RESOLUTION SCALABLE MOTION-COMPENSATED JPEG 2000

Robert A. Cohen and and John W. Woods

Center for Image Processing Research
Rensselaer Polytechnic Institute, Troy, NY 12180
cohen@cipr.rpi.edu, woods@ecse.rpi.edu

ABSTRACT

This paper proposes two motion-compensated extensions to JPEG 2000 for the efficient scalable compression of video. The first one is a rather conventional pyramid coder with MCTF on each spatial level and a closed loop coding structure. The second one is a novel open-loop coder that preserves the full scalability of the high-spatial level. We show that for certain reasonable values of high-resolution and low-resolution bitrates that both coders have approximately equal performance, with the new open-loop coder retaining full scalability advantages.

Index Terms— Scalable video coding, JPEG 2000, MCTF

1. INTRODUCTION

A recently developed international standard for wavelet-based image compression is JPEG 2000 [1]. It exhibits spatial and SNR scalability, but since it is only an intraframe coder, it does not take advantage of temporal dependencies to improve compression efficiency. Motion JPEG 2000 can be used to code video, but it is primarily a file format used for storing intraframe-coded images. For broadcast and transmission over the networks or channels commonly in use today, however, interframe compression is usually needed to allow reasonably good video quality given a limited bitrate.

Some previous interframe wavelet-based coders related to ours can be found in [2, 3]. These methods use wavelet-based coding both spatially and temporally to get good compression efficiency along with the functional benefits of high scalability. The goal of the present work is to combine ideas from these interframe coders that use motion-compensated temporal filtering (MCTF) with the intraframe JPEG 2000 coder to produce a motion-compensated coder based on this standard.

2. CLOSED-LOOP HIERARCHICAL MC-JPEG 2000

In this section we provide a description and analysis of a closed-loop system in which quantized low-resolution data is used to predict high-resolution subbands. The low-resolution data is coded into a base layer, and the prediction residuals along with unpredicted subbands are coded into an enhancement layer.

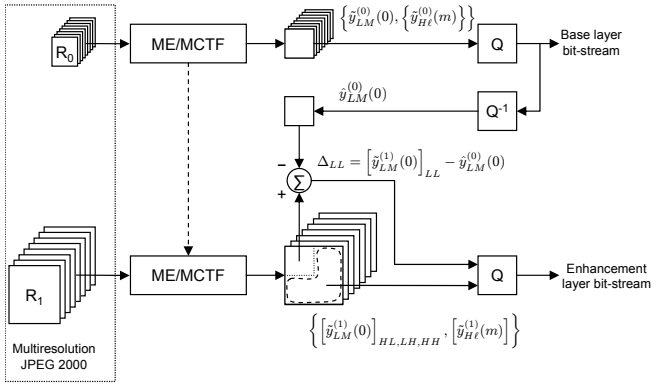

Fig. 1. Closed-loop hierarchical MC-JPEG 2000 encoder.

2.1. System description

A diagram of the closed-loop resolution scalable motion-compensated JPEG 2000 encoder is shown in Fig. 1. The sequence of frames for resolution level r input to our Haar-based MCTF is $\{y^{(r)}(0), y^{(r)}(1), \ldots, y^{(r)}(M)\}$, where M is the size of the group of frames. The MCTF output for the first pair of frames would be $\tilde{y}^{(r)}_{L0}(0)$ and $\tilde{y}^{(r)}_{H0}(0)$, corresponding to temporal level zero. To generate temporal level one, the MCTF operates on the lowpass temporal subbands of temporal level zero. For example, the first two lowpass subbands of temporal level zero $\tilde{y}^{(r)}_{L0}(0)$ and $\tilde{y}^{(r)}_{L0}(1)$ would be input to the MCTF, yielding $\tilde{y}^{(r)}_{L1}(0)$ and $\tilde{y}^{(r)}_{H1}(0)$. The lowest temporal subband along with the temporal high subbands are then quantized using a multicomponent JPEG 2000 encoder [4].

Once the low resolution $r = 0$ is processed, the motion estimation for the high resolution $r = 1$ can be initialized by scaling up the low-resolution motion vectors, followed by a small local search. Because the motion vectors in both resolutions are similar, we use $\tilde{y}^{(0)}_{LM}(0)$ as a prediction for the spatial-LL subband of $\tilde{y}^{(1)}_{LM}(0)$. The resulting prediction error Δ_{LL} along with the non-predicted temporal-high subbands are then quantized using multicomponent JPEG 2000.

