An Efficient Bit Allocation Algorithm in Dependent Coding Framework and One-way Video Applications

Seong Hwan Jang and Nikil Jayant

Multimedia Communication Lab.
Georgia Institute of Technology
{shjang, jayant}@ece.gatech.edu

Abstract- In this paper, we describe an efficient bit allocation algorithm using two-dimensional interpolation of rate-distortion (R-D) characteristics in a dependent coding environment for applications of one-way video. A lot of R-D based bit allocation approaches have been proposed to aim at the best possible video quality for given bandwidth and buffer constraints. Those approaches require measuring of actual R-D characteristics of the input data before making quantization decisions. However, the solution of the R-D approaches becomes exponentially complex as the dependency tree depths progress. To solve the dependent coding problem, we need to measure R-D data for all possible combination of quantization settings in the reference frame and the predictive frame. Two-dimensional interpolation of rate-distortion characteristics is performed with only a few R-D control points to reduce complexity without significant loss of performance. We also explain a pruning algorithm to further reduce the computational complexity in obtaining R-D characteristics. The cost function for each quantization pair is obtained using two dimensional R-D characteristics. The complexity is reduced by the 2D-interpolation and pruning algorithm without significant loss of optimality. Experiments with MPEG-2 show that the proposed algorithm provides improvement of PSNR by 0.9–1.8dB over the MPEG-2 TMS algorithm.

I. INTRODUCTION

Recently, one-way digital video services using pre-encoded video bit streams are becoming an increasingly important part of visual services and are widely available via internet or wireless channels. However, due to bandwidth and buffer constraints, compressed video quality is still inadequate. Bit allocation and quantization are of key importance to compression and visual quality. The problem of bit allocation is the efficient distribution of the given total bit budget to coding units, with a set of admissible quantization choices to minimize distortion [1]-[4].

Many efforts for optimal bit allocation have been proposed to increase video quality. Model based approaches have been used to deal with the bit allocation problem [6]. Many kinds of rate-quantization models have been introduced with appropriate coding parameters to reduce the complexity and avoid the need to measure the R-D data on the possible quantization settings [7]. However, the accuracy of the model depends on many parameters such as various image contents and bit rates. Therefore, the errors in these models can be still large according to image content especially in P and B pictures. Another approach is the operational R-D based framework [1]-[4]. In practice, classical rate-distortion theory cannot be directly applicable to a complex encoding system, since video sources are typically not well-characterized and the rate-distortion function \( R(D) \) is difficult, if not impossible, to determine. An operational R-D plot can be constructed by measuring the actual rate and distortion achieved by a specific encoder, since practical coding systems resort to a finite set of admissible quantization. The operational R-D framework can be used by the optimal bit allocation algorithm to minimize overall coding distortion subject to a total bit budget constraints [1]. In most of the research addressed in the literature, the input signal units have been considered to be coded independently. However, all standard video coding schemes involve a dependent coding framework such as DPCM and motion compensation. In the dependent coding framework, the set of available R-D operating points for some coding units depends on the particular choice of the R-D point for a previous coding unit [4]. The optimal solution for the dependent coding problem becomes exponentially complex, since the number of calculations for the set of quantization choices is unrealistic. A pruning algorithm using the monotonic property of dependent R-D curves has been proposed to reduce the complexity and obtain near-optimal solution [4]. However, the pruning algorithm is still complex, since the size of the quantization set is 31 in typical video standards. An interpolation of R-D characteristics was introduced to reduce the complexity without significantly reducing the optimization process [5]. However, a model for rate and distortion is still used in the P picture and the B picture. In this paper, we introduce a bit allocation algorithm using a two dimensional interpolation of R-D characteristic. In the two-dimensional R-D framework, one axis stands for quantization set of the reference frame. (e.g. Intra picture) The other axis stands for quantization set of the dependent frame. (e.g. Predictive picture) The interpolation is performed not only in the direction of reference frame but also in the direction of dependent frame. A two dimensional cost function is obtained using those two dimensional R-D characteristics. The proposed algorithm is suitable for one-way video application, since much more coding delay is allowed in one-way video. The remainder of this paper is organized as follows. In Sections II, III and IV, we describe the problem formulation in the dependent coding framework, 2-D interpolation of R-D characteristics, and the pruning algorithm, respectively. In Sections V and VI, we discuss experimental results and conclusions.

