G-CAST: Gradient Based Image SoftCast for Perception-Friendly Wireless Visual Communication

Ruiqin Xiong[1], Hangfan Liu[1], Siwei Ma[1], Xiaopeng Fan[2], Feng Wu[3] and Wen Gao[1]

[1]Institute of Digital Media, Peking University, Beijing 100871, China

[2]Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China

[3]Microsoft Research Asia, Beijing 100080, China

Email: rqxiong@pku.edu.cn

Abstract

Conventional image and video communication systems are usually designed with the objective being to maximize the fidelity of reconstructed images measured by mean square errors (MSE). It is well known that the fidelity metric MSE may not reflect the visual quality perceived by human eyes. Recent advancements in image quality assessment tell us that the structural similarity (SSIM), especially the gradient similarity, reveals the perceptual fidelity of images more reliably. Inspired by this observation, this paper proposes a new image communication approach, which conveys the visual information in an image by transmitting the image gradients and recovers the image from the received gradient data at decoder side using statistical image prior knowledge. In particular, we designed a gradient-based image SoftCast scheme for wireless scenarios. Experimental results show that the proposed scheme can produce reconstruction images with much better perceptual quality. The advantage in perceptual quality is verified by the quality improvement measured by the metrics SSIM and gradient signal-to-noise ratio (GSNR).

1 Introduction

Today, mean square error (MSE) is still widely used as the fidelity measure for the design and optimization of image communication systems. For example, to predict a pixel block from the neighboring pixels in already-reconstructed blocks, the prediction mode is usually selected in such a way that minimal square prediction error is achieved. Similarly, the rate-distortion optimization (RDO) is generally performed using MSE as the distortion measure. In addition, we may interpret the adoption of orthogonal (or nearly orthogonal) decorrelation transform (e.g. DCT or DWT) in the existing image and video coding schemes as an example of using the MSE metric implicitly. This is because an orthogonal transform keeps the MSE distortion unchanged so that a good approximation in the transform domain is guaranteed to be a good approximation in the signal domain, in terms of MSE.

MSE may be a very good distortion metric indeed for some signal processing tasks. However, it exhibits weak performance in some other applications and has been widely criticized for serious shortcomings, especially when dealing with perceptually important signals [1, 2]. Most pictures, still or moving, are meant to be viewed by people

This work was supported in part by the National Natural Science Foundation of China (61073083, 61370114, 61121002), Beijing Natural Science Foundation (4112026, 4132039) and Research Fund for the Doctoral Program of Higher Education (20100001120027, 20120001110090).

1068-0314/14 $31.00 © 2014 IEEE
DOI 10.1109/DCC.2014.55

IEEE computer society

ultimately. Accordingly, the fidelity metric for image communication should be tailored to the properties of human visual perception. Recently, a great deal of effort has been put into the development of image quality assessment (IQA) methods. The famous work of Wang *et. al.* [3] tells us that the structural similarity (SSIM) index is more correlated with the perceptual image quality than the widely used MSE metric. In SSIM, the distortion/similarity of two signals is modeled by three components, namely, the luminance similarity, contrast similarity and structural similarity. The more recent works [4] and [5] suggest that image gradients convey important visual information and hence gradient similarity is more relevant to image quality assessment.

Based on the advancements in image quality assessment, some perceptual image/video coding schemes were proposed in literatures (e.g. [6, 7]). However, these schemes usually focus on parameter selection or coding mode optimization, and do not change the coding framework and coding tools. Inspired by the gradient-based IQA metrics, this paper proposes a new image communication approach, which conveys the visual information in an image by transmitting the image gradients. The approach reconstructs the image from the noisy gradient data at the receiver side, using statistical prior knowledge about the gradient data or the image itself. In particular, we consider the wireless communication scenario and design a gradient-based image SoftCast scheme. The proposed scheme has the same advantages as SoftCast [8–10] that it achieves graceful quality transition for very wide channel SNR range, and it can serve multiple receivers with different channel qualities simultaneously using a single signal transmission. More importantly, we expect the proposed scheme to achieve better perceptual quality, since the scheme is designed with the objective being to reproduce (as accurately as possible) the image gradients, which is believed to highly relevant to the visual information perceived by human eyes.

The remainder of the paper is organized as follows. Section 2 describes the proposed gradient-based image SoftCast scheme. Section 3 describes the gradient-based image reconstruction algorithm. Experimental results are reported in Section 4 and Section 5 concludes the paper.

