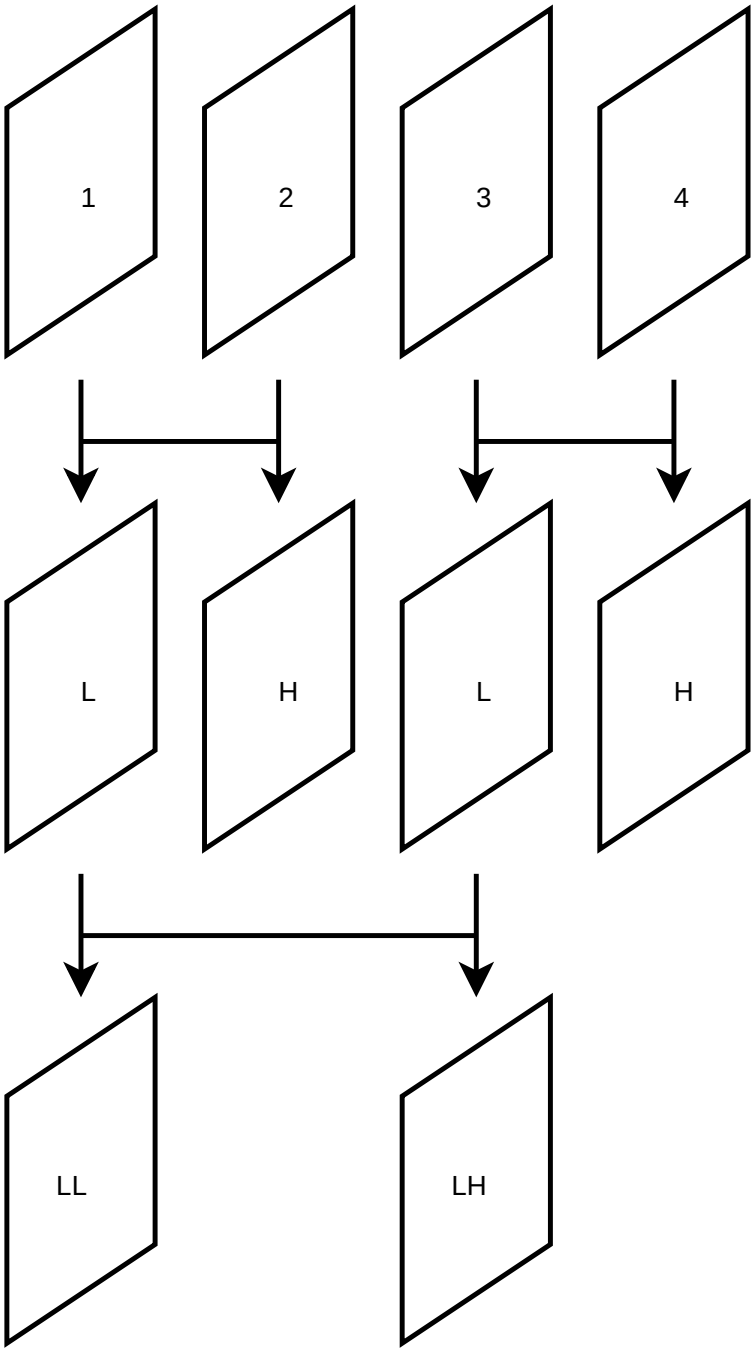



Figure 2.2: MCTF operating on a GOP by recursively processing lowpass frames.

## 2.3 Complex Wavelet Transforms

In order to overcome some of the shortcomings of the traditional critically-sampled DWT, Kingsbury [7] introduced the DDWT consisting of two trees of real wavelet filters operating on the same data in parallel, with the filters designed such that the two trees produce the real and imaginary parts of the complex-valued coefficients.While the DWT lacks shift invariance, the DDWT is approximately shift invariant and offers higher directional selectivity. However, $2^m$:1 redundancy is added for an $m$-dimensional signal.

Selesnick and Li [9] developed a 3D version of the DDWT to provide a useful representation for video. It turns out that the degree of redundancy can be reduced without sacrificing perfect reconstruction by simply discarding the complex parts of the coefficients, resulting in 4:1 redundancy. For this real-valued transform, four separable 3D DWTs based on Hilbert pairs are applied to the original signal, and only the real parts of the coefficients are retained. The four sets of transform data are then combined with linear operations to produce subbands that isolate features in a variety of orientations. The resulting DDWT subbands are arranged in four separate *transform combinations* with each combination having the same subband organization as would a 3D DWT of the original data but with each combination containing subbands of different orientation. For example, a 3D dyadic DWT consists of 7 highpass subbands at each resolution level, plus a single baseband at the lowest resolution. Consequently, at each resolution level, the corresponding 3D DDWT consists of $4 \times 7 = 28$ highpass subbands, except at the lowest resolution, which contains 4 baseband subbands.

Although the 3D DDWT produces four times the data that the 3D DWT does, the DDWT requires fewer critical coefficients to efficiently represent the underlying signal [10]. With this in mind, Reeves and Kingsbury [8] proposed deliberately reducing the

number of DDWT coefficients by discarding small-magnitude coefficients and refining the remaining coefficients to compensate. This "noise-shaping" procedure is an iterative projection of signals between the original-signal domain and the DDWT domain. On each iteration, the signal is thresholded in the DDWT domain to remove small coefficients, and the remaining coefficients are compensated by the original-signal-domain error induced by the thresholding. This noise-shaping procedure increases the sparsity of the representation to the point that the 3D DDWT typically requires fewer non-zero coefficients than the 3D DWT to achieve the same level of reconstruction quality for a video signal [10]. The DDWTVC coder [11, 12] extends this observation to actual coding results that include quantization and entropy coding.

## 2.4 DDWTVC

Because the 3D DDWT provides inherent motion-selectivity, Wang *et al.* [11, 12] developed the DDWTVC coder to avoid explicit motion estimation and compensation. DDWTVC is a bitplane coder that exploits cross-subband redundancy in the highpass bands of the 3D DDWT coefficients. While the 3D dyadic DWT results in 7 highpass subbands per level of decomposition, the corresponding 3D DDWT produces 28 highpass bands per level, due to the 4:1 redundancy of the transform. After the noise shaping of [8] is applied to the original video sequence, the significance states of the co-located coefficients in each of the 28 highpass bands are coded as a 28-bit vector using adaptive arithmetic coding in each bitplane separately. For the four lowpass bands, a 16-bit significance vector comprising $2 \times 2$ blocks of coefficients co-located in each of the bands is coded at the first bitplane; however, with each successive bitplane, previously significant coefficients are removed from the vector so that the dimensionality decreases. The sign information for the coefficients is predicted, and the prediction error is coded. Arithmetic coding of both the sign-prediction error and the

magnitude-refinement information is performed with context models in each subband individually. Wang *et al*. showed that the DDWTVC system exhibits rate-distortion performance superior to that of 3D-SPIHT [24] applied directly to the video sequence with no motion estimation or compensation.

In this chapter we have reviewed the work that serves as the foundation for the contributions of this thesis. In the next chapter, we will present a video coder that improves upon the MC-EZBC video-coding approach by adapting MCTF to the redundant-wavelet domain.

# CHAPTER III

# RWMH-EZBC

For an MCTF-based video-coding system to be effective, it is crucial that the motion estimation-procedure provides an accurate prediction of motion. By utilizing multiple predictions in determining the motion field, systems employing multi-hypothesis motion compensation can achieve a more accurate motion estimate. In this chapter, a video coder is introduced in which redundant wavelet multi-hypothesis (RWMH) is embedded within the MC-EZBC framework. We start with an examination of RWMH in Sec. 3.1, followed by a description of the proposed RWMH-EZBC video coder in Sec. 3.2. Lastly, experimental results produced by the video coder, which were first published in [6], are discussed in Sec. 3.3.

