Iterative Rate-Distortion Optimization of H.264 With Constant Bit Rate Constraint

Cheolhong An, *Student Member, IEEE*, and Truong Q. Nguyen, *Fellow, IEEE*

Abstract—**In this paper, we apply the primal-dual decomposition and subgradient projection methods to solve the rate-distortion optimization problem with the constant bit rate constraint. The primal decomposition method enables spatial or temporal prediction dependency within a group of picture (GOP) to be processed in the master primal problem. As a result, we can apply the dual decomposition to minimize independently the Lagrangian cost of all the MBs using the reference software model of H.264. Furthermore, the optimal Lagrange multiplier λ^* is iteratively derived from the solution of the dual problem. As an example, we derive the optimal bit allocation condition with the consideration of temporal prediction dependency among the pictures. Experimental results show that the proposed method achieves better performance than the reference software model of H.264 with rate control.**

Index Terms—**Bit allocation, H.264, primal-dual decomposition, rate control, rate-distortion (RD) optimization, subgradient.**

I. Introduction

After rate-distortion (RD) optimization is introduced for video compression using the Lagrange multiplier [1], [2], there are many methods to reduce the complexity in deciding macroblock (MB) modes, motion vectors (MVs) for a given Lagrange multiplier λ. Even though RD optimization method is not mandatory for standard video compression, such as H.264 [3], it is the main part of video coding to improve the coding efficiency [2], [4]. Therefore, we review the relation between RD optimization and previous works. RD optimization with inequality constraint in a frame is mathematically formulated as follows:

$$\min_{\mathbf{m}} \quad \sum_{n=1}^{N} d_n(\mathbf{m}_n) \tag{1}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} x_n(\mathbf{m}_n) \leq X_F \tag{2}$$

where $\mathbf{m}_n = (M_n, \mathbf{MV}_n, QP_n, \mathbf{Ref}_n)$ is a vector of MB mode, MVs, quantization parameter (QP) and reference frames for inter prediction. N is the number of MBs in a frame and X_F is the bit constraint of a frame. d_n and x_n are distortion and coded bits of the nth MB, respectively. The optimization problem (1) can be solved by the Lagrangian duality in order to obtain the optimal solution if the problem is a convex optimization problem and satisfies the Slater's condition [5]. The Slater's condition is easily satisfied since there are $\mathbf{m}$ vectors which make sum of coded bits less than X_F, but the problem (1) mathematically is not a convex optimization problem since distortion function $d_n(\mathbf{m}_n)$ is not a convex function [1] and a feasible set $\mathbf{m}_n$ is not a convex set [5]. However, the near optimal solution of the primal problem (1) can be obtained if duality gap is small [6]. Therefore, the Lagrange duality is applied and the dual function of the primal problem (1) is

$$q(\lambda) = \min_{\mathbf{m}} \quad \sum_{n=1}^{N} \left(d_n(\mathbf{m}_n) + \lambda x_n(\mathbf{m}_n) \right) - \lambda X_F \tag{3}$$

and its dual problem is

$$\max_{\lambda \geq 0} q(\lambda). \tag{4}$$

If we know the optimal solution of the dual problem, we can obtain the solution of the primal problem (1) after solving (3). However, in order to simplify the above optimization problems, the relation between λ and QP was derived in [2], [7]–[9] and estimation of X_F from QP was studied in [10], [11]. The reference software model of H.264 (simply denoted as JM model) [12] has the following relation:

$$\lambda = \kappa 2^{((QP-12)/3)} \tag{5}$$

$$X_F = aQP^{-1} + bQP^{-2} \tag{6}$$

where κ is a function of picture types (I, P, B), the number of referenced frames and QP, and a and b are estimated using the linear regression based on mean absolute difference (MAD) and target bits. Equations (5) and (6) give estimated solution for λ of the dual problem (4), that is, QP[1] is estimated from (6) for a given constraint X_F and then λ is induced from (5). Thus, JM model does not directly solve the dual problem (4). For a given λ, JM model minimizes the Lagrangian function, that is, solves the problem (3). However, if there is no bit constraint, we can just choose any QP to derive λ from (5). Consequently, JM model has two coding modes: one is a coding mode without a rate constraint and the other is with a rate constraint.

Without a rate constraint, users just specify any QP and group of picture (GOP) structure, and then JM model solves the problem (3). As a result, users do not know how many bits are generated after encoding. In this case, reference frames, QP, and λ are given, the optimization variables are MB modes and MVs for all MBs of a frame. This problem can be simplified

Manuscript received June 7, 2007; revised May 10, 2008. First published August 4, 2008; last published August 13, 2008 (projected). This work was supported in part by CWC and in part by the UC Discovery program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ricardo L. de Queiroz.

The authors are with the Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: chan@ucsd.edu; nguyent@ece.ucsd.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/TIP.2008.2001046

[1] For simplicity, we directly denote QP instead of Qstep in (6), and QP is derived from the mapping between QP and Qstep.

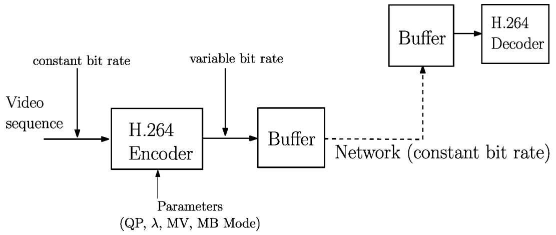

Fig. 1. Example of video streaming.

by the independent assumption among the MBs. Consequently, the problem (3) is

$$q(\lambda) = \sum_{n=1}^{N} \min_{\mathbf{m}_n} \left(d_n(\mathbf{m}_n) + \lambda x_n(\mathbf{m}_n)\right) - \lambda X_F \qquad (7)$$

where the optimization variables are MB mode and MVs for each MB. This optimization problem is solved by the following method. First, fix a MB mode of all the inter MB modes and then find optimal MVs with or without considering both residual bits and MV bits for the MB mode. Next, given the optimal MVs for inter MB modes, find the optimal MB mode which minimizes the Lagrangian cost $l_n(\mathbf{m}_n)$, that is, $d_n(\mathbf{m}_n)+\lambda x_n(\mathbf{m}_n)$ among the inter and the other MB modes such as intra MB and direct MB modes. In order to reduce the loss of coding efficiency of independent assumption, [1], [13], [14] solve the dependent optimization problem (3) using dynamic programming without considering frame-level dependency or λ. Reference [15] considers the frame-level dependent coding problem using the Viterbi algorithm (VA), but it considers that distortion and coded bits are only function of QP.

With a rate constraint, users specify the coded bit rate and GOP structure, and then JM model solves the problem (7) with independent assumption. Although λ and QP are obtained from (5) and (6), bit constraint X_F in (3) should be derived from user bit rate constraint because user bit rate constraint is average bits per second, but not MB-level or frame-level bit constraint. Therefore, MB-level or frame-level which are generalized as a basic unit (a group of MBs) [16], [17] and GOP-level bit constraints need to be derived from a given user bit rate constraint. JM model uses the basic unit for a bit constraint. Without loss of generality, the basic unit is considered as a MB or a frame in this paper. If we find target bits for a basic unit, the other parameters can be obtained from (5) and (6). References [16]–[19] show how to estimate target bits of a basic unit from user bit rate constraint, video frame rate, buffer fullness, picture type and some other information.

