PERCEPTUAL IMAGE CODING BASED ON A MAXIMUM OF MINIMAL STRUCTURAL SIMILARITY CRITERION

Zhou Wang[1], *Qiang Li*[1] *and Xinli Shang*[2]

[1]Dept. of Electrical Engineering, Univ. of Texas at Arlington, Arlington, TX 76019, USA
[2]Microsoft Corporation, Redmond, WA 98052, USA
Emails: zhouwang@ieee.org; qxl7033@exchange.uta.edu; xinlis@microsoft.com

ABSTRACT

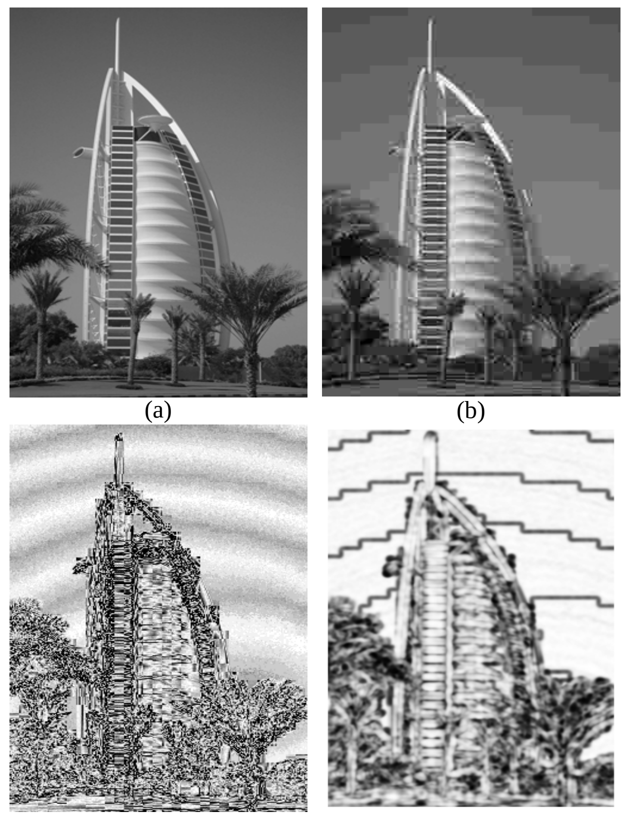
Perceptual image coding algorithms typically impose perceptual modeling in a preprocessing stage. A perceptual normalization model is often used to transform the original image signal into a perceptually uniform space, in which all the transform coefficients have equal perceptual importance. Standard coding schemes are then applied uniformly to all coefficients. Here we use a different approach, in which we iteratively reallocates the available bits over the image space based on a *maximum of minimal structural similarity criterion*. We demonstrate the proposed method by incorporating it with the bitplane coding scheme in the set partitioning in hierarchical trees algorithm.

Index Terms— perceptual image coding, image quality assessment, structural similarity (SSIM), bitplane coding, set partitioning in hierarchical trees (SPIHT)

1. INTRODUCTION

Image coding algorithms have been traditionally optimized to achieve the minimal mean squared error (MSE) under the constraint of a limited bit budget. However, MSE has been widely criticized for poorly correlating with visual perception of image quality [1]. An example is shown in Fig. 1, where a JPEG compressed image is evaluated locally to create the the absolute error map and the structural similarity (SSIM) index [2] map. Both maps use brighter pixels to indicate better quality, but they give substantially different evaluations. Careful inspection of the distorted image together with the quality maps concludes that absolute error (which is the basis for all Minkowski error metrics, including MSE) is not a good indicator of local image quality when compared with the SSIM index (e.g., the SSIM map clearly points out the annoying blocking artifacts in the sky).

