Modeling Video Rate as a Function of Frame Size, Frame Rate and Quantization Stepsize

Zhan Ma[†‡], Meng Xu[†] and Yao Wang[†]
[†]Dept. of ECE, Polytechnic Institute of NYU, Brooklyn, NY 11201
[‡]Samsung Telecommunications America, Richardson, TX 75082

***Abstract*—In this paper, we investigate the impacts of spatial, temporal and amplitude resolution (STAR) on the bit rate of a compressed video. We propose a rate model in terms of the quantization stepsize, frame size and frame rate. Experimental results reveal that the increase of the video rate as the individual resolution increases follows a power function. Hence, the proposed model expresses the rate as the product of power functions of the quantization stepsize, frame size and frame rate, respectively. The proposed rate model is analytically tractable, requiring only four content dependent parameters. Simulation results show that model predicted rates fit the measured data very well with high Pearson correlation (PC) and small relative root mean square error (RRMSE). The same model function works for different coding scenarios (including scalable and non-scalable video, temporal prediction using either hierarchical B or IPPP structure, etc.) with very high accuracy (average PC $>$ 0.99), but the values of model parameters differ.**

***Index Terms*—Rate model, spatial resolution, temporal resolution, quantization, H.264/AVC, SVC**

I. Introduction

A fundamental and challenging problem in video encoding is, given a target bit rate, how to determine at which spatial resolution (i.e., frame size), temporal resolution (i.e., frame rate), and amplitude resolution (usually controlled by the quantization stepsize or QS), to code the video. One may code the video at a high frame rate, large frame size, but high QS, yielding noticeable coding artifacts in each coded frame. Or one may use a low frame rate, small frame size, but small QS, producing high quality frames. These and other combinations can lead to very different perceptual quality. Ideally, the encoder should choose the spatial, temporal, and amplitude resolution (STAR) that leads to the best perceptual quality, while meeting the target bit rate. Optimal solution requires accurate rate and perceptual quality prediction at any STAR combination.

In this paper, we investigate how does the rate change as a function of the quantization stepsize q, frame size s and frame rate t. This work is extended from our previous paper [1], where we consider the impact of temporal and amplitude resolutions on the video rate. Rate modeling for video coding has been researched over decades. However, almost all of them consider the rate model with respect to the quantization only [2]–[5]. This work is the first one attempting to model the rate in terms of the complete STAR combination. Based on our extensive simulations, our proposed rate model is generally applicable to all coding scenarios, such as scalable or non-scalable (i.e., single layer) video, hierarchical B or IPPP structure, with or without QP cascading [6], etc.

The remainder of this paper is organized as follows: Section II presents the rate model considering the joint impact of spatial, temporal and amplitude resolutions, for videos coded with spatial and temporal scalability, but no amplitude scalability. We then validate the same rate model is applicable for other coding scenarios in Section III. Section IV concludes the current work and discusses the future research directions.

II. Rate Model For Spatial and Temporal Scalable Video

In this section, we develop a rate model $R(q, s, t)$, which relates the rate R with the quantization stepsize q, frame size s and frame rate t, based on the rates of video bitstreams generated using the spatial and temporal scalability of H.264/SVC at multiple fixed quantization parameters (QPs). The model is derived by recognizing

$$R(q,s,t) = \\ R_{\max} R_t(t; q_{\min}, s_{\max}) R_q(q; s_{\max}, t) R_s(s; q, t), \qquad (1)$$

where $R_{\max} = R(q_{\min}, s_{\max}, t_{\max})$ is the maximum bit rate obtained with a chosen minimal quantization stepsize $q_{\min}$, a chosen maximum frame size $s_{\max}$ and a chosen maximum frame rate $t_{\max}$;

$$R_s(s; q, t) = \frac{R(q, s, t)}{R(q, s_{\max}, t)}$$

is the normalized rate versus spatial resolution (NRS) under a certain q and t;

$$R_q(q; s_{\max}, t) = \frac{R(q, s_{\max}, t)}{R(q_{\min}, s_{\max}, t)}$$

is the normalized rate versus quantization stepsize (NRQ) for any t under the $s_{\max}$; and