Even though rate control (RC) induces loss of performance which will be shown in Section V, it is well known that RC algorithms are necessary in the video streaming applications to satisfy the network bit rate to avoid a buffer overflow or underflow. In this section, we briefly mention necessity of RC with a simple example of video streaming system in Fig. 1. The original video sequences have constant bit rate according to the frame rate and frame size but the output bit rate of video encoder becomes variable bit rate since intra and inter prediction errors of each frame highly depend on correlation among the frames and within a frame. Thus, the variable coded bits should be smoothed out to the constant or variable network through the buffer as shown in Fig. 1. The virtual buffer fullness which has negative fullness is illustrated in Fig. 2 after encoding with and without RC. The decrease of buffer fullness indicates input rate of the buffer is lower than output rate, otherwise, input rate is higher than output rate. The buffer fullness of encoding without RC, that is, with QP fixed for all frames highly fluctuates. The actual buffer fullness is around zero before 70th frame since there is no negative buffer fullness and then buffer fullness increases. This phenomenon results from increase of prediction errors due to the high motion and scene change after 70th frame. Thus, video streaming applications without RC require a larger buffer to compensate bit fluctuation. However, we do not know how large a buffer can compensate the bit fluctuation before encoding the whole sequence. Furthermore, different sequences which may never be coded before require different buffer size which is not known. Therefore, encoder needs to control bit rate to prevent a buffer overflow and underflow for given buffer limitation. Fig. 2 shows that the coded bits of encoder with RC fluctuates within the buffer limitation. Encoder with RC mainly changes QP in order to control bit rate which induces larger change of quantization distortion as shown in Fig. 3. Even though quality of coded sequences without RC is better due to the low variation of peak signal-to-noise ratio (PSNR), the buffer management is required and accomplished through RC for the bit constraint with minimum distortion. This operation is mathematically formulated as a RD optimization problem (1).

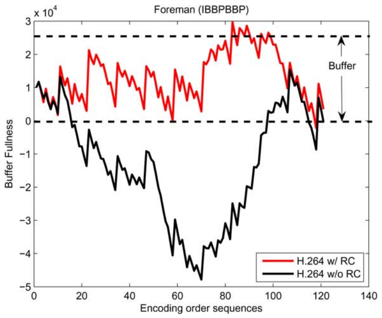

Fig. 2. Virtual buffer fullness with and without RC.

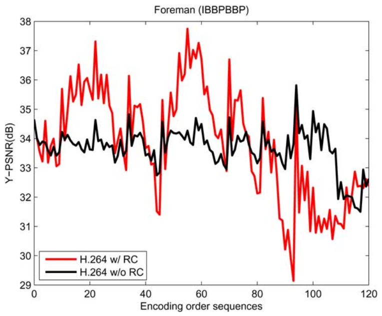

Fig. 3. PSNR of decoded sequences with and without RC.

However, JM model of H.264 mainly focuses on real-time or low complexity rate control scheme with the constant or variable bit rate constraint. Therefore, the rate control method induces loss of coding efficiency, and it cannot tightly satisfy the bit constraint which are shown in this paper. In case of nonreal time applications with the constant bit rate constraint, the loss can be reduced. In this paper, we apply the primal-dual decomposition and subgradient projection methods to solve directly the problem (1) with the constant GOP bit constraint. Although this method can be used for the optimal bit allocation of any basic unit with consideration of spatial and temporal prediction dependency, we show the frame-level bit allocation within a GOP with considering temporal prediction dependency as an example. Thus, we ignore spatial prediction dependency, that is, MBs which have intra and spatial direct modes are independent.

The rest of this paper is organized as follows. We introduce iterative RD optimization with geometric interpretation in Section II. In Section III, we explain primal-dual decomposition and subgradient projection, and we apply these methods for MB-level bit allocation of an intrasliced picture with independent assumption. In Section IV, frame-level bit allocation is considered with temporal dependency. Experimental results are shown in Section V. Section VI concludes the paper.

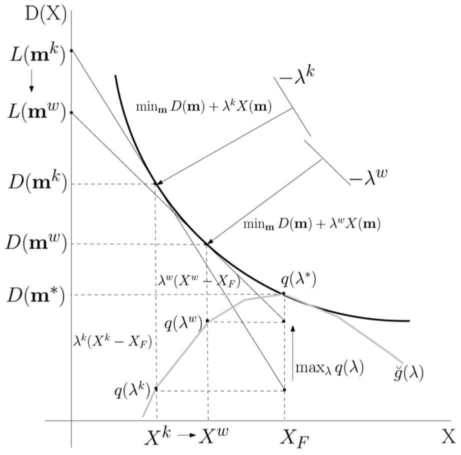

Fig. 4. Geometric interpretation of iterative RD optimization.

II. Iterative Rate-Distortion Optimization

In this section, we solve the problem (1) for a given frame bit constraint X_F using the iterative RD optimization. The procedure is explained with geometric interpretation which is shown in Fig. 4. Here, we assume that RD function is smooth and continuous for simplicity. In Fig. 4, let $\sum_n d_n(\mathbf{m}_n)$, $\sum_n x_n(\mathbf{m}_n)$ and $\sum_n l_n(\mathbf{m}_n)$ be $D(\mathbf{m})$, $X(\mathbf{m})$ and $L(\mathbf{m})$ for frame-level distortion, coded bits and the Lagrangian cost, respectively. k and w are iteration indices and X^k is $X(\mathbf{m}^k)$ where $\mathbf{m}^k$ is the solution of (3) at k iteration. For a given λ^k, JM model solves the problem (3) that is equivalent to finding $\mathbf{m}^k$ associated with the point ($X(\mathbf{m}^k)$, $D(\mathbf{m}^k)$) at which λ^k is tangent to the RD function [1]. It induces the minimum of $L(\mathbf{m})$ for a given λ^k. If the bit constraint X_F is $X(\mathbf{m}^k)$, $\mathbf{m}^k$ and λ^k are the optimal solutions of the primal problem (1) and dual problem (4), respectively. Thus, RD optimization without a bit constraint can always achieve the optimal solutions for a given λ with assumption that the coded bits are constraint bits. If there is a bit constraint and $X_F \neq X(\mathbf{m}^k)$, $\mathbf{m}^k$ and λ^k are not the optimal solutions of the primal problem (1) and dual problem (4). In this case, we can use iterative methods to find the optimal solutions. In Fig. 4, the dual function value $q(\lambda^k)$ of (3) which is marked on the line of X^k is $D(\mathbf{m}^k) + \lambda^k(X^k - X_F)$. If we mathematically know the dual function, it is easy to find the solution of dual problem (4) since the constraint is simple. From the optimal dual solution λ^*, JM model can find the optimal primal solution. However, it is difficult to find the dual function. Therefore, we try a different λ^w which increases the dual function value. Fig. 4 shows that if λ^w is a little bit smaller than λ^k, $q(\lambda^w)$ is larger than $q(\lambda^k)$ which means that λ^w is closer than λ^k to the optimal solution of the dual problem (4). On the contrary, λ^w decreases the frame-level Lagrangian cost from $L(\mathbf{m}^k)$ to $L(\mathbf{m}^w)$.

In Section III, we discuss how to decide next iteration λ^w from λ^k. After several iterations, if we find λ^* which maximizes the dual function, the optimal primal solution $\mathbf{m}^*$ can be found for a given λ^*. Then $q(\lambda^*)$ is $D(\mathbf{m}^*)$ since $X(\mathbf{m}^*)$ is X_F. It is well known result from the Karush–Kuhn–Tucker (KKT) condition [5], [6]. However, the practical RD function is not a convex function and not continuous. Therefore, we need to find operational RD function which consists of convex-hull points. Furthermore, there can be no feasible solution to satisfy a bit constraint X_F since $\mathbf{m}$ are discrete. Therefore, we allow constraint violation within some range and find the best solution around a bit constraint. After several iterations, we can linearly approximate the dual function $q(\lambda)$ as $\check{q}(\lambda)$ which is illustrated in Fig. 4. If λ^* is obtained, QP can be derived from (5) or QP can be an optimization variable in the problem (3) as discussed in [7] and [8].

