3D Video Coding with Redundant-Wavelet Multihypothesis

Yonghui Wang, *Member, IEEE*, Suxia Cui, *Member, IEEE*,
and James E. Fowler, *Senior Member, IEEE*

***Abstract*— Multihypothesis with phase diversity is introduced into motion-compensated temporal filtering by deploying the latter in the domain of a spatially redundant wavelet transform. The centerpiece of this redundant-wavelet approach to multihypothesis temporal filtering is a multiple-phase inverse transform that involves an implicit projection significantly reducing noise not captured by the motion model of the temporal filtering. The primary contribution of the work is a derivation that establishes analytically the advantage of the redundant-wavelet approach as compared to equivalent temporal filtering taking place in the spatial domain. For practical implementation, a regular triangle mesh is used to track motion between frames, and an affine transform between mesh triangles implements motion compensation within a lifting-based temporal transform. Experimental results reveal that the incorporation of phase-diversity multihypothesis into motion-compensated temporal filtering improves rate-distortion performance, and state-of-the-art scalable performance is observed.**

***Index Terms*— motion-compensated temporal filtering, multihypothesis motion compensation, redundant wavelet transform, scalable video coding**

I. Introduction

It has been generally recognized that the goal of highly scalable video representation is fundamentally at odds with the traditional motion-estimation/motion-compensation (ME/MC) feedback loop which hinders the achieving of a high degree of spatial, temporal, and fidelity scalability. Consequently, the use of 3D transforms, which break the ME/MC feedback loop, are becoming a preferred approach to full scalability, and a number of modern 2D still-image algorithms have been straightforwardly extended to the third dimension (e.g., 3D-SPIHT [1]) by employing separable 3D wavelet transforms. This approach usually involves a wavelet-packet subband decomposition wherein a group of frames is processed with a temporal transform followed by spatial decomposition of each frame. However, without MC, temporal transforms produce low-quality temporal subbands with significant "ghosting" artifacts [2] and decreased coding efficiency. Consequently, there has been significant interest in motion-compensated temporal filtering (MCTF) [2–16] in which it is attempted to have the temporal transform follow motion trajectories.

Y. Wang and S. Cui were with the Department of Electrical & Computer Engineering and the GeoResources Institute, Mississippi State University, Starkville, MS, and are currently with the Department of Engineering Technology at Prairie View A&M University in Prairie View, TX.

J. E. Fowler is with the Department of Electrical & Computer Engineering and the GeoResources Institute, Mississippi State University, Starkville, MS.

This work was funded in part by the National Science Foundation under Grants Nos. CCR-0310864 and ACI-9982344.

In this paper, we describe a video coder using a 3D wavelet transform with MCTF. The salient aspect of this coder lies in that we employ multihypothesis motion compensation (MHMC) within the MCTF to combat the uncertainty inherent in estimating motion trajectories for MCTF, thereby achieving rate-distortion performance significantly superior to the usual single-hypothesis MCTF approach. Although multihypothesis has been used in conjunction with MCTF before (e.g., [9–11, 15] propose both spatially and temporally diverse multihypothesis MCTF predictions), in our proposed system, we employ a new class of MHMC—phase-diversity multihypothesis [17, 18]. Specifically, phase-diversity MHMC is implemented by deploying MCTF in the domain of a spatially redundant wavelet transform such that multiple hypothesis temporal filterings are combined implicitly in the form of an inverse transform. While the overwhelming majority of previous MCTF techniques have deployed MCTF in the spatial domain, a few recent approaches (e.g., [9, 10, 12–15]) have used the shift invariance of spatially redundant transforms to enable wavelet-domain MCTF. In contrast to these techniques, our redundant-wavelet-multihypothesis (RWMH) approach to MCTF exploits the transform redundancy not only for its shift invariance, but for, more importantly, its potential for superior temporal filtering via phase-diversity multihypothesis.

In [18, 19], we established analytically the performance gain associated with RWMH when employed in the traditional hybrid coding architecture. That analysis considered a predictive feedback loop under the assumption of a simple motion model consisting of a translation plus noise. It was shown that, thanks to the well-known robustness of overcomplete transforms to noise in the transform domain, RWMH substantially reduced the variance of the noise not captured by the translational motion model, resulting in a significant reduction in the power of the prediction residual in the MC feedback loop. As a primary contribution of this paper, we apply this analysis to the MCTF setting, showing that essentially this same principle carries over to RWMH-based MCTF. Our analysis yields that RWMH can reduce both the variance of the entire MCTF highpass frame as well as the variance of the noise within the MCTF lowpass frame by up to 7 dB with respect to an equivalent system using spatial-domain MCTF.

To verify the analysis in a practical setting, we describe a 3D-RWMH implementation that combines the RWMH paradigm we introduced in [17, 18] with MCTF. We depart from the common approach to MCTF which relies on traditional block-based motion models, opting instead for the triangle-mesh lifting-based MCTF architecture pioneered in

[2–4]. This mesh-based approach facilitates subpixel accuracy as well as temporal filters longer than the Haar filter typically encountered in MCTF systems—both aspects of which we employ to enhance performance of our 3D-RWMH system. Experimental results demonstrate that the incorporation of phase-diversity multihypothesis into MCTF can achieve a significant gain in rate-distortion performance over a comparable single-hypothesis system, particularly so when there is substantial noise not captured by the MCTF process. Additionally, the results indicate that our 3D-RWMH coder usually outperforms bidirectional MC-EZBC [7], a prominent 3D video coder with state-of-the-art performance, and occasionally outperforms H.264 [20], the state of the art in non-scalable hybrid coding.

Below, we describe our approach in greater detail. We first review the RWMH technique from [17, 18] in Sec. II. Then, in Sec. III, we describe in detail our 3D-RWMH approach that incorporates RWMH into MCTF, including theoretical analysis of the associated performance gain. In Sec. IV, we consider some issues surrounding the implementation of 3D-RWMH using MCTF with triangle meshes and subpixel ME accuracy. We then present experimental observations in Sec. V, and, finally, we make some concluding remarks in Sec. VI.

II. Redundant-Wavelet Multihypothesis (RWMH)

MHMC [21] forms a prediction of pixel $s[x, y, t]$ in the current frame as a combination of multiple predictions in an effort to combat the uncertainty inherent in the ME process. Assuming that the combination of these hypothesis predictions is linear, we have that the prediction of frame $s[x, y, t]$ is

$$\tilde{s}[x, y, t] = \sum_i w_i[x, y, t]\tilde{s}_i[x, y, t], \qquad (1)$$

where the multiple predictions $\tilde{s}_i[x, y, t]$ are combined according to some weights $w_i[x, y, t]$. One approach to MHMC is to implement multihypothesis prediction spatially—the predictions $\tilde{s}_i[x, y, t]$ are culled from spatially distinct locations in the reference frame; e.g., subpixel-accurate ME/MC and overlapped block MC. Another approach is to deploy MHMC temporally by choosing predictions from multiple reference frames; e.g., B-frames and long-term-memory MC [22].

In [17, 18], we introduced a new class of MHMC—phase-diversity MHMC—in which the multihypothesis-prediction concept is extended into the transform domain. Specifically, we performed ME/MC in the domain of a redundant, or overcomplete, wavelet transform, and used multiple predictions that were diverse in transform phase. Our approach to phase-diversity multihypothesis, RWMH, takes place in the domain of the redundant discrete wavelet transform (RDWT)[1] which is an approximation to the continuous wavelet transform that, in essence, removes the downsampling operator from the traditional critically sampled transform to produce an overcomplete representation. As illustrated in Fig. 1, the size of each subband of an RDWT is the same as that of the input signal. Additionally, 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 outlined in [17, 18], 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. A multiple-phase inverse RDWT combines these multiple hypotheses into a single prediction. Specifically in reference to (1), for a J-scale RDWT, the reconstruction from DWT i of the RDWT is $\tilde{s}_i[x, y, t]$, $0 \leq i < 4^J$, while $w_i[x, y, t] = 4^{-J}$, $\forall i$.

In [17, 18], we described a video-coding system that incorporated RWMH into the MC feedback loop of the traditional hybrid, block-based video-coding architecture. In [18, 19], we presented an analytical derivation that quantifies the performance gain of this hybrid RWMH architecture over single-phase prediction. Key to this analysis was the fact that noise in the RDWT domain undergoes a substantial reduction in variance when the multiple-phase inverse RDWT is applied. This noise reduction was 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 MC feedback loop and higher coding efficiency. In fact, the analysis of [18, 19] predicted that, as ME becomes highly accurate, RWMH can reduce the prediction-residual variance by up to 7 dB regardless of the power of the noise.

III. Motion-Compensated Temporal Filtering with Redundant-Wavelet Multihypothesis

In this section, we introduce the RWMH concept into the MCTF framework, in effect eliminating the MC feedback loop from our RWMH system of [17, 18] and producing a fully scalable 3D video coder. In essence, this 3D-RWMH coder is based on the 3D coder of [2, 3] but with the key addition of multihypothesis via RWMH. We overview the 3D-RWMH system next, and then, in Sec. III-B, we apply the analysis from [18, 19] to the 3D MCTF setting to quantify the performance gain of the 3D-RWMH approach.

A. The 3D-RWMH System

As depicted in Fig. 2, the encoder of our 3D-RWMH video-coding system first performs a spatial RDWT on each frame and then performs MCTF in the redundant-wavelet domain. This is in contrast to many prior MCTF techniques [2–8] in which MCTF takes place in the spatial domain. Since MCTF is performed in the RDWT subbands, it is overcomplete spatially; consequently, before coding the temporal subbands, we remove this spatial redundancy by performing an inverse spatial RDWT on each frame. Intuitively, each RDWT phase in each frame can be considered to have viewed the MCTF from a different perspective and thus forms an independent hypothesis about the temporal filtering taking place. The inverse spatial RDWT implicitly combines these hypotheses

[1]The RDWT first appeared as the *algorithme à trous* [23, 24] and has subsequently been known by a variety of names including the undecimated DWT (UDWT), the overcomplete DWT (ODWT), and discrete wavelet frames (DWF).

into a multihypothesis estimate of what the true temporal filtering should be. After the inverse spatial transform, the temporally transformed frames are coded by a suitable 3D coder.[2]

B. Analysis of 3D-RWMH

In this section, we show analytically that the multihypothesis nature of the MCTF in the 3D-RWMH coder of Fig. 2 offers substantial performance gain over the usual approach of spatial-domain MCTF. For the moment, in order to render the analysis tractable, we assume that simple translational motion takes place between frames and employ simple MCTF—namely, a temporal Haar transform—implemented via lifting [2–4]. We will consider more sophisticated temporal filtering for the experimental results later.

