Motion Compensation Via Redundant-Wavelet Multihypothesis

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

***Abstract*— Multihypothesis motion compensation has been widely used in video coding with previous attention focused on techniques employing predictions that are diverse spatially or temporally. In this paper, the multihypothesis concept is extended into the transform domain by using a redundant wavelet transform to produce multiple predictions that are diverse in transform phase. The corresponding multiple-phase inverse transform implicitly combines the phase-diverse predictions into a single spatial-domain prediction for motion compensation. The performance advantage of this redundant-wavelet-multihypothesis approach is investigated analytically, invoking the fact that the multiple-phase inverse involves a projection that significantly reduces the power of a dense-motion residual modeled as additive noise. The analysis shows that redundant-wavelet multihypothesis is capable of up to a 7-dB reduction in prediction-residual variance over an equivalent single-phase, single-hypothesis approach. Experimental results substantiate the performance advantage for a block-based implementation.**

***Index Terms*— multihypothesis motion compensation, redundant wavelet transform, phase-diversity multihypothesis**

I. Introduction

Multihypothesis motion compensation (MHMC) [1] forms a prediction of pixel $s[\mathbf{p},t]$ at spatial location $\mathbf{p} = \begin{bmatrix} x & y \end{bmatrix}^T$ in the current frame at time t as a combination of multiple predictions in an effort to combat the uncertainty inherent in the motion-estimation (ME) process. Assuming that the combination of these hypothesis predictions is linear, we have that the prediction of frame $s[\mathbf{p},t]$ is

$$\tilde{s}[\mathbf{p},t] = \sum_i w_i[\mathbf{p},t]\tilde{s}_i[\mathbf{p},t], \tag{1}$$

where the multiple predictions $\tilde{s}_i[\mathbf{p},t]$ are combined according to some weights $w_i[\mathbf{p},t]$. A number of multihypothesis techniques for motion compensation (MC) have been proposed over the last decade. One approach to MHMC is to implement multihypothesis prediction in the spatial dimensions; i.e., the predictions $\tilde{s}_i[\mathbf{p},t]$ are culled from spatially distinct locations in the reference frame. Included in this class of MHMC would be subpixel-accurate MC [2] and overlapped block motion compensation (OBMC) [3,4]. Another approach is to deploy MHMC in the temporal dimension by choosing predictions $\tilde{s}_i[\mathbf{p},t]$ from multiple reference frames. Examples of this class of MHMC are bidirectional prediction (B-frames) as used in MPEG-2 and H.263 and long-term-memory motion compensation (LTMMC) [5]. Of course, it is possible to combine these two classes by choosing multiple predictions that are diverse both spatially and temporally [6]. In [7], we introduced a new class of MHMC by extending the multihypothesis-prediction concept 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. We coined the term redundant-wavelet multihypothesis (RWMH) to describe our approach.

In this paper, we present a thorough investigation of the performance advantage of RWMH over single-hypothesis techniques that base ME/MC on merely a single phase, the most prominent of these latter strategies being the system of Park and Kim [8]. The centerpiece of this investigation is an analytical derivation that quantifies the gain of RWMH over single-phase prediction under a model that partitions wavelet-domain motion into a global pan-motion translation as well as a dense motion field. Our RWMH technique substantially reduces the variance of the prediction residual due to the dense-motion component thanks to the robustness of overcomplete transforms to additive noise in the transform domain. This noise robustness leads to a theoretical reduction in the variance of the overall MC prediction residual by up to 7 dB with respect to the single-phase system. A block-based implementation, which approaches the analytical motion model as the block size decreases, exhibits similar variance reduction in experimental results and achieves substantial gains in actual coding performance over an equivalent single-hypothesis system.

The remainder of the manuscript is organized as follows. We start with Sec. II which overviews theory behind the redundant discrete wavelet transform (RDWT) [9–11] necessary to understanding the analysis to follow. In Sec. III, we overview RDWT-based video coding including our RWMH technique. In Sec. IV, we present the main contribution of the paper, an analytical derivation of the gain of RWMH over single-phase prediction. Afterward, we consider some issues relevant to practical implementation of RWMH in a block-based system in Sec. V. We follow with Sec. VI in which we present experimental results. Finally, we make some concluding remarks in Sec. VII.

J. E. Fowler is with the Department of Electrical & Computer Engineering and the GeoResources Institute, Mississippi State University, Starkville, MS 39762 USA (e-mail: fowler@ece.msstate.edu).

Y. Wang and S. Cui were with the Department of Electrical & Computer Engineering and the GeoResources Institute, Mississippi State University, Starkville, MS 39762 USA and are currently with the Department of Engineering Technology at Prairie View A&M University in Prairie View, TX 77446 USA (e-mail: yowang@pvamu.edu; sucui@pvamu.edu).

This work was funded in part by the National Science Foundation under Grants Nos. CCR-0310864 and ACI-9982344. Portions of this work were performed while J. E. Fowler was at École Normale Supérieure des Télécommunications (ENST), Paris, France, as a visiting professor funded in part by the French Centre National de la Recherche Scientifique (CNRS).

IEEE Transactions on Image Processing, vol. 15, pp. 3102-3113, October 2006.

II. The Redundant Wavelet Transform

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

$$\left\{S^{(k)}[\mathbf{p},t]\right\}_k = \mathcal{R}\Big[s[\mathbf{p},t]\Big]. \tag{2}$$

In essence, the RDWT [9–11] removes the downsampling operation from the traditional critically sampled discrete wavelet transform (DWT) to produce an overcomplete representation. While the well-known shift variance of the DWT arises from its use of downsampling, the RDWT is shift invariant to a single translation $\boldsymbol{\delta} = \begin{bmatrix}\delta_x & \delta_y\end{bmatrix}^T$ applied identically to each subband since the spatial sampling rate is fixed across scale. We observe that, if we resolve values off the integer-pixel sampling lattice using an interpolation operator $I(\cdot)$ which is linear and shift-invariant, the RDWT can be considered to be shift invariant even for non-integer translations $\boldsymbol{\delta}$ in the sense that

$$\mathcal{R}\Big[I\Big(s[\mathbf{p}-\boldsymbol{\delta},t]\Big)\Big] = \left\{I\left(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t]\right)\right\}_k. \tag{3}$$

There are several ways to implement the RDWT, and several ways to represent the resulting overcomplete set of coefficients. The most obvious implementation, direct implementation of the *algorithme à trous* [9,10] results in subbands that are exactly the same size as the original signal, as shown for a 2D signal in Fig. 1. The advantage of this "spatially coherent" representation is that each RDWT coefficient is located within its subband in its spatially correct position. As illustrated in Fig. 1, by appropriately subsampling each subband of an RDWT, one can produce exactly the same coefficients as does a critically sampled DWT applied to the same input signal. In fact, in a J-scale 2D RDWT, there exist 4^J distinct critically sampled DWTs corresponding to the choice between even- and odd-phase subsampling at each scale of decomposition.

On the other hand, the most popular coefficient-representation scheme employed in RDWT-based video coders [8,12–17] is that of a "coefficient tree," as illustrated in Fig. 2 for a 1D signal. This tree representation is easily created by employing filtering and downsampling as in the usual critically sampled DWT; however, all "phases" of downsampled coefficients are retained and arranged as "children" of the signal that was decomposed. The process is repeated on the lowpass bands of all nodes to achieve multiple decomposition scales. It is straightforward to see that each path from root to leaf in the RDWT tree constitutes a distinct critically sampled DWT, and there are 4^J such critically sampled DWTs in a J-scale 2D decomposition.

Focusing on the *algorithm à trous* coefficient representation, given subbands $S^{(k)}[\mathbf{p},t]$, there are several methods to invert the RDWT. The *single-phase inverse* consists of subsampling the RDWT coefficients to extract one critically sampled DWT from the RDWT and inverting using the corresponding inverse DWT; i.e.,

$$s'[\mathbf{p},t] = \mathcal{D}^{-1}\left[\downarrow\left\{S^{(k)}[\mathbf{p},t]\right\}_k\right], \tag{4}$$

where $\mathcal{D}^{-1}(\cdot)$ is the inverse DWT, and $\downarrow$ denotes subsampling to a single phase. Alternatively, one can employ a *multiple-phase inverse* which we denote as

$$s''[\mathbf{p},t] = \mathcal{R}^{-1}\left[\left\{S^{(k)}[\mathbf{p},t]\right\}_k\right]. \tag{5}$$

In such a multiple-phase inverse, one independently inverts each of the 4^J critically sampled DWTs constituting the J-scale 2D RDWT and averages the resulting reconstructions together, or one can equivalently employ a more computationally efficient filtering implementation [18]. In either case, we observe that the single-phase inverse (4) is generally not the same as the multiple-phase inverse (5), but the two do yield the same result if the argument is in the range space of $\mathcal{R}(\cdot)$; i.e.,

$$s[\mathbf{p},t] = s'[\mathbf{p},t] = s''[\mathbf{p},t], \tag{6}$$

as long as $\left\{S^{(k)}[\mathbf{p},t]\right\}_k$ in (4) and (5) is the RDWT of some $s[\mathbf{p},t]$. We note that this will not always be the case since the range space of the RDWT is larger than the original signal domain, the RDWT being overcomplete. Finally, like the forward transform, the multiple-phase inverse is shift invariant under linear fractional-pixel interpolation,

$$\mathcal{R}^{-1}\left[\left\{I\left(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t]\right)\right\}_k\right] = I\Big(s''[\mathbf{p}-\boldsymbol{\delta},t]\Big), \tag{7}$$

while the single-phase inverse is not.