$$R_t(t; q_{\min}, s_{\max}) = \frac{R(q_{\min}, s_{\max}, t)}{R(q_{\min}, s_{\max}, t_{\max})}$$

is the normalized rate versus temporal resolution (NRT) under the $q_{\min}$ and $s_{\max}$; As will be shown later by experimental data, $R_s(s; q, t)$ is actually quite independent of the q and t, which is denoted as $R_s(s)$; $R_q(q; s_{\max}, t)$ is independent of the t, denoted as $R_q(q)$, while $R_t(t; q_{\min}, s_{\max})$ can be noted as $R_t(t)$ for fixed $q_{\min}$ and $s_{\max}$.

Emails: zhan.ma@ieee.org, mxu02@students.poly.edu, yao@poly.edu.

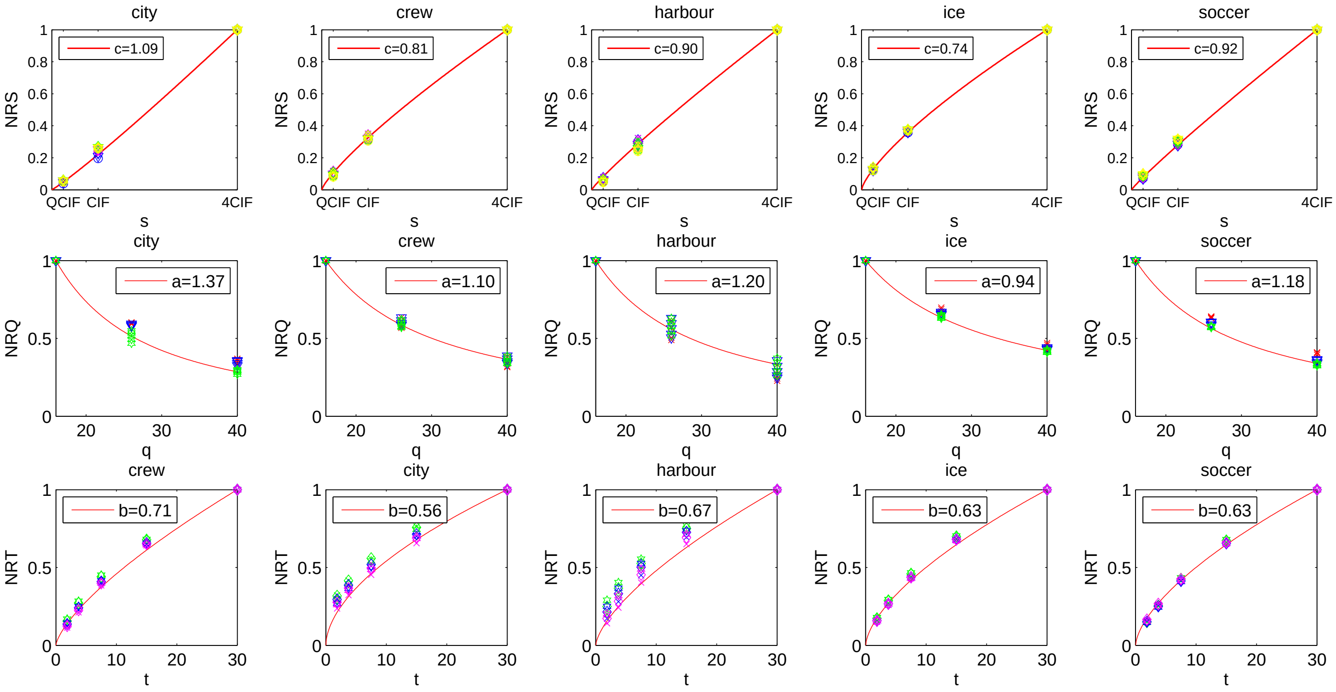

Fig. 1. Illustrations of NRS, NRQ and NRT for all combinations of q, s and t, where $q \in [64, 40, 26, 16]$, $t \in [1.875, 3.75, 7.5, 15, 30]$ and $s \in$ [QCIF, CIF, 4CIF]. Points are measured rates, curves are predicted rates using respective Eqs. (2) (3) and (4). NRS curves are fitted using all possible q and t; NRQ curves are fitted using all possible t at $s_{\max}$ (green hexagram markers), while NRT curves are obtained using points at $q_{\min}$ and $s_{\max}$ (magenta cross markers). In our case, $q_{\min}$ = 16 and $s_{\max}$ = 4CIF.

