A STATISTICAL MODEL FOR A MIXED RESOLUTION WYNER-ZIV FRAMEWORK

Bruno Macchiavello[1], Debargha Mukherjee[2] and Ricardo L. De Queiroz[1]

[1]Universidade de Brasilia, Departamento de Engenharia Eletrica, Brasilia, Brazil
[2]Hewlett Packard Laboratories, Palo Alto, CA, USA.

ABSTRACT

In this paper we present a rate distortion analysis and a statistical model in order to select coding parameters for memoryless coset codes, for a spatial scalability based mixed resolution Wyner-Ziv framework. The mixed resolution framework, used in this work, is based on full resolution coding of the key frames and spatial 2-layer coding of the intermediate non-reference frames where the spatial enhancement layer is Wyner-Ziv coded. The framework enables reduced encoding complexity through reduced spatial-resolution encoding of the non-reference frames. The quantized transform coefficients of the Laplacian residual frame are mapped to cosets and sent to the decoder. A correlation estimation mechanism that guides the parameter choice process is proposed based on extracting edge information and residual error rate in co-located blocks from the low resolution base layer.

Index Terms— Wyner-Ziv, reversed-complexity coding, spatial scalability

1. INTRODUCTION

Distributed coding has its roots in the information theory proofs of Slepian and Wolf 0 for the lossless case and Wyner and Ziv [2] for the lossy case. Recently, a kind of video coding referred to as reversed complexity coding, has been proposed based on these principles, where the encoder complexity is reduced by obviating the need for full motion search, but the performance loss is partially recovered by a more complex decoding process exploiting source statistics. A review of distributed video coding can be found in [3].

In [5]-[7], we proposed a mixed resolution framework that can be implemented as an optional coding mode in any existing video codec standard. In this framework, the reference frames are coded exactly as in a regular codec as *I*-, *P*- or reference *B*-frames, at full resolution. For the non-reference *P*- or *B*- frames, called non-reference Wyner-Ziv (NRWZ) frames, the encoding complexity is reduced by low resolution (LR) encoding. As shown in Figure 1, the non-reference frames are decimated and coded using decimated versions of the reconstructed reference frames in the frame store. Then the Laplacian residual, obtained by taking the difference between the original frame and an interpolated version of the LR layer reconstruction, is Wyner-Ziv coded to form the enhancement layer. Since the reference frames are regular coded, there are no drift errors. Ideally, the number of non-references frames and the decimation factor can be varied dynamically based on the complexity reduction target. At the decoder, a high quality version of the non-reference frames are generated by a multi-frame motion-based mixed super-resolution mechanism [5]-[9]. The interpolated LR reconstruction is subtracted form this frame to obtain the side-information Laplacian residual frame. Thereafter, the Wyner-Ziv layer is channel decoded to obtain the final reconstruction. Note that for encoding and decoding the LR frame, all reference frames in the frame store and syntax elements are first scaled to fit the non-reference LR coded frame. Related work [4] has also explored spatial reduction recently, but our mixed resolution approach while less aggressive in complexity reduction can achieve better compression efficiency.

In realistic usage scenarios for video communication using mobile power-constrained devices, it is not necessary for a video encoder to always transmit video to a more powerful machine or server. In the mixed resolution approach, the LR bit-stream can be decoded immediately for real-time communications albeit at lower quality. However, the main difference with other work in this area is that we do not employ a feedback channel for rate-estimation, thereby enabling the enhancement layer to be decoded offline. This requirement necessitates a sophisticated mechanism for estimating the correlation statistics at the encoder, followed by mapping the estimated statistics to actual encoding parameters. We present in this paper, as a continuation of previous works [5]-[7], a statistical model as well as a mechanism to estimate the model parameters, based on which optimal coding parameters [6][10] for memoryless coset codes can be selected.

