

Wyner-Ziv Video Coding with Multi-resolution Motion Refinement: Theoretical Analysis and Practical Significance

Wei Liu, Lina Dong, Wenjun Zeng
 Computer Science Department, University of Missouri, Columbia, MO 65211, USA

ABSTRACT

The accuracy of motion estimation (ME) plays an important role in improving the coding efficiency of Wyner-Ziv video coding (WZVC). Most existing WZVC schemes perform ME at the decoder. The unavailability of the current frame on the decoder side usually impairs the accuracy of ME, which also causes the degradation of coding efficiency of WZVC. To improve the accuracy of ME, some works in the literature assume the current frame can be progressively decoded, and the decoder iteratively refines the motion field based on each partially-decoded image. In this paper, we present an analytical model to estimate the potential gain by employing multi-resolution motion refinement (MMR), assuming the current frame is progressively decoded in the frequency domain. The theoretical results show that at high rates, WZVC with MMR falls about 1.5 dB behind the conventional inter-frame coding, but outperforms WZVC with motion extrapolation by 0.9 to 5 dB. Significant gain has also been observed in the simulations using real video data.

Keywords: Wyner-Ziv video coding, motion estimation, multi-resolution motion refinement

1. INTRODUCTION

In the last twenty years, video coding has achieved a tremendous success using the hybrid coding structure: inter-frame motion compensated prediction (MCP) plus intra-frame transform coding. Such a structure relies on a powerful encoder that can afford the heavy computational burden, and the decoders are relatively lightweight, working in a slave mode. In recent years, with the rapid development of wireless sensor networks (WSN), Wyner-Ziv video coding (WZVC) has become a hot research topic. WZVC assumes there are multiple encoders with limited processing capability and limited power supply, while the decoder is a powerful station. Therefore the computational burden needs to be shifted to the decoder. WZVC provides a useful alternative to the conventional hybrid coding, especially in the so-called “uplink” video transmission applications.

It has been widely known that the most computational-intensive job in decorrelating a video source is the inter-frame motion estimation (ME). Thus most existing WZVC schemes perform ME at the decoder. A significant difference between decoder-side ME and the conventional ME is that the decoder does not have access to the current frame. This hurts the accuracy of the estimated motion vectors (MV), and consequently, the coding efficiency of WZVC.

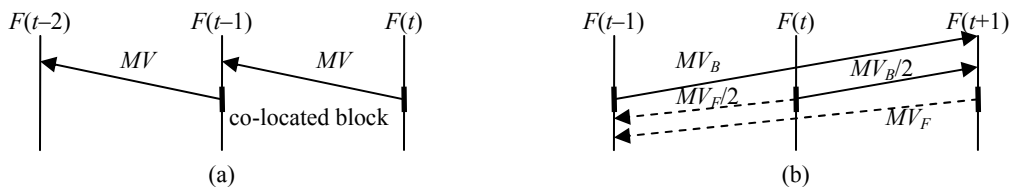


Fig. 1. 1-D illustration of (a) motion extrapolation and (b) motion interpolation at the decoder side

In [1], several approaches for decoder-side ME are reviewed and their performances are compared. One typical approach of motion extrapolation is to perform forward ME from $F(t-2)$ to $F(t-1)$; for a block in $F(t)$ ¹, the same MV of the co-located block in $F(t-1)$ is used to find its motion-compensated correspondence. A 1-D example of this approach is illustrated in Fig. 1(a). As for motion interpolation, a typical approach is illustrated in Fig. 1(b): the decoder finds the forward motion vector MV_F of the co-located block in $F(t+1)$ from $F(t-1)$, as well as the backward motion vector MV_B of

¹ Since most of our discussions are about the decoder, we simply use $F(t)$ to denote the decoder’s observation of the t -th frame, as opposed to the original, uncoded frame.

the co-located block in $F(t-1)$ from $F(t+1)$ then uses $MV_F/2$ and $MV_B/2$ as the forward / backward MV's of the current block, respectively.

Without any knowledge about the current frame, motion extrapolation / interpolation approaches basically assume that objects move at a constant speed and the estimated MV's reflect true motion, which is an over-simplification to the reality. As a result, the estimated side information (SI) is usually not a good approximation to the current frame. Analytical results in [2] suggest that WZVC could fall 6dB or more behind conventional video coding in the worst case due to the inaccurate MCP. Realizing this, some other works propose to use hash codes [3] or CRC codes [4] to facilitate the motion search. These approaches can be considered as having partial access to the current frame.

Approaches involving iterative motion refinement can be found independently in [5][6][7]. For example, in [6], the current frame is decoded progressively from the most-significant bit-plane to the least-significant bit-plane. Once a bit-plane is decoded, the motion search is refined to find better SI. Intuitively, having partial access to the current frame could improve the accuracy of ME, but not always, especially when the quality of the partially decoded frame is extremely low. Unfortunately, there have not been any analytical results to guide the design of such schemes. To the best of our knowledge, this paper is the first attempt.

In this paper, we consider the multi-resolution motion refinement (MMR): assume the current frame is progressively decoded in the frequency domain, from low-frequency components to high-frequency components. Once the decoder reconstructs a low-resolution version of the current frame, it uses this version for the MCP and reconstructs a higher-resolution image. The process iterates until the full resolution image is reconstructed. We provide both theoretical analysis and experimental results to show that the proposed architecture can significantly improve the MCP accuracy and, consequently, the coding efficiency of WZVC.

The rest of the paper is organized as follows. We first present a rate-distortion (R-D) analysis of this approach in Section 2. Then a case study of WZVC in the wavelet domain is carried out in Section 3. Section 4 concludes the paper.