2 Gradient Based Image SoftCast (G-Cast)

2.1 The G-Cast Sender

The proposed G-Cast scheme transmits an image over a noisy wireless channel using one base-layer and one enhancement layer. The purpose of the base layer is to deliver the DC and the low-frequency components of the image so that the receiver knows the global luminance of the whole image and the local luminance of each region. The base layer also provides a coarse reconstruction of the image which can be refined by the gradient information in the enhancement layer. The base layer will be coded into a binary representation of very low bit rate and sent out using digital transmission techniques with strong protection so that the receiver can get the base layer with very high probability, even when the channel SNR is very low. The purpose of the enhancement layer is to deliver the gradient information of the image so that a viewer

can observe the visual details. As stated before, the scheme strives to reproduce the image gradients instead of the pixel intensities, as accurately as possible at the receiver side. The enhancement layer will be represented by a stream of real numbers and sent out using semi-analog transmission techniques, which can achieve graceful quality transition when the channel SNR fluctuates.

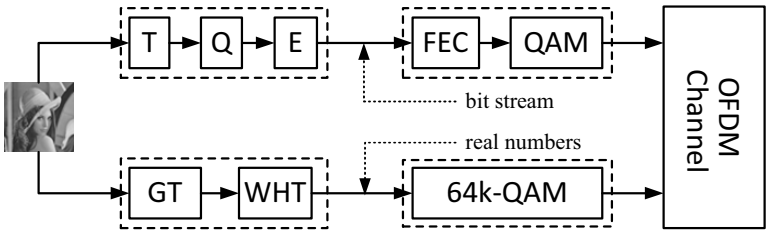

Figure 1: The G-Cast Sender.

The detail of the proposed scheme is illustrated in Fig. 1. In the base layer, the whole input image is first transformed into frequency domain using discrete Fourier transform (DFT) or discrete cosine transform (DCT). Then a small set, say $M \times M$, of low-frequency coefficients are retained while all the other coefficients are discarded. The retained coefficients are then quantized and encoded into a bitstream using entropy coding techniques. An alternative coding scheme for base layer, if DCT is used to produce the coefficients, is to convert the retained $M \times M$ low-frequency coefficients back to a small $(M \times M)$ image using an inverse DCT of size $M \times M$ (see [11, 12]), and then code the small image to a high quality using a state-of-the-art image coder (such as HEVC Intra). The coded bitstream is then sent to the OFDM module for transmission, using FEC codes for error protection and quadrature amplitude modulation (QAM) (such as BPSK, QPSK, etc.) for modulation.

In the enhancement layer, the image gradient is first extracted from the input image using a gradient transform (GT). Then the gradient image (as shown in Fig. 2) is processed by Walsh-Hadamard transform (WHT) to reduce the peak-to-mean ratio (PMR) of the gradient data stream. The WHT-transformed gradient data is directly modulated to a dense constellation (e.g. 64k-QAM) for raw OFDM transmission, in the same way as done in SoftCast [8–10]. For each OFDM sub-carrier, a pair of numbers is extracted from the WHT-transformed gradient data stream and mapped to a point in the constellation, using the two numbers as the I- and the Q- components respectively, which ultimately controls the amplitude and phase of the sub-carrier.

2.2 The G-Cast Receiver

Signals transmitted in the air will be influenced by the interferences from other transceivers. The G-Cast receiver recovers the image from the noisy OFDM signal via a base layer decoder and an enhancement layer decoder, as illustrated in Fig. 3. The base layer decoder reconstructs the DC and low-frequency coefficients of the image (or equivalently, the corresponding low-resolution image), by first recovering the base layer bitstream via demodulation and FEC decoding and then performing entropy decoding and dequantization. The enhancement layer decoder first retrieves

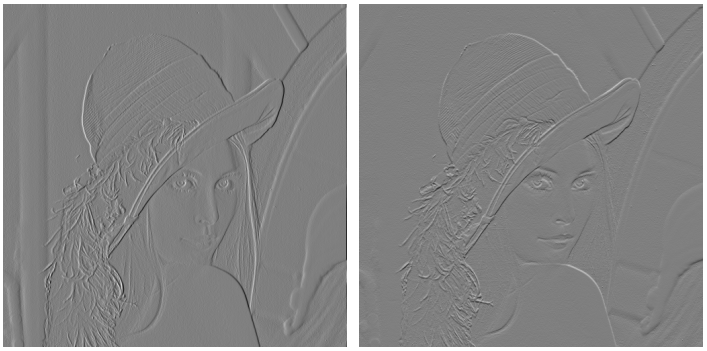

Figure 2: The gradient image of *Lena* (gray, 512×512). Left: the horizontal gradient. Right: the vertical gradient. The data for both images are shifted by +128 for display.

the gradient data from the noisy OFDM signal via demodulation and inverse WHT transform. It then creates a final estimation of the image via a gradient based reconstruction (GBR) procedure, utilizing both the gradient information at the enhancement layer and the low-frequency coefficients provided by the base layer. The GBR procedure will be described in the next section.

