Convolutional Neural Networks Based Soft Video Broadcast

Wenbin Yin[1(✉)], Xiaopeng Fan[1], and Yunhui Shi[2]

[1] Harbin Institute of Technology, Harbin, Heilongjiang, China
{ywb, fxp}@hit.edu.cn
[2] Beijing University of Technology, Beijing, China
syhzm@bjut.edu.cn

Abstract. Video broadcasting is becoming more and more popular in wireless networks. However, the existing digital coding and transmission approaches can hardly accommodate users with diverse channel conditions, which is called the cliff effect. Recently, a novel video broadcasting method called SoftCast has been proposed. It achieves graceful degradation with increasing noise by making the magnitude of the transmitted signal proportional to the pixel value and using a novel power allocation scheme. In this paper, we propose a novel video broadcast method that exploits deep convolutional networks and group based sparse representation. It utilizes the channel condition information generated from decoder to optimize the decoding process and reduce the various artifacts caused by source and channel coding. By utilizing soft video broadcast transmission, it achieves good broadcast performance and avoids the cliff effect. The experimental results show that the proposed scheme provides better performance compared with the traditional SoftCast with up to 1.5 dB coding gain.

Keywords: Video broadcasting · Convolutional neural networks
Soft video broadcast

1 Introduction

Wireless video broadcasting is becoming more and more popular in our daily life and its purpose is to transmit one video signal simultaneously to multiple receivers with different channel conditions. The main challenge we face is the difficulty to provide receivers with video quality that matches their channel conditions. The traditional wireless broadcasting design such as DVB-T standard [1] that combines a layered transmission scheme [2, 3] and scalable video coding (SVC) scheme [4, 5] is one of the typical wireless video broadcasting schemes. SVC encodes the video signal into one base layer (BL) and multiple enhancement layers (EL). In transmission, the hierarchical modulation (HM) [6] superimposes the multiple layer bits in one wireless symbol and allows the user to decode different numbers of layers according to their own channel condition. However, the layered scheme reduces both the compression efficiency and the transmission efficiency.

Recently, a novel solution of wireless video broadcasting called SoftCast [7] is proposed. The SoftCast consists of four steps: DCT transform, power allocation,

R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 641–650, 2018.
https://doi.org/10.1007/978-3-030-00764-5_59

Hadamard transform and direct dense modulation. DCT transform compresses the video frame by removing the spatial redundancy of a video frame. Power allocation reduces the total distortion by optimally scaling the DCT coefficients. Hadamard transform can make each packet with equal importance as a protect-coding. The most attractive difference between SoftCast and traditional approach is that SoftCast directly map the data into wireless symbols by a very dense Quadrature Amplitude Modulation (QAM). At decoder side, SoftCast uses Linear Least Square Estimator to reconstruct the video frame.

Although SoftCast achieves graceful degradation with increasing noise by making the magnitude of the transmitted signal proportional to the pixel value and using a novel power allocation scheme to against the channel noise, it still has room for improvement. Compression and transmission in its nature will introduce undesired complex artifacts, which will severely reduce the users' experience.

In recent years, several soft video broadcast schemes have been proposed [10, 11]. Meanwhile, a number of sparse coding based methods for image restoration [12, 13] have been developed and deep learning has shown impressive results on vision problems [8, 9]. In this paper, we utilize convolutional neural networks and sparse coding based representation to achieve a video multicast method with less compression and transmission artifacts. The encoder compresses the video frame by linear transformation and uses power allocation to minimize the distortion caused by channel noise. The decoder utilizes LLSE and inverse transformation to reconstruct the video frame. However, the decoded frame usually has some artifacts. The proposed scheme utilizes group based sparse representation to reduce the distortion produced by compression and CNN to reduce the distortion produced by transmission.

The rest of the paper is organized as follows. Section 2 describes the encoding and decoding process of the proposed scheme. The performance of our scheme is showed in Sect. 3, followed by concluding remarks in Sect. 4.

