Gradient Based Image/Video SoftCast with Grouped-Patch Collaborative Reconstruction

Hangfan Liu[†], Ruiqin Xiong[†], Siwei Ma[†], Xiaopeng Fan[‡], and Wen Gao[†]
[†]Institute of Digital Media, School of EE&CS, Peking University, Beijing 100871, China
[‡]Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China

***Abstract*—Inspired by the recent image quality assessment (IQA) studies which indicate that the image gradient data reflects the visual information more reliably than the image pixels, gradient based transmission scheme was recently proposed to pursue better perceptual quality for wireless visual communication. This paper develops an effective method to reconstruct high quality image from the received noisy gradient data. The proposed method utilizes both local correlation and non-local similarity within the image signal to regularize the reconstruction image. Principle component analysis (PCA) is employed to learn signal-adaptive two-dimensional (2D) transform basis, and 3D transform is performed on grouped similar patches to further decorrelate the coefficients. In this way, distortions can be effectively suppressed via adaptive collaborative shrinkage on the transform coefficients. Experimental results demonstrate that the proposed method improves the reconstruction performance remarkably compared with the existing schemes.**

***Index Terms*—Gradient-based image transmission, local correlation, nonlocal similarity, grouped similar patches, collaborative shrinkage**

I. Introduction

Most image communication and reconstruction schemes use mean square error (MSE) as the fidelity measurement when transmitting pixel values or DCT coefficients. Nevertheless, MSE of pixel values may not coincide with the visual quality perceived by human eyes. Recent researches on image quality assessment reveal that gradient similarity is highly correlated with perceptual image quality. Inspired by such observations, [1] proposed to convey the visual information by delivering gradient data with minimum distortion in order to achieve better perceptual reconstruction quality.

Although [1] has shown evident visual improvement over the compared anchor schemes, it merely uses a rather simple image prior model to suppress the influence of noise. Bayesian framework tells us that, as a way to depict the characteristic of the original image, the image priors play a significant role in image reconstruction. Widely used image priors include sparse representation in different domains, such as DCT domain [2] and Wavelet domain [3], or on a trained dictionary [4]. The transform or dictionary de-correlates the signals, concentrating the energy on a few coefficients. However, the use of fixed transform basis like DCT and wavelet ignores the fact that image signals are not stationary. Structural patterns in an image can be quite different, making it impossible to represent the whole image efficiently with a fixed basis. The principal component analysis (PCA) can be utilized to tackle this problem [5], which computes signal-adaptive basis in order to get the sparsest representation of image.

In order to achieve better reconstruction performance than [1], this paper takes advantage of both local correlation and nonlocal similarity. We adopt the total variation (TV) regularization to utilize local correlation, and perform three-dimensional PCA as well as collaborative shrinkage to make use of nonlocal similarity. Taken together, the TV regularization attempts to smooth out noise, while the usage of nonlocal similarity can preserve the finest details when suppressing distortions, so the cooperation of these two terms in our scheme may achieve excellent reconstruction performance.

The remainder of this paper is organized as follows. Section II briefly reviews the G-Cast scheme. Section III describes the proposed reconstruction framework for G-Cast and Section IV explains the numerical solution. Experimental results are reported in Section V and Section VI concludes the paper.

II. Gradient Based Image SoftCast

Based on recent researches showing that image gradients contain large amount of visual information, G-Cast advocates to transmit an image over wireless channel by delivering gradient data. At G-Cast sender, image gradients are generated by gradient transform (GT), and processed by Walsh-Hadamard transform (WHT) to reduce peak-to-mean ratio. Just as SoftCast [6], [7], the WHT transformed data are modulated into a dense constellation for raw OFDM transmission. In addition, a few low frequency components are also provided so as to tell the global and regional luminance of the image. For this purpose, the image is transformed into frequency domain and a small amount of data are extracted by low pass selection (LPS). These low frequency components are encoded into bitstream using variable length coding (VLC), then sent to OFDM module for transmission using FEC codes (for error protection) and quadrature amplitude modulation (QAM).

