J. Vis. Commun. Image R. 21 (2010) 917–929

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

A rate-distortion analysis on motion prediction efficiency and mode decision for scalable wavelet video coding

Chia-Yang Tsai, Hsueh-Ming Hang *

Department of Electronics Engineering, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 8 January 2010
Accepted 27 August 2010
Available online 6 September 2010

Keywords:
Interframe wavelet coding
Scalable wavelet video
Motion prediction efficiency
Motion information gain
Rate-distortion optimization
Prediction mode decision
Video coding bit allocation
Video coding rate control

ABSTRACT

A rate-distortion model for describing the motion prediction efficiency in interframe wavelet video coding is proposed in this paper. Different from the non-scalable video coding, the scalable wavelet video coding needs to operate under multiple bitrate conditions and it has an open-loop structure. The conventional Lagrangian multiplier, which is widely used to solve the rate-distortion optimization problems in video coding, does not fit well into the scalable wavelet structure. In order to find the rate-distortion trade-off due to different bits allocated to motion and textual information, we suggest a motion information gain (MIG) metric to measure the motion prediction efficiency. Based on this metric, a new cost function for mode decision is proposed. Compared with the conventional Lagrangian method, our experiments show that the proposed method is less extraction-bitrate dependent and generally improves both the PSNR performance and the visual quality for the scalability cases.

1. Introduction

Over the past few years, multimedia delivery becomes an important class of wireless/wired internet applications, for example, mobile video and digital TV broadcasting. To overcome the constraints on transmission bandwidth and receiver capability, the scalable coding technique was developed and adopted by the recent international video standards. There are two major approaches on scalable video coding: the DCT-based and the wavelet-based coding schemes. These two coding schemes share many similar coding concepts, especially in removing the temporal redundancy. The scalable video coding (SVC) extension of the H.264/AVC is a representative scheme of the DCT-based approach and has been accepted as the ITU/MPEG standards in 2007 [1]. On the other hand, the wavelet-based coding scheme is a relatively new structure and has its potential and advantages [2] as shown during the MPEG competition process for standardization.

Discrete wavelet transform (DWT) has been successfully applied to still image compression. By exploiting the inter-subband or intra-subband correlation, the DWT transformed image signal can be efficiently compressed by a context-based entropy coder, such as EZW [3], SPIHT [4], and EBCOT [5]. Different from the DCT-based JPEG image coding, the multiresolution property of wavelet transform provides a natural way in producing scalable bitstreams. It enables the spatial and the SNR scalability features in the well-known JPEG2000 image coding standard [6]. In addition to the spatial decomposition, DWT can also be applied along the temporal axis and decomposes video frames into temporal subband signals. Therefore, it provides the temporal scalability for videos. In the past 15 years, the temporal wavelet decomposition is refined by adopting the motion compensated temporal filtering (MCTF) technique. These schemes were proposed and improved by Ohm [7], Hsiang and Woods [8], Secker and Taubman [9], and Xu et al. [10]. MCTF can efficiently decompose video frames along the motion trajectories. After MCTF and spatial 2-D DWT, the original video frames are transformed to spatio-temporal subband signals and compressed by a context-based entropy coder [9,11]. This interframe wavelet video coding scheme can achieve temporal, spatial and SNR scalability goals simultaneously. Depending on the processing order in the spatio-temporal domain, the scalable wavelet coding methods can be classified to "t+2D" and "2D+t" structures [12]. In this paper, we will focus on the t+2D structure.

The rate-distortion analysis of a scalable interframe wavelet video coder is very different from that of a DCT-based coder owing to the following two issues: inter-scale coding and open-loop coding structure. In DCT-based video coders, such as MPEG-2 or H.264, use the hybrid coding technique; all the temporal and spatial prediction operations are basically block-based. Thus, it is quite straightforward to perform the rate-distortion analysis along the coding operation flow. On the other hand, in the interframe wavelet coders, the temporal MCTF is performed block-wise, but the spatial entropy coding is performed on the subbands. This

* Corresponding author.
E-mail addresses: cytsai.ee94g@nctu.edu.tw (C.-Y. Tsai), hmhang@mail.nctu.edu.tw (H.-M. Hang).

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.09.001

inconsistent data partition increases the rate-distortion analysis difficulty drastically. Wang and van der Schaar proposed a solution in [13] to analyze the rate-distortion behavior across different coding scales for wavelet video coder. The second issue is that the DCT-based video coder has a closed-loop coding structure. The prediction errors within the loop can be controlled by adjusting coding parameters [14]; thus, the optimal rate-constrained motion compensation can be adaptively adjusted [15,16]. But the interframe wavelet coding has an open-loop prediction structure and the quantization process is performed after all the encoding operations are completed. This open-loop scheme provides more flexibility on bitstream extraction and robustness to transmission errors, but it has no feedback path to provide useful information to adjust prediction parameters in the encoding process. Therefore, it is difficult to achieve the rate-distortion optimization target, especially in the case of allocating bits between the motion and the texture data at multiple operation points all at the same time. How to generate adequate amount of motion information and decide the best prediction modes for MCTF becomes a challenging problem in the scalable interframe wavelet video coding.

Our objective is to develop a suitable mode decision method to achieve the rate-distortion optimization goal in interframe wavelet coding. Our approach is to derive an analytical model that describes the trade-off between the motion compensation bits and the residual texture coefficients bits. We then allocate bits to each category properly at different scalability dimensions. We first examine the rate-distortion effect due to the increase or decrease of motion information bits. Extending our previous work in [17], we derive a quantitative expression to measure the motion prediction efficiency. Most significantly, we give a theoretical explanation to this metric from the entropy viewpoint. Based on this finding, a new cost function is proposed. By minimizing the proposed cost function, the best prediction mode is decided and the corresponding motion vectors are chosen for the MCTF operation. Compared with the mode decision procedure in the conventional scalable wavelet video coder, the proposed method shows a PSNR improvement for the combined SNR and temporal scalability cases.

The paper is organized as follows. Section 2 gives a brief review of interframe wavelet video and the rate-distortion mechanisms in video coding. In Section 3, we propose the motion information gain (MIG) metric to measure the motion prediction efficiency. According to our source model, the MIG metric is further discussed from the entropy viewpoint. Extending the MIG concept, we propose an MIG-based cost function to decide the best prediction mode in Section 4. Section 5 shows the experimental results and compares both the PSNR coding performance and visual quality with the conventional scheme. Finally, a conclusion is given in Section 6.

2. Rate control issues in scalable wavelet video coding

2.1. Brief introduction to interframe wavelet video coding

The most popular coding structure of interframe wavelet video codec is the so-called "t+2D" structure as shown in Fig. 1. The order of "t+2D" implies the encoding operation order: the temporal analysis first and then the spatial analysis. The temporal analysis employs the MCTF technique. It decomposes a group of pictures (GOP) into several temporal high-pass frames and one low-pass frame along the motion vector trajectories. The motion information portion is, in the conventional approach, non-scalable, which is denoted as v in Fig. 1. Then, the spatial decomposition operation (2-D DWT) is applied to the low-pass and high-pass frames to form subbands for further quantization and entropy coding. With the help of a scalable entropy coder, these spatio-temporal subbands are compressed to a scalable bitstream, denoted as s in Fig. 1. Therefore, the coded output bitstream consists of two parts, one is the scalable bitstream for the texture information (s) and the other is the non-scalable bitstream for the motion information (v); together, they are denoted as $\{s, v\}$. To fulfill the application requirements imposed on the video bitrates, image resolution, and frame rate, the texture bitstream is truncated accordingly but the motion bitstream remains intact. Therefore, the output bitstreams of the bitstream extractor are $\{s'_0, v\}, \{s'_1, v\}, \ldots, \{s'_n, v\}$ to match the scalable requirements $r_0, r_0, \ldots, r_n$, respectively, as shown in Fig. 1. The truncation mechanism is designed to collaborate with the scalable entropy coder.

The EBCOT [5] image coding algorithm is adopted by the JPEG2000 standard, and similar algorithms are widely adopted by the state-of-art wavelet video codecs [9,11]. The basic coding flow of an interframe wavelet video coder is as follows. After temporal and spatial analysis, each subband is partitioned into a number of code blocks, and the bitplanes of each block are processed by a few coding paths. The boundary between two consecutive coding paths is a truncation point. These truncation points are characterized by the slopes of the rate-distortion curves at the truncation point. These slope values are recorded and sent to the bitstream extractor. In one extraction unit, such as one GOP, the coding paths with similar slopes are grouped into the same coding layer. A permissible positive slope value is called a rate-distortion threshold. The coding layers with the absolute values of their slopes higher than the rate-distortion threshold are chosen to form an output bitstream. The sum of the bitrates of these chosen coding layers is calculated. If the calculated bitrate is less than the target bitrate, the rate-distortion threshold is adjusted to a smaller value so that more coding layers will be included and the total bitrate increases. On the other hand, the threshold value increases so as to discard some coding layers. By repeating the above operation, the bitrate of the truncated bitstream reaches the target value. Because each bitplane of a code block is split into three coding paths, the bitrate extraction can be quite accurate. Therefore, the bitrate of the texture bitstream can be precisely controlled by the bitstream truncation mechanism. But the non-scalable motion information imposes a constraint on bitstream scalability. The motion information is typically temporal scalable and can be adapted to different decoding frame rates. However, when the spatial scalability feature is turned on, the motion information is often not adjustable to different decoding picture size during the extraction. In the following sub-section, we will compare the rate-distortion optimization

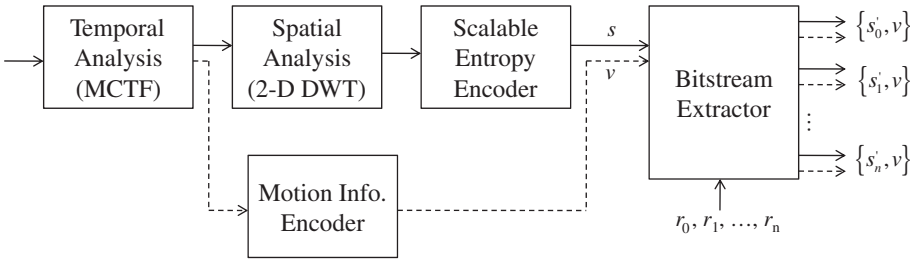

Fig. 1. The t+2D coding structure of interframe wavelet encoder. The solid line and dashed line show the data paths of the texture and motion information, respectively.

methods for the non-scalable and the scalable video cases, and then develop the methods in the next section to adjust the motion information bitrate.