Additional understanding of the significance of the inversion of the RDWT results from considering the theory of frames [19]. From this perspective, the RDWT is a frame operator, $\mathcal{R}$, while the multiple-phase inverse RDWT, $\mathcal{R}^{-1}$, is the corresponding pseudo-inverse operator [18]. As a pseudo-inverse, $\mathcal{R}^{-1}$ can be considered to consist of an orthogonal projection onto the range space of $\mathcal{R}$ followed by a mapping into the original signal domain. It has been observed [19] that the fact that an inverse frame operator includes a projection makes frame operators such as the RDWT robust to added noise. Specifically, noise in the RDWT domain, when mapped to the original signal domain via the pseudo-inverse, typically undergoes a significant reduction in variance since the component of the noise orthogonal to the range space of the RDWT is eliminated by the projection. Below, we will see that this projection property of the pseudo-inverse and its effect on noise lie at the heart of our proposed RWMH technique.

III. RDWT-Based Video Coding

The majority of prior work concerning RDWT-based video coding originates in the work of Park and Kim [8], in which the system shown in Fig. 3 was proposed. In essence, the system of Fig. 3 works as follows. An input frame is decomposed with a critically sampled DWT which is matched to an RDWT decomposition of the previous reconstructed frame. Since these reconstructed RDWT coefficients are arranged in the tree representation similar to Fig. 2, the ME procedure of this system amounts to identifying a particular critically sampled DWT in the reference-frame tree (a root-to-leaf path), and a displacement within that DWT. Subsequent work has offered refinements to the system depicted in Fig. 3, such as the deriving of motion vectors for each subband [12,15], or resolution [17], independently; sub-pixel accuracy ME [16]; resolution-scalable coding [13,15,17]; and ME/MC based on triangle meshes that are adapted to the contents of the current frame by exploiting the redundancy of the RDWT via a cross-subband correlation operator [20].

In most of the RDWT-based video-coding systems described above, the redundancy inherent in the RDWT is used exclusively to permit ME/MC in the wavelet domain by overcoming the well-known shift variance of the critically sampled DWT. In [7], we presented an entirely new use for the redundancy in the RDWT; specifically, we employed transform redundancy 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 significantly enhance coding performance. The encoder of the resulting RWMH system is depicted in Fig. 4.

Intuitively, we observe that each of the critically sampled DWTs within an RDWT will "view" motion from a different perspective. Consequently, if motion is predicted in the RDWT domain, the multiple-phase inverse RDWT, $\mathcal{R}^{-1}(\cdot)$, forms a multihypothesis prediction in the form of (1). Specifically, for a J-scale RDWT, the reconstruction from DWT i of the RDWT is $\tilde{s}_i[\mathbf{p},t]$, $0 \leq i < 4^J$, while $w_i[\mathbf{p},t] = 4^{-J}$, $\forall i$. In the next section, we establish more rigorously the performance advantages of RWMH.

IV. Analysis of RWMH

In this section, we show analytically that the multihypothesis nature of our RWMH prediction of Fig. 4 offers substantial performance gain over the single-hypothesis prediction of Fig. 3. In Fig. 4, the current and reference frames are transformed into RDWT coefficients, and MC takes place in the RDWT domain, yielding $\left\{\widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k$, the RDWT-domain prediction of the current frame. The spatial-domain prediction residual is then

$$\begin{aligned} r^{(\mathrm{MH})}[\mathbf{p},t] &= s[\mathbf{p},t] - \mathcal{R}^{-1}\left[\left\{\widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k\right] \\ &= \mathcal{R}^{-1}\left[\left\{S^{(k)}[\mathbf{p},t] - \widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k\right] \\ &= \mathcal{R}^{-1}\left[\left\{R^{(k)}[\mathbf{p},t]\right\}_k\right], \end{aligned} \tag{8}$$

where $\left\{R^{(k)}[\mathbf{p},t]\right\}_k = \left\{S^{(k)}[\mathbf{p},t] - \widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k$ is the RDWT-domain residual, and the superscript (MH) denotes that this is the prediction error for the multihypothesis coder of Fig. 4; we will determine an equivalent quantity for the single-hypothesis coder of Fig. 3 shortly. We note that a preliminary version of the subsequent analysis with a simpler motion model was first presented in [21].

A. Motion Model

In a general sense, motion of objects from one frame to the next as viewed in the RDWT domain will consist of some motions to which the multiple-phase inverse RDWT is invariant. That is, $\mathcal{R}^{-1}$ in (8) will be shift invariant in the sense of (7) to some of the motion between the two RDWT-domain frames. On the other hand, other motion between the two frames will not possess this invariance. With this observation in mind, we propose a simple RDWT-domain model of motion that is partitioned into shift-invariant and shift-variant components. Specifically, for spatial location $\mathbf{p}$ in subband k, we model the motion from time $t-1$ to time t with the vector

$$\mathbf{d}^{(k)}[\mathbf{p}] = \boldsymbol{\delta} + \boldsymbol{\xi}^{(k)}[\mathbf{p}]. \tag{9}$$

In (9), $\boldsymbol{\delta}$ represents a simple translation of all spatial locations in all subbands by the same vector; below, we will see that the multiple-phase inverse RDWT will be invariant to such global "pan" motion as prescribed by (7). On the other hand, $\boldsymbol{\xi}^{(k)}[\mathbf{p}]$ is a dense-motion field that varies with each spatial location in each subband, capturing more complex motion (zooms, rotations, luminance variations, etc.); the inverse RDWT will typically not be invariant to the translations of this dense-motion field.

The ME process of the RWMH coder of Fig. 4 can be considered to estimate the motion field of (9) as

$$\widehat{\mathbf{d}}[\mathbf{p}] = \widehat{\boldsymbol{\delta}} + \widehat{\boldsymbol{\xi}}[\mathbf{p}], \tag{10}$$

since the same motion estimate is used in each subband in order to reduce the motion-vector overhead (see Sec. V below). In subband k, the error in this motion estimate is then

$$\mathbf{d}^{(k)}[\mathbf{p}] - \widehat{\mathbf{d}}[\mathbf{p}] = \boldsymbol{\Delta} + \boldsymbol{\Xi}^{(k)}[\mathbf{p}], \tag{11}$$

where $\boldsymbol{\Delta}$ is the global pan-motion error,

$$\boldsymbol{\Delta} = \boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}, \tag{12}$$

and $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$ is the dense-motion error,

$$\boldsymbol{\Xi}^{(k)}[\mathbf{p}] = \boldsymbol{\xi}^{(k)}[\mathbf{p}] - \widehat{\boldsymbol{\xi}}[\mathbf{p}]. \tag{13}$$

We model $\boldsymbol{\delta}$, $\widehat{\boldsymbol{\delta}}$, $\boldsymbol{\xi}^{(k)}[\mathbf{p}]$, and $\widehat{\boldsymbol{\xi}}[\mathbf{p}]$ as random vector fields with the mean of $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$ being the zero vector. We argue that it is reasonable to assume that, although both $\boldsymbol{\xi}^{(k)}[\mathbf{p}]$ and its estimate $\widehat{\boldsymbol{\xi}}[\mathbf{p}]$ are likely to be highly correlated to the current frame, the difference between the two, $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$, is independent of $S^{(k)}[\mathbf{p},t]$. Finally, we assume that the dense-motion error of one subband is independent of that of another subband ($\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$ is independent of $\boldsymbol{\Xi}^{(l)}[\mathbf{p}]$ for $k \neq l$), the dense-motion error at one spatial position is independent of that at another spatial position ($\boldsymbol{\Xi}^{(k)}[\mathbf{p}_1]$ is independent of $\boldsymbol{\Xi}^{(k)}[\mathbf{p}_2]$ for $\mathbf{p}_1 \neq \mathbf{p}_2$), and the horizontal component of $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$ is independent of its vertical component.

