Dependent R/D Modeling Techniques and Joint T-Q Layer Bit Allocation for H.264/SVC

Yongjin Cho, Do-Kyoung Kwon, *Member, IEEE,* Jiaying Liu, *Member, IEEE,* and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*—We investigate dependent rate/distortion (R/D) modeling techniques for H.264/SVC videos. We introduce a self-domain (*S*-domain) analysis method for characterizing the dependent R/D behaviors, where the R/D characteristics of a base layer are employed as the observation domain for those of dependent layers. Based on *S*-domain observations, we propose empirical dependent R/D models and analyze physical implications of the proposed models. As an application of the proposed R/D models, we examine a joint temporal-quality layer bit allocation algorithm formulated as a Lagrange optimization problem. The proposed R/D models enable us to derive an analytical solution to the joint optimization problem. Finally, it is demonstrated by experimental results that our bit allocation algorithm outperforms JSVM benchmark by a significant margin (10%–20%) at various bit rates.**

***Index Terms*—Dependent R/D model, H.264/SVC, joint bit allocation, rate-distortion optimization, video coding.**

I. Introduction

IN VIDEO coding, analytical rate/distortion (R/D) models play an important role in the development of practical and efficient rate control algorithms in video encoders. Rate control algorithms are often formulated as an optimal bit allocation problem. Analytical R/D models not only provide the understanding of the source characteristics, but also they provide a simple solution to the complex bit allocation problems. We have various R/D models in the literature that model the R/D characteristics of a frame in various video coding standards of MPEG and H.26x. Among them, a quadratic model in the quantization domain (q-domain) analysis [1] and a linear source model in the ρ-domain analysis [2], [3] are most widely employed.

Conventional R/D models characterize the R/D behaviors of residual images after intra- or interprediction [1]–[10]. For this reason, the influence of the fidelity of references on the R/D characteristics of a macroblock or a frame cannot be properly understood with them. In this research, we investigate the influence of the fidelity of references on the R/D characteristics of predicted images, i.e., dependent R/D characteristics. Considering that a video is a sequence of images, it is desirable to understand the dependent R/D characteristics of an input video. Once we understand them, it becomes feasible to perform an optimal bit allocation for a number of frames, e.g., a group of pictures (GOP), which is known as a dependent quantization or dependent bit allocation problem in the literature.

Ramchandran *et al.* [11] studied a dependent quantization problem. They introduced monotonicity property that provides a general principle for a quantization decision between a reference and an intercoded frame. They could demonstrate an optimal solution to a dependent bit allocation problem. However, their solution is limited by the exponential complexity because it employs a dynamic programming method on a Trellis of real R/D data at each quantization step. That is, a number of frames have to be encoded with all possible combinations of quantization steps of all frames. As a result, the number of encoder runs for the data generation grows exponentially with the number of considered frames for a bit allocation. Lin *et al.* [12] proposed dependent R/D models of P-frames for MPEG-2 videos. Even though the proposed R/D models are quite accurate, they could not model the dependent R/D characteristics of B-frames properly. For this reason, it is difficult to apply their models to a number of frames with highly complex prediction structures, such as hierarchical B-pictures [13].

In this paper, we first investigate dependent R/D modeling techniques for H.264/SVC, a scalable extension of H.264/AVC [14]. As an extension of H.264/AVC, H.264/SVC provides three scalability dimensions of temporal (T), quality (Q), and spatial (S) scalability. The T scalability is realized by hierarchical frame structures, and layered architecture is employed for Q and S scalability [15].

Fig. 1 demonstrates a GOP structure of H.264/SVC, where the prediction structure is illustrated by arrows. To investigate the interdependence among scalable layers, we introduce a self-domain (*S*-domain) analysis. It is called an *S*-domain

Manuscript received November 9, 2011; revised March 22, 2012, June 7, 2012, and September 20, 2012; accepted December 26, 2012. Date of publication February 21, 2013; date of current version May 31, 2013. This work was supported in part by the National Natural Science Foundation of China, under Contract 61101078, and the Doctoral Fund of the Ministry of Education of China under Contract 20110001120117. This paper was recommended by Associate Editor P. L. Callet. (*Corresponding author:* J. Liu.)

J. Liu is with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: liujiaying@pku.edu.cn).

Y. Cho is with Samsung Electronics, Seoul 443742, Korea (e-mail: choyongjin@gmail.com).

D.-K. Kwon is with Texas Instruments, Dallas, TX 75243 USA (e-mail: d-kwon@ti.com).

C.-C. Jay Kuo is with the Signal and Image Processing Institute and the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: cckuo@sipi.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2248215

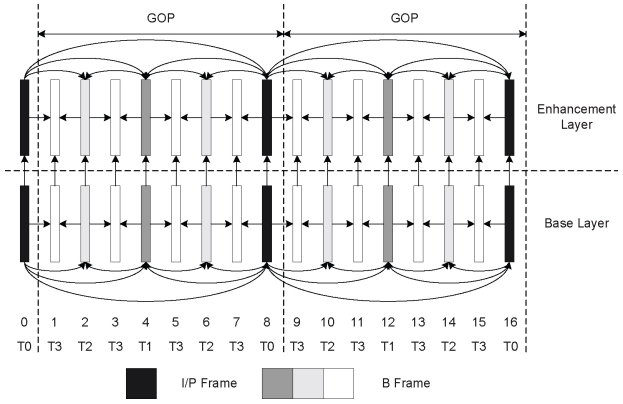

Fig. 1. Prediction structure within a GOP of H.264/SVC, where four temporal layers and two quality layers are shown.

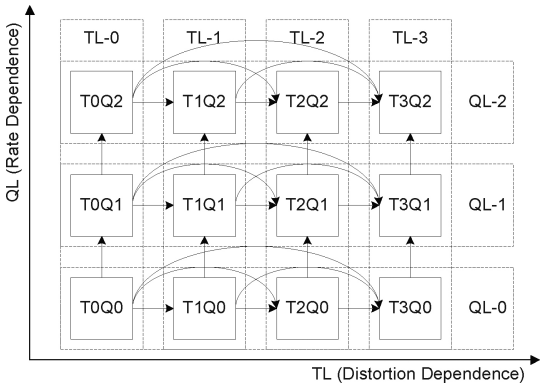

Fig. 2. H.264/SVC video with combined T-Q scalability, four TLs, and three QLs.

analysis because the R/D behaviors of dependent layers are characterized by those of their base layers. By the *S*-domain analysis, four important properties of the dependent R/D characteristics are identified, and the rate (or distortion) of a dependent layer is successfully expressed as a linear combination of the rate (or distortion) functions of base layers. As a result, complex multivariate R/D functions of dependent coding units are converted into a simple linear combination of univariate R/D functions of a base layer. In this paper, we focus on the dependence in the combined T and Q scalability. Dependent R/D models in the spatial scalability are already well studied in our previous work [16]. It is important to note that the proposed models characterize only the influence of the fidelity of references on the R/D behaviors of dependent scalable layers. For this reason, the proposed models have to be used with residual R/D to solve a dependent bit allocation problem.

Second, we examine a dependent bit allocation problem in the combined T-Q scalability of H.264/SVC as an application of the proposed dependent R/D models. Several layer-based rate control algorithms have been proposed for H.264/SVC [17]–[21]. Most of them are based on existing algorithms for previous video coding standards that do not consider dependent R/D characteristics. Even though Pranantha *et al.* [11] considered the interdependence issue, their solution directly follows the framework, and thus, its complexity requirement grows exponentially with the number of frames. The rate control algorithm in [17] is a single layer rate control algorithm, where the dependence among scalable layers is not considered. Even though Liu *et al.* [20] considered a T layer (TL) weighting in the weighting factors are identical regardless of source video temporal characteristics. TL differentiation is examined in [18] where different scaling factors are considered for hierarchical levels. However, scaling factors are determined mainly by the frame complexity that is heuristically defined as the product of bits and quantization step size of a frame. For this reason, their scaling factors cannot be justified to properly represent the temporal dependence in the hierarchical B-pictures.

In the joint T-Q layer bit allocation problem, a scalable block specified by a T and a Q layer (TL and QL) ID is chosen to be a bit allocation unit. We propose a bit allocation algorithm that efficiently allocates the bit budget to each bit allocation unit based on the proposed dependent R/D models. The problem is formulated by the Lagrange optimization method, and a fully analytical solution is derived using the proposed R/D models. The performance of the proposed algorithm is demonstrated in comparison with the JSVM FixedQPEncoder [22]. Significant coding gain is observed, and some of our preliminary results can be found in [23] and [24].

There are two major contributions in this research. First, we propose dependent linear R/D models for the joint T-Q scalability of H.264/SVC. Our *S*-domain based dependent R/D characteristics analysis allows greatly simplified dependent R/D models. To the best of our knowledge, they are the first dependent R/D models for scalable video. Second, we develop a low-complexity joint T-Q layer dependent bit allocation algorithm based on an analytical solution to a Lagrange equation. That is, the exponential complexity of the conventional dependent bit allocation algorithms [11], [12] is successfully reduced to linear complexity in the proposed bit allocation algorithm.

The remainder of this paper is organized as follows. We examine the R/D characteristics of dependent T and Q layers individually in Section II. Dependent R/D models are proposed using the *S*-domain analysis in Section III, and Section IV provides the analysis of the proposed models. We study the problem of joint T-Q layer bit allocation in Section V, and show that the proposed bit allocation algorithm enhances the overall performance by a significant margin compared to its benchmark. Finally, concluding remarks and future research directions are provided in Section VI.