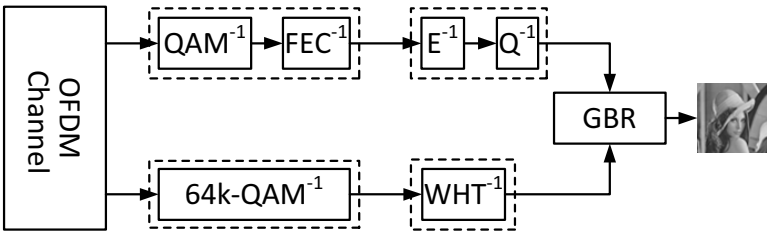

Figure 3: The G-Cast Receiver.

3 Gradient Based Image Reconstruction

3.1 The Overall Reconstruction Procedure

Suppose $\mathbf{u}$ is the lexicographically stacked representation of the original image, $\mathbf{g}^{(\mathrm{h})} = D^{(\mathrm{h})}\mathbf{u}$ and $\mathbf{g}^{(\mathrm{v})} = D^{(\mathrm{v})}\mathbf{u}$ are the horizontal and the vertical gradient of $\mathbf{u}$, respectively. The matrices $D^{(\mathrm{h})}$ and $D^{(\mathrm{v})}$ represent the gradient operators in the horizontal and the vertical directions, respectively. Suppose $\mathbf{m} = T^{(\mathrm{L})}\mathbf{u}$ denotes the low-frequency coefficients of $\mathbf{u}$, where T is the DFT or DCT transform matrix and $T^{(\mathrm{L})}$ is the $M \times M$ rows of T, corresponding to the retained $M \times M$ low-frequency coefficients. The G-Cast sender encodes $\mathbf{m}$ in the base layer and transmits $\mathbf{g}^{(\mathrm{h})}$ and $\mathbf{g}^{(\mathrm{v})}$ in the enhancement layer. The G-Cast receiver gets $\tilde{\mathbf{m}}$, $\tilde{\mathbf{g}}^{(\mathrm{h})}$ and $\tilde{\mathbf{g}}^{(\mathrm{v})}$ due to the quantization effect and the existence of channel noises. For the convenience of discussion, we write $D = [D^{(\mathrm{h})}\ D^{(\mathrm{v})}]$, $\mathbf{g} = [\mathbf{g}^{(\mathrm{h})}\ \mathbf{g}^{(\mathrm{v})}]$ and $\tilde{\mathbf{g}} = [\tilde{\mathbf{g}}^{(\mathrm{h})}\ \tilde{\mathbf{g}}^{(\mathrm{v})}]$. For any vector $\mathbf{v}$ (or matrix V), we use $\mathbf{v}_i$ (or V_i) to represent its i^{th} element (or row).

The GBR procedure recovers $\mathbf{u}$ from $\tilde{\mathbf{g}}^{(\mathrm{h})}$, $\tilde{\mathbf{g}}^{(\mathrm{v})}$ and $\tilde{\mathbf{m}}$. To reduce the influence of channel noise, this paper considers the statistical feature that gradient data usually

conforms to zero-mean Laplacian distributions. The optimization objective can be formulated as:

$$\min_{\mathbf{u}} \frac{\mu_1}{2}\sum_i \|D_i\mathbf{u}\|_1 + \frac{\mu_2}{2}\left(\left\|D^{(\mathrm{h})}\mathbf{u}-\tilde{\mathbf{g}}^{(\mathrm{h})}\right\|_2^2 + \left\|D^{(\mathrm{v})}\mathbf{u}-\tilde{\mathbf{g}}^{(\mathrm{v})}\right\|_2^2\right) + \frac{\mu_3}{2}\left\|T^{(\mathrm{L})}\mathbf{u}-\tilde{\mathbf{m}}\right\|_2^2 \tag{1}$$

with $\mu_1 = \frac{2\sqrt{2}}{\sigma_\Delta}$, $\mu_2 = \frac{1}{\sigma_n^2}$, $\mu_3 = \frac{12}{Q^2}$. Here σ_Δ^2 and σ_n^2 are the variance of gradient data and the channel noise, respectively, Q is the quantization step for coding $\mathbf{m}$.

