CLOSED LOOP OPTIMIZATION OF IMAGE CODING USING SUBJECTIVE ERROR CRITERIA

Minoru ETOH, Mitsuru KOBAYASHI, and Satoru ADACHI

Multimedia Laboratories, NTT DoCoMo, Inc.
3-5, Hikarinooka, Yokosuka, Kanagawa, Japan

ABSTRACT

This paper proposes a closed-loop optimization framework to improve image coding efficiency by searching DCT coefficients at the equivalent subjective quality to original coding result. The proposed framework shares a basic idea currently adopted in speech coding that searches optimal codes in closed-loop operation, evaluating the coded signal with perceptually weighted mean square error. To evaluate the perceptual quality in image coding, we introduce Masked PSNR that accounts masking effects, by which we apply the stepwise removal of subjectively negligible DCT coefficients. The result justifies the effectiveness of the proposed framework.

1. INTRODUCTION

For the recent two decades, moving image coding has been developed and optimized based on motion compensated hybrid DCT(MC-DCT) coding[1], where characteristics of human perception has been highly utilized for effective lossy coding in several ways such as exchange of spatial and amplitude resolutions by DCT, unequal quantization by a weighting matrix, high-luminance-low-chroma sampling resolution, and so on[2, 3]. Those are based on human perception models. From the viewpoint of speech coding, however, usage of perception model in image coding is different from speech coding. Speech coding with VSELP(Vector Sum Excited Linear Prediction) or CELP(Code Excited Linear Prediction), which searches optimal codes in closed-loop operation, evaluating the coded signal with perceptually weighted mean square error[4] , explicitly use the human perception model. The basic idea proposed in this paper is inspired by such a speech coding framework. In this context, an image coding that has a closed optimization loop can be viewed as in Fig. 1.

Here we have the following two issues when discussing image coder that has a closed loop to optimize subjective image quality.

subjective image quality measure : Subjective image quality is difficult to measure and quantify in general even if not impossible, since the quality is influenced by personal experience and other high-level factors outside of any simple physical description of the image. Nevertheless, we can use a more subjective measure rather than the mean square error metric (MSE) when seeing the success in speech coding. In that sense, as reported in many contributions[5, 6, 7, 8, 9] , there could be many candidates to the distortion evaluator depicted in Fig. 1. What should be the choice?

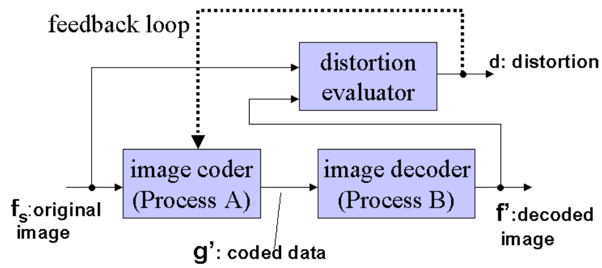


Fig. 1. Image encoder with a feedback loop.

optimized coding parameters : In the speech coding mentioned above, the optimal vector/code is to be selected in a vector quantization process after fixing linear prediction parameters. In MC-DCT image coding, on the other hand, coding parameters are motion vectors, mode selections, quantization steps. Here is a general question; can those parameters be optimized or is there any room to optimize?

In this paper, we use a weighted PSNR ("Masked PSNR"[10] introduced in the next section) for its availability. Please note that the optimality of that metric is beyond the scope of this paper. Preferably we will focus on the optimization framework introduced into image coding area, and point out its possibility to improve coding efficiency in light of the subjective quality. The scope of this paper is described as follows

- The image encoder assumed here is DCT-based. Not specifically, MPEG-4 intra coding [3] will be opti-

