JOINT SPATIO-SNR -TEMPORAL RATE CONTROL AND DISTORTION MODELING FOR QUALITY TRADEOFFS IN VIDEO TRANSCODING

Yong Ju Jung and Yong Man Ro

Multimedia Group, Information and Communications University (ICU)
Yusong, Daejeon, PoBox. 77, 305-732, Korea

ABSTRACT

In this paper, we present the multidimensional transcoding for MPEG compressed bitstream, specifically, a practical joint control of spatial, temporal, and SNR scaling by rate-distortion modeling. The objective of the joint control is to determine an optimal combination of transcoding operations. To allocate bits more efficiently, the multidimensional factors including spatial, SNR, and temporal and their tradeoffs are exploited in terms of the quality of content. We analyzed and solved the typical problem in video transcoding by rate-distortion (R-D). Based on the statistical analysis of dependency distortion, we propose a new distortion model and rate control algorithm.

1. INTRODUCTION

Content adaptation is an important technique for maximizing content accessibility, *e.g.*, it could provide pervasive and interactive functionalities. Since various networks might have different bandwidths, a gateway could include a transcoder to adapt the video bit rates in order to provide consistent video services to users. Namely content transcoding allows the multimedia content to be adapted to the wide diversity of client device capabilities in communication, processing, and display as well. Users could consume as best quality as the contents are available.

Typically, in the decision engine for the content adaptation, one should take into account information about multimedia content, network characteristics, device capabilities and user preferences. Its final objective is to provide the best presentation to user given a certain set of constraints. Basically, decision-making process is related with two following questions:
1. When should the content be adapted? 2. How should the content be adapted?

The motivation of our research is to find an optimal transcoding strategy for best perception with given content and resource constraints, *i.e.* to find a possible set of transcoding operators to meet the bit rate constrained by network or terminal for best quality of the adapted content.

Furthermore, in order to allocate bits more efficiently, the multidimensional, *e.g.*, spatio-SNR-temporal, tradeoffs should be exploited. For example, in case of reducing bitrate, it should be decided to transmit either more content with lower SNR quality or less content with higher quality [3] or either more content with smaller picture size or less content with larger one. Various tradeoffs should be considered to jointly control the spatial, temporal, and SNR scaling.

The paper is organized as follows: In Section 2, we describe briefly a review of video transcoding and formulize the problem. In Section 3, we discuss distortion modeling and a rate control algorithm based on R-D modeling. In Section 4, we show experimental results and the effectiveness of the proposed modeling. Finally, we draw conclusions in Section 5.

2. A REVIEW OF MULTIDIMENSIONAL VIDEO TRANSCODING

The objective of the proposed joint control is to determine an optimal combination of transcoding operations. So far, there are a few analytical approaches based on R-D optimization. Most previous works focused on either bit allocation over coded frames under a constant frame rate or frame rate control at a fixed spatial resolution. They did not deal with the decision considering fully spatio-temporal-SNR tradeoffs.

The tradeoff between spatial and temporal quality has been studied in [1], where the tradeoff was achieved with a simple parametric model. In [2], a multi-dimensional bit rate control for selecting video coding parameter was studied. However, they did not focus on model based rate control but operational one. Also in [3], authors dealt with the joint control of SNR-temporal factors by simply adding one more constraint to the conventional rate-quantizer (R-Q) model. In [4], Yin, et al. attempted to model distortion for multidimensional transcoding. However, they did not consider dependency among transcoding operations. So, we improve their approach.

In this paper, we focus on the multidimensional transcoding, specifically, practical joint control of spatial, temporal, and SNR scaling by rate-distortion modeling. In the following subsection, we formalize a general multidimensional transcoding problem.

2.1. Problem formulation

Before formulating the optimization problem, let's define some notations.

Then, the R-D optimization problem is to find optimum operation set, $\{\boldsymbol{RQ}^*, \boldsymbol{TS}^*, \boldsymbol{SS}^*\}$, so that the average distortion of the transcoded content is minimized.

$$\{\boldsymbol{RQ}^*, \boldsymbol{TS}^*, \boldsymbol{SS}^*\} = \arg \min_{RQ,TS,SS} \sum_{i=1}^{N} D_i(\boldsymbol{RQ}, \boldsymbol{TS}, \boldsymbol{SS}),$$

$$\text{subject to } \sum_{i=1}^{N} R_i(\boldsymbol{RQ}, \boldsymbol{TS}, \boldsymbol{SS}) < R_{\max}. \quad (1)$$

$\boldsymbol{RQ}=[RQ_1, RQ_2, \cdots, RQ_N], RQ_i \in [RQ_{\min}, RQ_{\max}],\ i=1,...,N,$ where RQ_i is the requantization parameter of i^{th} frame and N is the total number of frames in a temporal segment(e.g., scene).