Let $s[x, y, t]$ be a video sequence sampled spatially on an integer-pixel lattice and temporally at integer times and denote the 2D spatial RDWT of frame $s[x, y, t]$ at time t as the collection of subbands $S^{(k)}[x, y, t]$,

$$\Big\{S^{(k)}[x, y, t]\Big\}_k = \mathcal{R}\Big[s[x, y, t]\Big]. \tag{2}$$

Given subbands $S^{(k)}[x, y, t]$, we define the inverse RDWT as

$$s[x, y, t] = \mathcal{R}^{-1}\left[\Big\{S^{(k)}[x, y, t]\Big\}_k\right], \tag{3}$$

which is a multiple-phase inverse equivalent to inverting each of the 4^J critically sampled DWTs constituting the J-scale 2D RDWT and averaging the resulting reconstructions together.

Suppose $s[x, y, t-1]$ and $s[x, y, t]$ are two consecutive frames of a video sequence, and let $W_{t-1\to t}$ denote an operator that maps the frame at time $t-1$ onto the coordinate system of the frame at time t through the particular ME/MC scheme of choice. Assuming that the operator $W_{t-1\to t}$ is applied identically to each subband, Haar-based MCTF in the RDWT domain would be implemented via lifting [2–4] as

$$\Big\{H^{(k)}[x, y]\Big\}_k = \left\{\frac{1}{2}\Big(S^{(k)}[x, y, t] - W_{t-1\to t}\Big[S^{(k)}[x, y, t-1]\Big]\Big)\right\}_k, \tag{4}$$

$$\Big\{L^{(k)}[x, y]\Big\}_k = \Big\{S^{(k)}[x, y, t-1] + W_{t\to t-1}\Big[H^{(k)}[x, y]\Big]\Big\}_k, \tag{5}$$

where $\big\{L^{(k)}[x, y]\big\}_k$ and $\big\{H^{(k)}[x, y]\big\}_k$ are the lowpass and highpass frames, respectively, of the temporal transform.

Within RDWT subband k, we adopt the simple translational motion model from [25]. Specifically, we assume that the current frame at time t is a simple displacement of the previous frame plus residual noise not captured by the translational motion; i.e.,

$$S^{(k)}[x, y, t] = I\Big(S^{(k)}[x - d_x, y - d_y, t-1]\Big) + N^{(k)}[x, y, t], \tag{6}$$

where (d_x, d_y) is the unknown translation and $I(\cdot)$ is a linear interpolation operator used to resolve fractional-pixel values. Furthermore, let us assume that operators $W_{t-1\to t}$ and $W_{t\to t-1}$ estimate the motion (d_x, d_y) as $(\hat{d}_x, \hat{d}_y)$ such that

$$W_{t-1\to t}\Big[S^{(k)}[x, y, t-1]\Big] = I\Big(S^{(k)}[x - \hat{d}_x, y - \hat{d}_y, t-1]\Big), \tag{7}$$

$$W_{t\to t-1}\Big[S^{(k)}[x, y, t]\Big] = I\Big(S^{(k)}[x + \hat{d}_x, y + \hat{d}_y, t]\Big). \tag{8}$$

The high- and lowpass frames, (4) and (5), respectively, then become

$$\Big\{H^{(k)}[x, y]\Big\}_k = \left\{\frac{1}{2}I\Big(S^{(k)}[x - d_x, y - d_y, t-1]\Big) - \frac{1}{2}I\Big(S^{(k)}[x - \hat{d}_x, y - \hat{d}_y, t-1]\Big) + \frac{1}{2}N^{(k)}[x, y, t]\right\}_k, \tag{9}$$

$$\Big\{L^{(k)}[x, y]\Big\}_k = \left\{\frac{1}{2}S^{(k)}[x, y, t-1] + \frac{1}{2}I\Big(S^{(k)}[x + \hat{d}_x - d_x, y + \hat{d}_y - d_y, t-1]\Big) + \frac{1}{2}I\Big(N^{(k)}[x + \hat{d}_x, y + \hat{d}_y, t]\Big)\right\}_k. \tag{10}$$

In the system of Fig. 2, after MCTF takes place in the spatial RDWT domain, an inverse spatial RDWT is applied to combine the multiple MCTF hypotheses. Since the inverse RDWT is a pseudo-inverse, this is tantamount to a projection onto the range space of the RDWT following by a mapping back into the original spatial domain. Since the noise $N^{(k)}[x, y, t]$ not captured by the motion model of (6) is almost certainly not in the range space of the RDWT, the mapping of the RDWT-domain noise back to the spatial domain will result in a reduction in noise variance.

Let us first consider the noise variance in the highpass frame. In the spatial domain, the highpass frame (9) is

$$h[x, y] = \mathcal{R}^{-1}\left[\Big\{H^{(k)}[x, y]\Big\}_k\right] = \frac{1}{2}n[x, y, t] + \frac{1}{2}I\Big(s[x - d_x, y - d_y, t-1] - s[x - \hat{d}_x, y - \hat{d}_y, t-1]\Big), \tag{11}$$

where we define $n[x, y, t]$ to be

$$n[x, y, t] = \mathcal{R}^{-1}\left[\Big\{N^{(k)}[x, y, t]]\Big\}_k\right], \tag{12}$$

and invoke the fact that $\mathcal{R}^{-1}$ is shift-invariant under linear fractional-pixel interpolation. In [18, 19], we employed the analysis put forth by Girod [25, 26] to derive the variance of a prediction residual such as given by (11) to be

$$\nu_h = \frac{1}{4}\nu_n + \frac{1}{4}\Gamma_{ss}, \tag{13}$$

where ν_n is the variance of $n[x, y, t]$ and

$$\Gamma_{ss} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Phi_{ss}(\omega_x, \omega_y)\Big(2 - 2\Re\big[P(\omega_x, \omega_y)\big]\Big)\,d\omega_x\,d\omega_y. \tag{14}$$

[2]For many 3D coders, such as 3D-SPIHT [1], a spatial forward DWT (not shown in Fig. 2) is applied to each frame following the spatial inverse RDWT of the 3D-RWMH system, since the coefficients resulting from 3D-RWMH are in the DWT domain in only one dimension (the temporal dimension).

In (14), $\Phi_{ss}(\omega_x, \omega_y)$ is the 2D power spectral density of $s[x, y, t]$; $P(\omega_x, \omega_y)$ is the 2D Fourier transform of the probability density function of the displacement error, $(\Delta_x, \Delta_y) = (d_x, d_y) - (\hat{d}_x, \hat{d}_y)$; $\Re(\cdot)$ denotes the real part of a complex number; and we have assumed $I(\omega_x, \omega_y) = 1$ (i.e., sinc interpolation [26]) in order to simplify the analysis.

Now turning to the lowpass frame, in the spatial domain, (10) becomes

$$l[x,y] = \mathcal{R}^{-1}\left[\left\{L^{(k)}[x,y]\right\}_k\right] = \frac{1}{2}I\Big(n[x+\hat{d}_x, y+\hat{d}_y, t]\Big) + \frac{1}{2}s[x,y,t-1] + \frac{1}{2}I\Big(s[x-\Delta_x, y-\Delta_y, t-1]\Big), \quad (15)$$

where once again $n[x, y, t]$ is given by (12). A straightforward modification to our derivation in [18, 19] yields that the variance of the lowpass frame (15) is

$$\nu_l = \frac{1}{4}\nu_n + \frac{1}{4}\Lambda_{ss}, \quad (16)$$

where

$$\Lambda_{ss} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Phi_{ss}(\omega_x,\omega_y)\Big(2 + 2\Re\big[P(\omega_x,\omega_y)\big]\Big)\,d\omega_x\,d\omega_y. \quad (17)$$

We argue that, for effective compression, MCTF should serve to reduce the power in the highpass frame as much as possible; i.e., MCTF should minimize the variance ν_h of (13). Additionally, in the lowpass frame, the portion of the lowpass-frame variance due to failure of the motion model should be minimized, while the portion of the variance due to the video sequence itself should be maximized to exploit the energy-compaction property of the temporal transform to the largest extent possible. That is, in (16), Λ_{ss} should be maximized while ν_n should be minimized.

Let us use the preceding analysis to compare the performance of the 3D-RWMH system of Fig. 2 to an equivalent system with MCTF operating in the spatial domain. In order to make a quantitative evaluation, we assume that the noise $N^{(k)}[x, y, t]$ not captured by the motion model of (6) is zero-mean, white, and of variance ν_N. This white-noise model is an obvious oversimplification since real MC noise signals will typically possess a significant degree of correlation both between subbands as well as spatially within subbands. However, due to a result derived in [27], assuming white noise permits us to quantify the noise reduction due to the inverse RDWT of (12) as

$$\nu_n = \frac{\nu_N}{5}\left[1 + 4\left(\frac{1}{16}\right)^J\right], \quad (18)$$

assuming that the wavelet filters underlying the J-scale 2D RDWT are orthonormal. In Sec. V, we will return to this white-noise assumption to examine the noise-reduction capabilities of (12) on real MC noise signals. For now, however, let us assume white noise and that the motion model of the spatial-domain MCTF is as effective as its RDWT-domain counterpart. That is, the translational motion model of both temporal filterings fails to capture a residual noise of power ν_0. From (13), the difference (in dB) in variance of the highpass band between the RWMH and spatial-domain approaches is

$$\gamma_h = 10\log_{10}\left(\frac{\nu_n + \Gamma_{ss}}{\nu_0 + \Gamma_{ss}}\right), \quad (19)$$

where $\nu_N = \nu_0$ in (18). To quantify this difference, let us assume, as was done in [25, 26], an isotropic signal power spectrum,

$$\Phi_{ss}(\omega_x,\omega_y) = \frac{2\pi}{\omega_0^2}\left(1 + \frac{\omega_x^2+\omega_y^2}{\omega_0^2}\right)^{-\frac{2}{3}}, \quad (20)$$

where $\omega_0 = -\ln(0.93)$, and an isotropic Gaussian displacement-error density of variance ν_Δ such that

$$P(\omega_x,\omega_y) = \exp\left[-\frac{\nu_\Delta}{2}\left(\omega_x^2+\omega_y^2\right)\right]. \quad (21)$$

Under these models, we numerically evaluate (19) versus displacement-error accuracy β in Fig. 3 for several noise variances, where $\beta = \frac{1}{2}\log_2(12\nu_\Delta)$ such that $\beta = -1$ for half-pixel accuracy, $\beta = -2$ for quarter-pixel accuracy, etc. We assume that J is large. We see that as MCTF becomes highly accurate (β small), RWMH produces a 7-dB reduction in highpass-frame variance as compared to the spatial-domain system regardless of the strength of the noise not captured by the motion model. Additionally, we observe that, at moderate MCTF accuracy ($0 \lesssim \beta \lesssim 2$), the greater the noise variance, the greater is the variance reduction of RWMH over the spatial-domain system.

