JOINT SOURCE-CHANNEL DISTORTION MODELLING FOR MPEG-4 VIDEO

Muhammad F. Sabir, Robert W. Heath Jr. and Alan C. Bovik

Department of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, TX 78712-1084, USA.
Email: {mfsabir, rheath, bovik}@ece.utexas.edu

ABSTRACT

Joint source-channel coding is becoming more important for wireless multimedia transmission due to high bandwidth requirements of these multimedia sources. Design of all joint source-channel coding schemes require an estimate of distortion at different source coding rates and under different channel conditions. In this paper, we present one such distortion model for estimating distortion due to quantization and channel errors in a joint manner for MPEG-4 compressed video streams. This model takes into account important aspects of video compression such as transform coding, motion compensation, and variable length coding. Results show that our model estimates distortion within 1 *dB of actual simulation values in terms of peak-signal-to-noise-ratio.*

1. INTRODUCTION

Image and video communication are becoming common over wireless systems with the introduction of high data rates and bandwidth. Transmission of these sources, particularly digital videos with high fidelity require large amounts of bandwidth and high reliability. It is highly anticipated that with the introduction of multiple-input multiple-output (MIMO) systems, and hence higher data rates and better reliability, real-time image and video communication will become one of the major applications of next generation commercial wireless systems. As discussed by many researchers, to optimize the use of available bandwidth and data rate while still maintaining very good quality, it is prudent to use joint designs such as joint source-channel coding (JSCC) [1], joint source coding and transmit power management [2], and other such schemes for coding and transmission of digital images and videos.

In almost all such joint design techniques for digital images and videos, the goal is to minimize distortion with a constraint on a resource such as available data rate, bandwidth, transmission power, latency, etc. To be able to design joint coding schemes for real time image and video communication applications, it is necessary to have an estimate of distortion at different source coding rates and channel coding conditions, that can be used for coding and transmission in real time. Furthermore, the distortion estimation process needs to be computationally non-intensive for the joint design technique to be practical.

Distortion can either be estimated using simulations and operational rate-distortion curves, or it can be obtained using statistical distortion models at different source coding rates and channel conditions. Whereas the simulations based approaches tend to provide closer estimates of distortion, they usually are computationally intensive and hence cannot be used for real-time applications. The model based approaches on the other hand are not very accurate, however they are computationally non intensive while providing good enough estimates of distortion. For this reason, model based distortion estimation schemes are more suited for real time image and video communication applications. Most of the joint design schemes in literature have focused on simulation based design strategies, with not much work being done in the field of developing distortion models for practical image and video coding standards.

In [3], Kim and Kim presented a resource allocation scheme for video transmission. They modelled end-to-end distortion taking into account the effects of error propagation in motion compensated video. This distortion model is then used for allocation of resources in their proposed method. Ruf and Modestino presented a distortion model for discrete wavelet transform (DWT) compressed images in [4], which is then used for efficient joint allocation of source and channel bits. Appadwedula *et al.* derived an expression for the expected value of distortion for a general class of images in [5], and then applied it to different classes of source and channel coders. In our previous work in [6], we presented a joint source-channel distortion model for JPEG compressed images, which is then used for unequal power allocation for JPEG transmission over MIMO systems in [7]. This model estimates distortion for JPEG coded images due to quantization and channel errors at different source coding rates and bit error rates.

In this paper, we present a joint source-channel distortion model for MPEG-4 [8] compressed video streams. This model estimates the amount of distortion due to quantization and channel errors. MPEG-4 error resilient tools such as data partitioning and packetization are used to encode the video into different layers. The expected value of mean squared error (MSE) is then found as a function of source coding rate and channel bit error probability for each layer. The total distortion is the sum of distortions due to individual layers. This model takes into account motion compensation and prediction, discrete cosine transform (DCT) coding, variable length coding (VLC), error propagation, and estimates distortion due to errors in I and P frames. The parameters of our model are computed using a 'training' database of videos. Results show that our model predicts distortion within 1 dB of actual simulation values in terms of peak-signal-to-noise-ratio (PSNR) at all values of source coding rates and channel bit error rates (BER). While the expressions are derived explicitly for MPEG-4 video coded streams, this model can be extended to other similar video coding schemes that use transform coding, motion compensation and entropy coding.