II. Dependent R/D Characterization via Self-Domain Observation

Fig. 2 demonstrates a combined T-Q scalability plane with four TLs and three QLs, where arrows indicate prediction structure. We define a scalable block as a set of layer pictures identified by a coordinate of a TL and a QL ID (TID and QID). A scalable block is specified by $TiQj$ to indicate a coordinate of a TID and a QID as demonstrated in Fig. 2. Then, a scalable bit stream with a combined T-Q scalability is composed of a number of disjoint scalable blocks. Similarly, a TL or a QL is formed by a set of scalable blocks having an identical TID or QID. For example, TL-2 in Fig. 2 is composed of scalable

blocks with $TID = 2$, i.e., $T2Q0$, $T2Q1$, and $T2Q2$. Similarly, QL-2 is formed by scalable blocks with $QID = 2$, i.e., $T0Q2$, $T1Q2$, $T2Q2$, and $T3Q2$. Our goal in this research is to understand dependent R/D characteristics of scalable blocks in the combined T-Q scalability. To be specific, we would like to derive tractable functions of the dependent R/D characteristics with a scalable block as a basic modeling unit.

Dependent R/D functions are often expressed in multivariate functions. For example, the R/D functions of a dependent scalable block $T1Q1$ can be expressed as

$$f_{1,1}\left(x_{1,1}|x_{0,0}, x_{0,1}, x_{1,0}\right) \tag{1}$$

where $f_{i,j}$ represents a residual rate (or distortion) function of $TiQj$, and $x_{i,j}$ is an independent variable of $TiQj$ that determines rate (or distortion) of a residual scalable block (e.g., quantization step size or the number of nonzero coefficients). In (1), conditions in the functions represent the fact that the quantity of interest (rate or distortion) is dependent on that of its preceding scalable blocks. In this example, three scalable blocks of $T0Q0$, $T1Q0$, and $T0Q1$ have influences on the R/D characteristics of a scalable block $T1Q1$. Since we employ q-domain residual R/D models in this paper, the R/D functions of a dependent scalable block $TiQj$ can be expressed as

$$\begin{gathered} R_{i,j}(q_{i,j}|q_{0,0}, \cdots, q_{i-1,j}, q_{i,j-1}) \text{ and} \\ D_{i,j}(q_{i,j}|q_{0,0}, \cdots, q_{i-1,j}, q_{i,j-1}) \end{gathered} \tag{2}$$

where $q_{i,j}$ is the quantization step size of $TiQj$. Now, the conditional R/D functions can be written to more general forms of the multivariate functions

$$R_{i,j}(\mathbf{Q}_{i,j}) \text{ and } D_{i,j}(\mathbf{Q}_{i,j}) \text{ with } \mathbf{Q}_{i,j} = \begin{pmatrix} q_{0,0} & \cdots & q_{0,j} \\ q_{1,0} & \cdots & q_{1,j} \\ \cdots & \cdots & \cdots \\ q_{i,0} & \cdots & q_{i,j} \end{pmatrix} \tag{3}$$

where $\mathbf{Q}_{i,j}$ is the matrix of parameters (i.e., quantization step sizes).

To examine the R/D characteristics of dependent scalable blocks, we make observations of the dependent R/D behaviors by rate-to-rate and distortion-to-distortion plots, where rate (or distortion) of independent reference layers is on the horizontal axis (domain), and the vertical axis (range) takes values of dependent rate (or distortion). Due to these rate-to-rate and distortion-to-distortion plots, we called this approach a self-domain (S-domain) analysis, which is a collective name for a rate-domain and a distortion-domain analyses. If we can observe strongly consistent behaviors of the dependent R/D characteristics with respect to the independent R/D characteristics, the dependent R/D functions can be expressed by certain combinations of the independent R/D functions. Then, we can develop dependent R/D models that convert multivariate functional expressions in (3) into univariate functional expressions. Here, it is important to understand that the dependent R/D functions still remain as functional expressions (of univariate functions instead of multivariate functions) by the proposed modeling approach. Hence, we need residual R/D functions of the independent scalable blocks to derive closed-form expressions of the dependent R/D functions.

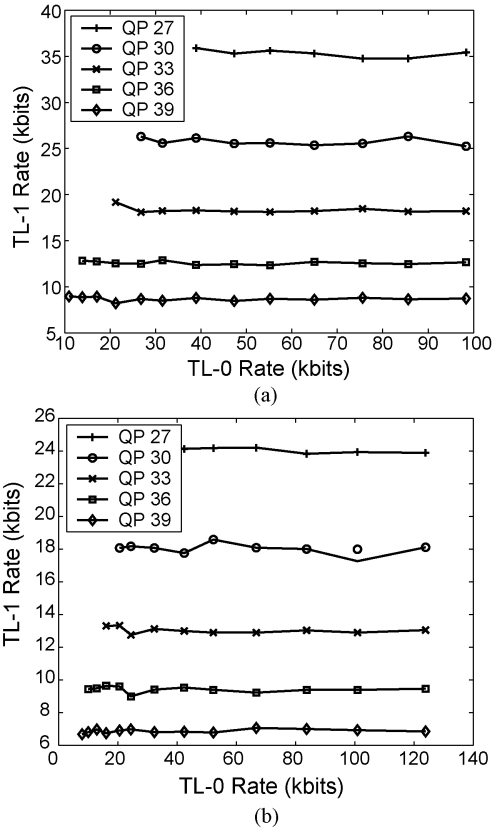

Fig. 3. TL rate dependency. A bi-variate rate function of a dependent scalable block is plotted as a function of the rate of the base TL block, i.e., $R_{0,0}(q_{0,0})$ versus $R_{1,0}(q_{0,0}, q_{1,0})$. In the legend, QP is a quantization parameter of a scalable block $T1Q0$ corresponding to $q_{1,0}$. (a) *Football*, QCIF. (b) *Foreman*, CIF.

A. S-Domain Observations

We begin with simple cases of a base TL and a base QL separately, i.e., TL-0 and QL-0. We first look at QL-0 ($T0Q0$, $T1Q0$,$\cdots$, and $TiQ0$) for the dependent R/D characterization in the T scalability (R/D dependency of TL blocks), and consider TL-0 ($T0Q0$, $T0Q1$, $\cdots$, and $T0Qj$) for the dependent R/D characterization in the Q scalability (R/D dependency of QL blocks). Hence, we have four cases of dependent rate and distortion in the T and Q scalability, respectively. We apply specific combinations of the quantization parameters (QPs) to the participating scalable blocks to observe the dependent R/D characteristics with respect to variations of those of the independent blocks.

1) *Case I: Rate dependency of a scalable block in the T scalability*: Fig. 3 demonstrates typical dependent rate characteristics of $T1Q0$ in the T scalability, where $(R_{0,0}(q_{0,0}), R_{1,0}(q_{0,0}, q_{1,0}))$ is plotted with the value of $q_{1,0}$ being constant for each curve. That is, rates of $T1Q0$ at fixed QPs in the legend of Fig. 3 are observed with respect to the variations of those of $T0Q0$. We see that the variations of the $T0Q0$ rates have almost no influence on the $T1Q0$ rates, i.e., $T1Q0$ rate is independent of $T0Q0$ rate. As a result, we can simplify the dependent rate function of $T1Q0$ to

$$R_{1,0}(q_{0,0}, q_{1,0}) \approx R_{1,0}(q_{1,0}) \tag{4}$$

where $q_{1,0}$ is the quantization step size of $T1Q0$.

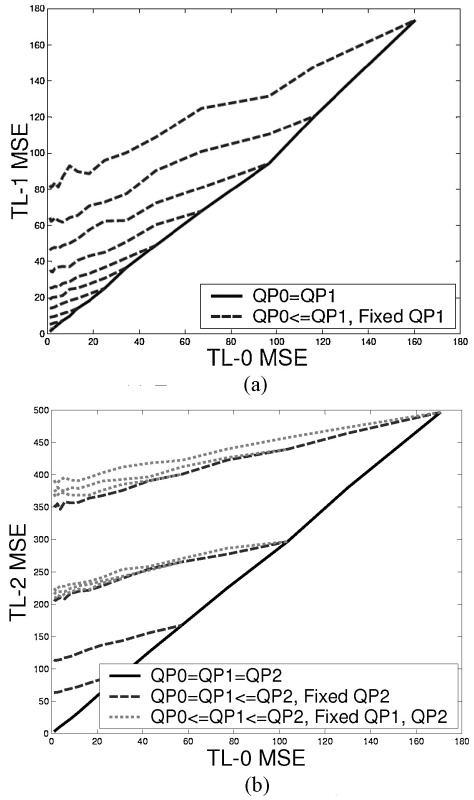

Fig. 4. TL distortion dependency. Dependent distortion functions of $T1Q0$ and $T2Q0$ ($D_{1,0}(q_{0,0}, q_{1,0})$ and $D_{2,0}(q_{0,0}, q_{1,0}, q_{2,0})$) are plotted with respect to variations of $T0Q0$ distortion ($D_{0,0}(q_{0,0})$). QPi is the quantization parameter of a scalable block $TiQ0$ corresponding to $q_{i,0}$. (a) *Foreman*, QCIF, TL-1. (b) *Football*, CIF, TL-2.

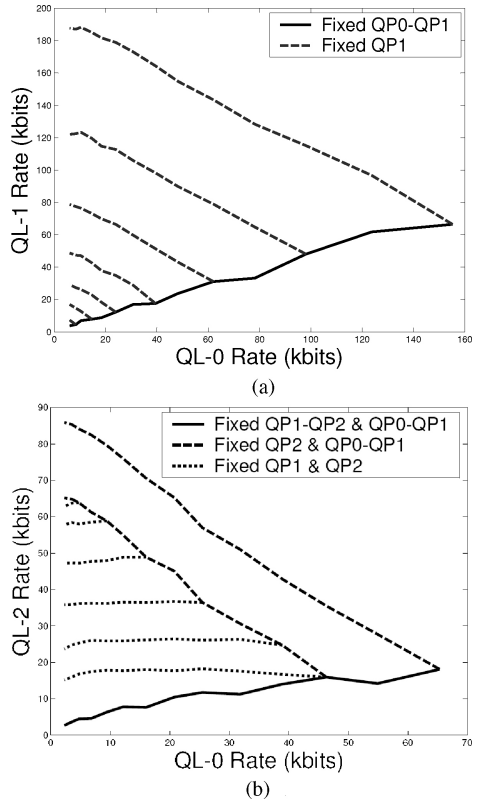

Fig. 5. QL rate dependency. Dependent rate functions of $T0Q1$ and $T0Q2$ ($R_{0,1}(q_{0,0}, q_{0,1})$ and $R_{0,2}(q_{0,0}, q_{0,1}, q_{0,2})$) are plotted with respect to variations of the rate of $T0Q0$ ($R_{0,0}(q_{0,0})$). QPj is the quantization parameter of $T0Qj$ corresponding to $q_{0,j}$. (a) *Soccer*, CIF. (b) *City*, QCIF.