This above optimization problem is complicated and difficult to solve directly. But it can be solved with the assistance of variable splitting and augmented Lagrangian methods (see [13] and the references therein). Using the auxiliary variable $\mathbf{g} = D\mathbf{u}$, (1) can be reformulated as

$$\min_{\mathbf{u},\mathbf{g}} \frac{\mu_1}{2}\sum_i \|\mathbf{g}_i\|_1 + \frac{\mu_2}{2}\|\mathbf{g}-\tilde{\mathbf{g}}\|_2^2 + \frac{\mu_3}{2}\left\|T^{(\mathrm{L})}\mathbf{u}-\tilde{\mathbf{m}}\right\|_2^2 \quad \text{s.t. } \mathbf{g} = D\mathbf{u} \tag{2}$$

The augmented Lagrangian function for (2) is

$$\begin{aligned} J(\mathbf{u},\mathbf{g}) = {} & \frac{\mu_1}{2}\sum_i \|\mathbf{g}_i\|_1 + \frac{\mu_2}{2}\|\mathbf{g}-\tilde{\mathbf{g}}\|_2^2 + \frac{\mu_3}{2}\left\|T^{(\mathrm{L})}\mathbf{u}-\tilde{\mathbf{m}}\right\|_2^2 \\ & + \sum_i\left(\frac{\beta}{2}\|D_i\mathbf{u}-\mathbf{g}_i\|_2^2 - \mathbf{v}_i^{\mathrm{T}}(D_i\mathbf{u}-\mathbf{g}_i)\right) \end{aligned} \tag{3}$$

Here the Lagrangian variable $\mathbf{v}$ has the same dimension with $\mathbf{g}$. The problem (2) can be solved by an iterative algorithm, minimizing (3) with respect to $\mathbf{u}$ and $\mathbf{g}$ and then updating $\mathbf{v}_i$ by $\mathbf{v}_i \leftarrow \mathbf{v}_i - \beta(D_i\mathbf{u}-\mathbf{g}_i)$ in each iteration. The minimization of (3) can be easily handled by solving the following two sub-problems.

3.2 The $\mathbf{g}$ *Sub-problem*

With $\mathbf{u}$ fixed, the problem (3) is reduced to

$$\min_{\mathbf{g}} \sum_i \left(\frac{\mu_1}{2}\|\mathbf{g}_i\|_1 + \frac{\mu_2}{2}\|\mathbf{g}_i-\tilde{\mathbf{g}}_i\|_2^2 + \frac{\beta}{2}\left\|D_i\mathbf{u}-\mathbf{g}_i-\frac{\mathbf{v}_i}{\beta}\right\|_2^2\right) \tag{4}$$

which has a closed-form minimizer

$$\mathbf{g}_i = \mathit{Shrink}\left(\frac{\mu_2\tilde{\mathbf{g}}_i + \beta(D_i\mathbf{u}-\frac{\mathbf{v}_i}{\beta})}{\mu_2+\beta}, \frac{\mu_1}{2(\mu_2+\beta)}\right) \tag{5}$$

Here the shrinkage operation is defined by $\mathit{Shrink}(\mathbf{v},s) = \max(\|\mathbf{v}\|-s,0)\cdot\frac{\mathbf{v}}{\|\mathbf{v}\|}$.

3.3 The $\mathbf{u}$ *Sub-problem*

With $\mathbf{g}$ fixed, the problem (3) is reduced to

$$\min_{\mathbf{u}} \frac{\mu_3}{2}\left\|T^{(\mathrm{L})}\mathbf{u}-\tilde{\mathbf{m}}\right\|_2^2 + \frac{\beta}{2}\sum_i\left\|D_i\mathbf{u}-\mathbf{g}_i-\frac{\mathbf{v}_i}{\beta}\right\|_2^2 \tag{6}$$

The second term in (6) is exactly $(\beta/2)\left\|D\mathbf{u}-\mathbf{g}-\mathbf{v}/\beta\right\|_2^2$. By writing $\mathbf{t}=\mathbf{g}-\mathbf{v}/\beta$ and reorganizing $\mathbf{t}_i$ (for all i) into $\mathbf{t}^{(\mathrm{h})}$ and $\mathbf{t}^{(\mathrm{v})}$, (6) can be reformulated as

$$\min_{\mathbf{u}} \frac{\mu_3}{2}\left\|T^{(\mathrm{L})}\mathbf{u}-\tilde{\mathbf{m}}\right\|_2^2+\frac{\beta}{2}\left(\left\|D^{(\mathrm{h})}\mathbf{u}-\mathbf{t}^{(\mathrm{h})}\right\|_2^2+\left\|D^{(\mathrm{v})}\mathbf{u}-\mathbf{t}^{(\mathrm{v})}\right\|_2^2\right) \tag{7}$$