With the true motion field given by (9), subband k of the current frame is obtained from the reference frame via

$$S^{(k)}[\mathbf{p},t] = I\Big(S^{(k)}\big[\mathbf{p} - \mathbf{d}^{(k)}[\mathbf{p}], t-1\big]\Big). \tag{14}$$

In [22], a frame displaced by a dense motion field was approximated using a first-order Taylor-series expansion. Applying this approach to the right side of (14), we have

$$\begin{aligned}
&I\Big(S^{(k)}\big[\mathbf{p}-\mathbf{d}^{(k)}[\mathbf{p}],t-1\big]\Big) = \\
&\quad I\Big(S^{(k)}[\mathbf{p},t-1]\Big) - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \mathbf{d}^{(k)}[\mathbf{p}] \\
&\quad = I\Big(S^{(k)}[\mathbf{p},t-1]\Big) - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\delta} \\
&\qquad - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\xi}^{(k)}[\mathbf{p}] \\
&\quad = I\Big(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t-1]\Big) \\
&\qquad - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\xi}^{(k)}[\mathbf{p}],
\end{aligned} \tag{15}$$

where $\nabla = \begin{bmatrix}\frac{\partial}{\partial_x} & \frac{\partial}{\partial_y}\end{bmatrix}^T$ is the spatial gradient operator, and $\bullet$ indicates a vector inner product. Similarly, subband k in the RDWT-domain prediction of the current frame using the estimated motion field of (10) is

$$\begin{aligned}
\widetilde{S}^{(k)}[\mathbf{p},t] &= I\Big(S^{(k)}\big[\mathbf{p}-\widehat{\mathbf{d}}[\mathbf{p}],t-1\big]\Big) \\
&= I\Big(S^{(k)}[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \widehat{\boldsymbol{\xi}}[\mathbf{p}].
\end{aligned} \tag{16}$$

Subband k of the RDWT-domain residual of (8) is then

$$\begin{aligned}
R^{(k)}[\mathbf{p},t] &= S^{(k)}[\mathbf{p},t] - \widetilde{S}^{(k)}[\mathbf{p},t] \\
&= I\Big(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t-1] - S^{(k)}[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) \\
&\qquad - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\xi}^{(k)}[\mathbf{p}] \\
&\qquad + \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \widehat{\boldsymbol{\xi}}[\mathbf{p}] \\
&= I\Big(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t-1] - S^{(k)}[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) \\
&\qquad - \nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\Xi}^{(k)}[\mathbf{p}].
\end{aligned} \tag{17}$$

We define

$$N^{(k)}[\mathbf{p},t] = -\nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)\bullet \boldsymbol{\Xi}^{(k)}[\mathbf{p}]. \tag{18}$$

In the spatial domain, the prediction residual of (17) is then

$$\begin{aligned}
&r^{(\mathrm{MH})}[\mathbf{p},t] = \\
&\quad \mathcal{R}^{-1}\Bigg[\Big\{I\Big(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t-1] - S^{(k)}[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) + \\
&\qquad\qquad N^{(k)}[\mathbf{p},t]\Big\}_k\Bigg],
\end{aligned} \tag{19}$$

which, due to (6) and (7), becomes

$$r^{(\mathrm{MH})}[\mathbf{p},t] = I\Big(s[\mathbf{p}-\boldsymbol{\delta},t-1] - s[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) + n^{(\mathrm{MH})}[\mathbf{p},t], \tag{20}$$

where

$$n^{(\mathrm{MH})}[\mathbf{p},t] = \mathcal{R}^{-1}\Big[\Big\{N^{(k)}[\mathbf{p},t]\Big\}_k\Big]. \tag{21}$$

The essence of this analysis is that the prediction residual of the RWMH system consists of two components—a residual due to the estimation of global pan motion as well as a residual due to the estimation of the dense-motion field. The former arises due to the fact that the multiple-phase inverse RDWT is invariant to both $\boldsymbol{\delta}$ and $\widehat{\boldsymbol{\delta}}$ such that such global pan translations in the RDWT domain produce identical translations in the spatial domain as seen in the first terms on the right of (19) and (20). However, the other residual, $N^{(k)}[\mathbf{p},t]$, results from the error in estimating the dense motion of the model, and the effect of this complex motion cannot be as easily characterized once it is mapped into the spatial domain by $\mathcal{R}^{-1}$. In the next section, we examine the nature of $N^{(k)}[\mathbf{p},t]$ and invoke a white-noise model for this dense-motion residual. This white-noise model, in turn, permits us to draw some quantitative observations on the benefit of RWMH later in Sec. IV-C.

B. Noise Model for the Dense-Motion Residual

For $\boldsymbol{\Xi}^{(k)}[\mathbf{p}] = \begin{bmatrix}\xi_x^{(k)}[\mathbf{p}] & \xi_y^{(k)}[\mathbf{p}]\end{bmatrix}^T$, subband k of the dense-motion residual of (18) is

$$\begin{aligned}
&N^{(k)}[\mathbf{p},t] = \\
&-\xi_x^{(k)}[\mathbf{p}]\frac{\partial}{\partial_x}I\Big(S^{(k)}[\mathbf{p},t-1]\Big) - \xi_y^{(k)}[\mathbf{p}]\frac{\partial}{\partial_y}I\Big(S^{(k)}[\mathbf{p},t-1]\Big).
\end{aligned} \tag{22}$$

Due to independence between $\boldsymbol{\Xi}^{(k)}[\mathbf{p},t]$ and $\nabla I\Big(S^{(k)}[\mathbf{p},t-1]\Big)$, it is straightforward to establish that $N^{(k)}[\mathbf{p},t]$ has zero mean. The cross-subband correlation between subbands k and l at spatial locations $\mathbf{p}_1$ and $\mathbf{p}_2$ is

$$\begin{aligned}
&E\Big[N^{(k)}[\mathbf{p}_1,t]N^{(l)}[\mathbf{p}_2,t]\Big] = \\
&E\Big[\xi_x^{(k)}[\mathbf{p}_1]\xi_x^{(l)}[\mathbf{p}_2]\Big]E\left[\frac{\partial}{\partial_x}I\Big(S^{(k)}[\mathbf{p}_1,t-1]\Big)\frac{\partial}{\partial_x}I\Big(S^{(l)}[\mathbf{p}_2,t-1]\Big)\right] + \\
&E\Big[\xi_x^{(k)}[\mathbf{p}_1]\xi_y^{(l)}[\mathbf{p}_2]\Big]E\left[\frac{\partial}{\partial_x}I\Big(S^{(k)}[\mathbf{p}_1,t-1]\Big)\frac{\partial}{\partial_y}I\Big(S^{(l)}[\mathbf{p}_2,t-1]\Big)\right] + \\
&E\Big[\xi_y^{(k)}[\mathbf{p}_1]\xi_x^{(l)}[\mathbf{p}_2]\Big]E\left[\frac{\partial}{\partial_y}I\Big(S^{(k)}[\mathbf{p}_1,t-1]\Big)\frac{\partial}{\partial_x}I\Big(S^{(l)}[\mathbf{p}_2,t-1]\Big)\right] + \\
&E\Big[\xi_y^{(k)}[\mathbf{p}_1]\xi_y^{(l)}[\mathbf{p}_2]\Big]E\left[\frac{\partial}{\partial_y}I\Big(S^{(k)}[\mathbf{p}_1,t-1]\Big)\frac{\partial}{\partial_y}I\Big(S^{(l)}[\mathbf{p}_2,t-1]\Big)\right].
\end{aligned} \tag{23}$$

As mentioned above, we assume that the horizontal component of the dense-motion error is independent of the vertical component, the dense-motion error at location $\mathbf{p}_1$ is independent of that at $\mathbf{p}_2$, and the dense-motion errors in different subbands are independent. Thus, the middle two terms on the right of (23) are zero, and we have that the variance of the dense-motion residual is

$$\begin{aligned}
\nu_N^{(k)} &= E\Big[N^{(k)}[\mathbf{p}_1,t]N^{(l)}[\mathbf{p}_2,t]\Big] \\
&= \begin{cases}\nu_{\xi_x}^{(k)}\nu_{\nabla S_x}^{(k)} + \nu_{\xi_y}^{(k)}\nu_{\nabla S_y}^{(k)}, & k=l \text{ and } \mathbf{p}_1=\mathbf{p}_2, \\ 0, & \text{else,}\end{cases}
\end{aligned} \tag{24}$$

where $\nu_{\xi_x}^{(k)}$, $\nu_{\xi_y}^{(k)}$, $\nu_{\nabla S_x}^{(k)}$, and $\nu_{\nabla S_y}^{(k)}$ are the variances of $\xi_x^{(k)}$, $\xi_y^{(k)}$, $\frac{\partial}{\partial_x}I\Big(S^{(k)}[\mathbf{p},t-1]\Big)$, and $\frac{\partial}{\partial_y}I\Big(S^{(k)}[\mathbf{p},t-1]\Big)$, respectively, assuming these latter quantities are spatially stationary.