2) *Case II: Distortion dependency of a scalable block in the T scalability*: Typical dependent distortion behaviors of scalable blocks in the T scalability are plotted with respect to variations of those of the base TL blocks ($T0Q0$) in Fig. 4. The distortion dependency of the TL blocks is more complicated than the rate dependency, and its approximation and simplification will be discussed in Sec. II-B.
3) *Case III: Rate dependency of a scalable block in the Q scalability*: Typical dependent rate behaviors of scalable blocks in the Q scalability are plotted with respect to the variations of those of the base QL blocks ($T0Q0$) in Fig. 5. Its approximation and simplification will be discussed in Sec. II-B.
4) *Case IV: Distortion dependency of a scalable block in the Q scalability*: Fig. 6 demonstrates behaviors of $D_{0,1}(q_{0,0}, q_{0,1})$ with respect to the variations of $D_{0,0}(q_{0,0})$. Similarly to Case I, the distortion of $T0Q0$ does not affect that of $T0Q1$. Hence, we can safely assume that $T0Q1$ distortion is a function of its own quantization step size

$$D_{0,1}(q_{0,0}, q_{0,1}) \approx D_{0,1}(q_{0,1}) \tag{5}$$

where $q_{0,j}$ is the quantization step size of $T0Qj$. With the QL distortion dependency, we have one more important observation in Fig. 7, where $T0Q2$ distortions, $D_{0,2}(q_{0,2})$, of various test sequences are plotted with respect to the $T0Q0$ distortion, $D_{0,0}(q_{0,2})$. As we can observe from the figure, $D_{0,2}(q_{0,2})$ is strongly correlated with $D_{0,0}(q_{0,2})$ when they have quantization step size in common. Therefore, we can simplify $T0Q2$ distortion function by

$$D_{0,2}(q_{0,0}, q_{0,1}, q_{0,2}) \approx D_{0,2}(q_{0,2}) \approx \mu_0^2 \cdot D_{0,0}(q_{0,2}) \tag{6}$$

where μ_0^2 is the distortion model parameter of $T0Q2$ representing the slopes of the lines in Fig. 7. Please note that subscripts and superscripts are used to indicate TL and QL indices. We will keep this notation in the following discussions as well.

From Cases I and IV, we could observe two important properties of the dependent R/D characteristics in the combined T-Q scalability:

1) *Property 1: Rate independence of a scalable block in the T scalability*: This property states that the rate of a scalable block is not influenced by the fidelity of TL references. For example, the rate of $T3Q2$ in Fig. 2 is not influenced by the rates of temporally preceding blocks $TiQj$ for $i < 3$. Hence, only the scalable blocks belonging to TL-3 in Fig. 2 need to be considered for the rate of $T3Q2$.
2) *Property 2: Distortion independence of a scalable block in the Q scalability*: Similarly to Property 1, this property states that the distortion of a scalable block is not

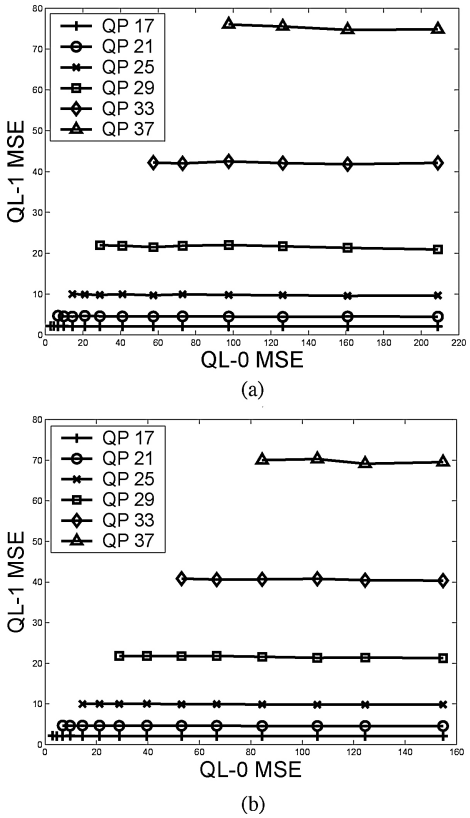

Fig. 6. QL distortion dependency. A bi-variate distortion function of a dependent scalable block is plotted as a function of the distortion of the base QL block, $D_{0,0}(q_{0,0})$ versus $D_{0,1}(q_{0,0}, q_{0,1})$. In the legend, QP is the quantization parameter for $T0Q1$ corresponding to $q_{0,1}$. (a) *City*. (b) *Soccer*. (a) *City*, QCIF. (b) *Soccer*, CIF.

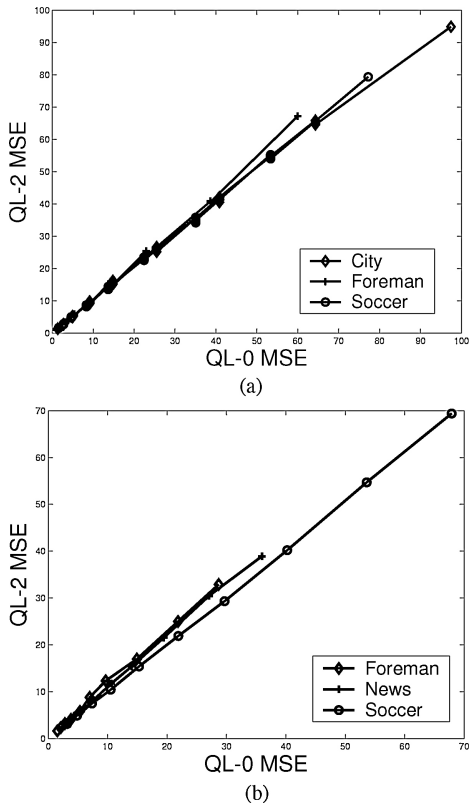

Fig. 7. Univariate distortion function, $D_{0,2}(q)$, of $T0Q2$ as a function of the distortion function, $D_{0,0}(q)$, of $T0Q0$. (a) QL-2 versus QL-0, QCIF. (b) QL-2 versus QL-0, CIF.

influenced by the fidelity of the QL references. That is, when we consider the distortion of $T3Q2$, we can safely ignore the influences of preceding scalable blocks in the Q scalability, i.e., $TiQj$ for $j < 2$. For this reason, only the scalable blocks belonging to QL-2 in Fig. 2 need to be considered for the distortion of $T3Q2$.

B. Interpretations of Cases II and III

From Properties 1 and 2, we know that Cases II and III (dependent distortion in the T scalability and dependent rate in the Q scalability) can be studied separately. We focus on the interpretations of Cases II and III that appear in Figs. 4 and 5. In this section, we change double subscripts (i, j) of the R/D function to a single subscript i or j depending on the discussion context for notational convenience. Because we made observations with QL-0 and TL-0, we will skip QL index for the analysis of the TL distortion dependency and TL index will be skipped for the analysis of QL rate dependency.

We choose variables (q_i's or q_j's) carefully in Figs. 4(b) and 5(b) so that the influence of these variables can clearly be identified. Both figures consist of solid, dashed, and dotted curves. The S-domain coordinate and q setting for each curve type are summarized in Table I with TL-2 and QL-2 scalable blocks ($T2Q0$ and $T0Q2$) as examples. We set q values such that only one variable is active with each curve. Then, we can isolate influences of individual variables on the dependent R/D behaviors. In the table, q is the only variable that determines function values, and q_i and q_j are constant values for each curve. For example, in a group of dashed curves in Figs. 4 and 5, q_2 for one dashed curve remains the same along the curve whereas values of q are varied to generate the curve on the plane. Δ in the table refers to a constant QP difference. To be precise, the notation has to be $q(QP - \Delta)$ based on the one-to-one relation between q and QP ($q(QP)$), but we use $q - \Delta$ for notational convenience.

We approximate R/D curves of Cases II and III in Figs. 4 and 5 based on two observed properties.

1) *Property 3: Linearity of the R/D characteristics of a scalable block in the T/Q scalability*: All the R/D curves in Figs. 4 and 5 can be linearly approximated with respect to their domains, i.e., independent rate or distortion function.
2) *Property 4: Parallelism of the R/D characteristics of a scalable block in the T/Q scalability*: All curves under the same q setting are approximately parallel. For example, all dashed and dotted curves form a group of parallel lines, respectively.