To see how quantization, frame size and frame rate respectively influence the bit rate, we first encode several test videos using the joint spatial and temporal scalability tool of JSVM [6] and measured the actual bit rates corresponding to different STARs. Specifically, five video sequences, “city”, “crew”, “harbour”, “ice” and “soccer”, at original 4CIF (704x576) resolutions, are encoded into 5 temporal layers using dyadic hierarchical prediction structure, with frame rates at 1.875, 3.75, 7.5, 15 and 30 Hz, respectively, and each temporal layer contains 3 dyadic spatial layers (i.e., QCIF, CIF and 4CIF). For simplicity, we constrain the same QP for all temporal and spatial layers (i.e., without using *QP cascading* [7]). To investigate the impact of QP, we have coded the video using QP ranging from 16 to 40. Here, we only present the results with QP = 40, 36, 32 and 28. Corresponding quantization stepsizes are 64, 40, 26 and 16, respectively. Other QPs have the similar performance according to our simulation results.

The bit rates of all layers are collected and normalized by the rate at the largest frame size, i.e., 4CIF, to find NRS points $R_s(s; q, t)$ for all q and t, which are plotted in the first row of Figure 1. As shown, the NRS curves obtained with different q and t overlap with each other, and can be captured by a single curve quite well. Similarly, the NRQ curves (middle row) are also almost invariant with the frame rate t, and vary slightly for different frame size s as shown in Figure 1; On the other hand, NRT curves (last row) are quite dependent on frame size and quantization as shown in Figure 1. These observations suggest that the effects of quantization q, frame size s and frame rate t on the bit rate can be captured using (1). Therefore, the overall rate modeling problem is divided into three parts, one is to devise an appropriate functional form for $R_s(s)$, so that it can model the measured NRS points for all q and t in Figure 1 accurately, the second one is to derive an appropriate functional form for $R_q(q)$ that can accurately model the measured NRQ points for all t at $s = s_{\max}$, and the third part is to provide a proper functional form for $R_t(t)$ that can accurately capture the measured NRT points at $q_{\min}$ and $s_{\max}$. The derivation of the models $R_q(q)$, $R_t(t)$ and $R_s(s)$ are explained in detail as follows.

A. Model for Normalized Rate v.s. Spatial Resolution $R_s(s)$

$R_s(s)$ is used to describe the reduction of the normalized bit rate as the frame size reduces. As we can see, the desired property for the $R_s(s)$ function is that it should be 1 at $s = s_{\max}$ and monotonically reduces to 0 at $s = 0$. We choose a power function to model the $R_s(s)$, i.e.,

$$R_s(s) = \left(\frac{s}{s_{\max}}\right)^c . \tag{2}$$

Fig. 1 shows the model curve using (2) along with the measured data (for all possible q and t). The parameter c is obtained by minimizing the RMSE between model predicted rates and actual measured rates, and characterizes the speed of bit rate reduction along when the frame size decreases. It can be seen that the model prediction fits the actual measurements very well.

B. Model for Normalized Rate v.s. Quantization $R_q(q)$

Analogous to the $R_s(s)$, $R_q(q)$ is used to describe the reduction of the normalized bit rate as the quantization stepsize increases at a fixed frame size $s_{\max}$. As shown in the data presented in Fig. 1, $R_q(q)$ is independent of the t. The desired property for the $R_q(q)$ function is that it should be 1 at $q = q_{\min}$ and monotonically reduces to 0 as q goes to infinity. Hence, we choose an inverse power function for $R_q(q)$, i.e.,

$$R_q(q) = \left(\frac{q}{q_{\min}}\right)^{-a}. \tag{3}$$

Fig. 1 shows the model curve using (3) along with the measured data. It can be seen that the model fits the measured data points accurately. The parameter a characterizes how fast the bit rate reduces when q increases. We note that the model in (3) is consistent with the model proposed by Ding and Liu [2] for non-scalable video where video frame size and frame rate are both fixed.