II. PROBLEM FORMULATION

In the budget-constrained bit allocation problem, quantization is the main coding parameter to be adjusted. The problem is minimizing overall distortion \( D \), subject to a bit budget constraint \( B \) on the number of used bits. In typical video coding standards, the quantization step size is determined for each macroblock. While our experiments are
based on the MPEG coder, the proposed algorithm is general enough to be applied to other video coding standards such as H.263 and H.264. Shoham et al. gave an efficient bit allocation algorithm based on the Lagrange-multiplier method [1]. In this method, the distortion term is weighted against a rate term, and the constrained problem is transformed to the following unconstrained problem:

**Problem 1.** Given a set of quantizers \( \{q_1, q_2, \ldots, q_M\} \), a sequence of coding units \( < x_1, x_2, \ldots, x_N > \), and a parameter \( \lambda \), determine an assignment of quantizers \( \mathcal{Q} = \{Q_1, Q_2, \ldots, Q_N\} \) to each coding unit that minimizes the cost function

\[
J(\mathcal{Q}) = \mathcal{D}(\mathcal{Q}) + \lambda R(\mathcal{Q}).
\]

In the MPEG encoder, one out of 31 possible quantization step size values is assigned to each macroblock. The bit allocation problem can be classified into two stages: frame level bit allocation for temporal bit allocation and macroblock level bit allocation for spatial bit allocation. Thus the coding units \( < x_1, x_2, \ldots, x_N > \) can be either frame or macroblock.

In the dependent coding framework, the bit allocation problem can be expressed as follows:

**Problem 2.** Determine an assignment of quantizers \( \mathcal{Q} = \{Q_1, Q_2, \ldots, Q_N\} \) that minimizes the cost function,

\[
\min_{\mathcal{Q}} \left\{ J_1(Q_1) + J_2(Q_2) + \ldots + J_N(Q_N) \right\}
\]

where \( J_i(Q_1, Q_2, \ldots, Q_N) = D_i(Q_1, Q_2, \ldots, Q_N) + \lambda R_i(Q_1, Q_2, \ldots, Q_N) \).

The optimal solution for the dependent problem becomes exponentially complex as the dependency tree depth progresses to the \( N \)-th stage. The number of stages is \( N^m \) to solve the problem. Since the number of calculations is unrealistic, we need a pruning algorithm to simplify the complexity. By using pruning algorithm, we can reduce the number of trellis nodes and the number of reference frames to be traced.

### III. 2-D INTERPOLATION OF R-D CHARACTERISTICS

An approximation of R-D characteristics is required to reduce complexity. However, modeling the R-D characteristics shows large errors since the parameters can be changed according to the input images. A curve fitting can be used to increase the accuracy of the models using only a small set of quantization step sizes called “control points”. Curve fitting is the art and science of creating a function definition that passes through a set of control points. We can interpolate R-D characteristics from the function obtained by curve fitting. Many kinds of curve fitting methods have been introduced. We used cubic spline interpolation method since it shows good performance compared to other methods [5].

The cubic spline interpolation is a piecewise continuous curve fitting. Many kinds of curve fitting methods have been introduced. We used cubic spline interpolation method since it shows good performance compared to other methods [5].

Many kinds of curve fitting methods have been introduced. We used cubic spline interpolation method since it shows good performance compared to other methods [5].

The cubic spline interpolation is a piecewise continuous curve fitting. Many kinds of curve fitting methods have been introduced. We used cubic spline interpolation method since it shows good performance compared to other methods [5].

In this method, the distortion term is weighted against a rate term, and the constrained problem is transformed to the following unconstrained problem:

\[
J(\mathcal{Q}) = \mathcal{D}(\mathcal{Q}) + \lambda R(\mathcal{Q}).
\]

The greater the number of control points, the more accurate is the approximation result. Fig 1 shows an experimental plot for MSE and quantization step size for the Flower Garden sequence. The figure explains the relation between the number of control points and the accuracy of interpolation by cubic spline curve fitting. The number of measured data was 25, shown as ‘+’ marks. The obtained approximated curve was almost perfect when 7 control points were used (shown as circled marks). Deviation is noticeable when 4 control points were used (shown as rectangular marks), but errors were less than 1 %. A 2-dimensional set of quantization scales can be constructed if we set x-axis and y-axis as quantization scales of I-picture and P-picture, respectively.

