SSIM-INSPIRED DIVISIVE NORMALIZATION FOR PERCEPTUAL VIDEO CODING

Shiqi Wang [1,2], *Abdul Rehman*[2], *Zhou Wang*[2], *Siwei Ma*[1], *Wen Gao*[1]

[1]Institute of Digital Media, Peking University, Beijing 100871, China
[2]Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

We propose a perceptual video coding framework based on an SSIM-inspired divisive normalization scheme as an attempt to transform the DCT domain frame prediction residuals to a perceptually uniform space before coding. Based on the residual divisive normalization process, we define a distortion model for mode selection and show that such a divisive normalization strategy largely simplifies the subsequent perceptual rate-distortion optimization procedure. Experiments demonstrate that the proposed scheme can achieve significant gain in terms of rate-SSIM performance in comparison with H.264/AVC.

Index Terms— SSIM index, rate distortion optimization, residual divisive normalization, H.264/AVC coding

1. INTRODUCTION

The main objective of video coding is to optimize the perceptual quality of the reconstructed video within available bit rate. Ideally, the distortion model used in the video coding framework should correlate perfectly with perceived distortion of the Human Visual System (HVS), which is the ultimate consumer of the video content. However, almost all existing video coding techniques use the Sum of Absolute Difference (SAD) or Sum of Square Difference (SSD) as the distortion model. It has been widely criticized in the literature that SAD and SSD measures correlate poorly with the HVS [1]. Fortunately, a lot of research has been done recently towards perceptual image quality assessment (IQA) models that perform significantly better than SSD or SAD in predicting perceptual image quality. Among them, the structural similarity (SSIM) index [1] is widely used in quantifying compression artifacts because of its accuracy, simplicity and efficiency. Recently, there have been a number of efforts to design video coding techniques based on the SSIM index, e.g., mode selection [2] and rate control [3].

Since the HVS has varying sensitivity to different frequencies, frequency weighting [4] has been incorporated in the quantization process in many picture coding standards, from JPEG to H.264/AVC high profile [5], [6]. However, in these standards, the quantization matrix is usually predetermined and is fixed once the coding process starts. More advanced perceptual models that take into account supra-threshold distortion criteria and masking effect are not considered.

In this paper, inspired by the SSIM index [1] and its derivation in DCT domain [7], we propose a joint residual divisive normalization and rate distortion optimization (RDO) scheme for video coding. The normalization factor is obtained from the prediction MB. As a result, the quantization matrix is determined adaptively and no side information is required to be transmitted from the encoder to the decoder. Furthermore, motivated by the SSIM index, we define a new distortion model and propose a perceptual RDO scheme for mode selection.

2. SSIM INSPIRED RESIDUAL DIVISIVE NORMALIZATION

Our work follows the predictive video coding framework, where previously coded frames are used to predict the current frame, and only the residuals after prediction is coded. Let $C(k)$ be the k^{th} DCT transform coefficient for residuals, then the normalized coefficient is computed as $C'(k) = C(k)/f$ where f is a positive normalization factor. The quantization of the normalized coefficients, for a given predefined Q_s, is performed as follows

$$\begin{aligned} Y(k) &= sign\{C'(k)\}round\{\frac{|C'(k)|}{Q_s}+p\} \\ &= sign\{C(k)\}round\{\frac{|C(k)|}{Q_s \cdot f}+p\} \end{aligned} \tag{1}$$

where p is the rounding offset in the quantization.

This divisive normalization scheme can be interpreted in two ways. One can apply an adaptive normalization factor, followed by quantization with a predefined fixed step Q_s. Alternatively, one can define an adaptive quantization matrix for each MB and thus each coefficient is quantized with a different quantization step $Q_s \cdot f$. By (1), we see that these two interpretations are equivalent.

In the context of still image processing and coding, several approaches have been used to derive the normalization factor, which can be defined as the sum of the squared neighboring coefficients plus a constant [8], or derived from a local statistical image model [9]. Since our objective here is to optimize

978-1-4577-1302-6/11/$26.00 ©2011 IEEE

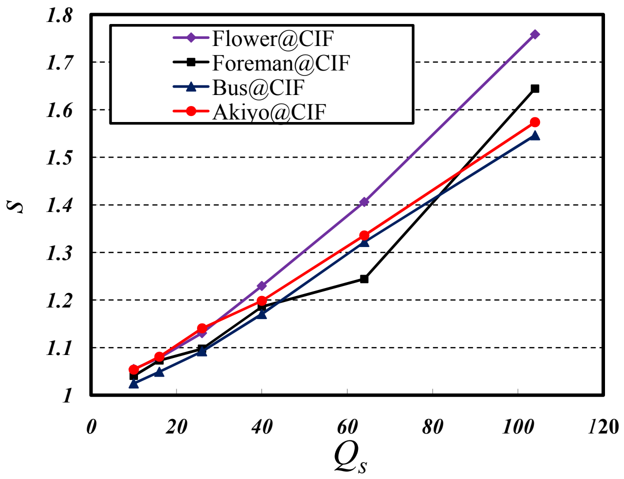

Fig. 1. Energy compensation factor s vs quantization step Q_s for different video sequences.

the SSIM index, we employ a convenient approach based on the DCT domain SSIM index.