2. PERFORMANCE ANALYSIS OF WZVC WITH MMR

As pointed out in [2], WZVC suffers from three types of performance loss in comparison with conventional hybrid video coding. The first is *system loss*, which is due to the rate loss in Wyner-Ziv coding. This loss is nonzero unless the sources are jointly-Gaussian and the distortion is measured by the mean-square-error (MSE). The second is *source coding loss*, due to the gap between the performance of an error correction code and the Shannon bound. The third is *video coding loss*, which is due to the imperfect MCP at the decoder side. The first two types of performance loss seem to be unavoidable in a distributed source coding system. But the MCP accuracy and, consequently, the coding efficiency can be potentially improved by having partial access to the current frame, especially in the areas with complex motion.

2.1 Efficiency of MCP coding

Theoretical analysis presented in [8] suggests that at high rates, the rate saving of MCP coding over intra-frame coding is

$$\Delta R = R_{MCP} - R_{intra} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - |P(\omega)|^2 \right] d\omega \quad (1)$$

where $\omega = (\omega_x, \omega_y)^T$ is the vector of spatial frequencies in radius, $P(\omega)$ is the Fourier transform of the probability density function (p.d.f.) of the ‘‘motion vector displacement’’ $p(\Delta d)$, where $\Delta d = (d_x, d_y)$ stands for the error between the estimated MV and the true MV. For example, in integer-pel-accuracy motion search, a coarse model of the displacement is uniform distribution within the range of $[-1/2, 1/2] \times [-1/2, 1/2]$. In reality, the saved rate depends more on the variance of Δd rather than the exact shape of the p.d.f.. In the literature, a zero-mean isotropic Gaussian distribution of Δd is usually assumed:

$$p(\Delta d) = \frac{1}{2\pi\sigma_{\Delta d}^2} \exp\left(-\frac{d_x^2 + d_y^2}{2\sigma_{\Delta d}^2}\right) \quad (2)$$

and (1) can be rewritten accordingly as

$$\Delta R = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp(-\omega^T \omega \sigma_{\Delta d}^2) \right] d\omega \quad (3)$$

Eq. (3) suggests that the coding gain of MCP depends exclusively on $\sigma_{\Delta d}^2$, the accuracy of motion estimation. This conclusion applies to both inter-frame video coding and WZVC.

In [2] the accuracy of motion extrapolation is estimated as

$$\sigma_{\Delta d}^2 = (1 - \rho^2) \sigma_d^2 \quad (4)$$

where ρ is the correlation between the motion fields in adjacent frames, and σ_d^2 is the variance of the motion vectors. Substituting (4) into (3) we get

$$\Delta R_{ext} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp(-\omega^T \omega (1 - \rho^2) \sigma_d^2) \right] d\omega \quad (5)$$

It is clear that the accuracy of the motion extrapolation is essentially content-dependent, and it becomes less efficient when the temporal correlation between the motion fields decreases. Such scenarios might include stereo image coding, where there is no previously-estimated “motion” available. In this case $\rho=0$ in (4), and only the reference frame itself can be used as the SI. In this case the performance of motion-extrapolation degrades to that of DPCM coding (coding of the frame difference without motion compensation). As another example, in real-time video communication, frame dropping is unavoidable. Motion extrapolation can only be applied to the nearest available frame. For a rough estimate, if the reference motion field is t frames away from the current one, the correlation becomes ρ^t in (4), which has significant impact on the accuracy of the estimated MV's.

The accuracy of motion search, however, is not related to the temporal correlation of the motion field. Rather, it depends on the motion search method itself, for example, block size used for matching, integer- or fractional-pel accuracy in the search, etc. Buschmann [9] proposes an excellent model in estimating $\sigma_{\Delta d}^2$. In the next subsection, the model is simplified and $\sigma_{\Delta d}^2$ is expressed as a function of block size and pixel accuracy.

2.2 Motion search accuracy using block matching

It is suggested in [9] that a block-matching motion search method can be described analytically by low-pass estimation filtering, sampling and quantization of an exact motion field, combined with subsequent reconstruction filtering. The estimation filtering is introduced because a local neighborhood is usually used for block-matching. The larger the block size is, the smaller the bandwidth of the low-pass filter is. Sampling and quantization operations account for the fact that the estimated motion field has limited spatial resolution and amplitude precision. The sampling rate is also determined by the block size, and the quantization depends on whether or not fractional-pel accuracy motion search is used. The reconstruction filter is also a low-pass filter, representing the spatial interpolation (e.g. nearest-neighbor interpolation) of the estimated motion field.

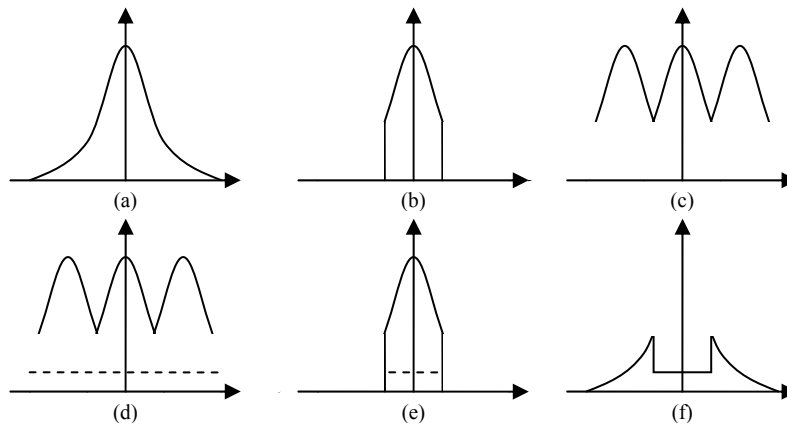


Fig. 2. A 1-D illustration of introducing motion vector displacement: (a) PSD of the original motion field; (b) PSD after estimation filtering; (c) PSD after sampling; (d) PSD after quantization; (e) PSD after reconstruction filter; and (f) PSD of the motion vector displacement.