2. SYSTEM MODEL

Due to the presence of entropy coding, motion compensation and predictive coding, the compressed image and video bitstreams are highly sensitive to channel errors. A single bit error can have catas-

This work was supported by the Texas Advanced Technology Program Grant No. $003658 - 0380 - 2003$.

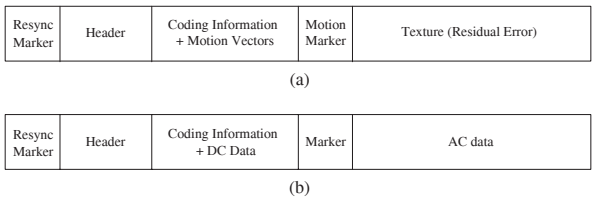


Fig. 1. MPEG-4 packet structure for (a) P frames, and (b) I frames.

trophic effect on the received image or video. Due to this reason, different error resilience tools are introduced in almost all the image and video coding standards. In this paper we use the MPEG-4 part 2 video coding standard with certain error resilience features. These error resilience features include packetization and data partitioning, as explained in Sec. 2.1. We explain our source coding model and channel model in the following sections.

2.1. The Source Coding Model

We use MPEG-4 part 2 (visual) for source coding. All the frames are either coded as I or P frames. I frames are first transformed into DCT coefficients, and then these coefficients are coded using VLC. For P frames, motion estimation and compensation is first performed, and the resulting motion vectors are coded using VLC along with DCT coefficients of residual error (texture). As a single bit error can propagate to different parts of a frame, and in subsequent frames causing large amounts of distortion, therefore, we use two of the error resilient tools that are a part of MPEG-4 video coding standard. These tools are packetization and data partitioning. In packetization, the bitstream is divided into packets, and differential coding, VLC and run-length coding are re-initialized for each packet. This prevents spatial propagation of errors. In data partitioning mode, data is divided into different partitions hence separating more important data from data that is less important. For I frames, DC coefficients and macroblock coding information are coded in a separate partition than AC coefficients within the same packet. For P frames, motion vectors and macroblock coding information are coded in a separate partition than texture DCT coefficients. Note that within a packet, differential coding, VLC and run-length coding are again re-initialized for different partitions. These partitions and packets are separated by uniquely decodeable headers and markers, which we assume to be transmitted error free. Simplified structures of MPEG-4 P and I frame packets are shown in Fig. 1 (a) and (b) respectively.

2.2. The Channel Model

We assume a binary symmetric channel (BSC) to derive the expressions for our distortion model for a given bit error probability. Given the bit error probabilities for any channel (AWGN, Rayleigh fading, etc) and the fact that the probability of making an error from 0 to 1 is the same as that of 1 to 0, that channel can be represented as a BSC. Therefore, the distortion model presented in this paper can be used to find the distortion curves for any channel that can be represented as a BSC, given that the source coding rate and the bit error rate are known. Hence our distortion model is independent of modulation type and channel coding.

3. DISTORTION MODEL FOR MPEG-4

We derive expressions for estimating distortion due to quantization and channel errors in MPEG-4 coded video in this section. MSE is used as our distortion metric. In the following sub-sections, we first outline our assumptions, and then derive MSE expressions separately for I and P frames.

3.1. Assumptions

The goal of our distortion model is to find expressions for MSE in the video sequence as a function of source coding rate and channel bit error probability. To find distortion expressions for practical video coding standards is a complicated task due to the presence of VLC, differential coding, run-length coding, and motion estimation and compensation. Even single bit errors can have catastrophic effects on a video frame and subsequent frames. Due to these reasons, we use error resilient tools in MPEG-4 and make certain simplifying assumptions. We assume that headers and markers are transmitted error free. We also assume that the decoder detects bit errors. These assumptions and their validity are discussed in detail in [6]. Also, we model the DCT coefficients and the pixel values in a packet as random variables.