C. Model for Normalized Rate v.s. Temporal Resolution $R_t(t)$

$R_t(t)$ is used to describe the reduction of the normalized bit rate as the frame rate reduces at $q_{\min}$ and $s_{\max}$. Therefore, the desired property for the $R_t(t)$ function is that it should be 1 at $t = t_{\max}$ and monotonically reduces to 0 at $t = 0$. We choose a power function to describe the $R_t(t)$, i.e.,

$$R_t(t) = \left(\frac{t}{t_{\max}}\right)^{b}. \tag{4}$$

Fig. 1 shows the model curve using this function along with the measured data. The parameter b is obtained by minimizing the squared error between the modeled rates and measured rates. It can be seen that the model fits the measured data (e.g., at $q_{\min}$ and $s_{\max}$) points very well. We also tried some other functional forms, including logarithmic and inverse falling exponential. We found that the power function yields the least fitting error.

D. The Overall Rate Model

Combining Eqs. (2), (4) and (3), we propose the following rate model

$$R(q,s,t) = R_{\max}\left(\frac{q}{q_{\min}}\right)^{-a}\left(\frac{t}{t_{\max}}\right)^{b}\left(\frac{s}{s_{\max}}\right)^{c}, \tag{5}$$

where $q_{\min}$, $s_{\max}$ and $t_{\max}$ should be chosen according to the underlying application, $R_{\max}$ is the actual rate when coding a video at $q_{\min}$, $s_{\max}$ and $t_{\max}$, and a, b and c are the model parameters. Here, we assume $R_{\max}$ can be estimated accurately. In practices, we can have the knowledge of $q_{\min}$, $s_{\max}$ and $t_{\max}$. For instance, almost for all video capable mobile handhelds, they are featuring video decoding or encoding at 720p (i.e., 1280x720) and 30 frame per second, noted as 720p@30 Hz. We can assume $s_{\max}$ = 720p and $t_{\max}$ = 30 Hz. $q_{\min}$ can be estimated by the maximum bit rate which is limited by the codec profile and level constraints. On the other hand, in real product, there is a $q_{\min}$ below which there is no visual difference. Hence, $q_{\min}$ can be determined from either maximum bit rate or visual difference.

The model parameters, a, b and c, are obtained by minimizing the RMSE between measured and predicted rates corresponding to all STARs. Table I lists the parameter values and model accuracy in terms of relative RMSE (i.e., RRMSE = RMSE/$R_{\max}$), and the Pearson correlation (PC) between measured and predicted rates. We see that the model is very accurate for all different sequences, with small RRMSE and high PC. We exemplify the actual rate data and corresponding estimated rates for all videos, via the proposed model (5) in Figure 2. Results show that our proposed model can predict the bit rate very well.

TABLE I
RATE MODEL PARAMETER AND ITS ACCURACY FOR SVC#1

	city	crew	harbour	ice	soccer	ave.
a	1.394	1.139	1.373	0.936	1.152	
b	0.547	0.702	0.640	0.628	0.635	
c	1.114	0.830	0.952	0.736	0.899	
RRMSE	1.12%	0.75%	0.94%	0.72%	0.41%	0.80%
PC	0.9985	0.9991	0.9985	0.9993	0.9997	0.990

III. MODEL VALIDATION FOR OTHER CODING SCENARIOS

The results presented in the previous section is for video coded using joint spatial and temporal scalability only. Different amplitude resolutions are fulfilled by encoding multiple joint spatial-temporal bitstreams using different quantization stepsize. In this section, we verify that our rate model works for other coding scenarios as well, i.e., scalable or non-scalable (i.e, single layer) video using hierarchical B or IPPP for temporal prediction, with or without QP cascading, etc.