2 Proposed Scheme

At the encoder side, video frames are coded by block based DCT transform to compress the data and the coded components are scaled by power allocation to minimize the distortion cause of channel noise, and then transmitted to users with different channel conditions. In the traditional digital video transmission method, cliff effect affects the users' decoding experience. In our method, the scaled coefficients are directly transmitted through soft broadcast without syndrome coding over a very dense constellation that avoids the cliff effect. At decoder side, it uses LLSE to reconstruct the video frame. Since group based sparse representation model can utilize the intrinsic local sparsity and non-local similarity of nature image at same time, we exploit group based sparse representation to reduce the blocking artifacts caused by block based video compression. With initial reconstruction, we exploit convolutional networks to reduce the distortion caused by soft video transmission, since convolutional networks can extract features formed by different quality of channel noise and restore the decoded frames for different channel conditions (Fig. 1).

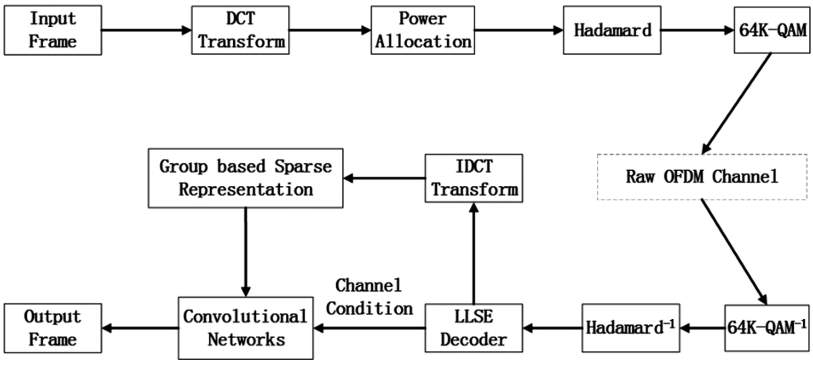


Fig. 1. Flow graph of the proposed scheme

2.1 Video Compression

Since video frames are relatively smooth and show spatial correlation. The proposed method exploits this property to compact the information in frames by taking block based DCT of pixel values. Traditional video coding scheme works with the assumption of a known channel, and encoder quantizing the DCT values as much as desired. This kind of quantization will force all receivers to the same reconstruction quality.

The proposed scheme divides the input frame into blocks, then it takes block-based DCT on this matrix to transform the frame from spatial domain into frequency domain. In general, the DCT components in the right bottom corner stand for high spatial frequencies and have low values, close or equal to zero. We can compress the video frame by discarding the zero value DCT components while these components have limited impact on the information in a frame. However, this kind of compression will cost a large amount of metadata to the decoder side to express the specific location of these discarded DCT components.

To reduce the metadata for the high frequencies DCT components, it divides the DCT values into bands and operates on bands. Specifically, we group DCT components in same position of each blocks into one band. Then we make one decision for all DCT values in a band. As we known, high frequencies components usually concentrated in same region, making one decision for a whole band can provide close performance with discarding individual DCT components. Since the proposed method has discarded few bands, it is much simpler to express the location of these bands than specific location of every discarded DCT components.

2.2 Power Allocation and Transmission

The power allocation can be treated as a protection method for each frequencies of the transmitted signal. Let P be the total power of transmission and g_{R_i} be the scaling factor of R_i that donates DCT components. According to [7], g_{R_i} is given by

$$g_{R_i} = \left(\frac{P}{\sqrt{\lambda_i} \sum_i^K \sqrt{\lambda_i}} \right)^{1/2} \tag{1}$$

where λ_i is the variance of i-th frequency DCT value and K is the total number of frequencies. We define a diagonal matrix $G = diag\{g_{R_1}, g_{R_2}, \ldots, g_{R_K}\}$, the signal M can be represent as $M = G \cdot R$. The λ_i needs to be transmitted as metadata to decoder side.

After the power allocation, we want to maximize proposed method's resilience to packet loss. We can generate packets with equally important by multiplying by a Hadamard matrix H, let, i.e.

$$U = H \cdot M = HG \cdot R = C \cdot R \tag{2}$$

In PHY layer, the metadata and DCT value are transmitted in different ways. Since the metadata needs to be transmitted without any error, we use conventional way to send the metadata. The encoder applies 8-bits scalar quantization on metadata and the quantization results are compressed by variable length coding (VLC). The compressed bit-stream is transmitted by the standard 802.11 PHY layer with FEC and modulation. To well protect the metadata, we use a 1/2 convolutional code and BPSK modulation.