$\boldsymbol{TS}=[TS_1, TS_2, \cdots, TS_N], TS_i \in \{0,1\},\ i=1,....,N,$ where $\boldsymbol{TS}$ is a set of temporal scaling operation and 1 means no-skipping and 0 does frame skipping.

$\boldsymbol{SS}=[SS_1, SS_2, \cdots, SS_N], SS_i \in (0,1],\ i=1,...,N,$ where $\boldsymbol{SS}$ is a set of spatial scaling operation and 1 means the original video and 0.5 means half spatial downscaling video.

s : original frame (non-coded).
$\tilde{s}$: coded frame.
$\hat{s}$: transcoded frame.
e_i : estimation error in time t_i ($= \tilde{s}_i - \hat{s}_i$).
D_i : distortion of i^{th} frame.
R_i : bitrate of i^{th} frame.
d_{SS_i} : distortion of i^{th} frame due to spatial interpolation.
d_{TS_i} : distortion of i^{th} frame due to temporal interpolation.
d_{RQ_i} : distortion of i^{th} frame due to requantization.

To simplify Eq. (1), we can consider two cases, coded frame and dropped frame. Similar to [5] and [6], the total distortion is defined as the sum of coded frame distortion and dropped frame distortion, *i.e.*

$$\sum_{i=1}^{N} D_i(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS}) = \sum_{i=1}^{N} \{D_i \mid (TS_i = 1) + D_i \mid (TS_i = 0)\}, \quad (2)$$

and the bit rate is $\sum_{i=1}^{N} R_i(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS}) = \sum_{i=1}^{N} \{R_i \mid (TS_i = 1)\}$. (3)

3. JOINT CONTROL FOR HYBRID TRANSCODING

Our approach is R-D model based approach motivated by practical issue. The main objective of R-D model is to evaluate the target bitrate and distortion before performing the actual transcoding process [4].

3.1. Rate-distortion model

In conventional R-D modeling, there are various models obtained by considering the statistical characteristics of source [7, 8]. For Laplacian source, quadratic R-D model was adopted in MPEG-4 [8]. The average number of bits per coded sample is

$$R(D) = \ln\left(\frac{1}{\alpha D}\right) \cong aD^{-1} + bD^{-2}. \quad (4)$$

In addition, since we deal with multi-dimension, not only SNR, we need to analyze and model the distortion of ***TS*** and ***SS*** operations and, if exist, dependency among operations. The final goal is to model the total distortion $D(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS})$ and the rate $R(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS})$.

In case of simple repeat interpolation (zero-order hold), the distortion generated by temporal interpolation is the variance for spatial gradients and motion vectors [6], *i.e.*

$$d_{TS_i} = \sigma_{x_i}^2 \sigma_{mv_x}^2 + \sigma_{y_i}^2 \sigma_{mv_y}^2. \quad (5)$$

At the same condition, in simple repeat interpolation, distortion generated by spatial interpolation is the variance of source, *i.e.* $d_{SS_i} = \sigma_{s_i}^2$. (6)

Here, it should be noted that there is dependency among transcoding operations or frames. In an independent case where there is no interframe and inter-operation dependency, $D_i(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS})$ and $R_i(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS})$ depend only on RQ_i, TS_i, and SS_i. Therefore, $D_i(\boldsymbol{RQ},\boldsymbol{TS},\boldsymbol{SS}) = D(RQ_i)+D(TS_i)+D(SS_i)$. However, in a dependent case, an operation can affect the distortion caused by the other operations. Namely, resolution downsampling affects the requantization distortion of the coded frame.

3.1.1. Statistical analysis and dependency distortion modeling

Figure 1 shows AC coefficients distribution for a transcoded frame.

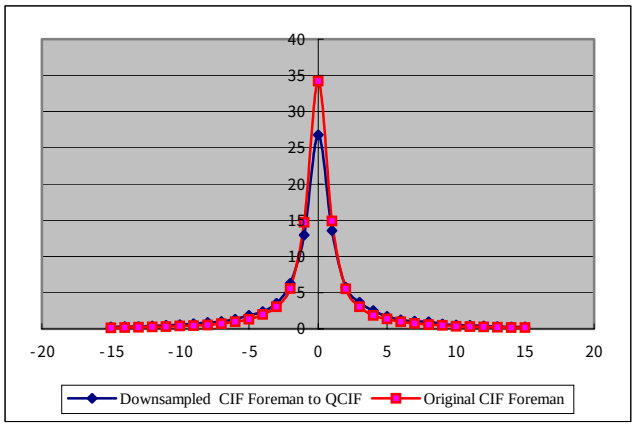

Fig. 1. The probability mass function of AC coefficients. Frame 0 of Foreman CIF originally coded at 30fps and 510kbps.

