MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Rate-Distortion Models for Video Transcoding

Peng Yin Anthony Vetro Minghui Xia Bede Liu

TR-2003-28 January 2003

Abstract

This paper addresses the problem of selecting a single transcoding method from multiple transcoding possibilities that satisfy a specified delivery constraint. We first discuss the rate-distortion modeling of DCT coefficients. Then, we quantify the distortion of the video transcoding output by examining the rate-distortion relationship of the popular transcoding techniques, including requantization, spatial downsampling and temporal downsampling. The use of our model is illustrated in some typical applications.

Published in SPIE Conf. on Image and Video Communications and Processing, Jan 2003.

Rate-Distortion Models for Video Transcoding

Peng Yin[a], Anthony Vetro[b], Minghui Xia[c] and Bede Liu[c]

[a]Corporate Research, Thomson Multimedia Inc., Princeton, NJ
[b]MERL - Mitsubishi Electric Research Laboratories, Murray Hill, NJ
[c]Princeton University, Department of Electrical Engineering, Princeton, NJ

ABSTRACT

This paper addresses the problem of selecting a single transcoding method from multiple transcoding possibilities that satisfy a specified delivery constraint. We first discuss the rate-distortion modeling of DCT coefficients. Then, we quantify the distortion of the video transcoding output by examining the rate-distortion relationship of the popular transcoding techniques, including requantization, spatial downsampling and temporal downsampling. The use of our model is illustrated in some typical applications.

Keywords: T ranscoding, Rate-distorion, Requantization, Spatial/Temporal downsampling

1. INTRODUCTION AND PREVIOUS WORK

A number of approaches have been proposed for video transcoding, including requantization, temporal downsampling (frame skipping), spatial downsampling and spatio-temporal downsampling. Given all these methods, there are often several ways that the video can be adapted to satisfy a specified delivery constraint. The delivery constraint itself may be a function of several variables, including the current network condition and the capabilities of a terminal. In this paper, we investigate techniques that can be use to adaptively select the best transcoding method under the current constraints.

Most previous attempts to address this issue have been based on an ad-hoc model approaches.[1] Here, we propose to model the rate-distortion of the video transcoding output analytically. This model will allow us to quantify the distortion produced by various transcoding methods. Under certain fixed constraints, the best transcoding method can then be determined. The transcoding methods we consider here are the most commonly used ones given above. For temporal downsampling, we consider a fixed rate output only, however this approach can also be extended to dynamic frame skipping.

We shall focus on the modeling of textures in this paper by examining the explicit relationship between bit rate and distortion. We assume a block-based DCT coding scheme, which is commonly used in many current coding methods. In order to save computation, it is typically preferable to avoid full decoding and performs most operations directly in the transform domain. Therefore, the proposed modeling is based on DCT coefficients as input.

The rest of the paper is organized as follows. We first discuss the rate-distortion modeling of DCT coefficients in Section 2. The estimation of the distortion due to requantization, spatial downsampling, temporal downsampling and spatio-temporal downsampling will be given in Section 3, where the discussion is focused on intra-frames only. Finally, a discussion of how to use these models in an application is given in Section 4 and the work is summarized in Section 5.

2. RATE-DISTORTION MODELING OF DCT COEFFICIENTS

For a memoryless zero-mean Gaussian source $N(0, \sigma_z^2)$, the rate-distortion bound is given by[2]

$$R(D)_G = max\{0, \frac{1}{2} log_2 \frac{\sigma_z^2}{D}\}, \tag{1}$$

and

$$D(R)_G = 2^{-2R}\sigma_z^2.$$

This work is partly supported by a NJ State R&D Excellence Grant. Contact info: yinp@tce.com, avetro@merl.com, {mxia,liu}@ee.princeton.edu

where bit rate, R, is bits per pixel and distortion, D, is Mean Square Error(MSE).

For a memoryless non-Gaussian source, the distortion bound can be approximated by[2]

$$D(R) = a \cdot 2^{-2R} \cdot \sigma_z^2, \qquad 0 < a \leq 1, \tag{2}$$

Equations 1 and 2 provide a rate-distortion bound for a memoryless source. We shall use them to approximate the R-D relationship of AC coefficients.