2.2. Analysis of closed-loop system

To simplify notation, the temporal-low subband for the low resolution will be denoted y_0, and the spatial-LL subband for the temporal-low subband of the high resolution will be denoted y_1. The error incurred by quantizing y_0 at rate R_0 is denoted q_0, and the quantization error for coding y_1 at rate R_1 is q_1. At the decoder, the decoded base-layer subband $\hat{y}_0$ is added to the decoded prediction error $\hat{\Delta}_{LL}$ resulting in the reconstructed subband $\hat{y}_1$. The reconstruction error for y_1 in the closed-loop system is therefore

$$\varepsilon_1 = y_1 - (\hat{y}_0 + \hat{\Delta}_{LL}) , \tag{1}$$

where

$$\begin{aligned} \hat{y}_0 &= y_0 + q_0 \\ \hat{\Delta}_{LL} &= \Delta_{LL} + q_1 \\ \Delta_{LL} &= y_1 - \hat{y}_0 \quad = \quad y_1 - y_0 - q_0 . \end{aligned}$$

Using a Gaussian $D(R)$ model for our quantizers, the reconstruction error distortion becomes

$$\sigma^2_{\varepsilon_1} = \sigma^2_{q_1} = \sigma^2_{\Delta_{LL}} 2^{-2R_1} = \sigma^2_{(y_1 - \hat{y}_0)} 2^{-2R_1} . \tag{2}$$

After applying the orthogonality principle, this reduces to

$$\sigma^2_{\varepsilon_1} = \left(\sigma^2_{y_0} + \sigma^2_{y_1} - 2\rho_{y_0 y_1}\sigma_{y_0}\sigma_{y_1} + \sigma^2_{q_0}\right) 2^{-2R_1} , \tag{3}$$

where $\rho_{y_0 y_1}$ denotes the correlation coefficient. Using a similar $D(R)$ model, we get

$$\begin{aligned} \sigma^2_{\varepsilon_1} &= \sigma^2_{y_0} 2^{-2R_{\text{tot}}} \\ &\quad + \left(\sigma^2_{y_0} + \sigma^2_{y_1} - 2\rho_{y_0 y_1}\sigma_{y_0}\sigma_{y_1}\right) 2^{-2R_{\text{tot}}} \cdot 2^{2R_0} \end{aligned} \tag{4}$$

for $|\rho_{y_0 y_1}\sigma_{y_0}\sigma_{y_1}| < \frac{\sigma^2_{y_0} + \sigma^2_{y_1}}{2}$, where $R_{\text{tot}} = R_0 + R_1$. Since the slope of the reconstruction error variance with respect to R_0 is always positive given the constraints of (4), the base-layer rate that minimizes the reconstruction error and the corresponding error variance are therefore

$$\begin{aligned} &R_{0_{\text{opt}}} = 0 \quad \text{bits/sample} \\ &\sigma^2_{\varepsilon_1}\Big|_{R_0 = R_{0_{\text{opt}}}} = \left(2\sigma^2_{y_0} + \sigma^2_{y_1} - 2\rho_{y_0 y_1}\sigma_{y_0}\sigma_{y_1}\right) 2^{-2R_{\text{tot}}} . \end{aligned} \tag{5}$$

A graph of this analytical result for a set of fixed total rates is included in Fig. 3. This behavior is consistent with results obtained with other closed-loop pyramid coding methods [5].

3. OPEN-LOOP HIERARCHICAL MC-JPEG 2000

We will now look at a novel open-loop system that eliminates the quantizer feedback loop of the closed-loop encoder. Analysis will show the performance trade-offs when allocating rates between the base and enhancement layers.

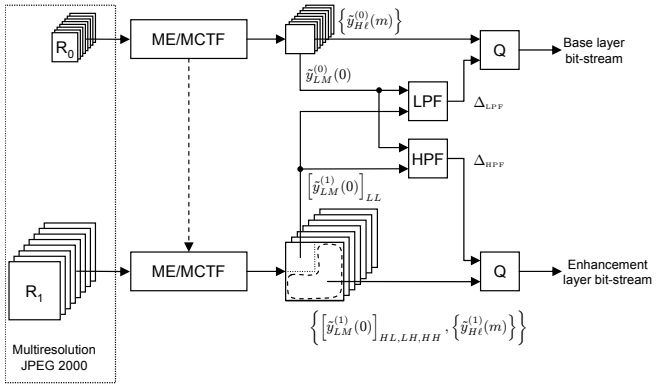

Fig. 2. Open-loop hierarchical MC-JPEG 2000 encoder.