2.2. Rate-distortion mechanism in video coding

According to the Shannon's source coding theory [18], the rate-distortion function can be derived from the probability model of a coding source. Based on the rate-distortion function and with the help of optimization methods, an optimal rate-distortion trade-off can be theoretically obtained for a given bitrate or distortion condition.

In a typical hybrid video coding scheme, the coding source is the transformed residual signal after inter or intra predictions. It is well-known that the probability distribution of the transformed coefficients can be closely approximated by the Laplacian distribution [21]

$$P(x) = \frac{\Lambda}{2}\exp\{-\Lambda|x|\}, \tag{1}$$

where Λ is the Laplacian parameter and can be estimated from the signal standard deviation σ by $\Lambda = \frac{\sqrt{2}}{\sigma}$. If the probability distribution of the transformed residual signal is a Laplacian source, its rate-distortion function with quantization distortion D and texture coding rate R was derived in [18]. In addition to the texture coding bitrate, the extra side information needed in a hybrid coder is mostly the motion information rate ΔR. According to the optimization theory, the best motion prediction mode can be obtained by minimizing the Lagrangian cost function defined by

$$J_{Mode} = D + \lambda_{Mode}(R + \Delta R), \tag{2}$$

where λ_{Mode} is the Lagrange parameter. For a fixed ΔR, λ_{Mode} can be theoretically derived for a well-defined rate-distortion function in (2). Both the theory and the real data show that the λ_{Mode} value is strongly related to the quantization step size, which controls the amount of distortion directly [22,23]. Different λ_{Mode} values are used by several popular reference encoders. These λ_{Mode} values are picked or derived based on their system characteristics and the experimental data [24]. The rate-constrained motion estimation is performed separately by using another Lagrangian cost function given by

$$J_{Motion} = FD + \lambda_{Motion}\Delta R, \tag{3}$$

where FD is a function of the frame difference between the original and the reconstructed image blocks. In many practical systems, FD is either SSD (sum of squared differences) or SAD (sum of absolute differences). In the MPEG reference encoder, λ_{Motion} is empirically chosen to be λ_{Motion} and $\sqrt{\lambda_{Mode}}$ for SSD and SAD, respectively [22].

From (2) and (3), λ_{Mode} is, clearly, an important factor that balances the weights of rate and distortion in the overall cost (J) and it thus affects the bitrates allocated to the texture and the motion information. As discussed earlier, λ_{Mode} depends on the source characteristics, the quantization step size and the bitrate. Several papers [19,20] show that the statistics of the texture are helpful in selecting the proper λ_{Mode} value. The key for solving the mode decision and bit allocation problem is to find the relationship between quantization step size, texture characteristics and bitrate.

Using only one fully self-embedded bitstream to satisfy different coding requirements simultaneously is the most attractive feature of the scalable video coding technique. In the scalable interframe wavelet coding, the bitstream generation process and bitstream extraction process are two separate, independent steps. The encoding process generates lossless compressed bitstream. After the encoding, the extractor truncates the lossless bitstream according to the bitrate requirement. In other words, the extractor plays the role of quantizer. This coding structure uses the input source frames, not the *reconstructed frames*, to predict the current frame. It is often referred as "open-loop structure" in the 3D wavelet coding literature [12]. It is very difficult to precisely control the prediction accuracy during the encoding process. Moreover, multiple bitstreams are to be extracted from the same coded bitstream. It is hard to adequately allocate the motion information bitrates at encoder (before the extractor) to satisfy all target operation points simultaneously. A theoretical treatment on the optimum trade-off between the motion information bitrate and the texture signal bitrate for a motion-compensated video codec was earlier explored by Girod [15] and will be discussed in the next section. In practice, most existing scalable wavelet video coding schemes still adopt the cost functions used in the hybrid video coding ((2) and (3)), but the Lagrange parameter in each temporal decomposition stage is manually selected empirically [25]. Because the target bitrate is given after the entire bitstream is coded, the pre-selected, fixed-value Lagrange parameter must be working for a range of bitrates. In other words, we hope it can provide a reasonable overall performance for all the bitrates of interest. The cost function defined by (2) determines the best motion prediction mode. If a total bitrate is given, we can follow the conventional approach to pick up the Lagrange parameter. But unfortunately, the bitrate is not known at the encoding stage for scalable wavelet video encoding.

To go one step further, we look into the role that the motion vectors play in scalable interframe wavelet coding. The MCTF unit performs the temporal decomposition operation along the motion trajectory; therefore, the accuracy of motion vectors is critical to their motion compensation performance. The low-pass frames produced by temporal filtering will be further decomposed at the next temporal level. Thus, the temporal decomposition layers form a hierarchical structure. The inefficiency in motion prediction propagates along the temporal hierarchy in the same GOP. Therefore, accurate motion vectors tend to decrease the overall distortion. But, a very accurate motion vector often requires more coding bits.

To sum up, the Lagrangian cost function is a very powerful tool in the conventional non-scalable coder. But due to the open-loop coding structure and the requirement of multiple operating points, the use of the Lagrangian cost function in scalable wavelet video coding becomes inadequate. The key problem is finding the proper trade-off between the motion information and the residual texture information for scalable wavelet video coder. The whole scenario becomes even more complicated when we consider the propagation of MCTF inefficiency along temporal hierarchy. Therefore, we propose another approach to replace the ordinary Lagrangian cost function for scalable wavelet video coding.

3. Motion information gain (MIG) metric

A typical extraction process in scalable wavelet coding truncates only the encoded texture bitstream and maintains the integrity of the entire encoded motion information. For a given bitrate condition, different amounts of motion information lead to different types of residual texture signals, and thus lead to different rate-distortion behavior. Although there are other approximate solutions [26,27] that select the scalable motion information to match certain very low-bitrate requirements, we focus on the pre-partitioned motion information solution in the following study. That is, the optimal amount of information bits is decided at the encoding stage.

We first analyze the rate-distortion behavior of the motion-predicted residual signals. Then, based on this rate-distortion relationship, we derive a quantitative metric that measures the coding efficiency of motion information. Also, a theoretical explanation from the entropy viewpoint is given to our coding efficiency metric.

3.1. Rate-distortion model of motion-compensated prediction

For a scalable wavelet video coder, theoretically, we can fix an extraction bitrate and then find the rate-distortion behavior due to the increase/decrease of motion information. In other words, at a given bitrate, if a certain amount of the texture bitrate is shifted to the motion information, will the reconstructed image distortion be reduced or increased? A solution to this problem is searching for the optimal motion information that leads to the optimal R-D performance at different bitrates. For example, is the block size or the motion vector accuracy more important in improving the coded image quality? Clearly, the answer depends on both picture content and bitrate.

Although the residual frames after MCTF will be further spatially decomposed by 2-D DWT, in this study we focus on the rate-distortion behavior of the texture information at the MCTF stage (not after 2-D DWT) because the motion information coding efficiency is our main concern. Because the consecutive frames are often very similar, the motion-predicted residual signals typically have zero-mean and nearly symmetrical distribution. The residual signals after motion prediction can be modeled as Laplacian sources. Because the temporal high-pass frame is essentially a weighted combination of the motion-predicted residual frames, we next try to construct the rate-distortion model of the motion-compensated residual signals.

When the residual texture signal is produced by the motion prediction operation, the rate-distortion behavior of this texture information portion is decided. That is, since the residuals are fixed after motion compensation, their rate and distortion trade-off due to quantization and entropy coding is also fixed. However, if we change the motion vectors (mv) used in motion prediction, the residual signals are different and thus, the texture rate-distortion function changes. We like to know the texture rate-distortion function variation before and after the motion prediction being applied to the same coding block.

For a motion-compensated video codec, Girod [15] pointed out that at a given total bitrate, the optimum trade-off point should locate at

$$\frac{\partial D}{\partial R_{texture}} = \frac{\partial D}{\partial R_{mv}}, \tag{4}$$

where the left-hand side is the distortion decrease due to texture rate increase and the right-hand is the distortion decrease due to motion information rate increase. Fig. 2 gives an illustration of this principle. We use the zero motion vector (no motion-compensation) case as a reference. In Fig. 2, $D_0(R)$ is the rate-distortion function of the residual signal produced by using the zero motion vector, and $D_{\mathbf{v}}(R)$ is the rate-distortion function of the residual signals produced with the motion vector set $\mathbf{v}$. From the bitrate viewpoint, an extra coding bitrate ΔR is needed for sending the motion vectors $\mathbf{v}$. Since the total target bitrate R_T is given, the bitrate available for the texture information is reduced to $R_T - \Delta R$. If this set of mv is beneficial for the overall performance, the quantization error (distortion) of the texture information with mv should be less than that without mv at the same target bitrate. Otherwise, the motion compensation is judged inefficient. Therefore, the distortion with motion prediction is smaller than that without motion prediction:

$$D_v(R_T - \Delta R) < D_0(R_T). \tag{5}$$

Conceptually, (5) is equivalent to (4) in [15]. But different from the motion region partition approach in [15], we try to find an instrumental trade-off measure and a design procedure for adjusting the mv bitrate.