III. Primal-Dual Decomposition and Subgradient Projection

In this section, we introduce the general framework to solve optimization problems using the primal-dual decomposition [6], [20], [21]. The primal decomposition corresponds to deciding the optimal bit constraint of a basic unit. Dual decomposition and the Lagrangian duality, as explained in previous section, are equivalent to obtaining the optimal primal and dual solution for the given optimal bit constraint as a result of the primal decomposition. For convenience, we simplify the notation of problem (1) as follows:

$$\min_{\mathbf{m}} \quad \sum_n d(\mathbf{m}_n), \text{ s.t. } \sum_n x_n(\mathbf{m}_n) \leq X \tag{8}$$

$$\begin{aligned} \min_{\mathbf{y}} \quad & \min_{\mathbf{m}} \sum_n d(\mathbf{m}_n) \\ \text{s.t.} \quad & x_n(\mathbf{m}_n) \leq y_n, \quad \sum_n y_n \leq X, \quad \forall\, n \end{aligned} \tag{9}$$

$$\min_{\mathbf{m}} \quad \sum_n d(\mathbf{m}_n), \text{ s.t. } x_n(\mathbf{m}_n) \leq y_n, \quad \forall n \tag{10}$$

$$\min_{\mathbf{y}} \quad q^*(\mathbf{y}), \text{ s.t. } \sum_n y_n \leq X \tag{11}$$

where $q^*(\mathbf{y}) = \min_{\mathbf{m}} \sum_n d(\mathbf{m}_n) + \lambda_n^*\{x_n(\mathbf{m_n}) - y_n\}$ which is the optimal value of the problem (10). The original problem (8) can be reformulated into the problem (9) by introducing slack variable $\mathbf{y}$. Then the problem (9) can be decomposed into two optimization problems (10) and (11) with respect to (w.r.t.) optimization variables $\mathbf{m}$ and $\mathbf{y}$, respectively. The decomposition from problem (8) to problem (11) is called as a master primal decomposition, and the decomposition from problem

(10) to problem (13) is the dual decomposition. Problem (10) is solved by the Lagrangian duality as follows:

$$q(\mathbf{y},\lambda) = \min_{\mathbf{m}} \sum_n d(\mathbf{m}_n) + \lambda_n\{x_n(\mathbf{m}_n) - y_n\} \quad (12)$$

$$= \sum_n \min_{\mathbf{m}_n} d(\mathbf{m}_n) + \lambda_n\{x_n(\mathbf{m}_n) - y_n\} \quad (13)$$

$$q^*(\mathbf{y}) = \max_{\boldsymbol{\lambda}\succeq 0} q(\mathbf{y},\lambda) \quad (14)$$

$$= \max_{\boldsymbol{\lambda}\succeq 0} \sum_n q_n(y_n,\lambda_n)$$

$$= \sum_n \max_{\lambda_n \geq 0} q_n(y_n,\lambda_n) \quad (15)$$

$$= \sum_n q_n^*(y_n) \quad (16)$$

where $q_n(y_n,\lambda_n) = \min_{\mathbf{m}_n} d(\mathbf{m}_n) + \lambda_n\{x_n(\mathbf{m}_n) - y_n\}$. Equations (13) and (15) are derived from independent assumption. Problem (13) is solved by the RD optimization which is implemented in JM model [12] and the dual problem (15) can be solved by the subgradient projection method [6] as follows:

$$\lambda_n^{k+1} = \left[\lambda_n^k + \eta^k g_n^k\right]^+ = \max\left(\lambda_n^k + \eta^k g_n^k, 0\right) \quad (17)$$

where η^k is a positive step size at iteration k and $[\cdot]^+$ denotes the projection onto the nonnegative orthant. The projection operation guarantees that the Lagrange multipliers λ_n satisfy their nonnegative conditions. The subgradient g_n^k of $q_n\left(\lambda_n^k, y_n\right)$ is $x_n^k - y_n$ which is derived in Appendix A. Therefore, the subgradient of $q_n\left(\lambda_n^k, y_n\right)$ is just the difference between coded bits and the constraint bits at iteration k. If we substitute g_n^k with $x_n^k - y_n$ in (17), (17) yields

$$\lambda_n^{k+1} = \max\left(\lambda_n^k + \eta^k\left(x_n^k - y_n\right), 0\right) \quad (18)$$

where x_n^k is the coded bits for a given λ_n^k. Equation (18) indicates that if coded bits x_n^k are smaller than the constraint bits y_n, the current Lagrangian multiplier λ_n^k decreases, otherwise, λ_n^k increases. From the RD optimization, smaller λ increases the coded bits. Therefore, the coded bits are getting close to the constraint bits after several iterations. The reason why we use the subgradient instead of gradient is that the exact RD function is not known and the operational RD function is a convex-hull of the RD function which is not a smooth function. Consequently, the step size η^k is carefully selected since the subgradient direction is not always in increasing direction for any step size [6].

The master primal problem (11) is also solved by the subgradient projection method. However, the constraint is not as simple as in (17). The solution of the problem (11) is obtained from two procedures. First, the optimization variables y_n are updated by the subgradient as follows:

$$\tilde{y}_n^{k+1} = y_n^k - \eta^k g_n^k \quad (19)$$

and then $\tilde{\mathbf{y}}$ is projected onto the feasible constraint set as

$$\min_{\mathbf{y}} \parallel \tilde{\mathbf{y}} - \mathbf{y} \parallel^2, \text{ s.t. } \sum_n y_n \leq X \quad (20)$$

which is formulated from the fact that the projected point $\mathbf{y}$ from $\tilde{\mathbf{y}}$ minimizes the distance between two points. This problem can be solved using a very efficient algorithm discussed in [22]. The subgradient g_n of $q_n^*(y_n)$ is shown in Appendix B. As a result, the subgradient g_n^k of $q_n^*(y_n)$ at y_n^k is $-\lambda_n^k$ where λ_n^k is the optimal dual variable of the sub-problem in (10), that is, the convergent solution of (18) for a given y_n^k after iterations. From Appendix B, the subgradient is the lower bound of sensitivity of the optimal dual function w.r.t. allocated bits y as follows:

$$0 \geq \frac{q^*(\tilde{y}) - q^*(\hat{y})}{\tilde{y} - \hat{y}} \geq -\hat{\lambda}$$

and from (19), $\tilde{y}_n^{k+1} = y_n^k + \eta^k\lambda_n^k$. Consequently, more bits are allocated to MBs or pictures (basic units) which have larger λ since larger λ implies that distortion of a MB or a picture decreases further according to the unit bit increment of the constraint. Fig. 5 shows the geometric interpretation of subgradient of $q^*(y)$. The optimal value $q^*(y)$ can be rewritten as follows for a given optimal dual variable $\hat{\lambda}^*$ (for simplicity, index n is omitted)

$$q^*(\hat{y}, \hat{\lambda}^*) = \min_x d(x) + \hat{\lambda}^*(x - \hat{y}). \quad (21)$$

[5] indicates that the optimal coded bits x^* of the problem (21) satisfies the complementary slackness condition, that is, $\hat{\lambda}^*(x^* - \hat{y}) = 0$. For $\hat{\lambda}^* > 0$, $x^* = \hat{y}$. Consequently, $q^*(\hat{y}, \hat{\lambda}^*) = d(x^*) = d(\hat{y})$. If we allocate more bits from $\hat{y}$ to $\tilde{y}$, x^* changes from $\hat{y}$ to $\tilde{y}$. If we set $\tilde{y} = \hat{y} + \varepsilon$, the sensitivity of the optimal value $q^*(y)$ is

$$\lim_{\varepsilon\to 0} \frac{q^*(\tilde{y}) - q^*(\hat{y})}{\tilde{y} - \hat{y}} = \lim_{\varepsilon\to 0} \frac{d(\hat{y} + \varepsilon) - d(\hat{y})}{\varepsilon} = -\hat{\lambda}^*. \quad (22)$$

Thus, λ indicates that MBs or pictures which have larger λ can reduce more distortion. As a result, if we reallocate bits, sum of distortion can be decreased. Furthermore, all the subgradients of MBs or pictures should be equal for the optimal bit allocation. Because the sensitivity of the optimal values of all the MBs or pictures are equal, there is no way to reallocate bits to decrease sum of distortion. This can be clearly observed from (19) since $\tilde{y}_n^{k+1}$ increase equally if their subgradient g_n^k and step size η^k are equal, and then $\tilde{y}_n^{k+1}$ are projected onto the feasible set shown in (20). The projected y_n^{k+1} are the same as y_n^k which results from [22].