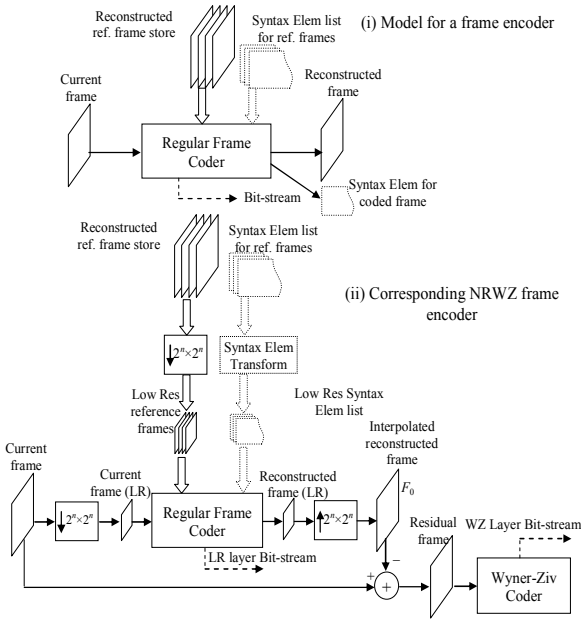

Figure 1. Non-reference Wyner-Ziv frames Encoder

2. WYNER-ZIV CODER

Our WZ coder operates on the Laplacian residual error frame in the block-transform domain. Let the random variable X denote the transform coefficients. Then, the quantization of X, with a dead-zone quantizer, yields a quantization index random variable Q: $Q = \phi(X, QP)$, QP being the quantization step-size. Next, cosets are computed to yield a random variable C: $C = \psi(Q, M) = \psi(\phi(X, QP), M)$, M being the coset modulus:

$$\psi(Q,M) = \begin{cases} Q - M\lfloor Q/M \rfloor & , \; Q - M\lfloor Q/M \rfloor < M/2 \\ Q - M\lfloor Q/M \rfloor - M & , \; Q - M\lfloor Q/M \rfloor \geq M/2 \end{cases} \quad (1)$$

If quantization bin q corresponds to interval $[x_l(q), x_h(q)]$, then the probability of the bin $q \in \Omega_Q$, and the probability of a coset index $c \in \Omega_C$ are given by the probability mass functions:

$$p_Q(q) = \int_{x_l(q)}^{x_h(q)} f_X(x)dx \quad (2)$$

$$p_C(c) = \sum_{q \in \Omega_Q : \psi(q,M)=c} p_Q(q) = \sum_{q \in \Omega_Q : \psi(q,M)=c} \int_{x_l(q)}^{x_h(q)} f_X(x)dx \quad (3)$$

The entropy coder that already exists in the regular coder can be reused for C, but a different entropy coder conditioned on M for each coefficient should yield better compression.

For decoding, existence of the corresponding side-information coefficient random variable Y is assumed. The minimum MSE reconstruction function $\hat{X}_{YC}(y,c)$ based on unquantized side information y and received coset index c, is given by:

$$\hat{X}_{YC}(y,c) = E(X/Y=y, C=c) = \frac{\sum_{q \in \Omega_Q : \psi(q,M)=c} \int_{x_l(q)}^{x_h(q)} x f_{X/Y}(x,y)dx}{\sum_{q \in \Omega_Q : \psi(q,M)=c} \int_{x_l(q)}^{x_h(q)} f_{X/Y}(x,y)dx} \quad (4)$$

The regularly coded reference frames in our framework are assumed to be coded with quantization step-size QP_t. Additionally, the LR layer of the NRWZ frames is also assumed to be coded with the same step-size QP_t. Therefore, the enhancement Wyner-Ziv layer for NRWZ frames should be ideally coded such that the distortion is about the same level as that obtained by regular coding with quantization step-size QP_t.

3. CHOOSING CODING PARAMETERS

In order to make an optimal choice of the quantization and modulus parameters $\{QP, M\}$, we assume a general enough statistical model: $Y = \rho X + Z$, where X is a Laplacian distributed transform coefficient, Z is additive Gaussian noise uncorrelated with X and $0 < \rho \leq 1$ is an attenuation factor expected to decay at higher frequencies. Note that while this is a generalization of the simpler model: $Y = X + Z$ dealt with in [6][10], since we can rewrite it as $Y/\rho = X + Z/\rho$, the same procedure [6][10] can be applied with simply $(\sigma_Z/\rho)^2$ replacing σ_Z^2 and Y/ρ replacing Y during decoding. In the rest of this section, we review the optimal parameter choice mechanism for the $Y = X + Z$ model, however to use for the $Y = \rho X + Z$ model, σ_Z needs to be replaced by (σ_Z/ρ).