This is a simple least square problem. Since both the operators $D^{(\mathrm{h})}$ and $D^{(\mathrm{v})}$ are convolutions, they can be efficiently calculated in the frequency domain. To be specific, if we use periodic extension at signal boundary, the convolution can be calculated by pointwise multiplication in the DFT domain. Also, we note that $T^{(\mathrm{L})}$ is essentially data masking in the frequency transform, the problem (7) can be solved in frequency domain, by

$$\mathbf{u}=\mathcal{T}^{(-1)}\left(\frac{\beta\mathcal{T}^{(*)}(D^{(\mathrm{h})})\circ\mathcal{T}(\mathbf{t}^{(\mathrm{h})})+\beta\mathcal{T}^{(*)}(D^{(\mathrm{v})})\circ\mathcal{T}(\mathbf{t}^{(\mathrm{v})})+\mu_3\mathcal{T}^{(*)}(T^{(\mathrm{L})})\circ\mathcal{T}(\tilde{\mathbf{m}})}{\beta\mathcal{T}^{(*)}(D^{(\mathrm{h})})\circ\mathcal{T}(D^{(\mathrm{h})})+\beta\mathcal{T}^{(*)}(D^{(\mathrm{v})})\circ\mathcal{T}(D^{(\mathrm{v})})+\mu_3\mathcal{T}^{(*)}(T^{(\mathrm{L})})\circ\mathcal{T}(T^{(\mathrm{L})})}\right) \tag{8}$$

Here $\mathcal{T}$ is the forward DFT transform and $\mathcal{T}^{(-1)}$ is the inverse DFT transform. $\mathcal{T}^{(*)}$ is conjugate of the forward transform. If we use DCT (instead of DFT) as the transform in base layer and employ symmetric extension for the convolution $D^{(\mathrm{h})}$ and $D^{(\mathrm{v})}$, the problem can be similarly solved in the DCT domain [14], with minor changes to (8).

4 Experimental Results

To evaluate the performance of our proposed scheme, we compare it with two anchor schemes. One anchor is the direct transmission (*DirectTx*) scheme, in which the pixels are directly sent out using analog transmission, without any transform and power allocation. This is similar to the traditional analog-TV. Another anchor is the *SoftCast* scheme in [8, 9], which employs DCT transform and power allocation. For our *G-Cast* scheme, we set $M=8$ (i.e., the base layer includes 8×8 DCT coefficients). To make the comparison fair, the transmission in *DirectTx* and *SoftCast* is performed twice (and averaged at the receiver side) so that they send the same amount of data as *G-Cast* does. The three schemes are tested under various channel SNR conditions, ranging from -3dB to 15dB. For simplicity, the OFDM transmission is simulated by an Additive White Gaussian Noise (AWGN) channel.

Fig. 4 shows the PSNR results of the reconstructed images. The proposed scheme turns out to be better than *DirectTx*, but inferior to *SoftCast*, in terms of PSNR values. This is not surprising because our scheme is not optimized w.r.t. MSE. In fact, we are more interested in the perceptual quality or structural similarity. Fig. 5 shows the SSIM results of the reconstructed images. Clearly, *G-Cast* outperforms both *SoftCast* and *DirectTx* for all CSNR conditions. The advantage of *G-Cast* over *SoftCast* is particularly significant at low CSNR conditions. To gain deeper insight, we also evaluate the schemes by gradient SNR (GSNR). In other words, we transform both the original image and the reconstructed image into gradient domain and measure the signal fidelity by the signal-to-noise ratio in the gradient domain. Fig. 6 shows the GSNR results of the reconstructed images. We can see that *G-Cast* outperforms both *SoftCast* and *DirectTx* in GSNR values, for all CSNR conditions.