To maintain a certain perceptual effect, the quantization step size of DC coefficients is always fixed and is usually very small. We can approximate $D_{0,0}$ by[2]

$$D_{0,0} = \frac{Q_{0,0}^2}{12}. \tag{3}$$

Since the DCT decorrelates a block of data, the DCT coefficients can be treated independently.[3] So

$$R = \frac{1}{N}(r_0 + \sum_{k=1}^{N-1} r_k), \tag{4}$$

where N is the number of coefficients in a block and r_k is the number of bits for the kth DCT coefficient. It should be noted that r_0 can be directly estimated from the original coded frame. For transcoding operations that maintain the same spatial resolution, the DC coefficients are unchanged; for transcoding operations that reduce the spatial resolution by a factor of 2 in each dimension, we have found that the required bits for DC coefficients is about $1/3 - 1/2.5$ that of original. From equations 1 and 2, we can deduce the following equations:

$$\begin{aligned} r_k &= max\{0, \frac{1}{2} log_2 \frac{\sigma_k^2}{\theta}\}, \; k \neq 0, \\ R &= \frac{1}{N}(r_0 + \sum_{k=1}^{N-1} max\{0, \frac{1}{2} log_2 \frac{\sigma_k^2}{\theta}\}), \\ D &= a \cdot \frac{1}{N}(D_{0,0} + \sum_{k=1}^{N-1} min\{\theta, \sigma_k^2\}), \end{aligned} \tag{5}$$

where the parameter θ is a threshold value. Those coefficients whose mean square error falls below θ are not coded. θ is determined by the targeted value of R, if it is known, or the targeted value of D, if it is known.

3. VIDEO TRANSCODING MODELING

Video transcoding is a second generation of the bit stream. The statistics of the transcoded bit stream can be derived from the first generation of the bit stream. For requantization, the model can be derived from the DCT coefficients modeling discussed in section 2. For downsampling (spatial and/or temporal), we will divide the transcoding operation into two approximately independent components to simplify the analysis. The first is requantization noise, and the second is downsampling error. The simulation is done for several sequences. However, due to space limitations, we only show the result for *Akiyo* and *Foreman* with 300 frames in each case. These results are representative of others sequences.

3.1. Requantization

We assume that the result of coding a video by first encoded with Q_1 and then transcoded with Q_2 is almost identical to that of coding the original raw video with Q_2. This is a reasonable approximation when the original video is coded with high quality, which is usually the case. In this case, the error d caused by requantization can be approximated as

$$d = X - \tilde{\tilde{X}} \approx \tilde{X} - \tilde{\tilde{X}}. \tag{6}$$

where X is raw bit stream, $\tilde{X}$ represents the original bit stream, and $\tilde{\tilde{X}}$ represents the transcoded bit stream. The distortion D_r can then be computed as in Section 2 using Equation 5. A more detailed analysis about the additional distortion caused by requantization can found in.[4]

To confirm this, we use fixed quantizers to code the first and second generation of a video. The first generation is coded with *qscale=3*, while the second generation is coded with *qscale=8* for the first 100 frames, *qscale=15* for the second 100 frames and *qscale=25* for the last 100 frame. Figure 1 shows the result of distortion of each frame. The distortion is estimated from the rate by Equation 5, where $a = 0.5$ for *Akiyo* and $a = 1$ for *Foreman*. In the figure, MSE_CODED denotes the true MSE for second generated video against first generated video, while $MSE\_ORIGINAL$ denotes MSE against original raw video and $MSE\_ESITMATE$ denotes the estimated result D_r. The figure shows that MSE_CODED is quite close to $MSE\_ORIGINAL$, except for the first 100 frames where the quantization is fine. We may conclude from these plots that the approximation of Equation 6 is useful, except for fine quantization.