Unlike the metadata, the signal consists of real values rather than binary values. In PHY layer, these real values are first mapped to complex symbol directly by 64K QAM constellation. Every two integers are quantized by an 8-bit quantizer and combined into one complex symbols as the output of the 64K QAM constellation. Given a set of complex time-domain samples, an inverse FFT is computed on each packet of symbols. The real and imaginary components are first converted to the analogue domain using D/A converters, the analogue signals are then used to modulate cosine and sine waves at the carrier frequency respectively. Then these signals are summed to give the transmission signal. With such aforementioned direct source and channel mapping method, it can let the reconstructed quality matching the channel condition in proposed method.

2.3 LLSE at Decoder

Here we define N as channel noise, and the received signal can be represented as

$$\hat{U} = U + N = HG \cdot R + N \tag{3}$$

And the received signal can be recovered by first LLSE estimator in transform domain as follows

$$\hat{R} = \Sigma_r C^T (C \Sigma_r C^T + \Sigma_N)^{-1} \hat{U} \tag{4}$$

where Σ_r and Σ_N are the covariance matrices of R and N. At high CSNR, the LLSE estimator simply inverts the encoder computation. At high CSNR, one can trust the measurement. At low CSNR, one cannot fully trust the measurements and it is better to re-adjust the estimate according to the statistics of the DCT components.

2.4 Group Based Sparse Representation for Deblocking

Due to block based DCT and power allocation, soft video broadcast usually results in visually annoying blocking artifacts in coded videos. Since the sparse representation performs well at removing the blocking artifacts and obtain visually acceptable quality for block based DCT coded videos. In the proposed method, we formulate the GSR based deblocking algorithm though maximum a posteriori (MAP) framework.

Here we define, that given first decoded video frame $\hat{x}$ and the input frame x, processed frame can be obtained by:

$$y = \arg\max \log(p(\hat{x}|x)) + \log(p(x)) \tag{5}$$

where the first term represents data-fidelity, and the second term corresponds to image priors. Inspired by the success of image group based sparse representation, the optimization problem for frame deblocking through MAP is formulated as

$$y = \arg\max \log(p(\hat{x}|x)) + \log(pGSR(x)) + \log(pQC(x)) \tag{6}$$

where $pGSR(x)$ and $pQC(x)$ stand for GSR prior and QC prior, respectively.

The decoded video frame contains transmission and compression noise. The GSR part focus on the compression noise which main causes the blocking artifacts. With the Gaussian compression noise model and compression noise variance σ_{com}^2, the first data-fidelity term can be formulated as

$$\log(p(\hat{x}|x)) = -\frac{1}{2\sigma_{com}^2}\|x - \hat{x}\|_2^2 \tag{7}$$

The group based sparse representation model [14] assumes that a few atoms of a dictionary can represent each group of image blocks. The sparse coding process of each group over dictionary is seek a sparse vector $x_{G_k} \approx D_{G_k}\alpha_{G_k}$. Then the whole image can be sparsely represented by the set of sparse codes $\{\alpha_{G_k}\}$ in the unit of group. So the second term in the Eq. (6) can be formulated as

$$\log(p_{GSR}(x)) = -\eta\|\alpha_G\|_0 \tag{8}$$

where α_G denotes the concatenation of all α_{G_k} and imposes the sparse codes vector α_G to be sparse. In order to incorporate QC prior, we define the indicator by Ω as

$$\psi(x) = \left\{ \begin{array}{ll} 0, & \text{if } x \in \Omega \\ +\infty, & \text{if } x \notin \Omega \end{array} \right\} \tag{9}$$

where Ω is the range of scaled DCT coefficients. The third term can be formulated as

$$\log(p_{QC}(x)) = -\psi(x) \tag{10}$$

Utilize the above priors, the deblocking minimization problem can be formulated as

$$(\hat{\alpha}_G, \hat{D}_G) = \arg\min_{\alpha_G, D_G} \frac{1}{2\sigma^2_{noise}} \|D_G \circ \alpha_G - \hat{x}\|_2^2 + \lambda \|\alpha_G\|_0 + \Psi(D_G \circ \alpha_G) \tag{11}$$

which can be solved by the framework of split Bergman iteration. Equation (11) is equivalently transformed into three iterative step and each separated sub-problem can acquire an efficient solution. After we get $\hat{\alpha}_G$ and $\hat{D}_G$ in hand, the de-blocked frame can be reconstructed by $y = \hat{D}_G \circ \hat{\alpha}_G$.

