SSIM based perceptual distortion rate optimization coding

Shiqi Wang*, Siwei Ma, Wen Gao
Institute of Digital Media, Peking University, Beijing, 100871, China

ABSTRACT

The current rate distortion optimization (RDO) coding schemes usually use the Sum of Absolute Difference (SAD) or Sum of Square Difference (SSD) as the distortion metric. However, neither SAD nor SSD correlates with the human visual system (HVS) very well. To develop a perceptual distortion based video encoder, we employ the Structural Similarity (SSIM) Index as the distortion metric and propose a SSIM based Lagrangian perceptual distortion rate optimization (PDRO) method in this paper. Furthermore, to adapt the different input sequences dynamically, we present an adaptive Lagrange multiplier selection scheme based on the properties of the input sequences. By modeling the transformed residuals with Laplace distribution, the statistical SSIM and rate models are deduced to obtain the adaptive Lagrange multiplier. Extensive experiments show that the proposed scheme can achieve better perceptual distortion rate performance and provide better visual quality than the SAD/SSD based RDO coding scheme.

Keywords : Rate distortion optimization, SSIM index, Lagrange multiplier, Laplace distribution.

1. INTRODUCTION

In the current video coding standards, usually more coding modes are designed to improve the coding efficiency, e.g. the coding modes of H.264 can vary in the mode sets {Intra16x16, Intra8x8, Intra4x4, Inter16x16, Inter16x8, Inter8x16, Inter8x8, Inter8x4, Inter4x8, Inter4x4, SKIP, DIRECT}. Which mode should be selected as the best mode is very important for the final coding performance. Actually the mode selection problem is an optimization problem to minimize the distortion $D$ for a given rate $R_c$, shown as:

$$\min \{ D \} \quad \text{subject to} \quad R \leq R_c \quad (1)$$

The constraint optimization problem can be solved with the Lagrangian method as:

$$\min \{ J \} \quad \text{where} \quad J = D + \lambda R \quad (2)$$

where $\lambda$ is the Lagrange multiplier. $J$ is called the rate distortion (RD) cost. The minimal $J$ is given by setting its derivation to $R$ to be zero,

$$\frac{dJ}{dR} = \frac{dD}{dR} + \lambda = 0 \quad (3)$$

yielding:

$$\lambda = - \frac{dD}{dR} \quad (4)$$

In the hybrid video coding standards such as H.263 and H.264, the Lagrange multiplier $\lambda$ is determined by experimental results and typical rate-distortion models [1]. However, this scheme didn’t consider the properties of input sequences. In order to make the rate distortion optimization (RDO) adapt to the various coding statistics and input sequences, some adaptive $\lambda$ estimation methods were proposed. In [2], the authors developed an adaptive $\lambda$ estimation algorithm in $\rho$–domain. In [3], the Laplace distribution based rate and distortion models were established to derive the $\lambda$ dynamically. These methods have been proved to improve the coding performance significantly.

In the traditional RDO coding scheme, the Sum of Absolute Difference (SAD) and Sum of Square Difference (SSD) are used as the distortion measures. However, they are widely criticized because the properties of human visual system...
The Structural Similarity (SSIM) index [5] is a newly developed metric which is well matched to the perceived quality. In [6], [7], the SSIM index was incorporated into the motion estimation and mode selection in hybrid video coding, but the Lagrange multiplier was derived experimentally so that the properties of input sequences were ignored in the rate distortion optimization.

In this paper, we employ SSIM as the distortion metric and propose a SSIM based perceptual distortion rate optimization coding method. The rate and distortion models are firstly established based on the Laplace distribution of transformed residuals. Then the Lagrange multiplier $\lambda$ is derived dynamically. Finally, the PDRO is performed with the Lagrange multiplier $\lambda$.

The paper is organized as follows: In Section 2, the statistical SSIM model and rate model are developed to deduce the Lagrange multiplier. Then the proposed SSIM based perceptual distortion rate optimization scheme is introduced in Section 3. Section 4 presents the experimental results. Finally, the conclusion is given in Section 5.

2. STATISTICAL SSIM AND RATE MODELS

The relationship between SSIM index of quantized images and the quantization step size was first developed by S. Channappayya et al. [8], but this scheme only gave the upper and lower bounds on SSIM index, which was inappropriate for RDO. In this study, we propose a novel SSIM-Q model which can accurately estimate the value of SSIM index.