## 3.1 Redundant-Wavelet Multihypothesis (RWMH)

The redundant discrete wavelet transform (RDWT) [25, 26] is an approximation to the continuous wavelet transform that removes the downsampling operation from the traditional critically sampled discrete wavelet transform (DWT) to produce an overcomplete representation. The well-known shift variance of the DWT arises from its use of downsampling, while the RDWT is shift invariant since the spatial sampling rate is fixed across scale. Numerous RDWT-based video-coding systems have been developed, originating with the work of Park and Kim [27]. In most of these systems, the redundancy inherent in the RDWT is used exclusively to permit motion estimation and compensation in the wavelet domain by overcoming the shift variance of the

critically sampled DWT. In [4], an entirely new use for the redundancy in the RDWT was presented; specifically, transform redundancy was employed to yield multiple predictions of motion that were combined into a single multihypothesis prediction. This approach represented a new paradigm in MHMC wherein diversity in transform phase yields multihypothesis predictions that enhance motion-compensation performance.

Fig. 3.1 shows how a $J$-scale RDWT can be considered to be composed of $4^J$ distinct critically sampled transforms, each corresponding to the choice between even- and odd-phase subsampling in both the horizontal and vertical directions at each scale of decomposition. In the RWMH paradigm, wherein motion estimation and compensation take place in the redundant-wavelet domain, each one of these critically sampled transforms "views" motion from a different perspective and thus forms an independent hypothesis of the true motion of the video sequence. After motion compensation is complete, a multiple-phase inverse RDWT combines these multiple hypotheses into a single prediction.

In [4], a video-coding system is described that incorporates RWMH into the motion-compensation feedback loop of the traditional hybrid, block-based video-coding architecture. An in-depth analysis [5] of this hybrid RWMH architecture reveals that the performance gains over single-phase prediction are largely based on the ability of RWMH to reduce the variance of the prediction residual. That is, noise in the RDWT domain undergoes a substantial reduction in variance when the multiple-phase inverse RDWT is applied, which is due to the well-known fact that the inverse RDWT is a pseudo-inverse operation and thereby consists of a projection onto the range space of the forward transform. Consequently, noise not captured by the motion model is greatly reduced in the hybrid RWMH system, leading to substantial reduction in the variance of the prediction residual in the motion-compensation feedback loop and higher coding efficiency. Additionally, in [28, 29], the RWMH concept was introduced into a general,

mesh-based MCTF framework to produce a fully scalable video coder (3D-RWMH). In this thesis, RWMH is deployed into the block-based MCTF framework of MC-EZBC, producing the proposed RWMH-EZBC system.

## 3.2 The RWMH-EZBC System

In the MC-EZBC system, motion estimation and temporal filtering take place with the video frames in the original spatial domain, resulting in a single-phase prediction of motion. In the RWMH-EZBC system, MCTF is instead performed in the redundant-wavelet domain in order to generate multiple predictions of motion that are diverse in transform phase. The block diagram for the encoder of the RWMH-EZBC system is shown in Fig. 3.2.

In Fig. 3.2, each frame of the input GOP is decomposed with a spatial RDWT, and the resulting frames of RDWT coefficients are used in a bidirectional block-matching motion-vector search. In a $J$-scale RDWT decomposition, each $B \times B$ block in the original spatial domain corresponds to $3J + 1$ blocks of the same size, one in each subband. We call the collection of these co-located blocks a $set$; each set contains all the different phases of RDWT coefficients. In the motion-estimation procedure, block matching is used to determine the motion of each set as a whole. As in MC-EZBC, HVSBM is used for motion estimation, with an additional cross-subband distortion measure as the matching criterion. Absolute errors for each block of the set are summed such that the coefficients from all phases in both the current and reference frames contribute to the distortion measurement. Specifically, the motion vector for the set located at $[x, y]$ is

$$(d_x, d_y) = \arg \min_{-W \leqslant d_x, d_y \leqslant W} \mathrm{MAE}(x, y, d_x, d_y), \tag{3.1}$$

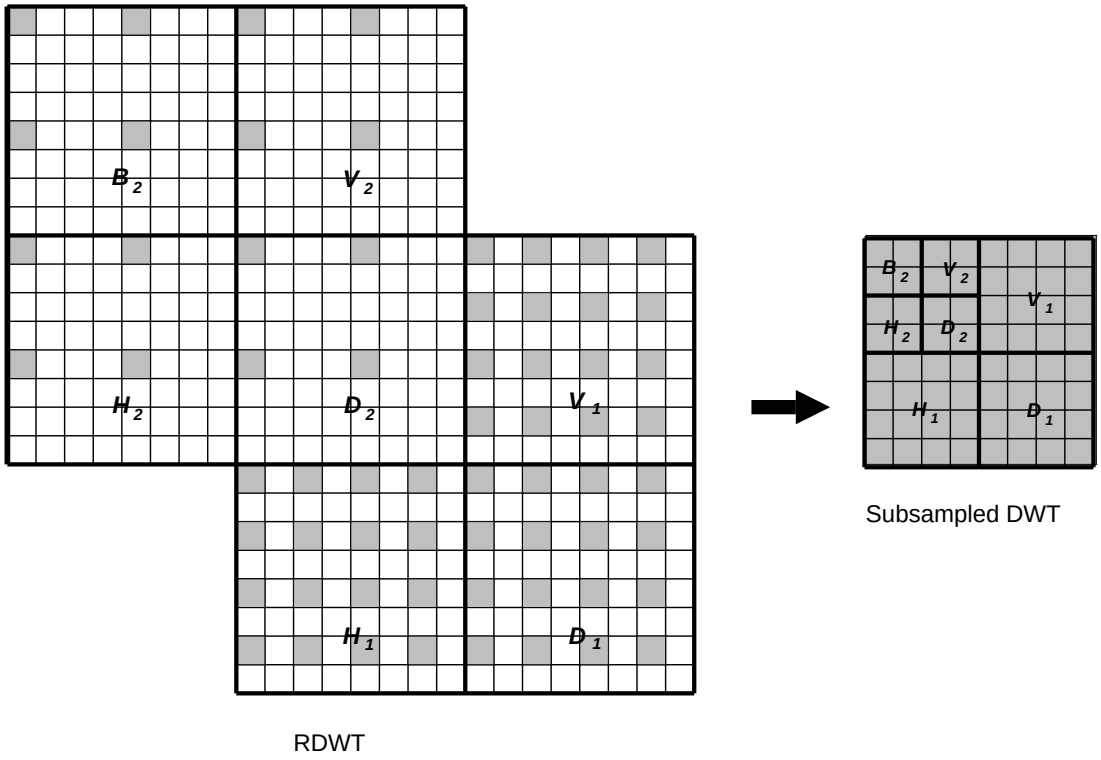



Figure 3.1: Two-scale 2D RDWT, with an example of subsampling recovering one of the $4^J$ critically sampled DWTs. $B_j$, $H_j$, $V_j$, and $D_j$ denote the baseband, horizontal, vertical, and diagonal subbands, respectively, at scale $j$.

where the mean absolute error (MAE) is

$$\mathrm{MAE}(x, y, d_x, d_y) = \frac{1}{B^2} \sum_{k=0}^{B-1} \sum_{l=0}^{B-1} \mathrm{AE}(x + k, y + l, d_x, d_y), \tag{3.2}$$

and the absolute error (AE) is

$$\begin{aligned} \mathrm{AE}(x, y, d_x, d_y) = 2^{-J} \Big| B_J[x, y, t] - B_J[x + d_x, y + d_y, t - 1] \Big| + \\ \sum_{j=1}^{J} 2^{-j} \Bigg( \Big| V_j[x, y, t] - V_j[x + d_x, y + d_y, t - 1] \Big| + \\ \Big| H_j[x, y, t] - H_j[x + d_x, y + d_y, t - 1] \Big| + \\ \Big| D_j[x, y, t] - D_j[x + d_x, y + d_y, t - 1] \Big| \Bigg). \end{aligned} \tag{3.3}$$

In the above equations, $B_j$, $H_j$, $V_j$, and $D_j$ are the baseband, horizontal, vertical, and diagonal RDWT subbands, respectively, at scale $j$. A window $[-W, W]$ is used for the block search.

