# Bit Allocation for Spatial Scalability Coding of H.264/SVC With Dependent Rate-Distortion Analysis

Jiaying Liu, *Student Member, IEEE,* Yongjin Cho, *Student Member, IEEE,* Zongming Guo, *Member, IEEE,* and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*—We propose a model-based spatial layer bit allocation algorithm for H.264/scalable video coding (SVC) in this paper. The challenge of this problem lies in the fact that the rate-distortion (R-D) behavior of an enhancement layer is dependent on its preceding layers because of inter-layer prediction. To solve it, we first focus on the case of two spatial layers, derive the distortion and rate models of the dependent layer analytically, and develop a low-complexity bit allocation algorithm. It is shown by experimental results that the proposed two-layer bit allocation algorithm can achieve the coding performance close to the optimal R-D performance based on the full search method. Then, we extend this result to multilayer bit allocation by performing the two-layer allocation scheme recursively. Finally, we compare the performance of group of pictures-based and frame-based spatial layer bit allocation schemes at a fixed temporal resolution. The superior performance of the proposed spatial layer bit allocation algorithm is demonstrated using Joint Scalable Video Model reference software algorithm and two prior H.264/SVC rate control algorithms as the benchmarks.**

***Index Terms*—Dependent layer, frame-based distortion and rate models, H.264/scalable video coding (SVC), spatial layer bit allocation.**

## I. Introduction

SCALABLE VIDEO coding (SVC) has recently been standardized to extend the capabilities of the H.264/advanced video coding (AVC) standard [1], [2]. It addresses the application need of a more flexible format of coded video in heterogeneous and time-varying environments. The fundamental principle of SVC is to generate a single compressed bit stream that can adapt to the varying bit rates of different transmission channels, display resolutions, and computational resource constraints of various receivers rapidly and easily.

Manuscript received August 5, 2008; revised May 25, 2009 and September 21, 2009. Date of publication March 18, 2010; date of current version July 16, 2010. This work was supported by the National Basic Research 973 Program of China, under Contract 2009CB320907, and the National Natural Science Foundation of China, under Contract 60902004, with additional support by the State Scholarship Fund from the China Scholarship Council. This paper was recommended by Associate Editor Y.-S. Ho.

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

Y. Cho and C.-C. J. Kuo are 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: yongjinc@usc.edu; 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.2010.2045924

To achieve the spatial scalability, H.264/SVC follows the conventional approach of multilayer coding, where each layer corresponds to a spatial resolution. In each spatial layer, the motion-compensated prediction and the intra-prediction are employed as those in the single-layer H.264/AVC coding. Furthermore, to improve coding efficiency with respect to simulcasting of different spatial resolutions, additional inter-layer prediction is adopted to exploit statistical dependences between different spatial layers [3]. Due to the inter-layer prediction, the problem of spatial bit allocation in H.264/SVC is very challenging. That is, we should consider the tradeoff in coding efficiency between the base layer (BL) and the enhancement layer (EL) simultaneously. Practically, H.264/SVC video can be conveniently delivered in heterogeneous networks with varying client display resolutions, transmission bandwidths, and network conditions. Although some target bit rates are desired for specific applications, most of them do have the flexibility of a bit rate range. Then, H.264/SVC bit allocation algorithm provides an efficient way to utilize the available resource and offer higher quality video to all clients of various spatial resolutions.

There is no previous work that handles the H.264/SVC bit allocation problem by taking into account the relationship between the dependent and reference layers. For example, the reference software [Joint Scalable Video Model (JSVM)] specifies a bottom-up approach to produce a scalable bit stream [2]. That is, the encoding process starts from the bottom-most BL and subsequent ELs are encoded in an ordered manner, where the bit budget of each layer is set individually. It is desirable to develop an optimized bit allocation scheme by considering the inter-layer dependence between the BL and ELs. This is the main objective of our current research.

Bit allocation for inter-frame dependence has been examined since MPEG-2 video. For example, Ramachandran *et al.* [4] studied the dependent bit allocation problem with a trellis-based solution. Although it can yield the optimal solution, the complexity of the solution grows exponentially as the number of dependent frames increases. For this reason, it can be used only as a performance benchmark rather than a practical solution. Lin and Ortega [5] sped up the dependent bit allocation solution by encoding the source video with a few quantization steps and using interpolation to find the

rate-distortion (R-D) values for other quantization steps. Their scheme used the spline interpolation for I frames and the piecewise linear interpolation for P frames. Liu and Kuo [6] investigated the dependent temporal-spatial bit allocation problem for H.263+/MPEG-4 simple profile with a variable frame rate. However, the complexity of these two algorithms is still high since the source video has to be encoded several times. Due to the complexity concern, these solutions cannot be practically applied to the dependent bit allocation problem involved with multiple layers in H.264/SVC.

Among bit allocation algorithms recently proposed for H.264/SVC, except for the highly complex trellis-based solution by Pranantha *et al.* [7], the inter-layer dependence is not well addressed in the problem formulation. For example, Xu *et al.* [8] proposed a rate control algorithm for spatial scalable coding in SVC by employing an improved TMN8 model based on the mode analysis of P/B frames. Liu *et al.* [9], [10] presented a rate control algorithm for the spatial and coarse-grain-SNR (CGS) scalability of H.264/SVC. Their algorithms operate on a fixed rate of each layer and implement an macroblock (MB)-layer bit allocation scheme. In other words, the dependence among spatial layers is not considered at all. To solve this inter-layer dependence issue, a group of pictures (GOP)-based bit allocation problem for the spatial scalability of H.264/SVC was studied in [11] with two spatial layers. In this paper, we attempt to generalize the result in [11] along two directions: 1) from the GOP-based to the frame-based scheme; and 2) from two to multiple spatial layers. They are detailed below.

Since the GOP-based bit allocation algorithm demands a longer delay in the encoding process, it is not suitable for real-time conversational applications. Here, we consider the problem of allocating the bit budget among dependent spatial layers in one frame as the basic coding unit. To provide an analytical solution to this problem, we first investigate the two-layer case. That is, we develop a scheme to decouple the influence of BLs quantization choice on the R-D characteristics of a dependent EL, and show that the statistics of the input signal to the EL quantizer can be modeled as a probability density function parameterized by the BL quantization step-size. As a result, the impact of the BL quantization to the EL quantization can be successfully isolated. Furthermore, we study the relationship between the BL rate and EL rate and propose a rate model of the dependent layer by fixing each EL quantization step-size and changing BL quantization step-sizes.

The major contribution of our paper is the proposal of a novel frame-based dependent distortion and rate model targeting at H.264/SVC spatial scalability. The dependent distortion and rate models are relatively simple yet accurate enough to provide good R-D performance tradeoff. Based on these distortion and rate models, the optimal bit allocation problem can be formulated using the Lagrangian multiplier approach and solved numerically. Furthermore, we extend the two-layer scheme to the multilayer scenario with a recursive process. It is demonstrated by experimental results that the proposed algorithm outperforms the JSVM FixedQPEncoder tool [12] and two previous H.264/SVC rate control schemes proposed in [8] and [10].

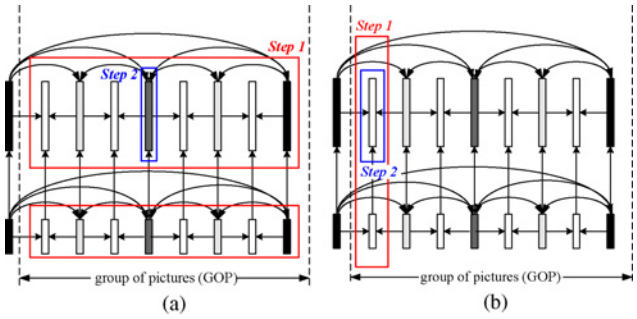


Fig. 1. Illustration of two bit allocation strategies within a GOP. (a) Strategy I: GOP-based strategy. (b) Strategy II: frame-based strategy.

Another contribution of our paper is to offer a low complexity algorithm for dependent spatial layer bit allocation using the R-D model. Since the proposed scheme provides a model-based bit allocation mechanism, it only takes a few encoding passes with several different quantization parameters to obtain the model parameters. For example, the number of the encoding pass for two-layer R-D model construction is equal to 3 (see discussion in Section III). After the target bit budget is allocated to each layer based on the estimated R-D model, a frame is encoded once to meet its individual target bit rate.

The rest of this paper is organized as follows. The frame-based spatial layer bit allocation problem is formulated in Section II. The dependent two-layer bit allocation problem is solved in Section III by analyzing and simplifying the dependent R-D models and a practical two-layer bit allocation algorithm is proposed. Then, the multilayer bit allocation problem is examined in Section IV. The performance of the proposed spatial bit location scheme in terms of coding efficiency and computational complexity is evaluated in Section V. Finally, concluding remarks and future research directions are given in Section VI.

## II. Problem Formulation

In the spatial scalability of H.264/SVC, a video signal with a high spatial resolution is encoded in such a way that the output bit stream provides multiple layers of various spatial resolutions. Suppose that the bit budget for one GOP of spatial layers in H.264/SVC is given. Within one GOP, there are two simple strategies to allocate bits as illustrated in Fig. 1. The first strategy is to allocate the bit budget to different spatial layers of the whole GOP, and then assign the bit budget to each frame within the same spatial layer. The second strategy is to allocate the bit budget to each frame first, where bits are allocated to different picture types (I, P, B). Then, for each frame, bit allocation is conducted for spatial layers. Since Strategy I has a longer delay (in the unit of one GOP), we focus on Strategy II in this paper. For more discussion about the comparison between these two strategies, refer to Section V.

When a bit budget on a full-resolution frame is given, an encoder has still to distribute this bit budget to different spatial layers for optimal coding efficiency. As shown in Fig. 2, a frame is employed as a basic bit allocation unit in this paper,

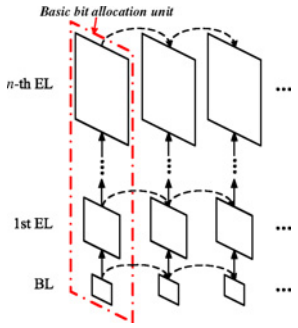


Fig. 2. Allocation of the bit budget of a full-resolution frame to one BL and several EL frames.

which consists of a BL frame and several EL frames. The rate and the distortion of a coded video stream are determined by the choice of the quantization step-sizes at different layers. In the following, we formulate a general dependent bit allocation problem, and show how this formulation is applicable to the H.264/SVC encoder.

It is worthwhile to mention that a practical rate control algorithm for H.264/SVC spatial scalability operates at two levels: 1) the layer-level bit allocation by selecting a quantization parameter for each spatial layer; and 2) the MB-level bit allocation by selecting a quantization parameter for each MB within the same spatial layer. This research has focused on the solution to the first level. As to the second-level, we use the JSVM QP assignment mechanism to select the quantization parameter for each MB for experiments.