From (24), we have that the dense-motion residual can be considered to be noise that is uncorrelated spatially as well as across subbands. Although (24) indicates that the variance of this dense-motion noise may vary from subband to subband (i.e., it is a function of k), we have observed empirically that the variances $\nu_{\xi_x}^{(k)}$ and $\nu_{\xi_y}^{(k)}$ of the dense-motion error tend to increase as the scale of the subband decreases, while, at the same time, the variances $\nu_{\nabla S_x}^{(k)}$ and $\nu_{\nabla S_y}^{(k)}$ of the spatial gradient tend to decrease. The decrease in the variance of the spatial gradient tends to counteract the increase in the dense-motion error variance, such that the resulting variance of the dense-motion residual tends to be somewhat level across the subbands.

For example, referring to (13), we apply simple block-based ME independently in each subband of the RDWT of two frames of a video sequence in order to estimate a motion field tailored to each individual subband; we take this estimate to be $\boldsymbol{\xi}^{(k)}[\mathbf{p}]$. We then estimate $\widehat{\boldsymbol{\xi}}[\mathbf{p}]$ by applying the block-based, cross-subband ME procedure described later in Sec. V to generate a single, cross-subband motion field. The difference between these two fields approximates the dense-motion error $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$, the variances of the horizontal and vertical components of which we plot in Figs. 5–7 for frames from several sequences. Additionally, we plot in these figures an estimate of the spatial-gradient variances $\nu_{\nabla S_x}^{(k)}$ and $\nu_{\nabla S_y}^{(k)}$ as calculated using the 5-tap multidimensional-derivative algorithm of [23] to provide a spatial gradient for each RDWT subband. Finally, we plot the resulting dense-motion residual variances $\nu_N^{(k)}$ as calculated via (24) from the estimated $\nu_{\xi_x}^{(k)}$, $\nu_{\xi_y}^{(k)}$, $\nu_{\nabla S_x}^{(k)}$, and $\nu_{\nabla S_y}^{(k)}$. Although we see that $\nu_N^{(k)}$ does tend to "peak" for one, sometimes two, subbands, it is more or less relatively constant for the other subbands. From our observations, the subbands at which these peaks in $\nu_N^{(k)}$ occur vary from sequence to sequence.

As a consequence of these observations, we adopt a simple model for the dense-motion residual that disregards exceptional peaks in $\nu_N^{(k)}$. That is, we model the $N^{(k)}[\mathbf{p},t]$ as zero-mean white noise of a single variance ν_N. Although we have argued that the dense-motion residual is reasonably uncorrelated both within and across subbands, this white-noise model is somewhat of an oversimplification since real motion residuals may have variance that varies somewhat between subbands, such as depicted in Figs. 5–7. However, the simplicity of this white-noise model permits us to analytically quantify a gain due to RWMH in the next section. Later, in Sec. VI, we will empirically evaluate this gain for real motion residuals.

C. Performance Gain

For a prediction residual such as given by (20), Girod [2,24] derives the 2D power spectral density to be

$$\Phi_{rr}^{(\mathrm{MH})}(\boldsymbol{\omega}) = \Phi_{nn}^{(\mathrm{MH})}(\boldsymbol{\omega})\left|I(\boldsymbol{\omega})\right|^2 + \Phi_{ss}(\boldsymbol{\omega})\left(1+\left|I(\boldsymbol{\omega})\right|^2 - 2\Re\left[I(\boldsymbol{\omega})P(\boldsymbol{\omega})\right]\right), \tag{25}$$

where $\Phi_{ss}(\boldsymbol{\omega})$ and $\Phi_{nn}^{(\mathrm{MH})}(\boldsymbol{\omega})$ are the 2D power spectral densities of $s[\mathbf{p},t]$ and $n^{(\mathrm{MH})}[\mathbf{p},t]$, respectively; $P(\boldsymbol{\omega})$ is the 2D Fourier transform of the probability density function of the global pan-motion error $\boldsymbol{\Delta}$ of (12); $I(\boldsymbol{\omega})$ is the frequency response of the interpolation filter used to resolve fractional-pixel values; $\boldsymbol{\omega} = \begin{bmatrix}\omega_x & \omega_y\end{bmatrix}^T$; and $\Re(\cdot)$ denotes the real part of a complex number.

As in [2], we focus on the prediction-residual variance to gauge coding performance. From (25), the prediction-residual variance is

$$\nu_r^{(\mathrm{MH})} = \nu_n^{(\mathrm{MH})} + \Gamma_{ss}, \tag{26}$$

where $\nu_n^{(\mathrm{MH})}$ is the variance of $n^{(\mathrm{MH})}[\mathbf{p},t]$,

$$\Gamma_{ss} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Phi_{ss}(\boldsymbol{\omega})\Big(2-2\Re\big[P(\boldsymbol{\omega})\big]\Big)\,d\boldsymbol{\omega}, \tag{27}$$

and we have assumed $I(\boldsymbol{\omega}) = 1$ (i.e., sinc interpolation [2]) in order to simplify the analysis.

The key to our RWMH technique is embodied by (21)—as we have discussed previously, the multiple-phase inverse RDWT is a pseudo-inverse frame operator which is tantamount to a projection onto the range space of the RDWT followed by a mapping back into the original spatial domain. Since the multiple-phase inverse RDWT is not invariant to the dense-motion component of the motion model, the dense-motion residual $N^{(k)}[\mathbf{p},t]$ of (18) is almost certainly not in the range space of the RDWT. Thus, the mapping of the dense-motion residual back to the spatial domain via (21) will result in a reduction in variance. Above, we modeled the dense-motion residual $N^{(k)}[\mathbf{p},t]$ as zero-mean white noise of variance ν_N; consequently, the following theorem quantifies the variance reduction resulting from the multiple-phase inverse RDWT.

Theorem 1: If noise present in the RDWT domain is zero-mean, white, and of variance ν_N, and the wavelet filters underlying the J-scale 2D RDWT are orthonormal, then the variance of the spatial-domain noise given by (21) is

$$\nu_n^{(\mathrm{MH})} = \frac{\nu_N}{5}\left[1+4\left(\frac{1}{16}\right)^J\right]. \tag{28}$$

The proof of Theorem 1 is given in [25]. We note that Theorem 1 will approximately hold if the wavelet filters underlying the RDWT are "near-orthonormal" biorthogonal filters, as is common in practice.

Now, suppose that, in Fig. 4, rather than mapping the RDWT-domain prediction to the spatial domain using the multiple-phase inverse RDWT, we instead use the single-phase inverse of (4). It is straightforward to see that this single-phase MC process is equivalent to the single-hypothesis system shown in Fig. 3. In this case, (8) becomes

$$\begin{aligned} r^{(\mathrm{SH})}[\mathbf{p},t] &= s[\mathbf{p},t] - \mathcal{D}^{-1}\left[\downarrow\left\{\widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k\right] \\ &= \mathcal{D}^{-1}\left[\downarrow\left\{S^{(k)}[\mathbf{p},t] - \widetilde{S}^{(k)}[\mathbf{p},t]\right\}_k\right], \end{aligned} \tag{29}$$

where last equality is due to (4) and (6). Furthermore, (19) becomes

$$r^{(\mathrm{SH})}[\mathbf{p},t] = \mathcal{D}^{-1}\left[\downarrow \left\{ I\Big(S^{(k)}[\mathbf{p}-\boldsymbol{\delta},t-1] - S^{(k)}[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) + N^{(k)}[\mathbf{p},t]\right\}_k\right], \quad (30)$$

which, through use of (3), can be expressed as

$$r^{(\mathrm{SH})}[\mathbf{p},t] = \mathcal{D}^{-1}\left[\downarrow \mathcal{R}\Big[I\Big(s[\mathbf{p}-\boldsymbol{\delta},t-1] - s[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) + N^{(k)}[\mathbf{p},t]\Big]\right]. \quad (31)$$

Applying (4), (6), and the linearity of $I(\cdot)$ then gives

$$r^{(\mathrm{SH})}[\mathbf{p},t] = I\Big(s[\mathbf{p}-\boldsymbol{\delta},t-1] - s[\mathbf{p}-\widehat{\boldsymbol{\delta}},t-1]\Big) + n^{(\mathrm{SH})}[\mathbf{p},t], \quad (32)$$

with

$$n^{(\mathrm{SH})}[\mathbf{p},t] = \mathcal{D}^{-1}\left[\downarrow \left\{N^{(k)}[\mathbf{p},t]\right\}_k\right]. \quad (33)$$

Since the DWT is a unitary transform, and the dense-motion residual is RDWT-domain noise equally present in all phases, the spatial-domain variance in this case is

$$\nu_n^{(\mathrm{SH})} = \nu_N, \quad (34)$$

assuming, as before, that the dense-motion residual in the RDWT domain is zero-mean, white, and of variance ν_N. The prediction-residual variance is then

$$\nu_r^{(\mathrm{SH})} = \nu_n^{(\mathrm{SH})} + \Gamma_{ss}, \quad (35)$$

where Γ_{ss} is once again given by (27).