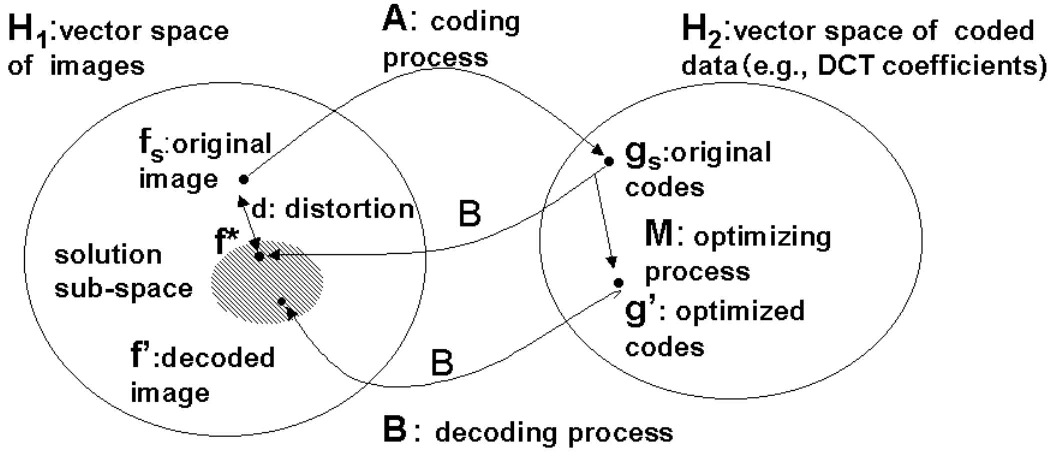


Fig. 2. Relationship between images and coded data.

mized for experiments. Optimization including inter frame coding is of further study.

- Only DCT coefficients are operated while the other parameters are left intact.

Fig. 2 depicts our formulation, where

$\mathbf{H_1}$: image space (N-dimensional vector space).

$\mathbf{H_2}$: coded data space (M-dimensional vector space). For simplification in this paper, suppose coded data $\mathbf{g} \in \mathbf{H_2}$ is identical to quantized DCT coefficients, thus $N = M$ since DCT is a linear orthogonal operator.

coding operator $\mathbf{A}$: $\mathbf{H_1} \rightarrow \mathbf{H_2}$

decoding operator B : $\mathbf{H_2} \rightarrow \mathbf{H_1}$

$\mathbf{f_s} \in \mathbf{H_1}$: original image.

$\mathbf{g_s} \in \mathbf{H_2}$: quantized DCT coefficients coded from $\mathbf{f_s}$.

$\mathbf{f}^*$: decoded image from $\mathbf{g_s}$. $\mathbf{f}^* \neq \mathbf{f_s}$ due to the quantization process in $\mathbf{A}$.

optimization operator M : $\mathbf{H_2} \rightarrow \mathbf{H_2}$

$\mathbf{g}'$: optimized from $\mathbf{g_s}$

$\mathbf{f}'$: decoded image with optimization.

d**()** : a distortion measure in $\mathbf{H_1}$.

$L(\mathbf{g})$: a generic code length function(i.e., a rate function) that gives amount of $\mathbf{g}$.

Given an appropriate distortion measure d, in our proposed framework, we try to seek a better DCT coefficients $\mathbf{g}'$ than the original $\mathbf{g_s}$, which minimizes

$$F(\mathbf{g}';\mathbf{f_s}) = d(\mathbf{f}' - \mathbf{f_s}) - d(\mathbf{f}^* - \mathbf{f_s}) + \lambda(L(\mathbf{g}') - L(\mathbf{g_s})), \tag{1}$$

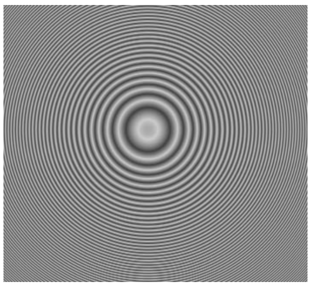
where λ denotes the ratio of distortion:code length. The formula "distortion + λ rate" is is well known in rate-distortion optimization problems, to which for example Ramchandran tried to optimize quantization parameters[11]. The difference is obvious, where we try directly to modify DCT coefficients to better one that maintains subjective image quality similar to the original image $\mathbf{f_s}$ without changing the quantization-step parameters. Suppose the distortion measure $d()$ is MSE, it always follows that any change by $\mathbf{M}$ from the original code causes additional MSE, since MSE itself is a Euclid distance in $\mathbf{H_1}$, the operators $\mathbf{A}$ and $\mathbf{B}$ are unitary transforms. It seems less possible to improve as far as we adhere to MSE. Instead we try to use an appropriate subjective quality measure for $d()$.