The transmission process is usually influenced by interferences in the air, which is modeled by additive Gaussian white noise. At the receiver side, it is the gradient values rather than pixel intensities that the G-Cast scheme reproduces for

Corresponding author: R. Xiong (rqxiong@pku.edu.cn). This work was supported in part by the National Natural Science Foundation of China (61370114, 61121002, 61073083), Beijing Natural Science Foundation (4132039, 4112026) and Research Fund for the Doctoral Program of Higher Education (20120001110090, 20100001120027).

978-1-4799-6139-9/14/$31.00 ©2014 IEEE

image reconstruction. The decoder first retrieves the gradients from the noisy OFDM signal by demodulation and inverse WHT transform, then creates an estimation of the image via a gradient based reconstruction (GBR) procedure using these gradients as well as the several low frequency components. The illustration of G-Cast transmission scheme is shown in Fig. 1. Please refer to [1] for further details.

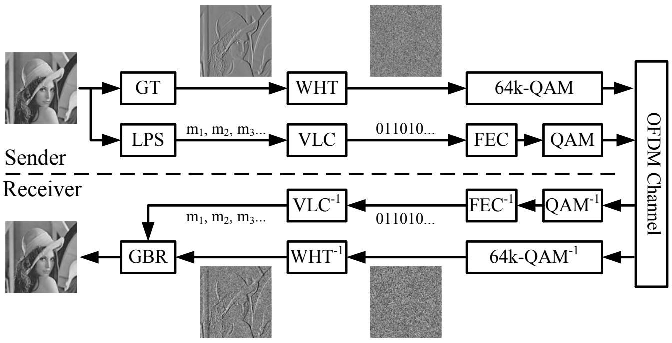

Fig. 1. Illustration of G-Cast Transmission Scheme

III. Image Prior for G-Cast Reconstruction

The image reconstruction scheme in [1] did not consider advanced methods to further improve the performance. Bayes rule tells us that, in order to achieve better reconstruction efficiency, a good prior model is indispensable. Obviously, the reconstructed image should share with natural images the characteristics of nonlocal similarity, which reflects the repetitiveness of the structural patterns in an image. The usage of nonlocal similarity in this paper is two-fold: finding data samples for principle component analysis (PCA), and three-dimensional (3D) collaborative shrinkage in transform domain.

Suppose $\mathbf{p}_i$ is a $k \times k$ patch treated as a column vector, its M most similar patches $\mathbf{p}_i^1, \mathbf{p}_i^2, \dots \mathbf{p}_i^M$ can be searched out by block matching. Let matrix $P_i = [\mathbf{p}_i^1, \mathbf{p}_i^2, \dots \mathbf{p}_i^M]$. We can estimate $\overline{\mathbf{p_i}}$, the expectation of $\mathbf{p_i}$, by averaging every row of $\mathbf{P_i}$. Then the centralized version of P_i is $\tilde{P}_i = [\mathbf{p}_i^1 - \overline{\mathbf{p}_i}, \mathbf{p}_i^2 - \overline{\mathbf{p}_i}, \dots \mathbf{p}_i^M - \overline{\mathbf{p}_i}]$. The covariance matrix of $\tilde{P}_i$ is symmetric and can be written as $\text{Cov}(\tilde{P}_i) = \Phi_i \Lambda_i \Phi_i^{\text{T}}$, where Φ_i is the orthonormal eigenvector matrix and Λ_i is the sorted diagonal eigenvalue matrix. Denote by $\Psi_i = \Phi_i^{\text{T}}$, then Ψ_i is the desired transform to fully de-correlate the signal $\mathbf{p}_i$.

To thoroughly decorrelate the coefficients in transform domain, we adopt a collaborative filtering approach, which is similar with BM3D [8]. For the current patch $\mathbf{p}_i$, we stack its M similar patches into a group, perform PCA patch-wise, i.e., multiply each patch by Ψ_i, then perform 1D DCT along the third dimension. Denote the resulted coefficient block by $\mathbf{Z}_i$.