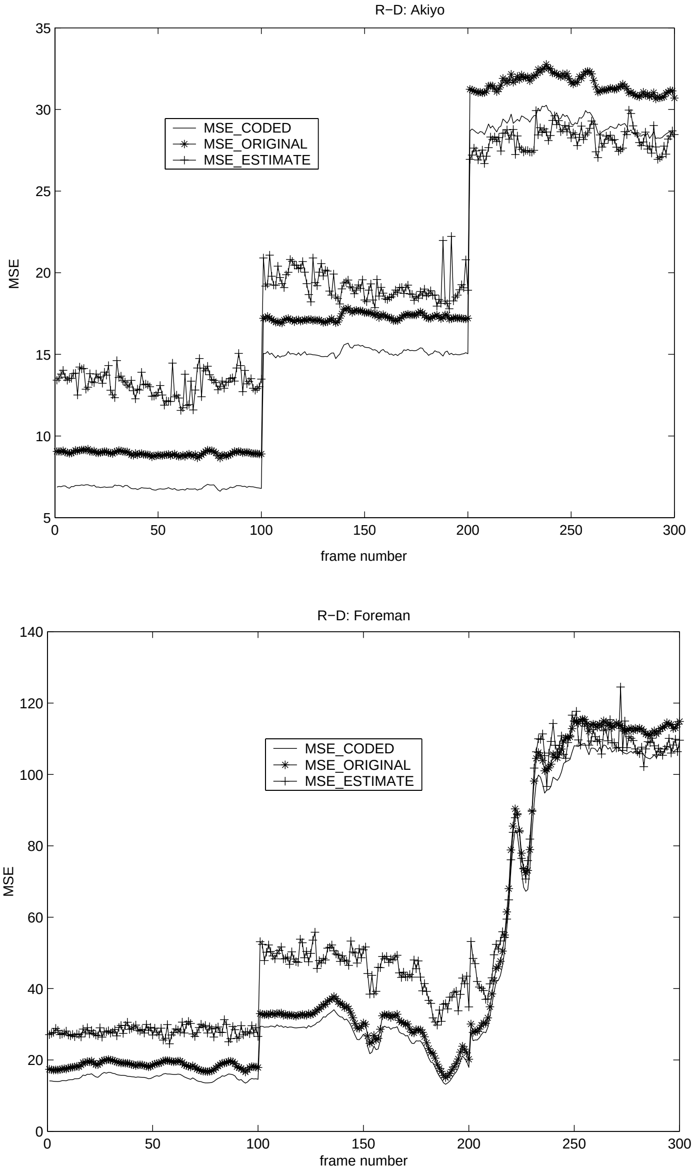

Figure 1. Experimental Result of Rate-Distortion for Requantization

3.2. Spatial Downsampling

For a fair comparison, we need to compare the spatial downsampled transcoded video against the original video. To do so, we need to restore the picture size by spatially upsampling the transcoded video. Here we only discuss the case of reducing the picture size by a factor of 2 in each dimension.

We use $\mathcal{D}$ to denote the down-sampling operation, $\mathcal{U}$ the up-sampling operation, variable X the full-sized image, and Y the reduced-sized image. The distortion incurred by downsampling is then

$$\begin{aligned} d &= X - \tilde{\tilde{X}} = X - \mathcal{U}(\tilde{\tilde{Y}}) \\ &= X - \mathcal{U}(Y) + \mathcal{U}(Y) - \mathcal{U}(\tilde{\tilde{Y}}) \\ &\approx (X - \mathcal{U}(Y)) + \mathcal{U}(\tilde{Y}) - \mathcal{U}(\tilde{\tilde{Y}})) \\ &= (X - \mathcal{U}(\mathcal{D}(X))) + \mathcal{U}(\tilde{Y} - \tilde{\tilde{Y}}) = d_{du} + d_{qu} \end{aligned} \tag{7}$$

The distortion can be broken into two parts: one due to down-and-upsampling d_{du} and the other due to quantization at the reduced-resolution d_{qu}.

If we assume these two quantities are independent, the downsampling distortion D can be approximated by

$$D_d = D_{du} + D_{qu} \tag{8}$$

To increase the speed of transcoding, the downsampling is performed in transform domain by operating on rows and columns of the matrix consecutively. $\mathcal{D}$ is implemented as

$$S = A\,P\,A^T \tag{9}$$

where S is the reduced block of size 8×8, P is the full-sized block of size 16×16 and A represents the downsampling filter with dimension 8×16.