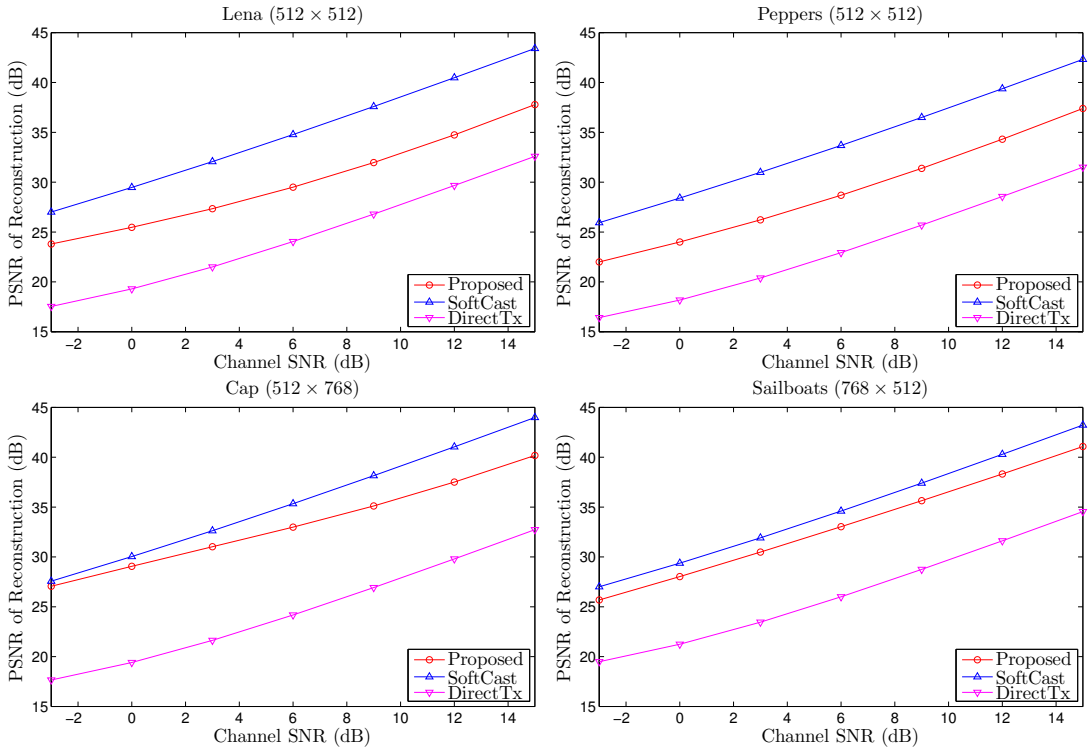

Figure 4: The PSNR results of reconstructed images.

That means, under the same channel condition, the *G-Cast* scheme preserves the gradient information more accurately than both *SoftCast* and *DirectTx*.

Perceptual quality is the ultimate target of our design. Fig. 7 shows the reconstructed images of the three tested schemes, under the channel condition CSNR=0dB. We see that the output of *DirectTx* contain white noises while the output of *SoftCast* contain low-frequency noises. The output of *G-Cast* are much clear, with most of image details reserved and most noises removed. Experiments on more test images give similar observations. Based on this, we argue that the proposed gradient-based image SoftCast scheme is desirable for perception oriented wireless visual communication.

5 Conclusions and Discussions

Recent advancements in image quality assessment indicate that the perceptual fidelity of images can be measured more reliably by the gradient similarity than the widely used conventional distortion metric mean square error. This suggests that the perceptible visual information of an image can be described by the image gradient more efficiently than the pixel intensity itself. Inspired by this observation, we proposed a gradient-based image transmission scheme which communicates an image signal by sending its gradient data with its achievable minimum distortion. Experimental results show that the proposed scheme provides very promising perceptual quality for

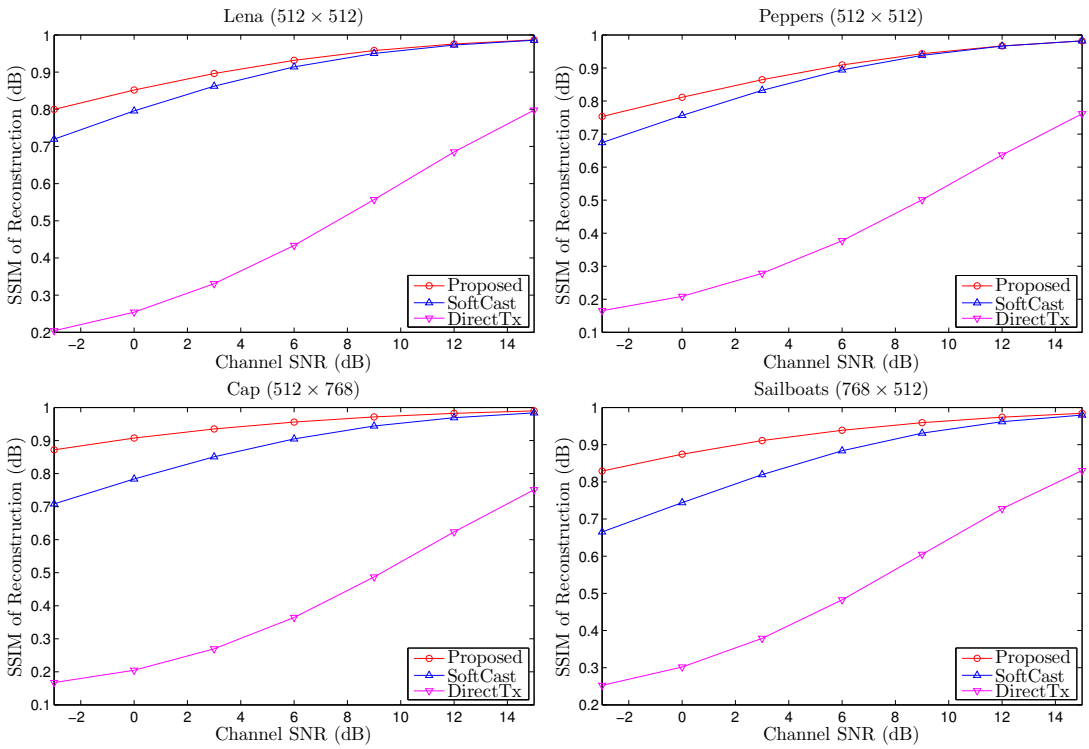

Figure 5: The SSIM results of reconstructed images.

wireless visual communication.