TABLE II
EXPERIMENTS WITH DIFFERENT CODING STRUCTURES

	QP Cascading		GOP	#SR	#AR: QP
	temporal	spatial			
SVC#1	NO	NO	HierB: 16	3	4: 28, 32, 36, 40
SVC#2	Yes	Yes	HierB: 16	3	3 : 16, 20, 24
SL#1	NO	NO	IPPP: 8	3	4: 24, 28, 32, 36
SL#2	Yes	Yes	HierB: 8	3	3: 16, 20, 24
CombS	Yes	Yes	HierB: 8	2	3: 16, 20, 24

We summarize the few scenarios considered in Table II. HierB stands for dyadic hierarchical B prediction structure with the number indicating the GOP length, and so is the IPPP structure. #SR is the number of spatial resolutions (SRs). For SVC cases, multiple SR and TR are obtained by using the spatial scalability and hierarchical B structure of the H.264/SVC standard. With single layer (SL) cases, to code a video at different SRs, we first down sample the original video to the desired SR using the filter suggested by [7] [1], and then code the video at that SR. For SL#2, multiple TRs are obtained using the HierB structure; whereas for SL#1, each TR is obtained by temporally down-sampling the original video to the desired TR. #AR is the number of amplitude resolutions (ARs), which is controlled by the QP. Without QP cascading, all pictures are coded using the same QP. But with QP cascading, the QP used for pictures at higher spatial and temporal resolutions are higher than those used for the pictures with lower spatial and temporal resolution (i.e. the base layer). The cited QP are those used at the base layer,

[1] JSVM also applies the same down-sample filter to generate spatial scalable streams.

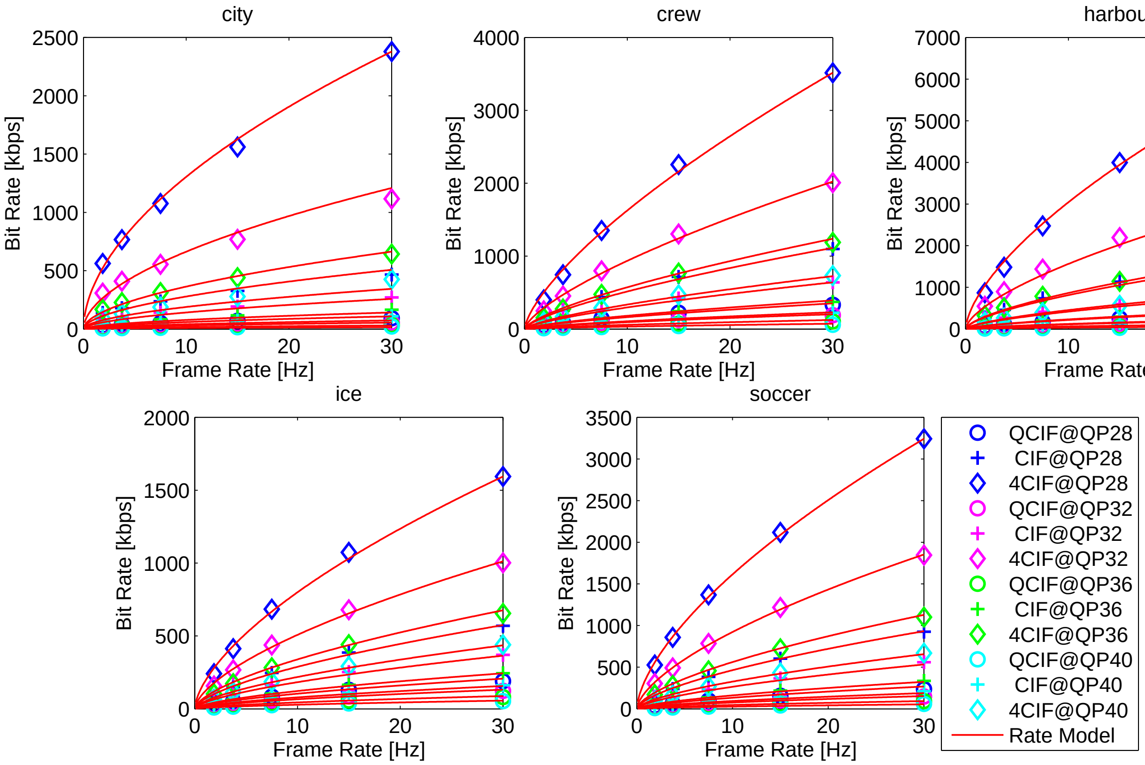

Fig. 2. Rate prediction using (5) for test sequences at all STAR combinations.