The corresponding upsampling is also performed in the transform domain. Thus, $\mathcal{U}$ is implemented as

$$\overline{P} = B\,S\,B^T, \tag{10}$$

where $\overline{P}$ is the upsampled block of size 16×16 and B represents the upsampling filter with dimension 16×8.

From Equations 9 and 10, we have

$$\overline{P} \;=\; B\,A\,P\,A^T\,B^T = (BA)\,P\,(BA)^T = C\,P\,C^T, \tag{11}$$

where C represents the successive downsampling and upsampling filtering operation with a matrix dimension of 16×16.

To simplify the analysis, we write the above two-dimensional transformation as a one-dimensional transformation by making use of their Kronecker product separability. Denoting the Kronecker product by $\otimes$, and mapping the matrices, $S, P, \overline{P}$, into row-ordered vectors, $s, p, \overline{p}$, we have,

$$s = \mathcal{A}\,p,\; \overline{p} = \mathcal{B}\,s,\; \overline{p} = \mathcal{C}\,p, \tag{12}$$

where

$$\mathcal{A} = A \otimes A,\; \mathcal{B} = B \otimes B,\; \mathcal{C} = C \otimes C, \tag{13}$$

Using the above notation, we can deduce that

$$\begin{aligned} D_{du} &= \frac{1}{256} E(||p - \overline{p}||^2) \\ &= \frac{1}{256} E\{tr[(p - \overline{p})(p - \overline{p})^T]\} \\ &= \frac{1}{256} tr[((\mathcal{I} - \mathcal{C})^{\mathcal{T}}(\mathcal{I} - \mathcal{C}))E(pp^T)] \\ &= \frac{1}{256} tr[\mathcal{F}\mathbf{R_p}], \end{aligned} \tag{14}$$

where $\mathcal{F}$ denotes the distortion matrix for down-and-upsampling, $\mathbf{R}$ is the correlation matrix, and $\mathbf{R_p}$ is approximated by $\mathbf{R}_{\tilde{\mathbf{p}}}$ in real application.

In the same way, we can arrive at

$$\begin{aligned} D_{qu} &= \frac{1}{256}\,E(||\tilde{p} - \tilde{\tilde{p}}||^2) \\ &= \frac{1}{256}\,E(||\mathcal{B}(\tilde{s} - \tilde{\tilde{s}})||^2) \\ &= \frac{1}{256}\,E(||\mathcal{B}q||^2) \\ &= \frac{1}{256}\,tr[(\mathcal{B}^{\mathcal{T}}\mathcal{B})E(qq^T)] \\ &= \frac{1}{256}\,tr[\mathcal{H}\mathbf{R_q}], \end{aligned} \tag{15}$$

where q represents the error caused by quantization in reduced image, so R_q is a diagonal matrix whose elements are quantization distortion. $\mathcal{H}$ denotes the distortion matrix for upsampling the quantization noise.

We notice that because q is actually the quantization error whose distortion can be deduced as in Section 2 by Equation 5, we need to compute the variance of DCT coefficients of reduced-resolution image. We can approximate it from full-resolution image

$$var(\tilde{s}) \quad = \quad \Lambda(\mathbf{R}_{\tilde{s}}) = \mathbf{\Lambda}(\mathcal{A}\mathbf{R}_{\tilde{\mathbf{p}}}\mathcal{A}^{\mathcal{T}}) \tag{16}$$

where var is the variance and Λ denotes the diagonal element of the matrix. We see that only $\mathbf{R}_{\tilde{\mathbf{p}}}$ needs to be calculated.

The model was tested with a simulation under the same condition as 3.1. The downsampling and upsampling filters proposed in[5] were used. The resulting $\mathcal{H}$ is a diagonal matrix with each diagonal element equal to 4, so D_{qu} equals to the quantization error in the reduced resolution. Figure 2 shows the result of distortion on each frame. In the figure, "*actual*" is the true MSE for the second generated video against original raw video and "*estimate*" is the estimated result using Equation 8, where D_{du} and D_{qu} are calculated using equations 14 and 15. From this figure, we can see that the estimated distortion is close to the true result.

3.3. Temporal Downsampling

For fair comparison among different transcoding methods involving temporal downsampling, we need to compare not only the coded frames, but also the skipped (non-coded) frames against the original video. Such a model has been investigated in.[6] For completeness, this work is briefly summarize here. We then discuss a different method for the estimation of non-coded frame distortion.