Supposing $x$ and $y$ to be the two image signal vectors (both in $\mathbb{R}^N$), the SSIM index between $x$ and $y$ could be simplified in the form [5]:

$$SSIM(x, y) = \left( \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right) \times \left( \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right)$$

where $\mu_x$, $\mu_y$ are the mean intensity of $x$ and $y$ respectively. $\sigma_x$, $\sigma_y$ are the standard deviation of $x$ and $y$ respectively, and $\sigma_{xy}$ is the cross-covariance between $x$ and $y$. The constants $C_1$ and $C_2$ are used to avoid instability when the means and variances become smaller.

To derive a statistical SSIM model for PDRO in the DCT domain, we firstly divide one frame into 4x4 non-overlapped blocks which is the smallest size in the mode selection. The size of the blocks is indicated by $N$. Then these blocks within one frame are classified into smooth ones and non-smooth ones by an adaptive threshold determined by the quantization step $Q_{step}$. The percentage of smooth blocks is represented by $p$. Finally, the average SSIM index between the block $x$ in the original frame and the distorted block $y$ in the decoded frame is approximated by:

$$E[SSIM(x, y)] \approx \left[ 1 - E[X(0) - Y(0)]^2 \times E[\frac{1}{2X(0)^2 + NC_1}] \times \left( 1 - p \times E[\frac{\sum_{k=1}^{N-1} X_s(k)^2}{\sum_{k=1}^{N-1} X_s(k)^2 + NC_2}] \right) \right]$$

$$- (1 - p) \times E[\sum_{k=1}^{N-1} (X_s(k) - Y_s(k))^2] \times E[\frac{1}{2 \sum_{k=1}^{N-1} X_s(k)^2 + NC_2}]$$

where $X(i)$ and $Y(i)$ are the DCT coefficients for $x$ and $y$. $X_s(i)$ represents the AC coefficients of the smooth blocks in $x$; $X(i)$ and $Y(i)$ indicate the AC coefficients of the non-smooth blocks in $x$ and $y$.

Proof:

In the DCT domain, the SSIM index between the block $x$ in the original frame and the distorted block $y$ in the decoded frame can be modeled as [8]:

$$SSIM(x, y) = M \cdot S, \text{ with } \quad M = 1 - D_1 \cdot F_1 \quad S = 1 - D_2 \cdot F_2 \quad D_1 = [X(0) - Y(0)]^2$$

$$D_2 = \sum_{k=1}^{N-1} [X(k) - Y(k)]^2 \quad F_1 = \frac{1}{X(0)^2 + Y(0)^2 + NC_1} \quad F_2 = \frac{1}{\sum_{k=1}^{N-1} (X(k)^2 + Y(k)^2) + NC_2}$$

(7)
Since the DCT coefficients have low correlation, the expectation of SSIM index is given by:

\[ E[\text{SSIM}(x, y)] = E(M)E(S) \]  

(8)

From Table 1, we can observe that the distortion of DC coefficients \( D_1 \) have weak correlation with \( F_1 \). Furthermore, as illustrated in Fig.1, the low frequency components are less distorted so that the DC coefficients of the original blocks and the distorted blocks are almost the same. Therefore, we can approximate \( Y(0) \) by \( X(0) \), yielding:

\[ E(M) \approx 1 - E[X(0) - Y(0)]^2 \times E\left[\frac{1}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right] \approx 1 - E[X(0) - Y(0)]^2 \times E\left[\frac{1}{2X(0)^2 + N \cdot C_1}\right] \]  

(9)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QP</th>
<th>I</th>
<th>P</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman@CIF</td>
<td>20</td>
<td>0.0368</td>
<td>0.0253</td>
<td>0.0405</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.0495</td>
<td>0.0777</td>
<td>0.0830</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.0807</td>
<td>0.1138</td>
<td>0.1298</td>
</tr>
</tbody>
</table>

Table 1. Correlation coefficients between \( D_1 \) and \( F_1 \)

Fig.1. Relationship between the DC coefficients of the original blocks \( x \) and the distorted blocks \( y \) (QP=40)