The poor performance of MSE motivated researchers to incorporate perceptual models in image coding [3]. Most perceptual coding methods first decompose the image signal using a linear transform (e.g., a DCT or a wavelet transform) and then normalize (rescale) each transform coefficient with a *perceptual weight* before a *uniform* quantization and entropy coding scheme is applied. These weights may be determined by a number of psychophysical features of the human visual system (HVS), typically including the contrast sensitivity function and the contrast masking effect [1, 3]. An equivalent method is to design a *nonuniform quantization* scheme, where the quantization steps of the transform coefficients are proportional to their perceptual weights. This general design prin-



(a) (b) (c) (d)

Fig. 1. (a) Original image; (b) distorted image (by JPEG compression); (c) absolute error map − brighter indicates better quality (smaller absolute difference); (d) SSIM index map − brighter indicates better quality (larger SSIM value).

 © IEEE 2007

ciple has been employed in many existing algorithms (e.g., [3–6]), with variations in the linear transforms being used and the way the perceptual weights are computed. It has also been used in the design of the JPEG quantization table [7] and the visual optimization tools in JPEG2000 [8, 9]. This approach is appealing because it completely separates perceptual modeling from the subsequent processes, making it convenient to implement. Nevertheless, its accuracy is questionable. For example, when the masking effect is considered at the encoder side, the perceptual weight of a given coefficient is computed from its original neighboring coefficients. However, after the subsequent nonlinear quantization process, the coefficient and all of its neighboring coefficients have been changed at the decoder side. As a result, the computed masking effect and the corresponding perceptual weight at the decoder would not be accurate anymore.

In this paper, we use a different approach. First, we use the SSIM index map as a local perceptual quality indicator, which, to the best of our knowledge, has not been used directly for image coding before. Second, we do not attempt to impose perceptual modeling using one single normalization process. Instead, we encode the image iteratively. Within each iteration, we operate on the bit allocation scheme that redistributes the available bits over the image space according to the SSIM quality map obtained from the last iteration. Third, our scheme aims to improve the worst case scenario, such that the quality at the lowest quality region in the image is enhanced. In other words, we use a *maximum of minimal structural similarity criterion* as our optimization goal. This is justified based on the observation that human visual attention is often attracted to the image regions with extremely annoying artifacts (very low quality) that could dominate the quality evaluation of the entire image.

2. METHOD

The central idea of our method is to iteratively redistribute the available bits based on local image quality measures. We find that an easy way to implement the idea is to incorporate it into an embedded bitplane coding algorithm. Embedded bitplane coding [8, 10, 11] has received wide acceptance in the past fifteen years. It encodes images into continuously scalable bit streams that can be truncated at arbitrary places to create multiple versions of decoded images with variable bit rate and quality. Moreover, the encoded information bits are naturally organized according to their importance. Figure 2(a) provides a simple illustration of a regular bitplane coding scheme. The image components (typically wavelet coefficients) are binary represented and aligned to bitplanes, and the bitplanes are scanned and coded from the most significant bitplane (MSB) to the least significant bitplane. The scanning and coding process may stop at any place when a target bit budget is reached. This is equivalent to setting all the remaining bits to zero.

The bitplane coding scheme is flexible in the sense that the importance of image components (coefficients) can be easily adjusted. There are two ways to accomplish this. The first approach emphasizes the important coefficients by shifting them up in the bitplane representation (or equivalently, shifting the unimportant coefficients down). Examples include the MaxShift [8] and the BbBShift [12] methods. The second approach, which we use in this paper, is the bitplane-trimming method demonstrated in Fig. 2(b), where the coefficients are trimmed from the bottom such that the bits below certain level are all set to zero. The trimming level is variable based on the importance of the coefficient. This is equivalent to quantizing the coefficient at that level. A regular bitplane coding scheme is applied to the trimmed coefficients, leading to a variable bit allocation over the image space. One advantage of this method is that the decoder does not need to reconstruct the trimming function, and thus no overhead is needed to encode the information about the trimming function.