We model the distribution of each coefficient by a separate zero-mean Laplacian distribution. Use a block $\mathbf{T}_i$ of the same size as $\mathbf{Z}_i$ to store the standard deviation values of these coefficients. These standard deviation values are estimated from the transform coefficients of the pilot image $\mathbf{u}'$. Specifically speaking, the same collaborative transform is carried out within $\mathbf{u}'$ and results in a coefficient block $\mathbf{Z}_i'$, in which each coefficient is treated as the sample of the corresponding coefficient in $\mathbf{Z}_i$. Since only one sample for each coefficient is available, we may just use the absolute value of the corresponding coefficient in $\mathbf{Z}_i'$ as the rough estimation of standard deviation, i.e. set $\mathbf{T}_i = |\mathbf{Z}_i'|$. Denote the prior by

$$\phi(\mathbf{u}) = \sum_i \left\| \frac{\sqrt{2}}{\mathbf{T}_i} \mathbf{Z}_i^u \right\|_1 . \tag{1}$$

IV. Numerical Algorithm for G-Cast Reconstruction

Let D^{v} and D^{h} be the vertical and horizontal finite difference operators, hence $D^{\text{v}}\mathbf{u}$ and $D^{\text{h}}\mathbf{u}$ are are respectively the vertical gradient picture and horizontal gradient picture. Let $\mathbf{m} = E \circ \mathcal{F}(\mathbf{u})$ be the low-frequency coefficients of $\mathbf{u}$, $\mathcal{F}$ is two-dimensional discrete Fourier transform, E represents the matrix to extract the $M \times M$ block at the top left corner from a matrix of the same size, and "$\circ$" denotes component-wise multiplication. Assume that the received gradient data $\mathbf{d}^{\text{v}}$ and $\mathbf{d}^{\text{h}}$ are polluted by Gaussian white noise:

$$\mathbf{d}^{\text{v}} = D^{\text{v}}\mathbf{u} + \mathbf{n}^{\text{v}}, \ \mathbf{d}^{\text{h}} = D^{\text{h}}\mathbf{u} + \mathbf{n}^{\text{h}}, \tag{2}$$

Write $D = [D^{\text{v}}; D^{\text{h}}]$, $\mathbf{d} = [\mathbf{d}^{\text{v}}; \mathbf{d}^{\text{h}}]$ and $\mathbf{n} = [\mathbf{n}^{\text{v}}; \mathbf{n}^{\text{h}}]$ for simplicity.

Based on TV regularization and the image prior (1), the MAP estimate of the reconstructed image is formulated as:

$$\min_{\mathbf{u}} \frac{\mu}{2}\|D\mathbf{u} - \mathbf{d}\|_2^2 + \frac{\sqrt{2}}{\sigma_\Delta}\sum_i \|D_i\mathbf{u}\| + \phi(\mathbf{u}) \text{ s.t. } E \circ \mathcal{F}(\mathbf{u}) = \mathbf{m}, \tag{3}$$

where σ_n^2 is the variation of noise and σ_Δ is the standard deviation of the gradient data. Resorting to variable splitting technique [9], [10] to change (3) into a constrained problem:

$$\begin{aligned} &\min_{\mathbf{u}} \frac{\mu}{2}\|\mathbf{w} - \mathbf{d}\|_2^2 + \frac{\sqrt{2}}{\sigma_\Delta}\sum_i \|w_i\| + \phi(\mathbf{x}) \\ &\text{s.t. } E \circ \mathcal{F}(\mathbf{u}) = \mathbf{u}, \mathbf{w} = D\mathbf{u}, \mathbf{x} = \mathbf{u}. \end{aligned} \tag{4}$$