However, for AC coefficients, there exists a strong correlation between \( D_2 \) and \( F_2 \), as illustrated in Table 2. This is mainly because in hybrid video coding, especially low bit rate video coding, the prediction techniques for AC coefficients is not as accurate as that for the DC coefficients. To solve this problem, we classify the blocks in a frame into smooth ones and non-smooth ones by setting a adaptive threshold \( T = (1 - \gamma)^2 Q_{\text{step}}^2 \) where \( \gamma \) is the rounding offset in the quantization. If the sum of square of the AC coefficients in a block is smaller than the threshold, i.e.,

\[ \sum_{k=1}^{N-1} X(k)^2 < (1 - \gamma)^2 Q_{\text{step}}^2 \]  

(10)

then the block is classified into a smooth block, otherwise it is recognized as a non-smooth block. The percentage \( p \) is computed as the percentage of smooth blocks in a frame.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QP</th>
<th>I</th>
<th>P</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman@CIF</td>
<td>20</td>
<td>-0.5081</td>
<td>-0.5479</td>
<td>-0.5043</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>-0.7284</td>
<td>-0.6828</td>
<td>-0.6428</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>-0.7761</td>
<td>-0.7481</td>
<td>-0.7238</td>
</tr>
</tbody>
</table>

Table 2. Correlation coefficients between \( D_2 \) and \( F_2 \)

In the smooth blocks, the probability of the predicted AC coefficients to be zero is high; then a large percent of AC coefficients are quantized into zero and \( S \) will be approximated by:

\[ E(S) \approx 1 - E\left[\frac{\sum_{k=1}^{N-1} X_k(k)^2}{\sum_{k=1}^{N-1} X_k(k)^2 + N \cdot C_2}\right] \]  

(11)

Proc. of SPIE Vol. 7744 774407-3
Furthermore, the correlation between $D_2$ and $F_2$ in non-smooth blocks is much smaller than that in smooth blocks, especially in low bit rate video coding. For instance, we encode the sequence of Foreman in CIF format at QP=40. The correlation coefficient between $D_2$ and $F_2$ in P frame is only -0.18 in the non-smooth blocks while it is -0.77 in smooth blocks. More results can be found in Table 3. Consequently, $S$ in non-smooth blocks can be approximated by:

$$E(S) \approx 1 - E[\frac{1}{N} \sum_{k=1}^{N} (X_i(k) - Y_i(k))^2] \times E[\frac{1}{2} \sum_{k=1}^{N} (X_i(k)^2 + Y_i(k)^2) + N \cdot C_2]$$  \hspace{1cm} (12)$$

From Fig. 2, we can observe that $Y(k)$ can also be approximated by $X(k)$ in non-smooth blocks, yielding:

$$E(S) \approx 1 - E[\frac{1}{N} \sum_{k=1}^{N} (X_i(k) - Y_i(k))^2] \times E[\frac{1}{2} \sum_{k=1}^{N} X_i(k)^2 + N \cdot C_2]$$  \hspace{1cm} (13)$$

From Equation (12) and (13), we will compute $E(S)$ as:

$$E(S) \approx (1 - p \times E(\frac{N}{N} \sum_{k=1}^{N} X_i(k)^2) - (1 - p) \times E(\frac{1}{N} \sum_{k=1}^{N} (X_i(k) - Y_i(k))^2) \times E(\frac{1}{2} \sum_{k=1}^{N} X_i(k)^2 + N \cdot C_2))$$  \hspace{1cm} (14)$$

Table 3. Correlation coefficients between $D_2$ and $F_2$ for smooth blocks and non-smooth blocks

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Block type</th>
<th>QP</th>
<th>I</th>
<th>P</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman@CIF</td>
<td>Smooth</td>
<td>20</td>
<td>-0.8077</td>
<td>-0.8247</td>
<td>-0.8379</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>-0.7812</td>
<td>-0.7047</td>
<td>-0.6648</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>-0.8375</td>
<td>-0.7739</td>
<td>-0.7226</td>
</tr>
<tr>
<td></td>
<td>Non-smooth</td>
<td>20</td>
<td>-0.4009</td>
<td>-0.4308</td>
<td>-0.3933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>-0.4376</td>
<td>-0.3722</td>
<td>-0.3454</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>-0.2049</td>
<td>-0.1762</td>
<td>-0.1367</td>
</tr>
</tbody>
</table>

