LETTER

Standard-Compliant Multiple Description Image Coding Based on Convolutional Neural Networks*

Ting ZHANG[†], Huihui BAI^{†a)}, *Nonmembers*, Mengmeng ZHANG^{††}, *Member*, and Yao ZHAO[†], *Nonmember*

SUMMARY Multiple description (MD) coding is an attractive framework for robust information transmission over non-prioritized and unpredictable networks. In this paper, a novel MD image coding scheme is proposed based on convolutional neural networks (CNNs), which aims to improve the reconstructed quality of side and central decoders. For this purpose initially, a given image is encoded into two independent descriptions by sub-sampling. Such a design can make the proposed method compatible with the existing image coding standards. At the decoder, in order to achieve high-quality of side and central image reconstruction, three CNNs, including two side decoder sub-networks and one central decoder sub-network, are adopted into an end-to-end reconstruction framework. Experimental results show the improvement achieved by the proposed scheme in terms of both peak signal-to-noise ratio values and subjective quality. The proposed method demonstrates better rate central and side distortion performance.

key words: image coding, multiple description (MD) coding, convolutional neural network

1. Introduction

Multiple description (MD) coding is a significant technique for robust transmission over error-prone channels. It can effectively combat packet loss without any retransmission, thus satisfying the demand of real time services and mitigating the impact of network congestion [1]. The classical scheme of MD coding is to encode one source into two bit streams (descriptions) which are mutually refinable and can be decoded independently. Those descriptions are then transmitted over separate channels. On the receiver side, if only one stream is received correctly, then the source can be reconstructed with a degraded but acceptable quality by side decoder, and the resultant distortion is called side distortion. Otherwise, if more than one stream is received, higher quality is obtained, with a so-called central distortion.

Many methods of multiple description image coding have been proposed, mainly including MD scalar quantizer [2], MD lattice vector quantizer [3], MD based on pairwise correlating transforms [4], and MD based on FEC [5], which are reported to have achieved better performance.

Manuscript received February 7, 2018.

Manuscript revised June 8, 2018.

Manuscript publicized July 19, 2018.

[†]The authors are with Beijing Jiaotong University, Beijing, 100044, China.

^{††}The author is with North China University of Technology, Beijing, 100144, China.

*This work was supported in part by Key Innovation Team of Shanxi 1331 Project (KITSX1331) and Fundamental Research Funds for the Central Universities (2018JBZ001).

a) E-mail: hhbai@bjtu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2018EDL8028

However, these works cannot fully consider the compatibility with the existing image coding standards, which limits their practical applications.

To overcome this limitation, some standard-compliant MD image coding schemes have been proposed and achieved promising results [1], [6], [7]. Because some image compression standards, such as JPEG and JPEG 2000, require the input image to be a rectangular grid of pixels, source splitting is adopted to generate multiple descriptions in those methods. For example, in [1], the source samples are split into odd- and even-indexed subsets which then encoded into two diverse descriptions. For side reconstruction, interpolation is used to decode from any subset of the descriptions. For central reconstruction, decoding from both channels requires the interleaving of the samples.

In recent years, deep learning not only have achieved great achievements in high-level computer vision tasks, but also been successfully utilized in low-level tasks. For example, Jiang *et al.* [8] propose an end-to-end compression framework using two CNNs, which have achieved high-quality image compression at low bit rates. In addition, Dong *et al.* [10] propose a CNN based image super-resolution (SR) method named SRCNN, which has achieved significant boost of performance in restoration quality by learning an end-to-end mapping between low and high-resolution images. Zhao *et al.* [11] propose a CNN-based depth image artifact removal scheme using two sub-networks. All these studies seem to illustrate the efficiency of adopting CNNs into image processing.

Motivated by the successful application of CNNs, we propose an end-to-end MD coding framework, called MD-CNN, which consists of three CNNs to improve the reconstructed quality of side and central decoders. An input image is split into two subsets by odd/even separation, which then can be encoded into two diverse descriptions by standard image codecs. If either of descriptions is lost, the received one will be fed into the side decoder sub-network to improve the quality of side reconstruction. Besides, if both descriptions are received, they can be processed by the central decoder sub-network. To the best of our knowledge, this is the first deep learning-based MD coding method. Compared with classical standard-compliant schemes, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values reconstructed by the proposed scheme can be improved significantly in both side and central channels.

The remainder of this paper is organized as follows. Section 2 elaborates the proposed scheme, including the ar-

chitectures of MDCNN. Experimental results are reported in Sect. 3. Finally, conclusions are drawn in Sect. 4.

2. Proposed Scheme

In this section, we first introduce the architecture of the proposed MD coding scheme based on CNNs and then present the details of the reconstruction networks.