We approximate the above operations in the frequency domain and illustrate the 1-D power spectrum density (PSD) of the motion field in Fig. 2. Here we assume the estimation filter and the reconstruction filter are both ideal low-pass filters, and the quantization step size is small. In this case we can see that the effects of sampling and reconstruction filtering are cancelled out by each other, and the PSD of Δd basically consists of two parts: inside the pass-band of the estimation filter it is white quantization noise, and outside the pass-band it equals the PSD of the motion field. Thus $\sigma_{\Delta d}^2$ can be calculated as:

$$\sigma_{\Delta d}^2 = \sigma_{\Delta d-lp}^2 + \sigma_{\Delta d-q}^2 \quad (6)$$

with

$$\sigma_{\Delta d-lp}^2 = \frac{1}{4\pi^2} \iint_{\Lambda(\omega_b)} S_{dd}(\omega) d\omega, \quad \sigma_{\Delta d-q}^2 = \frac{1}{4\pi^2} \int_{-\omega_b}^{\omega_b} \int_{-\omega_b}^{\omega_b} \frac{q^2}{12} d\omega = q^2 \omega_b^2 / 12\pi^2 \quad (7)$$

where $\sigma_{\Delta d-lp}^2$ and $\sigma_{\Delta d-q}^2$ denote the contribution to the motion vector displacement due to low-pass filtering and quantization, respectively, S_{dd} is the PSD of the motion field, q is the quantization step-size (1 for integer-pel accuracy search, etc.), ω_b is the cut-off frequency of the estimation filter (if B is the 1-D block size used for matching, we roughly have $\omega_b = \pi/B$), and $\Lambda(\omega_b)$ describes the stop-band of the estimation filter:

$$\Lambda(\omega_b) = \left\{ (\omega_x, \omega_y) : \max(|\omega_x|, |\omega_y|) > \omega_b \right\} \quad (8)$$

The autocorrelation function of the motion field is characterized as isotropic and exponentially decreasing in [9]. And its corresponding PSD is written as

$$S_{dd}(\omega) = \frac{2\pi\omega_0\sigma_d^2}{(\omega_0^2 + \omega_x^2 + \omega_y^2)^{-3/2}} \quad (9)$$

where ω_0 is a small constant. For relatively small block sizes, we have $\omega_b \gg \omega_0$, thus in the stop-band of the low-pass filter, the term ω_0^2 in the denominator of $S_{dd}(\omega)$ is omissible. In this case $\sigma_{\Delta d-lp}^2$ can be approximated as

$$\sigma_{\Delta d-lp}^2 \approx \frac{1}{4\pi^2} \iint_{\Lambda(\omega_b)} \frac{2\pi\omega_0\sigma_d^2}{(\omega_x^2 + \omega_y^2)^{-3/2}} d\omega = \frac{2\sqrt{2}\omega_0\sigma_d^2}{\pi^2} B \quad (10)$$

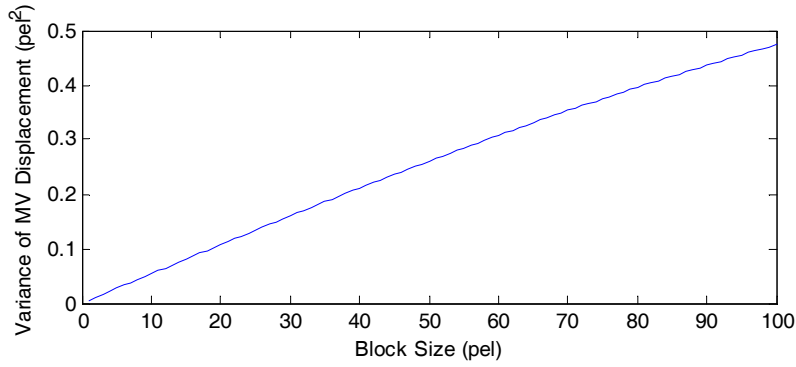


Fig. 3. Variance of motion vector displacement introduced by the estimation filter.

Eq. (10) suggests that for relatively small block sizes, the motion vector displacement introduced by low-pass filtering increases almost linearly with respect to the block size. In Fig. 3, we plot the exact value of $\sigma_{\Delta d-lp}^2$ when B varies from 2 to 100. It can be seen that this “near linearity” property holds for a large range of block sizes.

For the sake of simplicity, denote the factor $2\sqrt{2}\omega_0/\pi^2$ in (10) as k , also noticing that the error due to low-pass filtering and quantization will never exceed σ_d^2 , the overall motion vector displacement is modeled in a simplified function as

$$\sigma_{\Delta d}^2 = \min(q^2/12B^2 + k\sigma_d^2 \times B, \sigma_d^2) \quad (11)$$

For QCIF sequences, we can use $k \approx 0.005 \text{ pel}^{-1}$. The numerical output of (11) fits well with the result in [9], thus we treat it as a good approximation of the reality. The difference between (11) and the model used in [2] is that (11) takes into account the impact of block size and σ_d^2 . Using this result, for inter-frame coding, the rate saving is calculated as

$$\Delta R_{int} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp\left(-\omega^T \omega \left(q^2/12B^2 + k\sigma_d^2 \times B \right)\right) \right] d\omega. \quad (12)$$