For the Laplacian source described by (1), if the absolute-error distortion measurement is in use, (5) can be rewritten using the rate-distortion functions given in [18] as

$$\frac{1}{\Lambda_{\mathbf{v}}} \cdot 2^{-(R_T - \Delta R)} < \frac{1}{\Lambda_{\mathbf{0}}} \cdot 2^{-(R_T)}. \tag{6}$$

The Laplacian parameter $\Lambda_{\mathbf{v}}$ and $\Lambda_{\mathbf{0}}$ can be estimated from the residual signal variances, $\sigma^2_{\mathbf{v}}$ and $\sigma^2_{\mathbf{0}}$, respectively. That is, $\Lambda = \sqrt{2}/\sigma$. Thus, (6) becomes

$$\frac{\log_2(\sigma_{\mathbf{0}}) - \log_2(\sigma_{\mathbf{v}})}{\Delta R} > .1 \tag{7}$$

Let us define the function Φ to be the logarithm value of the signal standard deviation, and let $\Delta\Phi$ be

$$\Delta\Phi \equiv \Phi_{\mathbf{0}} - \Phi_{\mathbf{v}} = \log_2(\sigma_{\mathbf{0}}) - \log_2(\sigma_{\mathbf{v}}). \tag{8}$$

Then, (7) can be rewritten as

$$\frac{\Delta\Phi}{\Delta R} > 1. \tag{9}$$

From (5)–(9), we can see that the target bitrate term R_T is cancelled because it appears on both sides in (6). This target bitrate elimination gives us a big advantage in the rest of our rate-distortion analysis. Different from the conventional video coding, the target (extraction) bitrate is unknown during the scalable encoding process. In this formulation, the measurement of motion prediction efficiency is extraction bitrate irrelevant. This is true under the assumption that the residual signal probability distribution is Laplacian for both with and without motion-compensated prediction. This Laplacian model is not all accurate in real cases. Here, $\Delta\Phi$ and ΔR represent the variation of texture statistics and the bitrate cost of adopting motion estimation, respectively. We thus view $\Delta\Phi/\Delta R$ as a *gain* factor in measuring the motion prediction efficiency. Intuitively, the motion prediction operation is preferred if it reduces the texture variance significantly. Furthermore, (9) gives a quantitative metric and specifies a threshold of acceptable $\Delta\Phi/\Delta R$. This threshold is derived based on the Laplacian source assumption with absolute-error distortion definition.

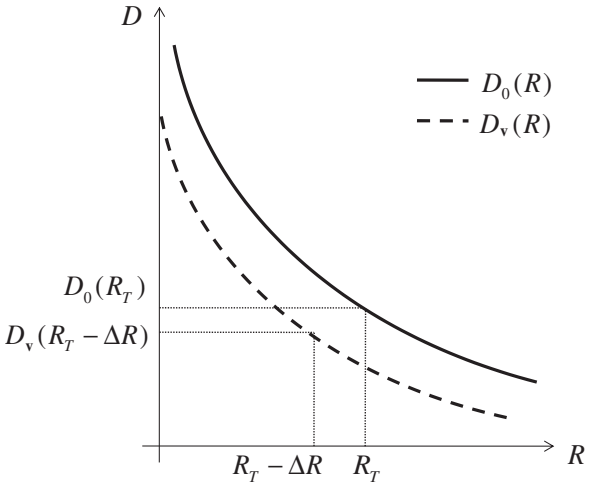

Fig. 2. Illustration of rate-distortion curves of texture residual signal before and after motion prediction.

3.2. Motion information gain (MIG)

According to the last sub-section, $\Delta\Phi$ represents the variation of texture statistics due to motion-compensated prediction. We are going to show next that $\Delta\Phi$ represents the difference between two differential entropies. For the Laplacian source X, its differential entropy $h(X)$ is given below [18].

$$h(X) = -\int_X P(X)\log_2(P(X))dx = -\int_{-\infty}^{\infty} \frac{\Lambda}{2} e^{-\Lambda|x|} \cdot \log_2\left(\frac{\Lambda}{2} e^{-\Lambda|x|}\right) dx = 1 + \log_2\left(\frac{e}{\Lambda}\right), \tag{10}$$

where Λ is the Laplacian parameter. Thus, the differential entropies of the residual signals $X_{\mathbf{0}}$ and $X_{\mathbf{v}}$ produced by the zero motion vector and the motion vector set $\mathbf{v}$ are, respectively,

$$h(X_{\mathbf{0}}) = 1 + \log_2\left(\frac{e}{\Lambda_{\mathbf{0}}}\right), \quad h(X_{\mathbf{v}}) = 1 + \log_2\left(\frac{e}{\Lambda_{\mathbf{v}}}\right). \tag{11}$$

Although the differential entropy does not represent the actual bitrate, the difference between two differential entropies represents the bitrate difference estimation of these two sources. Since the Laplacian parameter can be estimated from the signal variance, we thus obtain the following equation:

$$h(X_{\mathbf{0}}) - h(X_{\mathbf{v}}) = \log_2\left(\frac{\Lambda_{\mathbf{v}}}{\Lambda_{\mathbf{0}}}\right) = \log_2\left(\frac{\sigma_{\mathbf{0}}}{\sigma_{\mathbf{v}}}\right). \tag{12}$$

Comparing (12) with (8), as a consequence of rate-distortion theory on the Laplacian source, we find that these two equations are the same. Therefore, $\Delta\Phi$ represents the reduction of residual signal entropy in encoding the residual signals before and after motion-compensated prediction. Thus, the interpretation of $\Delta\Phi/\Delta R$ is as follows.

$$\frac{\Delta\Phi}{\Delta R} \sim \frac{\text{decrease in residual signal entropy}}{\text{increase in motion information bitrate}}. \tag{13}$$

From (13), we can see that $\Delta\Phi/\Delta R$ is the ratio of the "reward" and the "cost" due to the use of motion-compensated prediction. The "cost" is the extra bitrate for encoding the motion vectors, and the "reward" is the entropy reduction of the residual texture signals. Therefore, $\Delta\Phi/\Delta R$ is called the "motion information gain", abbreviated as MIG. It is thus used to measure the motion prediction efficiency. We denote this MIG function due to the motion vector set $\mathbf{v}$ by

$$\phi(\mathbf{v}) \triangleq \frac{\Delta\Phi}{\Delta R}. \tag{14}$$

This gain factor implicitly represents the trade-off between the residual signal bitrate and motion information bitrate. The fundamental concept behind (14) is similar to that (4) in [15] as discussed earlier. But through our preceding lengthy derivation, we show that the total target bitrate disappears in the final MIG expression. Thus, the MIG metric fits well for applying to the scalable wavelet video coding structure.

Let us extend the original criterion (9) to a more general form. When we consider the advantage of using motion prediction in scalable wavelet video coding, the MIG metric of the candidate motion vector set $\mathbf{v}$ should satisfy

$$\phi(\mathbf{v}) > C, \tag{15}$$

where C is a chosen threshold value. In the original derivation, C is 1. Here we investigate the range of C values in real video coding cases. Because a practical entropy coder cannot approach the entropy bound, both the compressed texture and the compressed motion information would need more bits to code. Therefore, the motion prediction is not as effective as (5) shows. The distortion reduction by the motion bitrate ΔR, measured in bits/pixel, is less than the expected value; that is, $D_{\mathbf{v}}$ should be larger in real cases. Therefore, (5) is modified to

$$a \cdot D_{\mathbf{v}}(R_T - \Delta R) < D_{\mathbf{0}}(R_T), \tag{16}$$

where $a > 1$. Using the above equation, we can follow the same derivation process in Section 3.1 to obtain the MIG lower bound. Consequently, an inequality similar (7) is derived:

$$\frac{\log_2(\sigma_{\mathbf{0}}) - \log_2(\sigma_{\mathbf{v}})}{\Delta R} > 1 + \frac{\log_2(a)}{\Delta R}. \tag{17}$$

Because $a > 1$, the right term of the above equation, the lower bound of C, is larger than 1. When ΔR is small or $\log_2(a)$ is large, C becomes much larger than 1.

4. Mig cost function and mode decision procedure

As we discuss in the previous sections, the Lagrangian multiplier approach works well for the single bitrate optimization but is not suitable for the multiple-rate scalable coding optimization. Because the MIG metric is independent of the target (extraction) bitrate and it is proportional to the cost and reward ratio when the motion-compensated prediction is activated, it can serve as an indicator for deciding the motion estimation mode and parameters. We are thus inspired to propose a new cost function and an associated mode decision procedure based on MIG for scalable wavelet video coding.

4.1. Properties of MIG

Since our motion mode and vector selection process is applied only to image blocks with non-zero optimal motion vectors, the denominator of (14) is non-zero. There are a few interesting properties associated with $\phi(\mathbf{v})$.

(1) $\phi(\mathbf{v}) \geqslant \mathbf{0}$. Clearly, we will not use an mv that produces a negative $\Delta\Phi$ value. For a given image block, if the zero mv is the best mv in the sense that any non-zero mv cannot reduce the residual signal variance, then the $\phi(\mathbf{v})$ value associated with this block is assigned to be 0 and the best coding mode is the one with the zero motion vector.