The constrained optimization problem for dependent bit allocation can be stated as follows. We seek the quantization step-size of each spatial layer so that the total distortion is minimized subject to the total bit budget constraint. Let $N$ be the number of spatial layers in a frame. $R_k(Q_1, \ldots, Q_k)$ and $D_k(Q_1, \ldots, Q_k)$ are the rate and the distortion model of the $k$th layer with respect to the vector of quantization step-sizes, denoted by $(Q_1, \ldots, Q_k)$. Given the bit budget $R_T$ of the current frame, the bit allocation problem can be formulated mathematically as

$$\mathbf{Q}^* = (Q_1^*, \ldots, Q_N^*) = \arg\min_{Q_k \in \mathcal{Q}} \sum_{k=1}^{N} \omega_k \cdot D_k(Q_1, \cdots, Q_k)$$
$$\text{subject to } \sum_{k=1}^{N} R_k(Q_1, \ldots, Q_k) \le R_T, \ \sum_{k=1}^{N} \omega_k = 1 \qquad (1)$$

where $\mathbf{Q}^* = (Q_1^*, \ldots, Q_N^*)$ is the optimal quantization vector, and $\mathcal{Q}$ is the set of all possible quantization candidates. $Q_1, \cdots, Q_{k-1}$ in the R-D functions of the $k$th layer indicate that the coding performance of the $k$th layer is dependent upon previously coded $(k-1)$ layers. Furthermore, $\omega_k$ is a weighting factor that indicates the importance of the $k$th layer. Thus, the total distortion is defined as a weighted sum of the distortion of each individual layer in (1).

The Lagrangian multiplier method can be used to map the constrained optimization problem in (1) to an equivalent unconstrained optimization problem by introducing the Lagrangian cost function as

$$\mathbf{Q}^* = \arg\min_{Q_k \in \mathcal{Q}} J(\mathbf{Q}, \lambda)$$

$$J(\mathbf{Q}, \lambda) = \sum_{k=1}^{N} \omega_k \cdot D_k(\cdot) + \lambda \cdot \left( \sum_{k=1}^{N} R_k(\cdot) - R_T \right) \qquad (2)$$

where $\lambda$ is the Lagrangian multiplier. To solve the problem given in (2), one solution is to conduct a full search over all possible combinations of admissible quantization choices. However, since the search space grows exponentially as the number of layers increases, the complexity of full search is prohibitively large. To address the complexity issue, we will model the R-D characteristics of dependent layers as elaborated in the next section.

## III. Bit Allocation Analysis for Two Spatial Layers

In this section, we consider the bit allocation problem for the two-layer case (i.e., $N = 2$). The solution will be generalized to the general multilayer case in Section IV. Mathematically, the Lagrangian cost function of a two-layer case can be expressed as

$$J(\mathbf{Q}, \lambda) = \omega_1 \cdot D_1(Q_1) + \omega_2 \cdot D_2(Q_1, Q_2) + \lambda \cdot (R_1(Q_1) + R_2(Q_1, Q_2) - R_T). \qquad (3)$$

In the following discussion, we assume the equal importance of these two layers; namely, $\omega_1 = \omega_2 = 0.5$, which can be easily generalized. In order to ease the computational burden of the full search method, we look for a solution method that avoids the need to collect all the R-D data while retaining some optimality. Consequently, we will adopt a model-based approach that analyzes the distortion and rate dependence between these two layers of H.264/SVC.

Generally speaking, the R-D characteristics of a dependent layer [i.e., $R_2(Q_1, Q_2)$ and $D_2(Q_1, Q_2)$] can be represented by a function of the quantization step-size of the reference layer $(Q_1)$ and the dependent layer $(Q_2)$. In this case, the dependent and reference layers indicate the EL and BL, respectively. For dependent R-D modeling, if we can convert the multi-variable rate and distortion functions into a number of independent single-variable functions, the solution to the bit allocation problem would be significantly simplified. We will propose a way to achieve this goal in the next two subsections.

### A. Distortion Modeling

Without loss of generality, we depict an exemplary H.264/SVC encoder in Fig. 3, whose input is a common intermediate format (CIF) sequence and output consists of two spatial layers, i.e., coded BL and EL. As shown in this figure, we first obtain a low frequency component of the input CIF video by the down-sampling process. The lowpass filtered video is fed into the BL encoder to produce the BL reconstruction, which corresponds to a quantized version of the low-frequency video using quantization step-size $Q_1$. The reconstructed BL is used as a basis to predict the low frequency component of the input to reduce inter-layer redundancy. Then, we use the differential video between the original and the interpolated BL signals as the input to the EL encoder.

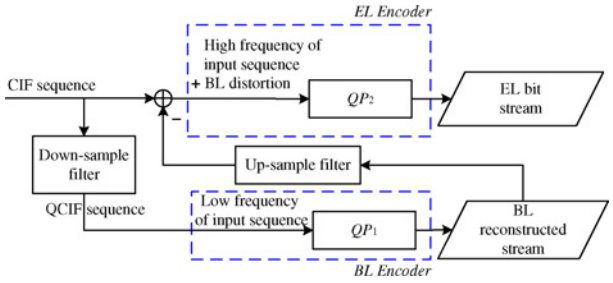


Fig. 3. Illustration of video layer decomposition in the H.264/SVC encoder for spatial scalability.

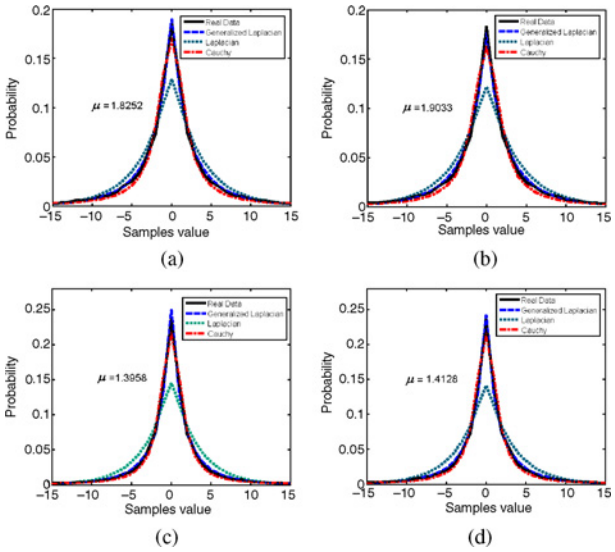


Fig. 4. Curve fitting results of DCT/AC coefficients of different input video signals to the BL quantizer for the *City* sequence. (a) CIF input and $Q_1 = 14$. (b) CIF input and $Q_1 = 32$. (c) 4CIF input and $Q_1 = 14$. (d) 4CIF input and $Q_1 = 24$.

This differential video actually consists of two parts: 1) the high frequency component; and 2) the distortion in the low frequency component due to the quantization effect by $QP_1$ in the BL, which is denoted as the "BL Distortion." The second part controls the coupling between BLs and ELs codings. If the BL distortion term is much smaller than the high frequency term, such a coupling effect can be ignored.

Since the frame is used as the basic unit, we use the distribution of DCT/AC coefficients of each frame to characterize the EL differential signal. It was stated in [13] that the zero-mean Cauchy density is more accurate in representing the distribution of AC coefficients than the traditional Laplacian density, while it is simpler than the generalized Laplacian density since it only has a single parameter. Thus, we approximate the distribution of DCT/AC coefficients of the differential EL sequence by the zero-mean Cauchy distribution of the following density function:

$$p(x) = \frac{1}{\pi} \cdot \frac{\mu}{\mu^2 + x^2}, \quad x \in \mathbf{R} \tag{4}$$

where parameter $\mu$ controls the thickness and the height of the pulse centered at the origin.

By conducting curve fitting with transform coefficients of the difference video sequence, we find that the zero-mean

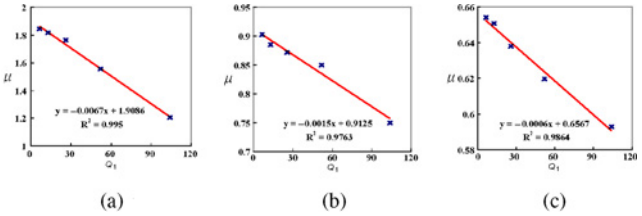


Fig. 5. Illustration of the affine relationship between $\mu$ and $Q_1$,with different spatial complexity sequences. (a) *City*. (b) *Foreman*. (c) *Akiyo*.

Cauchy density distribution is still valid. Fig. 4 shows the curve fitting results by the influence of different $Q_1$ values and different input video sizes. The two layers in Fig. 4(a) and (b) are the quarter common intermediate format (QCIF)–CIF pair while those in Fig. 4(c) and (d) are the CIF–4CIF pair. We see that both the generalized Laplacian density function and the zero-mean Cauchy density function provide an excellent match with the real data. Since the generalized Laplacian density has two parameters, we choose the Cauchy density function for simplicity. Next, we examine the relationship between parameter $\mu$ in the Cauchy pdf and the step-size, $Q_1$, of the BL quantization using three CIF sequences (i.e., *City*, *Foreman*, and *Akiyo*) of different content complexity. The following affine relationship is obtained from the results shown in Fig. 5:

$$\mu = \eta \cdot Q_1 + \varphi \tag{5}$$

where $\eta$ and $\varphi$ are two parameters of the affine model.

Besides the effect of the BL quantization step-size, $Q_1$, on the input video signal to the EL encoder, the output of the EL encoder is determined by the EL quantization step-size, $Q_2$. After the decomposition process, we treat the EL as a single layer. Since it is uniformly quantized by step-size $Q_2$, the distortion of the EL output can be estimated via

$$D_2(Q_2) = \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})Q_2}^{(i+\frac{1}{2})Q_2} |x - iQ_2|^2 p(x) dx. \tag{6}$$

It can be shown that the infinite sum in (6) converges and the converged value is bounded above by $Q_2^2/4$. For the Cauchy source, (6) can be simplified as

$$\begin{aligned} D_2(Q_2) = 2 \sum_{i=1}^{M} \Bigg[ & \frac{\mu Q_2}{\pi} - \frac{i\mu Q_2}{\pi} \ln \left( \frac{\mu^2 + (i+\frac{1}{2})^2 Q_2^2}{\mu^2 + (i-\frac{1}{2})^2 Q_2^2} \right) \\ & - \frac{\mu^2 - i^2 Q_2^2}{\pi} tan^{-1} \left( \frac{\mu Q_2}{\mu^2 + (i^2 - \frac{1}{4}) Q_2^2} \right) \Bigg] \\ & + \left[ \frac{\mu Q_2}{\pi} - \frac{2\mu^2}{\pi} tan^{-1} \left( \frac{Q_2}{2\mu} \right) \right]. \end{aligned} \tag{7}$$

Although (7) is complex, it can be approximated by an exponential form [13]

$$D_2(Q_2) \approx b \cdot Q_2^{\beta} \tag{8}$$

where $b$ is a parameter related to $\mu$ only while the value of $\beta$ is almost constant for a given frame. $b$ is computed using the least-square-errors solution. Several pairs of $b$ and $\mu$ values in the single layer are given in Table I. We observe an affine relationship between parameters $b$ and $\mu$ as shown in Fig. 6.