The distortion for coded frames denoted by D_c are just requantization error, which has the relationship that $D_c = D_r$. For skipped (non-coded) frames, the distortion D_n is broken into two parts: one due to the requantization of the reference D_r and the other due to the interpolation D_i.[6] Let $\hat{z}_k$ denote the interpolated frame, and $\hat{z}_i$ the last transcoded frame. We interpolate the skipped frame by simply repeating the last transcoded frame. Thus

$$\begin{aligned} d_n &= z_k - \tilde{\tilde{z}}_k \;=\; z_k - \tilde{\tilde{z}}_i \\ &= (z_k - z_i) + (z_i - \tilde{\tilde{z}}_i) \;=\; d_i + d_r. \end{aligned} \tag{17}$$

By assuming that these two parts of distortion are independent, we have

$$D_n = D_r + D_i. \tag{18}$$

D_r is the same as requantization error, and D_i is interpolation error. D_i can be approximated as in.[6] One potential drawback to the model introduced in this work is that it does not work well for high-motion sequences since it is derived based on a first-order Taylor expansion, which assumes small motion. Subjective factors aside, it was argued that in the case of higher motion, one would not want to skip frames anyway due to the large MSE that would be incurred, so the accuracy of the model for high motion sequences is less significant in this case.

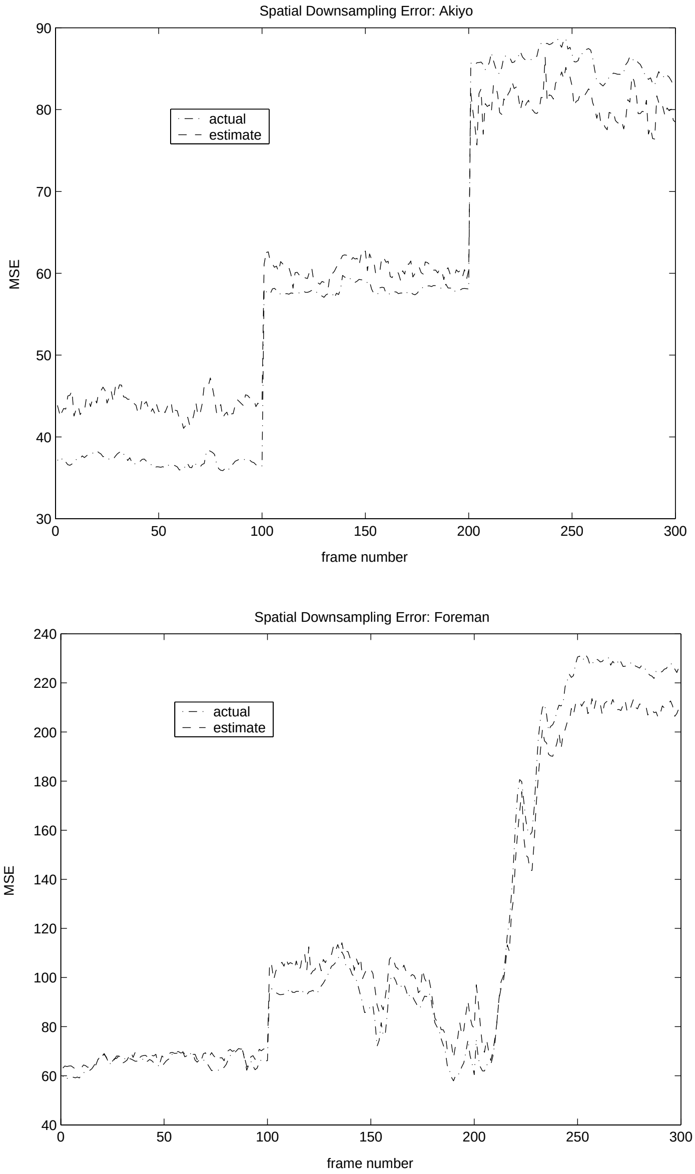

Figure 2. Experimental Result for Spatial Downsampling

An alternative model can be constructed by assuming that the images in the same scene are wide-sense-stationary. By taking squares and expectations of d_i, we have

$$D_i = 2\,(1-\rho)\,\sigma_z^2, \tag{19}$$

where ρ is the correlation coefficient between last coded reference and skipped frame.[2]