2.3 Efficiency of WZVC using MMR

As for the MMR, we consider an infinitely large frame which is sampled at the integer grid. The discrete time Fourier transform (DTFT) of the frame creates a periodic frequency spectrum. Due to the periodicity, we only have to consider the spectrum in one period $|\omega_x| < \pi, |\omega_y| < \pi$. Let $0 = \omega_0 < \omega_1 < \dots < \omega_N = \pi$ be a series of frequency points, and we construct N ideal band-pass filters so that the n^{th} filter encompasses the frequency band $\{\Lambda(\omega_{n-1}) \setminus \Lambda(\omega_n)\}$ for $n = 1, \dots, N$ (“\” denotes the set difference). If we apply the N filters to the original image, we get N band-pass images without overlapping with each other in the frequency domain. The rate to transmit the whole image equals the total rate to transmit the N band-pass images. Suppose at time n , the decoder has received the first n band-pass images, based on which a low-pass version of the frame can be reconstructed. Now, by assuming all of the band-pass images share the same motion field, the decoder can use this partially reconstructed frame to get a refined estimation of the motion field and do the MCP for the $(n+1)^{\text{th}}$ band-pass image. We are interested in how much we can benefit from this MMR.

According to the Nyquist sampling theorem, since the currently-available frequency band is $|\omega_x| < \omega_n, |\omega_y| < \omega_n$, the partially reconstructed image can be sampled at the rate of ω_n/π . This indicates that the block size used for matching should be ω_n/π times larger than that in the full-resolution image space to contain enough content for matching. From the result in (11) we can model the motion search accuracy in the n^{th} level as

$$\sigma_{\Delta d}^2(n) = \min\left(q^2 \omega_n^2 / 12\pi^2 B^2 + k\pi B \sigma_d^2 / \omega_n, \sigma_d^2\right). \quad (13)$$

Then the rate saving we can achieve in the $(n+1)^{\text{th}}$ band-pass image is

$$\Delta R_{n+1} = \frac{1}{8\pi^2} \iint_{\Lambda(\omega_n) \setminus \Lambda(\omega_{n+1})} \log_2 \left[1 - \exp\left(-\omega^T \omega \times \min\left(q^2 \omega_n^2 / 12\pi^2 B^2 + k\pi B \sigma_d^2 / \omega_n, \sigma_d^2\right)\right) \right] d\omega. \quad (14)$$

If ω_n and ω_{n+1} are close enough, we have $\omega_n \approx \max(|\omega_x|, |\omega_y|)$ in the $(n+1)^{\text{th}}$ frequency band. This approximation makes the term to be integrated in (14) independent of n . Thus we can sum up the rate saving in different frequency bands and get the total rate saving

$$\Delta R_{MMR} = \sum_{n=1}^N \Delta R_n = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp\left(-\omega^T \omega \times \delta(\omega)\right) \right] d\omega. \quad (15)$$

where $\delta(\omega) = \min\left(q^2 \max(\omega_x^2, \omega_y^2) / 12\pi^2 B^2 + k\pi B \sigma_d^2 / \max(\omega_x, \omega_y), \sigma_d^2\right)$ denotes the frequency-related MV displacement.

An important observation of (15) is that the rate saving no longer depends on the temporal correlation of the motion field. Similar to (12), ΔR_{MMR} is only related to the spatial complexity of the motion field (σ_d^2) as well as the motion search method itself (q and B).

In Fig. 4, we assume the block size $B = 4$, and study the performance of ΔR_{MMR} with respect to σ_d^2 , when integer-, half- and quarter-pel accuracy search is used. A comparison is also made with ΔR_{inter} with integer-pel search. Not surprisingly, both of them become less significant when the spatial variance of the motion field gets larger. As for the integer-pel accuracy search, *the gap between MMR and inter-frame coding is almost a constant of 0.25 bpp*, regardless of the motion field complexity. The rate-distortion theory also suggests that the R-D curves have a slope of 6.02 dB/bit for high-rate coding. Thus our result indicates that the PSNR performance of WZVC with hierarchical SI estimation falls about 1.5 dB behind the conventional inter-frame coding. The gain of using fractional-pel accuracy search in MMR is not as significant as that in inter-frame coding – about 0.05 bpp when $\sigma_d^2 = 1$, and diminishing as σ_d^2 increases.

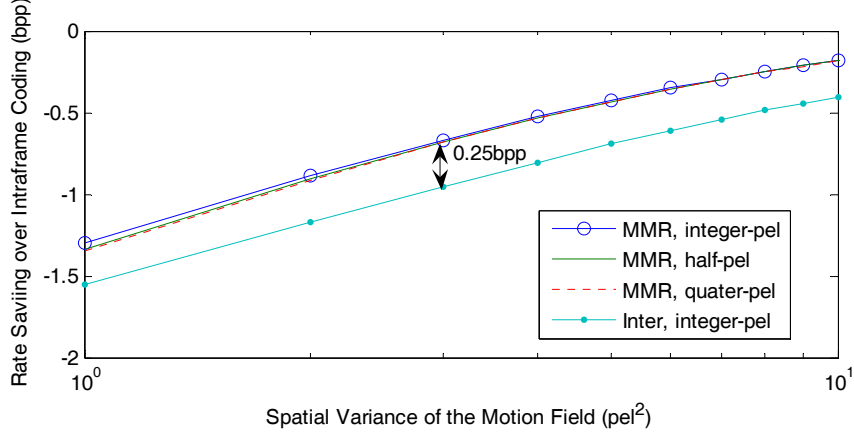


Fig. 4. Comparison of the rate saving by using inter frame coding and WZVC with MMR