As pertaining to the lowpass frame, it is straightforward to see that the integral in (17)—and consequently the Λ_{ss} component of ν_l in (16)—is maximized as ν_Δ goes to zero, as we would expect from intuition. The difference in the noise component of the lowpass-frame variance is

$$\gamma_l = 10\log_{10}\left(\frac{\nu_n}{\nu_0}\right). \quad (22)$$

With $\nu_N = \nu_0$, we have that $\gamma_l \approx -7$ dB. Thus, we conclude that RWMH can reduce both the highpass-frame variance and the noise component of the lowpass-frame variance by up to 7 dB as compared to an equivalent spatial-domain system.

The analysis presented in this section is built upon the assumptions of a simple translational motion model and MCTF using simple Haar temporal filtering. These assumptions render the analysis tractable; however, for a practical implementation, better performance can be had if we employ a more sophisticated MCTF that transcends these assumptions. In the next section, we describe how we have implemented 3D-RWMH in practice. We note that the measures γ_h and γ_l resulting from our analysis suggest, but do not guarantee, increased coding efficiency in a practical setting; however, we will see later in Sec. V that experimental results using the 3D-RWMH implementation described next do indeed support the main theoretical observations of this section.

IV. Implementation Issues

A. MCTF with Triangle Meshes

The assumption of a simple translational motion model in theory is usually reflected in practice by imposing simple

block-based MC. Consequently, the most common approach to MCTF combines traditional block-based MC with temporal filtering [4–10, 12]. These block-based techniques have encountered a number of drawbacks in that the rigid block-motion model fails to capture all aspects of the motion field, leaving a significant number of pixels "unconnected" between frames, while implementation of temporal filters other than the simple Haar is hindered by these numerous unconnected pixels. Additionally, in the case of 3D-RWMH, we wish to avoid filtering of the discontinuous blocking artifacts arising in block-based MCTF when inverting the RDWT to produce the multihypothesis MCTF. As a consequence, for the implementation of 3D-RWMH which we will experimentally evaluate below, we employ the lifting approach to MCTF of [2–4] that not only facilitates longer temporal filters but also permits ME/MC schemes more general than block displacement to be implemented in an easily inverted fashion. Specifically, we use the biorthogonal 5-3 filter formulated in [3, 4], which, when applied in the RDWT domain, is

$$H^{(k)}[x,y,t] = S^{(k)}[x,y,2t+1] - \frac{1}{2}\Big(W_{2t\to 2t+1}\Big[S^{(k)}[x,y,2t]\Big] + W_{2t+2\to 2t+1}\Big[S^{(k)}[x,y,2t+2]\Big]\Big), \quad (23)$$

$$L^{(k)}[x,y,t] = S^{(k)}[x,y,2t] + \frac{1}{4}\Big(W_{2t-1\to 2t}\Big[H^{(k)}[x,y,t-1]\Big] + W_{2t+1\to 2t}\Big[H^{(k)}[x,y,t]\Big]\Big), \quad (24)$$

and we drive the MCTF with triangle meshes similar to those of [2, 3], thus producing smooth motion fields more suitable to RWMH since they lack discontinuous block artifacts and unconnected pixels.

As originally proposed, the mesh-based MCTF of [2, 3] uses a uniform, regular triangle mesh resulting from the dividing of the frame into square blocks and the splitting of each block along its diagonal. The triangle vertices, or "control points," of this uniform mesh are tracked from one frame to the next via the iterative hexagonal-refinement optimization of [28]. In our implementation, we also use this regular triangle-mesh structure; however, we opt for the simpler block-based ME strategy of [29] to determine control-point motion. Specifically, motion into the next frame is estimated by centering a small block at each vertex in the first frame and finding the best matching block in the second frame. Motion of the control points from the second frame to the third frame is tracked in this same manner, and so on to subsequent frames. If the motion vectors have integer-pixel accuracy, the control points in every frame reside on the integer-pixel grid.

We search for the motion of the control points of the mesh by minimizing a distortion metric that spans across all subbands of the RDWT decomposition, as we did in [30]. Specifically, the motion vector, (d_x, d_y), for control point (x, y) in the reference frame is the vector in the search window about (x, y) in the current frame that minimizes the mean absolute error (MAE),

$$\mathrm{MAE}(x,y,d_x,d_y) = \frac{1}{B^2}\sum_{m=-\lfloor B/2\rfloor}^{\lfloor B/2\rfloor}\sum_{n=-\lfloor B/2\rfloor}^{\lfloor B/2\rfloor}\mathrm{AE}(x+m,y+n,d_x,d_y). \quad (25)$$

The absolute error (AE) is

$$\begin{aligned}\mathrm{AE}(x,y,d_x,d_y) = {} & 2^{-J}\Big|B_J[x,y,t] - B_J[x+d_x,y+d_y,t-1]\Big| + \\ \sum_{j=1}^{J} 2^{-j}\Bigg(&\Big|V_j[x,y,t] - V_j[x+d_x,y+d_y,t-1]\Big| + \\ &\Big|H_j[x,y,t] - H_j[x+d_x,y+d_y,t-1]\Big| + \\ &\Big|D_j[x,y,t] - D_j[x+d_x,y+d_y,t-1]\Big|\Bigg),\end{aligned} \quad (26)$$

where B_j, H_j, V_j, and D_j are the baseband, horizontal, vertical, and diagonal subbands, respectively, at scale j. We assume block size B is odd. In the search, motion vectors are chosen from a window of size $W > 0$ such that $-W \le d_x, d_y \le W$. The evolution of a regular triangle mesh over frames is illustrated in Fig. 4.

For an N-frame video sequence, this ME process results in $N-1$ motion fields regardless of the temporal filter used, as illustrated in Fig. 5. We note that this is the same number of motion fields produced by a traditional coder with a MC feedback loop. Since each of the $N-1$ motion fields are mesh-based and thus completely invertible, forward and backward motion fields between each pair of frames can be calculated from these $N-1$ fields. Using these forward and backward motion fields, affine transforms between the triangles of each pair of frames are used to implement a motion-compensated lifting-based filtering in the temporal direction. This temporal filtering proceeds by mapping each triangle in a reference frame into the current frame using an affine transform on each triangle in each subband separately. Bilinear interpolation between the surrounding four integer-pixel locations is used to resolve subpixel positions produced by the affine mapping.

We can construct affine transforms between any two frames by concatenating motion fields from the set of $N-1$ motion fields produced by the above ME process. Multiple-scale temporal transforms are thereby supported since these concatenated motion fields can be used for affine transforms at the higher temporal decomposition scales, as is illustrated in Fig. 5.

We have found it beneficial to periodically "reset" the triangle mesh rather than allow the ME of the control points to continue indefinitely. Specifically, we track control-point motion for N' frames and then reset the triangle mesh to the initial uniform mesh (again by diagonally splitting square blocks). We repeat this procedure for the next N' frames. Fig. 5 illustrates the motion-tracking and mesh-resetting procedure for $N' = 4$ and $N = 8$.

B. Subpixel Accuracy

The above ME procedure assumes integer-valued motion vectors, while MC in the form of the affine transform employs interpolation between integer-pixel values to resolve subpixel positions arising in the mapping. In this section, we describe modifications to the above ME/MC approach in order to accommodate subpixel accuracy.

When the motion-vector resolution is increased to half-pixel accuracy, the motion-vector search is first carried out as described above for integer-pixel accuracy. Then, the eight neighboring locations at a distance of $\left(\pm\frac{1}{2}, \pm\frac{1}{2}\right)$ from the best match location are searched to refine the motion vector to half-pixel resolution. We note that, in the case that a motion vector points to a location on the half-pixel grid in one frame, the initial integer-valued search for the motion of the control point into the next frame involves half-pixel locations. In this case, x and y in (25) and (26) will refer to half-pixel locations while d_x and d_y will be integer-valued. The subsequent refinement search will involve both integer- and half-pixel locations. This refinement process is recursively repeated to achieve quarter- and eighth-pixel ME accuracy.

Recall that, for ME/MC involving triangle meshes, subpixel accuracy is invoked in the affine-transform mapping of the MC process. In the subpixel 3D-RWMH system, we use values on the subpixel grid as the basis of the interpolation of the affine mapping. Specifically, the affine-transform mapping from one triangle to another uses bilinear interpolation applied to four nearest locations on the subpixel grid. In practice, we achieve subpixel accuracy for both ME and MC by interpolating the entire RDWT subband both horizontally and vertically. Afterward, ME of the control points and MC with the affine transform are carried out as if on the integer-pixel grid, and the resulting residual subbands are downsampled to their original size.

V. Experimental Results

We now present a series of experiments designed to investigate the performance of the 3D-RWMH technique in practice. In Sec. V-A, we first present simplified simulations designed to gauge the noise-reduction performance of RWMH predicted by the analysis of Sec. III-B and embodied in Fig. 3 when RWMH is applied to real MC noise signals. Then, in Sec. V-B, we turn our attention to real system performance. Throughout this section, we focus on grayscale sequences. Spatial RDWTs use the popular 9-7 biorthogonal filter with $J = 3$ decomposition scales and symmetric extension at the image boundaries. For control-point ME, a block size of $B = 17$ is used, and motion vectors are searched in a window of size $W = 15$. To code motion vectors, the H.261 variable-length-code (VLC) table for motion-vector data (MVD) is used for coding the integer part of the motion vectors, while any subpixel part of the vectors is sent by appending a fixed-length code to each Huffman codeword.