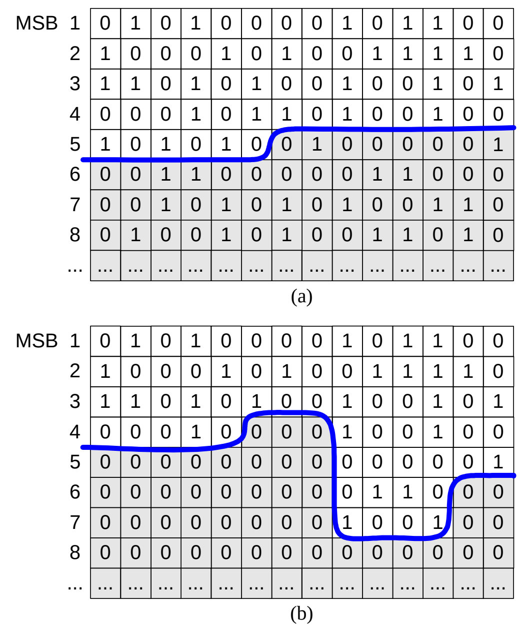


Fig. 2. (a) Regular bitplane coding. Bitplanes are scanned and coded until a target bitrate is reached. The result is equivalent to setting all bits in the gray region to zero; (b) Bitplane-trimming based coding. The gray region is set to all zero before regular bitplane coding.

Let $\mathbf{x}$ and $\mathbf{y}$ be the original and the decoded images respectively. Let $\mathbf{t}$ denotes the trimming function, i.e., it is a function of the coefficient index that defines the trimming level of all coefficients. Let $\mathbf{T}$ be the set of all possible trimming functions. Let $\mathbf{C}$ represent the entire image encoding and decoding operator that takes a given original image $\mathbf{x}$, a given bit rate R, and a given trimming function $\mathbf{t}$ as the input, and creates a decoded image $\mathbf{y}$ as the output:

$$\mathbf{y} = \mathbf{C}(\mathbf{x}, R, \mathbf{t})\,. \tag{1}$$

Let $S_{\mathbf{x},\mathbf{y}}$ denote the operator that computes the SSIM index map between $\mathbf{x}$ and $\mathbf{y}$, and thus $S_{\mathbf{x},\mathbf{y}}(n)$ represents the SSIM index value at spatial location n. Under the maximum of minimal structural similarity criterion, our task is to find the best trimming function that maximizes the minimal value in the SSIM index map $S_{\mathbf{x},\mathbf{y}}$. Mathematically, this can be expressed as

$$\mathbf{t}_{opt} = \underset{\mathbf{t}\in\mathbf{T}}{\operatorname{argmax}}\{\min_{n}[S_{\mathbf{x},\mathbf{C}(\mathbf{x},R,\mathbf{t})}(n)]\}\,. \tag{2}$$

Since the minimal SSIM value in an image has a highly nonlinear relationship with respect to the trimming function $\mathbf{t}$, it is difficult to find the optimal solution $\mathbf{t}_{opt}$ analytically. Here we propose an iterative approach given as follows:

1. Initiate the iteration number $i = 0$. For the given target bit rate R, create a constant initial trimming function $\mathbf{t}_0$, i.e., the trimming level (bitplane) is uniform for all coefficients. The trimming level should be high enough such that the bit rate needed to encode all bits above the level is lower than R.

2. Encode and decode the image to create $\mathbf{y}_i = \mathbf{C}(\mathbf{x}, R, \mathbf{t}_i)$.

3. Compute the SSIM map $S_{\mathbf{x},\mathbf{y}_i}$ between the original and the decoded image.

4. Find the minimal value and location in $S_{\mathbf{x},\mathbf{y}_i}$. If the minimal SSIM (Min-SSIM) value does not change for several iterations, stop the iteration and report $\mathbf{t}_i$ as the optimized trimming function. Otherwise, update $\mathbf{t}_i$ by adding one more bits for all the coefficients around the spatial location corresponding to the Min-SSIM value (In wavelet domain, these include a set of neighboring coefficients in all subbands). Let $i = i + 1$ and go to Step 2.

The process is guaranteed to converge when the trimming function goes deep enough to saturate all the bits available for a given bit rate. Figure 3 gives a demonstration about how the minimal SSIM value is updated over iterations and how the iterative algorithm converges.