It is observed that when there is very little information available in the partially reconstructed frames, the accuracy of the motion search degrades so that at some point, the temporal extrapolation of the motion field might be more accurate. Thus a smart decoder should be able to switch the prediction between MMR and motion extrapolation. For the $(n+1)^{\text{th}}$ frequency band $\{\Lambda(\omega_n) \setminus \Lambda(\omega_{n+1})\}$, MMR should be used if $\sigma_{\Delta d}^2(n) < (1 - \rho^2)\sigma_d^2$, or

$$k\pi B\sigma_d^2/\omega_n < (1 - \rho^2)\sigma_d^2. \quad (16)$$

Here we omit $\sigma_{\Delta d-q}^2$ because motion extrapolation also suffers from quantization error. We can imagine there is a “critical frequency”

$$\omega^* = k\pi B/(1 - \rho^2), \quad (17)$$

below which the motion extrapolation is used; otherwise MMR is used. The critical frequency is high, if the temporal correlation ρ of the motion field is high. Consider the specific settings with $\rho=0.95$, $B=4$, the resulting $\omega^* < \pi/4$, meaning that it is usually safe to perform MMR for a quarter-resolution images even for large ρ values.

The rate saving over motion extrapolation is obtained in the frequency area $\{\Lambda(\omega^*) \setminus \Lambda(2\pi)\}$:

$$\begin{aligned} \Delta R_{MMR+ext} = & \frac{1}{8\pi^2} \iint_{\Lambda(\omega^*) \setminus \Lambda(2\pi)} \log_2 [1 - \exp(-\omega^T \omega \times \delta(\omega))] d\omega \\ & + \frac{1}{8\pi^2} \int_{-\omega^*}^{\omega^*} \int_{-\omega^*}^{\omega^*} \log_2 [1 - \exp(-\omega^T \omega (1 - \rho^2)\sigma_d^2)] d\omega \end{aligned} \quad (18)$$

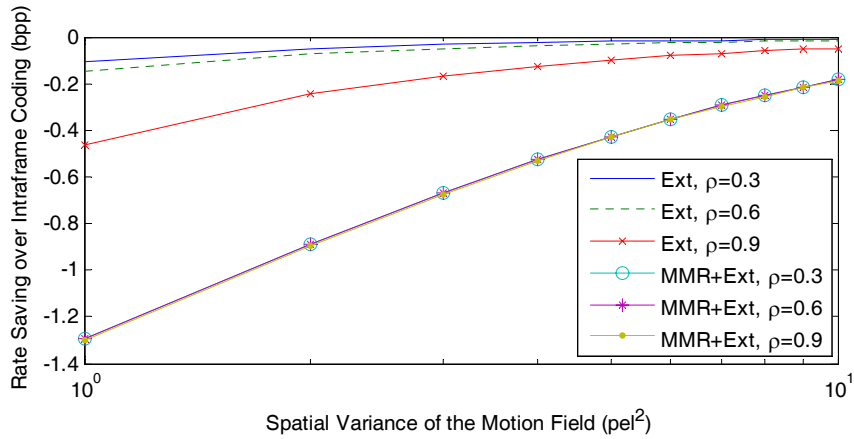


Fig. 5. Comparison of WZVC with motion extrapolation and WZVC with MMR and motion extrapolation

We plot in Fig. 5 the rate saving curves with respect to different motion field complexity σ_d^2 as well as the temporal correlation ρ . It is assumed that the block size is 4 and integer-pel accuracy search is used. We can see that the MMR combined with motion extrapolation method is insensitive to the temporal correlation ρ , because for most ρ values, the critical frequency ω^* is so small that most of the coding gain is obtained by using MMR. However, that is not the case for pure motion extrapolation. For medium or low temporal correlation values, the rate saving by using motion extrapolation could be poor (around 0.1 bpp). Even for high temporal correlation values, there is still a gap of 0.15 to 0.83 bpp away from MMR with motion extrapolation, resulting in 0.9 to 5 dB loss in the PSNR performance.

Finally we study the performance of MMR with multiple-hypothesis prediction (MHP). Using the results in [10], the rate saving using MHP is

$$\Delta R_{MHP} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 [1 - \mathbf{P}^H \mathbf{D}^{-1} \mathbf{P}] d\omega. \quad (19)$$

where $\mathbf{P} = [P_1(\omega), P_2(\omega), \dots, P_M(\omega)]^T$, M is the number of the predictors, the superscript H stands for the transposed conjugate, and \mathbf{D} is defined as

$$\mathbf{D} = \begin{bmatrix} 1 & P_1 P_2^* & \dots & P_1 P_M^* \\ P_2 P_1^* & 1 & \dots & P_2 P_M^* \\ \vdots & \vdots & \ddots & \vdots \\ P_M P_1^* & P_M P_2^* & \dots & 1 \end{bmatrix}. \quad (20)$$

If we assume that the variance of motion vector displacement is the same for each predictor, and the displacements of any two predictors are mutually independent of each other, (19) can be simplified to

$$\Delta R_{MHP} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \frac{M \exp(\omega^T \omega \sigma_{\Delta d}^2)}{1 + (M-1) \exp(\omega^T \omega \sigma_{\Delta d}^2)} \right] d\omega. \quad (21)$$

If $\sigma_{\Delta d}^2$ is replaced with $\delta(\omega)$ we can get the efficiency of MMR with MHP. In Fig. 6 we illustrate the rate saving by using different number of reference frames, where integer-pel accuracy search and $B = 4$ for the block size is assumed. We can see that whenever the number of hypotheses doubles, the rate saving is increased by 0.15 to 0.5 bpp, resulting in the coding gain in PSNR of around 0.9 to 3.0 dB. It is worth noting that in the above analysis, we assume the reference frames are all noise-free, which is usually not the case in reality. Thus in practice the PSNR gain cannot be arbitrarily large even if there are infinite number of reference frames.