Let us assume that the source statistics are Laplacian distributed $p(x) = \frac{\alpha}{2} e^{-\alpha|x|}$ *where* $-\infty < x < \infty$.

As seen from Fig. 1, downsampling affects the source distribution, *i.e.*, α. In previous analysis [8], a rate distortion function is derived as $R(D) = \ln\left(\frac{1}{\alpha D}\right)$, where R is bits per sample. That means R-D function is affected by the amount of dependency. Therefore, $\mathrm{R(D)} = \ln\left(\frac{1}{\alpha_r \mathrm{D}}\right)$ where $0 < \mathrm{D} < \frac{1}{\alpha_r}$, where α_r is the reduced α by the dependency. By Taylor series expansion, it can be written as $R(D) = a'D^{-1} + b'D^{-2}$, where parameter a' and b' are obtained from multiple empirical samples.

Based on the above observation, we can conclude as follows: Due to the dependency among transcoding operations, downsampling affects the traditional R-D model. In other words, R-D model depends on the amount of downsampling.

3.1.2. Distortion modeling for dropped frame

In this subsection, we model $D(RQ_k, SS_k \mid TS_k = 0)$ for the dropped frame $\tilde{s}_k$ at time t_k in Eq. (2). If we assume that a temporal interpolator repeats the previously coded frame $\tilde{s}_i$, the estimation error at t_k is $e_k = \tilde{s}_k - \hat{s}_k = \tilde{s}_k - \hat{s}_i$, where $\hat{s}_k = \hat{s}_i$ by simple interpolator. The estimation error can be summarized as;

(a) *if* $(SS_i = 1)$

$$e_k = \tilde{s}_k - \hat{s}_i = \tilde{s}_k - \tilde{s}_i + \tilde{s}_i - \hat{s}_i = \tilde{s}_k - \tilde{s}_i + D(RQ_i), \quad (7)$$

(b) *else if* $(SS_i \neq 1)$

$$e_k = \tilde{s}_k - \hat{s}_i = \tilde{s}_k - \tilde{s}_i + \tilde{s}_i - \hat{s}_i = \tilde{s}_k - \tilde{s}_i + D(RQ_i, SS_i). \quad (8)$$

In other words, this distortion is caused by temporal interpolation error of the dropped frame $\tilde{s}_k$, requantization error, and spatial resolution downsampling error of the previous coded frame $\tilde{s}_i$ [4].

3.1.3. Distortion modeling for resized frame

In this subsection, we model $D(RQ_i, SS_i)$ for the coded frame $\tilde{s}_i$ at time t_i in Eq. (8). If we assume SS_i is the same in a temporal segment of a video, $D(RQ_i, SS_i) = D(RQ_i, S_c)$, *i.e.* $SS_1 = SS_2 = \cdots = SS_N = S_c$, where $0 < S_c \leq 1$ and $S_c = \frac{n_1' \times n_2'}{n_1 \times n_2}$, where n_1 x n_2 is the spatial resolution of the original video frame, and n_1' x n_2' is the downscaled spatial resolution.

The estimation error is as follows:

$$e_i = \tilde{s}_i - \hat{s}_i = d_{RQ_i} + d_{SS_i} + d_{S_c} = d_{RQ_i, S_c} + d_{SS_i}, \quad (9)$$

where $\hat{s}_i$ is the spatially interpolated frame.

Therefore, this distortion is caused by requantization error, spatial interpolation error, and dependency error caused by the amount of interpolation, *i.e.* S_c. From the dependency analysis in 3.1.1, we can further reduce it as requantization error affected by S_c and spatial interpolation error.

By substituting Eq. (9) into Eq. (8), the distortion for the dropped frame can be further derived as follows:

$$D(RQ_k, SS_k \mid TS_k = 0) = d_{TS_k} + d_{SS_i} + d_{RQ_i, S_c}. \quad (10)$$

3.1.4. Distortion modeling for coded frame

In this subsection, we model $D(RQ_i, SS_i \mid TS_i = 1)$ for the coded frame $\tilde{s}_i$ at time t_i in Eq. (2). We assume that frame dropping is uniform, and $f_s = \frac{F_S}{F_T}$, where F_S is the original frame rate of the source video, and F_T is the frame rate of the transcoded video. Then, it can be said that $D(RQ_i, SS_i \mid TS_i = 1) = D(RQ_i, f_s, SS_i)$. Obviously, we do not consider inter-frame dependency here to focus on inter-operation dependency. If we consider inter-frame dependency, it will be $D(D_r, RQ_i, f_s, SS_i)$, where D_r is the distortion of the reference frame. This consideration will be on our future work.