By Properties 3 and 4, we approximate Figs. 4(b) and 5(b) by parallel line segments as shown in Fig. 8, which demonstrate idealized approximations of the dependent distortion of the TL-2 block ($T2Q0$) in the T scalability and the dependent rate of the QL-2 block ($T0Q2$) with three groups of line segments. Each group of parallel lines can be characterized

TABLE I
COORDINATES AND q SETTINGS IN THE S-DOMAIN PLOTS

	TL Dependent Distortion (Fig. 4(b))	QL Dependent Rate (Fig. 5(b))
Solid Curve	$(D_0(q), D_2(q, q, q))$	$(R_0(q), R_2(q, q-\Delta, q-2\Delta))$
Dashed Curve	$(D_0(q), D_2(q, q, q_2))$ with $q \le q_2$	$(R_0(q), R_2(q, q-\Delta, q_2))$ with $q-\Delta > q_2$
Dotted Curve	$(D_0(q), D_2(q, q_1, q_2))$ with $q \le q_1 \le q_2$	$(R_0(q), R_2(q, q_1, q_2))$ with $q > q_1 > q_2$

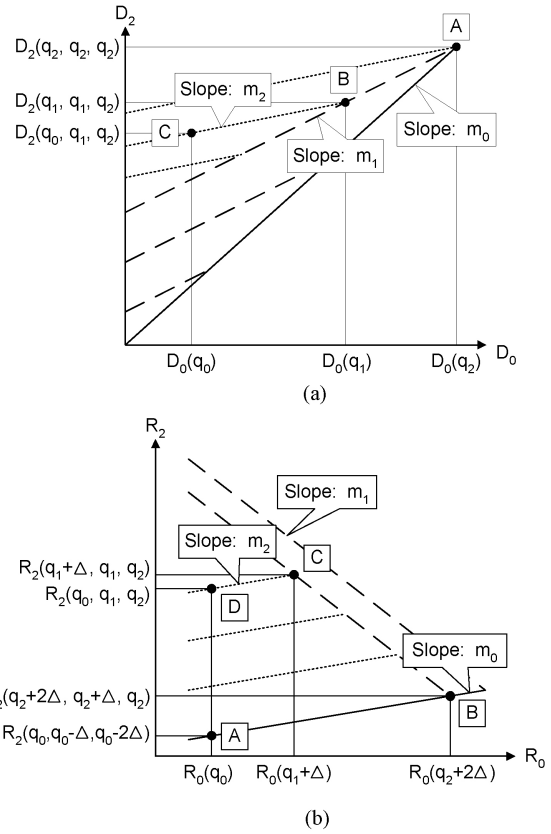

Fig. 8. Ideally approximated dependent R/D behaviors of scalable blocks. (a) Dependent distortion of a TL-2 block. (b) Dependent rate of a QL-2 block.

by a common slope of the group. That is, solid, dashed, and dotted line segments can be characterized simply by the slopes of m_0, m_1, and m_2 in Fig. 8.

The evaluation of each slope is straightforward if we assume that pivot points [A, B, and C in Fig. 8(a) and A, B, C, and D in Fig. 8(b)] are known. In the case of TL-2 block distortion function, the three slopes can be evaluated by

$$m_0 \approx \frac{D_2(q_2,q_2,q_2)}{D_0(q_2)} \quad (7)$$

$$m_1 \approx \frac{D_2(q_1,q_1,q_2)-D_2(q_2,q_2,q_2)}{D_0(q_1)-D_0(q_2)} \quad (8)$$

$$m_2 \approx \frac{D_2(q_0,q_1,q_2)-D_2(q_1,q_1,q_2)}{D_0(q_0)-D_0(q_1)}. \quad (9)$$

Now, we solve (7)–(9) for $D_2(q_0, q_1, q_2)$. Then, we have the tri-variate TL-2 block distortion function reduced to (11)

$$\begin{aligned} D_2(q_0, q_1, q_2) &= m_2 D_0(q_0) + (m_1 - m_2) D_0(q_1) + \\ &\quad (m_0 - m_1) D_0(q_2) \\ &= \zeta_{2,0} D_0(q_0) + \zeta_{2,1} D_0(q_1) + \zeta_{2,2} D_0(q_2) \end{aligned} \quad (10)$$

where $\zeta_{i,k}$ represents the kth model parameter of a scalable block at TL-i, and q_k is quantization step size of a scalable block at TL-k. Similarly, the tri-variate rate function of a scalable block at QL-2 can be expressed as

$$\begin{aligned} R_2(q_0, q_1, q_2) &= m_2 R_0(q_0) + (m_1 - m_2) R_0(q_1 + \Delta) + \\ &\quad (m_0 - m_1) R_0(q_2 + 2\Delta) + \eta^2 \\ &= \xi^{2,0} R_0(q_0) + \xi^{2,1} R_0(q_1 + \Delta) + \\ &\quad \xi^{2,2} R_0(q_2 + 2\Delta) + \eta^2 \end{aligned} \quad (11)$$

where $\xi^{j,k}$ is the kth model parameter of a scalable block at QL-j, η^j is another model parameter of a scalable block at QL-j, q_k is quantization step size of a scalable block at QL-k, and Δ is a predetermined constant that represents a fixed QP difference between two consecutive QLs.

We can get (11) by evaluating m_0, m_1, m_2, and η^2 from pivots A, B, C, and D in Fig. 8(b), and by solving (12)–(15) for $R_2(q_0, q_1, q_2)$

$$m_0 \approx \frac{R_2(q_0, q_0-\Delta, q_0-2\Delta)-R_2(q_2+2\Delta, q_2+\Delta, q_2)}{R_0(q_0)-R_0(q_2+2\Delta)} \quad (12)$$

$$m_1 \approx \frac{R_2(q_1 + \Delta, q_1, q_2) - R_2(q_2 + 2\Delta, q_2 + \Delta, q_2)}{R_0(q_1 + \Delta) - R_0(q_2 + 2\Delta)} \quad (13)$$

$$m_2 \approx \frac{R_2(q_0, q_1, q_2) - R_2(q_1 + \Delta, q_1, q_2)}{R_0(q_0) - R_0(q_1 + \Delta)} \quad (14)$$

$$\eta^2 \approx R_0(q_0, q_0 - \Delta, q_0 - 2\Delta) - m_0 \cdot R_0(q_0). \quad (15)$$

By the S-domain analysis, we could learn four important properties of the dependent R/D characteristics in the combined T-Q scalability. First, we could identify independent properties of the dependent scalable blocks, i.e., Properties 1 and 2, which lead to significant reduction in the number of variables, e.g., (4) and (6). More importantly, complex multivariate dependent R/D functions (TL dependent distortion and QL dependent rate functions) are decomposed into linear combinations of univariate R/D functions evaluated at participating layers' quantization step sizes, i.e., (11), and (11). Generally, we can express a multivariate R/D functions as

$$\begin{aligned} f_i(q_0, \cdots, q_i) &= f_i(q_i | q_0, \cdots, q_{i-1}) \\ &= c_0 \cdot f_0(q_0) + \cdots + c_i \cdot f_0(q_i) + c_{i+1} \end{aligned} \quad (16)$$

where c_i's are model parameters, $f_i(\cdot)$ is rate or distortion function of layer i and $f_0(\cdot)$ is a residual rate or distortion function of a base layer block. Now, the multivariate function given in the left-hand-side of (16) is successfully decomposed into a linear combination of the univariate functions in the right-hand-side of (16). Hence, we have derived tractable forms of the dependent R/D functions of scalable blocks.

TABLE II
MODELING ACCURACY (%) OF EACH SCALABLE BLOCK

Sequence	Format		T1Q0	T2Q0	T0Q1	T1Q1	T2Q1	T0Q2	T1Q2	T2Q2
City	QCIF	Rate	•	•	85.18	90.94	83.67	81.26	85.72	83.76
		MSE	97.14	96.32	95.33	96.56	93.40	86.07	86.30	85.60
	CIF	Rate	•	•	82.38	89.56	90.88	78.59	82.64	85.43
		MSE	98.63	97.66	93.92	95.07	96.57	85.10	85.60	85.16
Foreman	QCIF	Rate	•	•	86.32	91.12	86.94	83.47	90.96	85.05
		MSE	95.79	96.87	96.96	96.74	93.02	94.60	90.31	91.99
	CIF	Rate	•	•	80.63	89.61	87.24	81.97	89.96	92.73
		MSE	96.11	97.73	97.03	92.19	96.60	93.40	92.74	93.56
News	QCIF	Rate	•	•	82.24	87.90	86.31	90.09	92.08	91.87
		MSE	96.91	97.98	97.06	96.20	96.54	92.20	94.31	94.90
	CIF	Rate	•	•	80.21	78.34	90.44	92.27	72.91	80.94
		MSE	98.34	99.23	97.84	97.41	98.17	95.34	93.94	93.64
Soccer	QCIF	Rate	•	•	87.50	85.21	83.88	82.55	89.82	89.95
		MSE	96.65	97.22	94.86	95.28	92.67	90.19	94.59	90.79
	CIF	Rate	•	•	89.31	86.88	92.90	80.48	89.46	90.18
		MSE	95.52	96.71	94.52	97.05	97.70	78.06	87.28	89.09

III. JOINT TEMPORAL-QUALITY R/D MODELS FOR GOP

A. Model Derivation

By Properties 1 and 2, we can isolate a scalability dimension for the derivation of the dependent R/D functions of a scalable block $TiQj$. That is, we can safely ignore the influences of preceding TL blocks for a dependent rate function, and similarly, a dependent distortion function can be derived without considering preceding QL blocks' influences.

By (4) and (11), we can simplify the dependent rate function in (3) to

$$\begin{aligned} R_{i,j}(\mathbf{Q}_{i,j}) &= R_{i,j}(\mathbf{q}_i^t) \\ &= \xi_i^{j,0} R_{i,0}(q_{i,0}) + \xi_i^{j,1} R_{i,0}(q_{i,1}+\Delta) + \cdots \\ &\quad + \xi_i^{j,j} R_{i,0}(q_{i,j}+j\Delta) + \eta_i^j \\ &= \sum_{k=0}^{j} \xi_i^{j,k} R_{i,0}(q_{i,k}+k\Delta) + \eta_i^j \end{aligned} \tag{17}$$

where $\mathbf{q}_i$ is the ith row of $\mathbf{Q}_{i,j}$, $R_{i,0}(q)$ is a residual rate function of $TiQ0$, $q_{i,k}$ is the quantization step size of $TiQk$, and $\xi_i^{j,k}$ and η_i^j are rate model parameters of $TiQj$. The first reduction in (17), i.e., from a matrix of variables to a vector of variables, comes from Property 1 that all TL influences can be ignored for dependent rates of scalable blocks. Similarly, the distortion function in (3) reduces to

$$\begin{aligned} D_{i,j}(\mathbf{Q}_{i,j}) &= D_{i,j}(\mathbf{q}_j) \\ &= \zeta_{i,0}^j D_{0,j}(q_{0,j}) + \zeta_{i,1}^j D_{0,j}(q_{1,j}) + \cdots \\ &\quad + \zeta_{i,i}^j D_{0,j}(q_{i,j}) \\ &= \zeta_{i,0}^j \mu_0^j D_{0,0}(q_{0,j}) + \zeta_{i,1}^j \mu_0^j D_{0,0}(q_{1,j}) + \cdots \\ &\quad + \zeta_{i,i}^j \mu_0^j D_{0,0}(q_{i,j}) \\ &= \mu_0^j \sum_{k=0}^{i} \zeta_{i,k}^j D_{0,0}(q_{k,j}) \end{aligned} \tag{18}$$

where $\mathbf{q}_j$ is the jth column of $\mathbf{Q}_{i,j}$, $D_{0,0}(q)$ is a residual distortion function of $T0Q0$, $q_{k,j}$ is the quantization step size of $TkQj$, and $\zeta_{i,k}^j$ and μ_i^j are distortion model parameters of $TiQj$. Similarly to (17), Property 2 leads to the first reduction in (18) by allowing the cancelation of QL influences on the dependent distortions of scalable blocks. In (18), it is worthwhile noting that the dependent distortion model has one more reduction step than that of the dependent rate function by (6), i.e., $D_{0,j}(q) = \mu_0^j \cdot D_{0,0}(q)$.