Incorporating (9) and (14) into (8), we will obtain the approximation of SSIM index, as presented in Equation (6). In Equation (6), the average distortion of DC coefficients $D_1$ and AC coefficients $D_2$ of non-smooth blocks will be modeled by the distribution of transformed residuals. Other quantities will be obtained by discrete cosine transform of the original frame. In this way, the SSIM model could be achieved.
In this study, in order to achieve a good tradeoff between the fidelity and the computation complexity, we model the transformed residuals with the Laplace distribution, i.e.,

\[ f_{Lap}(x_r) = \frac{\Lambda}{2} e^{-\Lambda|x_r|} \]

(15)

where \( x_r \) represents the transformed residuals and \( \Lambda \) is the Laplace parameter. \( \sigma \) is the standard derivation of the transformed residuals. Then the distribution of DC coefficients and AC coefficients of the non-smooth blocks are characterized by \( \Lambda_{dc} \) and \( \Lambda_{act} \):

\[ f_{Lap}(x_{dc}) = \frac{\Lambda_{dc}}{2} e^{-\Lambda_{dc}|x_{dc}|} \]

(16)

\[ f_{Lap}(x_{act}) = \frac{\Lambda_{act}}{2} e^{-\Lambda_{act}|x_{act}|} \]

(17)

where \( x_{dc} \) represents the DC coefficients of the transformed residuals and \( x_{act} \) represents the AC coefficients of the transformed residuals in non-smooth blocks.

Then the distortion of the DC coefficients and the AC coefficients of non-smooth blocks will be approximated by summing up the distortion in every quantization interval:

\[ E(X(0) - Y(0))^2 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x_{dc}^2 f_{Lap}(x_{dc})dx_{dc} + 2 \sum_{n=1}^{\infty} \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x_{dc} - nQ)^2 f_{Lap}(x_{dc})dx_{dc} \]

(18)

\[ E(\sum_{k=1}^{N} (X(k) - Y(k))^2) = (n-1)(\int_{-(Q-\gamma Q)}^{Q-\gamma Q} x_{act}^2 f_{Lap}(x_{act})dx_{act} + 2 \sum_{n=1}^{\infty} \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x_{act} - nQ)^2 f_{Lap}(x_{act})dx_{act}) \]

(19)

Incorporating (18) and (19) into (6), we will obtain the final closed form of the SSIM-Q model. As presented in the later section, the novel adaptive model which is determined by the properties of the input sequences will accurately predict the SSIM index in the applications of video coding.

The rate model is derived based on two observations. Firstly, the skipping blocks should not be included in the rate model because the skipped blocks will not be entropy coded [3]. Secondly, in entropy encoding, the zero coefficients after the last non-zero coefficient within one block in the zigzag scan order will not be coded at all. So we extract the residual coefficients before the last non-zero coefficient in the non-skipped blocks and use the Laplace distribution described in (15) to model the residuals. We use \( x_{rc} \) to represent the extracted coefficients and \( \Lambda_r \) to represent the Laplace parameter, and then the probability of these transformed residuals is computed as:

\[ P_0 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} f_{Lap}(x_{rc})dx_{rc} \]

(20)

\[ P_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} f_{Lap}(x_{rc})dx_{rc} \]

(21)

The average rate can be approximated by the entropy, which is:

\[ R = (1 - p_s)H = (1 - p_s) \left( -P_0 \log_2 P_0 - \sum_{n=1}^{\infty} P_n \log_2 P_n \right) \]

(22)

where \( p_s \) indicates the percentage of non-encoded coefficients within one frame. Although the side information is important in inter frame video coding [9], it doesn’t change as much as the source information when QP changes. Therefore, we didn’t consider the side information in the rate model.
3. PROPOSED PERCEPTUAL DISTORTION RATE OPTIMIZATION

In section 2, the statistical rate and SSIM models were derived. To incorporate the SSIM index into the RDO process, the rate distortion cost is defined as [7]:

\[ J = (1 - SSIM) + \lambda \cdot R \]  

(23)

Since this rate distortion cost needs more computation time than the cost defined in H.264, we didn’t use it in motion estimation; otherwise it will bring a lot of additional complexities.