3.1. System description

For image and video coding, we know that subband/wavelet coders can provide scalability and energy compaction benefits over hybrid DPCM-like coders. Using these ideas, we can replace the differential predictor in our closed-loop MC-JPEG 2000 encoder with a wavelet-based predictor. Doing so may not always be as efficient as using a DPCM predictor, but it may be worthwhile because the feedback loop is eliminated, thereby reducing the complexity of the encoder and eliminating the decoder mismatch problem.

In the closed-loop system, a decoded low-resolution frame is subtracted from the LL subband of the high-resolution frame, resulting in the differential Δ_{LL}, that is quantized again. We will now replace this quantizer-cascade with a Haar transform, so we get the lowpass and highpass signals:

$$\Delta_{\text{LPF}} = \frac{1}{\sqrt{2}} (y_0 + y_1) ; \qquad \Delta_{\text{HPF}} = \frac{1}{\sqrt{2}} (y_0 - y_1) . \tag{6}$$

The lowpass signal Δ_{LPF}, which in effect is a scaled average between the low resolution and the LL subband of the high resolution, will be transmitted in the base layer bitstream along with the temporal high subbands. The highpass signal Δ_{HPF} will be coded in the enhancement layer along with all the other spatial and temporal high subbands needed to reconstruct the full resolution sequence. Hence, we have eliminated the predictive feedback loop. A diagram of the open-loop hierarchical MC-JPEG 2000 encoder is shown in Fig. 2.

3.2. Analysis of open-loop system

For a given fixed total channel or network rate R_{tot}, we would like to determine the low-resolution rate that gives the best high-resolution performance. The reconstruction error for y_1 in the open-loop system is

$$\varepsilon_1 = y_1 - \frac{\hat{\Delta}_{\text{LPF}} - \hat{\Delta}_{\text{HPF}}}{\sqrt{2}} . \tag{7}$$

For our model, the base and enhancement-layer quantizers are uncorrelated. Using (6) and the quantization errors q_0 and q_1, the reconstruction error variance becomes

$$\sigma_{\varepsilon_1}^2 = \frac{\sigma_{q_0}^2}{2} + \frac{\sigma_{q_1}^2}{2} \,. \tag{8}$$

Since the base-layer quantizer is quantizing Δ_{LPF}, and the enhancement layer corresponds to Δ_{HPF}, the quantizer variances approximated by a Gaussian $D(R)$ model become

$$\sigma_{q_0}^2 = \sigma_{\Delta_{\text{LPF}}}^2 2^{-2R_0} \,; \qquad \sigma_{q_1}^2 = \sigma_{\Delta_{\text{HPF}}}^2 2^{-2R_1} \,. \tag{9}$$

The variances of the output of the Haar synthesis filters of (6) are

$$\begin{aligned} \sigma_{\Delta_{\text{LPF}}}^2 &= E\left[\left(\tfrac{y_0+y_1}{\sqrt{2}}\right)^2\right] = \tfrac{\sigma_{y_0}^2}{2} + \tfrac{\sigma_{y_1}^2}{2} + \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} \\ \sigma_{\Delta_{\text{HPF}}}^2 &= E\left[\left(\tfrac{y_0-y_1}{\sqrt{2}}\right)^2\right] = \tfrac{\sigma_{y_0}^2}{2} + \tfrac{\sigma_{y_1}^2}{2} - \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} \,. \end{aligned} \tag{10}$$

After substituting (10) and (9) into (8), the reconstruction error becomes

$$\begin{aligned} \sigma_{\varepsilon_1}^2 = {} & \tfrac{1}{2}\left[\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} + \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right] 2^{-2R_0} \\ & + \tfrac{1}{2}\left[\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} - \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right] 2^{-2R_1} \,, \end{aligned} \tag{11}$$

for $|\rho_{y_0y_1}| < 1$. We can see here that the contribution to the reconstruction error by the base layer is the average variance *plus* the covariance between the low-resolution frame and the LL subband of the high-resolution frame, and the contribution of the enhancement layer is the average variance *minus* the covariance. This makes intuitive sense, because if y_0 is a good prediction of y_1, allocating more rate to the base layer will reduce the reconstruction error more quickly than increasing the enhancement-layer rate. If y_0 is a perfect prediction of y_1, i.e. $\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} = \sigma_{y_0}^2 = \sigma_{y_1}^2$, then the reconstruction error is

$$\sigma_{\varepsilon_1}^2 = \sigma_{y_0}^2 \, 2^{-2R_0} \,,$$

so allocating all the available rate to the base-layer will do the best job of reducing the reconstruction error. Note that we are referring only to the rate used for the predicted temporal low subbands of the MCTF. In the fully-implemented encoder, some rate is also used to code the temporal and spatial high subbands. If y_0 tells us nothing about y_1, i.e. $\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} = 0$, the reconstruction error is

$$\sigma_{\varepsilon_1}^2 = \frac{\sigma_{y_0}^2 + \sigma_{y_1}^2}{4}\left(2^{-2R_0} + 2^{-2R_1}\right)$$

In this case, the prediction is useless, so the Haar synthesis filter in the decoder needs equal contributions from both base and enhancement layers in order to reconstruct y_1.