In order to reconstruct the original image from the noisy gradient data, this paper utilized a very simple image prior model to reduce the influence of channel noises. In future works, we will consider more advanced image prior models to further improve the efficiency of the proposed scheme.

References

[1] B. Girod, "What's wrong with mean-squared error," *Digital Images and Human Vision*, pp. 207–220, MIT Press, Cambridge, MA, USA, 1993.

[2] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[3] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] J. Zhu and N. Wang, "Image quality assessment by visual gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 919–933, 2012.

[5] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[6] A. Rehman and Z. Wang, "Ssim-inspired perceptual video coding for hevc," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 497–502.

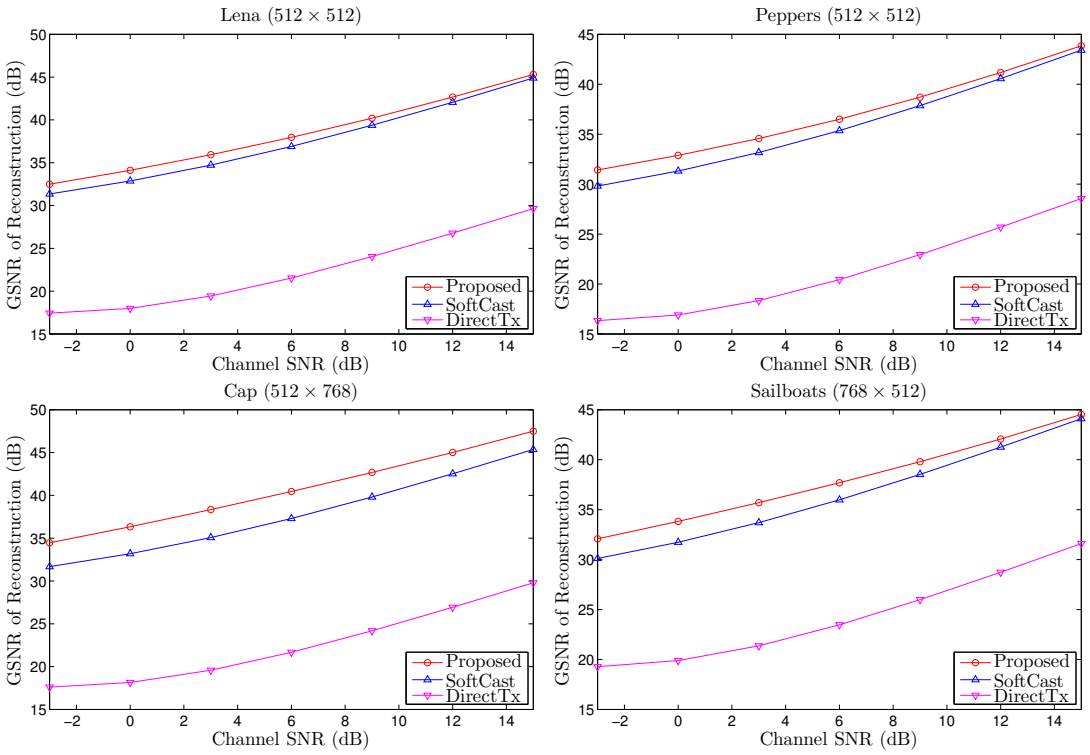

Figure 6: The GSNR results of reconstructed images.

[7] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on ssim-inspired divisive normalization," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1418–1429, 2013.

[8] S. Jakubczak, H. Rahul, and D. Katabi, "Softcast: One video to serve all wireless receivers," in *MIT Technical Report, MIT-CSAIL-TR-2009-005*, 2009.

[9] S. Jakubczak and D. Katabi, "A cross-layer design for scalable mobile video," in *International conference on Mobile computing and networking (MobiCom '11)*, New York, NY, USA, 2011, pp. 289–300.

[10] R. Xiong, F. Wu, J. Xu, and W. Gao, "Performance analysis of transform in uncoded wireless visual communication," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2013, pp. 1159–1162.

[11] K. N. Ngan, "Experiments on two-dimensional decimation in time and orthogonal transform domains," *Signal Processing*, vol. 11, no. 3, pp. 249–263, 1986.