The original contribution of this paper consists of
(1) a new framework that has a *feedback loop to optimize the DCT coefficients* and to attain an original image quality, and
(2) experimental justification of the proposed framework. As far as we know, similar challenge in the image coding field has not been reported.

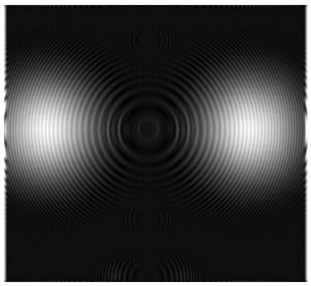
2. SUBJECTIVE QUALITY MEASURE: MPSNR

Although our proposed Masked PSNR is one of candidates, it is worth to describe how $d()$ works in our proposed framework. It is known that the presence of a background stimulus modifies the perception of a foreground stimulus: masking corresponds to a modification of the detection threshold of the foreground according to the local contrast of the background. Thus original image pattern can mask the distortion to some extent. We utilize the masking effect and realize it as a multichannel metrics where the original image and distortion are filtered independently by a collection of filters. Each filter is pooled to provide a separate channel. The channels are then weighted and pooled together with a different function. A collection of independent channels spans the frequency plane, partitioning the plane into frequency- and orientation selective bands. For implementation, we use a bank of four Gabor filters[12] tuned to two spatial frequencies (0 and 1/4 sampling frequency) and four orientations[10].

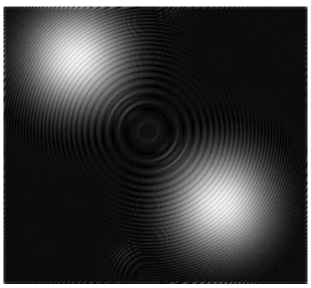
The use of Gabor filters is advantageous in terms of their phase invariance to edge position, because a Gabor filter is of quadrature mirror filters (consisting of a Hilbert's pair). A Gabor filter outputs localized directional spatial frequency. Fig. 3 shows three filters of our implementation.



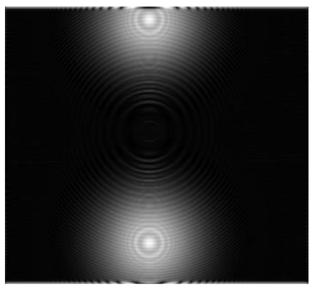
(a) Zone plate as input.



(b) Gabor response of 0 deg.



(c) Gabor response of 45 deg.



(d) Gabor response of 90 deg.

Fig. 3. Zone plate and its responses from Gabor filters.

By applying Gabor filters, we decompose the original input image as

$$\mathbf{f_s} \rightarrow \{\mathbf{m_0}, \mathbf{m_{45}}, \mathbf{m_{90}}, \mathbf{m_{135}}\}, \text{ where} \tag{2}$$

$\mathbf{m_0}, \mathbf{m_{45}}, \mathbf{m_{90}}, \mathbf{m_{135}}$ are band-passed response images with four orientations.

Letting $(\mathbf{f} - \mathbf{f_s})$ be $\mathbf{D}(\mathbf{f}; \mathbf{f_s})$, we also apply five Gabor filters to errors as

$$\mathbf{D}(\mathbf{f}; \mathbf{f_s}) \rightarrow \{\mathbf{D_{DC}}, \mathbf{D_0}, \mathbf{D_{45}}, \mathbf{D_{90}}, \mathbf{D_{135}}\}, \text{ where} \tag{3}$$

$\mathbf{D_{DC}}, \mathbf{D_0}, \mathbf{D_{45}}, \mathbf{D_{90}}, \mathbf{D_{135}}$ are component error images (i.e., zero frequency tuned and the previous fours) to the original image and satisfy

$$\mathbf{D}(\mathbf{f}; \mathbf{f_s}) = \mathbf{D_{DC}} + \mathbf{D_0} + \mathbf{D_{45}} + \mathbf{D_{90}} + \mathbf{D_{135}}. \tag{4}$$