The corresponding augmented Lagrange function reads

$$\begin{aligned} \mathcal{L}_A(\mathbf{u}, \mathbf{w}, \mathbf{x}) &= \frac{1}{2\sigma_n^2}\|\mathbf{w} - \mathbf{d}\|_2^2 \\ &+ \frac{\sqrt{2}}{\sigma_\Delta}\|\mathbf{w}\| + \frac{\beta}{2}\|\mathbf{w} - D^{\text{h}}\mathbf{u}\|_2^2 - \lambda^{\text{T}}(\mathbf{w} - D\mathbf{u}) \\ &+ \phi(\mathbf{x}) + \frac{\tau}{2}\|\mathbf{x} - \mathbf{u}\|_2^2 - \delta^{\text{T}}(\mathbf{x} - D\mathbf{u}) \\ &+ \frac{\gamma}{2}\|E \circ \mathcal{F}(\mathbf{u}) - \mathbf{m}\|_2^2 - \rho^{\text{T}}(E \circ \mathcal{F}(\mathbf{u}) - \mathbf{m}). \end{aligned} \tag{5}$$

where β, τ and γ are regularization parameters, λ, δ and ρ are Lagrange multipliers.

Then alternating direction technique [11], [12] can be used to decompose (5) into three sub-problems so as to conquer each of them efficiently.

A. $\mathbf{x}$-*problem*

With $\mathbf{u}$ and $\mathbf{w}$ fixed, the optimization problem associated with $\mathbf{x}$ can be written as:

$$\mathcal{L}_A(\mathbf{x}) = \phi(\mathbf{x}) + \frac{\tau}{2}\|\mathbf{x} - \mathbf{u} - \frac{\delta}{\tau}\|_2^2. \tag{6}$$

Let $\mathbf{r} = \mathbf{u} + \frac{\delta}{\tau}$, which can be regarded as a noisy observation of $\mathbf{x}$. Suppose N is the size of image, K is the total number of the groups generated by block matching over the whole image, and i is the group index. Assume that elements of $\mathbf{u} - \mathbf{r}$ conform to an i.i.d zero-mean distribution, and that every pixel appears in the groups equally frequently, then it can be inferred that

$$\|\mathbf{u} - \mathbf{v}\|_2^2 = \theta \sum_i^K \|\mathbf{Z}_i^u - \mathbf{Z}_i^v\|_2^2, \theta = \frac{N}{k^2 \times M}. \tag{7}$$

Then taking (1) into account, (6) is equivalent to

$$\mathcal{L}_A(\mathbf{Z}^x) = \sum_i^K \left(\left\| \frac{\sqrt{2}}{\mathbf{T}_i} \mathbf{Z}_i^x \right\|_1 + \frac{\theta \cdot \tau}{2} \|\mathbf{Z}_i^x - Z_i^r\|_2^2 \right). \tag{8}$$

The solution is a component-wise shrinkage operation:

$$Z_i^x = \max(|\mathbf{Z}_i^r| - \frac{\sqrt{2}}{\theta \cdot \tau \cdot \mathbf{T}_i}, 0) \cdot \mathrm{sgn}(\mathbf{Z}_i^r). \tag{9}$$

Then $\mathbf{x}$ can be obtained by taking inverse transform for every $\mathbf{Z}_i^x$, putting back the patches and performing weighted average.

B. $\mathbf{w}$-*problem*

With $\mathbf{u}$ and $\mathbf{x}$ fixed, the $\mathbf{w}$-problem is simplified as:

$$\mathcal{L}_A(\mathbf{w}) = \frac{\sqrt{2}}{\sigma_{\Delta_i}} \|\mathbf{w}\| + \frac{\beta + \eta}{2} \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2, \tag{10}$$

with $\eta = \frac{1}{\sigma_n^2}$, $\tilde{\mathbf{w}} = \frac{\beta(D^{\mathbf{u}} + \frac{\lambda}{\beta}) + \eta \mathbf{d}}{\beta + \eta}$. The solution is also a simple component-wise shrinkage operation:

$$\mathbf{w} = \max\left(|\tilde{\mathbf{w}}| - \frac{\sqrt{2}}{(\theta + \eta) \cdot \sigma_\Delta}, 0\right) \cdot \mathrm{sgn}(|\tilde{\mathbf{w}}| - \frac{\sqrt{2}}{(\theta + \eta) \cdot \sigma_\Delta}). \tag{11}$$