TABLE I

CORRESPONDING VALUES OF PARAMETER $b$ IN (8) AND PARAMETER $\mu$ IN (4)

| $\mu$ | $b$ |
|---|---|
| 1.5634 | 0.2501 |
| 1.6468 | 0.2356 |
| 2.3285 | 0.2223 |
| 2.8522 | 0.2087 |
| 3.0980 | 0.1976 |
| 3.2928 | 0.1889 |

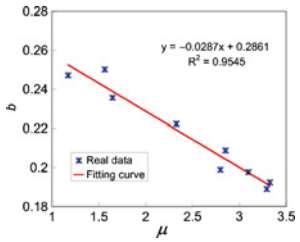


Fig. 6. Plot of parameter $b$ as a function of parameter $\mu$, which exhibits an affine relationship.

Based on the above analysis, we see that the two factors that affect the distortion of the dependent EL are decoupled clearly. That is, the BL quantization step-size, $Q_1$, determines the parameter, $\mu$, of the Cauchy distribution as shown in (5) while the single layer EL distortion is related to the EL quantization step-size, $Q_2$, as shown in (8) and the affine relationship between $b$ and $\mu$. Thus, we can obtain the following simplified distortion model for the dependent EL layer:

$$D_2(Q_1, Q_2) \approx (\zeta Q_1 + \upsilon) \cdot Q_2^{\beta} \tag{9}$$

where $\zeta$, $\upsilon$, and $\beta$ are model parameters, which are independent of $Q_1$ and $Q_2$.

We conducted experiments to verify the accuracy of the distortion model given in (9). In the experiments, we considered two layers of either QCIF–CIF at the frame rate of 15 frames/s or CIF–4CIF resolutions at the frame rate of 30 frames/s, and generated 500 distortion samples with various values of $Q_1$ and $Q_2$ for each test case. The parameter configure file was the scalable baseline profile. Table II lists the accuracy of the distortion model, which is defined as

$$\text{Accuracy} = \left(1 - \frac{|\text{Estimated value} - \text{Actual value}|}{\text{Actual value}}\right) \times 100\%. \tag{10}$$

We see that the proposed distortion model has an accuracy of about 84.84% on average.

### B. Rate Modeling

To derive the rate model of an EL, denoted by $R_2(Q_1, Q_2)$, we plot the rate of a dependent layer EL, with respect to the rate of its reference layer BL, in Fig. 7, where rate pairs $[R_1(Q_1), R_2(Q_1, Q_2)]$ are shown. There are two types of curves, reflecting two different settings of a two-variable rate function denoted by $R_2(Q_1, Q_2)$. The dashed curve on the diagonal plots the EL rate when $Q_1+6$ and $Q_2$ have the same values. For solid branches, the value of $Q_1$ varies whereas

TABLE II

VERIFICATION OF THE PROPOSED DISTORTION MODEL

| EL Resolution | Sequence | Accuracy (%) |
|---|---|---|
| CIF | *Flower* | 80.53 |
| | *Bridge-far* | 90.15 |
| | *Tempete* | 81.24 |
| | *News* | 87.82 |
| | *Stefan* | 91.47 |
| | *Carphone* | 86.26 |
| | *Salesman* | 84.30 |
| 4CIF | *Ice* | 83.02 |
| | *Harbour* | 81.11 |
| | *Soccer* | 78.48 |
| Average | | 84.84 |

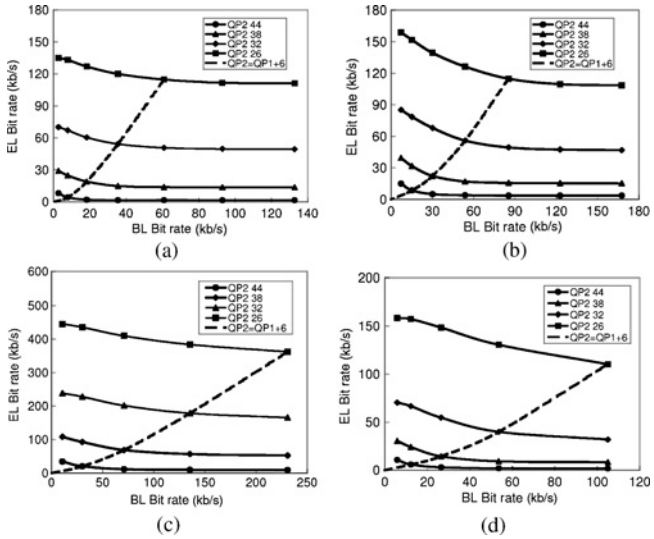


Fig. 7. Illustration of the proposed rate modeling results. (a) *City*, QCIF–CIF. (b) *Football*, QCIF–CIF. (c) *City*, CIF–4CIF. (d) *Crew*, CIF–4CIF.

the value of $Q_2$ is fixed at each branch. We see that, for each fixed $Q_2$, increasing $R_1(Q_1)$ (or decreasing $Q_1$) results in a roughly linear reduction in $R_2(Q_1, Q_2)$ for low BL bit rates. However, the EL rate $R_2$ becomes saturated and does not decrease furthermore beyond the point with $QP_2 = QP_1 + 6$.

Based on the above observation, the idealized rate characteristics are illustrated in Fig. 8. We conclude that the rate of a dependent spatial layer can be approximated as

$$R_2(Q_1, Q_2) = \begin{cases} r \cdot R_1(Q_1) + (s - r) \cdot R_1(Q_2/2) & Q_2 \le 2Q_1 \\ s \cdot R_1(Q_2/2) & Q_2 > 2Q_1 \end{cases} \tag{11}$$

where $s$ and $r$ are the slopes of the line when $QP_2 = QP_1 + 6$ and $QP_2 \le QP_1 + 6$, respectively. When $QP_2 = QP_1 + 6$, we have that the corresponding quantization step size is halved, i.e., $Q_2 = 2Q_1$. Although the dependent rate model is similar to those derived in [5] and [14], the context is different. That is, the dependent rate models in [5] and [14] were obtained for temporal dependence while ours arise from spatial dependence.

The accuracy of the proposed rate model given in (11) was verified experimentally with the EL of CIF at the frame rate of

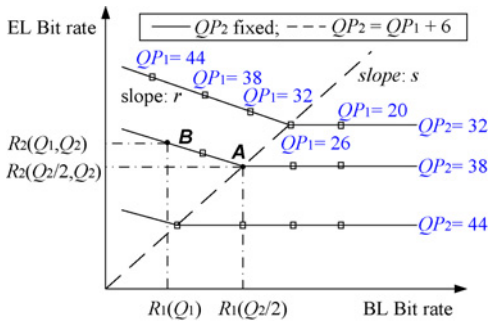


Fig. 8. Proposed rate model for the EL as a function of the BL.

TABLE III
VERIFICATION OF THE PROPOSED RATE MODEL

| EL Resolution | Sequence | Accuracy (%) |
|---|---|---|
| CIF | *Flower* | 82.75 |
| | *Bridge-far* | 96.13 |
| | *Tempete* | 93.61 |
| | *News* | 93.12 |
| | *Stefan* | 95.04 |
| | *Carphone* | 87.37 |
| | *Salesman* | 92.11 |
| 4CIF | *Ice* | 80.06 |
| | *Harbour* | 93.44 |
| | *Soccer* | 91.59 |
| Average | | 90.52 |

15 frames/s or 4CIF resolution at the frame rate of 30 frames/s. We generated 500 samples for each test sequence with various $Q_1$ and $Q_2$ values. The mean results in accuracy are shown in Table III, and the rate estimation accuracy is calculated as (10). We see that the proposed rate model has an accuracy of about 90.52% on average.

### C. Solution to the Lagrangian Formulation

The base layer of H.264/SVC is compatible with H.264/AVC, and the Cauchy-density-based R-D model has a good balance between complexity and estimation accuracy for frame-based R-D modeling H.264/AVC. Thus, for $R_1(Q_1)$ and $D_1(Q_1)$ of the BL, we adopt the following models [13]:

$$D_1(Q_1) = b \cdot Q_1^{\beta} \text{ and } R_1(Q_1) = a \cdot Q_1^{-\alpha} \tag{12}$$

where $a$, $b$, $\alpha$, and $\beta$ are model parameters. With the EL R-D models given in (9) and (11), and the BL R-D models in (12), we are ready to solve the bit allocation problem for H.264/SVC with two spatial layers. Since the proposed R-D models are defined by completely closed-form expressions, a numerical solution to the Lagrange formulation in (3) becomes feasible. Note also that the rate model in (11) can be further simplified when we impose the following constraint by taking account of the monotonicity condition:

$$0 \le QP_2 - QP_1 \le 6. \tag{13}$$

This constraint is usually met by the optimal solution $(Q_1^*, Q_2^*)$. It was also recommended generally by the Joint Video Team (JVT) [2] that the quantization parameter $QP_2$ for the enhancement layer was set to $QP_1+4$, with $QP_1$ being the quantization parameter for the base layer.

Under the above conditions, the Lagrangian cost function in (3) can be written as

$$J(\mathbf{Q}, \lambda) = \frac{1}{2}\left(b \cdot Q_1^{\beta_1} + (\zeta Q_1 + \upsilon) \cdot Q_2^{\beta_2}\right) + \lambda \cdot \left((1+r) \cdot aQ_1^{-\alpha} + (s-r) \cdot a(Q_2/2)^{-\alpha} - R_T\right) \tag{14}$$

To derive the optimal solution of the Lagrangian cost function, we take the partial derivatives with respect to $Q_1$, $Q_2$, and $\lambda$, which yields the following three equations:

$$\begin{aligned} b\beta_1 \cdot Q_1^{(\beta_1-1)} + \zeta \cdot Q_2^{\beta_2} - a\alpha(1+r)Q_1^{(-\alpha-1)} \cdot \lambda = 0 \\ \upsilon\beta_2 \cdot Q_2^{(\beta_2-1)} - 1/2a\alpha \cdot (s-r)(Q_2/2)^{(-\alpha-1)} \cdot \lambda = 0 \\ a \cdot (1+r)Q_1^{-\alpha} + a \cdot (s-r)(Q_2/2)^{-\alpha} - R_T = 0. \end{aligned} \tag{15}$$

Note that there are three variables $Q_1$, $Q_2$, and $\lambda$ in (15) while other parameters are determined in an earlier stage. To be more specific, the proposed algorithm consists of three stages: 1) pre-encoding for model parameter decision; 2) $Q$-decision to encode at a target bit rate; and 3) actual encoding based on the allocated rate. The model parameters are determined in the pre-encoding stage. Then, we compute $Q_1$ and $Q_2$ that optimize the Lagrangian cost function numerically. We can determine quantization parameters, $QP_1$ and $QP_2$, using the one-to-one correspondence between quantization step-size $Q$ and quantization parameter $QP$ [15]. Finally, each layer in the current basic coding unit is encoded to produce the final bit stream at the target bit rate.