After the motion field is generated, the number of unconnected pixels between two consecutive frames is calculated. If fewer than half the pixels are unconnected, then temporal filtering takes place between the RDWT frames, with each subband using the same motion field for motion compensation. The same process is carried out for each pair of frames in the GOP, after which it is performed recursively on the lowpass temporal subbands. Once the temporal decomposition of the GOP is complete, an inverse spatial RDWT is performed on each temporal subband, transforming the coefficients back into the spatial domain. Since each RDWT phase forms an independent hypothesis about the temporal filtering based on its unique perspective, the inverse RDWT implicitly combines these hypotheses into a multihypothesis estimate of what the true temporal filtering should be. At this point, the MC-EZBC system continues as

usual, with a 2D spatial wavelet decomposition of the temporal subbands followed by EZBC encoding of the resulting spatio-temporal coefficients.

## 3.3 Experimental Results

In these experiments, the grayscale video sequences shown in Table 3.1 are coded with both the MC-EZBC and RWMH-EZBC systems. Both systems use Haar filters for bidirectional MCTF, while RWMH-EZBC uses the popular 9-7 biorthogonal filter with symmetric extension to perform the spatial RDWT. A GOP size of 8 frames was used, allowing up to 3 levels of temporal decomposition if MCTF is performed at every level. Additionally, the RWMH-EZBC system uses a 1-level spatial RDWT decomposition. All sequences were coded with quarter-pixel motion accuracy.

Average PSNR results for all the test sequences at a fixed rate are provided in Table 3.1. In each sequence, the multihypothesis MCTF employed by the RWMH-EZBC system provided performance gains over MC-EZBC, albeit in varying degrees. The greatest gains were witnessed for sequences with fast or complex motion, such as the "Football" sequence, for which the average PSNR improved on the order of 0.5 dB. For sequences with little motion, such as the "Susie" sequence, the performance gains were not as substantial. The rate-distortion curves in Figs. 3.3–3.5 indicate that these observations hold over a range of rates.

In this chapter, we have discussed a video-coding system that enhances the MCTF process of the MC-EZBC coder by generating multiple motion predictions that are diverse in transform phase. In the next chapter, we present an alternative approach to wavelet-based video coding in which motion compensation is avoided altogether by relying on the motion-selective properties of the 3D dual-tree discrete wavelet transform.

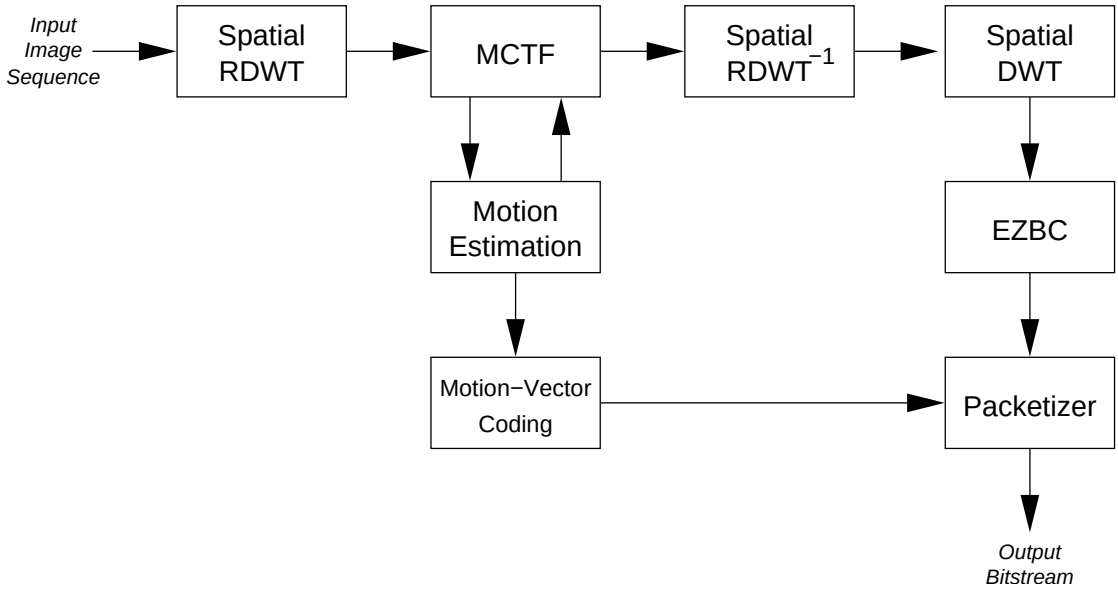



Figure 3.2: Block diagram of the RWMH-EZBC video-coding system.

Table 3.1: A performance comparison of RWMH-EZBC against MC-EZBC. Distortion averaged over all frames of the sequence for rate of 0.5 bpp.

| | PSNR (dB) | |
|---|---|---|
| | MC-EZBC | RWMH-EZBC |
| Football† | 29.7 | 30.2 |
| Table Tennis | 36.1 | 36.4 |
| Foreman | 39.6 | 40.0 |
| Susie† | 42.9 | 43.0 |
| Coastguard | 33.5 | 33.6 |
| NYC† | 40.4 | 40.6 |

Sequences are CIF ($352 \times 288$) at 30 Hz except †, SIF ($352 \times 240$).

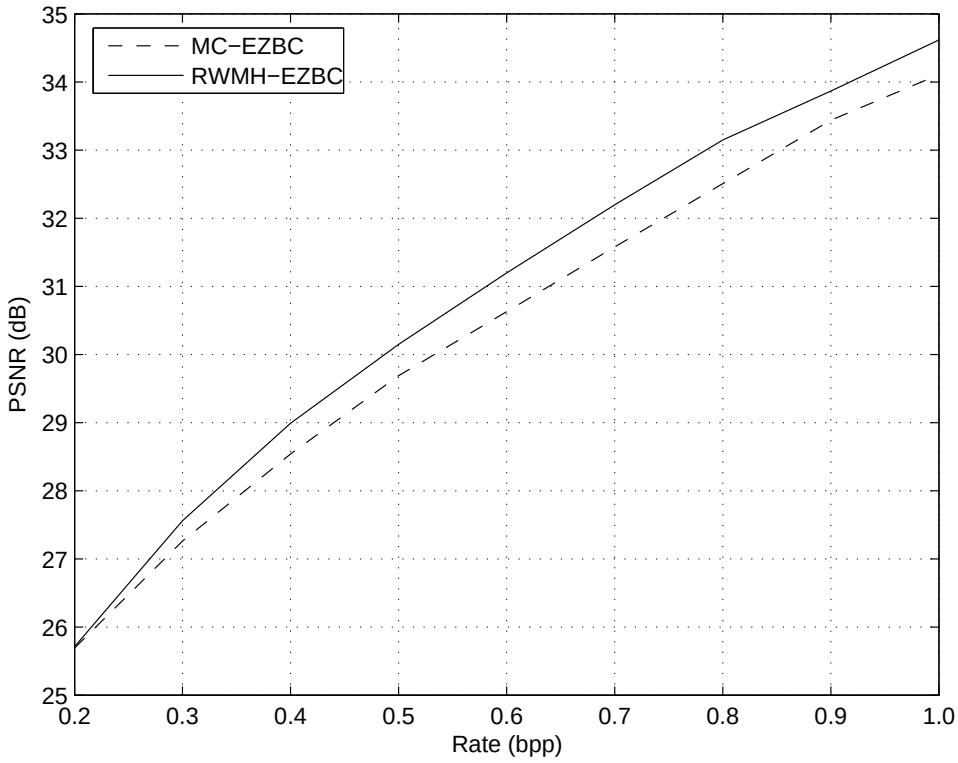



Figure 3.3: RWMH-EZBC and MC-EZBC rate-distortion performance for “Football” at 1/4 pixel accuracy