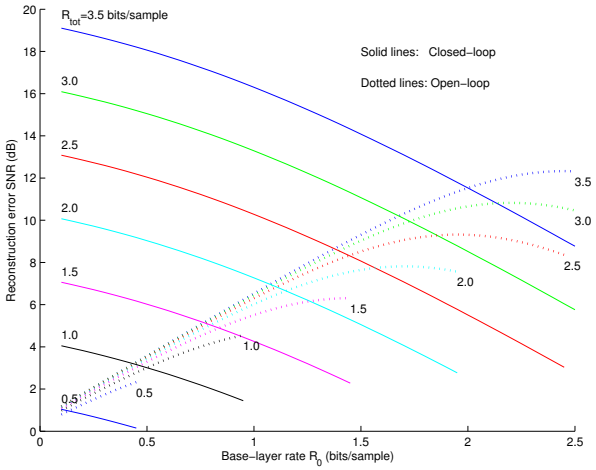

Fig. 3. Closed- and open-loop reconstruction error $\sigma_{\varepsilon_1}^2$ SNR for a set of fixed total rates, where the sources have unit variance and $\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} = 0.75$. The base-layer rate R_0 goes from 0.1 to 2.5 bits/sample.

For a fixed total rate $R_{\text{tot}} = R_0 + R_1$, we can take the derivative of (11) with respect to R_0 to determine the low-resolution rate that yields the best high-resolution performance. The resulting optimal rate and error variance [6] are

$$\begin{aligned} R_{0_{\text{opt}}} &= \tfrac{R_{\text{tot}}}{2} + \tfrac{1}{4}\log_2\left(\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2+2\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}}{\sigma_{y_0}^2+\sigma_{y_1}^2-2\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}}\right) \,, \\ \sigma_{\varepsilon_1}^2\Big|_{R_0=R_{0_{\text{opt}}}} &= \left[\left(\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} + \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right)\right. \\ &\quad \left.\cdot\left(\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} - \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right)\right]^{\frac{1}{2}} 2^{-R_{\text{tot}}} \,. \end{aligned} \tag{12}$$

The base-layer rate $R_{0_{\text{opt}}}$ that maximizes the high-resolution performance is therefore half the total rate plus a constant that depends on the correlation between low and high resolution, when $\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} \in \left[0, \tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2}\right)$. When $\rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} = \tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2}$, the optimal rate is $R_{0_{\text{opt}}} = R_{\text{tot}}$.

Fig. 3 shows the results for both the closed- and open-loop analytical models. For low base-layer rates, the closed-loop system clearly outperforms the open-loop system. As base-layer rate R_0 increases, after a certain point the open-loop system becomes better.

4. COMPARISON BETWEEN OPEN AND CLOSED-LOOP CODERS

We found analytically that the open-loop coder high-resolution performance was better than that of the closed-loop coder when the base-layer rate R_0 was above a certain value. To determine the crossing point we equate the open-loop and

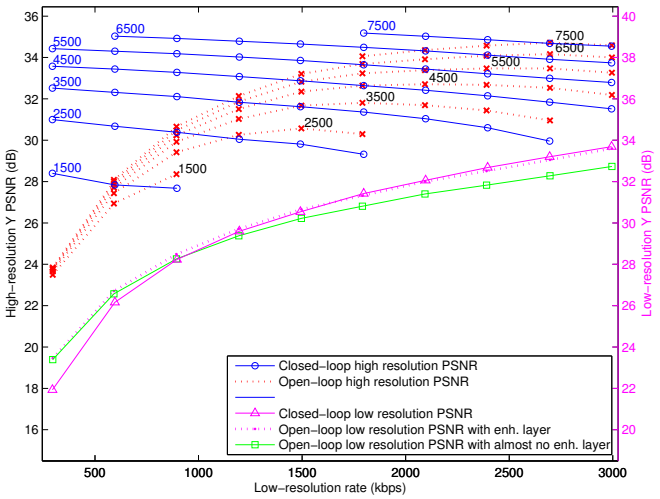

Fig. 4. Open and closed-loop combined PSNR values for *Harbour* (4CIF).