Finally, we can derive GOP dependent R/D models from (17) and (18) simply by adding rates and distortions of participating scalable blocks in a GOP of N_T TLs and N_Q QLs as

$$\begin{aligned} R_{GOP}(\mathbf{Q}) &= \sum_{i=0}^{N_T-1} \sum_{j=0}^{N_Q-1} R_{i,j} \\ &= \sum_{i=0}^{N_T-1} \sum_{j=0}^{N_Q-1} \left(\sum_{k=0}^{j} \xi_i^{j,k} R_{i,0}(q_{i,k}+k\Delta) + \eta_i^j \right) \end{aligned}$$

and

$$\begin{aligned} D_{GOP}(\mathbf{Q}) &= \sum_{i=0}^{N_T-1} \sum_{j=0}^{N_Q-1} D_{i,j} \\ &= \sum_{j=0}^{N_Q-1} \sum_{i=0}^{N_T-1} \left(\mu_0^j \sum_{k=0}^{i} \zeta_{i,k}^j D_{0,0}(q_{i,k}) \right) \\ &= \sum_{j=0}^{N_Q-1} \sum_{i=0}^{N_T-1} , \omega_{i,j} D_{0,0}(q_{i,j}) \end{aligned} \tag{19}$$

where $\omega_{i,j} = \mu_0^j \sum_{k=i}^{N_T-1} \zeta_{k,i}^j$ is the model parameter.

B. Model Verification

To verify the proposed dependent R/D models, estimated R/D values by the models are compared with the actual R/D values. In the combined T-Q scalability of three TLs and QLs, R/D samples are generated with various QP combinations to provide actual R/D values. Estimated R/D values corresponding to the QP combinations are computed from the R/D models. To get model parameters, we follow the steps in Fig. 8. We first generate pivot R/D points by actual

TABLE III
QL MODELING ACURACY (%)

	Format	QL-0		QL-1		QL-2	
		Rate	MSE	Rate	MSE	Rate	MSE
City	QCIF	94.44	97.56	91.16	95.71	92.42	88.41
	CIF	96.64	98.57	92.79	95.45	91.54	87.75
Foreman	QCIF	98.23	96.57	93.97	93.75	91.02	88.93
	CIF	97.16	97.76	89.95	96.57	88.65	94.55
News	QCIF	98.57	96.13	95.51	96.63	95.62	95.08
	CIF	97.87	99.30	95.98	98.05	96.60	95.32
Soccer	QCIF	99.12	96.28	92.42	91.39	90.24	92.10
	CIF	97.40	96.53	91.96	97.56	93.22	89.27

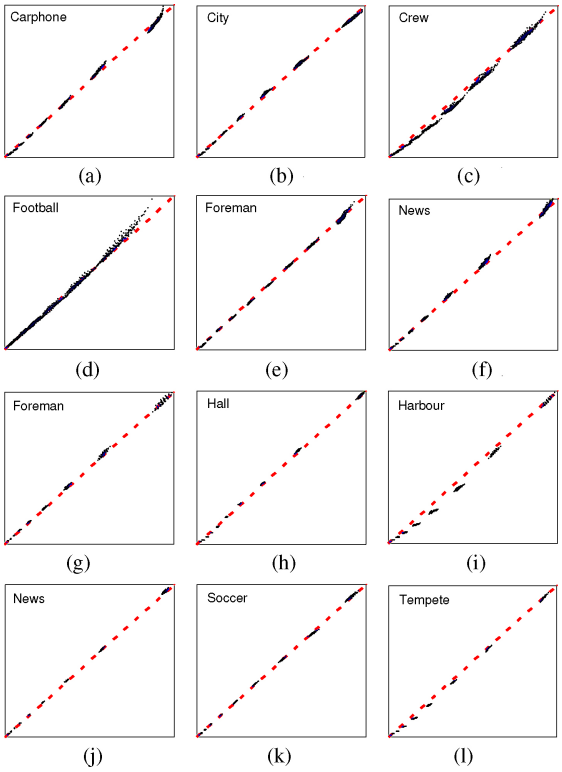

Fig. 9. Verification of the GOP-based (16 frames per GOP so that one GOP consists of 5 T layers) dependent distortion model in the T scalability, where the x-axis is the actual MSE of the GOP while the y-axis is the estimated MSE of the GOP by the proposed dependent distortion model. (a)–(f) QCIF. (g)–(l) CIF test sequences.

encoding, evaluated the slopes (m_i's) by (7)–(9) and (12)–(15), and finally, the model parameters are evaluated from m_i's with Δ set to two in our experiment.

We first show the modeling accuracy of each scalable block in Table II. Because QL-0 block rates serve as the independent basis functions for the dependent rates of scalable blocks, we do not have entries for $T1Q0$ and $T2Q0$ in Table II. The accuracy is computed by

$$\text{accuracy} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(1 - \frac{|s_i - \hat{s}_i|}{s_i}\right) \times 100 \qquad (20)$$

where N_s is the number of samples and s_i and $\hat{s}_i$ are the actual and the estimated R-D values, respectively. Table III shows QL R/D modeling results, where a QL is composed of three TLs.

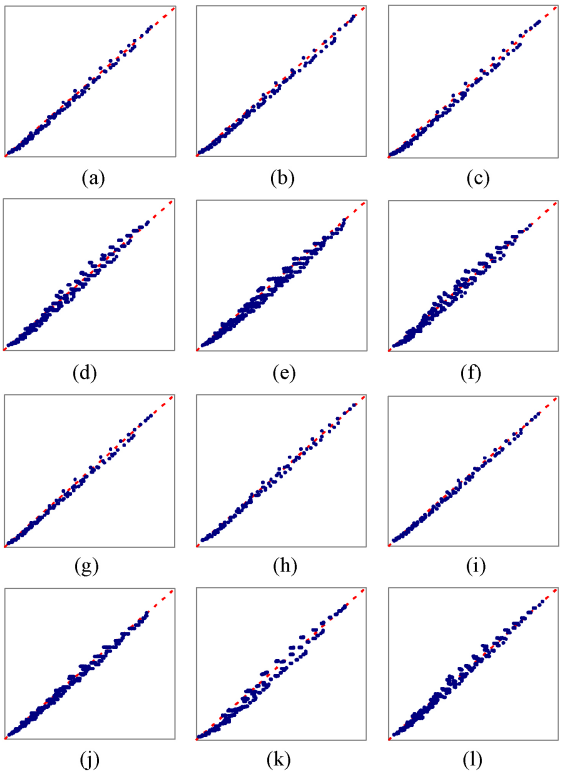

Fig. 10. Verification of the GOP-based dependent rate model in the Q scalability, where the x-axis is the actual rate of a QL, and the y-axis is the estimated rate of the QL by the proposed GOP-based QL rate model. (a)–(f) QCIF test sequences. (g)–(l) CIF test sequences.

We also provide graphical modeling accuracy verification in Figs. 9 and 10, where the estimated R-D values are plotted with respect to the actual R-D values with the identity ($y = x$) line on the diagonal. Fig. 9 demonstrates the GOP based TL distortion modeling results, where 2730 and 810 distortion samples are generated for each QCIF and CIF test sequence, respectively. Fig. 10 shows the QL rate modeling results, where 105 and 560 QL-1 and QL-2 R/D samples are generated for the the verification, respectively.

To summarize, the dependent rate model achieves the average accuracy in the range of 90%, and the average estimation accuracy of the dependent distortion model is greater than 90% for our test sequences.

IV. MODEL PARAMETERS ANALYSIS

In the proposed models, R/D functions of the dependent scalable blocks are represented by linear combination of the base scalable block R/D functions evaluated at the quantization step sizes of participating scalable blocks as in (17) and (18). Generally, weights in a linear combination refer to the contribution of the corresponding basis function to the target quantity of the decomposition. For this reason, in the proposed R/D models, the model parameters indirectly represent the influence of a quantization choice for each block with its independent basis function as an intermediate medium. Moreover, they make the optimization feasible by enabling partial differentiations of the dependent R/D functions with

TABLE IV
PARAMETERS OF DEPENDENT DISTORTION MODEL IN THE T SCALABILITY

Sequence	TL	QCIF					CIF				
		ζ_0	ζ_1	ζ_2	ζ_3	$\sum$	ζ_0	ζ_1	ζ_2	ζ_3	$\sum$
Hall	0	1	0	0	0	1	1	0	0	0	1
	1	0.985	0.015	0	0	1.000	0.924	0.082	0	0	1.006
	2	1.979	0.003	0.032	0	2.014	1.837	0.007	0.163	0	2.007
	3	3.948	0.010	0.010	0.047	4.015	3.679	0.033	0.027	0.291	4.030
	ω	7.912	0.028	0.042	0.047	8.029	7.440	0.122	0.190	0.291	8.043
Soccer	0	1	0	0	0	1	1	0	0	0	1
	1	0.495	0.547	0	0	1.042	0.642	0.466	0	0	1.108
	2	1.109	0.175	0.808	0	2.092	1.314	0.163	0.654	0	2.131
	3	2.322	0.398	0.366	0.899	3.985	2.664	0.377	0.186	0.976	4.203
	ω	4.926	1.120	1.174	0.899	8.119	5.620	1.006	0.840	0.976	8.442

respect to the individual variables

$$f_i(x_0, \cdots, x_i) = \sum_{k=0}^{i} c_k f_0(x_k) \;\Rightarrow$$
$$\frac{\partial}{\partial x_k} f_i(x_0, \cdots, x_i) = c_k \frac{d}{dx_k} f_0(x_k) \;\text{ for }\; k = 0, \cdots, i \tag{21}$$

where we assume that $f_0(x)$ is differentiable.