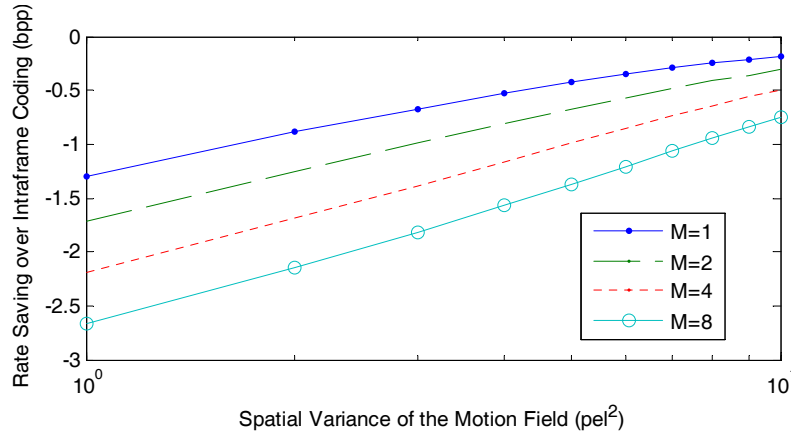


Fig. 6. Performance of WZVC with MMR using multiple hypothesis prediction

3. CASE STUDY: WAVELET DOMAIN WZVC WITH MMR

In this section, we describe a wavelet domain WZVC algorithm with MMR at the decoder. The diagram of this approach is illustrated in Fig. 7. Although in the following description only the previously-decoded frame is used for MCP, it can

be easily extended to multiple hypothesis MCP, by performing the same operation on other reference frames and calculating the average (or weighted average) of the results.

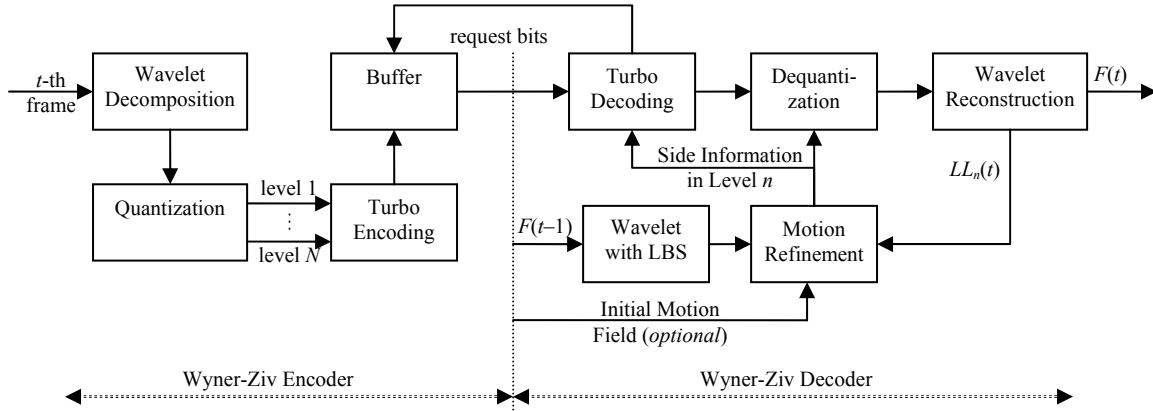


Fig. 7. Illustration of WZVC with multi-resolution motion refinement at the decoder

The encoder applies an N -level wavelet decomposition to each of the frames. For each frame, $(3N+1)$ subbands are produced, namely, HL_n , LH_n , HH_n , ($n = 1, 2, \dots, N$), and LL_N . For the sake of convenience, let LL_{n-1} denote the reconstructed subband from LL_n , HL_n , LH_n and HH_n , ($n = 1, 2, \dots, N$). Each of these subbands is Wyner-Ziv encoded and transmitted to the decoder at the necessary rate. To distinguish the subbands in different frames, we use notations such as $LL_n(t)$, where t is the frame index.

At the decoder side, supposed $F(t-1)$ is available, the decoding process of $F(t)$ can be described as follows:

- (1) Estimate $LL_N(t)$ from $LL_N(t-1)$ by applying MCP based on an initial motion field. The initial motion field could be the same one as $F(t-1)$, or could be set to zero if it is not available.
- (2) Reconstruct $LL_N(t)$ by applying Wyner-Ziv decoding.
- (3) Set $n = N$.
- (4) Refine the motion field by performing ME between $LL_n(t)$ and $LL_n(t-1)$.
- (5) Predict $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$ by copying the corresponding motion-compensated coefficients in $HL_n(t-1)$, $LH_n(t-1)$ and $HH_n(t-1)$.
- (6) Apply Wyner-Ziv decoding to reconstruct $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$.
- (7) Reconstruct $LL_{n-1}(t)$ based on $LL_n(t)$, $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$.
- (8) Reduce n by 1. If n reaches zero, the decoding is finished; otherwise go to step (4).

To overcome the shift-variance problem in the critically-sampled wavelet domain, $F(t-1)$ needs to be transformed to the over-complete wavelet domain. Therefore, the subbands of $F(t-1)$ are all in an over-complete form, derived from the so-called low-band-shift (LBS) decomposition [11]. Note that the difference between our approach and [11] is that we can only use the LL subbands for the ME. This approach significantly improves the efficiency of MCP, at the cost of extra memory consumption in the decoder. However this strategy still fits the “simple-encoding, complex-decoding” principle of WZVC.