2.5 Convolutional Networks for Artifacts Reduction

Since the input frame is compressed by the band based coding and transmitted though the OFDM channel, the reconstructed frames usually have some compression and transmission artifacts. Since deep learning has shown impressive results on low-level vision problems, convolutional networks can extract features formed by different quality of channel noise and restore the decoded frames for different channel conditions. We adopt convolutional networks to cope with the compression and transmission artifacts. The whole convolutional networks are shown in Fig. 2.

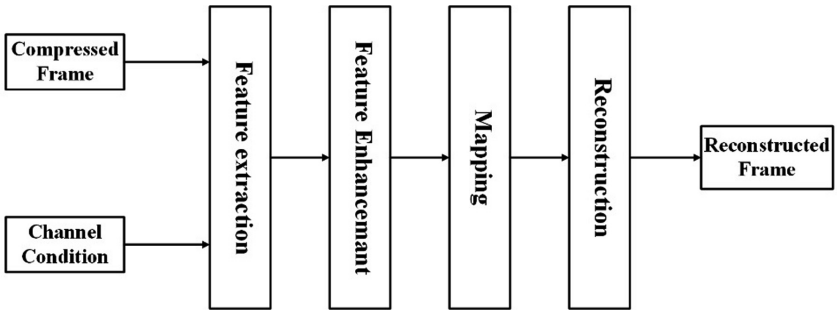


Fig. 2. Convolutional networks for video frame restoration

There are four layers in the restoration networks, each of which is responsible for a specific task. The first layer is used for patch and feature extraction, which extracts overlapping patches from the compressed frame and represents each patch as a high dimensional vector. The second layer can be seen as the feature enhancement layer which extract features from the n_1 feature maps of the first layer and form a new set of feature maps. After feature enhancement layer, at the third layer, we apply non-linear mapping layer to represents a high quality patch by a high dimensional vector. The last layer stands for reconstruction and it produce the final high resolution frames. The entire network can be express as:

$$\begin{aligned} &F_{\mathrm{i}}(y^{(c)}) = \max(0, W_i * y^{(c)} + B_i), i \in \{1, 2, 3\}; \\ &\hat{y}^{(c)} = W_4 * F_3(y^{(c)}) + B_4 \end{aligned} \tag{12}$$

where W_i and B_i represent the filters and biases of the *i-th* layer respectively. c represent the channel condition. F_i is the output feature map and $*$ means the convolution operation. The W_i has n_i filters with size of $n_{i-1} \times f_i \times f_i$ and n_0 represent the number of channels in the input frame. Rectified Linear Unit $(ReLU, max(0, x))$ is applied on the filter responses.

Here, we define the set of un-coded frame as ground truth and represented by $\{x_i\}$. The coded frames form a set called $\{y_i\}$ and each x_i has its corresponding y_i. We choose Mean Squared Error (MSE) as the loss function:

$$L(\Omega) = \frac{1}{n}\sum_{i=1}^{n}\left\|F(y_i^{(c)};\Omega) - x_i^c\right\|^2 \tag{13}$$

Here $\Omega = \{W_1, W_2, W_3, W_4, B_1, B_2, B_3, B_4\}$, n is the number of training samples. The loss is minimized using stochastic gradient descent with the standard back propagation.

3 Experimental Result

In experiments, we evaluate the performance of our proposed method in video unicast and multicast. We compare our scheme with SoftCast and H.264 which use standard 802.11 PHY layer with FEC and QAM modulations. The experiment method broadcasts the same video to users with different channel SNR.

We use over 400 images of size 180×180 for training. The training images are decomposed into 64×64 sub-images and then the compressed and transmitted samples are generated from the training samples with SoftCast decoder. A total of 204,800 patches are sampled with a stride of 20 on the training images. The learning rate is set as 10^{-5} in the last layer and 10^{-4} in the remaining layers. The convolutional network settings are $f_1 = 9, f_{1'} = 7, f_2 = 1, f_3 = 5, n_1 = 64, n_{1'} = 32, n_2 = 16, n_3 = 1$. A specific network is trained for each 5 dB CSNR range.