Table IV shows the values of distortion model parameters of selected test sequences. To understand the physical meaning of the parameters, we look at the slopes in Fig. 8(a) that demonstrate behaviors of the distortion of a dependent scalable block with respect to that of its base TL block. As we can observe from Fig. 4, a low motion sequence [*Foreman*, Fig. 4(a)] demonstrates steeper slopes than those of a high motion sequence [*Football*, Fig. 4(b)]. Because the model parameters are determined by the difference between the slopes except for the one corresponding to q_0, steeper slope indicates smaller model parameter values for dependent scalable blocks (ζ_1, ζ_2, and ζ_3 in Table IV), and more influence of TL-0 block (ζ_0 in Table IV).

This interpretation exactly coincides with the intuition that higher reference quality has more influence on the coding efficiency of low motion (i.e., more temporally dependent) than that of high motion sequences. For example, the GOP distortion model parameters (ω's) in Table IV demonstrate the contribution of each TL block distortion on the GOP distortion, and TL-0 block parameter (ω_0) of low motion sequences (e.g., *City* and *Hall*) are much higher than that of high motion sequences (e.g., *Football* and *Soccer*). More importantly, it has to be noted that the model parameters quantify the influence of references' quality on the coding efficiency of the dependent scalable blocks in the T scalability.

To understand the physical implications of QL dependent rate model parameters, we first look at the parameter values of the intracoded frames in Table V. Because the rate of a compressed frame is directly related to the entropy of a quantized residual image, we assume that the rate of a frame is equivalent to its entropy in the following discussion. That is

$$R(q_0, \cdots, q_j) \equiv H(X_{q_0}, \cdots, X_{q_j}) \;\text{ and}$$
$$R_j(q_j|q_0, \cdots, q_{j-1}) \equiv H(X_{q_j}|X_{q_0}, \cdots, X_{q_{j-1}}) \tag{22}$$

where $R(q_0, \cdots, q_j)$ is the rate of $(j+1)$ QLs, $R_j(q_j|q_0, \cdots, q_{j-1})$ is the rate of $(j+1)$st QL block, $H(\bullet)$ is the entropy of given random variables, and X_q refers to a residual image quantized by the quantization step size q.

Roughly speaking, we can assume that a residual image with finer quantization can be considered a super set of more coarsely quantized residual images of a frame with multiple QL blocks. Mathematically, we have

$$X_{q_0} \subset X_{q_1} \subset \cdots \subset X_{q_j}, \;\text{ for }\; q_0 > \cdots > q_j \tag{23}$$

where q_k is the quantization step size for k^{th} QL block. Given (23), the entropy relation between two QL blocks can be derived as

$$\begin{aligned} H(X_{q_0}, X_{q_1}) &= H(X_{q_0}) + H(X_{q_1}) - I(X_{q_0}; X_{q_1}) \\ &= H(X_{q_1}), \;\text{ and} \\ H(X_{q_1}|X_{q_0}) &= H(X_{q_1}) - I(X_{q_0}; X_{q_1}) \\ &= H(X_{q_1}) - H(X_{q_0}), \;\text{ for }\; q_0 > q_1 \end{aligned} \tag{24}$$

where $I(X_{q_0}; X_{q_1}) = H(X_{q_0})$ is the mutual information between two random variables X_{q_0} and X_{q_1}.

Without loss of generality, we can generalize (24) to

$$\begin{aligned} H(X_{q_0}, \cdots, X_{q_j}) &= H(X_{q_j}), \;\text{ and} \\ H(X_{q_j}|X_{q_0}, \cdots, X_{q_{j-1}}) &= H(X_{q_j}|X_{q_{j-1}}) \\ = H(X_{q_j}) - H(X_{q_{j-1}}), &\;\text{ for }\; q_0 > \cdots > q_j \end{aligned} \tag{25}$$

where the first equation is inferred from the inclusive relation between QL blocks in (23). In (25), the first equation explains that the sum of QL block rates of a frame is expected to be equal to the rate of the frame in a single layer encoding instance with the highest QL quantization step size, and the second equation explains that the rate of an enhancement QL block is determined by the difference in the information quantities that are covered by the quantization step sizes of the highest and the second highest enhancement QL blocks.

The model parameter values for intracoded frames in Table V follow the analysis in (25) very closely. First, the model parameter values of the highest and the second highest QL blocks are much larger than those of lower QL block parameters. This implies that the rate of a QL block is determined mainly by the quantization of the top two QL blocks, which is an identical conclusion to the second equation in (25). Second, the sum of model parameter values for each QL block is close to zero except for those of QL-2 block

TABLE V
PARAMETERS OF DEPENDENT RATE MODEL IN THE Q SCALABILITY

Intracoded Frames									
Sequence	QL	QCIF				CIF			
		ξ^0	ξ^1	ξ^2	η	ξ^0	ξ^1	ξ^2	η
City	0	1	0	0	0	1	0	0	0
	1	-0.96	1.16	0	1455.90	-0.96	1.15	0	4709.13
	2	0.01	-1.09	1.32	2115.37	0.00	-1.09	1.32	6749.59
	$\sum$	0.05	0.07	1.32	3571.27	0.04	0.06	1.32	11498.72
Football	0	1	0	0	0	1	0	0	0
	1	-0.95	1.12	0	2116.38	-0.98	1.15	0	5688.02
	2	-0.02	-1.06	1.28	2483.79	-0.03	-1.07	1.32	7016.81
	$\sum$	0.03	0.06	1.28	4600.17	-0.01	0.08	1.32	12704.83
Foreman	0	1	0	0	0	1	0	0	0
	1	-0.98	1.19	0	897.04	-1.01	1.24	0	1712.63
	2	0.01	-1.10	1.37	1421.92	0.03	-1.14	1.45	2061.73
	$\sum$	0.03	0.09	1.37	2318.96	0.02	0.10	1.45	3774.46
Intercoded Frames									
Sequence	QL	QCIF				CIF			
		ξ^0	ξ^1	ξ^2	η	ξ^0	ξ^1	ξ^2	η
City	0	1	0	0	0	1	0	0	0
	1	-0.71	1.18	0	421.97	-0.74	1.21	0	899.32
	2	0.06	-0.84	1.45	432.98	0.05	-0.89	1.47	1369.86
	$\sum$	0.35	0.34	1.45	854.95	0.31	0.32	1.47	2269.18
Football	0	1	0	0	0	1	0	0	0
	1	-0.91	1.14	0	2066.66	-0.92	1.17	0	3585.22
	2	-0.02	-0.99	1.28	2110.55	0.00	-1.01	1.35	4672.50
	$\sum$	0.07	0.15	1.28	4178.21	0.08	0.16	1.35	8257.72
Foreman	0	1	0	0	0	1	0	0	0
	1	-0.85	1.21	0	388.53	-0.90	1.25	0	551.35
	2	0.08	-0.97	1.42	522.92	0.09	-1.03	1.51	442.26
	$\sum$	0.23	0.24	1.42	911.45	0.19	0.22	1.51	993.61

parameters. This implies that the rate of a TL that is equivalent to the sum of participating QL blocks' rates is determined mainly by the rate of the highest QL block

$$R_i(q_{i,0}, \cdots, q_{i,j}) \approx \xi_i^{j,j} \cdot R_{i,0}(q_{i,j} + j \cdot \Delta) + \sum_{k=0}^{j} \eta_i^k \tag{26}$$

where $\sum_{k=0}^{j} \eta_i^k$ is the overhead of having QLs.

From Table V, it is worthwhile noting that $\xi_i^{j,j}$ takes values greater than one differently from the first equation in (25), which we can consider the compensation for the Δ term in the model. Interestingly, the difference between $\xi^{j,j}$'s of adjacent QLs is quite constant according to Table V, and the values of $\xi^{j,j} - \xi^{0,0}$ become proportional to j. That is

$$\begin{aligned} \xi^{j,j} - \xi^{j-1,j-1} \approx \xi^{2,2} - \xi^{1,1} \approx \xi^{1,1} - \xi^{0,0} \approx d \\ \Rightarrow \xi^{j,j} \approx \xi^{0,0} + d \cdot j \end{aligned} \tag{27}$$

which validates the above statements that the $\xi^{j,j}$ values greater than 1 compensate for the Δ term in the model.

Even though the model parameters of intracoded QL blocks are well explained by the above analysis, those of intercoded QL blocks show one disagreement from the analysis in that the sum of the rate model parameters of sub QL blocks is greater than zero. This is because the assumption in (23) does not hold among QL block residual images. That is, the mutual information between two consecutive QL blocks is not equal to the lower layer entropy, i.e., $I(X_{q_{j-1}}; X_{q_j}) \neq H(X_{q_{j-1}})$ because of motion compensated prediction of interframes. Hence, the approximation in (26) does not hold, and the rates of all participating QL blocks need to be considered to estimate the rate of a TL. However, all other explanations based on the entropy relation should still hold even for intercoded QL blocks, and it can be seen from the model parameter values in Table V.