3.1. Memoryless coset codes followed by minimum MSE recostrunction

The first step is to obtain expressions for expected rate and distortion functions for the memoryless coset codes described in Section 2, for a given $\{QP, M\}$ pair. Assuming an ideal entropy coder for the coset indices, the expected rate would be the entropy of the source C, given by:

$$\begin{aligned} E(R_{YC}) = H(C) &= -\sum_{c \in \Omega_C} p_C(c) \log_2 p_C(c) \\ &= -\sum_{c \in \Omega_C} \{ \sum_{q \in \Omega_Q : \psi(q,M)=c} \int_{x_l(q)}^{x_h(q)} f_X(x)dx \} \times \log_2 \{ \sum_{q \in \Omega_Q : \psi(q,M)=c} \int_{x_l(q)}^{x_h(q)} f_X(x)dx \} \\ &= -\sum_{c \in \Omega_C} \{ \sum_{q \in \Omega_Q : \psi(q,M)=c} [m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q))] \} \times \\ &\quad \log_2 \{ \sum_{q \in \Omega_Q : \psi(q,M)=c} [m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q))] \} \end{aligned} \quad (5)$$

where $m_X^{(i)}(x) = \int_{-\infty}^{x} x'^i f_X(x')dx'$.

Further, it can be shown [6][10] that the expected distortion D_{YC} for the minimum MSE reconstruction function (4), is:

$$E(D_{YC}) = \sigma_X^2 - \int_{-\infty}^{\infty} \{ \sum_{c \in \Omega_C} \frac{\left(\sum_{q \in \Omega_Q : \psi(q,M)=c} [m_{X/Y}^{(1)}(x_h(q),y) - m_{X/Y}^{(1)}(x_l(q),y)] \right)^2}{\left(\sum_{q \in \Omega_Q : \psi(q,M)=c} [m_{X/Y}^{(0)}(x_h(q),y) - m_{X/Y}^{(0)}(x_l(q),y)] \right)} \} f_Y(y)dy \quad (6)$$

where $m_{X/Y}^{(i)}(x,y) = \int_{-\infty}^{x} x'^i f_{X/Y}(x',y)dx'$.

A viable coding choice is to just use zero-rate coding, where no information is transmitted (i.e. QP→∞ or $M=1$). Then the rate will be 0 and it can be shown [6][10] that the expected distortion based on optimal reconstruction using Y alone is given by:

$$E(D_Y) = \sigma_X^2 - \int_{-\infty}^{\infty} m_{X/Y}^{(1)}(\infty, y)^2 f_Y(y)dy \quad (7)$$

3.2. Laplacian source with additive Gaussian noise

We now specialize for the case of Laplacian X and Gaussian Z, *i.e.*:

$$f_X(x) = \frac{1}{\sqrt{2}\sigma_X} e^{-\left|\frac{x\sqrt{2}}{\sigma_X}\right|}, \quad f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_Z} e^{-\frac{1}{2}\left|\frac{z}{\sigma_Z}\right|^2} \quad (8)$$

Closed form expressions for $m_x^{(0)}(x)$ and $m_x^{(1)}(x)$ can be readily evaluated. Defining:

$$\gamma_1(x) = erf\left(\frac{\sigma_X x - \sqrt{2}\sigma_Z^2}{\sqrt{2}\sigma_X\sigma_Z}\right), \quad \gamma_2(x) = erf\left(\frac{\sigma_X x + \sqrt{2}\sigma_Z^2}{\sqrt{2}\sigma_X\sigma_Z}\right) \quad (9)$$

where $erf(x) = (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$, and using $Y=X+Z$, we have:

$$f_Y(y) = \frac{1}{2\sqrt{2}\beta(y)\sigma_X} e^{\sigma_1^2/\sigma_Z^2} [\gamma_1(y) + 1.0 - \beta(y)^2(\gamma_2(y) - 1.0)] \quad (10)$$

$$f_{X/Y}(x,y) = \frac{\sqrt{2}\beta(y)}{\sqrt{\pi}\sigma_Z} \frac{e^{-\left|\frac{x\sqrt{2}}{\sigma_X}\right| \frac{1}{2}\left(\frac{y-x}{\sigma_Z}\right)^2 - \frac{\sigma_X^2}{\sigma_Z^2}}}{[\gamma_1(y) + 1.0 - \beta(y)^2(\gamma_2(y) - 1.0)]} \quad (11)$$

Given $f_{X/Y}(x,y)$, closed form expressions for $m_{X/Y}^{(0)}(x,y)$ and $m_{X/Y}^{(1)}(x,y)$ based on the *erf*() function (evaluated based on a polynomial approximation in [11]) can now be computed. All the expected rate and distortion values in Section 3.1 for a given $\{QP, M\}$ pair can now be evaluated based on these moments in conjunction with numerical integration with $f_Y(y)$.