The test sequences are 'foreman_cif.yuv', 'news_cif.yuv, 'mother_cif.yuv and 'bus_cif.yuv, respectively. The video frame rate is 30 Hz. The coded signal is transmitted over OFDM channel with AWGN.

We compare the proposed method with SoftCast and the conventional frameworks based on H.264. For conventional framework, we implement 4 recommended combination of channel coding and modulation of 802.11. We calculate the corresponding bit-rate according to the bandwidth for H.264 encoder. For the proposed method, there is no bit-rate but only channel symbol rate. The video PSNR of each framework under different channel SNR is given in Fig. 3 which shows that all the four conventional transmission approaches suffer from a very serious cliff effect. In contrast, the SoftCast and the proposed method do not suffer from the cliff effect. As the channel SNR increases, the reconstruction quality increases accordingly. Figure 5 gives the performance comparison on different video sequences. Since GSR based compression artifact reduction scheme and deep convolutional networks based transmission artifact reduction scheme performs well in decoded frame restoration. Figure 6(c) and (d) shows that under similar PSNR, the proposed method not only reduce most of the artifacts, but also provides better reconstruction on both edges and textures.

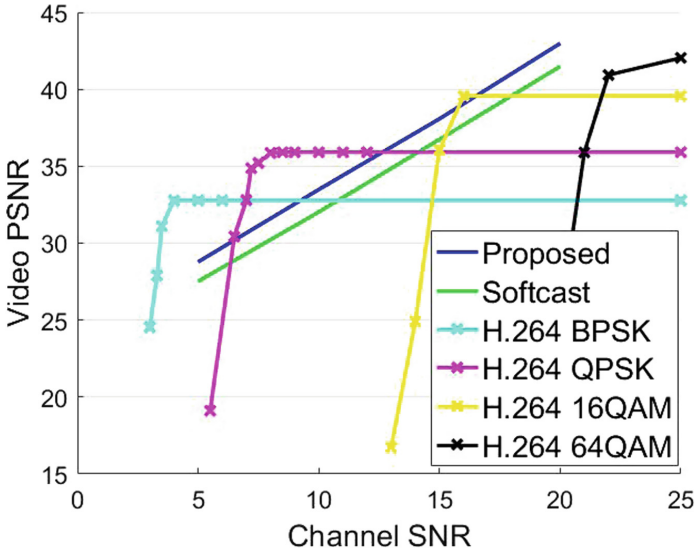


Fig. 3. Robustness comparison

We then let the frameworks serve a group of three receivers with diverse channel SNR. The channel SNR for each receiver is 5 dB, 10 dB and 20 dB. In conventional frameworks based on H.264, the server transmits the video stream by using BPSK. It cannot use higher transmission rate because otherwise the 5 dB user will not be able to decode the video. In both of SoftCast and the proposed method, the server can accommodate all the receivers simultaneously. Using our method, the 5 dB user will get slightly lower reconstruction quality than using H.264 based conventional frameworks. However, the 10 dB and 20 dB users get better reconstruction quality by using our method than conventional frameworks. The test result is given in Fig. 4.

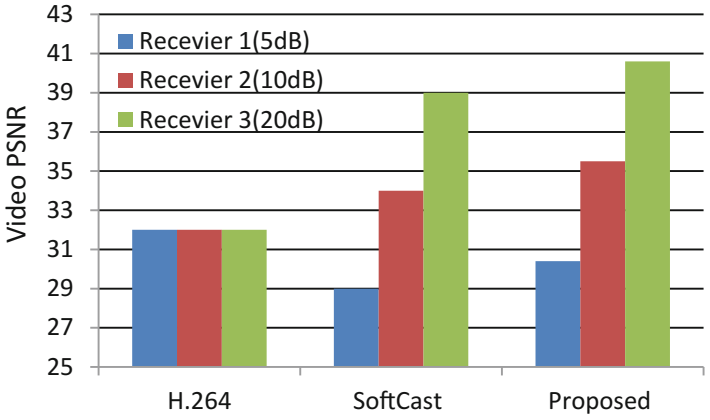
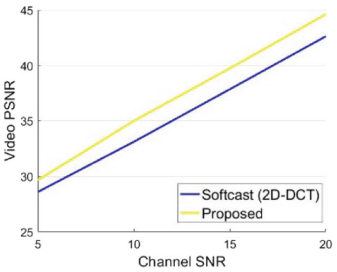
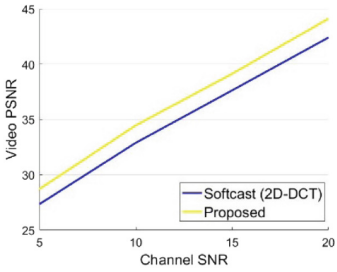