V. JOINT T-Q LAYER BIT ALLOCATION

In this section, we investigate a joint T-Q layer bit allocation problem as an application of the proposed dependent R/D models. We consider a combined T-Q scalability as demonstrated in Fig. 2, where a scalable block is specified as $Ti\,Q\,j$.

A. Problem Formulation

The joint T-Q bit allocation problem is formulated as an optimal QP (equivalently, q) decision problem, which minimizes the GOP distortion under a target bit rate for each QL of a GOP. Each scalable block in the T-Q plane (Fig. 2) is considered a bit allocation unit. Mathematically, we have

$$\begin{aligned} \mathbf{Q}^* = \underset{\mathbf{Q} \in \mathcal{Q}^{N_Q} \times \mathcal{Q}^{N_T}}{\arg\min} \; D_{GOP}(\mathbf{Q}) \text{ suject to } R_0(\mathbf{q}_0) \leq R_{T,0}, \\ R_1(\mathbf{q}_1) \leq R_{T,1}, \cdots, \text{and } R_{N_Q-1}(\mathbf{q}_{N_Q-1}) \leq R_{T,N_Q-1} \end{aligned} \tag{28}$$

where $R_j(\mathbf{q}_j)$ is the rate of QL-j, $\mathbf{Q}$ and $\mathbf{q_j}$ are the $N_Q \times N_T$ matrix and the $N_T \times 1$ vector whose elements are the quantization step sizes (i.e., q values) of the participating scalable blocks, $\mathcal{Q}$ is the space of all admissible quantization step sizes and $R_{T,j}$ is the target bit budget for QL-j.

The Lagrangian formulation of the constrained problem in (28) leads to the following unconstrained optimization problem:

$$\begin{aligned} J(\mathbf{Q}^*, \Lambda^*) &= \underset{\mathbf{Q}\in\mathcal{Q}^{N_Q}\times\mathcal{Q}^{N_T}, \Lambda\in\mathcal{R}^{N_Q}}{\arg\min} J(\mathbf{Q}, \Lambda) \\ &= \sum_{i=0}^{N_T-1}\sum_{j=0}^{N_Q-1} D_{i,j} \\ &\quad +\lambda_0\left(\sum_{i=0}^{N_T-1} R_{i,0} - R_{T,0}\right) + \cdots \\ &\quad +\lambda_{N_Q-1}\left(\sum_{i=0}^{N_T-1} R_{i,N_Q-1} - R_{T,N_Q-1}\right) \end{aligned} \tag{29}$$

where λ_j's are the Lagrange multipliers. With the proposed R/D models, we rewrite the Lagrange cost function in (29) as (30).

Finally, (30) can be written in a closed-form expression by the residual R/D models given in [7]

$$R(q) = a \cdot q^{-\alpha} \quad \text{and} \quad D(q) = b \cdot q^{\beta} \tag{31}$$

where a, b, α and β are model parameters. Finally, the optimization problem becomes (32), where i, j, and k are TL and QL scalable block indices, and thus, they are not used for any mathematical operation such as an exponent.

B. Solution to Lagrangian

The complex GOP R/D functions are decomposed into a linear sum of simple univariate functions by the proposed GOP R/D models. Hence, the optimization problem in (32) can be solved by deriving the partial derivatives with respect to $q_{i,j}$'s and λ_j's, which result in a system of nonlinear equations. The independence of the variables allows us to use the partial derivatives to find the solution. Mathematically, we have

$$\begin{aligned} \frac{\partial J(\mathbf{Q}, \Lambda)}{\partial q_{i,j}} &= \omega_{i,j}\cdot b\cdot\beta\cdot q_{i,j}^{\beta-1} \\ &\quad -\sum_{k=j}^{N_Q-1}\lambda_k\cdot\xi_i^{k,j}\cdot a_i\cdot\alpha_i\cdot q_{i,j}^{-\alpha_i-1} \\ &= 0 \quad \text{and} \\ \frac{\partial J(\mathbf{Q}, \Lambda)}{\partial \lambda_j} &= \sum_{i=0}^{N_T-1}\sum_{k=0}^{j}\left(\xi_i^{j,k}\cdot a_i\cdot q_{i,k}^{-\alpha_i}+\eta_i^k\right) - R_{T,j} \\ &= 0. \end{aligned} \tag{33}$$

Since the numbers of variables and equations are the same, $N_Q\times N_T+N_Q$, the solution to the system of nonlinear equations in (33) is feasible, and it is solved by the gradient method. To apply the gradient method, we define a new equation, where the sum of squares of all partial derivatives in (33) is set to zero, namely

$$g(\mathbf{Q}, \Lambda) = \sum_{i=0}^{N_T-1}\sum_{j=0}^{N_Q-1}\left(\frac{\partial J(\mathbf{Q}, \Lambda)}{\partial q_{i,j}}\right)^2 + \sum_{j=0}^{N_Q-1}\left(\frac{\partial J(\mathbf{Q}, \Lambda)}{\partial \lambda_j}\right)^2 = 0. \tag{34}$$

Then, the solution, ($\mathbf{Q}^*$ and Λ^*), is determined as the values that make $g(\mathbf{Q}, \Lambda)$ closest to zero.

C. Experimental Result

The performance of the proposed algorithm is examined on various test sequences in CIF (352x288), 480p (854x480), and 720p HD (1280x720) formats. As the performance benchmark, the FixedQpEncoder tool implemented in JSVM [22] is employed, which iterates encoding loop until layer target bit rates are satisfied. In the experiments, each output video contains three TLs and three QLs, where every TL-0 frame is encoded as a P-frame except for the first I-frame. The average Y-PSNR performance with respect to the bit rates in full T-Q resolution is provided in Fig. 11 and Table VI. Clearly, the proposed bit allocation algorithm outperforms the benchmark by a substantial margin in all cases.

The comparison of coding efficiency at each QL is provided in Table VII. We see that the proposed joint T-Q bit allocation algorithm can produce much more R-D efficient bit stream at each QL than that by the benchmark, which verifies the

$$\begin{aligned} J(\mathbf{Q}, \Lambda) &= \sum_{j=0}^{N_Q-1}\sum_{i=0}^{N_T-1}\omega_{i,j}D_{0,0}(q_{i,j}) + \lambda_0\left(\sum_{i=0}^{N_T-1}R_{i,0}(q_{i,0}) - R_{T,0}\right) + \cdots \\ &\quad +\lambda_{N_Q-1}\left(\sum_{i=0}^{N_T-1}\sum_{k=0}^{N_Q-1}\left(\xi_i^{N_Q-1,k}R_{i,0}(q_{i,k}) + \eta_k^j\right) - R_{T,N_Q-1}\right). \end{aligned} \tag{30}$$

$$\begin{aligned} J(\mathbf{Q}, \Lambda) &= \sum_{i=0}^{N_T-1}\sum_{j=0}^{N_Q-1}\omega_{i,j}\cdot b\cdot q_{i,j}^{\beta} + \lambda_0\cdot\left(\sum_{i=0}^{N_T-1}a_i\cdot q_{i,0}^{-\alpha_i} - R_{T,0}\right) + \lambda_1\cdot\left(\sum_{i=0}^{N_T-1}\sum_{k=0}^{1}\left(\xi_i^{1,k}\cdot a_i\cdot q_{i,k}^{-\alpha_i}+\eta_i^1\right) - R_{T,1}\right) \\ &\quad +\cdots+\lambda_{N_Q-1}\cdot\left(\sum_{i=0}^{N_T-1}\sum_{k=0}^{N_Q-1}\left(\xi_i^{N_Q-1,k}\cdot a_i\cdot q_{i,k}^{-\alpha_i}+\eta_i^k\right) - R_{T,N_Q-1}\right). \end{aligned} \tag{32}$$

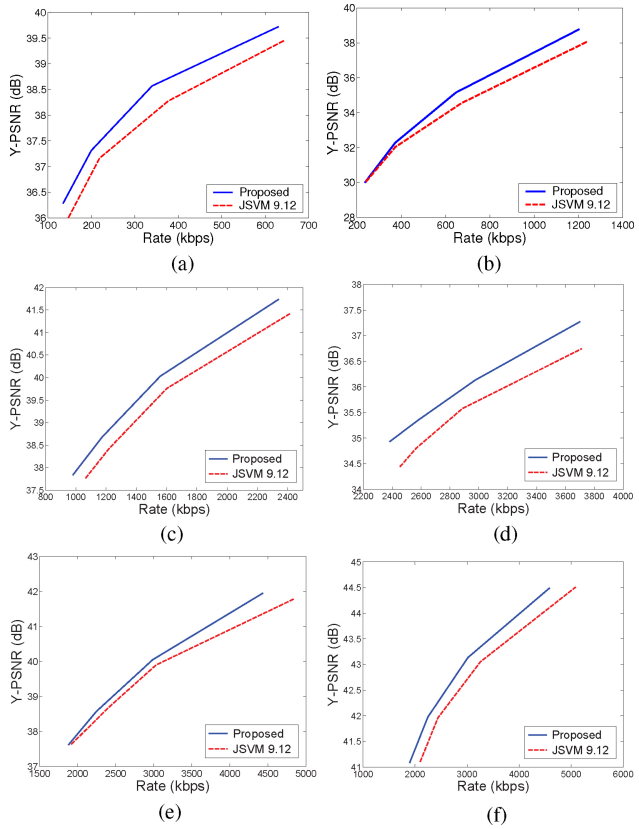

Fig. 11. Coding efficiency comparison (Y-PSNR versus Rate) for CIF, 480p and 720p HD test sequences. (a) *Hall*. (b) *Soccer*. (c) *Rush Hour*. (d) *Tractor*. (e) *Pedestrian Area*. (f) *Sunflower*.

importance of considering the interlayer dependence in the development of bit allocation algorithms. With the benchmark algorithm, the TL dependence is addressed by the QP cascading method that determines the QP difference from a preceding TL only by a T level regardless of temporal characteristics. Moreover, it does not consider the interlayer dependence. Hence, it cannot properly respond to different input video characteristics in terms of how the fidelity of references influences the coding efficiency of predicted frames resulting in suboptimal results.

