Adaptive Rate-Distortion Optimization using Perceptual Hints

Chun-Jen Tsai1+, Chih-Wei Tang2*, Ching-Ho Chen1, and Ya-Hui Yu1

1 Department of Computer Science and Information Engineering,
National Chiao Tung University, Hsinchu, Taiwan, R.O.C
2 Department of Electronics Engineering,
National Chiao Tung University, Hsinchu, Taiwan, R.O.C
+ cjtsai@csie.nctu.edu.tw, * chihwei.ee88g@nctu.edu.tw

Abstract

A novel video coding approach that performs adaptive rate-distortion optimization guided by perceptual hints is proposed in this paper. The key idea is to adaptively adjust the Lagrange multipliers of the RDO coder control module based on visual attention analysis. The observation is that human vision is sensitive to movement of well structured objects while tolerates large distortion in moving areas with random structure (texture-wise and motion-wise). The proposed algorithm analyzes permissible perceptual distortions and, accordingly, assigns larger Lagrange multipliers to regions that are perceptually less sensitive to distortion so that rate-reduction is weighted more than distortion reduction in these regions. Experiments show that this scheme achieves bit-rate saving of 5-7% with virtually same perceptual quality and is very promising for practical systems.

1. Introduction

Current video coder control modules typically select the encoding mode and the QP in the minimal mean absolute error (MAE) sense [1]. However, it is well recognized that minimization of pixel-wise MAE does not translate into minimal perceptual distortion [2]. A better approach is to use perceptual distortion metrics [2][3], instead of MAE, as the distortion measure for video signals. Even though there have been extensive researches on visual models for still images [4], perceptual distortion metrics for video signals involve more sophisticated psychological models which are not fully understood yet [5][6][7].

Given a distortion measure (based on either a perceptual model or simply MAE), rate-distortion optimization (RDO) is a common technique used in high performance video codecs to achieve the best quality under certain rate constraints [8]. In this approach, an encoder selects the coding modes and motion vectors (if any) that gives optimal rate-distortion tradeoff by solving a constrained optimization problem using Lagrangian formulation. Usually, the Lagrange multiplier $\lambda$ is defined as a function of the quantizer step size (QP) alone [8]. However, the optimal choice of $\lambda$ should depend on QP as well as on the perceptual characteristics of the video content. Therefore, it is more reasonable to make $\lambda$ a function of both QP (the rate constraint) and the perceptual characteristic of the target region. In this paper, a novel approach is proposed to perform fast but effective perceptual characteristic analysis on the input video signal. The outcome of the analysis is used together with the target QP to determine the proper Lagrange multiplier for RDO analysis. The proposed perceptual analysis algorithm is based on the observation that human vision is sensitive to movement of well structured objects while tolerates large distortion in moving areas with random structure (texture-wise or motion-wise). Therefore, in a region of random structure, rate-reduction should be weighted more than distortion-reduction during the RDO process. That is, a large $\lambda$ should be used in this case.

There have been previous efforts that try to incorporate perceptual distortion measure into video coding frameworks. For example, the rate control schemes in [9] and [10] use perceptual distortions to determine quantization parameters. In our approach, a “perceptual distortion masking measure” instead of a “perceptual distortion measure” is used. Furthermore, the measure is coupled into the RDO process, instead of the rate control process. In short, for image regions that can sustain large perceptual distortion, bitrate saving are achieved elegantly through wise selection of $\lambda$.

This paper is organized as follows. In section 2, H.264 is used as an example to discuss some problems with current RDO approaches. An efficient perceptual distortion tolerability measure is developed in Section 3. We then propose an adaptive RDO scheme in section 4 that achieves bitrate reduction with virtually no loss in perceptual quality. Experimental results and strong evidences are presented in section 5 to show the effectiveness of the proposed framework. Finally, some discussions and conclusive remarks are given in section 6.