The estimation error, $D(RQ_i, f_s, S_c)$ is derived as follows:

$$e_i = \tilde{s}_i - \hat{s}_i = d_{RQ_i} + d_{SS_i} + d_{S_c} + d_{f_s} = d_{RQ_i, S_c, f_s} + d_{SS_i}, \quad (11)$$

where $\hat{s}_i$ is the spatially interpolated frame.

This distortion is caused by requantization error, spatial interpolation error, and the dependency error caused by the amount of the spatial interpolation, *i.e.* S_c, and the temporal interpolation, *i.e.* f_s. From the dependency analysis in 3.1.1, we can deduce that the distortion of the coded frame is the requantization error affected by S_c and f_s, and the spatial interpolation error.

3.2. Rate control for hybrid transcoding

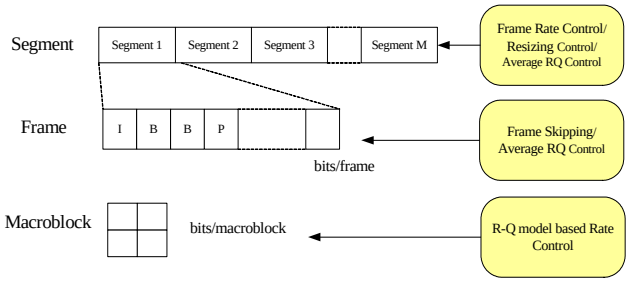

Fig. 2. The proposed rate control scheme.

Figure 2 shows the proposed rate control scheme which is a layered approach. This kind of scalable scheme has complexity scalability [9].

3.2.1. Segment level bit allocation

At first, among the operation sets, (***RQ***, ***TS***, ***SS***), we select the possible operation sets satisfying the given bitrate constraint for a segment. Then, we can select an operation from the possible sets so as to minimize the total distortion [4]. Especially, when a segment has similar source characteristics for each frame, the possible sets obtained are reasonable. In MPEG case, the number of each operation is generally finite. Therefore, the computational complexity is not high. Also, since we deal with not encoding but transcoding, the parameters to estimate rate and distortion, α_r, a, b, σ_{mv}^2, σ_s^2 and the average bits for each frame type, etc., are easily obtained during transcoding process. Figure 3 shows the proposed segment level rate control algorithm. Using this rate control algorithm, we can get frame rate, resizing amount, and average requantization amount for a segment by the finally selected operation set which minimizes the average distortion.

1. Calculate the target bit rate for a segment;
2. Estimate bitrate generated by using each operation in the operation set;
3. Select the possible operations which can generate the target rate;
4. Estimate the total distortion for the selected operations by the distortion measure, *i.e.* Eq. (5), Eq. (6), Eq. (10) and Eq. (11) modeled in the previous section;
5. Select an operation set (RQ, f_s, S_c) which minimizes the total distortion;

Fig. 3. The proposed rate control algorithm for a segment.

3.2.2. Frame level bit allocation

After downsampling, if needed, we can get easily each requantization parameter for each frame by using the conventional R-D model (4). In other words, since it is possible

to get parameters of model, *a* and *b*, from downscaled source, we can estimate R-Q model for each frame as in [8].

4. EXPERIMENTAL RESULTS

In our experiments, we have used Foreman video sequence which is originally encoded as MPEG-4 ASP format, 30f/s, and CIF resolution. It has IBBPBBPBBPBBPBBI structure. The used transcoding operations are as follows: requantization parameter RQ={1,..,31}, SS={1, 1/4}, and TS={1, 2/3, 1/2, 1/3, 1/5}. This means we transcode the bitstream as the type of {CIF,QCIF} and {30f/s, 20f/s, 15f/s, 10f/s, 5f/s}. We have employed the uniform frame dropping and spatial downsampling which is the same in a temporal segment. So, (RQ, f_s, S_c) pairs are selected as the result of joint control for hybrid transcoding.