For simplification, we assume ρ is related to the frame distance m in a monotonic relation by

$$\rho(m) = \rho_0^m, \tag{20}$$

or

$$\rho(m) = exp\{\alpha m\}, \tag{21}$$

or

$$\rho(m) = \rho_0 exp\{\alpha m\}. \tag{22}$$

The parameters in Equation 19 can be deduced in transform domain.[2] If we write DCT transformation in the form of Kronecker product as

$$\phi = \mathcal{W} z, \tag{23}$$

$$z = \mathcal{W}^{\mathcal{T}} \phi, \tag{24}$$

where $\mathcal{W}$ denotes the DCT transformation, z is the spatial signal and ϕ is the DCT coefficient. Then

$$\begin{aligned} \frac{1}{N}\sum_{n=0}^{N-1}\sigma_n^2 &= \frac{1}{N}\sum_{n=0}^{N-1}E(\phi_n^2) = \frac{1}{N}\sum_{n=0}^{N-1}E(\phi^T\phi) \qquad (25) \\ &= \frac{1}{N}\sum_{n=0}^{N-1}E(z^T\mathcal{W}^{\mathcal{T}}\mathcal{W}z) \\ &= \frac{1}{N}\sum_{n=0}^{N-1}E(z^Tz) = \frac{1}{N}\sum_{n=0}^{N-1}\sigma_z^2 = \sigma_z^2 \end{aligned}$$

In the same way, we can calculate ρ as

$$\rho = \frac{cov(z_k z_i)}{\sigma_z^2} = \frac{\frac{1}{N}\sum_{n=0}^{N-1} cov(\phi_{k,n}\phi_{i,n})}{\sigma_z^2} \tag{26}$$

The advantage of Equation 19 over[6] is that all the parameter can be estimated in the transform domain, instead of needing to compute the gradients in spatial pixel domain. The disadvantage is that on frame-basis Equation 19 can not track D_i as accurately as[6] does and data training is needed to estimate the parameters for $\rho(m)$.

Simulations were carried out using 150 frames of *Akiyo* and *Forman* in the same scene. Fixed quantizers are used to code the first and second generation of video. The first generation is coded with *qscale=3*, while the second generation is coded with *qscale=8* for the first 50 frames, *qscale=15* for the second 50 frames and *qscale=25* for the last 50 frames. We define "$frame\,skip\,number$ = 2" as the distance between two transcoded frames is 2. Figure 3 shows the result of average distortion of all frames over various frame skipping number based on Equation 19. In the figure, "actual" denotes the true MSE for the second generated video against original raw video, and "model1", "model2" and "model3" use Equation 20 to 22. The parameters for the model are trained based on the first 30 frames. These experiments shows that Equation 19 can approximate the distortion quite well.

3.4. Spatio-temporal Downsampling

The distortion for spatio-temporal downsampling is the combination of that of spatial downsampling and temporal downsampling. From the previous discussion, we have $D_c = D_d$ for coded frames. and

$$D_n = D_d + D_i, \tag{27}$$

for skipped frames, where D_d is estimated by Equation 8 and D_i is estimated by Equation 19.

Simulations were carried out in the same condition as in 3.3. Figure 4 shows the result of average distortion of all frames over various frame skipping number coupled with spatial downsampling. The experiment shows that Equation 27 can approximate the distortion quite well.

4. APPLICATIONS

The video transcoding models studied in this paper can be used in many applications. We now give an example of using the models to solve the problem discussed in Section 1, i.e., the selection of the best transcoding method which minimizes D among the candidates, for a given R.

