ABSTRACT
In Wyner-Ziv video coding system, motion estimation efficiency is much lower than that in conventional video coding system because current frame is not available when doing motion estimation. In this paper, we propose a successive resolution refinement algorithm to improve motion estimation efficiency. Based on our rate distortion analysis, we derive optimal downsample ratio for two-stage successive resolution refinement system. We also analyze the performance of multistage case, and find that it approaches the performance of ideal motion compensated Wyner-Ziv video coding system, with at most 2.17dB loss in PSNR. Experimental results demonstrate the correctness of the analysis and show that the proposed method out-performs original bit-plane refinement scheme up to 2.5dB, with much lower complexity.

1. INTRODUCTION
The basis of Wyner-Ziv video coding is a surprising conclusion proved by Slepian and Wolf in 1970s, that correlated sources can be encoded separately without losing coding efficiency. Wyner and Ziv extended this to lossy compression and proved the coding efficiency loss is zero in quadratic Gaussian case with MSE distortion measure[1].

Based on this, in Wyner-Ziv video coding, side information, i.e. predict frame, was no longer needed at encoder, so that motion estimation can be shift to decoder which form a low complexity encoder. But obviously it is difficult to perform motion search efficiently at decoder without knowing current frame. Initially it was proposed to use motion compensated interpolation(MCI) and extrapolation(MCE) to generate prediction frame based on neighboring reconstructed frames. But as analyzed in [2], although it achieved 6dB gain over intra frame encoder in PSNR, the performance is 6dB lower compared with using ideal motion compensation.

To conduct more accurate motion estimation as while keep low latency, joint decoding and motion estimation was proposed, for example the CRC check approach in [4], and the hash based approach [3]. Both of them provided better performance than MCI and MCE based Wyner-Ziv encoder. The idea of joint decoding and motion estimation was further developed recently in [5], which proposed to perform successive motion estimation via bitplane refinement. Up to 2/3dB gain were achieved over MCI/MCE based Wyner-Ziv video encoder respectively, with the cost of several times of motion estimation.

In this paper, we propose a novel successive refinement video coding scheme via resolution refinement. The rest of this paper is organized as follows. In section 2 our successive resolution refinement video coding scheme is introduced starting from a brief review of Wyner-Ziv successive refinement. Optimal downsample ratio and rate distortion performance are analyzed in section 3. Implementation detail and experimental results are given in section 4 followed by conclusion in section 5.

2. SUCCESSIVE REFINEMENT OF VIDEO

2.1. Wyner-Ziv Successive Refinement (WZSR)
Wyner-Ziv successive refinement [6], is to encode a source successively with corresponding side information at each refinement stage.

Fig 1 depicts a two-stage successive refinement system. Source $X$ is successively coded at a coarse stage and a refinement stage, where $\hat{X}_0$ and $\hat{X}_1$ are corresponding reconstruction. At coarse stage, only a coarse side information $Y_0$ is available to decoder, and at refinement stage a better side information $Y_1$ is available to decoder. We can strictly say $Y_1$ is better than $Y_0$, if $X \rightarrow Y_1 \rightarrow Y_0$ is a Markov chain. Denote $\vec{D} = (D_0, D_1)$ as distortion constraint for $X_0$ and $X_1$ respectively. If we have $D_1 \leq D_0$, then total rate $R^{SR}_{X_1Y_0,Y_1}(\vec{D})$
satisfies:

\[ R_{X|Y_1}(D_1) \leq R_{X|Y_0,Y_1}^S(D) \leq R_{X|Y_0}(D_1) \]  

where \( R_{X|Y_1}(D_1) \) denotes Wyner-Ziv rate using side information \( Y_1 \) at decoder.

### 2.2. Successive Refinement of Video

As is shown in Fig. 2, suppose \( Y_0 \) is side information generated at decoder without using any information of current frame \( X \), for example generated by MCI and MCE. Minimum rate for distortion \( D \) will be \( R_{X|Y_0}(D) \). Now suppose we encode source \( X \) in two stages: at first stage, \( X \) is encoded to meet distortion constraint \( D_0 \), using \( Y_0 \) as side information. Then the decoder estimate a new side information \( Y_1 \), with the help of the reconstruction \( \hat{X}_0 \). We can assert \( X \rightarrow Y_1 \rightarrow Y_0 \), because otherwise we can define a new random variable \( Y'_1 = (Y_0, Y_1) \) as new side information which definitely satisfy this Markov chain condition. At second stage, \( X \) is encoded to meet distortion constraint \( D \), using \( Y_1 \) as side information. From equation 1 we know that the rate for the new two-stage system will be lower than original Wyner-Ziv video coding system. And also it is not hard to understand that multiple stage successive refinement can perform better than two-stage.

Successive refinement of video can be implement by two approaches: The first is bitplane refinement which has been proposed in [5]. In this paper we propose a successive resolution refinement scheme, in which pixels of one frame were divided into several groups, and each group of pixels were encoded at one stage. Initially the first group of pixels are decoded by using initial side information, and then those decoded pixels are used in motion estimation to find better side information for remain pixels. Then second group of pixels are decoded by using the refined side information, and the decoder repeat motion estimation and Wyner-Ziv decoding alternatively until all pixels have been decoded.