(2) $\phi(\mathbf{v})$ is bounded. In digital image coding, the residual signal has a finite variance. The best non-zero mv can, at the best, reduce the residual variance to zero. The variance difference before and after employing mv is thus finite. In other words, the $\phi(\mathbf{v})$ value saturates and cannot be further improved when a proper mv is identified.

(3) In the following sections, we deal mainly with the case that $\phi_{max} \geqslant \phi(\mathbf{v}) > C$. That is, the useful mv, $\mathbf{v}$, should produce a $\phi(\mathbf{v})$ value greater than 0 and less than or equal to ϕ_{max}. Ideally, the parameter C is 1 and is independent of image contents and target bitrate if the Laplacian rate-distortion model holds. However, as discussed earlier, practically C is not 1 and is bitrate dependent.

4.2. The proposed mig cost function

Intuitively, the MIG metric $\phi(\mathbf{v})$ with the constraint, $\phi_{max} \geqslant \phi(\mathbf{v}) > C$, can be the cost function used for searching for the optimal mv. However, the C value is unknown and to be identified in real image coding. Thus, for the convenience in computation, we use the following equivalent form. We expand (15) with the aid of (8) and (14). The inequality becomes

$$\sigma_{\mathbf{0}}^2 > \sigma_{\mathbf{v}}^2 \cdot 2^{2 \cdot C \cdot \Delta R}. \tag{18}$$

A large MIG value implies a large $\Delta\Phi$ and/or a small ΔR. In (8), a large $\Delta\Phi$ value implies that the difference between $\sigma_{\mathbf{0}}$ and $\sigma_{\mathbf{v}}$ is large. Thus, the right term in (18), $\sigma_{\mathbf{0}}^2 > \sigma_{\mathbf{v}}^2 \cdot 2^{2 \cdot C \cdot \Delta R}$, should be as small as possible. Therefore, we propose a so-called "MIG cost function" to measure the prediction cost. For a coding source $\mathbf{s}$,

the motion vector set **v** produces the residual signals with variance $\sigma_{\mathbf{s}}^2(\mathbf{v})$ and its average information bitrate (for representing **v**) is $\Delta R(\mathbf{v})$. The MIG cost function J is defined as

$$J(\mathbf{s},\mathbf{v}|C) = \sigma_{\mathbf{s}}^2(\mathbf{v}) \cdot 2^{2\cdot C\cdot \Delta R(\mathbf{v})}, \tag{19}$$

where C is generally source and bitrate dependent. We include it explicitly in the argument of the J function to emphasize its role in our rate control algorithm. The problem now becomes looking for **v** that minimizes J.

We need to identify the value of C in (19). According to our previous discussions, the C value is decided by the coding system and the source signal **s** in (14). In practice, the source signal **s** is the temporal high-pass frames generated by MCTF. Indeed, the probability distributions of the different temporal layers have different shapes [28]. We conduct the following experiments to characterize J and also to identify the value of C.

We start with a fixed C value and simply use (19) as the cost function in performing motion estimation and mode decision in encoding. The detailed procedure of mode decision will be described in the next sub-section. After the encoding process is done, the encoded bitstream is truncated to a fixed bitrate, for example, 256 Kbps, and then we decode the truncated bitstream. The mean-squared error (MSE) between the decoded and the original images is calculated; thus, one test point of a MSE and C pair is obtained. The data are collected from 32 frames of the Mobile sequence at CIF resolution.

Repeating the above steps with different C values, we obtain a MSE vs. C curve at 256 Kbps as shown in Fig. 3(a). By changing the truncation bitrates settings, the MSE vs. C curves at 384 and 800 Kbps are obtained as shown in Fig. 3(c) and (d), respectively. Each of Fig. 3(a)–(c) shows that the MSE is minimal when C reaches a certain value. This is equivalent to the performance saturation phenomenon we discuss earlier. When C is large, only the very effective mv's can make positive contribution and their value is diminishing as C gets larger; and thus the MSE goes up again as shown in Fig. 3(a)–(c). Although the theory predicts that MIG is independent of bitrate, in reality, however, the coding system efficiency and the source probability distribution are bitrate and temporal-level dependent. Indeed, the best C value that leads to the minimum MSE tends to be smaller at higher bitrates. This is consistent with the known observation that the mathematical model matches the real rate-distortion relationship at higher rates. For example, the rate-distortion relationship of a quantizer approximates the asymptotical R-D function at high bitrates [18]. If the optimum C value does not change significantly, we prefer to use a constant C to cover the bitrates of our interests. We pick up 7

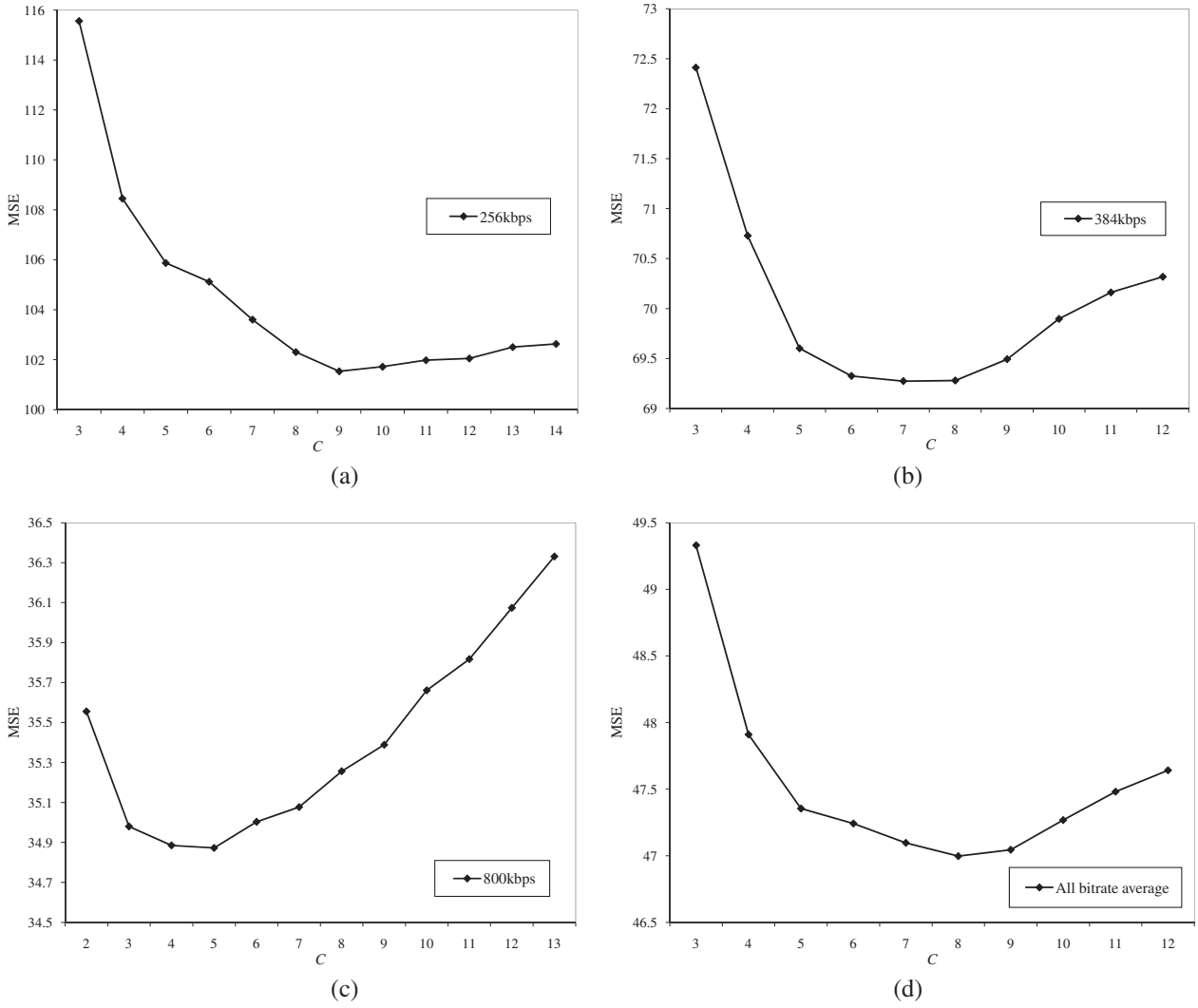

Fig. 3. MSE vs. C value in the MIG cost function at (a) 256 Kbps, (b) 284 Kbps, and (c) 800 Kbps truncation bitrates, and (d) the average MSE for 7 bitrates (Mobile, CIF resolution).

target bitrates, 256, 384, 512, 800, 1024, 1200, and 1500 k, and their average behavior (MSE vs. C) is shown in Fig. 3(d). In conclusion, the C value generally falls in the range of [4,10].