3.2. Error Concealment

We use a very simple form of error concealment. For the case of I frames, if an error occurs in the DC partition of a packet, all the data in that packet is discarded, and DC and AC coefficients of all the macroblocks in the packet are decoded as zeros. When an error occurs in the AC partition, only the AC coefficients of all the macroblocks in the packet are decoded as zeros. For the case of P frames, when an error occurs in the motion vector partition of a packet, the entire packet is discarded. To conceal this error, pixel values are copied from the previous frame at the exact spatial location. If an error occurs in the texture (residual error) partition of a packet, then that data is decoded as zero, and hence no texture is added to the predicted macroblocks.

3.3. Distortion Model for I Frames

Consider an I frame consisting of J number of packets. Let packet number j consist of K number of macroblocks, and each macroblock contains M number of 8×8 blocks. These blocks are luminance and chrominance blocks. Let $X_{u,m,k,j}$, $X^q_{u,m,k,j}$, $\widehat{X}^q_{u,m,k,j}$ be the unquantized, quantized and erroneous DCT coefficients corresponding to u^{th} subband, m^{th} block, k^{th} macroblock and j^{th} packet. Let L^j_{DC} and L^j_{AC} be the number of bits in the DC and the AC partitions of the packet j respectively, and p^j_e be the probability of bit error for packet j. The probability that at least one bit error occurs in the DC partition of packet j is $p^j_{DC} = 1 - \left(1 - p^j_e\right)^{L^j_{DC}}$. Then, the mean squared error in the entire I frame due to loss of all the DC coefficients in j^{th} packet MSE^j_0 can be written as

$$MSE^j_0 = \frac{1}{N} \sum_{k=1}^{K} \sum_{m=1}^{M} \left(X_{0,m,k,j} - \widehat{X}^q_{0,m,k,j} \right)^2 \cdot p^j_{DC}, \qquad (1)$$

where N is the total number of pixels in the frame. As the erroneous coefficients are decoded as zero, therefore $\widehat{X}^q_{0,m,k,j} = 0$. Also, quantization error and the quantized coefficients can be assumed to be uncorrelated. Hence,

$$MSE^j_0 = \frac{1}{N} \sum_{k=1}^{K} \sum_{m=1}^{M} \left[\left(X^q_{0,m,k,j} \right)^2 + \left(\xi_{0,m,k,j} \right)^2 \right] \cdot p^j_{DC}, \qquad (2)$$

where $\xi_{0,m,k,j}$ is the quantization error. To simplify our notation, let $\sigma^2_{0,j}$ and $\sigma^2_{\xi,0,j}$ denote the sample variance of quantized

DC coefficients and quantization error for packet j respectively; i.e. $\sigma^2_{0,j} = \frac{1}{MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(X^q_{0,m,k,j}\right)^2$ and $\sigma^2_{\xi,0,j} = \frac{1}{MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\xi_{0,m,k,j}\right)^2$, where we have assumed that both the DC coefficients and quantization error have zero mean.

Since a bit error in DC partition also results in the AC partition to be discarded, the distortion contribution due to the loss of AC coefficients, MSE^j_{1-63} (since there are 63 AC coefficients) is

$$MSE^j_{1-63}=\frac{1}{N}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{u=1}^{63}\left(X_{u,m,k,j}-\widehat{X}^q_{u,m,k,j}\right)^2\cdot p^j_{DC}, \quad (3)$$

We denote the sample variance for quantized AC coefficients and the corresponding quantization error for u^{th} subband of j^{th} packet with $\sigma^2_{u,j} = \frac{1}{MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(X^q_{u,m,k,j}\right)^2$ and $\sigma^2_{\xi,u,j} = \frac{1}{MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\xi_{u,m,k,j}\right)^2$ respectively. Hence, the total MSE (MSE^j_{0-63}) in the I frame due to an error in DC partition of j^{th} packet can be written as

$$MSE^j_{0-63} = \frac{(MK-1)}{N}\sum_{u=0}^{63}\left(\sigma^2_{u,j}+\sigma^2_{\xi,u,j}\right)\cdot p^j_{DC}. \quad (4)$$