Here, we show experimental results of primal-dual decomposition and subgradient projection with spatially independent assumption for MB-level bit allocation within a intra frame of QCIF size. Without a bit rate constraint, λ is decided from (5) for a given QP and then JM model of H.264 solves the problem (7) for all the MBs in a frame with independent assumption among the MBs. In this case, JM model solves the problem of (8) using a single Lagrange multiplier for all the MBs. With independent assumption, it is optimal since the sensitivity of all the MB (λ) is equal and as explained in Section II, the λ is optimal dual solution by assuming that the coded bits are equal to the constraint bits.

Now, we solve the same problem using the primal-dual decomposition and subgradient methods with the bit constraint. The bit constraint is given from coded bits of JM model after

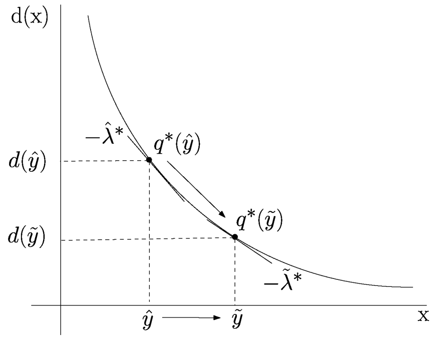

Fig. 5. Geometric interpretation of subgradient of $q^*(y)$.

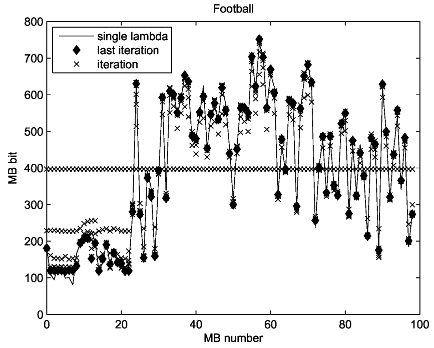

Fig. 6. MB bits of single λ versus multiple λs.

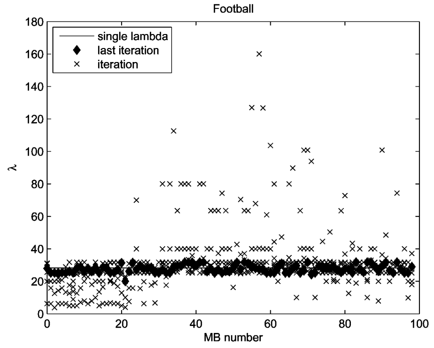

Fig. 7. Single λ versus multiple λs for all the MBs in a frame.

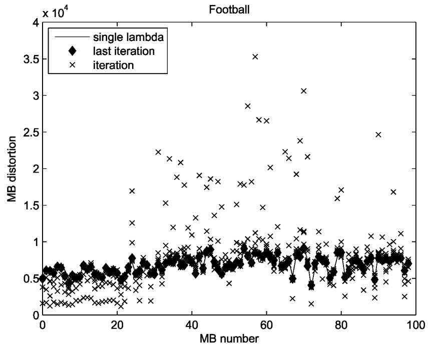

Fig. 8. MB distortion of single λ versus multiple λs.

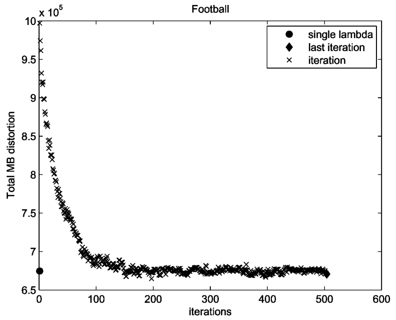

Fig. 9. Overall distortion of single λ versus multiple λs.

solving the problem of (7) (encoding a frame). Thus, we solve the same problem with different methods. Consequently, each MB has its own Lagrange multiplier λ from the problems (10) and (11). Fig. 6 shows that total bits are equally divided into each MB at the initial iteration. As a result, every MB has very different λ and distortion which are shown in Figs. 7 and 8. At the next iteration, more bits are allocated to MBs which have larger λ and on the other hand, fewer bits are allocated to MBs which have smaller λ shown in Figs. 6 and 7. Distortion and λ are smaller along with allocation of more bits. At the last iteration which is marked as a diamond symbol in the figures, MB bits, λ and distortion are almost equal between a single λ and multiple λs since we solve the same problem using different methods. The bits variation of MBs from 0 to 7 and MBs from 57 to 60 is almost equal from the initial to the last iteration (almost 300 bits are different) as shown in Fig. 6. However, Fig. 8 shows that decrease of distortion of MBs from 57 to 60 is much larger than increase of distortion of MBs from 0 to 7 because MBs which have larger λ, reduce their distortion more efficiently. Consequently, sum of MBs distortion decreases after iterations which is shown in Fig. 9, and overall distortion of primal-dual decomposition method is almost equal to a single λ.

IV. Frame-Level Optimization With Temporal Dependency

In the previous section, we assume that all the distortion and coded bit function are independent. In this section, independent assumption among the basic units is removed. Even though we only consider a frame-level optimization problem with temporal dependency among the frames within a GOP, there is no restriction in applying MB-level optimization with spatial dependency. However, we still assume that all the MBs which have spatial prediction dependency are independent for simplicity, but temporal coding dependency is considered among the basic units (frames). With this assumption, a single λ for all the MBs in a frame is optimal which is explained in the previous section.

Equation (8) is reformulated for the frame-level optimization with a GOP bit constraint as follows:

$$\begin{aligned} \min_{\mathbf{s}} \quad & \sum_{f=1}^{F} D_f\left(\mathbf{s_f}, \mathbf{x_{ref}^f}\right) \\ \text{s.t.} \quad & \sum_{f=1}^{F} X_f\left(\mathbf{s_f}, \mathbf{x_{ref}^f}\right) \le X_{\text{GOP}} \end{aligned} \qquad (23)$$

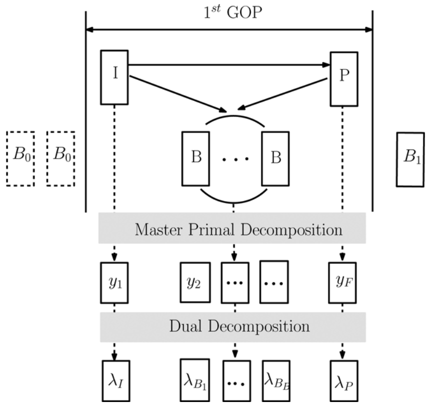

Fig. 10. Mapping between GOP and primal-dual decomposition.

where

$$D_f\left(\mathbf{s_f},\mathbf{x^f_{ref}}\right)=\sum_{n=1}^{N}d_n(\mathbf{m}_n)$$
$$X_f\left(\mathbf{s_f},\mathbf{x^f_{ref}}\right)=\sum_{n=1}^{N}x_n(\mathbf{m}_n)$$

and $\mathbf{s_f}=(\mathbf{M_f},\mathbf{MV_f},\mathbf{QP_f})$ where $\mathbf{M_f}$, $\mathbf{MV_f}$, $\mathbf{QP_f}$, and $\mathbf{x}^{\mathbf{f}}_{\mathrm{ref}}$ are MB modes, MVs, QPs, and bits of reference frames for all the MBs in a frame f, X_{GOP} is a GOP bit constraint and F is the number of frames within a GOP. Distortion D_f and bits X_f of a frame f depend on all the MB modes, MVs and QPs, as well as bits of reference frames. Therefore, every frame can not be optimized independently in the problem (23) due to the dependency of bits of reference frames $\mathbf{x}^{\mathbf{f}}_{\mathrm{ref}}$.