4.3. Temporal-level factor in MIG cost function

After motion-compensated prediction, the relationship between the pixels on the predicted and the reference frames can be classified to three types: *connected*, *unconnected*, and *multi-connected* [8]. During the MCTF process, because the temporal correlation between the low-pass frames at the deep temporal level is relatively small, the unconnected pixel percentage increases, which implies that the prediction effectiveness decreases. Furthermore, the connection relationship leads to the distortion propagation along the tree structure generated by the temporal filtering process after quantization, which is the so-called "quantization noise propagation" problem in MCTF [13,29]. Here we follow the notations defined by [13] in modeling the noise propagation process. The average distortions of the low-pass frame and the high-pass frame at temporal level t are denoted as $\bar{d}_L^{(t)}$ and $\bar{d}_H^{(t)}$, respectively. When the Harr wavelet filter is adopted in MCTF, Wang and van der Schaar [13] show that $\bar{d}_L^{(t)}$ and $\bar{d}_H^{(t)}$ are related to $\bar{d}_L^{(t-1)}$ by the following equation,

$$\bar{d}_L^{(t-1)} = \frac{1}{2}\bar{d}_L^{(t)} + \left(\frac{3}{4} - \frac{r_c}{4}\right)\cdot \bar{d}_H^{(t)}, \tag{20}$$

where r_c is the ratio of the connected pixels. It is obvious that r_c determines the severity of the distortion propagation problem. There are two major factors affecting the r_c value: the picture characteristics and the motion estimation method. By minimizing the MIG cost function with the pre-chosen $C^{(t)}$ parameter (the C value at the t temporal level), the best motion vector set $\mathbf{v}^{(t)}$ can be obtained, and thus r_c is decided. The frames are temporally decomposed along the $\mathbf{v}^{(t)}$ trajectory. Hence, $\bar{d}_L^{(t)}$ and $\bar{d}_H^{(t)}$ are the functions of $\mathbf{v}^{(t)}$. We rewrite (20) as

$$\bar{d}_L^{(t-1)} = \frac{1}{2}\bar{d}_L^{(t)}(\mathbf{v}^{(t)}|C^{(t)}) + \left(\frac{3}{4} - \frac{r_c(\mathbf{v}^{(t)}|C^{(t)})}{4}\right)\cdot \bar{d}_H^{(t)}(\mathbf{v}^{(t)}|C^{(t)}), \tag{21}$$

in which the notation $(.|C^{(t)})$ is inserted to emphasize the result depends on the $C^{(t)}$ value. Thus, in the Haar wavelet filter case, (21) shows that the rate-distortion behavior of the low-pass frame at temporal level $t-1$ is affected by the motion vectors at temporal level t.

Theoretically, to find the optimal solution of mv, the effects of the quantized/truncated residual signals at all the previous temporal levels have to be considered. Practically, because of the open-loop structure and the complexity of the inter-scale coding system, it is very difficult to construct an analytical model, or even an experimental model, to describe the relationship between the distortion propagation and the motion information. A feasible approach is to adjust the C value of (19) along with the increased temporal level. Also, this adjustment changes the values of $\sigma_{\mathbf{s}}^2(\mathbf{v})$and $\Delta R(\mathbf{v})$ according to their located MCTF decomposition layer and thus it can effectively compensate for the propagation distortion loss. Therefore, the MIG cost function of (19) is modified to

$$J(\mathbf{s},\mathbf{v}|C^{(t)}) = \sigma_{\mathbf{s}}^2(\mathbf{v})\cdot 2^{2\cdot C(t)\cdot \Delta R(\mathbf{v})}, \tag{22}$$

where the superscript t is the temporal level index in MCTF. It is shown that the statistical relationship between consecutive subband signals can be modeled by a hidden Markov model [30]. Similarly, a Markov-like relationship seems to exist between consecutive temporal decomposition layers. Thus, the optimally decided distortion values of these layers are correlated. Therefore,

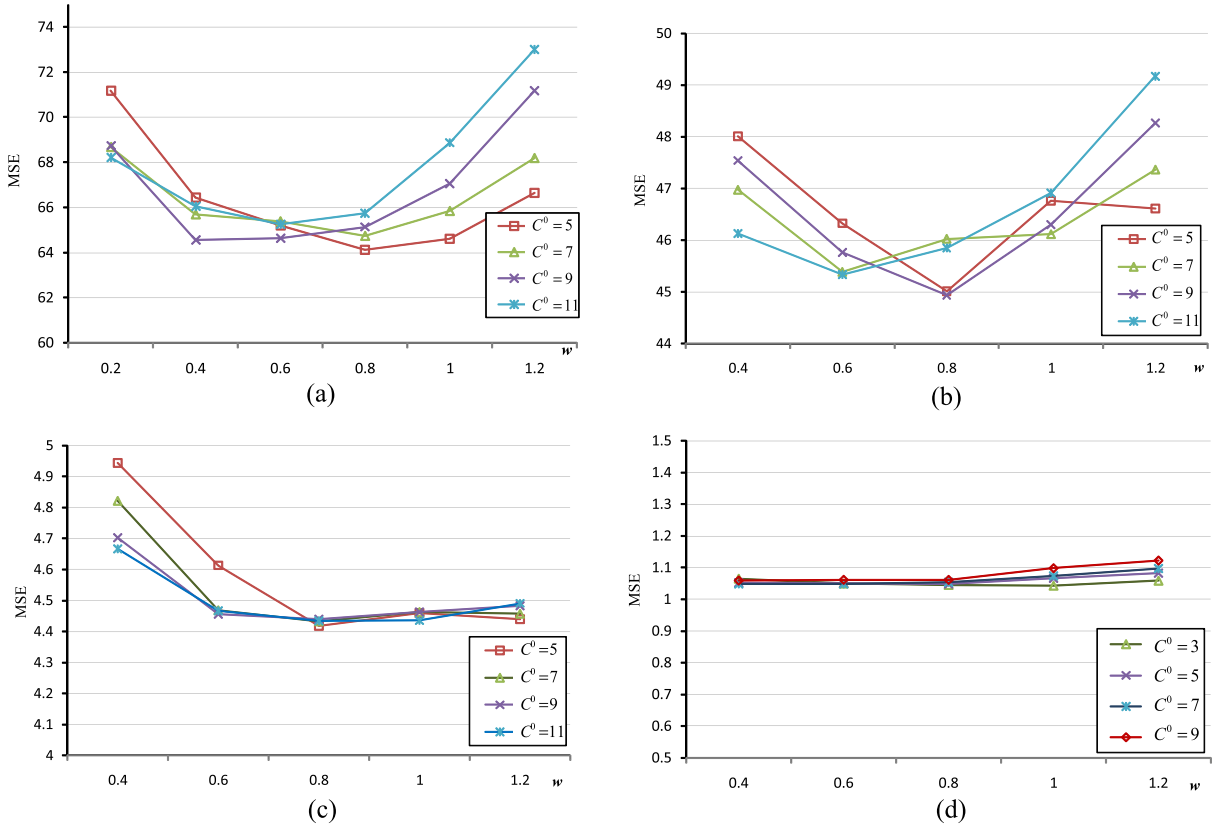

Fig. 4. MSE vs. w value with different $C0$ parameter settings in the MIG cost function: (a) Mobile, (b) Tempete, (c) Container, and (d) Akiyo, all in CIF resolution.

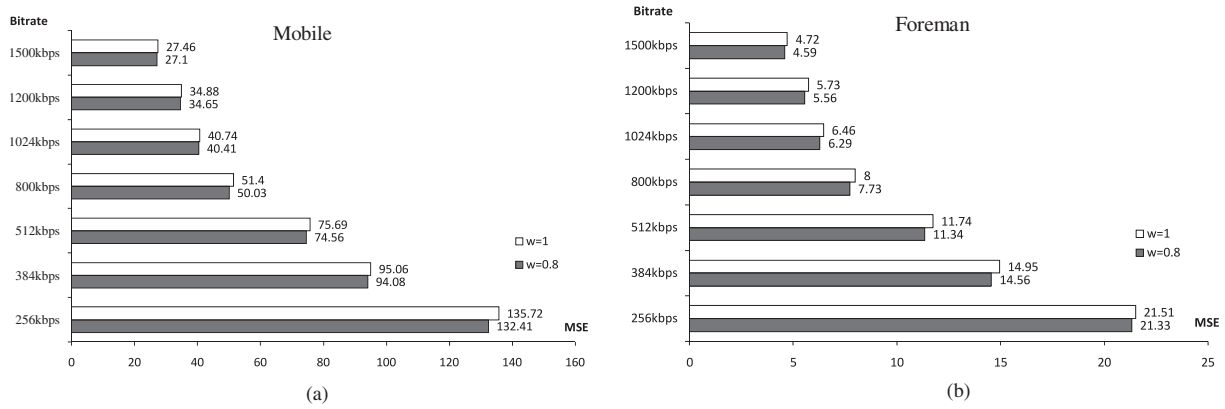

Fig. 5. The MSE comparison between the cases with temporal weighting, $w = 0.8$ and $w = 1$, in the MIG cost function at different truncation bitrates. Test sequences are (a) Mobile and (b) Foreman. (CIF resolution).

we conjecture that a simple linear predictor can describe the relationship of the C parameters among temporal layers. That is, for two consecutive temporal levels,

$$C^{(t)} = w \cdot C^{(t-1)}. \tag{23}$$

Consequently, if C^0 is given for the first temporal level, (23) becomes

$$C^{(t)} = w^{(t)} \cdot C^0. \tag{24}$$

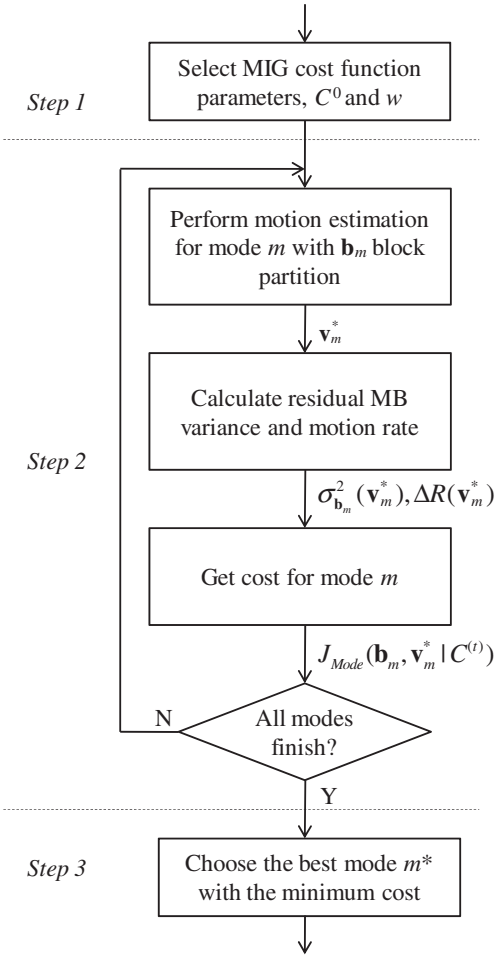

Fig. 6. Flow chart of the proposed mode decision procedure using the MIG cost function.