Due to the presence of prediction and motion compensation, this error will be propagated to subsequent frames till another I frame is encountered. Let T be the number of frames from an I frame to the last P frame before the next I frame, and $p(t)$ be the probability that a frame at distance t from the current frame is affected by an error in the current frame. Then the total MSE per pixel for the block of T frames (1 I and $T-1$ P frames) due to an error in the DC partition of the jth packet can be written as

$$MSE^j_{DC} = \frac{(MK-1)}{NT}\sum_{u=0}^{63}\left(\sigma^2_{u,j}+\sigma^2_{\xi,u,j}\right)\cdot p^j_{DC}\sum_{t=0}^{T-1}p(t). \quad (5)$$

When an error occurs in AC partition of a packet only, and the DC partition is received correctly, then, using similar notation as in the case of an error in DC partition, MSE can be written as

$$MSE^j_{AC} = \frac{(MK-1)}{NT}\left(\sigma^2_{\xi,0,j}+\sum_{u=1}^{63}\left(\sigma^2_{u,j}+\sigma^2_{\xi,u,j}\right)\right)\cdot p^j_{AC} \cdot(1-p^j_{DC})\sum_{t=0}^{T-1}p(t), \quad (6)$$

where $p^j_{AC} = 1-\left(1-p^j_e\right)^{L^j_{AC}}$. Combining (5) and (6), adding the quantization error variance for the case when there is no error, and summing for all the packets in the I frame, the total MSE (MSE_I) per pixel for T frames due to errors in DC and AC coefficients in J packets of the I frame can be written as

$$MSE_I=\sum_{j=1}^{J}\left(MSE^j_{DC}+MSE^j_{AC}+\sum_{u=0}^{63}\sigma^2_{\xi,u,j}(1-p^j_I)\right), \quad (7)$$

where $p^j_I = 1-\left(1-p^j_e\right)^{L^j_{AC}+L^j_{DC}}$.

3.4. Distortion Model for P Frames

For P frames, we have motion vectors (MV) and texture instead of DC and AC coefficients. Though texture information is coded as DCT coefficients, we will model distortion in sample domain to keep things simple. A sample can be either a luma or a chroma value. We will use similar notation as for I frames with slight modifications. Our MSE expressions for P frames will estimate distortion for the case when there is no error propagation due to errors in the DC coefficients of corresponding I frame. Propagation effects of errors in DC coefficients of I frames have already been taken care of in (7). Let $V_{i,m,k,j,n}$, $V^q_{i,m,k,j,n}$ and $\widehat{V}_{i,m,k,j,n}$ be the i^{th} unquantized, quantized and erroneous sample values in m^{th} block of k^{th} macroblock in packet number j of frame number n respectively. Suppose a bit error occurs in the MV partition of j^{th} packet, then both the motion vectors and texture data will be discarded, and the error will be concealed by copying data from previous frame at the exact spatial location. Hence, MSE due to an error in MV partition of packet j of frame number n can be written as

$$MSE^{j,n}_{MV} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{64}\Big(\left(V^q_{i,m,k,j,n}-V^q_{i,m,k,j,n-1}\right)^2 + \xi^2_{i,m,k,j,n}\Big)\cdot p^{j,n}_{MV}p^{'}_{DC}(t)\sum_{t=1}^{T-n}p(t), \quad (8)$$

where $\xi_{i,m,k,j,n}$ is the quantization error, $p^{j,n}_{MV} = 1-\left(1-p^{j,n}_e\right)^{L^{j,n}_{MV}}$, $L^{j,n}_{MV}$ is the number of bits in the MV partition, and $p^{'}_{DC}(t)$ is the probability that the data in current packet is free from propagation error effects of DC coefficients of I frame. Following similar notation as for I frame, let $\sigma^2_{j,n} = \frac{1}{64MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{64}\left(V^q_{i,m,k,j,n}-V^q_{i,m,k,j,n-1}\right)^2$ and $\sigma^2_{\xi,j,n} = \frac{1}{64MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{64}\left(\xi_{i,m,k,j,n}\right)^2$. Then, $MSE^{j,n}_{MV}$ becomes

$$MSE^{j,n}_{MV}=\frac{64MK-1}{NT}\left(\sigma^2_{j,n}+\sigma^2_{\xi,j,n}\right)p^{j,n}_{MV}p^{'}_{DC}(t)\sum_{t=1}^{T-n}p(t). \quad (9)$$