To quantity the gain of the multihypothesis approach over the single-hypothesis approach, let us assume, as was done in [2,24], an isotropic signal power spectrum,

$$\Phi_{ss}(\boldsymbol{\omega}) = \frac{2\pi}{\omega_0^2}\left(1 + \frac{\omega_x^2+\omega_y^2}{\omega_0^2}\right)^{-\frac{2}{3}}, \quad (36)$$

where $\omega_0 = -\ln(0.93)$, and an isotropic Gaussian density of variance ν_Δ for the global pan-motion error $\boldsymbol{\Delta}$ such that

$$P(\boldsymbol{\omega}) = \exp\left[-\frac{\nu_\Delta}{2}\left(\omega_x^2+\omega_y^2\right)\right]. \quad (37)$$

Under these models, we numerically evaluate $\nu_r^{(\mathrm{MH})}$ from (26) and $\nu_r^{(\mathrm{SH})}$ from (35) versus displacement accuracy β for several noise variances in Fig. 8, 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. [26]. It is evident in Fig. 8 that the multihypothesis approach results in a significant decrease in prediction-residual variance, particularly when ME is accurate (β small). We define the difference in prediction-residual variance in dB between the two approaches as

$$\gamma = 10\log_{10}\left(\frac{\nu_r^{(\mathrm{MH})}}{\nu_r^{(\mathrm{SH})}}\right) \quad (38)$$

and numerically evaluate γ as β varies in Fig. 9.

Examining the numerical results of Figs. 8 and 9 leads to the following important observations:

- RWMH systematically reduces the prediction-residual variance over that of single-phase ME/MC, although the performance gain may be small if ME is inaccurate, or the power of the dense-motion residual is low.
- As ME becomes highly accurate, RWMH produces a 7-dB reduction in prediction-residual variance as compared to single-phase ME/MC.
- For a given ME accuracy, the larger the power of the dense-motion residual, the more effective RWMH is at reducing the prediction-residual variance.

Finally, we note that the analysis here predicts that greater RWMH performance gains come from larger dense-motion residuals. On the other hand, if the dense-motion residual is zero, there will be no RWMH performance gain—this latter situation corresponds to the case in which motion from frame to frame is a simple global pan translation such that $\boldsymbol{\xi}^{(k)}[\mathbf{p}]$ in (9), and consequently $\boldsymbol{\Xi}^{(k)}[\mathbf{p}]$ in (18), is zero. Thus, we expect to see little or no gain in practice due to RWMH for sequences that consist largely of simple translational motion (e.g., camera pans), while sequences with complex motion are apt to see more substantial benefit.

V. Implementation of Block-Based RWMH

We now consider some issues related to implementation of RWMH as depicted in Fig. 4. We note that the motion model embodied by (9) consists of a single global pan-motion translation which is "refined" at each spatial location in each subband independently with a dense-motion field. A dense-motion field in each subband unfortunately entails an impractically large motion-vector coding overhead; thus, we invoke block-based MC using a single cross-subband motion-vector field. A block-based motion model will be equivalent to (9) for only the smallest block size; i.e., single-pixel blocks. On the other hand, in the extreme case of a very large block size encompassing the entire frame, the block-based model degenerates to merely global pan motion with the dense-motion component of (9) being zero—as discussed above, there is no RWMH gain for this extreme case. For MC with practical block sizes, we expect performance somewhere between the two extremes, with gains due to RWMH approaching those predicted by the analysis of Sec. IV-C as the block size decreases towards single-pixel blocks.

In implementing RWMH with ME/MC applied on a block-by-block basis, we note that, 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 ME procedure, block matching is used to determine the motion of each set as a whole. Specifically, a block-matching procedure uses a cross-subband distortion measure that sums absolute errors for each block of the set similar to the cross-subband ME procedure of [8]. However in our measure, the coefficients from all phases in both the current and reference frames contribute to the distortion measurement, in contrast to the measure of [8], in which only coefficients from a single critically subsampled DWT in the current frame contribute.

Specifically, the motion vector for the set located at $\mathbf{p}$ is

$$\widehat{\mathbf{d}}[\mathbf{p}] = \arg\min_{\mathbf{d}\in\mathcal{W}} \mathrm{MAE}(\mathbf{p},\mathbf{d}), \tag{39}$$

where the mean absolute error (MAE) is

$$\mathrm{MAE}(\mathbf{p},\mathbf{d}) = \frac{1}{B^2}\sum_{k=0}^{B-1}\sum_{l=0}^{B-1}\mathrm{AE}(\mathbf{p}+\begin{bmatrix}k & l\end{bmatrix}^T,\mathbf{d}), \tag{40}$$

and the absolute error (AE) is

$$\begin{aligned}\mathrm{AE}(\mathbf{p},\mathbf{d}) = 2^{-(2J+1)}\Big|B_J[\mathbf{p},t]-B_J[\mathbf{p}+\mathbf{d},t-1]\Big| + \\ \sum_{j=1}^{J}2^{-(2j+1)}\Bigg(\Big|V_j[\mathbf{p},t]-V_j[\mathbf{p}+\mathbf{d},t-1]\Big| + \\ \Big|H_j[\mathbf{p},t]-H_j[\mathbf{p}+\mathbf{d},t-1]\Big| + \\ \Big|D_j[\mathbf{p},t]-D_j[\mathbf{p}+\mathbf{d},t-1]\Big|\Bigg).\end{aligned} \tag{41}$$

In the above equations, B_j, H_j, V_j, and D_j are the baseband, horizontal, vertical, and diagonal subbands, respectively, at scale j. A window $\mathcal{W}$ of the range $[-W, W]$ both horizontally and vertically is used for the block search.

VI. Experimental Results

We now investigate RWMH through a series of empirical evaluations. Throughout this section, we consider grayscale sequences, and all wavelet transforms (DWT and RDWT) use the popular 9-7 biorthogonal filter with symmetric extension. ME is full search with a window of $W = 15$, and subpixel accuracy is implemented via a multiple-step interpolation procedure consisting of filtering and bilinear interpolation as in MPEG-4 [27]. We first focus on an experimental investigation of the theoretical analysis embodied by Fig. 9 before turning our attention to real coding performance.

A. Experimental Investigation of Theoretical Analysis

We have observed empirically that the prediction residual in our RWMH technique does in fact undergo a substantial reduction in variance as suggested by Fig. 9. For example, in Table I, we consider a single frame from a video sequence and use the previous frame to create a MC prediction of the first frame in the RDWT domain. We then invert the resulting RDWT-domain residual using both the single-phase inverse of (4) and the multiple-phase inverse of (5), calculating γ via (38) using the resulting spatial-domain variances $\nu_r^{(\mathrm{MH})}$ and $\nu_r^{(\mathrm{SH})}$. In order to divorce the effects of ME from those of single- and multiple-phase MC, we fix the motion-vector field for both inverse transforms to be the same field, estimated separately in the spatial domain between the two original frames. The results are shown in Table I for several block sizes B, several scales J of transform decomposition, and several subpixel accuracies β.

We see empirically in Table I that RWMH does, in fact, reduce the prediction-residual variance over that of single-phase MC in much the same way as predicted by the analysis of Sec. IV-C. Specifically, as the number of scales of decomposition is varied with a fixed block size, we see an exponential relationship between γ and J as indicated by (28)—as J increases, γ decreases by a decreasing amount. Additionally, we observe that the reduction in prediction-residual variance increases with increasing ME accuracy, again as predicted by Fig. 9.

Most saliently, however, we see that as J is held fixed, the gain γ varies significantly with the block size used in MC. This effect is anticipated, since the smaller the block size, the closer the block motion model embodies the combined pan-motion/dense-motion model of (9). In fact, we see that, for extremely small MC block sizes, the reduction in prediction-residual variance is approximately 6 dB, which is rather close to the maximum gain of 7 dB as anticipated by Fig. 9. The larger block sizes, being closer to simple global pan motion alone, result in reduced RWMH performance gain. Thus, we conclude that RWMH does indeed lead to variance reduction for a real MC residual in much the same way as indicated by the analysis of Sec. IV-C, which can be considered to represent an ideal performance bound approached by block-based ME/MC as the block size decreases.