Fig. 10 illustrates mapping between a GOP structure and primal-dual decomposition. As a specific example, we only consider the first GOP structure which starts with the instantaneous decoder refresh (IDR) frame (first B frames which are denoted as B_0 are included from the second GOP) and the close GOP which is that the last P frame within a GOP is not used for a prediction of B_1. However, the open GOP and any number and prediction dependency of B and P frames within a GOP are not limited to the frame-level optimization with dependency. Here, all the B frames within a GOP are predicted from the same reference frames I and P, and the P frame is predicted from the I frame. When we perform the master primal decomposition, dependency among the frames are considered. As in Section III, slack variables y_f are introduced for each frame bits. Consequently, the problem in (23) is decomposed into one master primal problem (25) and F sub-problems (24) which are solved by the Lagrangian duality

$$\begin{aligned}&\min_{\mathbf{s}} && D_f\left(\mathbf{s_f},\mathbf{y^f_{ref}}\right) && (24)\\ &\text{s.t.} && X_f\left(\mathbf{s_f},\mathbf{y^f_{ref}}\right)\le y_f,\quad f\in\{1,\ldots,F\}\\ &\min_{\mathbf{y}} && \sum_f Q^*_f\left(y_k,\mathbf{y^f_{ref}}\right)\\ &\text{s.t.} && \sum_f y_f\le X_{\mathrm{GOP}},\quad f\in\{1,\ldots,F\} && (25)\end{aligned}$$

where $Q^*_f\left(y_f,\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}\right)$ are the optimal values of sub-problems (24) for a given $\mathbf{y}$, and the reference frame bits are $\mathbf{y}^{\mathbf{w}}_{\mathrm{ref}}=(y_1,y_F)$ where $w\in\{2,\ldots,F-1\}$ for B frames, $\mathbf{y}^{\mathbf{F}}_{\mathrm{ref}}=y_1$ for the P frame and $\mathbf{y}^{\mathbf{1}}_{\mathrm{ref}}=\varnothing$ since the I frame has no reference frames. Comparing (23) with (24), we can recognize the main benefit from the primal decomposition. In the formulation (23), the reference frame bits $\mathbf{x}^{\mathbf{f}}_{\mathrm{ref}}$ prevents independent optimization, but in the formulation (24), $\mathbf{x}^{\mathbf{f}}_{\mathrm{ref}}=\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}$ because given y_f, $X_f(\mathbf{s_f}^*)$ are y_f due to the complementary slackness condition [5]. Independent optimization is clearer from the following equations:

$$\begin{aligned}\min_{\mathbf{s}}&\sum_f D_f\left(\mathbf{s_f},\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}\right)-\lambda_f\left(X_f\left(\mathbf{s_f},\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}\right)-y_f\right)\\ &=\sum_f\min_{\mathbf{s_f}}D_f\left(\mathbf{s_f},\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}\right)-\lambda_f\left(X_f\left(\mathbf{s_f},\mathbf{y}^{\mathbf{f}}_{\mathrm{ref}}\right)-y_f\right).\end{aligned}$$

Therefore, we reuse the same reference software model of H.264 to minimize the Lagrangian cost in problem (24) with the consideration of dependency. Reference [15] also solves the frame-level dependent coding problem using the VA, but it considers that distortion and coded bits are only functions of QP. Therefore, the problem can be trackable, but if distortion and coded bits are a function of QP as well as MB modes, MVs and λ, the number of states of the VA are too large which is not trackable. In this paper, we do not separately consider effects of all the parameters but are interested in allocating bits among the frames. Given a bit constraint, all the parameters are only optimization variables to satisfy the bit constraint. Especially, temporal dependency among the frames only depends on the prediction quality. Thus, the dependency among the frames is processed in the master primal problem as shown in (25) which is a much simpler optimization problem.

In order to solve the problem in (24), we use the Lagrangian duality and subgradient projection which are explained in the Section III. Therefore, we only discuss the master primal problem (25) in this section. From Appendix C, the subgradients of $\sum_{f=1}^{F}Q^*_f\left(y_f,y^f_{\mathrm{ref}}\right)$ w.r.t. B, P, and I pictures y_f at $\hat{y}_f$ are

$$\begin{aligned}&-\hat{\lambda}_k,\quad k\in\{2,\ldots,F-1\},\\ &-\left(\hat{\lambda}_F-\sum_{f=2}^{F-1}\min_{\mathbf{s_f}}L'_f(\mathbf{s_f},\hat{y}_F)\right)\\ \text{and }&-\left(\hat{\lambda}_1-\sum_{f=2}^{F}\min_{\mathbf{s_f}}L'_f(\mathbf{s_f},\hat{y}_1)\right)\end{aligned}$$

where $L'_f\left(\mathbf{s_f},\hat{y}^f_{\mathrm{ref}_i}\right)=\left(\partial L_f\left(\mathbf{s_f},y^f_{\mathrm{ref}_i}\right)/\partial y^f_{\mathrm{ref}_i}\right)\Big|_{y^f_{\mathrm{ref}_i}=\hat{y}^f_{\mathrm{ref}_i}}$, $\hat{y}^f_{\mathrm{ref}}\in\{\hat{y}_F,\hat{y}_1\}$ which shows that the variation of Lagrangian cost of a frame f w.r.t. the bit variation of its reference frame. Thus, $L'_f\left(\mathbf{s_f},\hat{y}^f_{\mathrm{ref}}\right)$ is generally negative because increasing of reference frame bits induces decreasing of the Lagrangian cost of the frame f. As a result, the subgradients of referenced frames which are used as reference frames for prediction are smaller than independent frames for given equal λs. It means that more bits are allocated to referenced frames from (19). This result

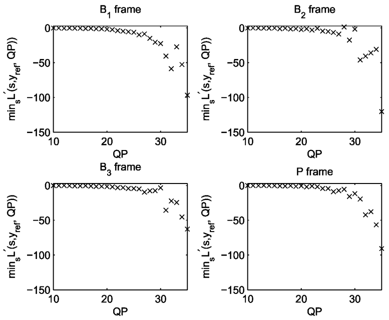

Fig. 11. $\min_{\mathbf{s_f}} L'_f\left(\mathbf{s_f}, \hat{y}^f_{\text{ref}}\right)$ of three B pictures and P picture at $\hat{\lambda}(QP)$.

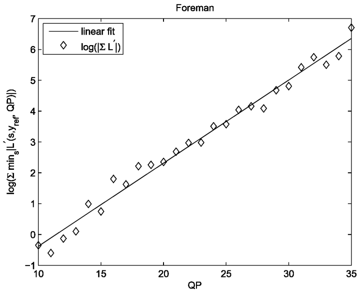

Fig. 12. $\log\left|\sum_f \min_{\mathbf{s_f}} L'_f\left(\mathbf{s_f}, \hat{y}^f_{\text{ref}}\right)\right|$ and its linear fit at $\hat{\lambda}(QP)$.

matches with intuition, that is, referenced frames are more important than nonreferenced frames because they are used for prediction. As explained in Section III, the subgradients of all the frames are equal for the optimal bit allocation. Therefore, the relation among the λ of pictures is derived as follows:

$$\lambda_I - \sum_{f=2}^{F} \min_{\mathbf{s_f}} L'_f(\mathbf{s_f}, y_I) = \lambda_P - \sum_{f=2}^{F-1} \min_{\mathbf{s_f}} L'_f(\mathbf{s_f}, y_P) = \lambda_B \tag{26}$$

where $\lambda_I = \lambda_1$, $\lambda_P = \lambda_F$ and $\lambda_B = \lambda_k, k \in \{2, \ldots, F-1\}$. Consequently, $\lambda_I \leq \lambda_P \leq \lambda_B$. This result explains the reason why JM model uses different κ of λ in (5) for different picture types as well as the number of prediction dependency. Furthermore, if B frames are used for prediction, the referenced B frames have different λs from nonreferenced B frames. However, current JM model [12] uses the same default value κ for I and P pictures.