### D. Experimental Results: Bit Allocation With Two Spatial Layers

The proposed two-spatial-layer bit allocation algorithm was implemented with JSVM 9.6 using the scalable baseline profile. Since there is no spatial layer bit allocation algorithm in the current version of JSVM, we compare the performance of the proposed two-layer bit allocation algorithm against that of the full search (FS) method. With the FS, the input video sequence is encoded with all possible $Q_1$ and $Q_2$ values, which are set constant for the whole sequences in the FS algorithm, and selects the one with the best R-D performance, showing the minimum value of the average distortion of BL and EL, while meeting the requirement of the total target bit rate. Then, the optimal solution is determined as the $(Q_1, Q_2)$ pair that provides the minimum average distortion while satisfying the target bit rate constraint. Clearly, the R-D curve obtained by the FS provides the optimal R-D tradeoff among all bit allocation schemes since the solution is determined based on the real R-D data.

One performance benchmark was obtained using the reference JSVM FixedQPEncoder tool under the SVC test conditions as defined in JVT-Q205 [16]. To employ the JSVM FixedQPEncoder tool, each layer is first assigned an initial QP. Then, the encoder performs the encoding process with a trial QP in each iteration, where the generated bit rate value is further used as the feedback information to adjust the trial QP value in the next iteration. The iteration terminates when the generated bit rate is within the acceptable mismatch range of the target bit rate or exceeds the predefined maximal threshold.

TABLE IV
TWO-LAYER SETTING AND THE EXPERIMENTAL CONFIGURATION IN SPATIAL SCALABLE CODING

| Scenario | Layer No. | Format | Frame Rate | Initial QP |
|---|---|---|---|---|
| I | 1 | QCIF | 15 | 32 |
| | 2 | CIF | 15 | 32 |
| II | 1 | CIF | 30 | 32 |
| | 2 | 4CIF | 30 | 32 |
| Profile | Scalable | SearchMode | | FastSearch |
| | Baseline | SearchRange | | 16 |

Otherwise, the QP is further computed and updated using the logarithmic search method. The number of multiple encoding passes is called the iteration number.

Two test scenarios and corresponding SVC configurations are given in Table IV. For Scenario I, two layers are of QCIF–CIF format with four test sequences (*Bus*, *Football*, *Foreman*, and *Mobile*) at the frame rate of 15 frames/s. For Scenario II, the two layers are of CIF–4CIF format with another four test sequences (*City*, *Crew*, *Soccer*, and *Harbour*) at a frame rate of 30 frames/s. These test sequences have different spatial complexities. Layer 1 is the base layer, which is encoded without any inter-layer prediction. Layer 2 is the spatial enhancement layer using adaptive inter-layer prediction from the base layer. For the FixedQPEncoder tool, the initial QP value is set to 32 in both layers. The GOP size is set to 1 to provide the IPPP structure.

The frame-by-frame peak signal-to-noise ratio (PSNR) performance comparison of three spatial bit allocation algorithms is shown in Figs. 9 and 10 for Scenarios I and II, respectively. From these figures, we see the proposed bit allocation method outperforms the JSVM FixedQPEncoder tool significantly. While the gap between the optimal performance (obtained by FS) and the proposed scheme is within 1 dB in both scenarios. It is worthwhile to point out that the performance of the proposed two-layer bit allocation scheme is not sensitive to frame rates or spatial resolutions of the underlying video.

We summarize the coding results using the proposed bit allocation algorithm and JSVM reference software algorithm in Tables V and VI for Scenarios I and II, respectively. The rate control method using the proposed bit allocation algorithm achieves an average of 1.82 dB and 1.38 dB PSNR gains over the JSVM in Scenarios I and II, respectively. Both methods can yield the desired rate with a small deviation (less than 3% of the target rate). We also show the iteration numbers of the JSVM FixedQPEncoder and the proposed scheme in Tables V and VI. The proposed bit allocation algorithm demands a fixed number of iteration. We have to determine the parameters of the distortion model and the rate model of BL and EL. Parameters $b$ and $\beta_1$ of BLs distortion model and parameters $a$ and $\alpha$ of BLs rate model can be calculated in the first two encoding passes. Parameters $\zeta$, $\upsilon$, and $\beta_2$ of ELs distortion model can be calculated in the first three encoding passes. The rate model of ELs rate model is determined by two slope values, which demands three encoding passes. To summarize, we need three encoding iterations to build all required R and D models. Furthermore, we need to encode a frame once to meet the target bit based on these R and D models. As a result, the total iteration number is four. In contrast, JSVM reference software algorithms demand a much higher number of iteration, which implies a higher computational complexity. These results demonstrate the effectiveness and the robustness of the proposed bit allocation algorithm for video sequences of various spatial characteristics.

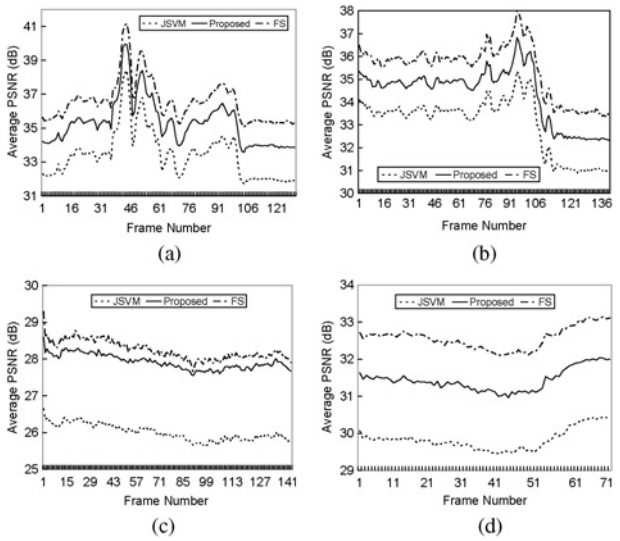


Fig. 9. PSNR value as a function of the frame number with the proposed, JSVM and FS bit allocation schemes for Scenario I. (a) *Football*, $R_T$ = 768 kb/s. (b) *Foreman*, $R_T$ = 192 kb/s. (c) *Mobile*, $R_T$ = 256 kb/s. (d) *Bus*, $R_T$ = 384 kb/s.

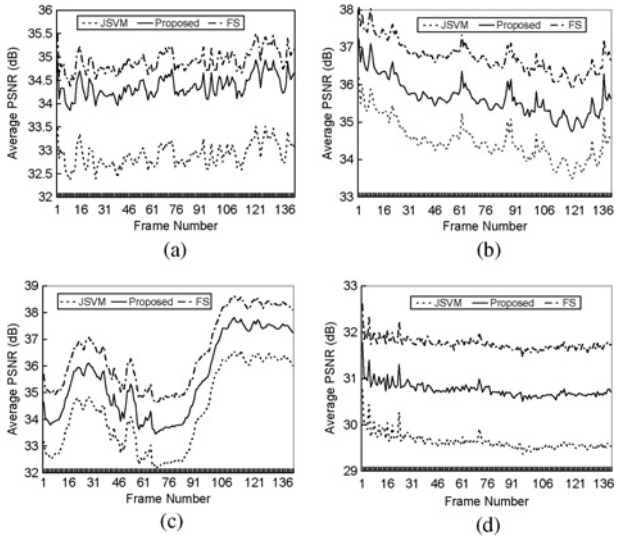


Fig. 10. PSNR value as a function of the frame number with the proposed, JSVM and FS bit allocation schemes for Scenario II. (a) *City*, $R_T$ = 1024 kb/s. (b) *Crew*, $R_T$ = 1536 kb/s. (c) *Soccer*, $R_T$ = 1536 kb/s. (d) *Harbour*, $R_T$ = 1536 kb/s.

For further evaluation, we compare the proposed algorithm with two previous SVC rate control algorithms proposed by Xu *et al.* [8] and Liu *et al.* [10]. The rate control algorithm in [8] targeted at spatial and CGS scalable coding in SVC. First, a rate-distortion model extended from TMN8 was applied to I/P frames. Then, a two-pass QP refinement process was adopted. In [10], a linear R-Q model estimation of texture bits was used for rate control. By exploiting the correlations between

TABLE V

PERFORMANCE OF THE PROPOSED ALGORITHM AND JSVM METHOD FOR QCIF−CIF TWO LAYERS IN TERMS OF OUTPUT RATE, PSNR, PSNR GAIN, Δ RATE, AND ITERATION NUMBER

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | PSNR Gain (dB) | Rate (kb/s) | Δ Rate (kb/s) | Iter. |
|---|---|---|---|---|---|---|---|
| *Bus* | 384 | Proposed | 31.43 | 1.6 | 376.45 | −7.55 | 4 |
| | | JSVM | 29.83 | | 386.97 | +2.97 | 57 |
| | 512 | Proposed | 32.57 | 0.83 | 495.14 | −16.86 | 4 |
| | | JSVM | 31.74 | | 527.28 | +5.28 | 40 |
| | 768 | Proposed | 33.98 | 1.8 | 764.58 | −3.42 | 4 |
| | | JSVM | 32.18 | | 788.86 | +0.32 | 45 |
| *Football* | 768 | Proposed | 34.87 | 1.89 | 779.1 | +10.9 | 4 |
| | | JSVM | 32.98 | | 776.04 | +8.04 | 57 |
| | 1024 | Proposed | 37.24 | 2.25 | 1020.88 | −3.12 | 4 |
| | | JSVM | 34.99 | | 1036.89 | +12.89 | 26 |
| | 1536 | Proposed | 38.53 | 3.51 | 1524.47 | −11.53 | 4 |
| | | JSVM | 35.02 | | 1519.54 | −15.46 | 31 |
| *Foreman* | 192 | Proposed | 35.01 | 1.66 | 194.78 | +2.78 | 4 |
| | | JSVM | 33.35 | | 192.59 | +0.59 | 27 |
| | 256 | Proposed | 36.85 | 1.95 | 254.42 | −1.58 | 4 |
| | | JSVM | 34.90 | | 254.57 | −1.43 | 36 |
| | 384 | Proposed | 38.12 | 1.18 | 388.23 | +4.23 | 4 |
| | | JSVM | 36.94 | | 380.66 | −3.34 | 37 |
| *Mobile* | 256 | Proposed | 27.76 | 1.94 | 257.11 | +1.11 | 4 |
| | | JSVM | 25.82 | | 250.26 | −5.74 | 36 |
| | 384 | Proposed | 29.04 | 1.39 | 382.65 | −1.35 | 4 |
| | | JSVM | 27.65 | | 372.88 | −11.12 | 27 |
| | 512 | Proposed | 31.24 | 1.88 | 520.13 | +8.13 | 4 |
| | | JSVM | 29.36 | | 514.19 | +2.93 | 39 |
| Average PSNR gain (dB) | | | | 1.82 | | | |

TABLE VI

PERFORMANCE OF THE PROPOSED ALGORITHM AND JSVM METHOD FOR CIF−4CIF TWO LAYERS IN TERMS OF OUTPUT RATE, PSNR, PSNR GAIN, Δ RATE, AND ITERATION NUMBER