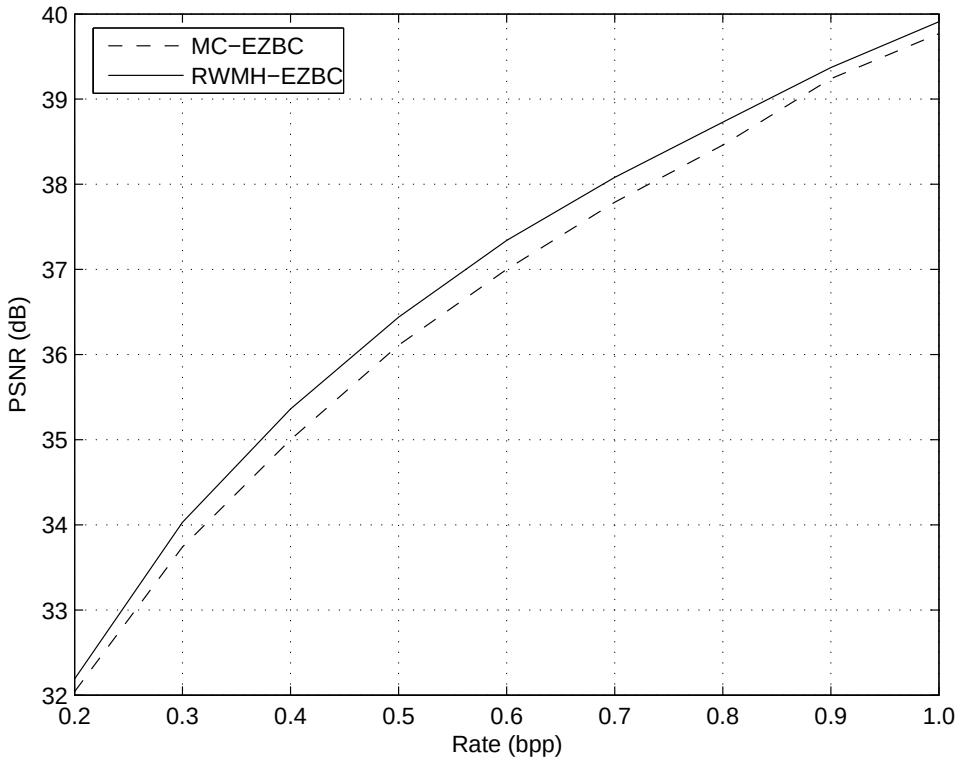



Figure 3.4: RWMH-EZBC and MC-EZBC rate-distortion performance for “Table Tennis” at 1/4 pixel accuracy

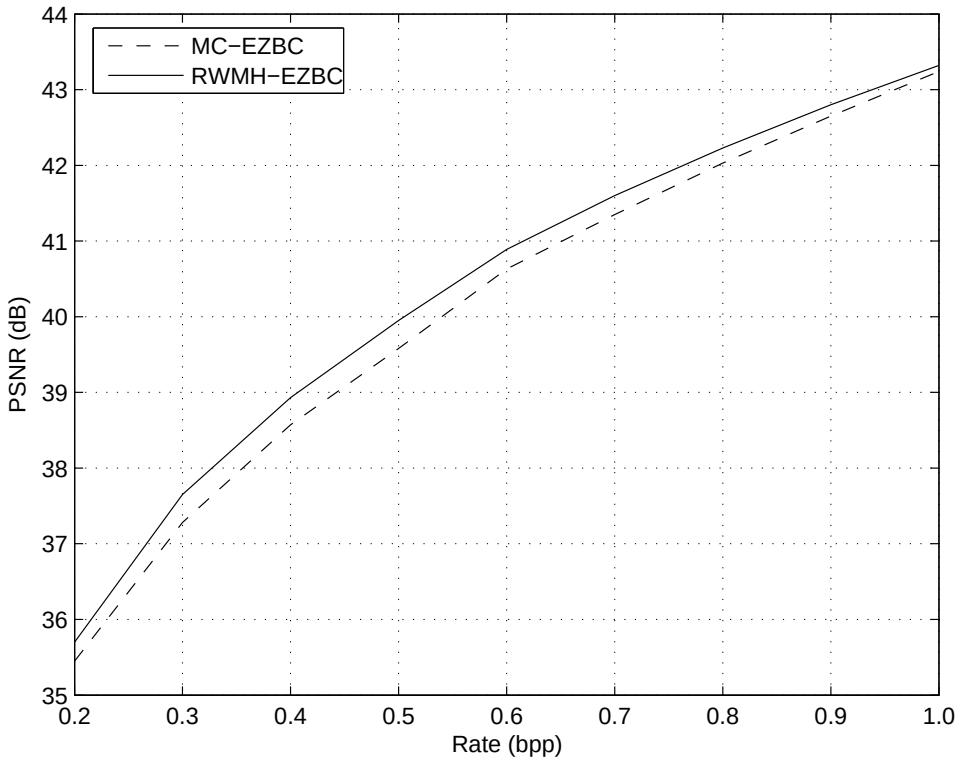



Figure 3.5: RWMH-EZBC and MC-EZBC rate-distortion performance for “Foreman” at 1/4 pixel accuracy

# CHAPTER IV

# DDWT-BISK

While the use of motion compensation can greatly improve the temporal decorrelation of a video signal, it does have some drawbacks. Motion estimation procedures are often the most computationally complex part of a video-coding system. Furthermore, some types of motion compensation can introduce artifacts into the reconstructed video. Recent research [9, 10] has shown that the directionally oriented subbands of the 3D DDWT can be used to isolate moving edges in video data, leading to the development of 3D DDWT based video coders that do not perform explicit motion compensation, such as the DDWTVC coder [11, 12]. In this chapter, a video coder is proposed that combines the 3D DDWT with the BISK embedded wavelet coding algorithm [13–15]. Next, an in-depth discussion of the DDWT-BISK video coder is presented in Sec. 4.1, afterwhich experimental results, previously published in [16, 17], are provided in Sec. 4.2.

## 4.1 The DDWT-BISK System

In the DDWTVC coder [11, 12], correlation between transform coefficients is exploited *across* subbands in the form of 28-bit significance vectors; however, the spatiotemporal coherency of insignificant-coefficient regions *within* a given subband is not exploited despite the fact that this coherency must necessarily be substantial due to the sparsity ensured by the noise-shaping process. In order to efficiently code spatiotemporally coherent regions of DDWT coefficients, the BISK algorithm [13–

15] can be modified to operate on the redundant coefficient set. BISK performs bitplane coding in which significant coefficients are located by recursive spatiotemporal partitioning. Specifically, $k$-d trees are used to split sets of coefficients into two subsets of roughly equal size. Once a significant coefficient is located, its sign information is coded, and its magnitude is refined on successive passes. Significance, sign, and magnitude-refinement information are all coded with adaptive arithmetic coding.

In the proposed DDWT-BISK coder, the noise shaping of [8] is applied to produce the sparse DDWT coefficients. Then, a modified version of the 3D-BISK algorithm [14, 15] operates on the transform coefficients to produce the final coded bitstream. First, coefficients are grouped into 4-dimensional vectors, where each vector consists of the four coefficients at the same spatiotemporal location in the same subband from each of the four DDWT transform combinations, as illustrated in Fig. 4.1. These coefficient vectors are then assembled into sets spanning the entire spatiotemporal subbands, producing 7 sets of vectors at each resolution level (4 sets of vectors at the baseband level), assuming a dyadic decomposition structure (other decompositions are discussed below). All the sets are placed in the list of insignificant sets (LIS).