A. Results in Support of the Analysis

In Sec. III-B, in order to quantify the gain due to RWMH in the form of γ_h and γ_l, we made the assumption that the noise not captured by the motion model, $N^{(k)}[x, y, t]$, was white, permitting the invocation of (18) to quantify the reduction in noise variance due to the projection implicit in the multiple-phase inverse RDWT. The assumption of white noise is an obvious oversimplification since real MC noise signals will likely possess a significant degree of correlation, both between subbands as well as spatially within each subband.

However, we have observed empirically that $N^{(k)}[x, y, t]$ does in fact undergo a substantial reduction in variance in practice due to the inverse transform despite this oversimplification. For example, Table I gives γ_h and γ_l values calculated on real MC noise signals between two frames of the "Football" sequence. To measure the values in Table I, a uniform triangular mesh is created in the reference frame by diagonal subdividing of $D \times D$ blocks, and the control points of this mesh are tracked into the current frame, where this ME process is conducted with both reference and current frames in the spatial domain. The resulting meshes are then used to map between the two frames via affine triangular transforms as described in Sec. IV-A. By using the same meshes for both the spatial-domain and RDWT-domain MC processes to follow, we effectively divorce the effects of ME from that of RWMH.

To empirically estimate γ_h, a motion-compensated frame is calculated in the RDWT domain and subtracted from the RDWT-domain current frame to create an RDWT-domain highpass frame as indicated by (4). The equivalent process, using the same ME meshes, is carried out in the spatial domain, creating a spatial-domain highpass frame. The RDWT-domain highpass frame is then mapped to the spatial domain via the multiple-phase inverse RDWT, and γ_h is calculated via (19) using the ratio of the variances of the two highpass frames. Table I shows the resulting γ_h values for a variety of values of D and ME accuracy β.

Empirically estimating γ_l is somewhat less straightforward. First, the RDWT-domain lowpass frame is created via the update step of (5) using the RDWT-domain highpass frame and the same meshes as before. An equivalent spatial-domain lowpass frame is created through the same process applied in the spatial domain. (16) indicates that the variance of the lowpass frame is due to both the MC noise signal as well as the video signal itself. Consequently, we estimate γ_l as

$$\gamma_l \approx 10 \log_{10} \left(\frac{\nu_l^{(\mathrm{RD})} - \nu_l^{(\mathrm{SD})}}{\nu_0} \right), \tag{27}$$

where $\nu_l^{(\mathrm{RD})}$ is the variance of the RDWT-domain lowpass frame after having been mapped to the spatial domain, $\nu_l^{(\mathrm{SD})}$ is the spatial-domain lowpass-frame variance, and ν_0 is an estimate of the variance of the spatial-domain noise signal, taken as the variance of the spatial-domain highpass frame calculated above. In (27), we are assuming that the component of lowpass-frame variance due to the video signal itself (Λ_{ss} in (16)) is the same for both the RDWT-domain and spatial-domain lowpass frames such that the difference of the variances rejects it from the calculation of γ_l. Table I shows the resulting γ_l values for a variety of values of D and ME accuracy β.

Table I offers a rough confirmation of the analysis of

Sec. III-B and Fig. 3. Specifically, we see that RWMH does in fact reduce the noise variance in both the highpass and lowpass frames despite the fact that the white-noise assumption made in Sec. III-B does not strictly hold in practice. We see that decreasing D (i.e., increasing the number of mesh triangles) results in a significant decrease in noise variance (smaller γ_h and γ_l); this is as expected since we anticipate that $N^{(k)}[x, y, t]$ will be more noise-like, at least spatially, the finer the triangle mesh becomes. However, we have observed that the mesh-based MC process maintains a significant degree of correlation structure in $N^{(k)}[x, y, t]$ despite the size of D which we believe is the reason that the γ_h and γ_l values in Table I do not quite reach the 7-dB performance limit predicted by Fig. 3. We see, however, that the reduction in highpass-frame variance does tend to increase as the ME process becomes more accurate (β decreases), which is consistent with Fig. 3.

Before moving on to consider real system performance, we note that, although Table I suggests we use as small a D as possible for maximum RWMH variance reduction, small values of D (i.e., a large number of triangles in the mesh) are impractical due to the large motion-vector overhead and dramatically increased computational complexity they entail. Consequently, for the remainder of this section, we focus on meshes with $D = 16$, corresponding roughly to the number of motion vectors typically used in block-based systems in which a macroblock size of 16×16 pixels is common.

B. System Performance

It is clear that the reduction in highpass and lowpass noise variances, γ_h and γ_l, do not necessarily translate into improved rate-distortion performance in practice. Therefore, we now verify RWMH performance with actual coding results. We use the grayscale sequences shown in Table II in our experiments, and we focus on temporal filtering with the lifting 5-3 biorthogonal filter of (23) and (24) with $J = 3$ decomposition scales and symmetric extension at each end of the video sequence. The triangle mesh driving this temporal filtering is reset to the uniform mesh every $N' = 4$ frames; this uniform mesh is the result of diagonal subdividing of 16×16 blocks. Since 3D-SPIHT [1]—the core compression engine in the 3D-RWMH system—produces an embedded coding, the sequence is coded at exactly the specified target rate.

Initially, we focus on ME with integer-pixel accuracy. We compare the rate-distortion performance of the 3D-RWMH system to an equivalent spatial-domain MCTF system. The spatial-domain technique (denoted "SD-MCTF") performs MCTF in the spatial domain and then subsequently employs a critically sampled spatial transform and embedded coding. In this system, a triangle-mesh ME procedure identical to that of the 3D-RWMH coder is employed, and, like the 3D-RWMH system, temporal decomposition takes place with 5-3 biorthogonal lifting with symmetric extension. This SD-MCTF system is essentially a single-hypothesis version of the 3D-RWMH coder and corresponds roughly to the system of [2], except that the ME process is somewhat different, and 3D-SPIHT, rather than JPEG-2000, is used to code the wavelet coefficients.

The analysis derived in Sec. III-B and embodied by Fig. 3 predicts that the performance gain of the multihypothesis MCTF of 3D-RWMH over the spatial-domain system depends on how much residual noise is not captured by the triangular-mesh motion model. The empirical results presented in Table II bear out this analysis—for sequences in which the motion model is sufficient for nearly all motion present (e.g., the "Foreman" sequence), a modest gain on the order of 0.1 dB is observed; however, for the sequences with more complex motion (e.g., the "Football" sequence), the power of the noise not captured by the mesh model is relatively large, and a more significant gain on the order of 0.5 dB over the spatial-domain system is seen. Figs. 6 and 7 illustrate that these observations hold over a range of rates.

The triangular-mesh motion model used in these empirical results is quite powerful and leaves only a modest amount of residual noise uncaptured for many sequences. However, the motion model does fail in certain situations. In Fig. 8, we have contrived an example of such model failure. Specifically, we have interleaved 16-frame subsequences from the "Football" and "Susie" sequences to simulate scene changes for which the mesh model results in substantial residual noise. For this interleaved sequence, we see that the gain of 3D-RWMH approaches 1 dB or more, again bearing out the analysis of Sec. III-B.

We next gauge the performance of the 3D-RWMH system against two state-of-the art coders. Specifically, we compare to the bidirectional MC-EZBC system from [7] as well as to H.264 [20]. Bidirectional MC-EZBC is a 3D video coder employing traditional block-based MCTF in the spatial domain and is largely considered the state-of-the-art for such fully scalable coders. Temporal filtering is essentially a bidirectional version of the Haar filter, with a lifting implementation providing $\frac{1}{8}$-pixel accuracy for ME/MC and appropriate measures to compensate for "unconnected" pixels. On the other hand, H.264 is the latest standard coder and is considered state-of-the-art for non-scalable coding based on the traditional hybrid architecture; we use H.264 JM 9.2 operating at High Profile, Level 4 with all advanced coding modes activated, and a frame-coding pattern of *IBPBPB*.... Experimental results for these two systems, along with those for 3D-RWMH using half-pixel accuracy in the ME of mesh control points, are presented in Table II. Although the results are mixed, we observe that 3D-RWMH with half-pixel accuracy usually offers performance slightly better than that of MC-EZBC and occasionally outperforms H.264.

As a final body of results, we compare the spatial scalability of 3D-RWMH and MC-EZBC, both fully scalable coders. Although our 3D-SPIHT implementation is not technically resolution-scalable, we simulate reduced spatial resolution in these experiments by decoding only the bits from the compressed bitstream necessary for reconstruction at the particular spatial resolution of interest; that is, only some subset of the spatial subbands is reconstructed. Then, the inverse MCTF is performed on only the available subbands. A coefficient coder with greater scalability support, such as JPEG-2000, would permit achieving such scalable reconstruction in practice. To quantitatively measure performance at reduced spatial resolu-

tion, we calculate a PSNR with respect to a reduced-resolution version of the original sequence obtained via discarding subbands from a spatial 2D DWT of each frame of the sequence. Although this is not perhaps the best manner in which to create a reduced-resolution sequence—the spatial DWT offering less than ideal anti-aliasing performance—one would expect that wavelet-based scalable coders such as 3D-RWMH and MC-EZBC would not perform any "better" than such a sequence at reduced-resolution, and so it constitutes a reasonable benchmark with which to compare the relative performance of the two algorithms. The resulting rate-distortion performance of 3D-RWMH with half-pixel-accurate ME and MC-EZBC is graphed in Fig. 9 for decoding taking place at full-SIF, quarter-SIF (QSIF), and quarter-quarter-SIF (QQSIF) resolutions. We observe that the spatial-scalability performance of 3D-RWMH is roughly comparable to that of MC-EZBC, with 3D-RWMH outperforming MC-EZBC somewhat at lower rates. Sample reconstructed images are shown in Fig. 10, and we observe that, perceptually, the two techniques perform similarly at reduced spatial resolution.