The remain problem is how to estimate the quantity of $\min_{\mathbf{s_f}} L'_f\left(\mathbf{s_f}, \hat{y}^f_{\text{ref}}\right)$. We experimentally estimate it as shown in Figs. 11 and 12. Fig. 11 represents the Lagrangian cost variation of three B pictures and one P picture according to the variation of I reference frame bits for a given $\hat{\lambda}$. Even though the Lagrangian cost does not monotonically decrease, it mainly exponentially decreases and sum of the variation of the Lagrangian cost is closer to exponential decrement which is shown in Fig. 12. Exponential decrement explains that bits of reference frames (quality of reference frames) are more important at low bit rate, that is, at large λ. As a result, $\sum_f \min_{\mathbf{s_f}} L'_f\left(\mathbf{s_f}, \hat{y}^f_{\text{ref}}\right)$ is modeled as $-\exp^{(\alpha QP + \beta)}$ where α and β are constants.

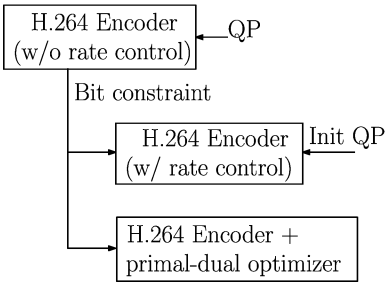

Fig. 13. Experimental environment.

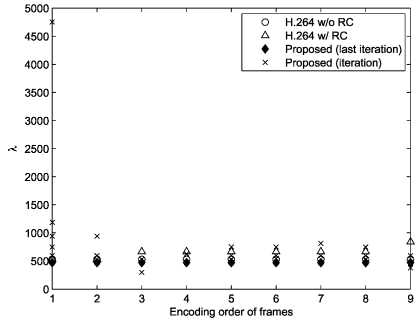

Fig. 14. λ of frames with independent assumption.

V. Experimental Results

In this section, we compare the performance of two coding modes of JM model [12] with the proposed method (H.264 Encoder + primal-dual optimizer) which is illustrated in Fig. 13. We set $QP = 35$ in JM model without Rate Control (RC) and then the coded bits after encoding are used for the constraint bits. Initial QP is set 35 to JM model with RC and proposed encoder as shown in Fig. 13. In JM model with RC, the initial QP is used to code I and P pictures in first GOP such that RC algorithm is only applied for B pictures. However, proposed encoder ignores initial QP.

1) Experiment 1: We assume all the frames are independent. Therefore, a global λ for all the frames within a GOP induces an optimal bit allocation since all the frames have equal λ. In order to compare the performance, we set equal κ of (5) of JM model for all the picture types according to independent assumption. A GOP consists of one I and P pictures and seven B pictures with the GOP structure of Fig. 10. Fig. 14 shows that in JM model without RC, every frame has equal λ and the proposed method also has equal λ after iterations, but JM model with RC has different λ, especially at the last B frame. Figs. 15 and 16 illustrate coded bits of each frame and its Y-PSNR(dB), respectively. JM model without RC and the proposed method show almost the same performance since they solve the same problem. However, JM model with RC predicts encoded bits and QP to satisfy bit constraints. Its dual variable λ is derived from (5). Consequently, it has some performance degradation. In order to compare exact performance, the proposed method does not perform QP optimization to minimize the Lagrangian cost. QP is derived from λ using (5). Thus, the proposed method uses the identical optimization variables (MVs, MB modes) which are the same in JM model without RC. The proposed method derives the similar results of JM model without RC. However,

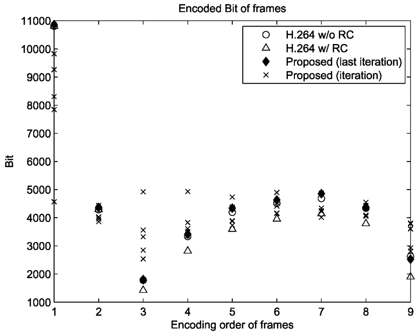

Fig. 15. Encoded frame bits with independent assumption.

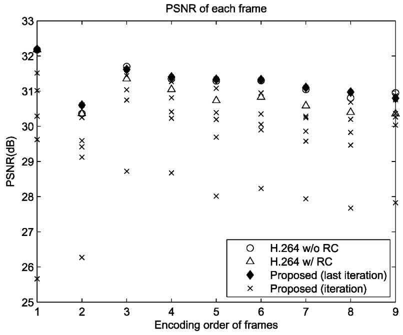

Fig. 16. Y-PSNR(dB) of frames with independent assumption.

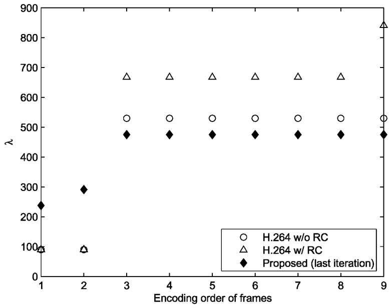

Fig. 17. λ of frames with temporal dependency.

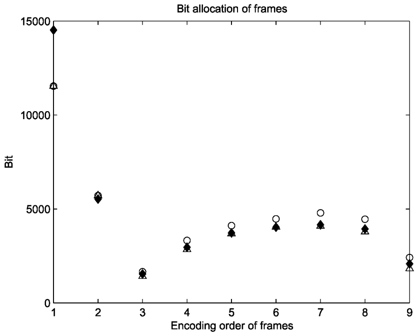

Fig. 18. Encoded frame bits with temporal dependency.

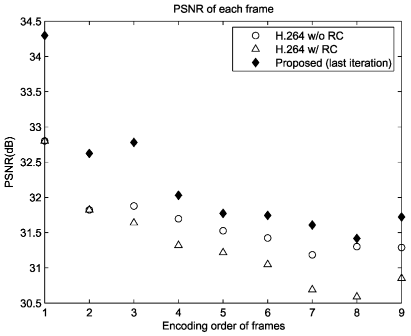

Fig. 19. Y-PSNR(dB) of frames with temporal dependency.

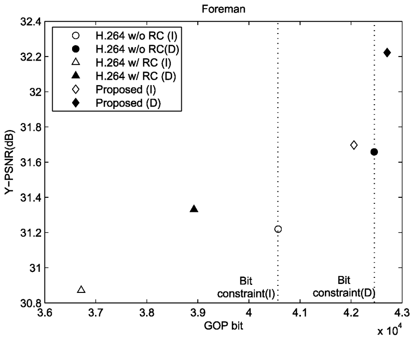

Fig. 20. PSNR versus bit with dependent and independent cases.

Fig. 16 shows that proposed method has higher PSNR than JM model with RC.

2) Experiment 2: In this experiment, we consider temporal dependency among the frames within a GOP. Therefore, κ of (5) of JM model is set to have different values [12]. Proposed method performs QP optimization within ± 1 at center QP which is derived from λ which gives more achievable bit region. Here, we only show the last iteration. Fig. 17 shows the different λs among I, P, and B pictures. All the B pictures have the same λ except JM model with RC. Thus, bit allocation of JM model with RC for B frames is not optimal. Proposed method uses different λs for I and P pictures as a result of (26), but JM model without RC uses the same λ for I and P pictures. Figs. 18 and 19 show that proposed method uses similar total bits and achieves higher PSNR than JM model with RC. Fig. 20 illustrates the overall encoded bits and Y-PSNR(dB). The bits of JM model without RC are constraints to JM model with RC and the proposed method. The dependent constraint bits (coded bits of JM model without RC) increase since λs of I and P pictures become smaller. Due to RD optimization, smaller λ increases coded bits. In the independent experiment, proposed method allows a constraint violation within 5% of allocated bits for I picture and 2% for P and B pictures. In the dependent case, 2%, 1%, and 1% constraint violations are allowed for I, P, and B pictures, respectively. Therefore, the dependent experiment meets more tightly bit constraint. Small bit constraint violation is allowed to consider the convex-hull point around the constraint. JM model with RC does not satisfy the constraint well in addition to having lower PSNR. These experiments are applied to various sequences which have different motion activity and different resolutions. Tables I and II show that proposed method achieves similar coded bits and PSNR with JM model without RC which is optimal results with independent assumption. JM model with RC has larger loss at high motion sequences such as Football and Bus and high resolution sequences of Paris and Tempete. This phenomena is similar to results with temporal