In section 2, the closed form of rate and SSIM models were derived. Then the Lagrange multiplier is also calculated by setting the derivation of \( J \) to zero,

\[ \frac{dJ}{dR} = \frac{dSSIM}{dR} + \lambda = 0 \]  

(24)

which yields:

\[ \lambda = \frac{dSSIM}{dQ} = \frac{dSSIM}{dR} \frac{dR}{dQ} \]  

(25)

The Lagrange multiplier for the SSIM based perceptual distortion rate optimization is obtained by incorporating (6) and (22) into (25). Finally, the PDRO is performed with the Lagrange multiplier \( \lambda \) derived in Equation (25). Unfortunately, the closed form of the Lagrange multiplier derived is too long to be presented.

It should be noted that the Lagrange parameter should be determined before coding the current frame. However, the parameters \( \Lambda_{bc}, \Lambda_{act}, \Lambda_t \) and \( p_s \) are not available at that time. So we estimate them by the average values in three previously coded frames with the same coding type.

Furthermore, before coding the current frame, we should perform discrete cosine transform on the original frame to obtain the quantities needed in Equation (6). However, the additional complexity brought by the proposed method is ignorable compared to the other techniques used in video coding such as motion estimation.

4. EXPERIMENTAL RESULTS

Experiments have been conducted to verify the validity of our proposed SSIM-Q model. Three sequences with CIF format: Mobile, Flower and Foreman are encoded by H.264 software JM 15.1 with fixed QP. The first frame is I-frame while all the rest are P-frames. The SSIM index of each frame is obtained by averaging all of the 4x4 blocks. We didn’t use the deblocking filter in this experiment. From Fig.3 we can observe that the SSIM-Q curves estimated by the proposed method fit the actual curves very well.

![Fig. 3. Estimation of SSIM-Q curves by the proposed model](image)

To perform the SSIM based PDRO, we encode the sequences at different QP values from 28 to 40. The common conditions are used in testing: one I frame followed by 99 P frames, high complexity RDO and five reference frames.
The experimental results are listed in Table 4. The R-D performance curves are also shown in Fig. 4. On average, 0.0095 in SSIM or 11.53% rate reduction is achieved. For the peak gain, 15.91% rate reduction is achieved for Mother_daughter.

Although there is some PSNR loss by the proposed algorithm, the subject quality has been improved. Two frames coded with almost the same bits and bit rate are presented in Fig. 5. The SSIM index of the proposed method is higher because of the proposed SSIM based PDRO. As can be seen from Fig. 5, the proposed method has better visual quality than the traditional RDO coding scheme, although the PSNR of the proposed method is lower.

Table 4. Rate distortion performance of the proposed algorithm (Compared to H.264 for CIF format)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Type</th>
<th>ΔSSIM</th>
<th>AR</th>
<th>Sequence</th>
<th>Type</th>
<th>ΔSSIM</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman</td>
<td>IPP..</td>
<td>0.0019</td>
<td>-2.89%</td>
<td>Flower</td>
<td>IPP..</td>
<td>0.0088</td>
<td>-14.13%</td>
</tr>
<tr>
<td>Mobile</td>
<td>IPP..</td>
<td>0.0104</td>
<td>-12.17%</td>
<td>Bus</td>
<td>IPP..</td>
<td>0.0165</td>
<td>-11.78%</td>
</tr>
<tr>
<td>Paris</td>
<td>IPP..</td>
<td>0.0102</td>
<td>-12.30%</td>
<td>Mother_daughter</td>
<td>IPP..</td>
<td>0.0089</td>
<td>-15.91%</td>
</tr>
</tbody>
</table>

Fig. 4. Performance comparison on test sequences Foreman, Mobile, Paris, Flower, Bus and Mother_daughter in CIF format (IPP…)

![Graphs showing performance comparison for different sequences](image-url)
5. CONCLUSION

In this paper, a SSIM based perceptual distortion rate optimization method is presented. Based on the Laplace distributions of transformed residuals, statistical SSIM and rate models are firstly derived. Then the adaptive Lagrange multiplier is determined according to the properties of input sequences. Extensive experimental results show that the proposed scheme outperforms the traditional RDO in H.264/AVC reference model and a peak gain up to 15.91% rate reduction is achieved.

ACKNOWLEDGEMENT

This work was supported in part by National Science Foundation (60833013, 60803068) and National Basic Research Program of China (973 Program, 2009CB320903). The authors would like to thank Prof. Zhou Wang for the valuable suggestions.

REFERENCE