In practice, the weighting factor w can be found by extensive experiments. We start with a pair of C^0 and w values and use (22) to perform motion search and mode decision. Repeating the same experimental steps for Fig. 3(d) with different w values, we obtain the MSE vs. w curves using different C^0 values. The experimental results are shown in Fig. 4. Because the motion information percentage in fast-motion pictures is larger than that in the slow-motion pictures, the error propagation problem is severe. Hence, the benefit of using our temporal weighting adjustment is more significant in the fast-motion cases. Fig. 4(a) and (d) shows the results of Mobile and Akiyo test sequences, respectively. Compared with Akiyo, Mobile is a relatively fast-motion test sequence, and thus the distortion in Fig. 4(a) is more sensitive to the w value than that in Fig. 4(d). In contrast, the temporal weight adjustment makes little difference in MSE for the Akiyo test sequence. Fig. 4 shows that the average MSE is a convex function in w and the minimal MSE appears at around [0.6, 0.9]. According to the collected data, $w = 0.8$ seems to be a good value for most cases. To verify the effectiveness of our chosen temporal weighting factor, we tested Mobile and Foreman videos and adopted the MIG cost function with weightings, $w = 0.8$ and $w = 1$. In these simulations, the C^0 parameter is set to 7. Fig. 5 shows that applying the temporal weighing factor can improve the overall MSE at different bitrates.

4.4. Block-based mode decision procedure

The MIG cost function can be used to decide the coding mode. It tells us the trade-off between the motion information and the texture information. Based on MIG, we develop a mode decision procedure. In a conventional non-scalable video coder, the best motion vector and coding mode are decided by minimizing the Lagrangian cost function ((2) and (3)) for a given single bitrate. As discussed in the previous sub-sections, with the MIG cost function we are able to choose the most appropriate coding mode (including mv) by minimizing its value. The basic steps in the

Table 1
The optimized default encoder parameters settings [32] for different MCTF temporal level t in the VidWav reference software [25].

Temporal level	Motion search range (pixel)	Motion vector accuracy (pixel)		Lagrange parameter	
		CIF	4CIF	CIF	4CIF
$t = 0$	32	1/4	1/4	16	16
$t = 1$	64	1/2	1/2	32	50
$t = 2$	128	1/2	1	64	150
$t = 3$	128	1/2	1	64	150
$t = 4$	128	1/2	1	64	150

proposed mode decision procedure are similar to that in the conventional scheme. In the existing scalable wavelet video coding schemes, the *mv* search is block-based and the variable block-size motion compensation technique is used to find the best macroblock coding mode. Each macroblock coding mode represents a partition of macroblock into a certain combination of sub-blocks.

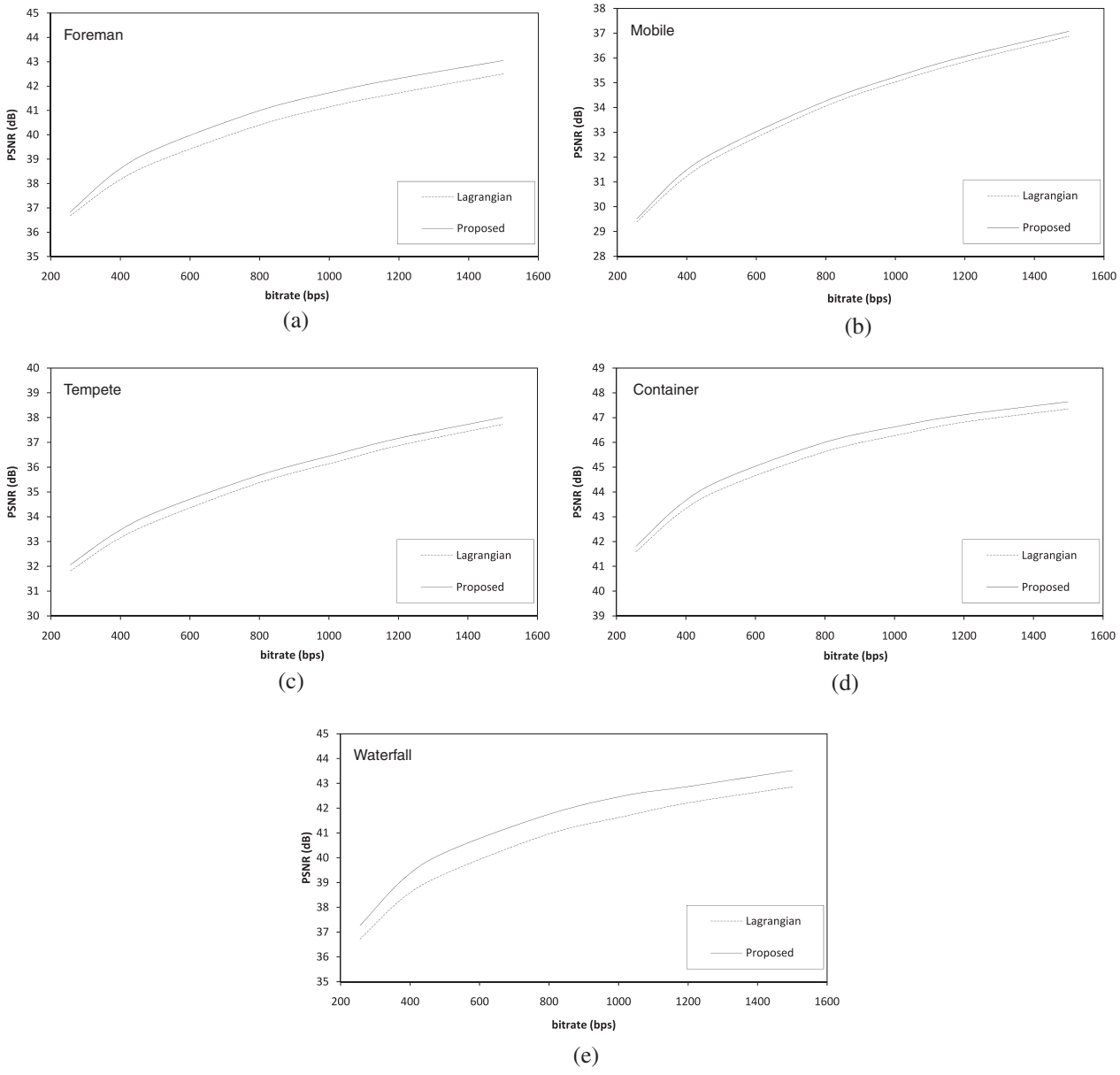

Fig. 7. Rate-distortion performance comparison between the Lagrangian method (dashed line) and the proposed MIG cost function method (solid line). The *x*-axis is the truncation bitrate (Kbps), and the *y*-axis is the decoded video PSNR (dB). The tested video sequences are (a) Foreman, (b) Mobile, (c) Tempete, (d) Container, and (e) Waterfall (CIF resolution).

Table 2
PSNR comparison between the Lagrangian and the proposed methods in the combined temporal and SNR scalabilities tests.

Sequence (4CIF)	GOP size	Cost Function	750 Kbps 15 fps	1024 Kbps 15 fps	1200 Kbps 30 fps	1500 Kbps 30fps	2048 Kbps 60fps	3000 Kbps 60fps
City	32	Lagrangian	36.39	37.33	37.42	37.98	38.49	39.33
		Proposed	36.73	37.69	37.80	38.40	38.85	39.62
Crew	32	Lagrangian	36.39	37.30	36.74	37.34	37.18	38.20
		Proposed	36.46	37.39	36.92	37.52	37.33	38.28
Harbour	32	Lagrangian	33.91	34.97	34.96	35.59	36.25	37.50
		Proposed	33.94	35.01	34.98	35.65	36.29	37.52
Soccer	32	Lagrangian	36.28	37.22	36.92	37.61	38.00	39.20
		Proposed	36.51	37.50	37.19	37.95	38.23	39.43
Ice	16	Lagrangian	40.51	41.65	41.25	42.00	42.41	43.62
		Proposed	40.89	42.05	41.76	42.51	42.87	44.08

Table 3
The average prediction error (squared error) and motion information bits by "MB" mode prediction at first temporal level. All sequences are CIF resolution video with 30 fps.

Test sequence	Cost function	Prediction error/pixel	Motion info. bits/macroblock
Container	Lagrangian	8.36	8.8
	Proposed	3.39	12.4
Foreman	Lagrangian	13.87	13.87
	Proposed	11.07	15.79
Akiyo	Lagrangian	5.1	9.0
	Proposed	3.5	11.4
Mobile	Lagrangian	57.11	16.3
	Proposed	42.23	15.1

Fig. 6 illustrates the proposed mode decision procedure, which consists of three steps as described below.

4.4.1. Step 1: select the appropriate MIG cost function parameters

The proposed MIG cost function contains one parameter, $C^{(t)}$ in (22), and it can be further separated into two parameters, C^0 and w in (24). According to our previous discussions, C^0 and w can be empirically chosen from the intervals, [4,10] and [0.6, 0.9], respectively for the CIF resolution videos.