The DCT domain SSIM index was first presented by Channappayya *et al.* [7].

$$SSIM(\mathbf{x},\mathbf{y}) = \{1 - \frac{(X(0)-Y(0))^2}{X(0)^2+Y(0)^2+N\cdot C_1}\}\times \{1-\frac{\frac{\sum_{k=1}^{N-1}(X(k)-Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\} \tag{2}$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients for the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. C_1 and C_2 are constants used to avoid instability when the means and variances are close to zero and N denotes the block size. This equation shows that the SSIM index is composed of the product of two terms, which are the normalized squared errors of DC and AC coefficients, respectively. Moreover, the normalization is conceptually consistent with the light adaptation (luminance masking) and contrast masking effects of the HVS [10].

We divide each MB into l sub-MBs for DCT transform. Normalization factors for DC and AC coefficients in each MB are desired to be

$$f_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2+Y_i(0)^2+N\cdot C_1}}{E(\sqrt{X(0)^2+Y(0)^2+N\cdot C_1})} \tag{3}$$

$$f_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2+Y_i(k)^2)}{N-1}+C_2}}{E(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2})} \tag{4}$$

where $X_i(k)$ denotes the k^{th} DCT coefficient in the i^{th} sub-MB and E represents the mathematical expectation operator.

These normalization factors would need to be computed at both the encoder and the decoder. The difficulties are that the distorted MB is not available at the encoder before it is coded, and the original MB is completely inaccessible at the decoder. Fortunately, for each mode, the prediction MB is available at both encoder and decoder sides. Assuming that the properties of the prediction MB are similar to those of the original and distorted MBs, we can approximate the normalization factor as

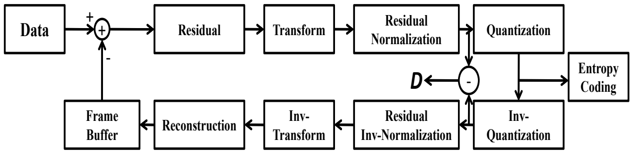

Fig. 2. Diagram of the proposed scheme.

$$f'_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2+N\cdot C_1}}{E(\sqrt{2Z(0)^2+N\cdot C_1})} \tag{5}$$

$$f'_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(Z_i(k)^2+s\cdot Z_i(k)^2)}{N-1}+C_2}}{E(\sqrt{\frac{\sum_{k=1}^{N-1}(Z(k)^2+s\cdot Z(k)^2)}{N-1}+C_2})} \tag{6}$$

where $Z_i(k)$ is the k^{th} DCT coefficient of the i^{th} prediction sub-MB for each mode. For intra mode, we use the MB at the same position in the previous coded frame.

Since the energy of AC coefficients may be lost due to quantization, we use a compensation factor s to bridge the difference between the energy of AC coefficients in the prediction MB and the original MB,

$$s = \frac{E(\sum_{k=1}^{N-1}X(k)^2)}{E(\sum_{k=1}^{N-1}Z(k)^2)} \tag{7}$$

As depicted in Fig. 1, s exhibits an approximately linear relationship with Q_s, which can be modeled empirically as

$$s = 1 + 0.005\cdot Q_s \tag{8}$$

Finally, analogous to [11], we define the quantization matrix for 4x4 DCT transform coefficients as

$$WS_{ij} = 16\cdot\begin{bmatrix} f'_{dc} & f'_{ac} & f'_{ac} & f'_{ac} \\ f'_{ac} & f'_{ac} & f'_{ac} & f'_{ac} \\ f'_{ac} & f'_{ac} & f'_{ac} & f'_{ac} \\ f'_{ac} & f'_{ac} & f'_{ac} & f'_{ac} \end{bmatrix} \tag{9}$$

3. PERCEPTUAL RATE DISTORTION OPTIMIZATION

The RDO process in video coding can be expressed by minimizing the perceived distortion D with the number of used bits R subjected to a constraint R_c. This can be converted to an unconstrained optimization problem as

$$\min\{J\} \quad \text{where} \quad J = D + \lambda\cdot R \tag{10}$$

where J is called the Rate Distortion (RD) cost and λ is known as the Lagrange multiplier which controls the trade-off between R and D.