2. Issues with current RDO approach

In order to explain the issues with current RDO approach, the JM model 7.3 of H.264 is used as an example. MPEG-4 AVC/H.264 is a new video coding standard jointly developed by ITU-T VCEG and ISO/IEC MPEG. This video codec
substantially outperforms previous ones such as MPEG-2, H.263 and MPEG-4 by employing a fine motion compensation model and arithmetic coding. The H.264 reference software adopts the Lagrangian optimization to determine a set of coding parameters in the rate-distortion sense under some coding constraints. For one image region, the encoder decides an optimal set of coding parameters to minimize the distortion $D$ subject to a rate constraint $R$. In terms of the Lagrange formula, the optimization is defined as

$$\min \{J\}, \text{ where } J = D + \lambda R.$$  

(1)

The Lagrange cost $J$ is minimized for a given Lagrange multiplier $\lambda$, and $\lambda \geq 0$. The multiplier controls the tradeoff between distortion and rate for different encoding modes and motion vectors.

In the motion estimation stage, RDO minimizes

$$J_{\text{motion}} = D_{\text{DFD}} + \lambda_{\text{motion}} R_{\text{motion}},$$  

(2)

where $D_{\text{DFD}}$ is the motion compensation prediction error, $\lambda_{\text{motion}}$ is the Lagrange multiplier, and $R_{\text{motion}}$ is the number of bits for encoding the motion vector.

In the mode-decision stage, RDO minimizes

$$J_{\text{mode}} = D_{\text{REC}} + \lambda_{\text{mode}} R_{\text{REC}}.$$  

(3)

where $D_{\text{REC}}$ is the distortion between the reconstructed and the original macroblocks, $\lambda_{\text{mode}}$ is the Lagrange multiplier, $R_{\text{REC}}$ is the total bits for mode information, motion vectors and the transform coefficients. In JM 7.3 of H.264,

$$\lambda_{\text{mode}} = 0.85 \cdot \beta^{(12-QP)/3},$$  

(4)

where $QP$ is the quantization parameter. And, $\lambda_{\text{motion}}$ and $\lambda_{\text{mode}}$ are related by

$$\lambda_{\text{motion}} = \sqrt{\lambda_{\text{mode}}}. $$  

(5)

One critical problem with this optimization is that $\lambda$ depends only on $QP$. As discussed in section I, human vision tends to focus on well structured moving objects and ignore moving objects with random texture and/or motion. However, the latter objects are more difficult to encode and bits are wasted for these perceptually insignificant objects. For example, we are more sensitive to the tennis player than the background audience in the STEFAN sequence (Fig. 1).

To illustrate the problem with fixed $\lambda$, for RDO, the RD curves of possible coding modes corresponding to two macroblocks in Fig. 1 are shown in Fig. 2. Each curve in Fig. 2 is generated with five different QPs, and the coding modes including INTER-16x16, INTER-16x8, INTER-8x16, INTER-8x8, INTRA-16x16 and INTRA-4x4. Figures 3 and 4 show the partial RD curves corresponding to $QP = 11$ in Fig. 2. It can be observed that for the 257th macroblock containing the tennis player’s leg, the mode decision is mode $A$ (with larger distortion) rather than mode $B$ in Fig. 3. Nevertheless, since distortion in the tennis player will be noticed more easily, mode $B$ should be preferred over $A$.

On the contrary, for the 330th macroblock on the tennis court, the decision is mode $C$ (with smaller distortion) but not mode $D$ in Fig. 4. In this case, mode $D$ should be a better choice since this region is perceptually less important. Thus, in the next section, a simplified human visual attention analysis scheme is developed to improve the visual quality of RDO by adapting $\lambda$ to the content.
3. Complexity analysis for visual attention

In this section, a fast perceptual distortion masking measure for image regions in a video sequence is developed. This measure is based on the statistics of the co-located image pixel differences of the previous and the target frames. The statistics consists of the mean and deviation values. While moving regions usually gives larger mean values, random-textured moving regions typically produces smaller deviation values. The proposed measure is the multiplication of these two values.

The proposed measure is calculated as follows.

1. For some target frame with frame number \( n_t \), compute the difference image of the luma components between the \( n_t \)th and \((n-1)th\) source frames.

2. Assuming there are \( N \) macroblocks in each frame, for each \( M \times M \) macroblock \( b \) of the difference image, compute its mean \( m_{ab} \) and deviation \( v_{ab} \) as

\[
m_{ab} = \frac{1}{M^2} \sum_{i,j} d_{ij}
\]

and

\[
v_{ab} = \sum_{i,j} |d_{ij} - m_{ab}|
\]

where \( d_{ij} \) is the luma value of pixel \( i,j \) and \( b = 1,2,\ldots,N \).