B. Coding Performance

It is well-known that a reduction in prediction-error variance, γ, does not necessarily portend an increase in rate-distortion performance in a real video-coding system. Consequently, we now investigate whether the RWMH performance gain predicted by Fig. 9 and exhibited empirically in Table I results in improved rate-distortion performance during actual coding of MC residuals. In the following simulation results, we code grayscale sequences with the first frame intra-coded (I-frame) while all subsequent frames use ME/MC (P-frames). Since SPIHT [28], used as the core compression engine in all experiments, produces an embedded coding, each frame of the sequence is coded at exactly the specified target rate. We consider three systems: the RWMH system of Fig. 4; the "RDWT Block" system of Fig. 3; and a third system, "Spatial Block," which refers to block-based MC in the spatial domain, the traditional method employed in video-coding standards, followed by an entire-image DWT and then SPIHT coding of the DWT coefficients. All three system implementations can be found within the QccPack [29] software package.[1]

Table I indicates that we should choose the MC block size to be as small as possible in order to maximize the gain of the RWMH technique. However, Table I neglects the ramifications of the rate overhead needed to transmit the motion-vector field in any real system. Of the block sizes considered in Table I, only $B = 8$ and $B = 16$ are practical choices as the motion-vector overhead is likely to be too great for the other values. Consequently, from this point on, we focus on a block size of $B = 8$; we have verified empirically that $B = 8$ is the better performing block size of the two in the RWMH system used

[1] see `http://qccpack.sourceforge.net`

in this section. Additionally, Table I indicates that we should choose J as big as possible. Unfortunately, the RDWT Block system of Fig. 3 is constrained such that $B \geq 2^J$ in order that each cross-scale block used in ME/MC includes coefficients from all subbands. Consequently, with $B = 8$, we use $J = 3$ henceforth.

Initially, in order to separate the effects of ME from those of MC, we use the same motion-vector fields for all three systems; these fields are calculated independently from the systems using full-search ME on the original frames in the spatial domain. The resulting gain in average PSNR for RWMH over the other systems operating at a fixed rate is tabulated in Table II. Table II indicates that, even when all systems use the same motion vectors, the RWMH system outperforms both the Spatial Block and RDWT Block systems on the order of 0.5–1 dB. Additionally, we see that the performance gain increases with increasing ME accuracy as was predicted by both Fig. 9 as well as Table I.

More realistic results are obtained when all systems estimate their own ME fields and incur the overhead of motion vectors in their bitstreams. Under these conditions, we have observed the highest PSNR for quarter-pixel accuracy ($\beta = -2$), the relatively large overhead of eighth-pixel accuracy resulting in diminished rate-distortion performance. Consequently, we present average PSNR figures at a fixed rate for $\beta = -2$ in Table III, and rate-distortion performance for several sequences over a range of rates in Figs. 10–12. Table III and Figs. 10–12 illustrate that multihypothesis prediction in the form of RWMH achieves a significant gain—on the order of 1.0 to 1.5 dB—over single-phase prediction for sequences with complex motion, which make up a majority of the sequences listed in Table III. The variance of the dense-motion residual is relatively large for these sequences. As foretold by Fig. 9, RWMH offers substantial gain over single-phase prediction for these sequences, even at relatively low ME accuracy, since ν_N is large. On the other hand, more modest gains over the single-phase system are exhibited for the sequence "Mother & Daughter" which consists of simpler motion for which ν_N is relatively small, an observation again in line with Fig. 9.

Although the focus of this paper is to investigate analytically and empirically the performance gain of RWMH over single-phase prediction, we also include in Table III performance figures for H.264/AVC [30], representative of the state of the art in hybrid, predictive-feedback-loop video coding. We configure H.264/AVC encoding so as to duplicate the conditions used for the other coders as closely as possibly. Specifically, we use H.264/AVC JM 9.5 operating in baseline profile using 8×8 ME/MC inter blocks without B-frames or rate-distortion optimization. As can be seen, the RWMH implementation used here yields rate-distortion performance quite close to that of H.264/AVC for most of the sequences. We note that H.264/AVC is capable of substantially superior rate-distortion performance when its advanced coding modes are employed, and we anticipate that RWMH could benefit from many of the same advanced coding techniques (LTMMC, multiple MC block sizes, etc.) as well. Such an involved implementation, however, lies largely outside the scope of the analytical discussion which is our focus here.

As a final comment concerning our empirical results, let us observe that, in comparing Fig. 4 to Fig. 3, one would expect RWMH to be somewhat more computationally complex than a corresponding single-phase implementation due to the presence of multiple redundant transforms which are substantially more expensive to compute than the usual critically sampled transform. Furthermore, the RWMH ME process is driven by a multiple-phase error measure (40) which involves substantially more coefficients than the single-phase measure employed in [8]. That said, we note that the RWMH implementation we use here actually runs much faster than the corresponding RDWT Block implementation. This is due to the fact that, within the RDWT Block coder, the RDWT is initially calculated in the spatially coherent representation of Fig. 1 before being subsampled into the tree representation of Fig. 2, and this subsampling procedure is extremely time-consuming. While it would be possible to produce the tree representation directly, the spatially coherent representation greatly facilitates the interpolation process necessary for subpixel accuracy. Furthermore, it is unclear that such direct production of the tree representation would result in any net gain in execution speed, since the overhead necessary for coefficient addressing in the ME distortion measure would be substantial for the tree representation, particularly at subpixel accuracy. Again, however, these practical concerns are largely outside the scope of our current considerations.

VII. Conclusions

In this paper, we have examined a new class of MHMC—phase-diversity multihypothesis—by analyzing a system which deploys MHMC in the domain of a redundant wavelet transform. Recognizing that RDWT coefficients with different phases view motion from different perspectives, this RWMH system treats each critically sampled DWT within the RDWT as a separate hypothesis prediction. A multiple-phase inverse RDWT operation implicitly combines the multiple predictions into a single spatial-domain prediction. The primary contribution of the work we report here is an analytical derivation that quantifies the performance gain of RWMH over single-phase prediction. Key to this analysis is that noise in the RDWT domain undergoes a substantial reduction in variance when the multiple-phase inverse RDWT is applied due to the fact that this pseudo-inverse contains a projection onto the range space of the forward transform. Our analysis adopts a motion model partitioned into a global pan-motion component and a dense-motion component. The variance of the prediction residual due to this latter component is greatly reduced in an RWMH system, leading to substantial reduction in the overall prediction-residual variance and higher coding efficiency. In fact, our analysis predicts that RWMH can reduce the prediction-residual variance by up to 7 dB as ME becomes highly accurate. Our experimental observations support the analysis in that our block-based implementation of RWMH, which can be considered to approach the analytical motion model as the block size decreases, significantly outperforms the single-phase system of [8] for sequences with complex motion which produce a substantial dense-motion residual.

We emphasize that the goal of this paper is not to produce a system with performance competitive with state-of-the-art hybrid video coding. Rather, our aim is to further the understanding of the RDWT and its use in ME/MC. We believe that the analysis we present here represents a significant theoretical advance in the field of wavelet-based video coding; as pertaining to practice, we anticipate that the RWMH paradigm can contribute to further improving existing hybrid video-coding structures when combined with a number of other advanced ME/MC strategies. However, our recent focus has been on incorporating RWMH into emerging 3D wavelet-based coders that eliminate the MC feedback loop in favor of motion-compensated temporal filtering (MCTF) in order to provide full fidelity, spatial, and temporal scalability. Indeed, our recent work [31,32] has produced 3D RWMH- and MCTF-based video coders with rate-distortion performance competitive with that of H.264/AVC but with the added advantage of full scalability.

REFERENCES

[1] 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.

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

[3] S. Nogaki and M. Ohta, "An overlapped block motion compensation for high quality motion picture coding," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 1, San Diego, CA, May 1992, pp. 184–187.

[4] M. T. Orchard and G. J. Sullivan, "Overlapped block motion compensation: An estimation-theoretic approach," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 693–699, September 1994.

[5] 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.

[6] M. Flierl, T. Wiegand, and B. Girod, "Rate-constrained multihypothesis prediction for motion compensated video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 11, pp. 957–969, November 2002.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] M. J. Shensa, "The discrete wavelet transform: Wedding the à trous and Mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, October 1992.

[12] H. S. Kim and H. W. Park, "Wavelet-based moving-picture coding using shift-invariant motion estimation in wavelet domain," *Signal Processing: Image Communication*, vol. 16, no. 7, pp. 669–679, April 2001.

[13] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelius, "Scalable wavelet video-coding with in-band prediction—Implementation and experimental results," in *Proceedings of the International Conference on Image Processing*, vol. 3, Rochester, NY, 2002, pp. 729–732.