In conventional RDO schemes, distortion models such as SAD and SSD are used in actual implementations. Here we replace them with a new distortion model that is consistent with the residual normalization process. As illustrated in Fig. 2, the distortion model is defined as the SSD between the normalized coefficients, which is expressed by

$$D = \frac{(X(0) - Y(0))^2}{f'^2_{dc}} + \frac{\sum_{N=1}^{N-1}(X(k) - Y(k))^2}{f'^2_{ac}} \tag{11}$$

Based on (10), the RDO problem can be approximated as

$$\min\{J\} \quad \text{where} \quad J = \frac{(X(0) - Y(0))^2}{f'^2_{dc}} + \lambda_{dc} \cdot R_{dc} + \frac{\sum_{N=1}^{N-1}(X(k) - Y(k))^2}{f'^2_{ac}} + \lambda_{ac} \cdot R_{ac} \tag{12}$$

The Lagrange parameter λ_{dc} for DC coefficient is obtained by calculating the derivative of J with respect to R_{dc}, then setting it to zero. Solving λ_{dc} for

$$\frac{dJ}{dR_{dc}} = \frac{d\frac{(X(0)-Y(0))^2}{f'^2_{dc}}}{dR_{dc}} + \lambda_{dc} = 0, \tag{13}$$

yields

$$\lambda_{dc} = -\frac{d\frac{(X(0)-Y(0))^2}{f'^2_{dc}}}{dR_{dc}} = -\frac{d(X(0) - Y(0))^2}{dR_{dc} \cdot f'^2_{dc}} \tag{14}$$

In H.264, the Lagrange multiplier derived for optimizing MSE is given by

$$\lambda_{H.264} = -\frac{dD_{MSE}}{dR} = c \cdot Q_s^2 \tag{15}$$

and the quantization step after applying the quantization matrix can be expressed as

$$Q_d = Q_s \cdot f'_{dc} \tag{16}$$

Combining (14) to (16), the Lagrange parameter for DC coefficient is computed as

$$\lambda_{dc} = \frac{c \cdot Q_d^2}{f'^2_{dc}} = c \cdot Q_s^2 = \lambda_{H.264} \tag{17}$$

This suggests that we can use the Lagrange multiplier derived with the predefined quantization step in our perceptual RDO scheme. The Lagrange multiplier for AC coefficients, λ_{ac}, can be derived in a similar fashion.

From the residual normalization point of view, the distortion model calculates the SSD between the normalized original and quantized coefficients, as shown in Fig. 2. In this way, the normalized residuals are quantized with the predefined constant quantization step, which also explains why we can use $\lambda_{H.264}$ in our perceptual RDO scheme.

4. IMPLEMENTATION AND EXPERIMENTS

Since DCT is an orthogonal transform that obeys Parseval's theorem, we have

$$\mu_x = \frac{\sum_{i=0}^{N-1} x(i)}{N} = \frac{X(0)}{\sqrt{N}} \tag{18}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{N-1} X(i)^2}{N-1} \qquad \sigma_{xy} = \frac{\sum_{i=1}^{N-1} X(i)Y(i)}{N-1} \tag{19}$$

Therefore, although our algorithms are derived in DCT domain, in actual implementations, it is not necessary to perform actual DCT transform for each block in order to perform normalization.

The proposed scheme has been implemented on the H.264/AVC reference software JM15.1. The common coding configurations are set as follows: only 4x4 DCT transform is enabled; all available inter and intra modes are enabled; five reference frames; one I frames followed by 99 P frames; high complexity RDO and the fixed quantization parameters (QP). We employ the method proposed in [12] to calculate the differences between two R-D curves. Furthermore, we use two different sets of QP values in the experiments: QP$_1$= $\{18, 22, 26, 30\}$ and QP$_2$=$\{26, 30, 34, 38\}$, where QP$_1$ represents a high bit-rate coding configuration.

From Table 1, it can be observed that over a wide range of test sequences, our proposed scheme achieves average rate reduction of 15.11% for QP$_1$ and 17.23% for QP$_2$ for fixed SSIM values, and the maximum coding gain is 37%. It is observed that our scheme performs better when there exist significant statistical differences in the same frame, for example, in sequences $Bridge$ and $Flower$. The rate-distortion performance of $Flower$ is shown in Fig. 3. It is also observed that the gains become more significant at middle bit-rates. This may be explained as follows. At high bit rate, the quantization step is relatively smaller and thus the differences of quantization steps among the MBs are not significant. At low bit rate, since the AC coefficients are severely distorted, the normalization factors derived from the prediction frame do not precisely represent the properties of original frame.

5. CONCLUSION

We propose an SSIM-inspired novel joint residual divisive normalization and rate distortion optimization scheme. The novelty of the scheme lies in normalizing the transform coefficients based on the DCT domain SSIM index and defining a new distortion model based on the divisive normalization approach. The proposed scheme demonstrates superior performance as compared to the state-of-the-art H.264 video codec by offering significant rate reduction, while keeping the same level of SSIM values.