C. $\mathbf{u}$-*problem*

Fixing $\mathbf{x}$ and $\mathbf{w}$, the $\mathbf{u}$-problem becomes:

$$\begin{aligned} \mathcal{L}_A(\mathbf{u}) = & \frac{\gamma}{2} \|E \circ \mathcal{F}(\mathbf{u}) - (\mathbf{m} + \frac{\rho}{\gamma})\|_2^2 \\ & + \frac{\beta}{2} \|D\mathbf{u} - (\mathbf{w} - \frac{\lambda}{\beta})\|_2^2 + \frac{\tau}{2} \|\mathbf{u} - (\mathbf{x} - \frac{\delta}{\tau})\|_2^2 \end{aligned} \tag{12}$$

Considering that D can be seen as a convolution operator while $\mathbf{m}$ happens to be a block of Fourier coefficients, the least square problem can be efficiently solved in the Fourier transform domain:

$$\mathbf{u} = \mathcal{F}^{-1} \left(\frac{\mathcal{F}^*(D) \circ \mathcal{F}(\mathbf{w} - \frac{\lambda}{\beta}) + \frac{\tau}{\beta} \mathcal{F}^*(\mathbf{I}) \circ \mathcal{F}(\mathbf{x} - \frac{\delta}{\tau}) + \frac{\gamma}{\beta}(\mathbf{m} + \frac{\rho}{\gamma})}{\mathcal{F}^*(D) \circ \mathcal{F}(D) + \frac{\tau}{\beta} \mathcal{F}^*(\mathbf{I}) \circ \mathcal{F}(\mathbf{I}) + \frac{\gamma}{\beta} \cdot E} \right), \tag{13}$$

here "$*$" denotes complex conjugacy, $\mathbf{I}$ is the identity matrix. Both the multiplication and the division are component-wise.

V. Experimental Results

This section presents the experimental results to evaluate the performance of the proposed G-Cast reconstruction scheme. As a reconstruction approach for wireless visual communication scheme, it should be tested under a wide CSNR range, which is not feasible for standard coding methods, so we use SoftCast [6] and TV based G-Cast [1] as the anchor schemes. To make the comparison fair, the transmission in SoftCast is performed twice (and averaged at the receiver side) so that they send the same amount of data as G-Cast does.

In the implementation, patch size is set to be 8×8 and the number of similar patches M is set to be 50. We set $m = 8$, i.e. the base layer of G-Cast remains 8×8 low-frequency coefficients. The three schemes are tested by 15 images. Besides the widely used metrics SSIM and PSNR, we also measure the signal fidelity in gradient domain by gradient SNR (GSNR). Since G-Cast is designed for perception oriented visual communication, we put more emphasis on SSIM and GSNR.

As can be seen from Table I, in terms of SSIM and GSNR, the proposed scheme outperforms TV based G-Cast substantially, and TV based G-Cast is much better than SoftCast. The average performance of the 15 tested images exhibits similar results, for example, when CSNR $= 0$dB, the GSNR of our proposed scheme is 1.52dB higher than TV based G-Cast and 3.59dB higher than SoftCast; the SSIM of the proposed method is 0.04 higher than TV based G-Cast and 0.13 higher than SoftCast. In terms of PSNR, the proposed scheme also has evident gain over TV based G-Cast, and is superior to SoftCast in many cases but inferior in others, which is not strange because G-Cast is not optimized w.r.t MSE.

Furthermore, portions of two reconstructed images are shown in Fig. 2. The proposed scheme has remarkable improvement over the other two schemes with respect to visual quality. Other experimental results show similar observations and are not displayed here due to lack of space.