Table 1 shows the possible operations which can generate the target rate 80Kbps from CIF Foreman originally coded at 510Kbps. In our experiment, we have used scene unit as segment for segment level rate control. Foreman sequence consists of 3 scenes. This result is for the first scene which is from frame 1 to frame 170. As seen from Table 1, the best operation set which minimizes the distortion is (18, 1, 1/4). To verify the effectiveness of our joint control, we examine the actual rate of the transcoded bitstream, the estimated distortion, and the finally selected operation pair satisfying each target rate. Table 2 shows the result of that for hybrid transcoding. Also, we examine the actually measured root mean square error (RMSE). Figure 4 shows that the distortion is estimated well.

5. CONCLUSION

In this paper, we have proposed the practical joint control of multidimensional transcoding by rate-distortion model based approach without transcoding a video for actual R-D curve. Also, we have discussed tradeoffs in quality among transcoding operations to allocate bits more efficiently. By describing statistical analysis of dependency distortion, we have proposed a new distortion model.

In our future work, we will incorporate the perceptual model based on the concept of distortion masking into our R-D model, *i.e.* HVS based optimal decision will be exploited.

6. REFERENCES

[1] F.C. Martins, W. Ding, and E. Feig, "Joint control of spatial quantization and temporal sampling for very low bit rate video," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 2072-2075, May 1996.

[2] E.C. Reed, and J.S. Lim, "Optimal Multidimensional Bit-Rate Control for Video Communication," *IEEE Trans. Image Processing*, vol. 11, no. 8, pp. 873-885, Aug. 2002.

[3] A. Vetro, H. Sun, and Y. Wang, "Object-Based Transcoding for Adaptable Video Content Delivery," *IEEE, Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 387-401, Mar. 2001.

[4] P. Yin, A. Vetro, M. Xia, and B. Liu, "Rate-Distortion Models for Video Transcoding," in *Proc. SPIE-IS&T Electronic Imaging*, vol. 5022, pp. 479-488, Jan. 2003.

[5] S. Liu, and C.-C. J. Kuo, "Joint Temporal-Spatial Rate Control for Adaptive Video Transcoding," in *Proc. Int. Conf. Multimedia and Expo*, vol. 2, July 2003.

[6] J.-W. Lee, A. Vetro, Y. Wang, and Y.-S. Ho, "Bit Allocation for MPEG-4 Video Coding With Spatio-Temporal Tradeoffs," *IEEE, Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 488-502, June 2003.

[7] N.S. Jayant, and P. Noll, "Digital Coding of Waveforms," Prentice-Hall Inc., 1984.

[8] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE, Trans. Circuits Syst. Video Technol.*, Feb. 1997.

[9] H.J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE, Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 878-894, Sept. 2000.

Table 1. The estimated operations which can generate the target rate 80Kbps from CIF Foreman originally coded at 510Kbps.

RQ	f_s	S_c	The estimated RMSE
7	1/5	1/4	14.60
10	1/3	1/4	13.46
12	1/2	1/4	11.87
14	2/3	1/4	10.26
18	1	1/4	7.00
21	1/5	1	12.02
30	1/3	1	10.76

Table 2. The result of joint control for hybrid transcoding of CIF Foreman originally coded at 510Kbps.

Target rate (Kbps)	The actual transcoded rate (Kbps)	The estimated RMSE	The finally selected operation (RQ, f_s, S_c)	The actual RMSE
30	31	11.09	(25, 1/5, 1/4)	11.35
40	43	9.63	(30, 1/2, 1/4)	9.15
60	73	7.11	(28, 1, 1/4)	8.25
80	87	7.00	(18, 1, 1/4)	7.17
100	103	6.92	(13, 1, 1/4)	6.69
200	278	4.01	(22, 1, 1)	4.98
300	324	3.90	(15, 1, 1)	4.63
400	420	3.78	(11, 1, 1)	3.80
500	454	3.68	(9, 1, 1)	3.68

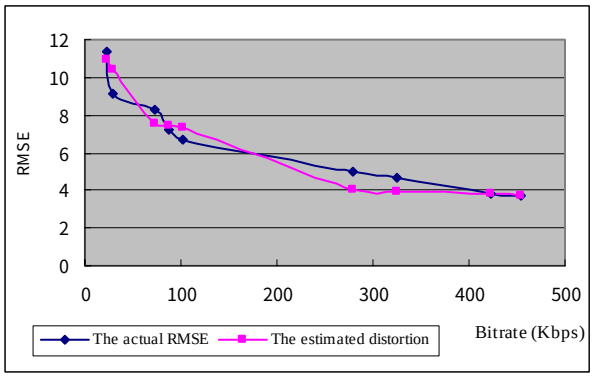

Fig. 4. The comparison of the estimated distortion and the actually measured root mean square error.