We denote four candidate transcoding methods as $\mathcal{T} = \{\mathcal{RQ}, \mathcal{SD}, \mathcal{TD}, \mathcal{STD}\}$, which corresponds to requantization, spatial downsampling, temporal downsampling and spatio-temporal downsampling, respectively. Suppose the original frame rate is 30 fps, and the set of transcoded frame rates include $\{30, 15, 10, 7.5, 5\}$. The procedure is as follows:

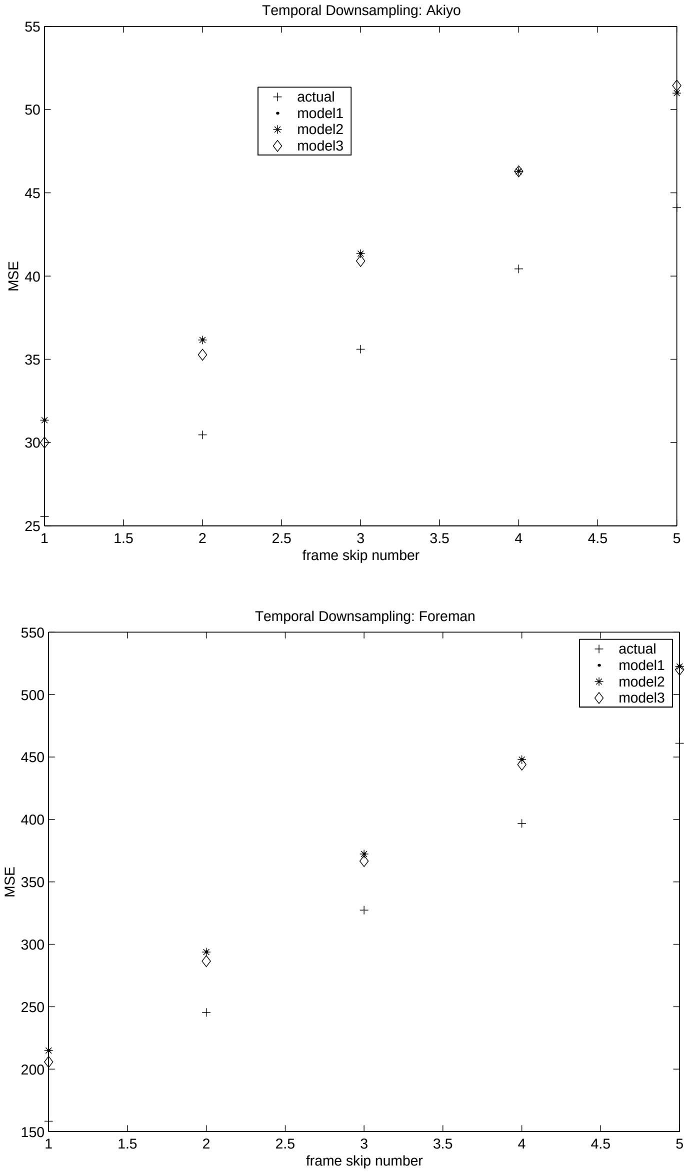

Figure 3. Experimental Result for Temporal Downsampling

- select possible transcoding methods from $\mathcal{T}$ which can satisfy the bandwidth constraint to form a new set $\mathcal{T}'$;
- compute distortion of transcoding methods in the set $\mathcal{T}'$ using video transcoding models;
- select transcoding method in the set $\mathcal{T}'$ which minimizes the distortion.

The first step is done by setting $qscale = 31$ and training of first $10 - 30$ frames, while the second step is computed by video transcoding modeling above. To save time, the model is computed for a scene, because the statistics within the same scene are similar. Fast algorithms for temporal segmentation can also be used.

From the simulation result in Section 3, we observe that for some sequences, the objective quality of spatial downsampling $\mathcal{SD}$ is always better than that of temporal downsampling $\mathcal{TD}$. But perceptually, human eyes can tolerate the video quality with a lower frame rate; also, temporal downsampling is much easier to implement. This implies that objective quality should not be the only criteria in selecting the optimal transcoding method. A better