VI. Conclusion

This paper presents an effective image reconstruction scheme for gradient based image transmission. Exploiting the local and nonlocal correlation in images, we stack the similar patches into a group, and perform PCA transform in spatial domain as well as a transform along the third dimension so as to get a sparser representation. Then perform adaptive shrinkage to the transform coefficients. The intensity of shrinkage is determined by the energy of each coefficient. After estimate of each patch is generated by inverse transform, we can get the reconstructed image by putting back all the patches and averaging overlaps. Experimental results show that the proposed method outperforms SoftCast and TV based G-Cast dramatically in terms of both objective and visual quality.

References

[1] R. Xiong, H. Liu, S. Ma, X. Fan, F. Wu, and W. Gao, "G-cast: Gradient based image softcast for perception-friendly wireless visual communication," *IEEE Data Compression Conference (DCC'14)*, pp. 133–142, March 2014.

TABLE I
SSIM, GSNR (dB) AND PSNR (dB) COMPARISON

	Images	*Cameraman (256 × 256)*			*Sailboats (768 × 512)*			*Window (512 × 768)*			*House (256 × 256)*		
CSNR	*Schemes*	*SoftCast*	*G-Cast(TV)*	*Proposed*	*SoftCast*	*G-Cast(TV)*	*Proposed*	*SoftCast*	*G-Cast(TV)*	*Proposed*	*SoftCast*	*G-Cast(TV)*	*Proposed*
0dB	*SSIM*	0.5814	0.7854	**0.8129**	0.7563	0.8564	**0.8960**	0.7817	0.8673	**0.9180**	0.7671	0.8648	**0.9057**
	GSNR	4.45	7.59	**8.52**	4.91	7.30	**9.03**	5.29	7.01	**9.38**	6.06	7.09	**9.31**
	PSNR	24.59	26.84	**28.63**	29.22	28.60	**29.93**	**29.14**	26.86	29.04	28.81	28.85	**31.76**
3dB	*SSIM*	0.6727	0.8365	**0.8657**	0.8297	0.9004	**0.9265**	0.8473	0.9119	**0.9449**	0.8425	0.9053	**0.9309**
	GSNR	6.39	9.59	**10.53**	6.89	9.16	**10.68**	7.25	9.05	**11.29**	8.05	8.96	**10.91**
	PSNR	27.09	29.24	**30.69**	31.76	31.10	**32.14**	**31.71**	29.55	31.65	31.45	31.54	**33.94**
6dB	*SSIM*	0.7640	0.8834	**0.9055**	0.8911	0.9347	**0.9497**	0.9016	0.9439	**0.9631**	0.9018	0.9382	**0.9513**
	GSNR	8.68	11.78	**12.63**	9.19	11.22	**12.43**	9.52	11.27	**13.20**	10.32	11.08	**12.62**
	PSNR	29.75	31.68	**32.75**	34.44	33.73	**34.48**	**34.41**	32.41	34.27	34.21	34.35	**36.12**
9dB	*SSIM*	0.8452	0.9216	**0.9359**	0.9359	0.9594	**0.9669**	0.9416	0.9656	**0.9756**	0.9430	0.9622	**0.9675**
	GSNR	11.25	14.10	**14.82**	11.74	13.48	**14.36**	12.05	13.63	**15.17**	12.83	13.43	**14.52**
	PSNR	32.55	34.19	**34.92**	**37.25**	36.52	36.97	**37.23**	35.39	36.85	37.04	37.27	**38.34**
12dB	*SSIM*	0.9073	0.9501	**0.9579**	0.9646	0.9760	**0.9793**	0.9675	0.9796	**0.9843**	0.9688	0.9781	**0.9797**
	GSNR	14.00	16.55	**17.11**	14.48	15.93	**16.51**	14.77	16.13	**17.27**	15.53	15.97	**16.67**
	PSNR	35.43	37.04	**36.78**	**40.14**	39.37	39.41	**40.12**	38.43	39.46	39.94	40.25	**40.69**