VI. Conclusions

In this paper, we presented a system that introduced phase-diversity multihypothesis into MCTF by deploying MCTF in the domain of a redundant wavelet transform and exploiting the transform redundancy to provide multiple hypothesis temporal filterings that were diverse in transform phase. The centerpiece of the work we reported here was an analysis of the performance of the proposed RWMH-based MCTF under the assumption of a simple translational motion model and simple Haar-based MCTF. Key to this analysis was the fact that noise in the RDWT domain undergoes a substantial reduction in variance when the multiple-phase inverse RDWT at the heart of RWMH is applied due to the well-known fact that this pseudo-inverse contains a projection onto the range space of the forward transform. Consequently, noise not captured by the motion model is greatly reduced in an RWMH system, leading to substantial reduction in both the overall variance of the MCTF highpass frame and the variance of the component of the lowpass frame due to noise. In fact, our analysis predicted that MCTF using RWMH can reduce both of these variances by up to 7 dB compared to those of an equivalent spatial-domain system.

For implementation purposes, we departed from the block-based motion models commonly employed in MCTF by implementing temporal filtering with mesh-based lifting. In essence, this 3D-RWMH implementation combined the flexibility and scalability of the mesh-based lifting MCTF of [2, 3] with the performance gains associated with RWMH. Additionally, performance was enhanced by using half-pixel accuracy for both describing the motion of the mesh as well as within the affine transforms implementing the MCTF.

Experimental results presented support the theoretical analysis and demonstrate state-of-the-art coding performance. Specifically, it was observed that the proposed 3D-RWMH system outperformed an equivalent system using spatial-domain MCTF, particularly so as noise not captured by the motion model increased, as predicted by our analysis. Additionally, we observed the 3D-RWMH system to usually perform slightly better than MC-EZBC [7], a state-of-the-art fully scalable coder with MCTF operating in the spatial domain. Additionally, 3D-RWMH occasionally outperformed H.264 [20], the current state of the art in non-scalable hybrid coding. Finally, we observed that 3D-RWMH offered quantitative and perceptual performance for scalable decoding at reduced spatial resolution equivalent to that of MC-EZBC.

Acknowledgment

The authors thank Z. Xiong for providing the implementation of 3D-SPIHT from [1] and J. B. Boettcher for aid in obtaining experimental results.

References

[1] 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.

[2] 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.

[3] ——, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1530–1542, December 2003.

[4] ——, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proceedings of the International Conference on Image Processing*, vol. 2, Thessaloniki, Greece, October 2001, pp. 1029–1032.

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

[6] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 155–167, February 1999.

[7] 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.

[8] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Salt Lake City, UT, May 2001, pp. 1793–1796.

[9] D. S. Turaga and M. van der Schaar, "Wavelet coding for video streaming using new unconstrained motion compensated temporal filtering," in *Proceedings of the 2002 Tyrrhenian International Workshop on Digital Communications (IWDC 2002): Advanced Methods for Multimedia Signal Processing*, Capri, Italy, September 2002, pp. 41–48.

[10] M. van der Schaar and D. S. Turaga, "Unconstrained motion compensated temporal filtering (UMCTF) framework for wavelet video coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Hong Kong, April 2003, pp. 81–84.

[11] D. S. Turaga, M. van der Schaar, Y. Andreopoulos, A. Munteanu, and P. Schelkens, "Unconstrained motion compensated temporal filtering (UMCTF) for efficient and flexible interframe wavelet video coding," *Signal Processing: Image Communication*, vol. 20, pp. 1–19, January 2005.

[12] J. C. Ye and M. van der Schaar, "Fully scalable 3-D overcomplete wavelet video coding using adaptive motion compensated temporal filtering," in *Visual Communications and Image Processing*, T. Ebrahimi and T. Sikora, Eds. Lugano, Switzerland: Proc. SPIE 5150, July 2003, pp. 1169–1180.

[13] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelius, "Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MTCF," in *Visual Communications and Image Processing*, T. Ebrahimi and T. Sikora, Eds. Lugano, Switzerland: Proc. SPIE 5150, July 2003, pp. 719–731.

[14] A. Munteanu, Y. Andreopoulos, M. van der Schaar, P. Schelkens, and J. Cornelius, "Control of the distortion variation in video coding systems based on motion compensated temporal filtering," in *Proceedings of the International Conference on Image Processing*, vol. 2, Barcelona, Spain, September 2003, pp. 61–64.

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

[16] A. V. Golwelkar and J. W. Woods, "Scalable video compression using longer motion compensated temporal filters," in *Visual Communications and Image Processing*, T. Ebrahimi and T. Sikora, Eds. Lugano, Switzerland: Proc. SPIE 5150, July 2003, pp. 1406–1416.

[17] 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.

[18] ——, "Motion compensation via redundant-wavelet multihypothesis," *IEEE Transactions on Image Processing*, submitted March 2004, revised February 2005, December 2005.

[19] J. E. Fowler, "Analysis of redundant-wavelet multihypothesis for motion compensation," in *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. Cohn, Eds., Snowbird, UT, March 2006.

[20] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T, May 2003, ITU-T Recommendation H.264.

[21] 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.

[22] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, February 1999.

[23] 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.

[24] 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.

[25] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 7, pp. 1140–1154, August 1987.

[26] ——, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Transactions on Communications*, vol. 41, no. 4, pp. 604–612, April 1993.

[27] J. E. Fowler, "The redundant discrete wavelet transform and additive noise," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 629–632, September 2005.

[28] Y. Nakaya and H. Harashima, "Motion compensation based on spatial transformation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 3, pp. 339–366, June 1994.

[29] M. Eckert, D. Ruiz, J. I. Ronda, and N. Garcia, "Evaluation of DWT and DCT for irregular mesh-based motion compensation in predictive video coding," in *Visual Communications and Image Processing*, K. N. Ngan, T. Sikora, and M.-T. Sun, Eds. Proc. SPIE 4067, June 2000, pp. 447–456.

[30] S. Cui, Y. Wang, and J. E. Fowler, "Mesh-based motion estimation and compensation in the wavelet domain using a redundant transform," in *Proceedings of the International Conference on Image Processing*, vol. 1, Rochester, NY, September 2002, pp. 693–696.

PLACE PHOTO HERE

Yonghui Wang received the B.S. degree in technical physics from Xidian University, Xi'an, China, and the M.S. degree in electrical engineering from Beijing Polytechnic University, Beijing, China, in 1993 and 1999, respectively. In December 2003, he received the Ph.D. degree in computer engineering from Mississippi State University, Starkville, MS. From 1993 to 1996, he worked as an engineer at the 41st Electrical Research Institute, Bengbu, China. From 2000 to 2003, he was a research assistant in the Visualization, Analysis, and Imaging Laboratory (VAIL) within the GeoResources Institute (GRI) at Mississippi State. He is currently an assistant professor in the Department of Engineering Technology at Prairie View A&M University in Prairie View, TX. His research interests include image- and signal-processing, image- and video-coding.

PLACE PHOTO HERE

Suxia Cui was born in Beijing, China. She received the B.S. degree and the M.S. degree in electrical engineering from Beijing Polytechnic University, Beijing, China, in 1996 and 1999, respectively. She received the Ph.D. degree in computer engineering from Mississippi State University, Starkville, MS, in August 2003. From 2000 to 2003, she was a research assistant in the Visualization, Analysis, and Imaging Laboratory (VAIL) within the GeoResources Institute (GRI) at Mississippi State. She is currently an assistant professor in the Department of Engineering Technology at Prairie View A&M University in Prairie View, TX. Her research interests include image processing, video coding, and wavelets.

PLACE PHOTO HERE

James E. Fowler received the B.S. degree in computer and information science engineering and the M.S. and Ph.D. degrees in electrical engineering in 1990, 1992, and 1996, respectively, all from the Ohio State University. In 1995, Dr. Fowler was an intern researcher at AT&T Labs in Holmdel, NJ, and, in 1997, he held an NSF-sponsored postdoctoral assignment at the Université de Nice-Sophia Antipolis, France. In 2004, he was a visiting professor in the Département TSI at École Nationale Supérieure des Télécommunications (ENST), Paris, France. He is currently an associate professor in the Department of Electrical & Computer Engineering at Mississippi State University in Starkville, MS, and is also a researcher in the Visualization, Analysis, and Imaging Laboratory (VAIL) within the GeoResources Institute (GRI) at Mississippi State. Dr. Fowler is an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.

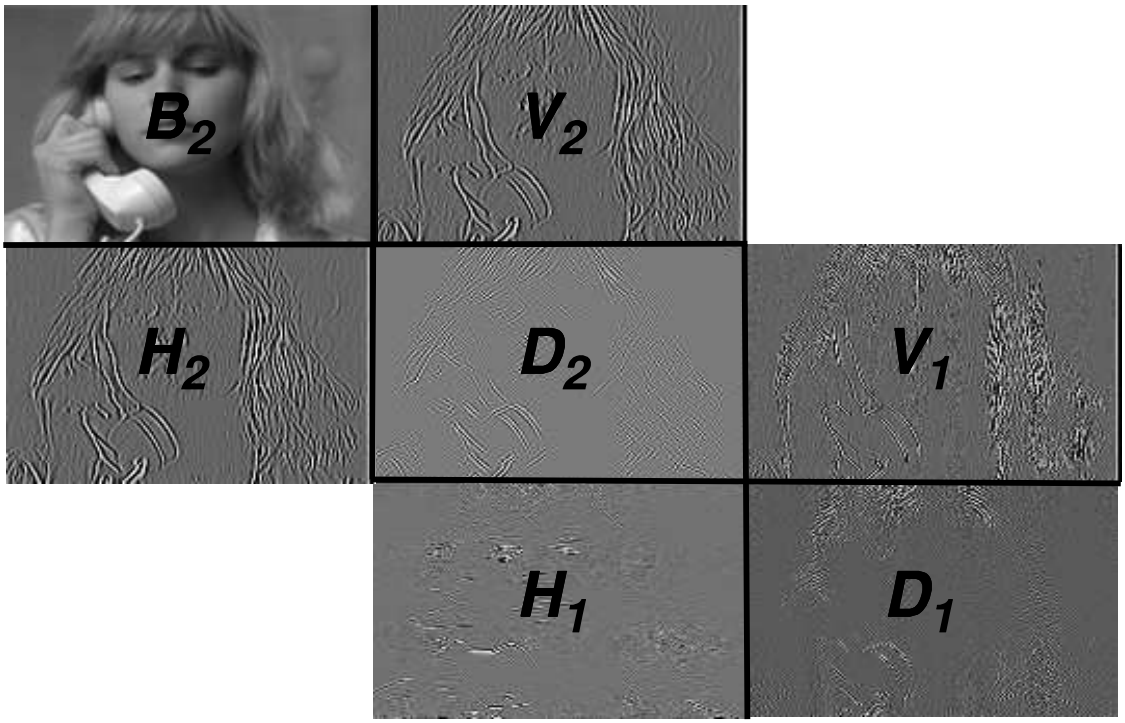

Fig. 1. A two-scale RDWT of a 2D image.