In the following simulations, the QCIF sequences Carphone, Foreman, Miss America and Suzie @15fps are used for testing. The encoder decomposes each frame into 3 levels using the Daubechies 9/7 bi-orthogonal filter bank. Referred to the discussions about the “critical frequency”, 3 levels of decomposition is usually “safe” for most QCIF sequences, and more decomposition levels only lead to marginal gain. At the decoder, overlapped block matching with full search is used, and the block size is chosen to be 8×8 . The approaches shown in Fig. 1 are treated as benchmarks. And in the MMR approaches, the LL_3 subbands of the SI are simply copied from the benchmark results.

3.1 MSE reduction and high-rate performance

In the first simulation, we choose small quantization step sizes such that the frames are coded at very high quality (PSNR around 60dB). Thus the prediction residual is almost purely due to motion mismatch (e.g. no quantization error propagation). We are going to show that MMR at the decoder provides significantly better performance by reducing the MSE of the estimated SI.

Table. 1. MSE reduction and corresponding bit-rate saving by using MMR.

			P Frames			B Frames		
			Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Car-phone	MSE	Benchmark	49.5	54.8	45.8	24.2	29.2	40.7
		Proposed	15.1	19.4	24.3	10.9	13.6	16.1
	Saved rates (bpp)		0.86	0.75	0.46	0.58	0.55	0.67
			Overall: 0.80			Overall: 0.57		
Foreman	MSE	Benchmark	38.8	64.4	58.0	19.7	34.9	37.1
		Proposed	16.0	25.7	31.3	10.6	18.2	24.4
	Saved rates (bpp)		0.64	0.66	0.44	0.45	0.47	0.30
			Overall: 0.62			Overall: 0.44		
Miss America	MSE	Benchmark	4.31	5.03	2.86	2.06	2.29	1.22
		Proposed	2.27	2.56	1.64	1.36	1.30	0.66
	Saved rates (bpp)		0.46	0.49	0.40	0.30	0.41	0.44
			Overall: 0.46			Overall: 0.32		
Suzie	MSE	Benchmark	13.1	16.8	16.9	7.33	8.95	9.49
		Proposed	6.6	7.4	11.2	4.22	4.40	6.64
	Saved rates (bpp)		0.49	0.59	0.30	0.40	0.51	0.26
			Overall: 0.50			Overall: 0.41		

Since the residual frame is not a white signal, we would rather compare the energy distribution of the residual frame in each subband. In Table 1 we actually assume the wavelet coefficients are i.i.d. in the subbands HL_n , LH_n and HH_n , thus only the MSE in each level is shown. Under this assumption, the MSE values provide a reasonable estimate on the necessary coding rate for each level according to the rate-distortion theory. For example, reducing the MSE by half roughly means 0.5bpp saving in high bit-rates. The overall bit-rate saving is calculated as the weighted average of those in different subbands (the weighting factors are 3/4, 3/16 and 3/64 for level 1, 2 and 3, respectively). From Table 1 we can see that, for sequences @15fps, motion extrapolation / interpolation is usually not an efficient way to estimate the motion, due to the reduced temporal domain motion correlation. For those sequences listed in Table 1, 0.46 to 0.80 bpp rate saving can be obtained for P frames (which translates into as much as 5 dB PSNR gain at high rates); and the rate saving is 0.32 to 0.57 bpp for B frames.

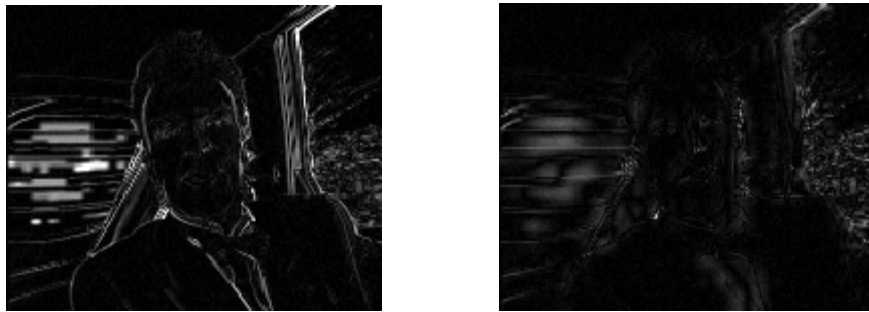


Fig. 8. Residual frame of the side information using bidirectional motion interpolation (left) and bidirectional MMR (right).

Sample images of the residual frame are shown in Fig. 8. Significant improvement is observed.

3.2 Rate-distortion performance in WZVC

In the second simulation, we observe the general R-D performance of WZVC using MMR and compare it to other coding paradigms. The odd-indexed frames are encoded using JPEG 2000, and the even-indexed frames are encoded as WZ frames. To provide a better visualization for the improvement, only the R-D curves of the WZ frames are given below.