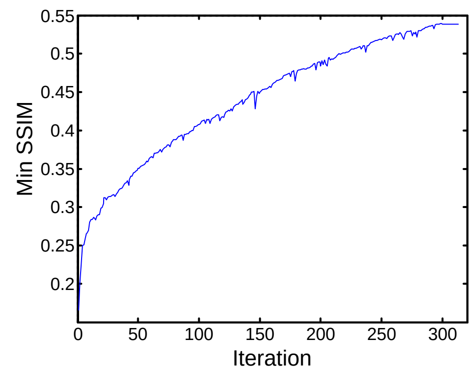


Fig. 3. Min-SSIM as a function of iteration for the "Lighthouse" image coded at 0.5bits/pixel.

3. TEST

We test the proposed approach using 8bits/pixel gray scale images. The set partitioning in hierarchical trees (SPIHT) algorithm is used as the basic bitplane encoding and decoding operator $\mathbf{C}$. The images are coded to a range of bit rates, from 0.2 to 0.9 bits/pixel using both SPIHT and the proposed method. The Min-SSIM results for the "Lighthouse" image are shown in Fig. 4. It can be seen that the proposed method achieves significantly higher Min-SSIM values than SPIHT over a wide range of bit rates. Similar results are also obtained for the other images being tested.

In Fig. 5, we compare the coding results of the "Lighthouse" image provided by SPIHT and the proposed algorithms at 0.2bits/pixel, respectively. The SSIM maps indicate that the quality of the image coded by the proposed method is more uniformly distributed over the image space than that of the SPIHT coded image. Since the proposed method mainly focuses on the worst case scenario, the regions with the worst quality in the SPIHT coded image obtain the most improvement. For better visualization, we also enlarged two regions in the images. It can be observed that more detailed structures are exhibited in the image coded by the proposed method.

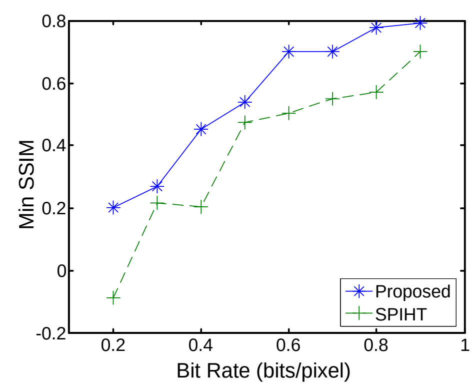


Fig. 4. Min-SSIM comparison of SPIHT and the proposed method for "Lighthouse" image coded at different bit rates.

4. CONCLUSION

We propose a novel perceptual coding method that incorporates a maximum of minimal SSIM criterion into bitplane coding through an iterative optimization process. The test results show that the proposed method significantly improves the worst case scenario (worst quality region in the image) and the coded image appears to have more uniform quality over the image space. Although the proposed scheme is currently implemented with the SPIHT algorithm only, the general design principle may be generalized for other bitplane