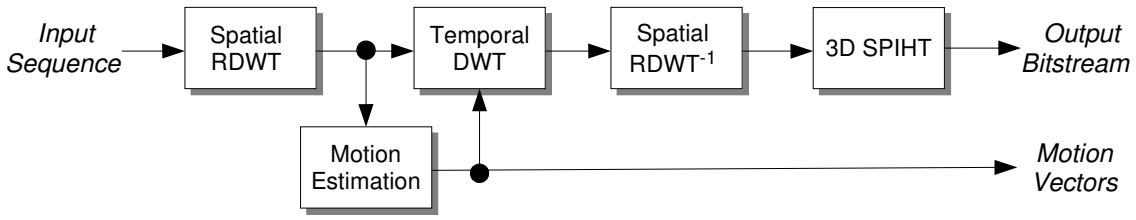

Fig. 2. The 3D-RWMH video-coding system.

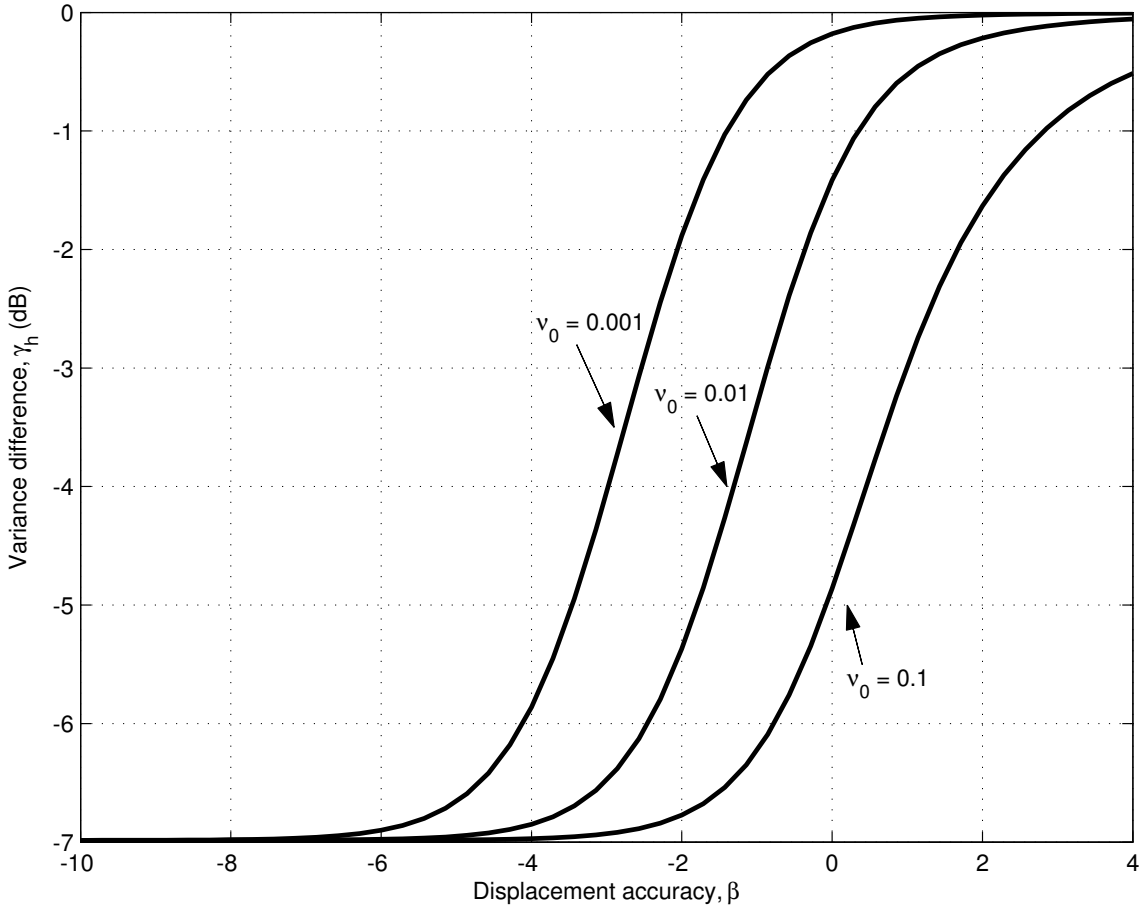

Fig. 3. Difference, γ_h, in highpass-frame variance between the RWMH and spatial-domain systems as the displacement accuracy β varies at various noise variances ν_0.

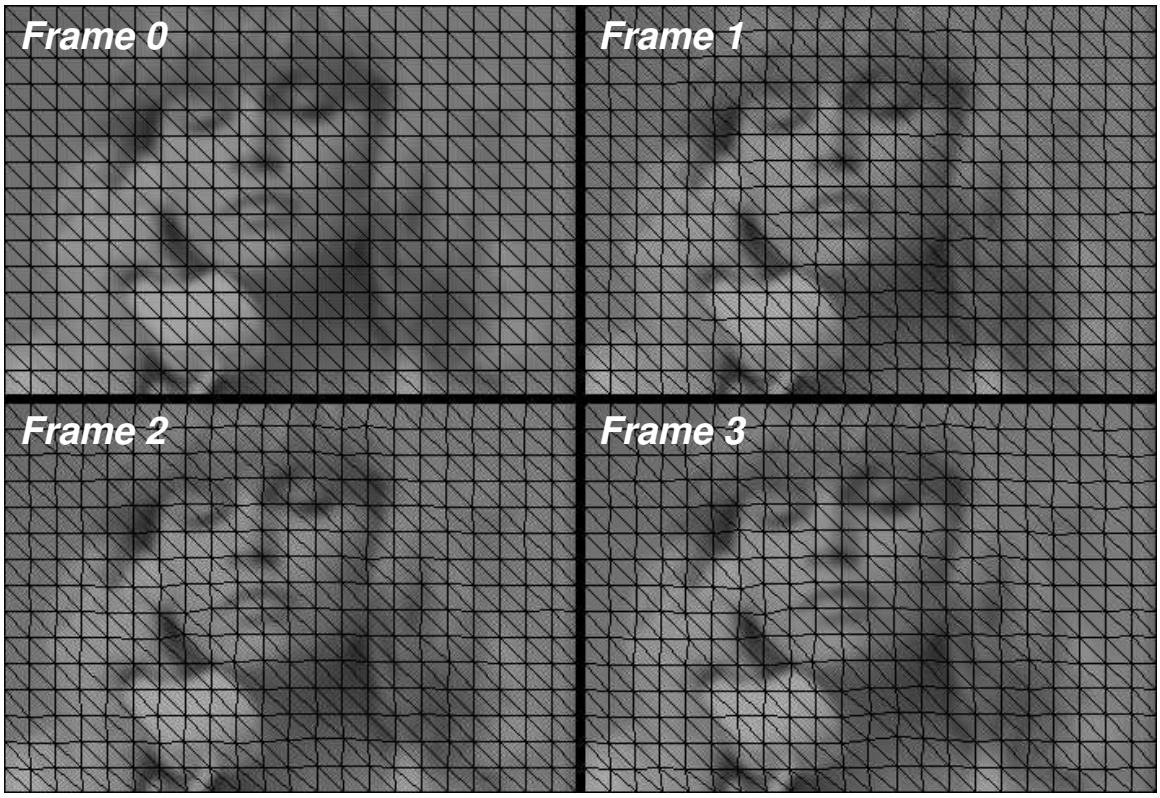

Fig. 4. The regular triangle mesh in the first frame and its evolution over subsequent frames. Only the basebands of the frames are shown.

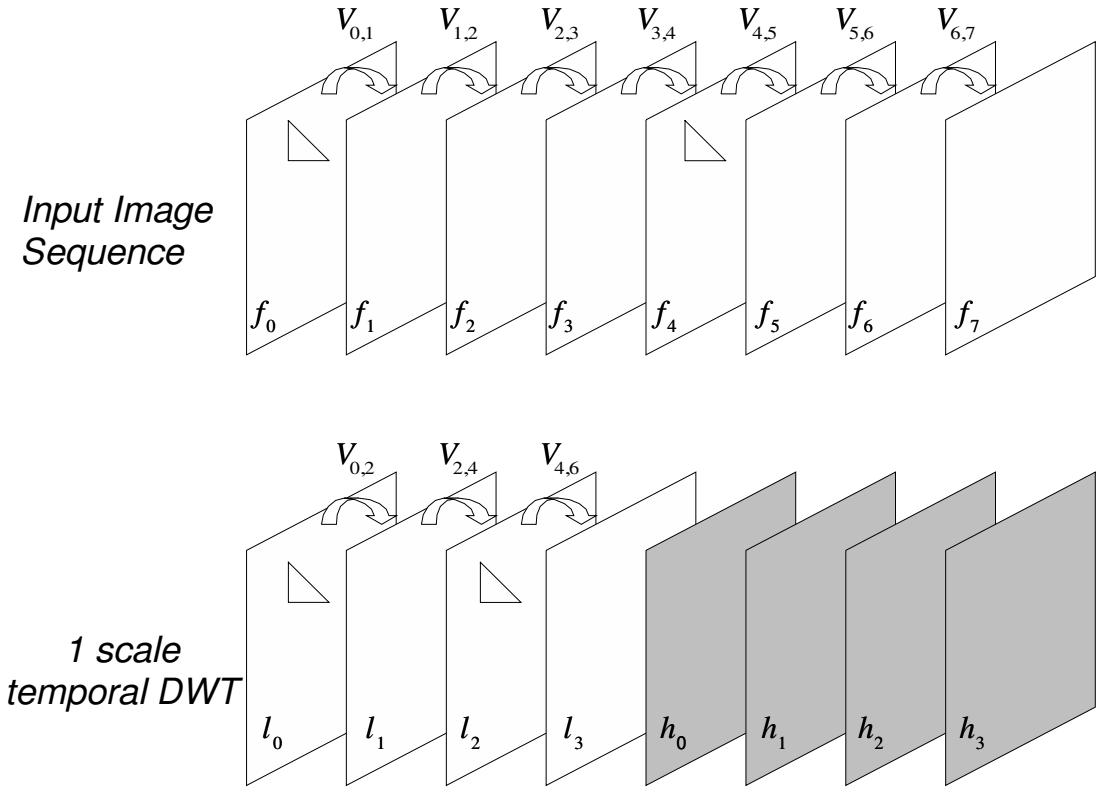

Fig. 5. Motion fields for MCTF in the 3D-RWMH system for $N = 8$ and $N' = 4$. A small triangle in a frame indicates where the triangle mesh is reset to the uniform mesh. Field $V_{i,j}$ maps frame i to frame j; concatenated fields are $V_{0,2} = V_{0,1} + V_{1,2}$, $V_{2,4} = V_{2,3} + V_{3,4}$, and $V_{4,6} = V_{4,5} + V_{5,6}$.