While there are many different combinations of $\{QP, M\}$ that can be chosen, only those combinations that yield R-D points on the Pareto frontier are optimal ones. A bunch of R-D points are evaluated offline by varying M and QP at small increments, for a given $\{\sigma_X, \sigma_Z\}$. The sub-optimal choices for $\{QP, M\}$ combination are pruned out by finding the Pareto-Optimal set P, wherein each point representing a code, is such that no other code is superior to it. For a given target

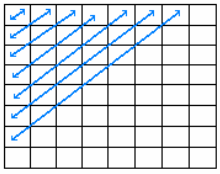

Figure 2. First seven frequency bands in an 8×8 block

distortion upper bound D_t, the optimal {QP, M} combination is selected by picking the code from the optimal set P that yields the closest distortion to D_t, not exceeding it.

3.3. Distortion target matching

Finally, we note that it is advantageous in our framework to specify D_t in terms of a target quantization step-size QP_t for regular coding. The expected distortion from regular encoding followed by MSE reconstruction (without side-information) is given by:

$$E(D_Q) = \sigma_X^2 - \sum_{q \in \Omega_Q} \frac{\left(m_X^{(1)}(x_h(q)) - m_X^{(1)}(x_l(q))\right)^2}{\left(m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q))\right)} \quad (12)$$

This function computed for a given QP_t yields D_t. Thereafter, a sorted set P can be searched for the optimal code with distortion closest to D_t not exceeding it.

In practice, this mapping from QP_t to {QP, M} can be pre-computed and stored in a *normalized* table for a range of QP_t values, for a given σ_Z assuming $\sigma_X=1$. To use for an arbitrary σ_X, the values of QP_t and QP in a normalized table need to be appropriately scaled before and after table-lookup. A limited set of normalized tables can be stored in a codec for a range of σ_Z values at small steps.

Reverting back to our $Y = \rho X + Z$ model, if the model parameters {ρ, σ_X, σ_Z} are known, then in order to find the {QP^*, M} combination corresponding to a given QP_t^*, we simply have to evaluate $\sigma_Z/(\rho.\ \sigma_X)$ to find the normalized table to use for look-up, find the closest entry in it corresponding to target $QP_t = QP_t^*/\sigma_X$, read off QP and M, and finally scale to obtain the final $QP^* = QP.\ \sigma_X$.

4. CORRELATION STATISTICS ESTIMATION

In this Section, we will propose a mechanism to estimate the parameters {ρ, σ_X, σ_Z} for our $Y = \rho X + Z$ model within the proposed spatial scalability framework. These parameters are to be next used for parameter selection as described in the previous section.

The model parameters need to be specialized for each frequency band (*FB*) within a block, where the FB is defined as diagonals in a transform block as shown in Figure 2. Also note that the correlation is obviously dependent on the quantization step-size QP_t for the reference frame and the LR layer. Besides, other vital information can be extracted from the LR layer to direct the estimation process as described below. Note that since any data from the LR layer is available at both decoder and encoder, no overhead bits need to be transmitted to convey this information. An alternative approach may transmit explicitly some statistical information, but in this work, we adopt a no-overheads approach.

To generate the estimation models we use a training based approach where X (transform coefficients of Laplacian residual of original frame) and Y (transform coefficients of residual after multi-frame processing) data for each FB is collected for a set of training video sequences for varying values of QP_t along with the corresponding values of additional information extracted from the LR layer.