Table 1. Performance of the proposed scheme (anchor: H.264/AVC video coding).

Sequence	QPs=(18,22,26,30)			QPs=(26,30,34,38)		
	$\Delta SSIM$	ΔR	$\Delta PSNR$ (dB)	$\Delta SSIM$	ΔR	$\Delta PSNR$(dB)
Akiyo(QCIF)	0.0033	-18.50%	-0.02	0.0085	-14.13%	0.29
Bridge_close(QCIF)	0.0063	-30.03%	-0.58	0.0242	-37.47%	0.50
News(QCIF)	0.0025	-11.62%	-0.69	0.0054	-9.58%	-0.27
Suzie(QCIF)	0.0024	-8.82%	-0.58	0.0040	-6.27%	-0.32
Flower(CIF)	0.0035	-23.61%	-1.98	0.0101	-20.16%	-1.29
Bus(CIF)	0.0041	-13.19%	-1.98	0.0183	-21.33%	-1.28
Waterfall(CIF)	0.0036	-12.90%	-0.32	0.0111	-8.20%	-0.11
Mobile(CIF)	0.0014	-8.04%	-1.21	0.0045	-12.20%	-0.74
Parkrun(720p)	0.0075	-12.70%	-2.29	0.0287	-31.80%	-1.86
Night(720p)	0.0028	-11.73%	-1.65	0.0058	-11.21%	-0.91
Average	0.0037	-15.11%	-1.13	0.0121	-17.23%	-0.60

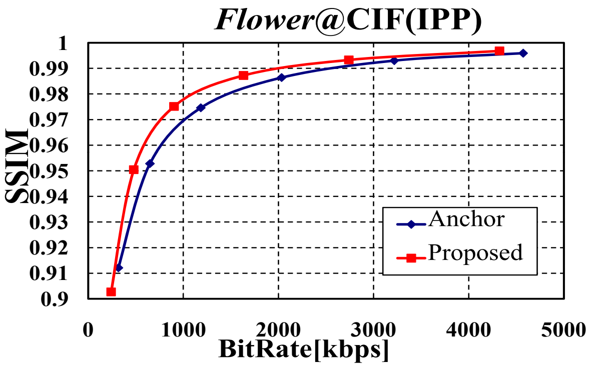

Fig. 3. Rate-SSIM performance comparison of the proposed and H.264/AVC coding schemes.

6. ACKNOWLEDGEMENT

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada, Ontario Early Researcher Award program, National Basic Research Program of China (973 Program, 2009CB320903), National Science Foundation (60833013, 60803068) of China and the State Scholarship Fund from the China Scholarship Council, which are gratefully acknowledged.

7. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[2] S. Wang, S. Ma, and W. Gao, “SSIM based perceptual distortion rate optimization coding,” *Visual Communications and Image Processing(VCIP)*, Jul. 2010.

[3] T. Ou, Y. Huang, and H. Chen, “A perceptual-based approach to bit allocation for H.264 encoder,” *SPIE Visual Communications and Image Processing*, Jul. 2010.

[4] J. Chen, J. Zheng, and Y. He, “Macroblock-level adaptive frequency weighting for perceptual video coding,” *IEEE Trans. on Consumer Electronics*, vol. 53, pp. 775–781, May. 2007.

[5] T. Suzuki, P. Kuhn, and Y. Yagasaki, “Quantization tools for high quality video,” *Joint Video Team of ISO/IEC MPEG and ITU-T VCEG JVT-B067*, Jan. 2002.

[6] T. Suzuki, K. Sato, and Y. Yagasaki, “Weighting matrix for jvt codec,” *Joint Video Team of ISO/IEC MPEG & ITU-T VCEG JVT-C053*, May. 2002.

[7] S. Channappayya, A. C. Bovik, and Jr. R. W. Heathh, “Rate bounds on SSIM index of quantized images,” *IEEE Trans. on Image Processing*, vol. 17, pp. 1624–1639, Sep. 2008.

[8] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, “Non-linear image representation for efficient perceptual coding,” *IEEE Trans. on Image Processing*, vol. 15, pp. 68–80, Jan. 2006.

[9] M. J. Wainwright and E. P. Simoncelli, “Scale mixtures of gaussians and the statistics of natural images,” *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 855–861, 2000.

[10] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual neuroscience*, vol. 9, no. 2, pp. 181–197, aug 1992.

[11] Toshiba, “Adaptive quantization matrix selection,” in *ITU WP3/SC16 Delayed contribution 267, T05-SG16-060403-D-0266*, Geneva, Apr. 2006.

[12] G. Bjontegaard, “Calculation of average PSNR difference between RD curves,” *Proc. ITU-T Q.6/SG16 VCEG 13th Meeting, Austin, TX*, Apr. 2001.