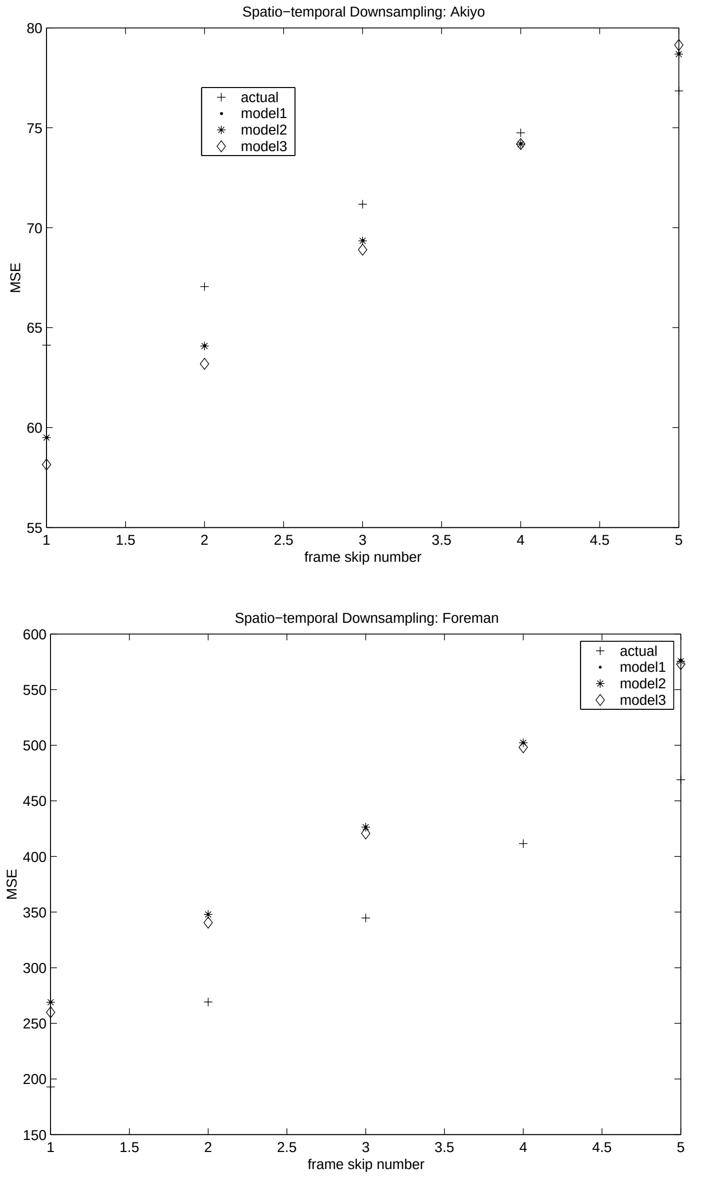

Figure 4. Experimental Result for Spatio-temporal Downsampling

objective function should include a measure of transcoding complexity for instances when the resources of a server are constrained. Additionally, the distortion should represent a subjective assessment of the video quality to account for differences in rendering the transcoded video at different spatio-temporal scales. However, the solution to this problem would be quite involved and is beyond the scope of this paper. We hope though that this work provides a good framework for further study. It is an important area for those working in the area of multimedia signal processing to account for human factors in the design and development of multimedia systems and applications.

5. SUMMARY

In this paper, we first discussed the *rate-distortion* modeling of DCT coefficients. Then, an analytic model was proposed to quantify the intra-frame distortion caused by various transcoding methods. This enabled us to compare the distortion resulting from different transcoding methods, which allows for one to consider an appropriate transcoding strategy among multiple candidate methods.

REFERENCES

1. F.C.M. Martins, W. Ding and E. Feig, “Joint Control of Spatial Quantization and Temporal Sampling for Very Low Bit Rate Video,”, *ICIP*, pp. 2071-2075, 1996.
2. N.S. Jayant, P. Noll, “Digital Coding of Waveforms,” Prentice-Hall Inc., 1984.
3. A. K. Jain, “Fundamental of Digital Image Processing,” Prentice Hall Inc., 1989.
4. O. Werner, “Requantization for Transcoding of MPEG-2 Intraframes,” *IEEE Trans. On Image Processing*, pp. 179-191, Feb., 1999
5. P. Yin, A. Vetro, H. Sun and B. Liu, “Drift Compensation Architectures and Techniques for Reduced Video Transcoding”, *SPIE*, vol. 4076, pp. 180-191, 2002
6. A. Vetro, “Object-based Encoding and Transcoding”, *Ph.D Thesis*, Polytechnic University, New York, 2001