Now consider the case when there is an error in the texture partition, but the MV partition is error free. In this case, the texture (residual error) will be lost, and the predicted pixel values will be displayed. Let $V^{'q}_{i,m,k,j,n}$ be the predicted sample value, $\Delta_{i,m,k,j,n} = V^q_{i,m,k,j,n} - V^{'q}_{i,m,k,j,n}$ be the texture (residual error), $\sigma^2_{\Delta,j,n} = \frac{1}{64MK-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{64}\Delta^2_{i,m,k,j,n}$, and $\sigma^2_{\xi,j,n}$ be the quantization error variance. Then, MSE due to error in texture partition of j^{th} packet of n^{th} frame can be written as

$$MSE^{j,n}_{\Delta} = \frac{64MK-1}{NT}\left(\sigma^2_{\Delta,j,n}+\sigma^2_{\xi,j,n}\right) \cdot p^{j,n}_{\Delta}(1-p^{j,n}_{MV})p^{'}_{DC}(t)\sum_{t=1}^{T-n}p(t). \quad (10)$$

where $p^{j,n}_{\Delta} = 1-\left(1-p^{j,n}_e\right)^{L^{j,n}_{\Delta}}$, and $L^{j,n}_{\Delta}$ is the number of bits in the texture partition of j^{th} packet of n^{th} frame. Hence, by combining (9) and (10), and summing for all the packets, we obtain total MSE per pixel over T video frames due to quantization and channel errors in n^{th} frame:

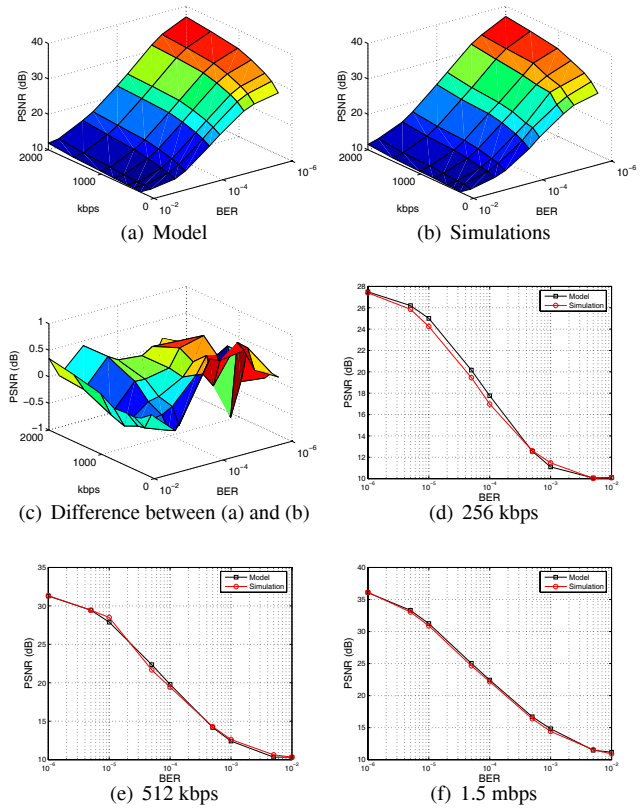
$$MSE^n_P=\sum_{j=1}^{J}\left(MSE^{j,n}_{MV}+MSE^{j,n}_{\Delta}+\sigma^2_{\xi,j,n}\left(1-p^{j,n}_P\right)\right), \quad (11)$$

where $p^{j,n}_P = 1-\left(1-p^{j,n}_e\right)^{L^{j,n}_{MV}+L^{j,n}_{\Delta}}$.

3.5. Total Distortion

The total distortion in the video sequence is the sum of distortions due to errors in both I and P frames. This can be expressed by adding (7) and (11):

$$MSE = MSE_I + \sum_{n=1}^{T}MSE^n_P. \quad (12)$$



(a) Model (b) Simulations

(c) Difference between (a) and (b) (d) 256 kbps

(e) 512 kbps (f) 1.5 mbps

Fig. 2. PSNR vs BER and kbps curves for the model and simulations for 'walk' video sequence.