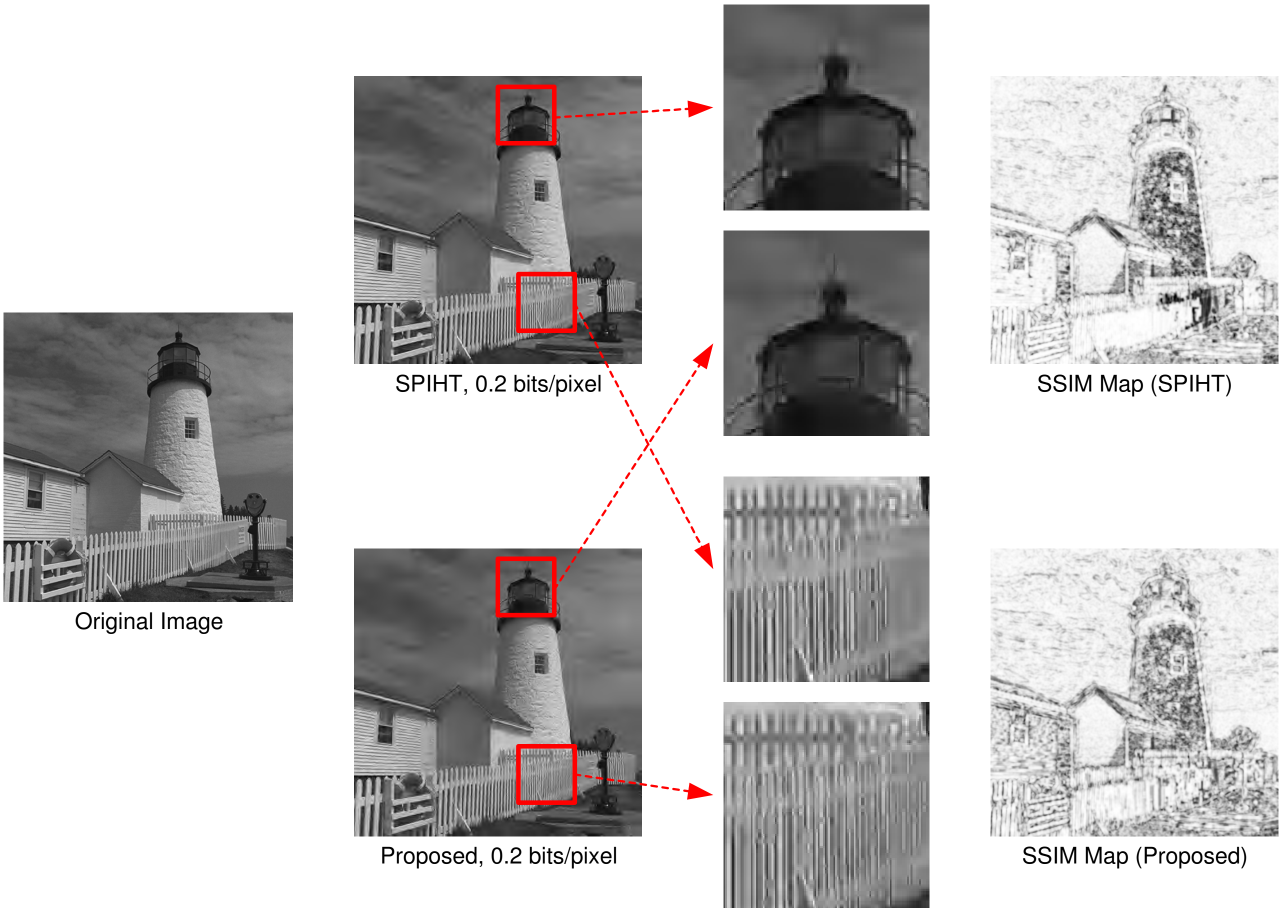


Fig. 5. Comparison of coding results by SPIHT and the proposed algorithms at 0.2bits/pixel.

coding schemes. The increased computational complexity due to repeated coding may be reduced by effectively estimating the trimming function before the initial iteration.

5. REFERENCES

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool Publishers, Mar. 2006.

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

[3] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.* (A. Bovik, ed.), 2nd ed., Academic Press, 2005.

[4] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 1945–1948, May 1989.

[5] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, Aug. 1997.

[6] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in *Human Vision and Electronic Imaging VII, Proc. SPIE*, vol. 4662, pp. 742–753, Jan. 2002.

[7] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. Kluwer Academic Publishers, 1992.

[8] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[9] W. Zeng, S. Daly, and S. Lei, "Visual optimization tools in JPEG 2000," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 37–40, Oct. 2000.

[10] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[11] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, pp. 243–250, June 1996.

[12] Z. Wang and A. C. Bovik, "Bitplane-by-bitplane shift (BbB-Shift) - A suggestion for JPEG 2000 region of interest coding," *IEEE Signal Processing Letters*, vol. 9, pp. 160–162, May 2002.