#### 3. RATE DISTORTION ANALYSIS

In this section, we derive the rate distortion function for proposed WZSR, using residue variance rather than the PSD model used in [2], because we focus on pixel domain Wyner-Ziv coding which ignores intra correlation.

### 3.1. Two-stage Refinement and Optimal Down-sample Ratio

Suppose we encode \( m \) pixels at first stage with side information \( Y_0 \) at decoder, and then encode remain \( k - m \) pixels at second stage with refined side information \( Y_1 \), where \( k \) is total pixels number in one frame. Let \( \sigma_0^2 = E(X - Y_0)^2 \) and \( \sigma_1^2 = E(X - Y_1)^2 \) be the residue variance for coarse stage and fine stage side information respectively. Define down-sample ratio \( r \) as: \( r = \frac{m}{k} \), the bitrate (per pixel) will be:

\[ R_{SR}(D, r) = \frac{1}{2} \log_2 \left( \frac{\sigma_0^2}{D} \right) + \frac{1}{2} \left( 1 - \frac{1}{r} \right) \log_2 \left( \frac{\sigma_1^2}{D} \right) \]  

if residue’s distribution is assumed to be Gaussian. Compared with a one stage system using side information \( Y_0 \), the rate difference is

\[ \Delta R(r) = R_{SR}(D, r) - R_{WZ}(D) = \frac{1}{2} (1 - \frac{1}{r}) \log_2 \left( \frac{\sigma_1^2}{\sigma_0^2} \right) \]  

which is independent of \( D \) and is negative generally, since the residue variance will reduce after refinement, i.e. \( \sigma_1^2 < \sigma_0^2 \).

Because down-sample ratio \( r \) will affect motion estimation efficiency for second stage, \( \sigma_1^2 \) is a function of \( r \). We will prove elsewhere that \( \sigma_1^2 \) is approximately linear to \( \sqrt{r} \) and can be expressed as: \( \sigma_1^2 \approx A \sqrt{r} + N \), where the values of \( A \) and \( N \) depend on particular video sequence. Due to page restriction, here we only give in Fig.3 the average residue variance curve collected from 700 frames of 7 different 15Hz QCIF sequences. The top straight line is DPCM residue which can be considered as infinite down-sample. Notice that \( \sqrt{r} = 8 \) means only 1 pixel in \( 8 \times 8 \) block is used in motion estimation.
Defining $r_{\infty} = (\frac{N^2}{A})^{\frac{1}{2}}$ and $\rho = \frac{N}{m^2}$, equation (3) becomes:

$$\Delta R(r) = \frac{1}{2} (1 - \frac{1}{r}) \log_2 \left( \frac{\sqrt{r} + \rho}{\sqrt{r_{\infty}} + \rho} \right)$$

(4)

which is negative for $r \in (1, r_{\infty})$. $\Delta R(r)$ is 0 at point $r = 1$ and point $r = r_{\infty}$, which is consistent with the fact that both encoding all pixels at first stage and at second stage are equivalent to original Wyner-Ziv video coding system.

To save complexity, at first stage we use reference frame directly as side information $Y_0$. Therefore $\sqrt{r_{\infty}}$ is at the cross point of the two line in Fig.3, and $\sqrt{r_{\infty}} \approx 12$. We find numerically the optimal $r^*$ for different $\rho$ and plot it in Fig.4 for $r_{\infty} = 144$. From the figure we can find that the optimal down-sample ratio is between 4 ~ 8. Since we can hardly have arbitrary down-sample ratio, we only try down-sample ratio 2,4,8 and 16, and find out that $r = 8$ has best performance.

![Fig. 4: Optimal down-sample $r^*$ and $\Delta R(r^*)$ for $r_{\infty} = 144$](image)

(a) Optimal down-sample ratio $r^*$  (b) Rate difference $\Delta R(r^*)$

3.2. Multistage Refinement

For multiple stage refinement, suppose at each stage $\frac{1}{n}$ of pixels is sent, then after $m$ times transition there are $\frac{m}{n}$ of pixels available to decoder, so down-sample ratio $r = \frac{m}{n}$ and residue variance will be $A \sqrt{\frac{m}{n}} + N$. Compared with using ideal motion search, whose down-sample ratio is 1 and residue variance is $A + N$, the rate redundancy for pixels encoded at next stage are $\frac{1}{2} \log_2 \left( \frac{\sqrt{r_{\infty}} + \rho}{\sqrt{r} + \rho} \right)$, so that the rate redundancy for all pixels is

$$\Delta R = \frac{1}{n^2} \log_2 \left( \frac{\sqrt{r_{\infty}} + \rho}{1 + \rho} \right) + \sum_{m=1}^{n-1} \frac{1}{n^2} \log_2 \left( \frac{\sqrt{m} + \rho}{1 + \rho} \right)$$

(5)

which is a monotonic decreasing function of $\rho$. When $\rho \gg 1$, $\Delta R = 0$; When $\rho \ll 1$ we have:

$$\Delta R \approx \frac{1}{4n} \log_2 \left( \frac{n^n}{(n-1)!} \right) \approx \frac{1}{4} \log_2 e = 0.36 \text{bits/pixel}$$

(6)

1This comes from Stirling formula and is already very accurate when $n = 8$. Actually $n$ can not be too big in our system, or it will reduce coding length.