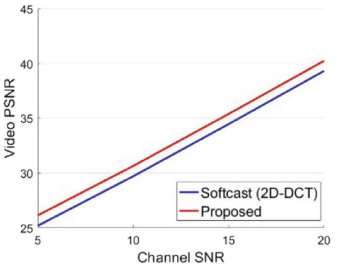
Fig. 4. Multicast comparison



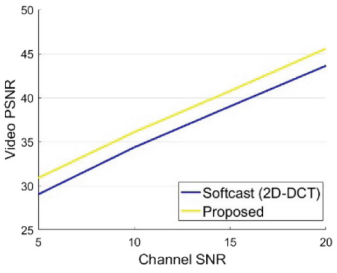
(a) foreman



(b) news



(c) bus



(d) waterfall

Fig. 5. Broadcast performance on different sequences

(a) SoftCast (Same CSNR)

(b) Proposed (Same CSNR)

(c) SoftCast (Similar PSNR)

(d) Proposed (Similar PSNR)

Fig. 6. Visual quality of 'foreman_cif'

4 Conclusion

The proposed scheme in this paper provides a novel method for video broadcasting. By utilizing band-based coding, power allocation, group based sparse representation and convolutional networks, it fully exploits the ability of deep learning and sparse coding to deal with vision problems and effectively reduces the artifacts caused by compression and transmission. By utilizing soft broadcast, it achieves good broadcast performance and avoids the cliff effect. Finally, it achieves wireless video broadcast system which matches modern wireless video broadcast demand perfectly.

References

1. Digital Video Broadcasting (DVB). http://www.etsi.org/deliver/etsi_en/300700_300799/300744/01.06.01_60/en_300744v010601p.pdf
2. Shacham, N.: Multipoint communication by hierarchically encoded data. In: International Conference on Computer Communications, vol. 3, pp. 2107–2114 (1992)
3. McCanne, S., Jacobson, V., Vetterli, M.: Receiver-driven layered multicast. In: Conference Proceedings on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 117–130. ACM (1996)
4. Wu, F., Li, S., Zhang, Y.Q.: A framework for efficient progressive fine granularity scalable video coding. IEEE Trans. Circuits Syst. Video Technol. **11**, 332–344 (2001)
5. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circuits Syst. Video Technol. **17**, 1103–1120 (2007)
6. Ramchandran, K., Ortega, A., Uz, K., Vetterli, M.: Multiresolution broadcast for digital hdtv using joint source-channel coding. In: IEEE International Conference on Communications, vol. 1, pp. 556–560 (1992)
7. Jakubczak, S., Katabi, D.: A cross-layer design for scalable mobile video. In: International Conference on Mobile Computing and Networking, pp. 289–300. ACM (2011)
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super resolution using deep convolutional networks. arXiv:1501.00092 (2014)
9. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
10. Fan, X., Wu, F., Zhao, D.: D-cast: DSC based soft mobile video broadcast. In: International Conference on Mobile Ubiquitous Multimedia, pp. 226–235 (2012)
11. Peng, X., Xu, J., Wu, F.: Line-cast: line-based semi-analog broadcasting of satellite images. In: International Conference on Image Processing, pp. 2929–2932 (2012)
12. Jung, C., Jiao, L., Qi, H., Sun, T.: Image deblocking via sparse representation. Signal Process. Image Commun. **27**, 663–677 (2012)
13. Liu, X., Wu, X., Zhao, D.: Sparsity based soft decoding of compressed images in transfer, domain. In: International Conference on Image Processing, pp. 563–566 (2013)
14. Zhang, J., Zhao, D., Gao, W.: Group based sparse representation for image restoration. IEEE Trans. Image Process. **4**, 1–2 (2014)