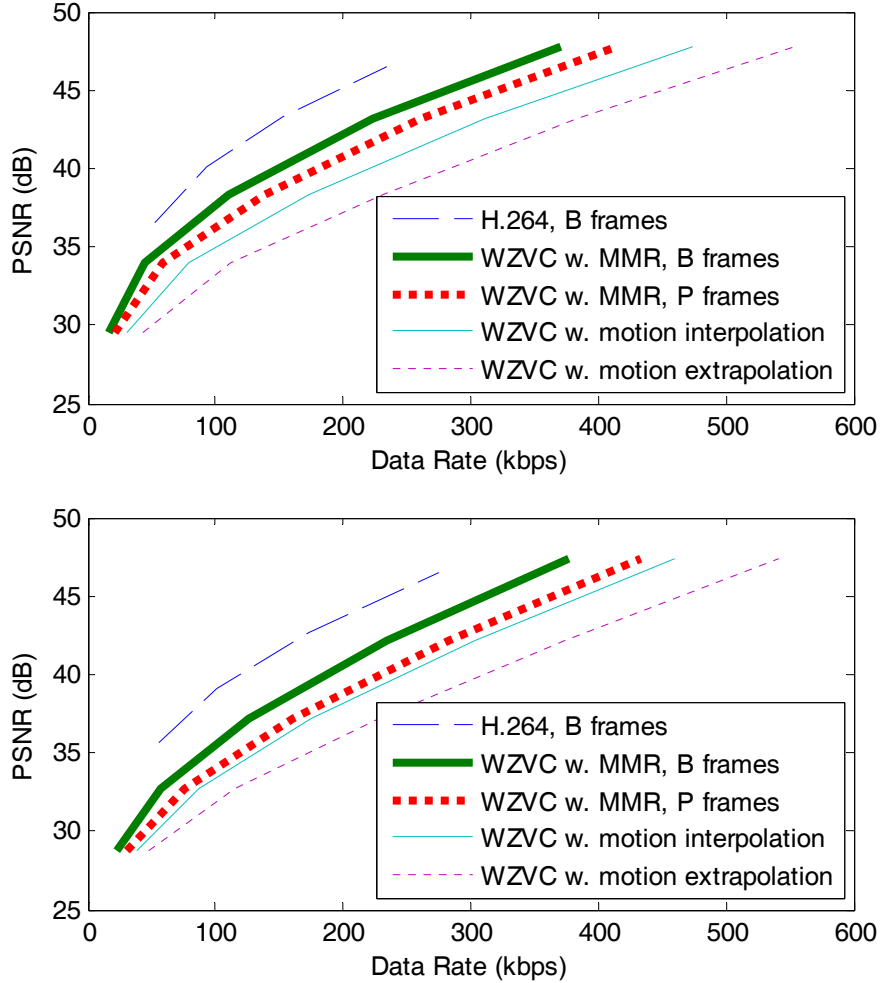


Fig. 9. R-D performance for the QCIF Foreman sequence (upper) and the QCIF Carphone sequence (lower), WZ frames only.

The results for the Foreman sequence and the Carphone sequence are illustrated in Fig. 9. We can see that for P frames, WZVC with MMR outperforms motion extrapolation by 3 to 4 dB at relatively high rates, which is consistent with the analytical result shown in Fig. 5 for medium spatial motion complexity and low temporal motion correlation. We also notice that when the number of reference frames increases from 1 to 2 (P frames to B frames), an additional coding gain of about 1 to 1.5 dB is obtained, which is also predicted by Fig. 6. Finally, WZVC with MMR still falls about 2.5 dB behind H.264 B frames, which is 1 dB higher than the rate loss we predicted in Fig. 4. That is probably due to the system loss and the source coding loss mentioned in [2]. Better correlation modeling for WZVC with MMR may help to mitigate the loss. It is also noted that in practical applications we can only perform limited levels of motion refinement.

4. CONCLUSION

The bottleneck in improving the coding efficiency of WZVC is the performance of ME at the decoder side. The ME accuracy can be improved if the decoder has partial access to the current frame. In this paper, we provide an analytical study on the rate-distortion performance of WZVC in a particular structure, where low-resolution images are iteratively decoded and used for the motion refinement. We show that in the ideal case, the performance gap between the conventional inter-frame coding and WZVC with MMR is about 1.5 dB, regardless of the temporal domain motion correlation and spatial domain motion complexity; and WZVC with MMR can outperform WZVC with motion extrapolation by 0.9 dB to 5 dB. A wavelet domain WZVC with MMR is also proposed based on the theoretical analysis. Significant improvement is observed.

ACKNOWLEDGEMENT

This research is supported in part by NSF grant CNS-0423386, and in part, by a grant from the University of Missouri System Research Board.

REFERENCES

- ¹ Z. Li and E. J. Delp, "Wyner-Ziv side estimator: conventional motion search methods revisited", *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sep. 2005.
- ² Z. Li, L. Liu and E. J. Delp, "Rate Distortion Analysis of Motion Side Estimation in Wyner-Ziv Video Coding", *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 98-113, Jan. 2007.
- ³ A. Aaron, S. Rane and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver", *Proc. IEEE Int. Conf. Image Processing*, Singapore, Oct. 2004.
- ⁴ R. Puri and K. Ramchandran, "Prism: An uplink-friendly multimedia coding paradigm", *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, Sep. 2003.
- ⁵ X. Artigas and L. Torres, "Iterative generation of motion-compensated side information for distributed video coding", *Proc. IEEE Int. Conf. Image Processing*, Genova, Italy, Sep. 2005.
- ⁶ J. Ascenso, C. Brites, and F. Pereira, "Motion compensated refinement for low complexity pixel based distributed video coding", *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance*, Como, Italy, Sep. 2005.
- ⁷ Z. Li, *New methods for motion estimation with applications to low complexity video compression*, Ph.D. dissertation, Purdue Univ., West Lafayette, IN, 2005.
- ⁸ B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas. Commun.*, vol. 5, no. 8, pp. 1140-1154, Aug. 1987.
- ⁹ R. Buschmann, "Efficiency of displacement estimation techniques", *Signal Processing: Image Communication*, vol. 10, pp. 43-61, 1997.
- ¹⁰ B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding", *IEEE Trans. Image Processing*, vol. 9, no. 2, pp. 173-183, Feb. 2000.
- ¹¹ H.-W. Park and H.-S. Kim, "Motion estimation using low-band-shift method for wavelet-based moving-picture coding", *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 577-587, 2000.