4.4.2. Step 2: search for the best motion vector set for each block mode

There are many possible sub-block combinations for motion compensation in one macroblock. For example, a typical 16×16 size macroblock has 16×16, 16×8, 8×16, and 8×8 block modes; and each 8×8 block can be further partitioned to 8×4, 4×8, and 4×4 sub-blocks. Assuming that a macroblock can be partitioned to N_m sub-blocks for mode m, the mv's ($\mathbf{v}_i$) associated with all sub-blocks (b_i) form two N_m-tuple vectors, $\mathbf{v}_m$ and $\mathbf{b}_m$, respectively, where

$$\begin{aligned} \mathbf{v}_m &= (\mathbf{v}_1, \ldots, \mathbf{v}_{N_m}), \\ \mathbf{b}_m &= (b_1, \ldots, b_{N_m}). \end{aligned} \tag{25}$$

For each sub-block, to find the best mv, all the mv candidates within the search range S are examined. These candidate motion vectors can have forward, backward or bi-directional prediction directions. By minimizing the MIG cost function in (22), the best motion vector v_i^* for sub-block b_i is obtained. Mathematically, it is identified by performing the following optimization procedure.

$$\begin{aligned} &v_i^* = \arg\min_{v \in \mathbf{S}} \{J_{Motion}(b_i, v|C^{(t)})\} \\ &\text{with } J_{Motion}(b_i, v|C^{(t)}) = \sigma_{b_i}^2(v) \cdot 2^{2 \cdot C(t) \cdot \Delta R(v)}. \end{aligned} \tag{26}$$

Then, the best mv for the macroblock is the collection of all the best motion vectors for mode m; i.e.,

$$\mathbf{v}_m^* = (v_1^*, \ldots, v_{N_m}^*). \tag{27}$$

The residual signal is modeled as a Laplacian source with zero-mean. After all the sub-blocks finish the motion estimation process for mode m, the residual variance $\sigma_{\mathbf{b}_m}^2(\mathbf{v}_m^*)$ and the average motion information bitrate $\Delta R(\mathbf{v}_m^*)$ of a macroblock can be, respectively, expressed as

$$\begin{aligned} \sigma_{\mathbf{b}_m}^2(v_m^*) &= \frac{1}{N_m} \sum_i^{N_m} \sigma_{bi}^2(v_i^*), \\ \Delta R(v_m^*) &= \frac{1}{N_m} \sum_i^{N_m} \Delta R(v_i^*) + r_m, \end{aligned} \tag{28}$$

where r_m is the average extra bits needed to record the coding mode information. Both ΔR and r_m are in bits/pixel.

4.4.3. Step 3: choose the best block mode with the minimum MIG cost

Assuming that the block mode m in *Step 2* belongs to the mode set $\mathbf{M}$, which contains all possible block modes, the MIG cost function in (22) is used again to choose the best macroblock mode. Hence, the best block mode is decided by minimizing the MIG cost function:

$$\begin{aligned} &m^* = \arg\min_{m \in \mathbf{M}} \{J_{Mode}(\mathbf{b}_m, \mathbf{v}_m^*|C^{(t)})\} \\ &\text{with } J_{Mode}(\mathbf{b}_m, \mathbf{v}_m^*|C^{(t)}) = \sigma_{\mathbf{b}_m}^2(v_m^*) \cdot 2^{2 \cdot C(t) \cdot \Delta R(\mathbf{v}_m^*)} \end{aligned} \tag{29}$$

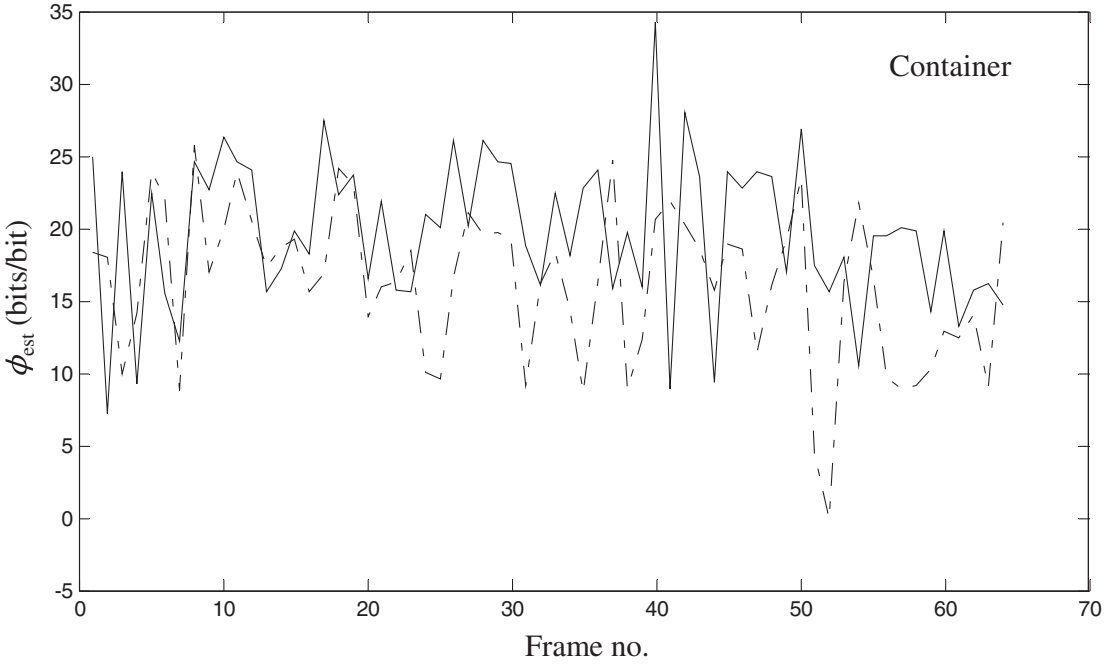

Fig. 8. The estimated MIG values (ϕest) based on the residual frames at the first temporal level for Container and sequence (CIF resolution with 30 fps and 128 frames). The solid line and dash line represent the proposed and Lagrangian approaches, respectively.

Therefore, the best block mode and its associated motion vectors of a macroblock are obtained.

5. Experimental results and discussions

To evaluate the coding performance of the proposed mode decision procedure, our algorithm is implemented on the VidWav reference software. The VidWav software was developed by Microsoft [31] aiming at the MPEG scalable video coding competition in 2004. The default parameter values included in the software were well tuned by the proponents in order to produce good picture quality in competition. It adopted the interframe wavelet video coding structure and a later version was described in [25]. We examine our proposed scheme for two scenarios: (1) the SNR scalability requirement for the CIF (352 × 288, 30 fps) test sequences, and (2) the combined temporal and SNR scalability requirement for the 4CIF (704 × 576, 60 fps) test sequences. To fulfill the scalability requirements, each test sequence is encoded into only one bitstream, and the bitstream is truncated to different bit sizes according to the test conditions. For both scenarios, the PSNR results of two mode decision methods, the conventional Lagrangian and our proposed MIG cost function, are compared. The temporal wavelet filter is the Daubechies 5/3 filter [6].

Here are the details of the encoder parameters in our experiments. First, to obtain a creditable control experimental result, we use the optimized default Lagrange parameters, which were designed by Microsoft in the aforementioned VidWav contest [32]. Table 1 lists the Lagrangian mode decision method parameters. Second, we employ the fast-motion search algorithms included in the VidWav software for both the conventional Lagrangian-based method and our proposed MIG-based method. Third, the motion search range and motion vector accuracy settings, described in Table 1, are the same for both the conventional and our proposed methods. Finally, to choose a proper GOP size, we first find the best GOP size, usually 16 or 32 frames, for each sequence using the conventional Lagrangian method, and then we use the same GOP size for our proposed scheme.

In the first scenario, our objective is to compare the coding performance at different bitrate conditions for the SNR scalability. To evaluate the performance, we test 5 video sequences at the CIF resolution and 30 fps: Foreman, Mobile, Tempete, Container, and Waterfall. In this scenario, the C^0 and w in (24) is set to 7 and 0.8 empirically as discussed in Section 4. Fig. 7 shows the rate-distortion curve comparison between these two coding schemes for the 5 test sequences. Compared to the Lagrangian mode decision, the proposed method shows a PSNR improvement from 0.1 to 0.9 dB depending on sequences. It indicates that our proposed mode decision method generally result in better coding performance in this SNR scalability scenario.

Because the high resolution video transmission becomes more and more popular recently, in the second scenario we test the combined temporal and SNR scalabilities at higher picture resolutions. The video sequences, City, Crew, Harbour, Soccer and Ice, have 4CIF picture size and 60 fps. The coding structure has a three-level

(a) Original raw

(b) Lagrangian

(c) The proposed

Fig. 9. (a) The original image, (b) coded picture using the Lagrangian method, and (c) coded picture using the proposed method. All three pictures are captured and magnified from the 41st frame of the Mobile sequence (CIF, 30 fps). The extraction bitrate is 256 Kbps. The PSNR of (b) and (c) are 30.17 and 30.53 dB, respectively.

temporal decomposition and coding rates from mid bitrate to high bitrates. Table 2 shows the PSNR results of the Lagrangian and the proposed methods. The proposed mode decision method has better performance (0.1–0.5 dB PSNR) in all 30 scalability test points.