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | PSNR Gain (dB) | Rate (kb/s) | ΔRate (kb/s) | Iter. |
|---|---|---|---|---|---|---|---|
| *City* | 1024 | Proposed | 34.37 | 1.47 | 1027.55 | +3.55 | 4 |
| | | JSVM | 32.90 | | 1019.50 | −5.50 | 24 |
| | 1536 | Proposed | 35.42 | 1.85 | 1519.14 | −16.86 | 4 |
| | | JSVM | 33.84 | | 1541.56 | +5.56 | 7 |
| | 2048 | Proposed | 36.98 | 2.37 | 2041.23 | −6.77 | 4 |
| | | JSVM | 34.61 | | 1987.71 | −60.29 | 36 |
| *Crew* | 1536 | Proposed | 35.65 | 1.16 | 1512.36 | −23.64 | 4 |
| | | JSVM | 34.49 | | 1485.36 | −50.64 | 39 |
| | 2048 | Proposed | 37.25 | 1.17 | 2033.16 | −14.84 | 4 |
| | | JSVM | 36.08 | | 2041.31 | −6.69 | 17 |
| | 3072 | Proposed | 38.68 | 1.66 | 3044.37 | −27.63 | 4 |
| | | JSVM | 37.02 | | 2967.51 | −104.49 | 31 |
| *Soccer* | 1536 | Proposed | 35.54 | 1.24 | 1532.88 | −3.12 | 4 |
| | | JSVM | 34.30 | | 1518.50 | −17.50 | 30 |
| | 2048 | Proposed | 37.51 | 1.50 | 2047.16 | −0.84 | 4 |
| | | JSVM | 36.01 | | 2041.29 | −6.71 | 9 |
| | 3072 | Proposed | 38.99 | 1.32 | 3079.64 | +6.64 | 4 |
| | | JSVM | 37.67 | | 3035.01 | −26.99 | 28 |
| *Harbour* | 1536 | Proposed | 30.76 | 1.12 | 1522.49 | −13.51 | 4 |
| | | JSVM | 29.64 | | 1542.01 | +6.01 | 13 |
| | 2048 | Proposed | 31.94 | 1.06 | 2043.11 | −4.89 | 4 |
| | | JSVM | 30.88 | | 2053.93 | +5.93 | 10 |
| | 3072 | Proposed | 33.37 | 1.02 | 3088.57 | +16.57 | 4 |
| | | JSVM | 32.35 | | 2977.31 | −94.69 | 51 |
| Average PSNR gain (dB) | | | | 1.38 | | | |

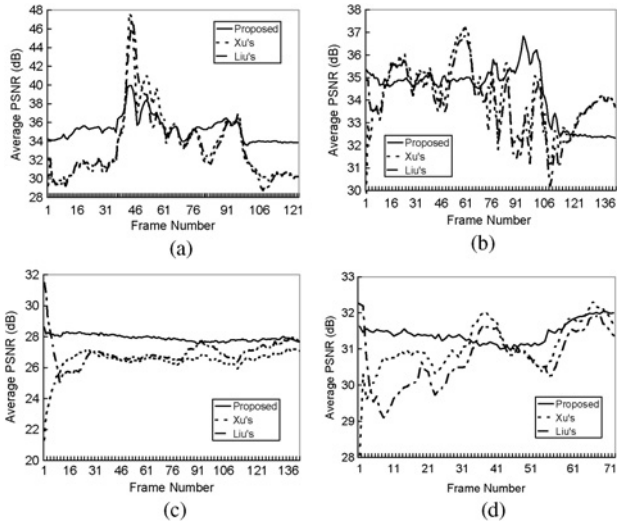


Fig. 11. Performance comparison of three algorithms for the averaged PSNR value of the BL and the EL as a function of the frame number in Scenario I. (a) *Football*, $R_T$ = 768 kb/s. (b) *Foreman*, $R_T$ = 192 kb/s. (c) *Mobile*, $R_T$ = 256 kb/s. (d) *Bus*, $R_T$ = 384 kb/s.

different layers, a switched model was proposed to predict the mean absolute difference (MAD) of residual texture based on the available MAD information of the previous frame in the same EL or the same frame in its BL. Consequently, abrupt MAD fluctuations in the EL could be predicted properly.

The results are summarized in Tables VII and VIII. For Scenario I, the proposed algorithm achieves an averaged PSNR gain of 1.14 dB and 1.27 dB over Liu's algorithm and Xu's algorithm, respectively. For Scenario II, the proposed algorithm achieves an averaged PSNR gain of 0.96 dB and 1.06 dB over Liu's and Xu's algorithms, respectively. The averaged PSNR value of the BL and the EL is plotted as a function of the frame number in Figs. 11 and 12. We see that the frame quality using our algorithm is the best among the three most often. Besides, the proposed algorithm has the lowest PSNR variation across frames. Since these two approaches are all model-based rate control algorithms, their complexities are also low. For Xu's algorithm, the iteration number is four because of the use of the two-pass encoding at each layer. For Liu's algorithm, since there is no any feedback used to adjust the current bit allocation scheme, it is a one-pass encoding process and its iteration number is two.

We concentrate on the optimal bit allocation scheme by considering layer dependence. Fig. 13 shows the optimal ratio of the BL rate to the sum of the BL rate and the EL rate, with *Foreman* and *Football* sequences. It is observed that the ratio for frames of higher spatial complexity (or, equivalently, frames with a large amount of higher frequency components) is smaller. This is consistent with our signal decomposition analysis in Section III; namely, the ratio is smaller for frames that contain a large amount of high frequency components. For example in Fig. 13, the difference of the ratios is clear depending on the picture complexity. With the camera changing from running football players to the background grass, the ratio increases from 0.5 on average in a period to 0.7. With the scene change in *Foreman* sequence, the ratio curve

TABLE VII

PERFORMANCE COMPARISON OF THE PROPOSED ALGORITHM, LIU'S ALGORITHM AND XU'S ALGORITHM FOR THE QCIF−CIF TWO LAYERS IN TERMS OF THE OUTPUT RATE, PSNR AND Δ RATE

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | Rate (kb/s) | ΔRate |
|---|---|---|---|---|---|
| *Bus* | 384 | Proposed | 31.43 | 376.45 | −7.55 |
| | | Liu's | 30.56 | 384.43 | +0.43 |
| | | Xu's | 30.56 | 387.28 | +3.28 |
| | 512 | Proposed | 32.57 | 495.14 | −16.86 |
| | | Liu's | 32.06 | 512.15 | +0.15 |
| | | Xu's | 31.91 | 515.38 | +3.38 |
| | 768 | Proposed | 33.98 | 764.58 | −3.42 |
| | | Liu's | 33.57 | 769.08 | +1.08 |
| | | Xu's | 33.89 | 771.54 | +3.54 |
| *Football* | 768 | Proposed | 34.87 | 779.10 | +10.90 |
| | | Liu's | 33.21 | 768.84 | +0.84 |
| | | Xu's | 33.35 | 771.38 | +3.38 |
| | 1024 | Proposed | 37.24 | 1020.88 | −3.12 |
| | | Liu's | 34.95 | 1027.34 | +3.34 |
| | | Xu's | 35.09 | 1027.49 | +3.49 |
| | 1536 | Proposed | 38.53 | 1524.47 | −11.53 |
| | | Liu's | 37.65 | 1538.75 | +2.75 |
| | | Xu's | 37.74 | 1539.63 | +3.63 |
| *Foreman* | 192 | Proposed | 35.01 | 194.78 | +2.78 |
| | | Liu's | 33.81 | 192.39 | +0.39 |
| | | Xu's | 33.73 | 195.23 | +3.23 |
| | 256 | Proposed | 36.85 | 254.42 | −1.58 |
| | | Liu's | 35.11 | 256.66 | +0.66 |
| | | Xu's | 35.10 | 259.23 | +3.23 |
| | 384 | Proposed | 38.12 | 388.23 | +4.23 |
| | | Liu's | 36.97 | 384.65 | +0.65 |
| | | Xu's | 36.99 | 387.25 | +3.25 |
| *Mobile* | 256 | Proposed | 27.76 | 257.11 | +1.11 |
| | | Liu's | 26.77 | 256.67 | +0.67 |
| | | Xu's | 25.81 | 259.25 | +3.25 |
| | 384 | Proposed | 29.04 | 382.65 | −1.35 |
| | | Liu's | 28.52 | 384.85 | +0.85 |
| | | Xu's | 27.90 | 387.28 | +3.28 |
| | 512 | Proposed | 31.24 | 520.13 | +8.13 |
| | | Liu's | 29.83 | 512.41 | +0.41 |
| | | Xu's | 29.30 | 515.29 | +3.29 |

decreases and fluctuates as shown in Fig. 13. Fig. 14 shows the buffer occupancy for the proposed bit allocation scheme. We see from these plots that the proposed rate control can maintain suitable buffer occupancy levels. In other words, the proposed bit allocation algorithm can prevent the buffer from overflow or underflow.

We also provide the comparison of the visual quality of reconstructed video sequences using these three bit allocation schemes. We show the ninth frame of the *Foreman* sequence with two layers (of CIF and QCIF resolutions) in Fig. 15. Since the spatial scalability is our main concern, we pay attention to the relatively still region of the textured background such as the walls of the building. Some regions are highlighted by rectangles for ease of comparison. We could also observe that the visual quality by the proposed algorithm outperforms that by JSVM for the test sequences. More importantly, it is very close to that of the optimal solution by the FS method.

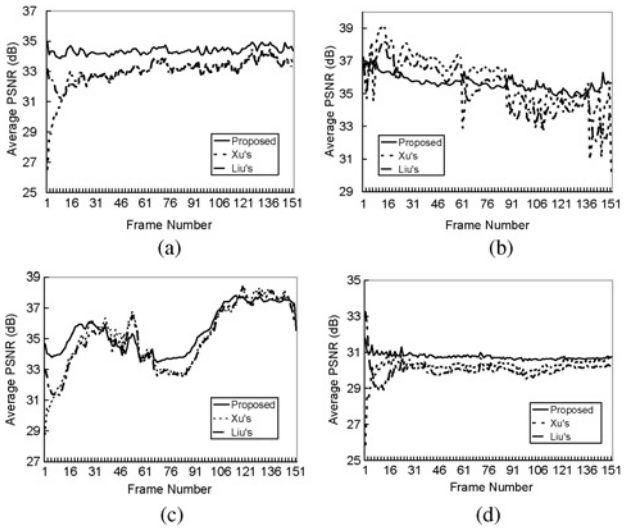


Fig. 12. Performance comparison of three algorithms for the averaged PSNR value of BL and EL as a function of the frame number in Scenario II. (a) *City*, $R_T$ = 1024 kb/s. (b) *Crew*, $R_T$ = 1536 kb/s. (c) *Soccer*, $R_T$ = 1536 kb/s. (d) *Harbour*, $R_T$ = 1536 kb/s.