Figure 2 shows the 3-D plots of MSE for each set of quantization scales (from 4 to 30). The z-axis stands for measured MSE for the Flower Garden sequence. If we select some set of measured data as control points, we can interpolate in the y-axis direction as well as in the x-axis direction as follows:

\[
SY_y = a_{x,y} (x - Q_{x,y})^2 + b_{x,y} (x - Q_{x,y}) + c_{x,y} (y - Q_{y,y}) + d_{x,y},
\]

for \( y \in [Q_{y,y}, Q_{y+1,y}] \). (4)

\[
\frac{\partial SY_y}{\partial y} = 2a_{x,y} (x - Q_{x,y}) + b_{x,y} + c_{x,y} (y - Q_{y,y}),
\]

for \( x \in [Q_{x,y}, Q_{x+1,y}] \). (5)

\[
\frac{\partial^2 SY_y}{\partial y^2} = 2a_{x,y},
\]

for \( x \in [Q_{x,y}, Q_{x+1,y}] \). (6)

\[
SX_x = a_{x,y} (x - Q_{x,y})^2 + b_{x,y} (x - Q_{x,y}) + c_{x,y} (y - Q_{y,y}) + d_{x,y},
\]

for \( x \in [Q_{x,y}, Q_{x+1,y}] \). (7)

\[
\frac{\partial SX_y}{\partial x} = 2a_{x,y} (x - Q_{x,y}),
\]

for \( x \in [Q_{x,y}, Q_{x+1,y}] \). (8)

\[
\frac{\partial^2 SX_y}{\partial x^2} = 2a_{x,y},
\]

for \( x \in [Q_{x,y}, Q_{x+1,y}] \). (9)

**IV. PRUNING ALGORITHM**

Since the number of quantization scales is 31, the possible number of choices is 31x31=961. The complexity exponentially increases as the dependency tree depth progresses. Therefore, the computational load is still unrealistic, even though an interpolation scheme is used for a reduction in measuring R-D characteristics. A pruning algorithm has been proposed to obtain a fast solution to the complex dependent allocation problem using the monotonicity property [3]. However, reduction by the pruning algorithm alone is not sufficient to implement a realistic system. In this paper, we propose a joint interpolation as smooth as possible, the first and second derivatives should be continuous as follows:

\[
\frac{dS_{y+1}}{dx}(Q_y) = \frac{dS_y}{dx}(Q_y),
\]

(2)

\[
\frac{d^2S_{y+1}}{dx^2}(Q_y) = \frac{d^2S_y}{dx^2}(Q_y).
\]

(3)
interpolation-pruning algorithm. The monotonicity property can be expressed as follows:

\[ J_2(Q_1, Q_2) \leq J_2(Q_1', Q_2) \quad \text{for} \quad Q_1 \leq Q_1' \]  

(10)

The monotonicity property can also be well depicted by Fig 4. Fig 4. shows R-D plots of P-picture for various choice of quantization scale at I-pictures. The lines lie on the convex hull of R-D plots, signifying the optimal solution of Lagrangian cost function. This can also be explained by noting that a better quality of the reference frame lead to better coding performance in predictive coding. However, the bit budget of the following predictive frame will be affected by the amount of increased bits of the better reference frame.

In the proposed algorithm, the candidate solutions of cost function for B-pictures are calculated with the conventional R-D framework, since the B-pictures are not used as reference frames. The joint interpolation-pruning algorithm is as follows:

**Step 1.** Calculate the minimum cost function \( J_1(Q_1) \) in I-picture. Make two more reference I-frames by coding at \( Q_1 - 2 \) and \( Q_1 + 2 \).

**Step 2.** Calculate the minimum cost function \( J_2(Q_1, Q_2) \). Measure R-D data from the three reference frames, at each \( Q_1 \), \( Q_2 \), \( Q_2 - 2 \), and \( Q_2 + 2 \) constructing 5 control points in the x-direction and y-direction. (Refer to Fig 5). Prune out all the quantization nodes for \( Q_1 > Q_1 + 2 \) and \( Q_2 > Q_2 + 2 \).