which are the first frame in each GOP at QCIF resolution in our case. In our simulation, QP cascading is applied using default settings recommended by [7]. To provide multiple amplitude resolutions, both the SVC and SL cases code a video using different base layer QPs. CombS refers to combined scalability of H.264/SVC, which can provide different STAR combinations within a single scalable stream. Because of the limitation of JSVM implementation, we have experimented the combined scalability using two spatial layers, three amplitude layers and four temporal layers as shown in Figure 3.

TABLE III
RATE MODEL PARAMETER AND ITS ACCURACY FOR SVC#2

	city	crew	harbour	ice	soccer	ave.
a	1.342	1.20	1.171	0.952	1.092	
b	0.329	0.538	0.508	0.496	0.454	
c	0.806	0.533	0.646	0.537	0.642	
RRMSE	1.03%	1.26%	1.60%	1.19%	1.14%	1.24%
PC	0.9985	0.9974	0.9956	0.9979	0.9980	0.9975

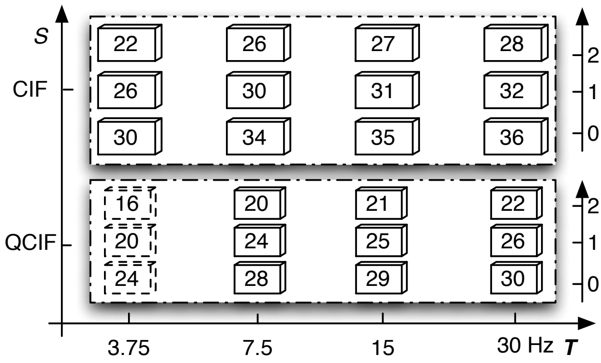

Fig. 3. Illustrative layered structure for CombS: $A = 0$ is the amplitude base layer. Different QPs are applied to temporal/spatial enhancement layers by enabling QP cascading. Delta QP is 4 and 6 for successive amplitude and spatial layers, respectively.

Please note that Table I is the experimental results for simulation SVC#1. Table III presents the prediction accuracy and model parameters for SVC#2, while Table IV and V show the results for SL#1 and SL#2. Model accuracy for CombS is given in Table VI. We can see that our proposed rate model is generally applicable regardless the coding structures, having high PC and small RRMSE for all coding scenarios. We also notice that model parameters are highly content dependent, and their values vary among different coding scenarios as well. We can see that QP cascading brings slower rate dropping with respect to frame rate or frame size with smaller b and c. This is mainly due to the less bits spent for pictures at higher spatial and temporal resolutions after applying larger QP. Because of the inter-layer prediction between successive amplitude resolutions (to remove more residual redundancy), CombS has smaller a indicating the slower rate dropping in terms of the quantization compared with SVC#2.

TABLE IV
RATE MODEL PARAMETER AND ITS ACCURACY FOR SL#1

	city	crew	harbour	ice	soccer	ave.
a	1.935	1.362	1.23	1.12	1.38	
b	0.836	0.828	0.795	0.679	0.711	
c	1.301	0.881	0.895	0.729	0.992	
RRMSE	0.76%	0.93%	0.95%	1.15%	0.81%	0.92%
PC	0.9992	0.9988	0.9983	0.9983	0.9991	0.9987

TABLE V
RATE MODEL PARAMETER AND ITS ACCURACY FOR SL#2

	city	crew	harbour	ice	soccer	ave.
a	1.333	1.054	1.149	0.851	1.037	
b	0.242	0.491	0.422	0.454	0.403	
c	0.479	0.266	0.361	0.239	0.40	
RRMSE	1.26%	1.24%	1.98%	1.19%	1.19%	1.37%
PC	0.9974	0.9970	0.9924	0.9971	0.9975	0.9963