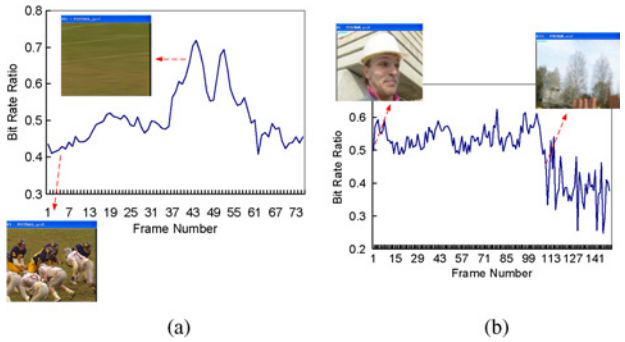


Fig. 13. Optimal bit rate ratio obtained by the proposed algorithm, $R_1/(R_1+R_2)$, as a function of the frame number. (a) *Football*, QCIF–CIF. (b) *Foreman*, QCIF–CIF.

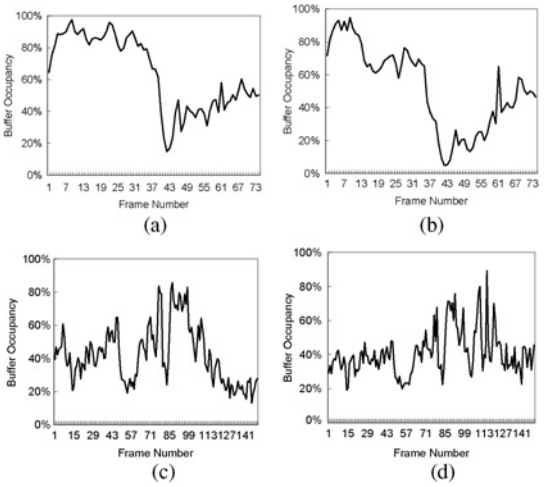


Fig. 14. Illustration of the buffer occupancy for each layer as a function of the frame number. (a) *Football*, BL. (b) *Football*, EL. (c) *Foreman*, BL. (d) *Foreman*, EL.

TABLE VIII
PERFORMANCE COMPARISON OF THE PROPOSED ALGORITHM, LIU'S ALGORITHM AND XU'S ALGORITHM FOR THE CIF−4CIF TWO LAYERS IN TERMS OF THE OUTPUT RATE, PSNR, AND Δ RATE

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | Rate (kb/s) | Δ Rate |
|---|---|---|---|---|---|
| *City* | 1024 | Proposed | 34.37 | 1027.55 | +3.55 |
| | | Liu's | 32.86 | 1025.51 | +0.51 |
| | | Xu's | 32.71 | 1030.38 | +6.38 |
| | 1536 | Proposed | 35.42 | 1519.14 | −16.86 |
| | | Liu's | 34.31 | 1537.83 | +1.83 |
| | | Xu's | 34.39 | 1542.48 | +6.47 |
| | 2048 | Proposed | 36.98 | 2041.23 | −6.77 |
| | | Liu's | 35.33 | 2050.80 | +2.80 |
| | | Xu's | 35.47 | 2054.47 | +6.47 |
| *Crew* | 1536 | Proposed | 35.65 | 1512.36 | −23.64 |
| | | Liu's | 35.25 | 1536.52 | +0.52 |
| | | Xu's | 34.17 | 1542.45 | +6.45 |
| | 2048 | Proposed | 37.25 | 2033.16 | −14.84 |
| | | Liu's | 36.41 | 2048.63 | +0.63 |
| | | Xu's | 35.52 | 2054.48 | +6.48 |
| | 3072 | Proposed | 38.68 | 3044.37 | −27.63 |
| | | Liu's | 37.97 | 3072.89 | −0.89 |
| | | Xu's | 37.29 | 3068.35 | −3.65 |
| *Soccer* | 1536 | Proposed | 35.54 | 1532.88 | −3.12 |
| | | Liu's | 34.89 | 1538.77 | +0.77 |
| | | Xu's | 34.77 | 1542.49 | +6.48 |
| | 2048 | Proposed | 37.51 | 2047.16 | −0.84 |
| | | Liu's | 36.08 | 2052.57 | +4.57 |
| | | Xu's | 36.04 | 2054.50 | +6.50 |
| | 3072 | Proposed | 38.99 | 3079.64 | +6.64 |
| | | Liu's | 37.77 | 3076.91 | −26.99 |
| | | Xu's | 37.89 | 3078.67 | +6.67 |
| *Harbour* | 1536 | Proposed | 30.76 | 1522.49 | −13.51 |
| | | Liu's | 29.99 | 1536.44 | +0.44 |
| | | Xu's | 30.31 | 1542.44 | +6.44 |
| | 2048 | Proposed | 31.94 | 2043.11 | −4.89 |
| | | Liu's | 31.15 | 2048.13 | +0.13 |
| | | Xu's | 31.64 | 2054.48 | +6.48 |
| | 3072 | Proposed | 33.37 | 3088.57 | +16.57 |
| | | Liu's | 32.86 | 3070.30 | −1.70 |
| | | Xu's | 33.49 | 3078.63 | +6.63 |

## IV. BIT ALLOCATION FOR MULTIPLE SPATIAL LAYERS

Although the discussion in Section III is restricted to the case of two-layer bit allocation, we can generalize its solution to the case of multiple spatial layers by recursion. Our idea is illustrated in Fig. 16. We consider an example of three layers consisting of QCIF, CIF, and 4CIF resolutions as shown in the top row of this figure. At the first decomposition stage, we apply the two layer decomposition and obtain the BL of the CIF resolution and the EL of the 4CIF resolution. The bit budgets for the BL and the EL will be assigned. Then, we perform the second stage decomposition on the BL video from the previous stage so that the new BL is of the QCIF resolution, the new EL of the CIF resolution and the total bit budget has to be the same as that assigned to the BL in the first stage. The general N-layer case is given in the second row of the figure. The shaded and the regular blocks in Fig. 16 represent the BL and the EL, respectively.


Fig. 15. Visual quality comparison among the proposed, JSVM and FS bit allocation schemes encoded by two spatial layers (QCIF–CIF), where the ninth frame of the *Foreman* sequence is shown. (a) FS, CIF. (b) FS, QCIF. (c) Proposed, CIF. (d) Proposed, QCIF. (e) JSVM, CIF. (f) JSVM, QCIF.

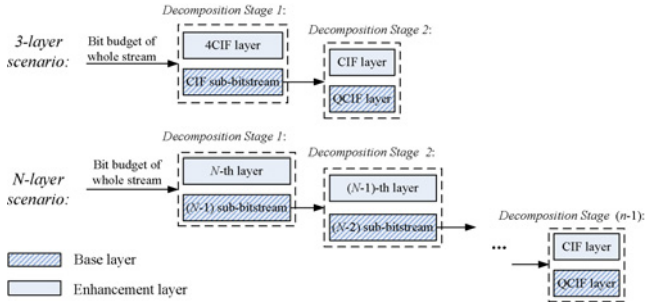


Fig. 16. Decomposition of a multilayer bit allocation problem into a sequence of two-layer bit allocation problems by recursion.

To evaluate the bit allocation algorithm for SVC with the multiple spatial layers, three layers were simulated with the setting and SVC configuration given in Table IX. Layer 1 is the base layer. Layers 2 and 3 are spatial enhancement layers, both of which are encoded using adaptive inter-layer prediction from their corresponding base layers (Layer 1 is the BL of Layer 2, and the integration of Layers 1 and 2 forms the base layer of Layer 3). The initial QP value is set to 32 and the frame rate is 30 frames/s. We tested four SVC sequences (*City*, *Crew*, *Harbour*, and *Soccer*) of various spatial complexities. The GOP size is set to 1 to provide IPPP structure. We compare the coding performance of three schemes; namely, the FS scheme, the proposed bit allocation scheme and the JSVM FixedQPEncoder.

The frame-by-frame PSNR performance of the three bit allocation methods is compared in Fig. 17. Again, we see that the proposed bit allocation scheme outperforms the JSVM significantly while the FS method provides the optimal results. We also compare the R-D performance of the proposed algorithm with that of the simulcast in Fig. 18, where two or more single-layer streams are transmitted together to provide the same functionality as a scalable one. We show results for the two-layer case in Fig. 18(a) and (b), where the R-D performance of the bit stream of QCIF resolution is represented by a triangle curve, which is the same for both the proposed and the simulcast schemes. The R-D performance of the proposed algorithm and the simulcast method is represented by the circle and the square curves, respectively. The bit streams of these two curves can provide coded video of the QCIF and the CIF resolutions at the same time. We see clearly that the proposed

TABLE IX
SETTING OF THREE-LAYER AND CONFIGURATION IN SPATIAL SCALABLE CODING

| Layer No. | Format | Frame Rate | Initial QP |
|---|---|---|---|
| 1 | QCIF | 30 | 32 |
| 2 | CIF | 30 | 32 |
| 3 | 4CIF | 30 | 32 |
| Profile | Scalable | SearchMode | FastSearch |
| | Baseline | SearchRange | 16 |

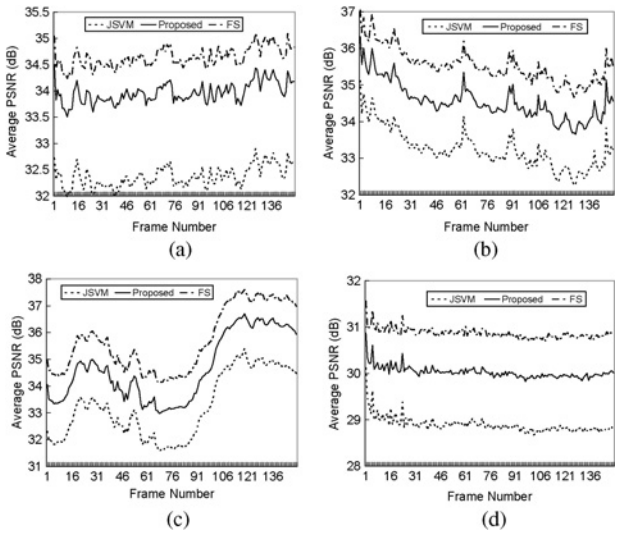


Fig. 17. PSNR value as a function of the frame number with the proposed, JSVM and FS bit allocation schemes for the three-layer case. (a) *City*, $R_T$ = 1024 kb/s. (b) *Crew*, $R_T$ = 1536 kb/s. (c) *Soccer*, $R_T$ = 1536 kb/s. (d) *Harbour*, $R_T$ = 1536 kb/s.

algorithm outperforms the simulcast method substantially. This is especially true for video with a lot of motion. Results for the three-layer case are shown in Fig. 18(c) and (d), where the comparison is made for three scenarios: 1) the base layer (QCIF) only; 2) the bottom two layers (QCIF and CIF); and 3) all three layers (QCIF, CIF and 4CIF). Again, we see that the spatial scalable coding of the proposed algorithm outperforms the simulcast in coding efficiency for the later two scenarios.