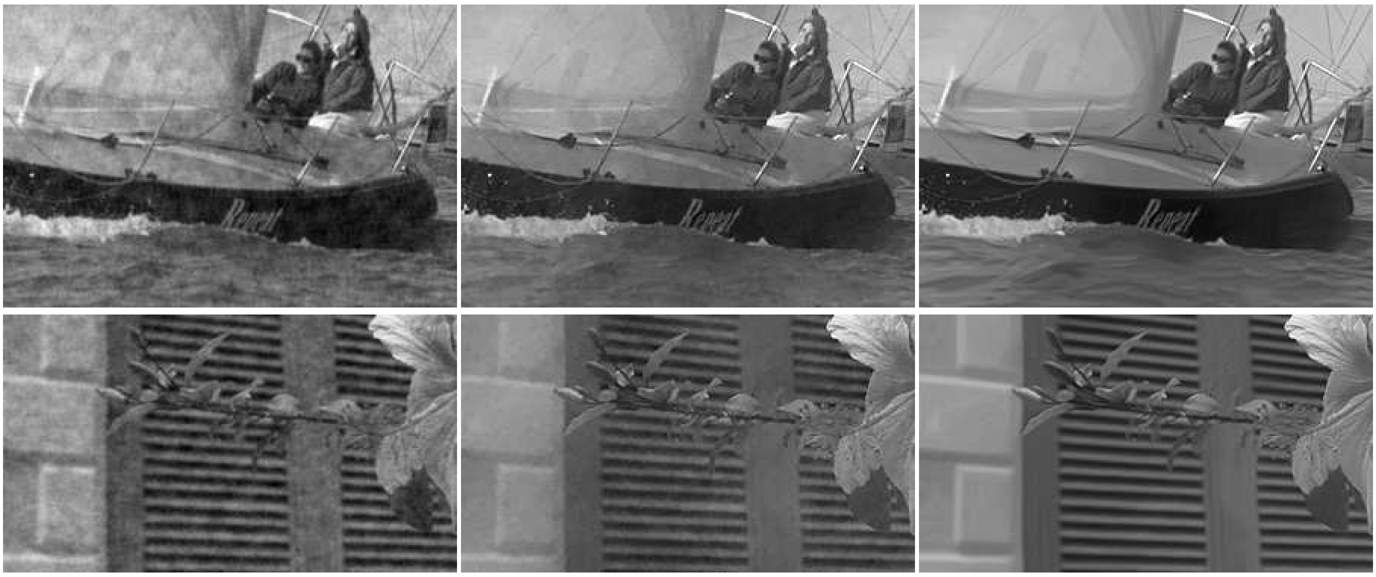

Fig. 2. Visual comparison of the reconstructed images with CSNR=0dB. From left to right: (a) SoftCast; (b) TV-based G-Cast; (c) Proposed.

[2] A. Foi, V. Katkovnik, and K. Egiazarian, “Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images,” *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1395–1411, May 2007.

[3] D. Donoho, “De-noising by soft-thresholding,” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[4] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[5] D. Muresan and T. Parks, “Adaptive principal components and image denoising,” in *2003 International Conference on Image Processing*, vol. 1, Sept 2003, pp. I–101–4 vol.1.

[6] 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.

[7] S. Jakubczak and D. Katabi, “A cross-layer design for scalable mobile video,” in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 289–300.

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-d transform-domain collaborative filtering,” *Image Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 2080–2095, Aug 2007.

[9] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, “Fast image recovery using variable splitting and constrained optimization,” *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2345–2356, Sept 2010.

[10] T. Goldstein and S. Osher, “The split bregman method for l1-regularized problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[11] C. Li, W. Yin, and Y. Zhang, “Tval3: Tv minimization by augmented lagrangian and alternating direction algorithms,” 2009.

[12] S. Chretien, “An alternating l_1 approach to the compressed sensing problem,” *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 181–184, Feb 2010.