The algorithm then performs bitplane coding with *sorting* and *refinement* passes. In the sorting pass, sets in the LIS are tested against the current threshold to determine the significance of the set as a whole—if the magnitude of any coefficient in the set is above the threshold, the set is significant. Significance sets are split in two along the longest dimension of the set. The resulting subsets are added back to the LIS as two new sets to be recursively tested and split if necessary. Eventually, a significant set will be reduced to a single four-coefficient vector in which at least one of the four coefficients will be significant. At this point, the vector is removed from the LIS, and a significance symbol is output to denote which coefficients in the vector are significant and which are not. The significant coefficients from the vector are then added to the list of significant

pixels (LSP), while the insignificant coefficients are added to the list of insignificant pixels (LIP). After each sorting pass, the LIP is processed by comparing each coefficient to the current threshold and outputting the significance state. If a coefficient in the LIP becomes significant, it is transferred to the LSP. The refinement pass then processes each coefficient in the LSP and outputs the current bitplane value of the coefficient magnitude. Sorting and refinement passes continue until the target bitstream length has been reached. This set-partitioning procedure is illustrated in Fig. 4.2.

In the DDWT-BISK system, two types of wavelet decomposition structures have been considered for the 3D DDWT as illustrated in Figs. 4.3(a) and (b). The first, shown in Fig. 4.3(a), is a traditional dyadic decomposition wherein the wavelet transform is applied to only the lowpass band at each successive level of decomposition. The second structure considered is a wavelet-packet decomposition, shown in Fig. 4.3(b) as the "anisotropic" structure, in which a full $J$-scale 1D wavelet transform is applied to each dimension of the 3D dataset separately. This anisotropic transform structure generates a greater number of subbands than does a dyadic structure for the same number of decomposition levels; in the context of the 3D DDWT, these additional subbands can provide additional directional orientations and can thus increase the degree of motion selectivity. However, a dyadic decomposition was used in the original development of the 3D DDWT [9] and in the DDWTVC coder [11, 12]. Although the anisotropic DDWT was discussed in [12] wherein it was demonstrated that the anisotropic structure provided significantly better reconstruction quality after noise-shaping than the dyadic DDWT for the same number of retained coefficients, the actual DDWTVC system is tied to the dyadic DDWT due to its use of 28-bit vectors for arithmetic coding.

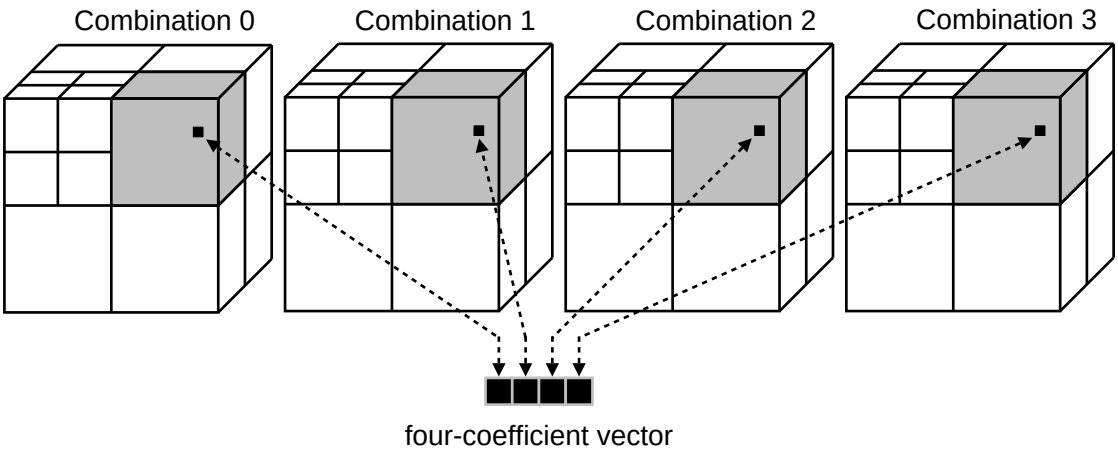



Figure 4.1: A DDWT formed from four transform combinations produced from dyadic DWTs. Co-located coefficients in each of the four transform combinations form a four-coefficient vector.

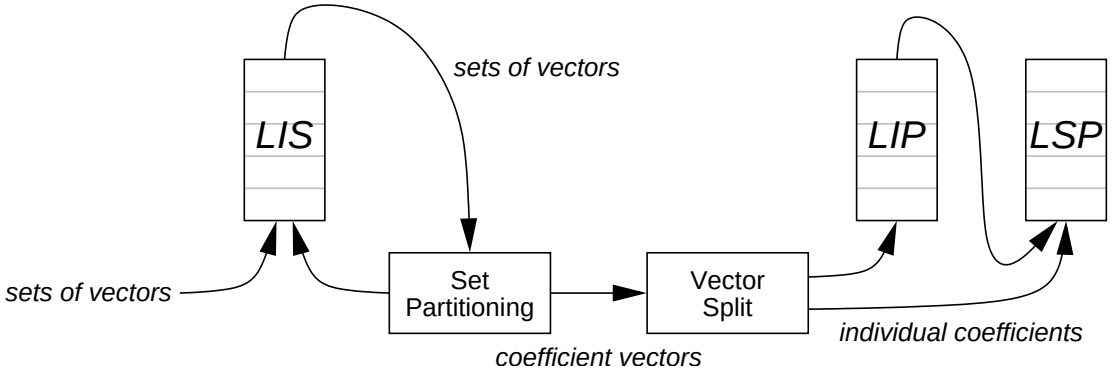



Figure 4.2: The set-partitioning process of the DDWT-BISK coder. The LIS processes sets of coefficient vectors. Once vectors leave the LIS, they are split into individual coefficients—significant coefficients go to the LSP, insignificant coefficients go to the LIP.

## 4.2 Experimental Results

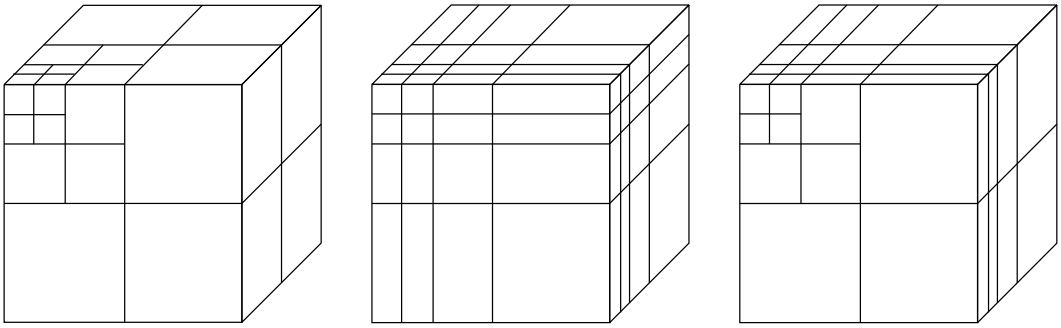
In these experiments, the grayscale sequences shown in Table 4.1 are coded with DDWT-BISK using both the dyadic and anisotropic transform structures. The iterative noise-shaping procedure employed is identical to that used in [12]—the initial and final thresholds driving the iterations were selected from a small set of possible values, with the selection optimized for each given bitrate. The DDWT-BISK results are compared against those provided in [12] for the DDWTVC coder. All coders applied three levels of wavelet decomposition in each dimension.

DDWT-BISK is also compared to JPEG2000 as a state-of-the-art coder using a traditional real-valued critically sampled DWT and no motion estimation or compensation. JPEG2000 results use extensions in Part 2 of the JPEG2000 standard [30] to produce either the anisotropic decomposition of Fig. 4.3(b) or a wavelet-packet transform consisting of a 1D temporal transform followed by a 2D dyadic spatial transform such as illustrated in Fig. 4.3(c) (this latter decomposition is referred to as the “packet” decomposition after terminology in [15, 24]).