TABLE X
PERFORMANCE OF THE PROPOSED AND JSVM SCHEMES FOR THREE-LAYER IN TERMS OF OUTPUT RATE, PSNR, Δ RATE AND ITERATION NUMBER

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | Rate (kb/s) | Δ Rate (kb/s) | Iter. |
|---|---|---|---|---|---|---|
| *City* | 1024 | Proposed | 33.96 | 1027.55 | +3.55 | 6 |
| | | JSVM | 32.37 | 1013.67 | −10.33 | 53 |
| | 1536 | Proposed | 35.42 | 1519.14 | −16.86 | 6 |
| | | JSVM | 33.84 | 1539.76 | +3.76 | 44 |
| | 2048 | Proposed | 36.78 | 2041.23 | -6.77 | 6 |
| | | JSVM | 35.04 | 1959.94 | −88.06 | 53 |
| *Crew* | 1536 | Proposed | 34.56 | 1512.36 | −23.64 | 6 |
| | | JSVM | 33.24 | 1519.13 | −13.87 | 79 |
| | 2048 | Proposed | 35.70 | 2033.16 | −14.84 | 6 |
| | | JSVM | 34.37 | 2057.61 | +9.61 | 27 |
| | 3072 | Proposed | 37.43 | 3044.37 | −27.63 | 6 |
| | | JSVM | 36.13 | 2913.41 | −158.59 | 25 |
| *Harbour* | 1536 | Proposed | 30.02 | 1532.88 | −3.12 | 6 |
| | | JSVM | 28.89 | 1542.67 | +6.67 | 61 |
| | 2048 | Proposed | 31.67 | 2047.16 | −0.84 | 6 |
| | | JSVM | 30.14 | 2014.38 | −33.62 | 52 |
| | 3072 | Proposed | 32.97 | 3079.64 | +6.36 | 6 |
| | | JSVM | 31.88 | 2976.06 | −95.94 | 62 |
| *Soccer* | 1536 | Proposed | 34.66 | 1532.88 | −3.12 | 6 |
| | | JSVM | 33.24 | 1521.85 | −15.15 | 59 |
| | 2048 | Proposed | 36.21 | 2047.16 | −0.84 | 6 |
| | | JSVM | 34.77 | 2041.40 | −6.60 | 47 |
| | 3072 | Proposed | 38.99 | 3079.64 | +6.36 | 6 |
| | | JSVM | 36.69 | 3057.04 | −14.86 | 31 |

Coding results of the proposed algorithm and JSVM are summarized in Table X, where we show the averaged PSNR value, the deviation from the target bit rate and the number of iteration required. The proposed bit allocation algorithm has an average PSNR gain of 1.48 dB over JSVM. The proposed algorithm has a fixed number of iteration whereas that by the JSVM FixedQPEncoder is determined by the iteration stopping criteria introduced in Section III-D. For the first two-layer decomposition, we need three encoding passes to build the R and D models as stated in Section III-D. We need two more encoding passes to obtain parameters $\zeta$ and $\upsilon$ of the new EL. Since parameter $\beta_2$ of the two EL models is almost same, it is only computed once. One additional encoding pass is needed to encode the whole frame according to these R and D models. Thus, the total number of iteration is equal to 6. The JSVM FixedQPEncoder has a much higher iteration number, which implies a higher computational complexity.

The frame-to-frame quality comparison among the proposed algorithm, the rate control algorithms proposed by Xu *et al.* [8] and Liu *et al.* [10] is given Fig. 19. It is clear that the proposed method achieves better performance. More R-D comparison data are shown in Table XI. The proposed algorithm achieves an averaged PSNR gain of 1.55 dB and 1.65 dB over Liu's algorithm and Xu's algorithm, respectively.

Finally, we compare the visual quality of reconstructed video using the three bit allocation schemes for the three-layer case. We show the 5th frame of the *City* sequence of 4CIF resolution in Fig. 20. We see that the proposed bit allocation algorithm gives a better result than JSVM and provides a result close to the FS scheme. In particular, we could observe that regions with high spatial complexity (those enclosed by rectangles) are well preserved by the proposed algorithm whereas they are not very clear by the JSVM benchmark.

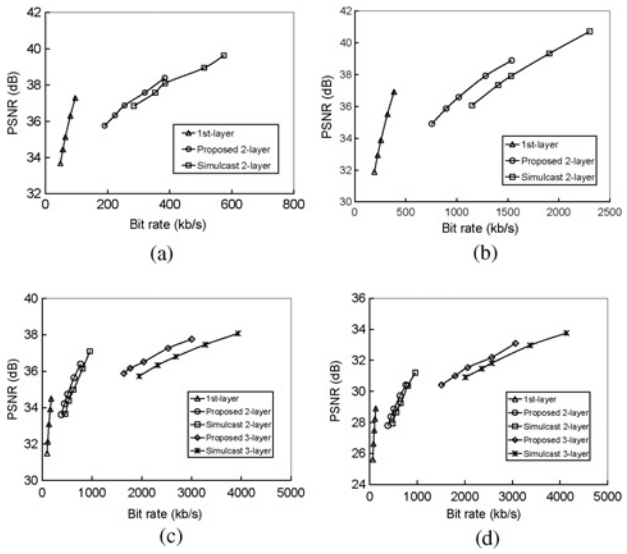


Fig. 18. Illustration of the R-D performance comparison between simulcast and the proposed algorithm. (a) *Foreman*, two-layer. (b) *Football*, two-layer. (c) *Harbour*, three-layer. (d) *Crew*, three-layer.

TABLE XI
PERFORMANCE OF THE PROPOSED, LIU'S AND XU'S BIT ALLOCATION ALGORITHMS FOR THREE-LAYER IN TERMS OF THE OUTPUT RATE, PSNR, AND Δ RATE

| Seq. | Target Rate (kb/s) | Method | PSNR (dB) | Rate (kb/s) | Δ Rate |
|---|---|---|---|---|---|
| *City* | 1024 | Proposed | 33.96 | 1027.55 | +3.55 |
| | | Liu's | 33.13 | 1025.66 | +1.66 |
| | | Xu's | 33.13 | 1034.55 | +10.55 |
| | 1536 | Proposed | 35.42 | 1519.14 | −16.86 |
| | | Liu's | 34.65 | 1537.89 | +1.66 |
| | | Xu's | 34.85 | 1546.64 | +10.64 |
| | 2048 | Proposed | 36.78 | 2041.23 | −6.77 |
| | | Liu's | 35.73 | 2051.02 | +3.02 |
| | | Xu's | 36.00 | 2058.55 | +10.55 |
| *Crew* | 1536 | Proposed | 34.56 | 1512.36 | −23.64 |
| | | Liu's | 32.73 | 1536.54 | +0.54 |
| | | Xu's | 32.43 | 1546.62 | +10.62 |
| | 2048 | Proposed | 35.70 | 2033.16 | −14.84 |
| | | Liu's | 34.01 | 2048.63 | +0.63 |
| | | Xu's | 33.87 | 2058.54 | +10.54 |
| | 3072 | Proposed | 37.43 | 3044.37 | −27.63 |
| | | Liu's | 35.88 | 3072.48 | +0.48 |
| | | Xu's | 35.83 | 3071.84 | −0.16 |
| *Harbour* | 1536 | Proposed | 30.02 | 1532.88 | −3.12 |
| | | Liu's | 28.55 | 1536.44 | +0.44 |
| | | Xu's | 28.60 | 1546.68 | +10.68 |
| | 2048 | Proposed | 31.67 | 2047.16 | −0.84 |
| | | Liu's | 29.68 | 2048.86 | +0.84 |
| | | Xu's | 29.94 | 2058.76 | 10.76 |
| | 3072 | Proposed | 32.97 | 3079.64 | +6.36 |
| | | Liu's | 31.37 | 3072.35 | +0.35 |
| | | Xu's | 31.79 | 3074.32 | +2.32 |
| *Soccer* | 1536 | Proposed | 34.66 | 1532.88 | −3.12 |
| | | Liu's | 33.13 | 1540.81 | +4.80 |
| | | Xu's | 32.40 | 1546.62 | +10.62 |
| | 2048 | Proposed | 36.21 | 2047.16 | −0.84 |
| | | Liu's | 34.46 | 2053.77 | +5.76 |
| | | Xu's | 33.84 | 2058.58 | +10.58 |
| | 3072 | Proposed | 38.99 | 3079.64 | +6.36 |
| | | Liu's | 36.40 | 3080.38 | +8.38 |
| | | Xu's | 35.87 | 3076.12 | +4.12 |

## V. COMPARISON BETWEEN GOP-BASED AND FRAME-BASED BIT ALLOCATION

Two bit allocation strategies for the spatial scalability were discussed in Section II. Strategy I uses a GOP as the optimization unit, where all frames of different types in each spatial layer are treated as a combined unit. For one GOP, we decouple the dependent relation between spatial layers and allocate bits to these layers in the first step. Then, we optimize the quantization parameter for all frames in one layer in the second step. For this reason, we call Strategy I as the GOP-based bit allocation with respect to the spatial scalability. The advantage of the GOP-based approach is that it can compensate the discrepancy of the spatial layer R-D model in the first step with the frame-based R-D model in the second step. On the other hand, it demands larger delay in the encoding process. The GOP-based approach was examined before by authors in [11]. In contrast, Strategy II allocates the bit budget to each frame set and then perform the bit allocation between spatial layers within one frame. Thus, we call Strategy II the frame-based bit allocation. The frame-based approach can reduce the encoding latency and is suitable for conversational applications.

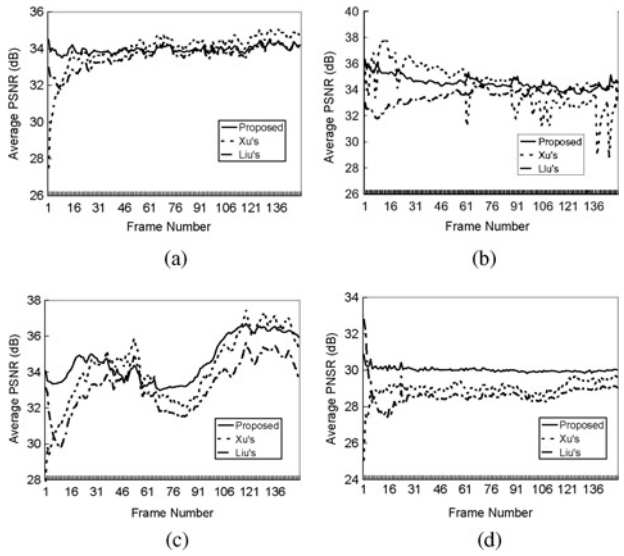


Fig. 19. Comparison of the frame-to-frame PSNR values of the proposed, Xu's and Liu's bit allocation algorithms for the three-layer case. (a) *City*, $R_T$ = 1024 kb/s. (b) *Crew*, $R_T$ = 1536 kb/s. (c) *Soccer*, $R_T$ = 1536 kb/s. (d) *Harbour*, $R_T$ = 1536 kb/s.