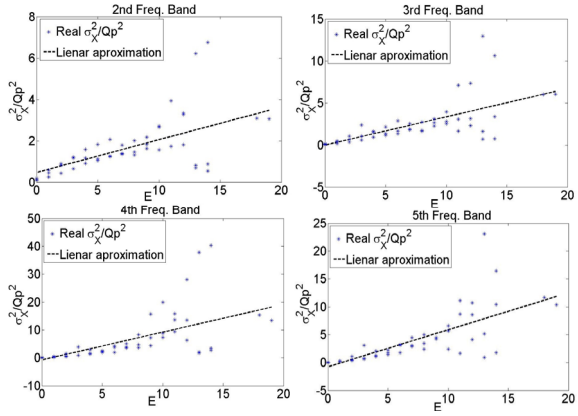

Figure 3. Real σ_X^2 / QP_t^2 *vs.* E and linear approximation

4.1. Estimation of σ_X^2- variance of Laplacian residual coefficients

The variance of a Laplacian residual coefficient (σ_x^2) will not be the same in every block of a coded frame. It will not only depend on QP_t and *FB*, but also on the high frequency content of the block. If the original frame has a high edge content it is likely that the error between the decimated-interpolated version and the original one would be larger. Even though the exact high frequency content in an original frame is not available at the decoder, we can use an edge activity measure of the reconstructed LR block as a parameter to estimate σ_X^2. It is intuitive to think that the edge activity in the LR block will be correlated with the energy of the high frequency coefficients of the Laplacian residual, while the energy at the lower frequencies in the Laplacian residual will be more related to QP_t. The edge activity, denoted E, is computed as the accumulate sum of the difference between the lines and columns of a macroblock in the reconstructed version of the interpolated LR frame. Then, the estimated σ_X^2 is modeled as a function of QP_t, *FB* and E. That is:

$$\sigma_X^2 = f_1(QP_t, FB, E) \quad (13)$$

We next assume σ_X^2 to be proportional to QP_t^2. Further, after processing the training data we find that it is enough to model the remaining part linearly for each *FB*, so that:

$$\sigma_X^2 = (k_{1,FB}E + k_{2,FB})QP_t^2 \quad (14)$$

where $k_{i,FB}$ are constants that vary for each frequency band. In Figure 3, we show the linear approximations used for σ_X^2 / QP_t^2 vs. E, compared to the real training data for some of the frequency bands.

4.2. Estimation of the correlation parameter

To estimate the ρ parameter, we use a simplified model assuming that it only depends on QP_t and *FB*:

$$\rho = f_2(QP_t, FB) \quad (15)$$

Note, that a better estimation of ρ could be the subject of future improvements. In this case, if T_{FB,QP_t} represents the training data set for QP_t and FB, we estimate:

$$\rho_{FB,QP_t} = \arg\min_p \sum_{(X,Y) \in T_{FB,QP_t}} \left(\left\| Y - \rho X \right\|^2 \right) \quad (16)$$

Then, after ρ_{FB,QP_t} is calculated for the entire training set, a linear approximation is used to model it as:

$$\rho_{FB,QP_t} = (k_{3,FB}QP_t + k_{4,FB}) \quad (17)$$

Note that with higher QP the variables X and Y are less correlated then $k_{3,FB}$ is a negative constants.

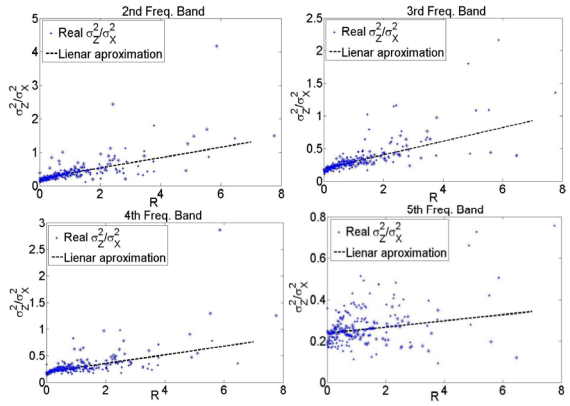

Figure 4. Real σ_Z^2/σ_X^2 *vs.* R_n and linear approximation.

4.3. Estimation of the variance of the Gaussian noise

To estimate σ_Z^2 from the training data set, we first calculate $Z = Y - \rho_{FBQP}X$. Further, we conjecture that σ_z^2 for a macroblock in the enhancement layer will depend on the residual error rate R used to code a co-located 8×8 block in the LR base layer along with QP_t, FB and E. A higher rate in the LR base layer indicates greater inaccuracy of motion estimation at reduced resolution, and therefore the multi-frame super-resolution process is also likely to yield more inaccurate estimate of the high-resolution frame at the decoder, leading to increase in σ_Z^2. However, since R depends also on QP_t we use normalized rate $R_n = R \times QP_t^2$ in order to remove the effect of QP_t. Now, we can model σ_Z^2 as:

$$\sigma_Z^2 = f_3(QP_t, FB, E, R_n)\ . \qquad (18)$$

We next assume σ_Z^2 to be proportional to σ_X^2 for a given FB and R_n, and the effect of QP_t and E to be subsumed within σ_X^2. Further, the remaining part is modeled linearly for each FB, such that:

$$\sigma_Z^2 = (k_{5,FB}R_n + k_{6,FB})\sigma_X^2 \qquad (19)$$

In Figure 4 we show the linear approximations used for σ_Z^2/σ_X^2 vs. R_n, compared to the real training data, for some of the frequency bands.