TABLE I

γ_h AND γ_l CALCULATED BETWEEN FRAMES 50 AND 51 OF THE "FOOTBALL" SEQUENCE USING A UNIFORM MESH OF BLOCK SIZE $D \times D$ IN THE REFERENCE FRAME AND ME ACCURACY OF β.

	γ_h (dB)				γ_l (dB)			
D	$\beta = 0$	$\beta = -1$	$\beta = -2$	$\beta = -3$	$\beta = 0$	$\beta = -1$	$\beta = -2$	$\beta = -3$
2	-1.55	-1.45	-1.72	-1.84	-4.27	-3.51	-4.57	-4.96
4	-1.07	-1.08	-1.29	-1.33	-2.31	-2.31	-2.73	-2.97
8	-0.75	-0.75	-0.87	-0.92	-1.98	-1.95	-2.30	-2.37
16	-0.44	-0.46	-0.51	-0.53	-1.06	-1.07	-1.15	-1.19

TABLE II

DISTORTION AVERAGED OVER ALL FRAMES OF THE SEQUENCE.

	PSNR (dB)				
	SD-MCTF integer pixel	3D-RWMH integer pixel	3D-RWMH 1/2 pixel	MC-EZBC 1/8 pixel	H.264
Football	29.3	**29.8**	30.0	29.6	**33.8**
Mother & daughter†	46.7	**47.1**	**47.5**	**47.5**	47.3
Susie	42.0	**42.3**	43.1	43.0	**44.1**
NYC	39.6	**39.8**	41.0	41.2	**42.5**
Foreman†	39.9	**40.0**	41.0	40.7	**43.7**
Coastguard†	**34.7**	**34.7**	35.2	34.6	**35.5**

Sequences are SIF (352×240, 30 Hz) at 1267 kbps, except †which are CIF (352×288, 30 Hz) at 1520 kbps.

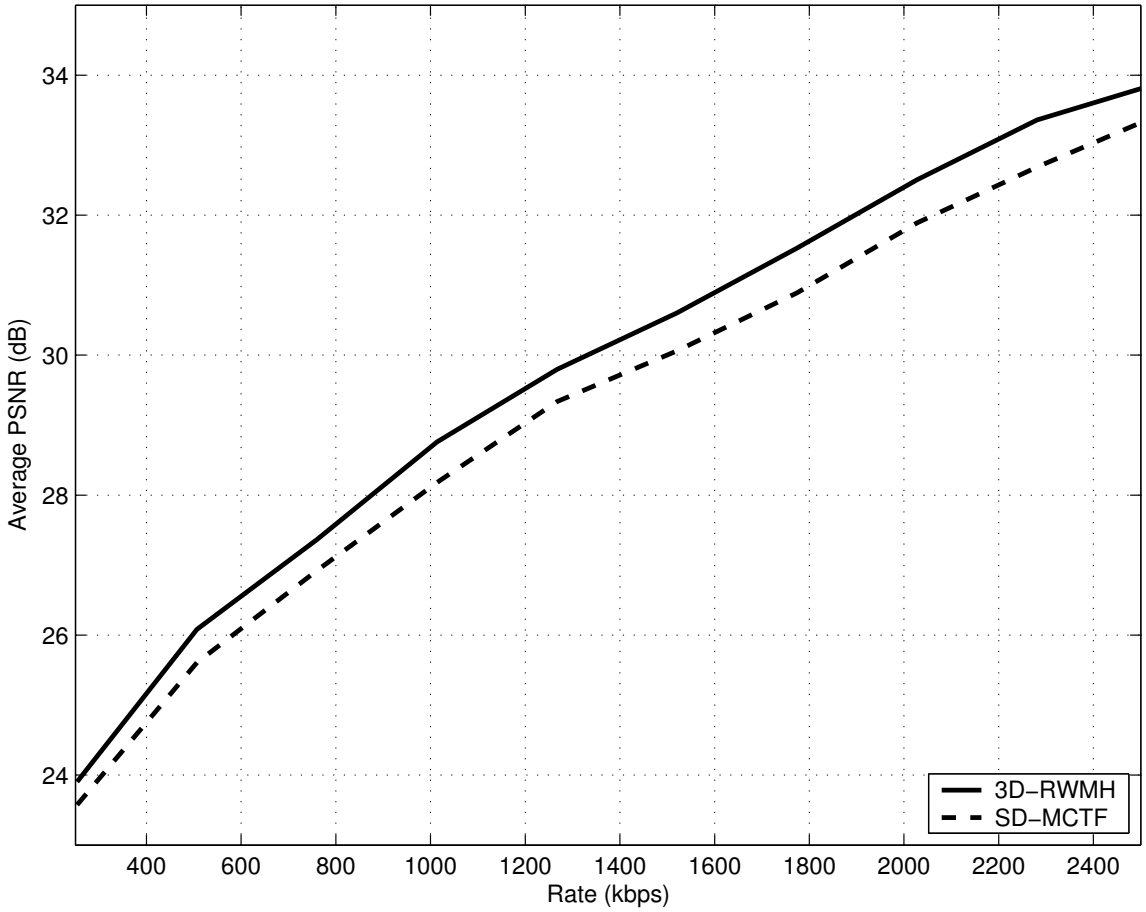

Fig. 6. Rate-distortion performance for "Football."

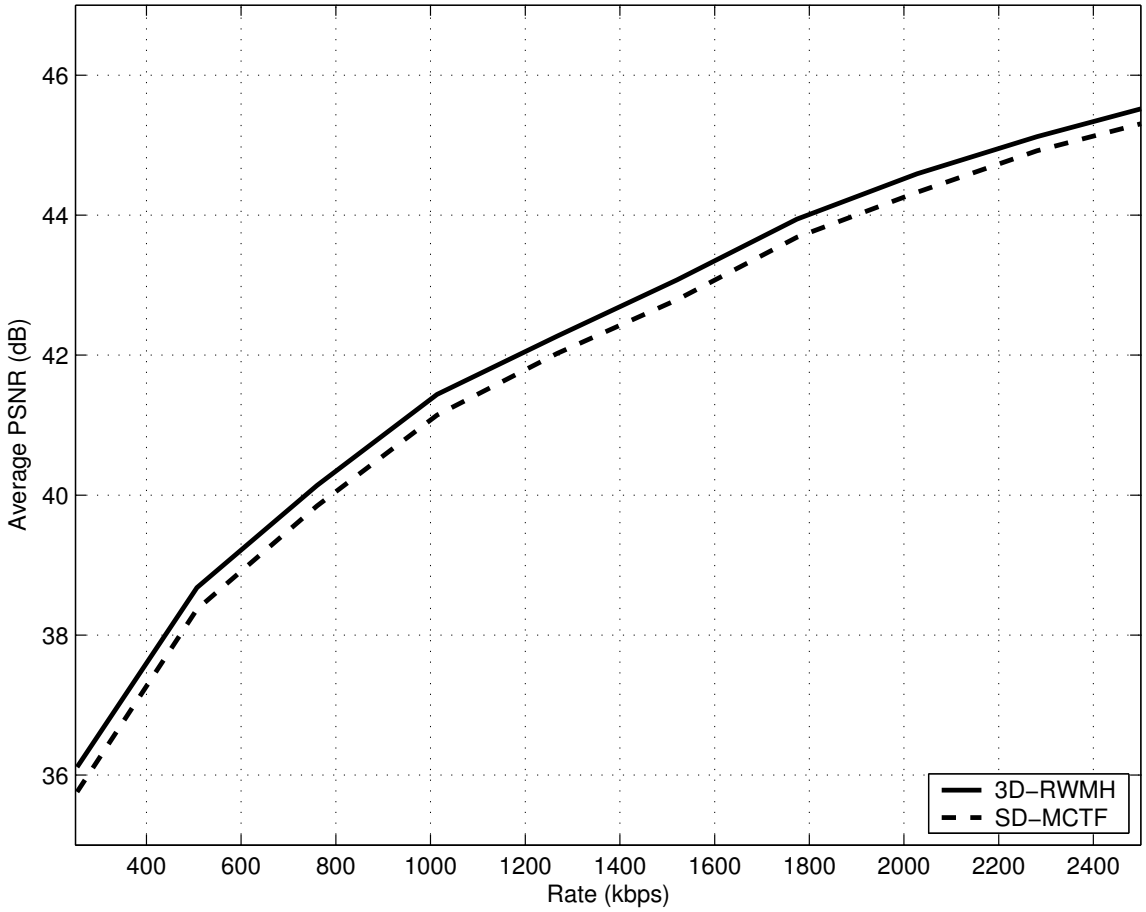

Fig. 7. Rate-distortion performance for "Susie."