In a uniform gray field, we do not need to model visual masking for $\mathbf{D}_{DC}$. For the others, masking occurs when a high contrast pattern is on or near the stimulus. Here we assume that each of $\mathbf{m_0}, \mathbf{m_{45}}, \mathbf{m_{90}}, \mathbf{m_{135}}$ is "maskers" to the distortion that has same localized directional spatial frequency. Thus each of $\mathbf{D_0}, \mathbf{D_{45}}, \mathbf{D_{90}}, \mathbf{D_{135}}$ should be attenuated by the ratio of $\mathbf{m_0}, \mathbf{m_{45}}, \mathbf{m_{90}}, \mathbf{m_{135}}$. While letting us set aside the detail[10], we calculate the suppression weight image as $\{\bar{\mathbf{m_0}}, \bar{\mathbf{m_{45}}}, \bar{\mathbf{m_{90}}}, \bar{\mathbf{m_{135}}}\}$ as the inverse ratios.

$$\begin{gathered}\text{Masked } \mathbf{D}(\mathbf{f}; \mathbf{f_s}) = \mathbf{D_{DC}} + \bar{\mathbf{m_0}} \otimes \mathbf{D_0} + \bar{\mathbf{m_{45}}} \otimes \mathbf{D_{45}} \\ + \bar{\mathbf{m_{90}}} \otimes \mathbf{D_{90}} + \bar{\mathbf{m_{135}}} \otimes \mathbf{D_{135}}, \text{ where}\end{gathered} \tag{5}$$

$\otimes$ is a pixel-by-pixel attenuation operation for each image. Letting

$$d(\mathbf{f} - \mathbf{f_s}) = |\text{Masked } \mathbf{D}(\mathbf{f}; \mathbf{f_s})|^2, \tag{6}$$

we can define Masked PSNR as

$$\text{Masked PSNR} = 10log_{10}\frac{255^2}{d(\mathbf{f} - \mathbf{f_s})/N}. \tag{7}$$

3. OPTIMIZATION PROCESS AND EXPERIMENT

Now let us discuss the optimization operator $\mathbf{M}$. Similar to vector quantization in the speech coding, improvement of the rate-distortion characteristics could be achieved by *searching* coefficients that significantly minimizes $F(\mathbf{g}'; \mathbf{f_s})$ in Eq. 1. One of simple operations for the rate-distortion optimization in DCT-based coding is to carefully remove a higher DCT coefficient which is unnecessary to maintaining the image quality, in a step-by-step manner (i.e., hill climb manner). It is a heuristic algorithm, assuming that higher frequency coefficients considered to have low impacts for image quality. The optimization process can be summarized as follows. λ in Eq. 1 is obtained beforehand through the experiments. λ corresponds to the slope of MPSNR curve as shown in Fig. 5. For a decoded image $\mathbf{f}$, $d(\mathbf{f} - \mathbf{f_s})$ in Eq. 6 is first calculated. Then quantized DCT coefficients are processed in the closed loop for every 8x8 pixel DCT block. In that process, one of DCT coefficients are tentatively removed, and its effect on the number of bits and $d(\mathbf{f} - \mathbf{f_s})$ of the reconstructed image are evaluated. If such operation improves the rate-distortion criterion in Eq. 1, the removal is adopted. This optimization process is repeated as long as such profitable removal is possible. The above reduction of the bits that represents subjectively less important image is performed to all of the 8x8 DCT blocks of the image repeatedly.

The removal of the DCT coefficients must introduces the additional disturbance to the original image quality. Fig. 4 shows the additional distortion introduced by the optimization process. Please note that the additional distortion is not noticeable because the distortion has same localized directional spatial frequency to the background stimulus, i.e. the maskers. In this experiment, for example we attained approximately 7% improvement of the bit rate from the original one at the same MPSNR; 18520 bit at 36.96dB was reduced to 17208 bit at 36.97dB. Fig. 5 shows the rate-distortion improvement.