5. RESULTS AND CONCLUSIONS

This optimal parameter choice was applied to the proposed mixed resolution framework using H.263+. In Figure 5 and Figure 6 we compare the performances of 1. a regular H.263+ codec working in *IBPB* mode; 2. the LR base where the *B*-frames are encode at quarter resolution (indicated as *Z* frames) and simply interpolated at the decoder; 3. the results from decoding both layers of the reversed complexity codec without optimal parameter choice (fixing *{QP,M}* for each frequency band); and 4. the results from decoding both layers using the proposed estimation of the correlated statistic for optimal parameter choice. It can be observed that a better statistical model and parameter choice mechanism improve significantly the performance of the *WZ*-codec. As future work, this model will be implemented in the latest video codec standard H.264.

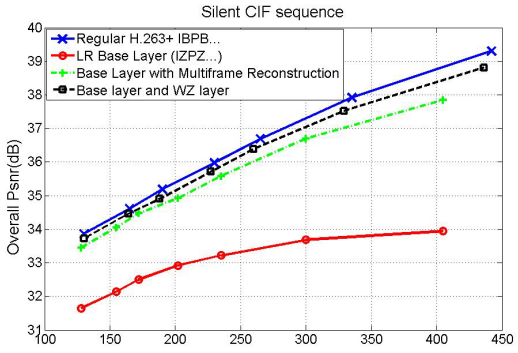

Figure 5. Performance for H.263+ (Silent CIF sequence).

6. REFERENCES

[1] J. D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 471–480, July 1973.

[2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[3] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding," *Proc. of the IEEE*, vol. 93, no. 1, pp. 71-83, January 2005.

[4] M. Wu, G. Hua, C. W. Chen, "Syndrome-based lightweight video coding for mobile wireless application," *Proc. Int. Conf. on Multimedia and Expo*, 2006, pp. 2013-2016.

[5] D. Mukherjee, "A robust reversed complexity Wyner-Ziv video codec introducing sign-modulated codes," *HP Labs Technical Report*, HPL-2006-80.

[6] D. Mukherjee and B. Macchiavello and R. L. de Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework", *Proc. of SPIE Visual Com. and Img. Proc., vol 6508, pp. 1Y1-1Y12,* Jan 2007.

[7] B. Macchiavello, R.L de Queiroz and D. Mukherjee, "Motion-based side-inforamtion generation for a scalable Wyner-Ziv video Coding," *to appear in IEEE Int. Conf. on Img. Proc.*, San Antonio, 2007.

[8] Z. Li, L. Liu, and E. J. Delp, "Rate Distortion Analysis of Motion Side Estimation in WynerZiv Video Coding," *IEEE Trans on Img. Proc.*, vol. 16, no. 1, pp. 98–113, Jan 2007.

[9] L.W. Kang and C. S. Lu, "Wyner-ziv video coding with coding mode-aided motion compensation," *In Proc. IEEE Int. Conf. on Img. Proc.*, pp. 237–240, 2006.

[10] D. Mukherjee, "Optimal parameter choice for Wyner-Ziv coding of Laplacian sources with decoder side-information," *HP Labs Technical Report*, HPL-2007-34.

[11] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.

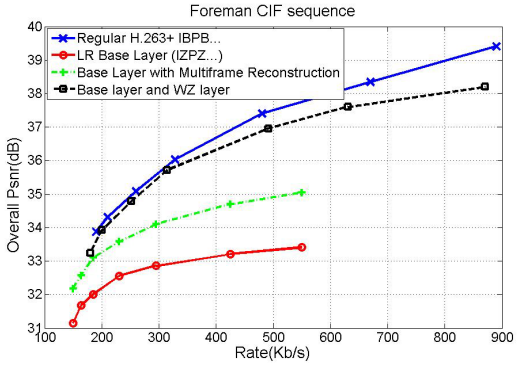

Figure 6. Performance for H.263+ (Foreman CIF sequence)