Table 4.1 provides average PSNR results for all the test sequences at a fixed rate. The results show that DDWT-BISK with the dyadic transform is generally competitive with JPEG2000 using the packet transform. However, when the anisotropic transform structure is used, DDWT-BISK consistently shows substantial gains, while JPEG2000 shows little change in PSNR. Figs. 4.4 and 4.5 include the DDWTVC coder along with DDWT-BISK and JPEG2000 for rate-distortion comparison. As both plots indicate, the DDWTVC and DDWT-BISK systems perform similarly when using the dyadic transform structure. However, DDWT-BISK with the anisotropic transform achieves significantly higher PSNR levels than those of the other methods.

As a final comparison relating the two contributions of this thesis, Figs. 4.6 and 4.7 provide rate distortion data for both of the proposed video coding systems, along with several benchmark video coders. In these results, the top performance is achieved by the H.264 video coder [31], which exhibits state-of-the-art compression but does not produce a fully scalable encoding. Below the H.264 curve, we see RWMH-EZBC offering moderate improvement over MC-EZBC for most bitrates. As expected, the DDWT-BISK, DDWTVC, and JPEG2000 coders are the bottom performers due to their lack of motion compensation.

We have now presented in-depth discussion and experimental results for both of the main contributions of this thesis. In the final chapter, we provide an analysis of the performance of the RWMH-EZBC and DDWT-BISK video coders and draw conclusions from these experimental results.



(a) (b) (c)

Figure 4.3: Three-level wavelet decomposition for (a) “dyadic,” (b) “anisotropic,” and (c) “packet” decompositions.

Table 4.1: A performance comparison of DDWT-BISK against JPEG2000. Distortion averaged over all frames of the sequence for rate of 0.5 bpp (1520 kbps).

| | PSNR (dB) | | | |
|---|---|---|---|---|
| | DDWT-BISK dyadic | DDWT-BISK anisotropic | JPEG2000 packet | JPEG2000 anisotropic |
| Stefan | 30.4 | 31.5 | 30.5 | 30.7 |
| Mobile | 29.4 | 30.5 | 28.0 | 27.7 |
| Foreman | 38.2 | 38.8 | 38.1 | 37.9 |
| Coastguard | 32.1 | 33.9 | 32.8 | 33.1 |
| Table Tennis | 33.7 | 35.6 | 35.1 | 35.2 |

Sequences are CIF ($352 \times 288$) with 80 frames at 30 Hz.

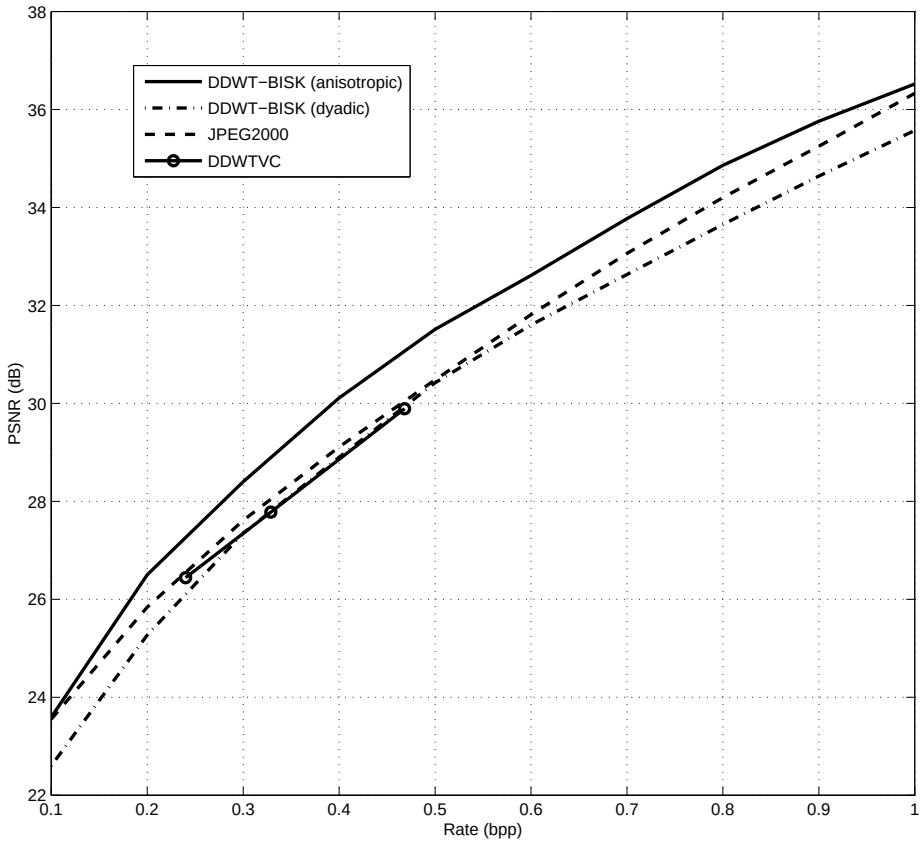



Figure 4.4: DDWT-BISK, DDWTVC, and JPEG2000 rate-distortion performance for "Stefan"

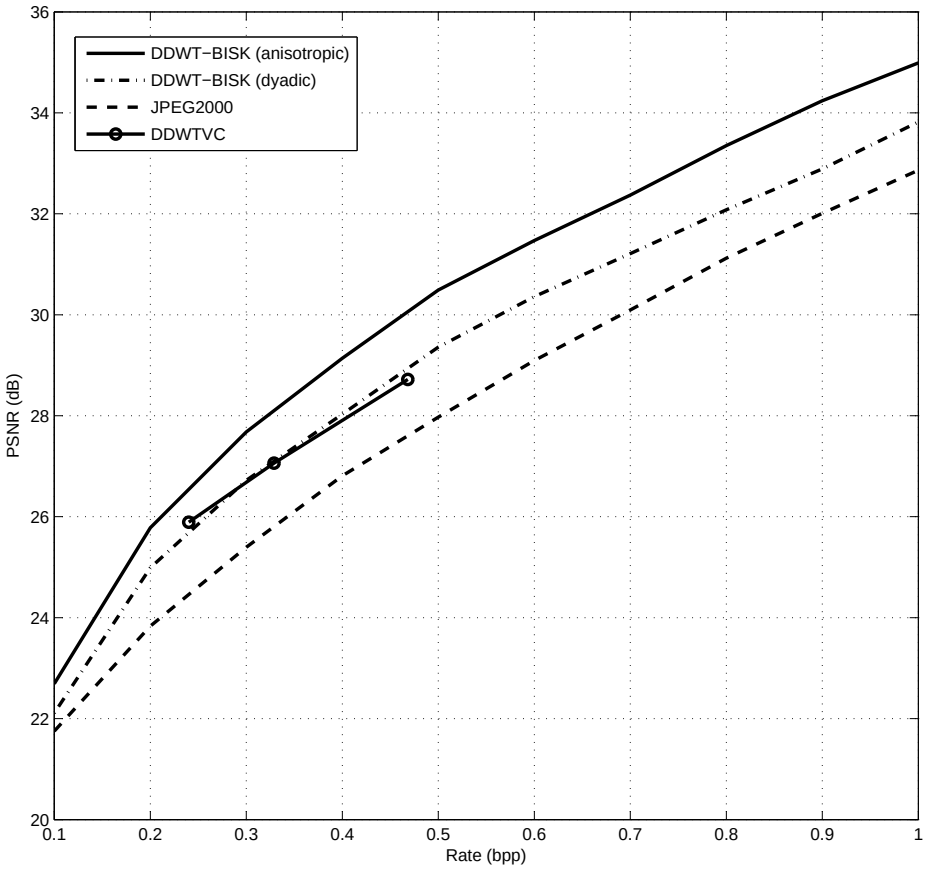



Figure 4.5: DDWT-BISK, DDWTVC, and JPEG2000 rate-distortion performance for "Mobile-Calendar"