VI. Conclusion and Future Work

In this research, motivated by the highly involved interdependence among scalable layers, we proposed dependent R/D characteristics models of dependent TL and QL blocks of H.264/SVC. The introduction to the *S*-domain analysis realized the successful decomposition of the dependent R/D functions into the weighted linear sum of independent R/D functions. The performance of the proposed R/D models was verified by comparing estimated R/D values with actual R/D values of various types of test sequences. The successful dependent R/D modeling enabled the proposed joint T-Q layer bit allocation algorithm to operate at significantly lower complexity than those of conventional dependent bit allocation algorithms [11], [12]. In other words, a simple analytical solution could be achieved by the successful isolation of individual function parameters. The proposed algorithm implemented highly efficient bit allocation scheme, which outperformed the benchmarks by a significant margin. Moreover, the coding efficiency of each QL could be greatly enhanced by the proposed bit allocation algorithm.

We have two different future research directions following this work. First, we would like to provide some practical guidelines based on the proposed R/D models such that dependent bit allocation algorithms at constant complexity could be achieved. The provision of the practical guideline is very important because the current algorithm is not yet fully practical in that it still requires multiple pre-encoding passes

TABLE VI
Experimental Result: Global Bit Stream Coding Efficiency

Sequence	R_T	JSVM 9.12 [22]		Proposed	
		Rate (kbps)	Y-PSNR (dB)	Rate (kbps)	Y-PSNR (dB)
Foreman (CIF)	144	146.13	32.39	42.41	32.54
	216	221.07	34.92	209.79	35.15
	360	372.88	38.14	346.31	38.24
	648	645.20	41.19	629.37	41.75
News (CIF)	144	134.36	35.39	138.65	35.60
	216	215.93	37.52	204.39	37.74
	360	367.83	40.24	338.35	40.40
	648	639.70	43.13	628.68	43.41
Pedestrian Area (480p)	1000	933.41	35.70	953.91	35.81
	1200	1225.60	36.84	1140.89	36.84
	1600	1552.55	38.43	1509.30	38.40
	2400	2374.21	40.46	2235.64	40.63
Sunflower (480p)	1000	1091.40	39.16	947.04	39.21
	1200	1229.10	40.04	1137.18	40.21
	1600	1599.60	41.42	1525.84	41.52
	2400	2470.88	43.46	2282.43	43.29
Rush Hour (720p)	2000	2027.56	39.55	1945.84	39.77
	2400	2392.66	40.37	2322.85	40.55
	3200	3385.88	41.51	3093.40	41.63
	4800	4799.59	42.79	4668.05	43.02
Tractor (720p)	3200	3139.69	33.50	3163.29	34.36
	3600	3280.19	33.86	3538.44	34.93
	4400	4435.87	35.09	4318.04	35.89
	6000	5311.17	36.25	5845.20	37.37

TABLE VII
QL Rates and Y-PSNR Average

Sequence	QL	R_T	JSVM 9.12 [22]		Proposed	
			Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)
Crew (CIF)	0	416	426.24	37.39	422.06	37.95
	1	832	841.42	39.96	833.06	40.13
	2	1248	1269.23	41.56	1236.49	41.31
	Average		·	39.23	·	39.57
Tempete (CIF)	0	224	206.34	30.80	230.42	31.63
	1	448	425.78	32.62	457.88	33.13
	2	672	652.97	33.91	678.73	34.06
	Average		·	32.26	·	32.82
Rush Hour (480p)	0	600	607.75	37.67	597.91	38.10
	1	900	907.26	38.45	886.99	38.75
	2	1200	1219.42	39.50	1173.94	39.27
	Average		·	38.42	·	38.67
Tractor (480p)	0	1500	1528.50	34.62	1508.91	35.40
	1	2250	2170.50	35.62	2240.45	36.28
	2	3000	3885.17	36.96	2972.80	36.84
	Average		·	35.58	·	36.13
Pedestrian Area (720p)	0	2400	2370.01	40.74	2321.55	41.24
	1	3600	3594.96	41.90	3433.55	42.13
	2	4800	4842.04	43.23	4437.36	42.60
	Average		·	41.78	·	41.95
Sunflower (720p)	0	1200	1262.82	41.03	1173.55	41.45
	1	1800	1853.19	42.02	1725.15	42.11
	2	2400	2439.51	43.27	2244.74	42.41
	Average		·	41.95	·	41.97

for the model parameter decision. Second, we are interested in developing efficient transmission algorithms of H.264/SVC videos. Originally, the video scalability is motivated by the application requirement of bit stream flexibility, and it is an important issue how to control the flexibility of scalable videos in the practical application scenarios. For this reason, we will investigate the video packet prioritization and its signaling for the QoS control in the future.

REFERENCES

[1] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. Circcuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, Feb. 1997.

[2] Z. He and S. K. Mitra, "A unified rate distortion analysis framework for transform coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1221–1236, Dec. 2001.

[3] Z. He and S. K. Mitra, "A linear source model and a unified rate control algorithm for DCT video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 11, pp. 511–523, Jun. 2002.

[4] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[5] A. Vetro, H. Sun, and Y. Wang, "MPEG 4 rate control for multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 186–199, Feb. 1999.

[6] J. Ribas-Corbera and S. Lei, "A frame-layer bit allocation for H.263+," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1154–1158, Oct. 2000.

[7] N. Kamaci, Y. Altinbasak, and R. M. Mersereau, "Frame bit allocation for H.264/AVC video coder via Cauchy-density-based rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[8] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application for rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1533–1544, Dec. 2005.

[9] Z. Chen and K. N. Ngan, "Toward rate-distortion tradeoff in real-time color video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 158–167, Feb. 2007.

[10] J. Sun, W. Gao, D. Zhao, and W. Li, "On rate-distortion modeling and extraction of H.264/SVC fine-granular scalable video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 323–336, Mar. 2009.

[11] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[12] L.-J. Lin and A. Ortega, "Bit-rate control using piecewise approximated rate-distortion characteristics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 4, pp. 446–459, Aug. 1998.

[13] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B-pictures and MCTF," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 1929–1932.

[14] *Joint Scalable Video Model*, document JVT-X202, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Jul. 2007.

[15] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[16] J. Liu, Y. Cho, and C.-C. J. Kuo, "Bit allocation for spatial scalability coding of H.264/SVC with dependent rate-distortion analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 7, pp. 967–981, Jul. 2010.

[17] L. Xu, S. Ma, D. Zhao, and W. Gao, "Rate control for scalable video model," in *Proc. SPIE Visual Commun. Image Process.*, Jul. 2005, pp. 525–534.

[18] L. Xu, W. Gao, X. Ji, and D. Zhao, "Rate control for hierarchical B-picture coding with scaling-factors," in *IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 49–52.

[19] D. Pranantha, M. Kim, S. Hahm, B. Kim, K. Lee, and K. Park, "Dependent quantization for scalable video coding," in *Proc. IEEE Int. Conf. Adv. Commun. Tech.*, Feb. 2007, pp. 222–227.

[20] Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control of H.264/AVC scalable extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 116–121, Jan. 2008.

[21] Y. Pitrey, M. Babel, O. Deforges, and J. Vieron, "ρ-domain based rate control schemes for spatial, temporal and quality scalable video coding," in *Proc. SPIE Visual Commun. Image Process.*, Jan. 2009, pp. 1–8.

[22] *Joint Draft ITU-T Rec. H.264 | ISO/IEC 14496-10 Amd.3 Scalable Video Coding*, document JVT-X201, Joint Video Team (JVT) ISO/IEC MPEG and ITU-T VCEG, Jul. 2007.

[23] Y. Cho, J. Liu, D.-K. Kwon, and C.-C. J. Kuo, "H.264/SVC temporal bit allocation with dependent distortion model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 641–644.

[24] Y. Cho, J. Liu, D.-K. Kwon, and C.-C. J. Kuo, "Joint quality-temporal (Q-T) bit allocation for H.264/SVC," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2009, pp. 2361–2364.

Yongjin Cho received the M.S. degree in computer science and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2003 and 2010, respectively.

He was a Research Assistant at the Signal and Image Processing Institute, USC, from January 2005 to December 2009. Currently, he is with Samsung Electronics, Seoul, Korea, as a Senior Engineer at the Multimedia Platform Laboratory. His current research interests include multimedia signal processing with a special focus on video compression, video quality assessment, and rate control.

Do-Kyoung Kwon (M'11) received the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2002 and 2006, respectively.

He was a Research Assistant at the Signal and Image Processing Institute, USC, from August 2003 to December 2006. Since 2007, he has been a Member of Technical Staff with the Systems and Applications Research and Development Center, Texas Instrument, Dallas, TX, USA. His current research interests include a wide range of issues related to video compression, including High Efficiency Video Coding (HEVC), scalable video coding, 3-D video coding and quality assessment, and rate control.

Jiaying Liu (S'09–M'10) received the B.E. degree in computer science from Northwestern Polytechnic University, Xian, China, in 2005, and the Ph.D. degree with the Best Graduate Honor in computer science from Peking University, Beijing, China, in 2010.

From 2007 to 2008, she was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA. She is currently an Associate Professor with the Institute of Computer Science and Technology, Peking University. Her current research interests include image processing, sparse signal representation, and video compression.

C.-C. Jay Kuo (F'99) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1985 and 1987, respectively, all in electrical engineering.

He is currently the Director of the Media Communications Laboratory and a Professor of electrical engineering, computer science, and mathematics with the Department of Electrical Engineering and Integrated Media Systems Center, University of Southern California, Los Angeles, CA, USA, and the President of the Asia-Pacific Signal and Information Processing Association. He is the co-author of 200 journal papers, 850 conference papers, and ten books. His current research interests include digital image/video analysis and modeling, multimedia data compression, communication, and networking.

Dr. Kuo is a fellow of the American Association for the Advancement of Science and the International Society for Optical Engineers.