[14] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, J. Barbarien, P. Schelkens, and J. Cornelius, "Wavelet-based fine granularity scalable video coding with in-band prediction," ISO/IEC JTC1/SC29/WG11, MPEG2002/M7906, Jeju Island, South Korea, March 2002.

[15] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelius, "Wavelet-based fully-scalable video coding with in-band prediction," in *Proceedings of the* 3^{rd} *IEEE Benelux Signal Processing Symposium*, Leuven, Belgium, March 2002, pp. 217–220.

[16] X. Li, L. Kerofsky, and S. Lei, "All-phase motion compensated prediction in the wavelet domain for high performance video coding," in *Proceedings of the International Conference on Image Processing*, vol. 3, Thessaloniki, Greece, October 2001, pp. 538–541.

[17] X. Li and L. Kerofsky, "High-performance resolution-scalable video coding via all-phase motion-compensated prediction of wavelet coefficients," in *Visual Communications and Image Processing*, C.-C. J. Kuo, Ed. Proc. SPIE 4671, January 2002, pp. 1080–1090.

[18] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1998.

[19] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.

[20] 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.

[21] 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, pp. 352–361.

[22] P. Moulin, R. Krishnamurthy, and J. W. Woods, "Multiscale modeling and estimation of motion fields for video coding," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1606–1620, December 1997.

[23] H. Farid and E. P. Simoncelli, "Differentiation of discrete multidimensional signals," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 496–508, April 2004.

[24] 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.

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

[26] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *IEEE Transactions on Image Processing*, vol. 9, no. 2, pp. 173–183, February 2000.

[27] M. Wollborn, I. Moccagatta, and U. Benzler, "Natural video coding," in *The MPEG-4 Book*, F. Pereira and T. Ebrahimi, Eds. Upper Saddle River, NJ: Prentice-Hall, 2002, ch. 8, pp. 293–382.

[28] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243–250, June 1996.

[29] J. E. Fowler, "QccPack: An open-source software library for quantization, compression, and coding," in *Applications of Digital Image Processing XXIII*, A. G. Tescher, Ed. San Diego, CA: Proc. SPIE 4115, August 2000, pp. 294–301.

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

[31] Y. Wang, S. Cui, and J. E. Fowler, "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.

[32] 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.

TABLE I

γ CALCULATED ON THE MC RESIDUAL BETWEEN FRAMES 52 AND 53 OF THE "NYC" SEQUENCE FOR MC BLOCKS OF SIZE $B \times B$, RDWT OF J DECOMPOSITION SCALES, AND ME ACCURACY OF β.

		γ (dB)		
J	B	$\beta = -1$	$\beta = -2$	$\beta = -3$
1	2	-4.29	-4.45	-4.65
2	2	-5.20	-5.27	-5.39
3	2	-5.68	-5.75	-5.84
4	2	-5.96	-6.01	-6.09
3	2	-5.68	-5.75	-5.84
3	4	-4.86	-5.29	-5.50
3	8	-2.02	-2.25	-2.35
3	16	-0.92	-0.99	-1.05

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, an Associate Editor for the INTERNATIONAL JOURNAL OF IMAGE & VIDEO PROCESSING, and a member of the program committee for the DATA COMPRESSION CONFERENCE.

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

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 41^{st} 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.

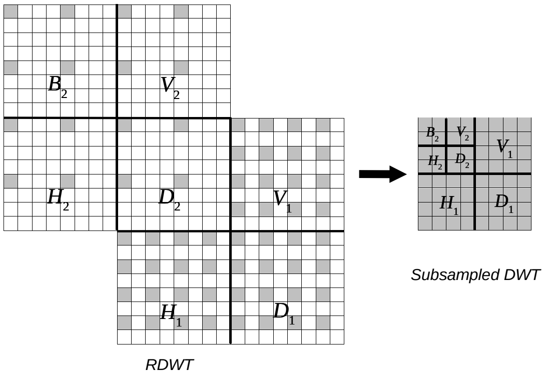

Fig. 1. Spatially coherent representation of a two-scale 2D RDWT. Coefficients retain their correct spatial location within each subband, and each subband is the same size as the original image. B_j, H_j, V_j, and D_j denote the baseband, horizontal, vertical, and diagonal subbands, respectively, at scale j. Shown is subsampling to recover one of the 4^J critically sampled DWTs.

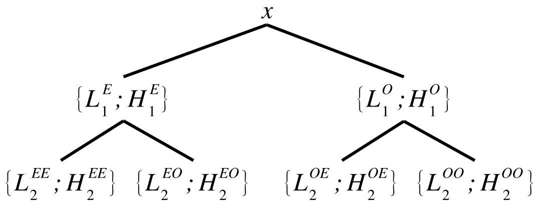

Fig. 2. Tree representation of a two-scale RDWT of 1D-signal x. Approximation and detail coefficients at scale j are L_j and H_j, respectively. E = even-phase subsampling; O = odd-phase subsampling.

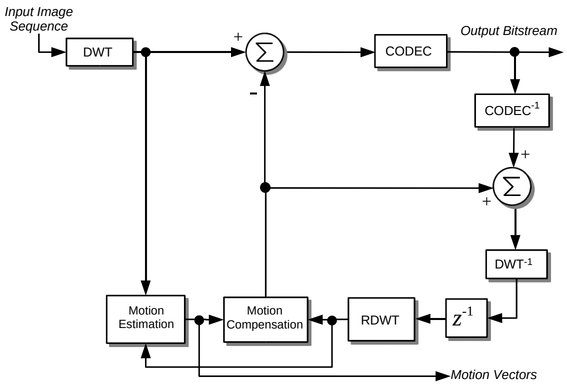

Fig. 3. The single-phase, RDWT-based video coder of [8]. z^{-1} = frame delay, *CODEC* is any still-image coder operating in the critically-sampled-DWT domain.

IEEE Transactions on Image Processing, vol. 15, pp. 3102-3113, October 2006.

TABLE II

Distortion relative to RWMH averaged over all frames of the sequence for rate of 0.5 bpp.

PSNR (dB) relative to RWMH system

	Spatial Block			RDWT Block		
	$\beta=-1$	$\beta=-2$	$\beta=-3$	$\beta=-1$	$\beta=-2$	$\beta=-3$
Football†	-0.6	-0.9	-1.1	-0.8	-1.1	-1.4
NYC†	-0.4	-0.5	-0.7	-0.8	-0.9	-1.1
Coastguard	-0.2	-0.3	-0.4	-0.5	-0.6	-0.8
Table Tennis‡	-0.3	-0.4	-0.6	-0.6	-0.9	-1.2
Foreman$^\sharp$	-0.5	-0.8	-1.1	-0.6	-0.9	-1.2
Susie†*	-0.4	-0.7	-1.1	-0.7	-0.9	-1.2
Mother-Daughter	0.0	-0.4	-0.7	-0.6	-0.9	-1.0

Sequences are 100 frames long, CIF (352×288) at 30 Hz except: †SIF (352×240), ‡25 Hz, *75 frames, $^\sharp$90 frames. Transforms use $J = 3$ scales of decomposition. ME with block size of $B = 8$ performed separately on original frames in spatial domain with all systems using the same motion-vector fields; motion-vector overhead not included in rate.

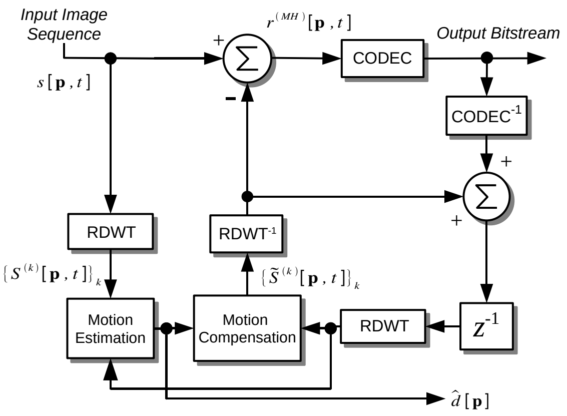

Fig. 4. The RWMH coder. z^{-1} = frame delay, *CODEC* is any still-image coder operating in the spatial domain. RDWT^{-1} denotes the multiple-phase inverse RDWT of (5).