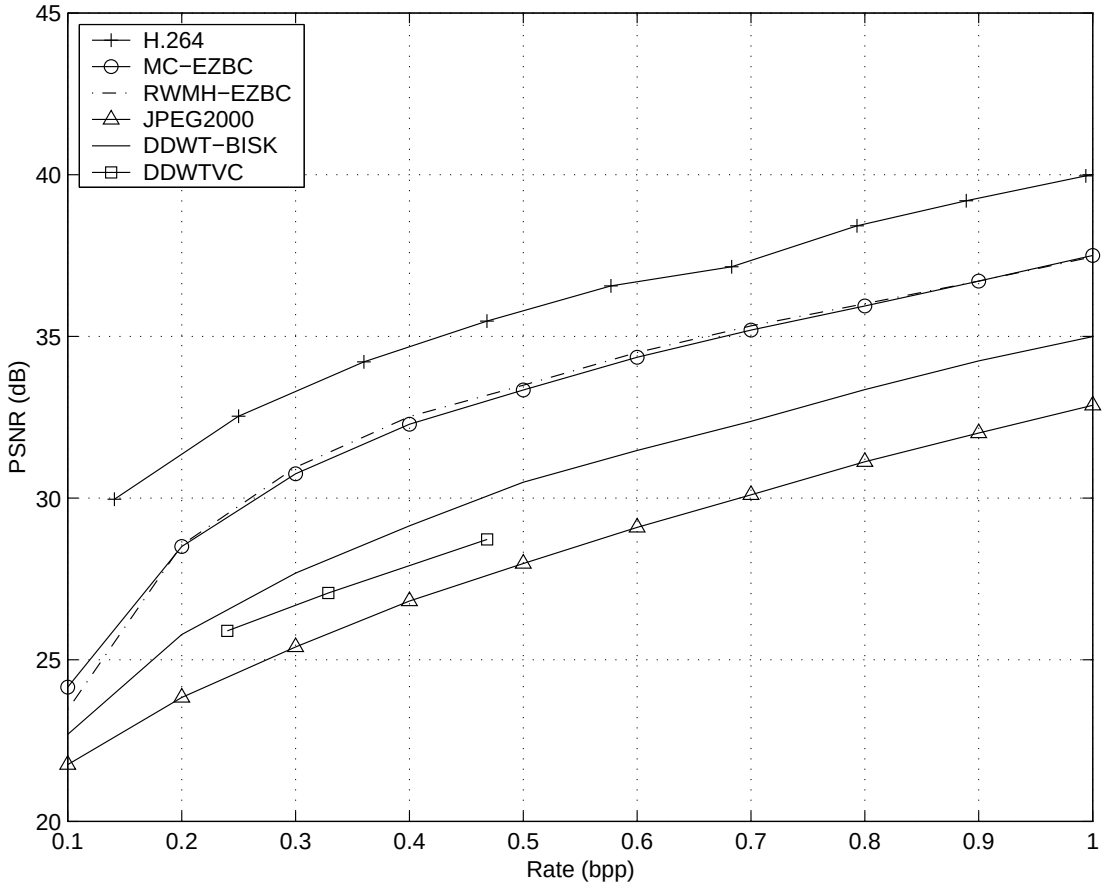



Figure 4.6: Rate-distortion performance of all video coders for “Stefan”

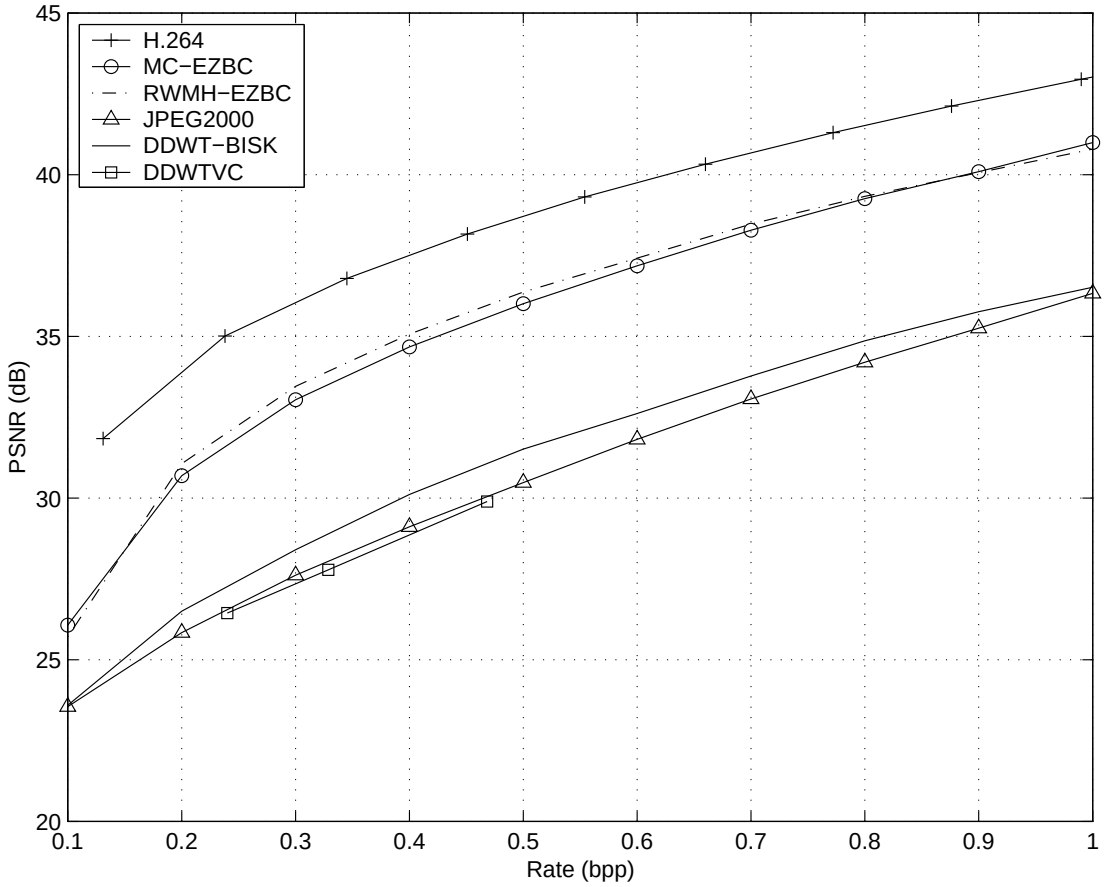



Figure 4.7: Rate-distortion performance of all video coders for “Mobile-Calendar”

# CHAPTER V

# CONCLUSIONS

In the first contribution of this thesis, RWMH [4, 5] is deployed into the prominent fully scalable MC-EZBC [3] video coder by performing MCTF in the domain of the redundant, or overcomplete, wavelet transform. In doing so, the redundancy of the transform is utilized to provide multiple predictions of motion that are diverse in transform phase, with each phase "viewing" motion from a different perspective. After MCTF is performed in the redundant wavelet domain, an inverse RDWT transforms coefficients back into the spatial domain, implicitly combining the multiple temporal filterings into a single, multihypothesis prediction of the true temporal filtering.

Experimental results show that the proposed RWMH-EZBC system improves upon the original MC-EZBC. Although average PSNR improved for all test sequences, the performance gains associated with RWMH-EZBC are more substantial when the video sequence contains fast or complex motion. This is as expected, as RWMH was more effective for these sequences in the systems of [4, 5, 28, 29] as well. The analysis of [5, 29] reveals that the noise reduction provided by the multiple-phase inverse RDWT of the RWMH process is more effective for these sequences since more noise is left uncaptured by the motion model when motion is fast or complex.