**Step 3.** Calculate interpolated R-D data at 20 positions to select the minimum cost function using cubic spline method and parallel shift.

**Step 4.** Determine the minimum cost function \( J_2(Q_1', Q_2) \). If the reference frame does not exist, make the reference frame.

**Step 5.** Repeat from Step 1 to select the cost function of next P-picture \( J_1(Q_1', Q_2', Q_3) \) using the selected reference frame with \( Q_2, Q_2 - 2 \), and \( Q_2 + 2 \).

**Step 6.** The cost functions for B-pictures are calculated by the conventional R-D framework after deciding \( Q_1' \) and \( Q_2' \). This dependent coding path is refreshed at the start of next GOP.

V. EXPERIMENTS

We have experimented the proposed 2-D interpolation of R-D characteristics and pruning algorithms. Motion estimation was based on the original frames to avoid re-computation of motion vectors after the quantization scale of the reference frame was changed. We used the MPEG-2 algorithm as a platform since MPEG-2 is the most popular technology for one-way video applications such as DVD or VOD. According to the experiments, the error between measured data and results of spline interpolation was less than 1% when just one or two data are interpolated in a line segment. We used a pruning algorithm to reduce complexity without severe loss of optimality. Two points were interpolated in both the x-direction and y-direction. Using monotonicity property, we pruned out all the quantization nodes with cost functions greater than the obtained minimum cost function+2. According to the results of simulation with Football and Flower Garden sequences (at 1 Mbps of CIF image and 4 Mbps of SIF image), the proposed algorithm shows improvement of PSNR by 0.9 ~1.8 dB, on average. Table I shows the average PSNR comparison between TM5 and the proposed algorithm. Fig 6 shows the comparison of the PSNR between the proposed algorithm and the TM5 algorithm at 1Mbps. We simplified the algorithm in B-pictures, because the B-picture does not affect the quality of other frames. In the 2-D view of reference quantization scale and predictive quantization scale, we select 5 control points to reduce complexity. However, due to the effects of pruning algorithm and accuracy of interpolation, the optimality is not severely sacrificed.

VI. CONCLUSION

In this paper, we have proposed an efficient bit allocation algorithm using 2-D interpolation of R-D characteristics together with a pruning scheme in the dependent quantization environment, for applications of one-way video. In the dependent coding framework as in current video coding methods, the optimal solution of the R-D approaches becomes exponentially complex as the dependency tree depths progress. So we have to find a sub-optimal solution without severely sacrificing the optimality. Specifically, we proposed 2-D interpolation and pruning algorithms for efficient bit allocation in the dependent coding framework.

To reduce the number of measured data, we have used a cubic spline method. According to the experiments, the cubic spline interpolation scheme performs well with a small number of intervals. We have also proposed an efficient pruning algorithm to reduce the number of calculations. Based on experiments with MPEG-2 video platform, the proposed algorithm shows improvement of PSNR. The perceptual quality has also been improved. Therefore, we have shown that one-way video coding performance can be increased by the proposed bit allocation algorithm.

REFERENCES


**TABLE I. COMPARISON OF AVERAGE PSNR BETWEEN TM5 AND PROPOSED ALGORITHM (dB)**

<table>
<thead>
<tr>
<th></th>
<th>TM5 1Mbps</th>
<th>TM5 4Mbps</th>
<th>Proposed Algorithm 1Mbps</th>
<th>Proposed Algorithm 4Mbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football</td>
<td>29.36</td>
<td>32.62</td>
<td>31.16</td>
<td>33.74</td>
</tr>
<tr>
<td>Flower Garden</td>
<td>25.81</td>
<td>29.75</td>
<td>26.71</td>
<td>30.56</td>
</tr>
</tbody>
</table>

![Fig. 1. The relationship between number of control points and accuracy of interpolation.](image1)

![Fig. 2. 3-D plot of MSE for each 2-D quantization set.](image2)

![Fig. 3. 3-D plot of measure bits for each 2-D quantization set.](image3)

![Fig. 4. R-D plots of P-picture for various quantization scale at I-picture.](image4)

![Fig. 5. measured points and interpolated points](image5)

![Fig. 6. PSNR comparison between proposed algorithm and TM5 algorithm at 1Mbps (Flower Garden)](image6)