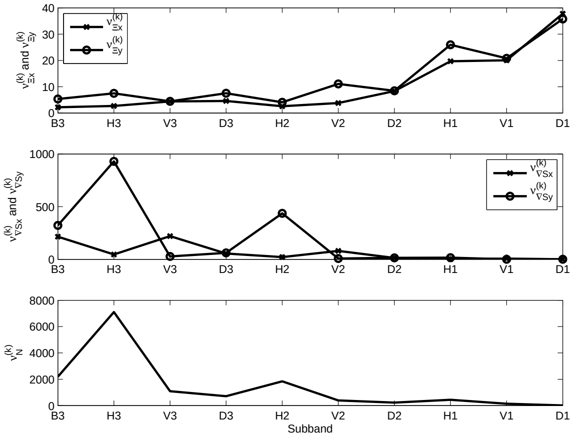

Fig. 6. Variances of the dense-motion error, the spatial gradient, and the dense-motion residual as estimated between frames 30 and 31 of the "Table Tennis" sequence.

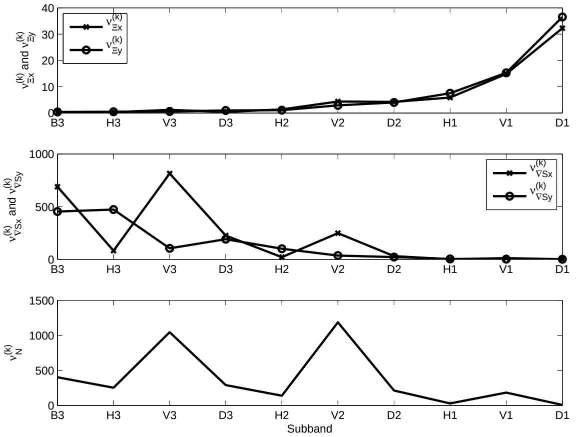

Fig. 5. Variances of the dense-motion error, the spatial gradient, and the dense-motion residual as estimated between frames 52 and 53 of the "NYC" sequence.

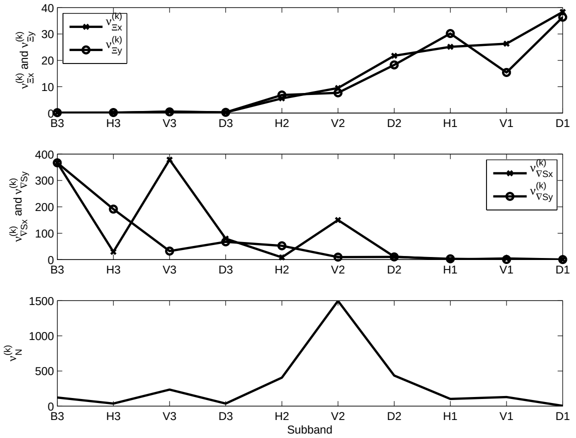

Fig. 7. Variances of the dense-motion error, the spatial gradient, and the dense-motion residual as estimated between frames 30 and 31 of the "Susie" sequence.

TABLE III

DISTORTION AVERAGED OVER ALL FRAMES OF THE SEQUENCE FOR RATE OF 0.5 BPP.

	PSNR (dB)			
	Spatial Block	RDWT Block	RWMH	H.264
Football	28.8	28.8	30.3	30.5
NYC	39.0	38.6	40.0	40.6
Coastguard	32.4	32.1	33.3	33.4
Table Tennis	35.2	34.7	36.1	37.0
Foreman	39.1	39.5	40.8	41.9
Susie	40.9	41.0	42.3	42.4
Mother-Daughter	45.0	45.3	45.9	46.3

Sequences are the same as in Table II. Each system uses $J = 3$, $\beta = -2$, and its own best-matching full-search ME with $B = 8$; motion-vector overhead included in rate.

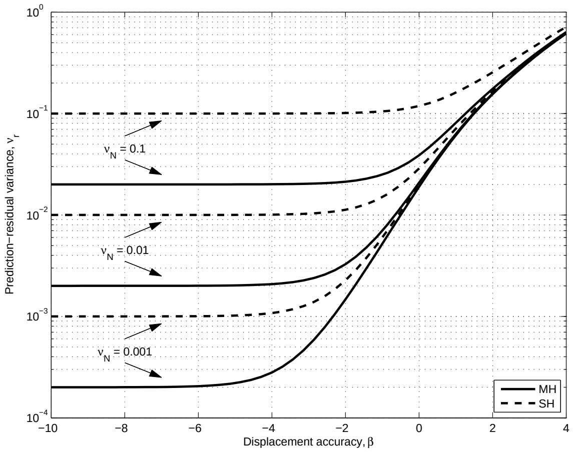

Fig. 8. Prediction-residual variances $\nu_r^{(\mathrm{MH})}$ and $\nu_r^{(\mathrm{SH})}$ for the multihypothesis and single-hypothesis systems, respectively, as the global displacement accuracy β varies at various dense-motion-residual variances ν_N. The number of transform scales is $J = 3$.

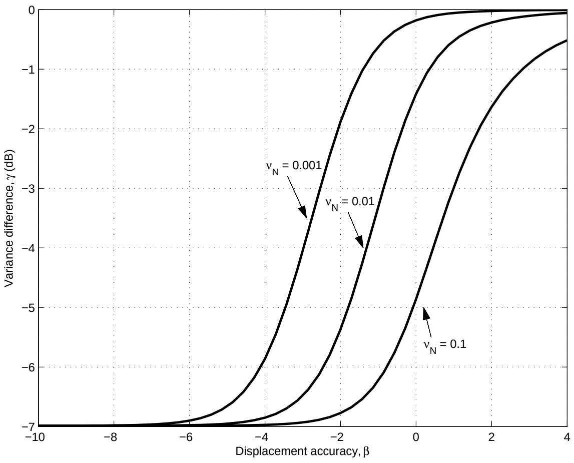

Fig. 9. Difference, γ, in prediction-residual variance between the multihypothesis and single-hypothesis systems as the global displacement accuracy β varies at various dense-motion-residual variances ν_N. The number of transform scales is $J = 3$.

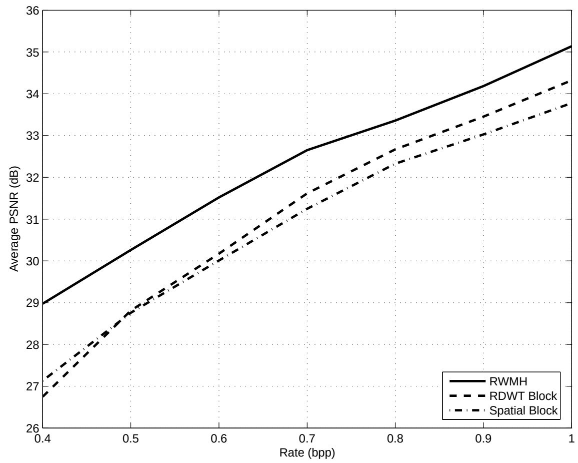

Fig. 10. Rate-distortion performance for "Football" for quarter-pixel accuracy ($\beta = -2, B = 8, J = 3$).

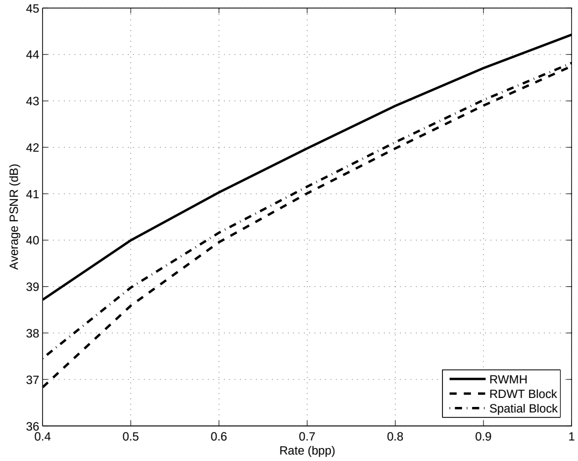

Fig. 11. Rate-distortion performance for "NYC" for quarter-pixel accuracy ($\beta = -2, B = 8, J = 3$).

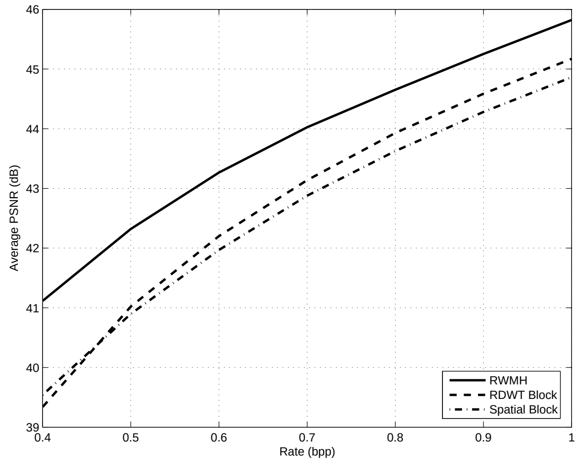

Fig. 12. Rate-distortion performance for "Susie" for quarter-pixel accuracy ($\beta = -2, B = 8, J = 3$).