TABLE I
CODED BITS AND THEIR DIFFERENCE RATIO (%)
WITH INDEPENDENT ASSUMPTION

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	125374	-11.45	-1.08
Bus	QCIF	81569	-11.67	-1.90
Foreman	QCIF	40572	-9.51	+7.37
Mobile	QCIF	51493	-5.05	+0.69
News	QCIF	20598	-3.56	+1.11
Carphone	QCIF	20028	-6.38	+3.76
Paris	CIF	120936	-7.37	+1.54
Tempete	CIF	125409	-9.57	+0.84

TABLE II
PSNR (DECIBELS) AND THEIR DIFFERENCE WITH INDEPENDENT ASSUMPTION

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	28.26	-0.55	-0.07
Bus	QCIF	28.29	-0.40	+0.04
Foreman	QCIF	31.22	-0.35	+0.57
Mobile	QCIF	27.43	-0.11	+0.01
News	QCIF	31.37	-0.06	+0.07
Carphone	QCIF	31.80	-0.15	+0.22
Paris	CIF	30.08	-0.17	+0.06
Tempete	CIF	28.73	-0.16	0.00

TABLE III
CODED BITS AND THEIR DIFFERENCE RATIO (%)
WITH TEMPORAL DEPENDENCY

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	129448	-10.49	+1.08
Bus	QCIF	82603	-13.99	-3.54
Foreman	QCIF	42453	-8.32	+0.60
Mobile	QCIF	56008	-4.06	-0.80
News	QCIF	22583	-2.44	+0.58
Carphone	QCIF	22004	-9.39	-0.49
Paris	CIF	128667	-5.69	+0.04
Tempete	CIF	139088	-6.89	-1.17

TABLE IV
PSNR (DECIBELS) AND THEIR DIFFERENCE WITH TEMPORAL DEPENDENCY

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	28.54	-0.53	+0.12
Bus	QCIF	28.78	-0.37	+0.27
Foreman	QCIF	31.66	-0.33	+0.56
Mobile	QCIF	28.29	-0.12	+0.14
News	QCIF	32.36	-0.06	+0.24
Carphone	QCIF	32.69	-0.16	+0.13
Paris	CIF	30.76	-0.16	+0.41
Tempete	CIF	29.56	-0.14	+0.10

TABLE V
CODED BITS AND THEIR DIFFERENCE RATIO (%)
WITH TEMPORAL DEPENDENCY

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	199320	+0.14	+0.34
Bus	QCIF	104637	+1.87	+0.76
Foreman	QCIF	46374	-0.74	+0.23
Mobile	QCIF	87382	+2.66	-0.20
News	QCIF	25580	+6.11	+0.23
Carphone	QCIF	26308	+9.19	-0.71
Paris	CIF	164918	+4.27	+0.64
Tempete	CIF	230810	+1.71	+0.18

TABLE VI
PSNR (DECIBELS) AND THEIR DIFFERENCE WITH TEMPORAL DEPENDENCY

Sequence	Resolution	H.264 w/o RC	H.264 w/ RC	Propose
Football	QCIF	28.87	-0.03	+0.08
Bus	QCIF	28.74	+0.05	+0.09
Foreman	QCIF	31.68	-0.16	+0.09
Mobile	QCIF	27.39	+0.24	+0.10
News	QCIF	31.89	+0.19	+0.28
Carphone	QCIF	32.40	+0.37	+0.00
Paris	CIF	30.51	+0.15	+0.01
Tempete	CIF	29.27	+0.14	+0.11

dependency which are shown in Tables III and IV. The proposed method achieves higher PSNR within the bit violation constraint. The different GOP structure which consists of one I and 15 P pictures is tested and the results are presented in Tables V and VI. JM model with RC achieves better results in I and P GOP than in I, B, and P GOP structure since reference frames are close to the current frame. However, the proposed method consistently shows good results.

VI. CONCLUSION

In this paper, we propose a general framework to solve a RD optimization problem by using the primal-dual decomposition and subgradient projection methods. As a result of primal decomposition, we can use the same reference software model of H.264 to solve the sub-optimization problems with consideration of prediction dependency. Furthermore, optimal bit allocation condition is derived under prediction dependency. Experimental results show that the proposed method is promising in compensating loss of coding gain for nonreal time application with the constant bit rate. However, the complexity of encoder increases proportionally to the number of iterations and the number of iterations highly depend on the initial values. Therefore, JM model with RC can cooperate with the proposed method for initial values of iteration which reduces the number of iterations. This work will be considered as future work. In this paper, we derive the optimal bit allocation condition but need further research how to estimate variation of the Lagrangian cost according to the reference bits. As a result, we can adaptively change weight of λ to both JM model without RC and JM model with RC. It will increase coding efficiency without consideration of iteration.

APPENDIX A

Subgradient of $q_n(\lambda_n, y_n)$*:* It is reformulated from [6] for this paper's notation

$$\begin{aligned} q_n(\tilde{\lambda}_n, y_n) &= \min_{x_n}\{d(x_n) + \tilde{\lambda}_n(x_n - y_n)\} \\ &\le d(\hat{x}_n) + \tilde{\lambda}_n(\hat{x}_n - y_n) \\ &= d(\hat{x}_n) + \hat{\lambda}_n(\hat{x}_n - y_n) + (\hat{x}_n - y_n)(\tilde{\lambda}_n - \hat{\lambda}_n) \\ &= q_n(\hat{\lambda}_n, y_n) + (\hat{x}_n - y_n)(\tilde{\lambda}_n - \hat{\lambda}_n), \quad \forall\, \tilde{\lambda}_n \end{aligned}$$

where $\hat{x}_n = \arg\min_{x_n}\{d(x_n) + \hat{\lambda}_n(x_n - y_n)\}$. Thus, the subgradient of $q_n(\lambda_n, y_n)$ at $\hat{\lambda}_n$ is $\hat{x}_n - y_n$.

APPENDIX B

Subgradient of $q_n^*(y_n)$*:* It is reformulated from [6] for this paper's notation

$$\begin{aligned} q_n^*(\tilde{y}_n) &= \max_{\lambda_n \ge 0} q_n(\lambda_n, \tilde{y}_n) \\ &= \max_{\lambda_n \ge 0} \min_{x_n}\{d(x_n) + \lambda_n(x_n - \tilde{y}_n)\} \\ &\ge \min_{x_n}\{d(x_n) + \hat{\lambda}_n(x_n - \tilde{y}_n)\} \\ &= \min_{x_n}\{d(x_n) + \hat{\lambda}_n(x_n - \hat{y}_n) - \hat{\lambda}_n(\tilde{y}_n - \hat{y}_n)\} \\ &= q_n^*(\hat{y}_n) - \hat{\lambda}_n(\tilde{y}_n - \hat{y}_n), \forall\, \tilde{y}_n \end{aligned}$$

where $\hat{\lambda}_n = \arg\max_{\lambda_n}\min_{x_n}\{d(x_n) + \lambda_n(x_n - \hat{y}_n)\}$. The subgradient of optimal dual function $q_n^*(y)$ at $\hat{y}_n$ is $-\hat{\lambda}_n$.