(a) Original result(QP=10) (B) Optimized result(QP=10)

(C) Added distortion to original result (magnified 5times)

Fig. 4. Original result, and its optimized one, and their difference.

4. CONCLUSION

To conclude this paper, let us question ourselves "Is our coding result of DCT coefficients optimal?, especially in light of subjective quality?". This paper tried to point out that the further optimization is plausible with the closed-loop optimization framework. On the other hand, many issues are left intact. One is an optimization process. A hill-climb method is inevitable in our formulation, since Eq. 1 can not be differentiated with respect to DCT coefficients. Further study is needed. Alternatives to MPSNR, obviously, should be developed and MPSNR itself is not sufficient to model the human perception. In this paper, we just started a closed-loop optimization in image coding. By incorporating more sophisticated spatio-temporal masking effects with a more effective optimization process, encoding techniques such as

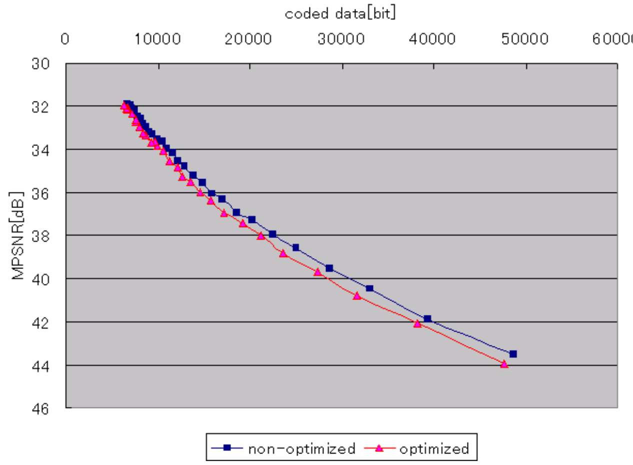


Fig. 5. Rate-distortion curve in MPSNR.

pre-processing and quantization schedule can be treated in the same framework.

5. REFERENCES

[1] H. Li, A. Lundmark, and R. Forchheiner, "Image sequence coding at very low bitrates:a review," *IEEE Trans. on Image Processing*, vol. 3, no. 5, pp. 589–609, Sept. 1994.

[2] D. Le Gall, "MPEG:a video compression standard for multimedia applications," *Commun. ACM*, vol. 34, no. 4, pp. 46–58, 1991.

[3] MPEG-4 visual, "Final draft international standard 14496-2," ISO/IEC JTC1/SC29/WG11 N2502, Oct. 1998.

[4] B. Atal and M Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-27, pp. 247–254, 1979.

[5] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Proc. SPIE*, 1992, vol. 1616.

[6] P. Teo and D. Heeger, "Perceptual image distortion," in *Proc. ICIP*, Nov. 1994, number lS-16.

[7] S Hangai, K. Suzuki, and K. Miyauchi, "Advanced wsnr for coded monochrome picture evaluation using fractal dimension," in *PCS'94*, 1994, pp. 92–95.

[8] J. Farrell, X. Zhang, C. van den Branden Lambrecht, and D. Silverstein, "Image quality metrics based on single and multi-channel models of visual processing," in *IEEE Proc. COMPCON '97*.

[9] A. Rohaly, P. Corriveau, L. John, A. Webster, V. Baroncini, J. Beerends, J. Blin, L. Contin, T. Hamada, D. Harrison, A. Hekstra, J. Lubin, Y. Nishida, R. Nishihara, J. Pearson, A. Pessoa, N. Pickford, A. Schertz, M. Visca, A. Watson, and S. Winkler, "Video quality experts group:current results and future directions," in *Proc.SPIE Visual Communications and Image Processing*, 2000, vol. 4067.

[10] M. Takahashi, "Quality improvement of wartermarked digital images by masking effect on visual disturbances," M.S. thesis, Nara Institute of Science and Technology, 2000, in Japanese.

[11] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and mpeg video coders," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 533–545, Sept. 1994.

[12] D. Gabor, "Theory of communication," *J. Institute of Elec. Engr.*, 1947.