Regardless of encoding time delay, it is interesting to compare the coding performance of these two bit allocation strategies for the spatial scalability of H.264/SVC. Here, we choose the two-layer case in the test with Scalable Baseline Profile, where Layer 1 is the BL and Layer 2 is EL using adaptive inter-layer prediction from BL. We have encoded three video sequences with low to high spatial complexity. They are: *Akiyo*, *Football* and *City*. The GOP size is set to 8 and 16, respectively.

The performance of these two strategies is compared in Fig. 21. When the target bit rate is lower, we see from result figures that the GOP-based scheme achieves better R-D performance than the frame-based scheme. When the bit budget increases, the difference between these two strategies becomes smaller. For the video sequences with high spatial complexity such as *City*, the GOP-based scheme takes full advantage of the temporal dependence between adjacent frames within the same spatial layer. Consequently, it has a better result than the frame-based scheme. These results indicate that coding efficiency can be improved by increasing the GOP size at the tradeoff of longer encoding/decoding delay.

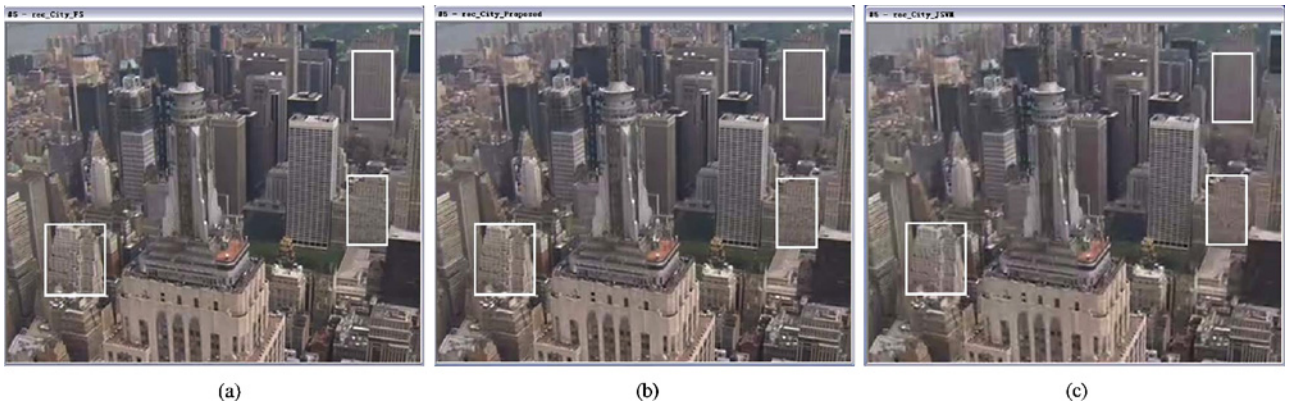

Fig. 20. Visual quality comparison among the proposed, JSVM and FS bit allocation schemes encoded by three spatial layers (QCIF–CIF–4CIF), where the fifth frame of the *City* sequence is shown. (a) FS. (b) Proposed. (c) JSVM.

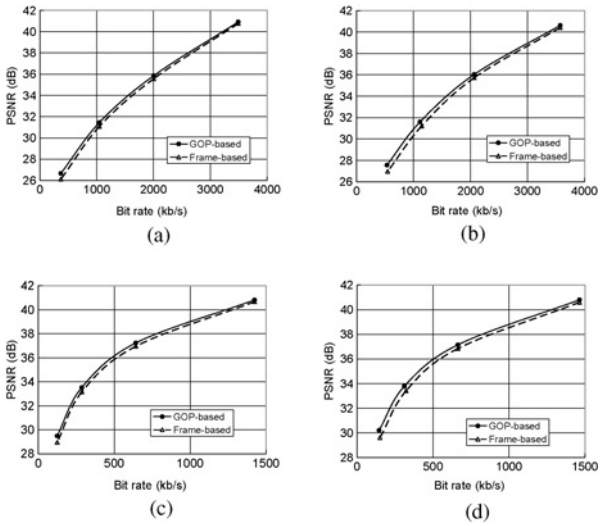


Fig. 21. R-D performance comparison between GOP-based and frame-based bit allocation schemes. (a) *Football*, GOP = 8. (b) *Football*, GOP = 16. (c) *City*, GOP = 8. (d) *City*, GOP = 16.

## VI. Conclusion and Future Work

We proposed a model-based spatial layer bit allocation algorithm for H.264/SVC in this paper. We first focused on the case of two spatial layers, derived the rate and distortion models analytically, and developed a low-complexity (in terms of a smaller iteration number) bit allocation algorithm by considering the dependent layer in the spatial scalability. Then, we extended this result to multilayer bit allocation by performing the two-layer allocation scheme recursively. Finally, we compared the performance of GOP-based and frame-based spatial layer bit allocation schemes at a fixed temporal resolution. The superior performance of the proposed spatial layer bit allocation algorithm was demonstrated using the reference software JSVM and two prior H.264/SVC rate control algorithms as the benchmarks.

Although the reconstructed lower-resolution video represents the low frequency information in spatial scalable video, it may not necessarily be the most suitable reference for inter-layer prediction. Actually, the spatial predictor has to compete with the temporal predictor. For example, for video sequences with slow motion and high spatial detail, the temporal prediction may be more effective than the spatial prediction. It is a challenging and open problem to solve the joint spatial-temporal bit allocation problem for H.264/SVC, which has been first considered with H.263+ in [17]. It is desirable to reduce these two-dimension dependent relations to further improve the R-D performance of the SVC bit stream and reduce the quality fluctuation between frames.

## Acknowledgment

The authors would like to thank the assistance of Dr. L. Xu and Dr. Y. Liu in offering their source codes for performance benchmarking in the experiments. They would also like to thank the reviewers for their careful reviews and valuable comments.

## References

[1] *Amendment 3 to ITU-T Rec. H.264 (2005) ISO/IEC 14496-10: 2005*, Scalable Video Coding, Jul. 2007.
[2] 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.
[3] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sep. 2007.
[4] 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.
[5] 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.
[6] S. Liu and C.-C. J. Kuo, "Joint temporal-spatial bit allocation for video coding with dependency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 15–26, Jan. 2005.
[7] 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. Technol.*, Feb. 2007, pp. 222–227.
[8] L. Xu, W. Gao, X. Ji, D. Zhao, and S. Ma, "Rate control for spatial scalable coding in SVC," in *Proc. Picture Coding Symp.*, Nov. 2007.
[9] Y. Liu, Y. C. Soh, and Z. G. Li, "Rate control for spatial/CGS scalable extension of H.264/AVC," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 1746–1750.
[10] Y. Liu, Z. 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.

[11] J. Liu, Y. Cho, Z. Guo, and C.-C. J. Kuo, "Bit allocation for spatial scalability in H.264/SVC," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2008, pp. 278–283.
[12] *Joint Scalable Video Model Software 9.6*, Joint Video Team of ITU-T VCEG and ISO/IEC MPEG [Online]. Available: ftp://garcon.ient.rwth-aachen.de
[13] 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.
[14] S. Liu and C.-C. J. Kuo, "Complexity reduction of joint temporal-spatial bit allocation with R-D models," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, pp. 729–732.
[15] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression Video Coding: Video Coding for Next-generation Multimedia*. Chichester, U.K.: Wiley, 2003.
[16] *Testing Conditions for SVC Coding Efficiency and JSVM Performance Evaluation*, document JVT-Q205.doc, Joint Video Team, Oct. 2005.
[17] H. Song and C.-C. J. Kuo, "Rate control for low-bit-rate video via variable-encoding frame rates," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 4, pp. 512–521, Apr. 2001.

**Jiaying Liu** (S'09) received the B.E. degree in computer science from Northwestern Polytechnic University, Xi'an, 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 Student with the University of Southern California, Los Angeles, invited by Prof. C.-C. Jay Kuo. She is currently a Faculty Member with the Institute of Computer Science and Technology, Peking University. Her current research interests include scalable video coding, rate-distortion analysis, and visual signal processing.

**Yongjin Cho** (S'08) received the B.S degree in electronic engineering from Yonsei University, Seoul, Korea, in 2001, and the M.S. degree in computer science from the University of Southern California (USC), Los Angeles, in 2003. He is currently pursuing the Ph.D. degree in electrical engineering from the Ming Hsieh Department of Electrical Engineering, USC.

His current research interests include rate-distortion optimal scalable video coding and efficient delivery of digital video.

**Zongming Guo** (M'09) received the B.S. degree in mathematics, and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively.

He is currently a Professor with the Institute of Computer Science and Technology, Peking University. His current research interests include video coding and processing, watermarking, and communication.

Dr. Guo is the executive member of China-Society of Motion Picture and Television Engineers. He received the First Prize of the State Administration of Radio Film and Television Award, the First Prize of the Ministry of Education Science and Technology Progress Award, and the Second Prize of the National Science and Technology Award, in 2004, 2006, and 2007, respectively. He received the Wang Xuan News Technology Award and the Chia Tai Teaching Award, in 2008, and received Government Allowance granted by the State Council, in 2009.

**C.-C. Jay Kuo** (S'83–M'86–SM'92–F'99) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively.

He is currently the Director of the Signal and Image Processing Institute, and a Professor of electrical engineering and Computer Science with the Ming Hsieh Department of Electrical Engineering and Integrated Media Systems Center, University of Southern California (USC), Los Angeles. He is the co-author of about 170 journal papers, 800 conference papers, and ten books. He has guided about 100 students to their Ph.D. degrees and supervised 20 post-doctoral research fellows. Currently, his research group at USC has around 30 Ph.D. students (http://viola.usc.edu), which is one of the largest academic research groups in multimedia technologies. He has delivered over 440 invited lectures in conferences, research institutes, universities, and companies. His current research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and biological signal/image processing.

Dr. Kuo is a Fellow of the International Society for Optical Engineers. He is the Editor-in-Chief for the *Journal of Visual Communication and Image Representation* (an Elsevier journal), and an Editor for the *LNCS Transactions on Data Hiding and Multimedia Security* (a Springer journal), the *Journal of Advances in Multimedia* (a Hindawi journal), and the *EURASIP Journal of Advances in Signal Processing* (a Hindawi journal). He was on the editorial board of the IEEE SIGNAL PROCESSING MAGAZINE from 2003 to 2004. He served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1995 to 1998, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 1995 to 1997, and the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2001 to 2003. He received the National Science Foundation Young Investigator Award and the Presidential Faculty Fellow Award in 1992 and 1993, respectively. He received the Northrop Junior Faculty Research Award from the USC Viterbi School of Engineering in 1994. He received the Best Paper Award from the Multimedia Communication Technical Committee of the IEEE Communication Society in 2005, from the IEEE Vehicular Technology Fall Conference in 2006, and from the IEEE Conference on Intelligent Information Hiding and Multimedia Signal Processing in 2006. He was an IEEE Signal Processing Society Distinguished Lecturer in 2006, the recipient of the Okawa Foundation Research Grant in 2007, and the recipient of the Electronic Imaging Scientist of the Year Award in 2010. He is an Advisor to the Society of Motion Picture Television Engineers-USC Student Chapter.