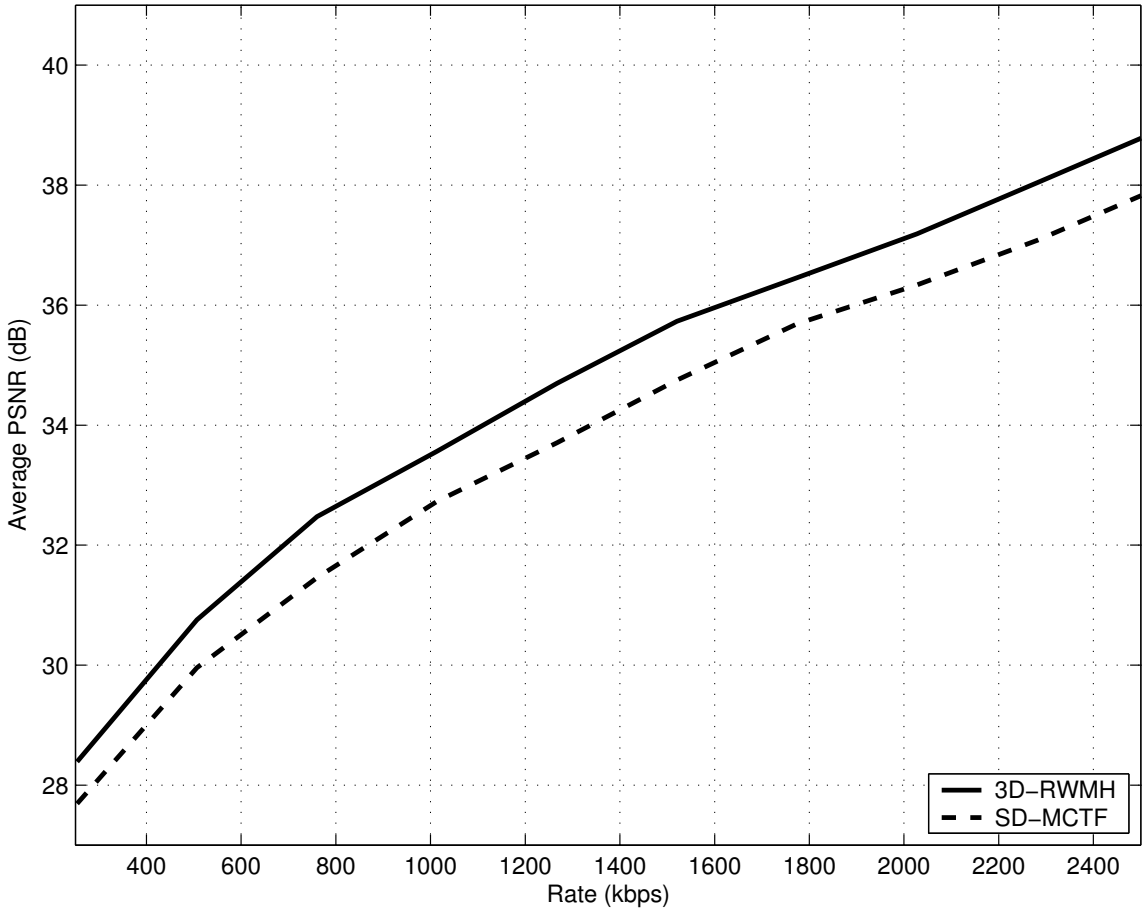

Fig. 8. Rate-distortion performance for "Football/Susie" interleaved sequence.

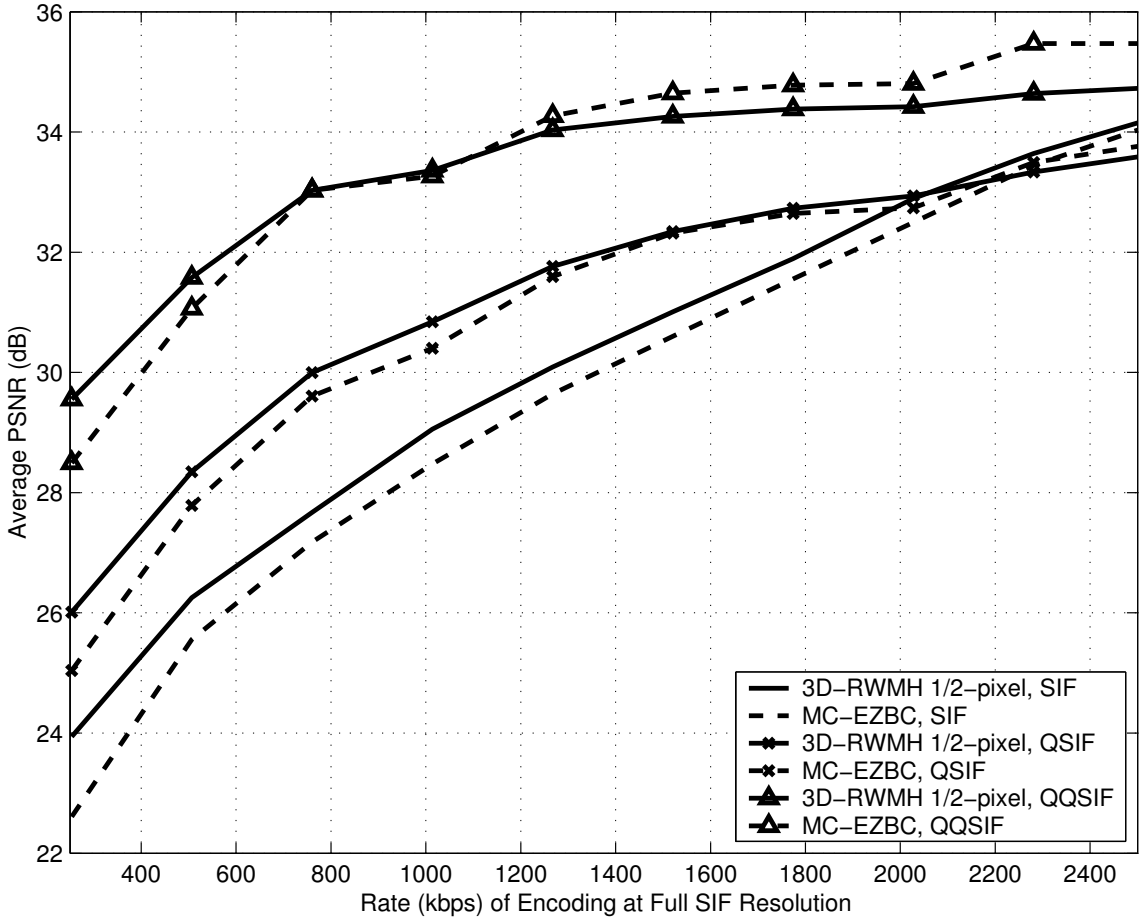

Fig. 9. Rate-distortion performance for "Football" decoded at SIF, QSIF (176×120), and QQSIF (88×60) resolutions.

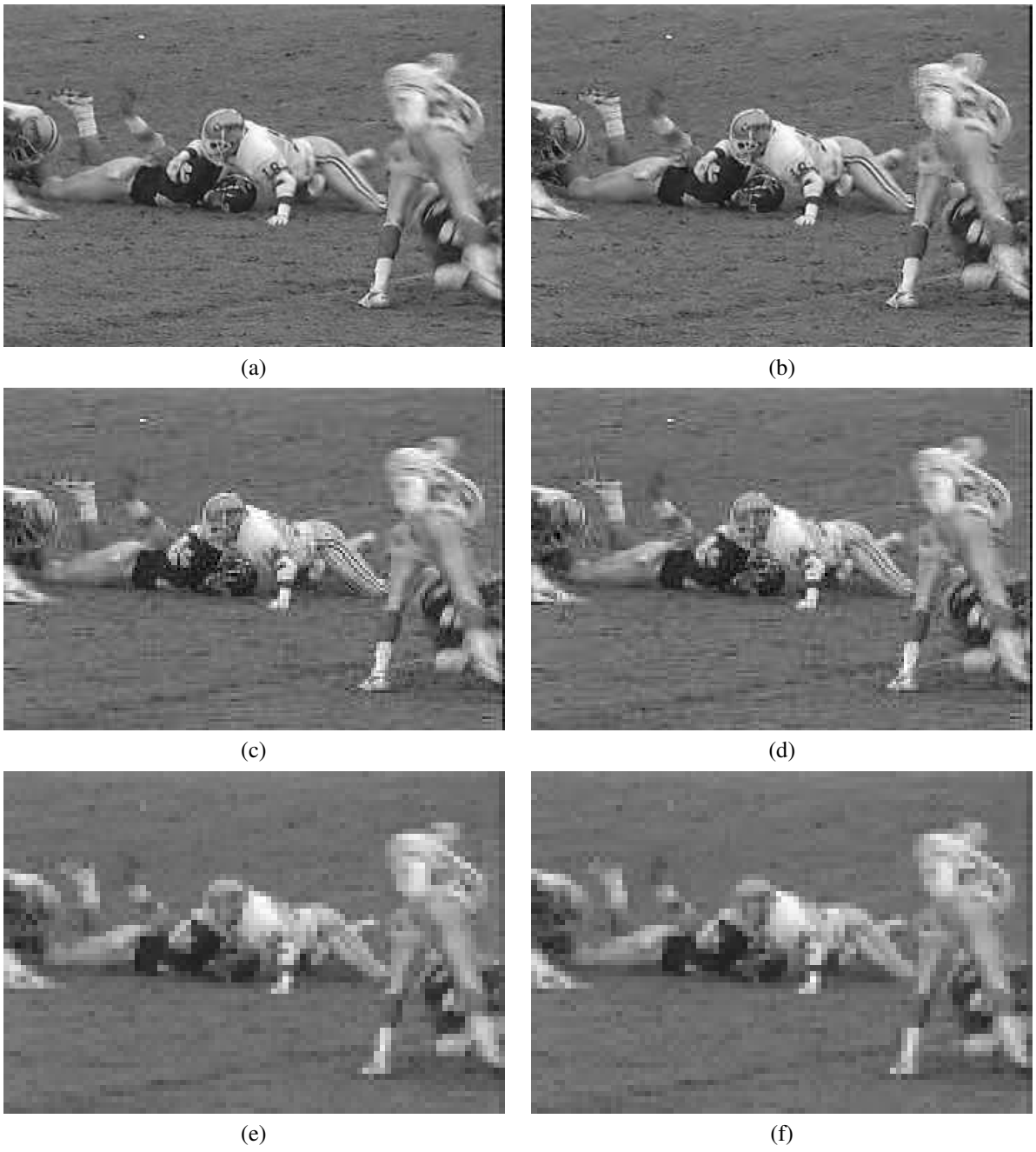

Fig. 10. Reconstructed images for frame 5 of "Football" encoded at 1267 kbps and decoded at reduced spatial resolution. (a) 3D-RWMH 1/2-pixel, SIF; (b) MC-EZBC, SIF; (c) 3D-RWMH 1/2-pixel, QSIF; (d) MC-EZBC, QSIF; (e) 3D-RWMH 1/2-pixel, QQSIF; (f) MC-EZBC, QQSIF.