We are going to take a close look, at the macroblock (MB) level, at the differences between the Lagrangian method and our method. Our method tends to select the prediction modes that produce large MIG values; that is, a high ratio of the residual signal reduction to the motion information bitrate. We collect various MB statistics including MB mode, prediction errors, and etc. Three types of motion compensation block sizes are in use: 16×16, 16×8, and 8×16. Table 3 shows the average prediction error and the motion information bits at the first temporal level, collected over the entire video sequence. In the case of Foreman test sequence, when compared with the conventional Lagrangian method, our proposed method needs additional 1.92 bits (15.79–13.87) per macroblock, which is about 11 Kbps in total. But the prediction error is reduced by an amount of 2.8 (13.87–11.07) per pixel, or about 0.98 dB PSNR gain in total. This average prediction error reduction is calculated using all macroblocks with MB prediction. In the Mobile case, the proposed method can, in fact, reduce both the prediction errors and the motion information bits at the same time. Table 3 shows 1.2 bits per macroblock reduction and 14.88 per pixel prediction error reduction, or about 1.31 dB PSNR gain. Generally, this trend is found at all temporal levels.

In addition, we examine the MIG values of the coded sequences. As discussed earlier that $\phi(\mathbf{v})$ in (14) represents the ratio between the residual texture entropy reduction and the motion bitrate increase. Hence, in general, a higher $\phi(\mathbf{v})$ is preferred. Here, $H(X)$ denotes the entropy of the motion-compensated residual signal, X. That is,

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_2(p(x)), \tag{30}$$

where $p(x)$ is the probability of x value. In addition, $\mathbf{v}^*$ and $\mathbf{v}_p$ are the searched motion vector set and initial search point for a macroblock, respectively, and thus $\Delta R(\mathbf{v}_p)$ is zero. The MIG value estimated from data is denoted as ϕ_{est}, which is defined below.

$$\phi_{est}(\mathbf{v}^*) = \frac{H(X_{\mathbf{v}^*}) - H(X_{\mathbf{v}_p})}{\Delta R(\mathbf{v}^*)}. \tag{31}$$

Fig. 8 shows the ϕ_{est} per frame results for both the conventional Lagrangian method and the proposed mode decision method at the first temporal level. It can be seem that the proposed MIG approach generally produces higher ϕ_{est} values than the conventional Lagrangian approach. However, a few residual frames may not be modeled well by the Laplacian distribution, and this inaccurate modeling may lead to a poor MB mode selection. Therefore, not all the frames coded using our method have the highest MIG values. From the above experiments and discussions, the proposed mode decision method shows its benefits in the scalable wavelet video coding. Particularly, our scheme has better coding performance on the combined temporal and SNR scalabilities.

Fig. 9 shows the visual quality comparison. The PSNR difference between Fig. 9(b) and (c) is about 0.36 dB. However, one can easily see that the proposed method has a better subjective quality on, for example, the numbers on the calendar (high contrast edges) and the rotating red ball (high motion).

6. Conclusions

In this paper, we propose a rate-distortion model for measuring the motion prediction efficiency and we also develop a mode decision procedure based on this model for interframe wavelet video coding. Because the motion information is encoded once and generates a non-scalable compressed bitstream, it is very difficult to satisfy the multiple bitrates requirements in the scalable interframe wavelet video coding. The key concept in our model is the so-called "motion information gain" (MIG) that represents the cost and reward trade-off between the texture and the motion information. Based on the MIG concept, the MIG cost function is derived to select motion vectors and decide motion prediction mode. We show that ideally the MIG-based cost function is bitrate independent. Therefore, different from the conventional Lagrangian method, the proposed mode decision method can deal with multiple bitrate conditions and is not constrained by single bitrate condition in theory. In practice, we introduce a temporal-level dependent parameter to the MIG cost function to compensate the distortion propagation effect in MCTF. Moreover, to match the real encoding situation, we identify the parameters in our algorithm from the experimental data. By adopting the proposed mode decision procedure, the simulation results show promising PSNR improvements for both the SNR scalability cases and the combined SNR and temporal scalability cases. When we examine the image blocks coded using the Lagrangian method and our method, the experimental data show that the proposed method gives better rate vs. distortion trade-off.

References

[1] ISO/IEC 14496-10/Amd.3 Scalable Video Coding, ITU-T SG16 Q.6, JVT-X201, 2007.
[2] R. Leonardi, T. Oelbaum, J.-R. Ohm, Status report on wavelet video coding exploration, ISO/IEC JTC1/SC29/WG11 MPEG, N8043, 2006.
[3] J.M. Shapiro, An embedded wavelet hierarchical image coder, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), San Francisco, CA, 1992, pp. 657–660.
[4] A. Said, W.A. Pearlman, A new fast and efficient image codec based on set partitioning in hierarchical trees, IEEE Trans. Circuits Syst. Video Technol. 6 (June) (1996) 243–250.
[5] D. Taubman, High performance scalable image compression with EBCOT, IEEE Trans. Image Process. 9 (7) (2000) 1158–1170.
[6] D. Taubman, M.W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, Kluwer Academic Publishers, Boston, 2002.
[7] J.-R. Ohm, Three-dimensional subband coding with motion compensation, IEEE Trans. Image Process. 3 (5) (1994) 559–571.
[8] S.-T. Hsiang, J.W. Woods, Embedded video coding using invertible motion compensated 3-D subband wavelet filter bank, Signal Process. Image Commun. 16 (May) (2001) 705–724.
[9] A. Secker, D. Taubman, Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression, IEEE Trans. Image Process. 12 (12) (2003).
[10] J. Xu, R. Xiong, B. Feng, G. Sullivan, M.C. Lee, F. Wu, S. Li, 3D Subband Video Coding Using Barbell Lifting, ISO/IEC JTC1/SC29/WG11 MPEG, M10569, 2004.
[11] J. Xu, Z. Xiong, S. Li, Y.-Q. Zhang, 3-D embedded subband coding with optimal truncation (3-D ESCOT), J. Appl. Comput. Harmonic Anal. 10 (May) (2001) 290–315.
[12] W.-H Peng, C.-Y. Tsai, T. Chiang, H.-M. Hang, Knowledge-Based Intelligent Information and Engineering Systems Berlin, Advances of MPEG scalable video coding standards, vol. 3684, Springer, Germany, 2005, pp.889–895 (Chapter. 3).
[13] M. Wang, M. van der Schaar, Operational rate-distortion modeling for wavelet video coders, IEEE Trans. Signal Processing 54 (9) (2006) 3505–3517.
[14] B. Girod, The efficiency of motion-compensating prediction for hybrid coding of video sequences, IEEE J. Sel. Areas Commun. SAC-5 (August) (1987) 1140–1154.
[15] B. Girod, Rate-constrained motion estimation, in: Proc. Int. Symp. Visual Commun. Image Processing, 1994, pp. 1026–1034.
[16] M.C. Chen, A.N. Willson Jr., Rate-distortion optimal motion estimation algorithms for motion-compensated transform video coding, IEEE Trans. Circuits Syst. Video Technol. 8 (2) (1998) 147–157.
[17] C.-Y. Tsai, H.-M Hang, Rate-distortion model for motion prediction efficiency in scalable wavelet video coding, Packet Video Workshop, 2009.
[18] T. Berger, Rate Distortion Theory, Prentice Hall, Englewood Cliffs, NJ, 1984.
[19] J. Ribas-Corbera, S. Lei, Rate control in DCT video coding for lowe-delay communications, IEEE Trans. Circuits Syst. Video Technol. 9 (February) (1999) 172–185.
[20] Z. He, S.K. Mitra, Optimum bit allocation and accurate rate control for video coding via rho-domain source modeling, IEEE Trans. Circuits Syst. Video Technol. 12 (10) (2002) 840–849.
[21] E.Y. Lam, J.W. Goodman, A mathematical analysis of the DCT coefficient distributions for images, IEEE Trans. Image Process. 9 (October) (2000) 1661–1666.

[22] G.J. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, IEEE Signal Processing Mag. 15 (November) (1998) 74–90.
[23] T. Wiegand, B. Girod, Lagrangian multiplier selection in hybrid video coder control, in: Proc. ICIP 2001, Thessaloniki, Greece, 2001.
[24] T. Wiegand et al., Rate-constrained coder control and comparison of video coding standards, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 688–703.
[25] R. Xiong, X. Ji, D. Zhang, J. Xu, Vidwav wavelet video coding specifications, ISO/IEC JTC1/SC29/WG11 MPEG, M12339, 2005.
[26] S.S. Tsai, H.-M. Hang, Motion information scalability for MC-EZBC, Signal Process. Image Commun. 19 (7) (2004) 675–684.
[27] A. Secker, D. Taubman, Highly scalable video compression with scalable motion coding, IEEE Trans. Image Process. 13 (8) (2004) 1029–1041.
[28] C.-Y. Tsai, H.-M. Hang, ρ-GGD source modeling for wavelet coefficients in image/video coding, in: Proc. IEEE Int. Conf. Multimedia and Expo (ICME), Hannover, Germany, 2008, pp. 601–604.
[29] T. Rusert, K. Hanke, J. Ohm, Transition filtering and optimized quantization in interframe wavelet video coding, in: Proc. SPIE Visual Communications and Image Processing (VCIP), vol. 5150, 2003, pp. 682–693.
[30] M.S. Crouse, R.D. Nowak, R.G. Baraniuk, Wavelet-based statistical signal processing using hidden markov models, IEEE Trans. Signal Process. 46 (4) (1998) 886–902.
[31] J. Xu, R. Xiong, B. Feng, G. Sullivan, M.C. Lee, F. Wu, S. Li, 3D subband video coding using barbell lifting, ISO/IEC JTC1/SC29/WG11 MPEG, M10569, 2004.
[32] R. Xiong, J. Xu, F. Wu, Coding performance comparison between MSRA wavelet video coding and JSVM, ISO/IEC JTC1/SC29/WG11 MPEG, M11975, 2005.