4. SIMULATIONS AND RESULTS

In this section we discuss our simulation details, and compare our model's prediction of MSE with simulations. We convert MSE to PSNR using the simple relation PSNR $= 10\log_{10}\frac{255^2}{MSE}$, since PSNR is a commonly used metric for video and image quality assessment.

4.1. Simulation Details

A training database of 20 352×288 $4:2:0$ (CIF) format videos with 25 frames per second is used to find the parameters for our model. The parameters $p(t)$ and $p_{DC}^{'}(t)$ are obtained using this training database via simulations. The number of P frames between I frames is varied from 20 to 200. Different source coding rates from 256 kilo bits per second (kbps) to 2 mega bits per second (mbps), and different packet sizes are used to keep the model parameters as generic as possible. These model parameters are then used to find the MSE using our model and actual simulations for different test video sequences. Source coding rates from 256 kbps to 2 mbps are used for the test video sequences, and BER is varied from 10^{-2} to 10^{-6}. It is assumed that the headers and markers are transmitted error free separately.

4.2. Results and Discussion

PSNR values are obtained using our model and simulations for different source coding rates and BERs, as shown in Fig. 2. These results are for a video sequence titled 'walk'. This video sequence consists of 105 frames, with 5 I frames and 20 P frames between the I frames. Packet size of 2000 bits is used, and 200 iterations are performed at each source coding rate and BER. Fig. 2 (a) and (b) show the PSNR curves obtained using our model and simulations (200 iterations) respectively, and Fig. 2 (c) shows their difference. Fig. 2 (d), (e) and (f) show overlapped slices of Fig. 2 (a) and (b) at 256 kbps, 512 kbps and 1.5 mbps respectively.

As can be seen from Fig. 2, our model predicts PSNR within 1 dB of the actual simulation values at all source coding rates and BERs. Similar PSNR curves are also obtained for 15 other test sequences with different combinations of I and P frames and different packet sizes, however, results for only one video sequence are shown here due to lack of space.

5. CONCLUSION

In this paper, we presented a model for estimating the distortion introduced in MPEG-4 compressed video stream due to quantization and channel errors. This model takes into account the effects of motion estimation and prediction, transform coding, and entropy coding, and uses different error resilience tools of MPEG-4 video coding standard. Simulation results show that the PSNR predicted by our model is accurate within 1 dB of the actual PSNR values obtained via simulations. Though this model is fine tuned for MPEG-4, it can be used for any video coding scheme that uses motion compensation, transform coding and entropy coding with slight modifications. This model can be used to design efficient joint source-channel coding and unequal error protection schemes for real time video communication applications. We are currently working on designing such schemes for wireless video communication.

6. REFERENCES

[1] H. Gharavi and S. M. Alamouti, "Multipriority video transmission for third-generation wireless communication systems," *Proceedings of the IEEE*, vol. 87, pp. 1751–1763, Oct. 1999.

[2] Y. Eisenberg, C. Luna, T. Pappas, R. Berry, and A. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 411–424, Jun 2002.

[3] I.-M. Kim and H.-M. Kim, "A new resource allocation scheme based on a PSNR criterion forwireless video transmission to stationary receivers over Gaussian channels," *Wireless Communications, IEEE Transactions on*, vol. 1, no. 3, pp. 393–401, 2002.

[4] M. J. Ruf and J. W. Modestino, "Operational rate-distortion performance for joint source and channel coding of images," *IEEE Trans. Image Processing*, vol. 8, pp. 305–320, Mar 1999.

[5] S. Appadwedula, D. Jones, K. Ramchandran, and L. Qian, "Joint source channel matching for wireless image transmission," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, vol. 2, pp. 137–141 vol.2, 1998.

[6] M. F. Sabir, H. R. Sheikh, R. W. Heath Jr. and A. C. Bovik, "A joint source-channel distortion model for JPEG comrpessed images," *accepted for publication in IEEE Trans. Image Processing*.

[7] M. F. Sabir, R. W. Heath, and A. C. Bovik, "An unequal power allocation scheme for JPEG transmission over MIMO systems," *To appear in proceedings of Asilomar Conference on Signals, Systems and Computers, 2005*, 2005.

[8] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression, Video Coding for Next-generation Multimedia*. Wiley, 2003.