3. Finally, the perceptual distortion masking measure \( x_{ab} \) for macroblock \( b \) is given by

\[
x_{ab} = m_{ab} \times v_{ab}.
\]

Later, this measure is used to indicate whether an image region allows more perceptual distortion in the adaptive RDO framework.

4. The proposed adaptive RDO technique

Based on the frame complexity measure for human visual attention as described in Section 3, we assign smaller Lagrange multipliers to the regions attracting more human attentions. As a result, more bits, if considered necessary by the RDO algorithm, will be allocated to these regions.

The Lagrange multipliers of coding mode selection and motion estimation in rate distortion optimization are adaptively adjusted as follows.

1. For a target frame with frame number \( n_t \), generate the histogram of collected values \( x_{ab} \) from \( N \) macroblocks, where each bin size is \((\max(x_{ab}) - \min(x_{ab}))/20\).

2. The complex correction value \( \Delta x_{ab} \) for the \( j \)th bin \( x_{ab} \) in the histogram is given by

\[
\Delta x_{ab} = -S \times (x_{ab} - \bar{x}_{nj})/\bar{x}_{nj},
\]

where \( \bar{x}_{nj} \) is the bin value corresponding to the peak of the histogram, and \( j = 1,2,\ldots,20 \). Then, \( \Delta x_{ab} \) is further bounded by parameter \( S \) as

\[
\Delta x_{ab} = \begin{cases} S, & \text{if } \Delta x_{ab} > S. \\ -S, & \text{if } \Delta x_{ab} < -S. \\ \Delta x_{ab}, & \text{otherwise.} \end{cases}
\]

3. For each macroblock in this target frame, the modified Lagrange multiplier \( \lambda'_{MODE} \) is given by

\[
\lambda'_{MODE} = \begin{cases} \alpha \times \lambda_{MODE}, & \text{if } \Delta x_{ab} > 0, \\ \lambda_{MODE}, & \text{otherwise.} \end{cases}
\]

where \( \alpha \) is a scaling parameter, and the initial Lagrange multiplier \( \lambda_{MODE} \) is related to the quantization parameter by Eq.(4). Accordingly, the modified Lagrange multiplier \( \lambda'_{MOTION} \) for motion estimation is also adjusted by Eq.(5).

After determining the Lagrange multipliers for each macroblock, the rate distortion optimization including motion estimation by Eq.(2) and coding mode selection by Eq.(3) works as usual except the modified multipliers.

5. Experimental results

JM 7.3 of H.264 is used to conduct the experiments in this section. The configuration is as follows. The Hadamard transform, CABAC and reconstruction filter are enabled. No B frame is inserted. The encoded sequence is the CIF version of STEFAN at 30 fps. The visual qualities of the reconstructed frames between the original and adaptive RDOs when \( \alpha = 8 \) in Eq. (11) for a given QP = 28 are compared in this section.

The bit rate is 1148K bps, overall PSNR is 33.77dB while they are 1239K bits and 35.58dB by the original RDO. That is, the bit rate reduction is 7%. Even though there seems to be a PSNR decrease of 1.8dB, visually, the difference is not observable since most of the PSNR loss is in the random-textured audiences (as shown in Fig. 6). The PSNR and bitrate differences for each macroblock inside two rectangular regions in Fig. 5 are shown in Fig. 6. We observe that for most of the macroblocks containing the tennis player’s body, both the PSNR and bit rate increase. That is, the visual quality is improved in these regions. On the other hand, the number of
bits for the macroblocks containing the audience is reduced since the perceptual distortions are less noticeable. Therefore, the perceptual quality is actually improved even though the PSNR for the whole video sequence degrades.

6. Conclusions

In this paper, we propose an approach to select content-adaptive Lagrange multipliers for rate distortion optimization under the perceptual distortion criteria. The experiments show that this simple scheme preserves the perceptual quality while reduces the bitrate. We are currently investigating an advanced approach to extend this idea for high perceptual quality video codecs. Early experiments show that the application of proper perceptual model to the rate control module can cut the bitrate in half with visually no degradation at all.

7. References