![Fig. 5: Down-sample patterns: left is 8-Queen pattern, middle is 4-Queen, right is 4-stage successive down-sample](image)

Now we can conclude that successive refinement system can approach ideal motion compensated Wyner-Ziv system very closely with only $0 \sim 2.17 \text{dB}$ loss. Notice that the loss of MCI or MCE based Wyner-Ziv video coding system can be up to 6dB according to the theoretical analysis in [2], which means that successive refinement system can gain up to 4dB over MCI or MCE based WZ coding system.

4. IMPLEMENTATION AND EXPERIMENTAL RESULT

The way we encode pixels is similar to other Wyner-Ziv video encoder, but difference is that all pixels are divided into groups and each group is encoded at each refinement stage. Pixels encoded at same stage are quantized to form biplanes which are encoded by adaptive LDPC code [7] with input length 3168 and density function $\lambda(x) = 0.321x^2 + 0.456x^3 + 0.010x^5 + 0.174x^7 + 0.039x^8$. At decoder, biplanes are decoded from MSB to LSB by iterative decoding starting from the intrinsic log-likelihood ratio (LLR):

$$LLR = \log \left( \frac{P(b_i = 0, b^- | y)}{P(b_i = 1, b^- | y)} \right) = \log \left( \frac{\int_{x^+} f_{X|y}(x)dx}{\int_{x^-} f_{X|y}(x)dx} \right)$$

(7)

where $b_i$ is current binary to be decoded, $b^-$ are binaries which have been decoded, $f_{X|y}$ is the pdf of $X$ given $Y$ which is assumed to be Lapacion in our experiment, and $x^- = \min(x|b^-)$, $x^+ = \max(x|b^-)$, $t = \min(x|b^-), b_i = 1) - 0.5$.

As to down-sample pattern, i.e. which pixels in a block will be encoded at each stage, for our two-stage refinement system, the problem is equivalent to pixel decimation in fast motion estimation, so that N-Queen pattern [8] in Fig.5 is used to get better motion estimation performance. And for multistage successive refinement, the down-sample pattern is also given in Fig.5 where the numbers on each pixels are encoding order.

To improve motion estimation efficiency, the mean of four neighbor blocks’ SAD, also contribute to current block’s SAD through:

$$SAD' = \lambda SAD + (1 - \lambda)NSAD$$

(8)

And for $n = 8$, even QCIF sequence can have an input coding length of 3168 which only lose coding efficiency very slightly.
where $\overline{NSAD}$ is the mean of neighbor SAD, and $\lambda$ is set to be 0.5. For multistage refinement, after decoding all bitplanes of all pixels, motion is refined again. Finally each pixel $x$ is reconstructed by MMSE estimation: $\hat{x} = E(x|y, Q(x))$.

The experiment focus on low delay low complexity system, where only 1 previous reference frame is available, so that at first stage reference frame are used directly as side information. First 100 frames of QCIF sequences at 15Hz is encoded and GOP structure is 'IPPP' with GOP length 8. In all schemes, intra frames are encoded by H.263+. As is shown in Fig.6, the proposed multistage resolution refinement(MRR) scheme performs only slightly better than proposed two-stage resolution refinement.

We also implement multistage bitplane refinement(MBR) in our system and the results indicate that the performance is lower than both of our two schemes. Our multistage resolution refinement gains 0.4~2.5dB over MBR. Notice that MBR is reported to gain 2~3dB over MCI and MCE based Wyner-Ziv system. As to complexity, all 3 approaches have equally low complexity at encoder as other pixel domain Wyner-Ziv video encoders. And at decoder, the proposed two-stage resolution refinement approach has lowest complexity because only 12.5% pixels are used in motion estimation. Our MRR has much lower complexity than MBR, because in MRR SAD of lower stages can be reused by higher stages.

Compared to Wyner-Ziv system with ideal motion compensation, proposed MRR lose no more than 2dB which is consistent with our rate distortion analysis.

For mobile sequence, all 3 successive refinement schemes outperform H.263+ at high bitrate mainly because of longer coding length and more accurate probability model and reconstruction function. But at low bitrate they are outperformed by H.263+ because of the contribution of DCT. For football, intra frame encoder performs the best because of fast scene change.

5. CONCLUSION

Successive resolution refinement scheme was proposed in this paper. Experimental result showed that our scheme had better RD performance than bitplane refinement scheme, with much lower complexity. Optimal down-sample ratio was derived for two-stage refinement. Side information loss of proposed multistage resolution refinement scheme was proved to be only up to $\frac{1}{4} \log_2 e$ bit/pixel, or equivalently 2.17dB, which was consistent with experimental results.

6. ACKNOWLEDGEMENT

The authors would like to thank David Varodayan and Anne Aaron (Information Systems Lab. of Stanford University), for their generous sharing and supporting of their low-density parity-check (LDPC) accumulate codes [7].