[12] S. Martucci, "Image resizing in the discrete cosine transform domain," in *International Conference on Image Processing*, vol. 2, 1995, pp. 244–247, vol.2.

[13] J. Zhang, R. Xiong, S. Ma, and D. Zhao, "High-quality image restoration from partial random samples in spatial domain," in *IEEE Visual Communications and Image Processing (VCIP)*, 2011, pp. 1–6.

[14] S. Martucci, "Symmetric convolution and the discrete sine and cosine transforms," *IEEE Transactions on Signal Processing*, vol. 42, no. 5, pp. 1038–1051, 1994.

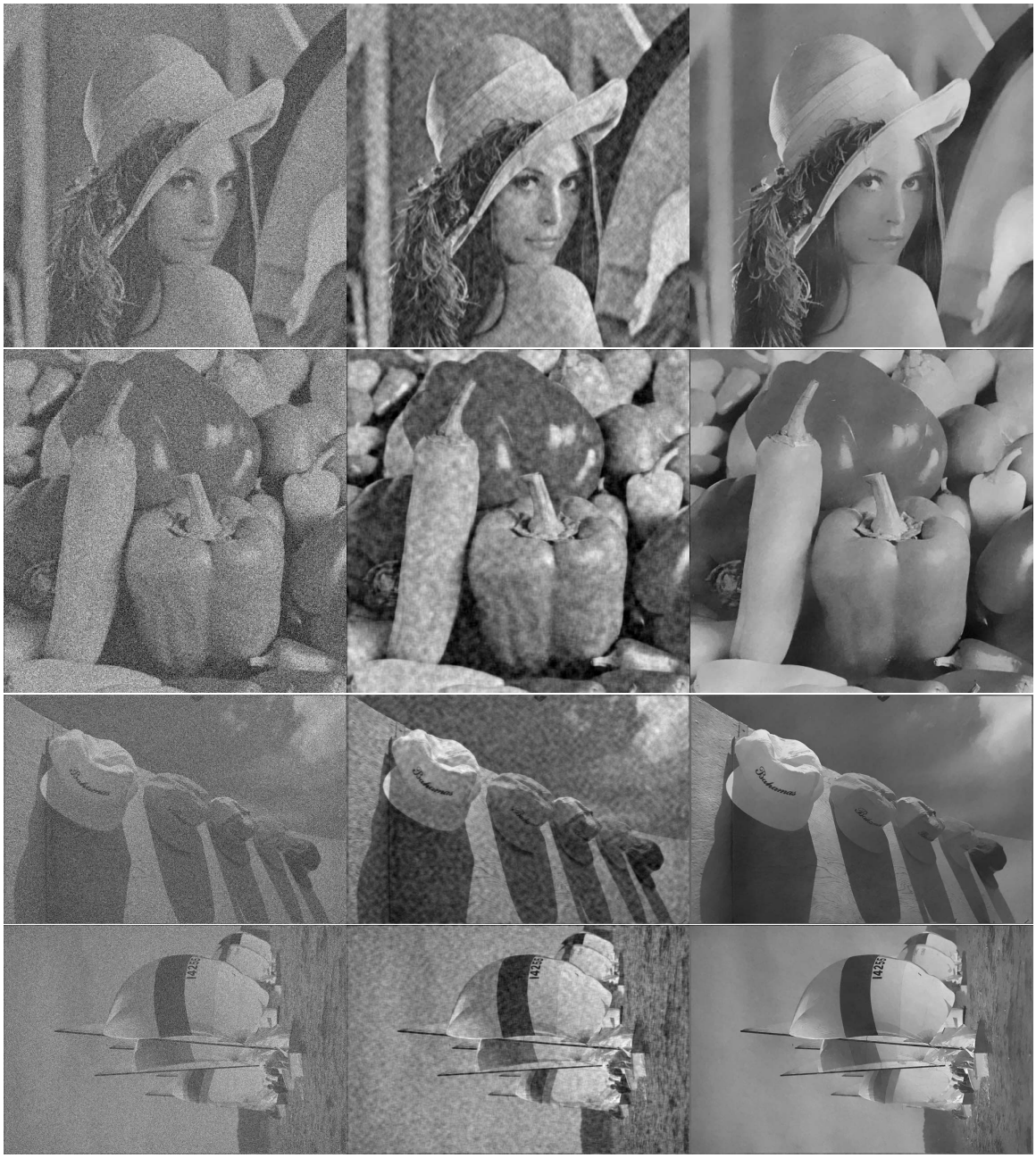

Figure 7: Perceptual quality comparison for the three evaluated schemes (CSNR=0dB). Left: the *DirectTx* scheme. Middle: the *SoftCast* scheme. Right: the *G-Cast* scheme. Enlarge the figures for visual details.