APPENDIX C

Subgradient of $\sum_f q_f^*(y_f, y_{\text{ref}}^f)$*:* (see the equation shown at the bottom of the page) where $L_f\left(\mathbf{s_f}, \mathbf{y_{ref}^f}\right) = D_f\left(\mathbf{s_f}, \mathbf{y_{ref}^f}\right) + \hat{\lambda}_f X_f\left(\mathbf{s_f} \mathbf{y_{ref}^f}\right)$, $\mathbf{y_{ref}^f} \in \left\{\hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}} \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right\}$ and $L'_f\left(\mathbf{s_f}, \hat{y}_{\text{ref}_i}^f\right) = \left(\partial L_f\left(\mathbf{s_f}, y_{\text{ref}_i}^f\right) / \partial y_{\text{ref}_i}^f\right)\Big|_{y_{\text{ref}_i}^f = \hat{y}_{\text{ref}_i}^f}$ and $\hat{\lambda}_f = \arg\max_{\lambda_f \ge 0}\min_{\mathbf{s_f}}\left\{D\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) + \lambda_f\left(X_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{y}_f\right)\right\}$ and R is the number of reference frames. In this paper, bits of the reference frames of B picture $\mathbf{y_{ref}^f}$ are y_1 and y_F and the reference frame bit of P picture is y_1 from Fig. 10. Consequently, the following equation is derived:

$$\begin{aligned} &\sum_{f=1}^{F} Q_f^*\left(\tilde{y}_f, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) \\ &\quad \ge \sum_{f=1}^{F} Q_f^*\left(\hat{y}_f, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \left(\hat{\lambda}_F - \sum_{f=2}^{F-1} \min_{\mathbf{s_f}} L'_f(\mathbf{s_f}, \hat{y}_F)\right) \\ &\quad \times (\tilde{y}_F - \hat{y}_F) \\ &\quad - \sum_{f=2}^{F-1} \hat{\lambda}_f(\tilde{y}_f - \hat{y}_f) - \left(\hat{\lambda}_1 - \sum_{f=2}^{F} \min_{\mathbf{s_f}} L'_f(\mathbf{s_f}, \hat{y}_1)\right) \\ &\quad \times (\tilde{y}_1 - \hat{y}_1). \end{aligned}$$

$$\begin{aligned} \sum_f Q_f^*\left(\tilde{y}_f, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) &= \sum_f \max_{\lambda_f \ge 0} \min_{\mathbf{s_f}} \left\{D_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) + \lambda_f\left(X_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \tilde{y}_f\right)\right\} \\ &\ge \sum_f \min_{\mathbf{s_f}} \left\{D_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) + \hat{\lambda}_f\left(X_f\left(\mathbf{s_k}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \tilde{y}\right)\right\} \\ &= \sum_f \min_{\mathbf{s_f}} \Bigg\{D_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) + \hat{\lambda}_f\left(X_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{y}_f\right) \\ &\qquad - \hat{\lambda}_f\left(\tilde{y}_f - \hat{y}_f\right) + D_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - D_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) \\ &\qquad + \hat{\lambda}_f\left(X_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - X_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right)\right)\Bigg\} \\ &= \sum_f \Bigg\{Q_f^*\left(\hat{y}_f, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{\lambda}_f\left(\tilde{y}_f - \hat{y}_f\right) \\ &\qquad + \min_{\mathbf{s_f}}\left\{D_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - D_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) + \hat{\lambda}_f\left(X_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - X_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right)\right)\right\}\Bigg\} \\ &= \sum_f \left\{Q_f^*\left(\hat{y}_f, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{\lambda}_f\left(\tilde{y}_f - \hat{y}_f\right) + \min_{\mathbf{s_f}}\left\{L_f\left(\mathbf{s_f}, \tilde{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - L_f\left(\mathbf{s_f}, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right)\right\}\right\} \\ &\ge \sum_f \left\{Q_f^*\left(\hat{y}_f, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{\lambda}_f\left(\tilde{y}_f - \hat{y}_f\right) + \min_{\mathbf{s_f}} \sum_i^R L'_f\left(\mathbf{s_f}, \hat{y}_{\text{ref}_i}^f\right)\left(\tilde{y}_{\text{ref}_i}^f - \hat{y}_{\text{ref}_i}^f\right)\right\} \\ &\ge \sum_f \left\{Q_f^*\left(\hat{y}_f, \hat{\mathbf{y}}_{\mathbf{ref}}^{\mathbf{f}}\right) - \hat{\lambda}_f\left(\tilde{y}_f - \hat{y}_f\right) + \sum_i^R \min_{\mathbf{s_f}} L'_f\left(\mathbf{s_f}, \hat{y}_{\text{ref}_i}^f\right)\left(\tilde{y}_{\text{ref}_i}^f - \hat{y}_{\text{ref}_i}^f\right)\right\} \end{aligned}$$

REFERENCES

[1] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression: An overview," *IEEE Signal Process. Mag.*, vol. 15, no. 11, pp. 23–50, Nov. 1998.

[2] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 11, pp. 74–99, Nov. 1998.

[3] Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, 2003.

[4] A. Puria, X. Chenb, and A. Luthrac, "Video coding using the h.264/ mpeg-4 avc compression standard," *Signal Process.: Image Commun.*, vol. 19, pp. 793–849, Oct. 2004.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[6] D. P. Bertsekar, *Nonlinear Programming*, 2nd ed. New York: Athena Scientific, 2003.

[7] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," presented at the IEEE Int. Conf. Image Proceesing, 2001.

[8] K. Takagi, Lagrange Multiplier and rd-Characteristics, JVT-C084, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, 2002.

[9] T. Weigand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.

[10] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 246–250, Feb. 1997.

[11] H. Kim, "Adaptive rate control using nonlinear regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 5, pp. 432–439, May 2003.

[12] H.264/AVC Reference Software (JM11.0) HHI [Online]. Available: http://iphome.hhi.de/suehring/tml/download/

[13] T. Wiegand, M. Lightstone, T. G. Campbell, and S. k. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging h.263 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, no. 4, pp. 182–190, Apr. 1996.

[14] Y. Yang and S. S. Hemami, "Generalized rate-distortion optimization for motion-compensatedvideo coders," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 9, pp. 942–955, Sep. 2000.

[15] 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. 9, pp. 533–545, Sep. 1994.

[16] Z. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, Adaptive Basic Unit Layer Rate Control for JVT, JVT-G012-r1, ISO/IEC JTC1/ SC29/WG11 and ITU-T SG16 Q.6, 2003.

[17] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu, and Y. Lu, "Adaptive rate control for h.264," *Vis. Commun. Image Represent.*, vol. 17, pp. 376–406, Apr. 2006.

[18] H. Lee, T. Chiang, and Y. Zhang, "Scalalble rate control for mpeg-4 video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 9, pp. 878–894, Sep. 2000.

[19] Z. Li, C. Zhu, N. Ling, X. Yang, G. Feng, S. Wu, and F. Pan, "A unified architecture for real-time video-coding systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 472–487, Jun. 2003.

[20] D. Palomar and M. Chiang, *Alternative Distributed Algorithms for Network Utility Maximization: Framework and Applications*, 2007, to be published.

[21] B. Johansson and M. Johansson, "Primal and dual approaches to distributed cross-layer optimization," presented at the 16th IFAC World Congr., Prague, Czech Republic, 2005.

[22] D. Palomar, "Convex primal decomposition for multicarrier linear mimo transceivers," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4661–4674, Dec. 2005.

Cheolhong An (S'07) received the B.S. and M.S. degrees in electrical engineering from Pusan National University, Busan, Korea, in 1996 and 1998, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering at the University of California at San Diego, La Jolla.

His research interests include resource allocation for video communication and statistical learning and optimization-based video coding.

Truong Q. Nguyen (F'06) is currently a Professor at the Electrical and Computer Engineering Department, University of California at San Diego, La Jolla. His research interests are video processing algorithms and their efficient implementation. He is the coauthor (with Prof. G. Strang) of the textbook *Wavelets and Filter Banks* (Wellesley-Cambridge, 1997), and author of several Matlab-based toolboxes on image compression, electrocardiogram compression, and filter bank design. He has authored over 200 publications.

Prof. Nguyen received the IEEE TRANSACTIONS ON SIGNAL PROCESSING Paper Award (Image and Multidimensional Processing area) for the paper he co-wrote with Prof. P. P. Vaidyanathan on linear-phase perfect-reconstruction filter banks (1992). He received the National Science Foundation Career Award in 1995 and is currently the Series Editor (Digital Signal Processing) for Academic Press. He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1994 to 1996, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1996 to 1997, and the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2004 to 2005.