closed-loop reconstruction distortions of (11) and (4). To facilitate the derivation, these equations can be rewritten as

$$\begin{aligned}
\text{Open-loop:} \quad \sigma_{\varepsilon_1}^2 &= a_0 2^{-2R_0} + a_1 2^{-2R_{\text{tot}}} 2^{2R_0} \\
\text{Closed-loop:} \quad \sigma_{\varepsilon_1}^2 &= \sigma_{y_0}^2 2^{-2R_{\text{tot}}} + 4a_1 2^{-2R_{\text{tot}}} 2^{2R_0} \\
\text{where } a_0 &= \tfrac{1}{2}\left(\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} + \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right) \\
a_1 &= \tfrac{1}{2}\left(\tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2} - \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1}\right) \\
& 0 \le \rho_{y_0y_1}\sigma_{y_0}\sigma_{y_1} < \tfrac{\sigma_{y_0}^2+\sigma_{y_1}^2}{2}\,.
\end{aligned} \tag{13}$$

Equating both distortions yields a quadratic equation with respect to 2^{2R_0}, whose real root is

$$\eta = \frac{1}{6a_1}\left[\sqrt{(\sigma_{y_0}^2)^2 + 12a_0a_1\cdot 2^{2R_{\text{tot}}}} - \sigma_{y_0}^2\right]. \tag{14}$$

The rate and distortion at which the closed- and open-loop reconstruction errors cross are therefore

$$\begin{aligned}
R_0 &= \frac{1}{2}\log_2\eta \quad \text{bits/sample}\,, \qquad \eta > 0 \\
\sigma_{\varepsilon_1}^2 &= \left(\sigma_{y_0}^2 + 4a_1\eta\right) 2^{-2R_{\text{tot}}}\,.
\end{aligned} \tag{15}$$

We next show some experimental results for both coders. Both the closed- and open-loop high-resolution performance for the MPEG test clip *Harbour* (4CIF) are combined with the low-resolution results in Fig. 4. The left axis indicates the PSNR for the high-resolution decoded sequence, and the right axis corresponds to low-resolution PSNR. Similar results for three other test clips are in [6].

If we want to view the high-resolution video at a total rate of 3500 Kbps, we can see that the closed-loop coder performs best until the R_0 reaches approximately 1500 Kbps, after which the open-loop system performs better. So suppose that we have a server that operates at this crossover point. If a neighbor on a peer-to-peer or multicast network decides they want to view a higher quality low-resolution sequence, the closed-loop system will suffer a penalty if we use this as the base layer, due to both the mismatch between the closed-loop encoder's embedded decoder and the client decoder, and also due to the decreased performance of the closed-loop coder as R_0 increases. In the open-loop case however, the high-resolution client will benefit from this additional low- resolution rate. As the total base-layer rate becomes too high, both coders are penalized; the closed-loop due to the allocation removed from R_1, and the open-loop due to the excessive allocation in R_0 not improving the high-resolution result. One may conclude that the closed-loop system is only well-matched to low base-layer rates and networks with low heterogeneity in its high-resolution clients.

Fig. 4 also shows the effects of the enhancement-layer bitstream on the low resolution PSNR. As the base-layer rate increases, the open-loop performance matches that of the closed-loop system if all the enhancement layer is included when decoding the low resolution.

5. CONCLUSIONS

We presented an analysis and results for a closed-loop motion-compensated JPEG 2000 coder that used differential prediction of MCTF subbands between resolution levels. We then introduced a new resolution-scalable video coder architecture capable of matching or exceeding closed-loop performance, and we showed that the experimental results behaved as predicted by the analysis.

6. REFERENCES

[1] "ISO/IEC 15444-1 JPEG 2000 Part 1 020719 (Final Publication Draft," ISO/IEC JTC1/SC29 WG1 N2678, July 2002.

[2] Shih-Ta Hsiang and John W. Woods, "Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank," *Signal Processing: Image Communication*, vol. 16, pp. 705–724, 2001.

[3] John W. Woods and Gary Lilienfield, "A resolution and frame-rate scalable subband/wavelet video coder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 9, pp. 1035–1044, September 2001.

[4] D. S. Taubman and M. W. Marcellin, *JPEG2000: image compression fundamentals, standards, and practice*, Kluwer, Massachusetts, 2002.

[5] Uwe Horn, Thomas Wiegand, and Bernd Girod, "Bit allocation methods for closed-loop coding of oversampled pyramid decompositions," in *Proc. International Conference on Image Processing*, October 1997, vol. 2, pp. 17–20.

[6] Robert A. Cohen, *Hierarchical Scalable Motion-Compensated Video Coding*, Ph.D. dissertation, Rensselaer Polytechnic Institute, `http://www.cipr.rpi.edu/ftp_pub/personal/cohen/cohen_phd.pdf`, January 2007.