In the second contribution of this thesis, a video coder based on the 3D DDWT is proposed. Prior work [9, 10] has indicated that the 3D DDWT can provide a video-signal representation with properties useful for video compression. Those promising

findings led to the development of the proposed DDWT-BISK coder, in which the BISK algorithm [13–15] is adapted for the coding of DDWT coefficients. The modified BISK coding scheme is applied to video data after it undergoes an iterative projection-based noise-shaping procedure [8] to reduce the number of non-zero DDWT coefficients. Whereas the existing DDWTVC coder [11, 12] exploits correlation as it exists across subbands in the DDWT, the DDWT-BISK coder additionally exploits coherent regions of insignificant coefficients that occur within subbands, a coherence that must necessarily be substantial due to the sparsity imposed by the noise-shaping process. When the DDWT uses a dyadic wavelet decomposition, our DDWT-BISK coder provides rate-distortion performance similar to both the DDWTVC coder as well as JPEG2000 applied using a temporal DWT and no motion estimation or compensation. However, the real advantage of the DDWT-BISK system is revealed when a DDWT with an anisotropic wavelet decomposition is used. While JPEG2000, when using this anisotropic transform, yields more or less unchanged performance as compared to the dyadic transform, DDWT-BISK consistently achieves substantial gains, with PSNR levels often 1 dB or more over both JPEG2000 as well as DDWTVC.

The anisotropic decomposition has many more subbands than does the dyadic transform, increasing the directionality of the decomposition, while at the same time decreasing the size of the subbands. The increased directionality of the decomposition appears to increase the sparsity of the significant coefficients. On the other hand, the anisotropic subbands are much smaller than their dyadic counterparts, reducing the capability of set-partitioning algorithms like DDWT-BISK to group large spatiotemporally contiguous regions of insignificant coefficients together into sets. The experimental results here suggest that the benefits of the first effect (increased directionality) must substantially outweigh the detriments of the second (decreased subband size).

While an anisotropic decomposition would appear to increase directionality within a critically sampled DWT as well, this additional directionality apparently has little impact on DWT-based coders since the DWT has only limited directionality to start with. Furthermore, while the anisotropic transform proves to be beneficial for our DDWT-BISK coder, it is unclear whether similar gains would be seen for other DDWT-based coders. DDWTVC, for example, is confined to using the dyadic DDWT; on the other hand, the DDWT-BISK coder can be applied effectively to any subband tiling.

Both of the main contributions of this thesis—the RWMH-EZBC system as well as the DDWT-BISK coder—were developed with the goal of increasing performance of alternatives to the traditional hybrid architecture that has dominated video coding in the form of the MPEG and H.26x standards for the last 20 years. While wavelets are currently the preferred approach to still-image coding, it has proven more difficult to adapt efficient wavelet performance to the video arena. The contributions of this thesis are steps toward a potential next-generation of video coding based on wavelets, and ultimately, we hope, an increase in video-compression performance beyond today's state of the art.

REFERENCES


[1] A. Secker and D. Taubman, "Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation," in *Proceedings of the International Conference on Image Processing*, vol. 3, Rochester, NY, September 2002, pp. 749–752.

[2] G. J. Sullivan, "Multi-hypothesis motion compensation for low bit-rate video coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, Minneapolis, MN, April 1993, pp. 437–440.

[3] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 10, pp. 1183–1194, October 2004.

[4] S. Cui, Y. Wang, and J. E. Fowler, "Multihypothesis motion compensation in the redundant wavelet domain," in *Proceedings of the International Conference on Image Processing*, vol. 2, Barcelona, Spain, September 2003, pp. 53–56.

[5] J. E. Fowler, S. Cui, and Y. Wang, "Motion compensation via redundant-wavelet multihypothesis," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3102–3113, October 2006.

[6] J. B. Boettcher and J. E. Fowler, "Video coding with MC-EZBC and redundant-wavelet multihypothesis," in *Proceedings of the International Conference on Image Processing*, vol. 3, Genoa, Italy, September 2005, pp. 229–232.

[7] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied Computational Harmonic Analysis*, vol. 10, pp. 234–253, May 2001.

[8] T. H. Reeves and N. G. Kingsbury, "Overcomplete image coding using iterative projection-based noise shaping," in *Proceedings of the International Conference on Image Processing*, vol. 3, Rochester, NY, September 2002, pp. 597–600.

[9] I. W. Selesnick and K. Y. Li, "Video denoising using 2D and 3D dual-tree complex wavelet transforms," in *Wavelets: Applications in Signal and Image Processing X*,

M. A. Unser, A. Aldroubi, and A. F. Laine, Eds. San Diego, CA: Proc. SPIE 5207, August 2003, pp. 607–618.

[10] B. Wang, Y. Wang, I. Selesnick, and A. Vetro, "An investigation of 3D dual-tree wavelet transform for video coding," in *Proceedings of the International Conference on Image Processing*, vol. 2, Singapore, October 2004, pp. 1317–1320.

[11] ——, "Video coding using 3-D dual-tree discrete wavelet transforms," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Philadelphia, PA, March 2005, pp. 61–64.

[12] ——, "Video coding using 3-D dual-tree wavelet transform," *EURASIP Journal on Image and Video Processing*, vol. 2007, 2007, article ID 42761, 15 pages.

[13] J. E. Fowler, "Shape-adaptive coding using binary set splitting with $k$-d trees," in *Proceedings of the International Conference on Image Processing*, vol. 2, Singapore, October 2004, pp. 1301–1304.

[14] J. T. Rucker and J. E. Fowler, "Coding of ocean-temperature volumes using binary set splitting with $k$-d trees," in *Proceedings of the International Geoscience and Remote Sensing Symposium*, vol. 1, Anchorage, AK, September 2004, pp. 289–292.

[15] ——, "Shape-adaptive embedded coding of ocean-temperature imagery," in *Proceedings of the* $40^{th}$ *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 2006, pp. 1887–1891.

[16] J. B. Boettcher and J. E. Fowler, "Video coding using a complex wavelet transform and set partitioning," *IEEE Signal Processing Letters*, vol. 14, no. 9, pp. 633–636, September 2007.

[17] ——, "A modified BISK algorithm for 3D dual-tree wavelet transform coding," in *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2007, p. 377, poster presentation.

[18] J. E. Fowler, J. B. Boettcher, and B. Pesquet-Popescu, "Image coding using a complex dual-tree wavelet transform," in *Proceedings of the European Signal Processing Conference*, Poznań, Poland, September 2007, to appear.

[19] J. B. Boettcher, Q. Du, and J. E. Fowler, "Hyperspectral image compression with the 3D dual-tree wavelet transform," in *Proceedings of the International Geoscience and Remote Sensing Symposium*, Barcelona, Spain, July 2007, to appear.

[20] J.-R. Ohm, "Temporal domain subband video coding with motion compensation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, March 1992, pp. 229–232.

[21] ——, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, September 1994.

[22] 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.

[23] ——, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 3, Geneva, Switzerland, May 2000, pp. 662–665.

[24] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374–1387, December 2000.

[25] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets: Time-Frequency Methods and Phase Space*, J.-M. Combes, A. Grossman, and P. Tchamitchian, Eds. Berlin, Germany: Springer-Verlag, 1989, pp. 286–297, Proceedings of the International Conference, Marseille, France, December 14–18, 1987.

[26] P. Dutilleux, "An implementation of the "algorithme à trous" to compute the wavelet transform," in *Wavelets: Time-Frequency Methods and Phase Space*, J.-M. Combes, A. Grossman, and P. Tchamitchian, Eds. Berlin, Germany: Springer-Verlag, 1989, pp. 298–304, Proceedings of the International Conference, Marseille, France, December 14–18, 1987.

[27] H.-W. Park and H.-S. Kim, "Motion estimation using low-band-shift method for wavelet-based moving-picture coding," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 577–587, April 2000.

[28] Y. Wang, S. Cui, and J. E. Fowler, "3D video coding using redundant-wavelet multihypothesis and motion-compensated temporal filtering," in *Proceedings of the International Conference on Image Processing*, vol. 2, Barcelona, Spain, September 2003, pp. 755–758.

[29] ——, "3D video coding with redundant-wavelet multihypothesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 166–177, February 2006.

[30] *Information Technology—JPEG 2000 Image Coding System—Part 2: Extensions*, ISO/IEC 15444-2, 2004.

[31] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003.