TABLE VI
RATE MODEL PARAMETER AND ITS ACCURACY FOR COMBS

	city	crew	harbour	ice	soccer	ave.
a	0.881	0.69	0.768	0.647	0.771	
b	0.254	0.536	0.471	0.486	0.441	
c	0.902	0.605	0.808	0.669	0.799	
RRMSE	2.19%	2.67%	1.89%	2.24%	1.52%	2.10%
PC	0.9968	0.9942	0.9971	0.9962	0.9983	0.9965

IV. DISCUSSION AND CONCLUSION

In this paper, we propose an analytical rate model considering the joint impact of the spatial, temporal and amplitude resolutions. The overall rate model is the product of power functions of frame rate, frame size and quantization stepsize. Our proposed analytical rate model is applicable to all coding scenarios, such as scalable or non-scalable video using hierarchical B or IPPP temporal prediction, with or without QP cascading, etc. Although we only show results for video at CIF resolution, we have also verified that our model is accurate for other spatial resolutions (e.g., 720p, WVGA, etc), which are not included here to save the space.

In applications of streaming pre-coded video, model parameters can be pre-calculated based on the actual rates at different STAR, while for encoder optimization (such as rate control), we need to estimate them accurately. Experimental data show that model parameters are highly content dependent. As a future study, we will investigate the model parameter prediction using content features. On the other hand, we have applied a prior version of the rate model and a quality model, where we only consider the impact of temporal and amplitude resolutions on the rate and quality respectively, to do frame rate adaptive rate control [8] and scalable video adaptation [1]. As a future work, we will also apply our complete R-STAR model, together with our proposed quality model as a function of STAR [9] to do perceptual encoder rate control and scalable video adaptation , where we maximize the video quality by choosing appropriate frame size, frame rate and quantization stepsize under the bit rate constraint.

REFERENCES

[1] Y. Wang, Z. Ma, and Y.-F. Ou, “Modeling rate and perceptual quality of scalable video as functions of quantization and frame rate and its application in scalable video adaptation,” in *Proc. of Packet Video*, Seattle, WA, 2009.

[2] W. Ding and B. Liu, “Rate control of MPEG video coding and recoding by rate-quantization modeling,” *IEEE Trans. Circuit and Sys. for Video Technology*, vol. 6, pp. 12–20, Feb. 1996.

[3] T. Chiang and Y.-Q. Zhang, “A new rate control scheme using quadratic rate distortion model,” *IEEE Trans. Circuit and Sys. for Video Technology*, vol. 7, no. 2, pp. 246–250, Feb. 1997.

[4] J. Ribas-Corbera and S. Lei, “Rate control in DCT video coding for low-delay communications,” *IEEE Trans. Circuit and Sys. for Video Technology*, vol. 9, no. 2, pp. 172–185, Feb. 1999.

[5] Z. He and S. K. Mitra, “A novel linear source model and a unified rate control algorithm for H.264/MPEG-2/MPEG-4,” in *Proc. of Intl. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, May 2001.

[6] Joint Scalable Video Model (JSVM), *JSVM Software*, Joint Video Team, Doc. JVT-X203, Geneva, Switzerland, June 2007.

[7] Joint Scalable Video Model, *JSVM Encoder Description*, Joint Video Team, Doc. JVT-X202, Geneva, Switzerland, June 2007.

[8] Z. Ma, M. Xu, K. Yang, and Y. Wang, “Modeling rate and perceptual quality of video and its application to frame rate adaptive rate control,” in *Proc. of IEEE ICIP*, Brussels, Belguim, Sept. 2011.

[9] Y.-F. Ou, Y. Xue, Z. Ma, and Y. Wang, “A perceptual video quality model for mobile platform considering impact of spatial, temporal, and amplitude resolutions,” in *Proc. of IEEE IVMSP*, Ithaca, NY, June 15-17 2011.