2.1 Overview of Proposed Scheme

As shown in Fig. 1 (a), the proposed multiple description scheme consists of image codecs and three CNNs. In order to be compatible with the existing image coding standards, many sub-sampling methods can be adopted to generate multiple description. Since our research focuses on improving the quality of side and central image reconstruction, we choose a more traditional and simple sub-sampling method, namely odd and even separation, to obtain two subsets. Specifically, for the input image X , odd- and even-numbered samples are separated horizontally to generate two subsets which then can be transmitted at two channels. Any standard image codec can be used to compress the sub-sampled images with a half size to produce *Description 1* and 2. In the decoder, we propose a novel reconstruction network based on convolutional neural networks named MDCNN. If only *Description 1* or 2 can be received, the side decoder sub-networks of D_1 and D_2 can be used to improve the quality of side reconstruction respectively. Otherwise, when both representations are received, the central decoder sub-network D_0 will be used to reconstruct the original image with better quality than side decoder.

Fig. 1 (a) Block diagram of our proposed scheme; (b) the architecture of the proposed MDCNN.

2.2 Reconstruction Network

Given *Description 1* and 2, our goal is to reconstruct the input image X whether one or two descriptions received at the decoder side. Meanwhile, the more the descriptions are received, the higher quality of the reconstruction can be achieved. As illustrated in Fig. 1 (b), the proposed MDCNN consists of three sub-networks, which are explained in the following sections.

2.2.1 Side Decoder Sub-Networks

For the two decoded images X_1 and X_2 , we first upscale it to the desired size using bicubic interpolation. Then, our goal is to learn two mappings φ_1 and φ_2 which can be formulated as:

$$Y_1 = \varphi_1(X_1; \Theta_1), Y_2 = \varphi_2(X_2; \Theta_2) \quad (1)$$

where Y_1 and Y_2 represent the side decoder results, Θ_1 and Θ_2 are the parameters of side decoder sub-networks D_1 and D_2 respectively.

Since the two side decoder sub-networks are the same, we will take D_1 as an example to describe the network architecture. As shown in Fig. 1 (b), the D_1 consists of five convolutional layers. The first and last layers are similar to SRCNN [10], which are adopted to generate feature representations and reconstruct the source image respectively. The middle three layers are used as non-linear mapping layers. Considering the degradation of input images quality caused by compression, we design a reasonably filter with size 3×3 for the second and forth layers to grasp richer structural information, which in turn facilitate the accurate reconstruction of high-quality images. In order to find a trade-off between the compression performance and the computational complexity of the mapping, the third layer is retained with kernel size 1×1 compared with the SRCNN [10]. The convolutional operations are set with stride 1 and padding 1 which are used to keep the size of reconstruction image same as the ground truth image. ReLU [15] is utilized as an activation function after the convolutional layers to realize nonlinear mapping, except the last layer.

2.2.2 Central Decoder Sub-Network

The central decoder sub-network D_0 is designed to enhance the quality of the central decoder when both descriptions are received. The two side reconstruction images are fed into D_0 which are concatenated on the channel's dimension. The aim of D_0 is to learn a mapping φ_0 which can be formulated as:

$$Y_0 = \varphi_0(Y_1, Y_2; \Theta_0) \quad (2)$$

where Y_0 represents the central decoder result, and Θ_0 is the parameters of D_0 . The structure of D_0 is similar to the D_1 or D_2 , except the width (i.e., the number of filters in a layer) of

the layers considering the capacity of the network.

2.2.3 Loss Function

All of the sub-networks in MDCNN are designed to make the reconstructed image as similar as possible to the original image. Therefore, a corresponding optimization goal is needed for each sub-network by minimizing a distance l_p ($p = 1$ or $p = 2$) between the original and reconstructed image as follows:

$$\begin{aligned} L_p(X, Y) &= \frac{1}{M \times N} \sum_{i=0}^2 l_p(X, Y_i) \\ &= \frac{1}{M \times N} \sum_{i=0}^2 \|Y_i - X\|_p^p \end{aligned} \quad (3)$$

where Y_1 or Y_2 represents the reconstruction of side decoder, and Y_0 represents the reconstruction results of central decoder, and $M \times N$ is the size of the input image patch of X .

In [16], it has been reported that l_2 loss always leads to unnatural blurriness of the output image. Considering that l_1 loss function is more robust and is generally less sensitive to outliers than l_2 loss, we choose l_1 loss in our network to keep the reconstructed image closer to the original image, and maintain the image detail well.

3. Experimental Results

To evaluate the performance of the proposed scheme, we conduct experimental comparisons against the conventional method which uses interpolator and odd/even interleaving in the side and central decoder respectively. The training details and experimental results are shown in the following sections.

3.1 Implementation Details

Our training datasets consist of 91 images from Yang *et al.* [14], 200 images from the training set of Berkeley Segmentation Dataset [12], and 100 images from General-100 dataset [13]. A total of 362856 patches of size 32×32 are sampled with a stride of 20 on the 391 images as well as their augmentation (horizontal flipping and vertical flipping). The training is performed using TensorFlow [17] on GPU. The weights are optimized by the Adam [18] solver, where $\beta_1 = 0.5$, $\beta_2 = 0.999$ and the learning rate is set to be $2e-4$. For testing, two datasets of SET5 [19] and SET14 [20] are used to prove the generalization ability of the scheme.

3.2 Objective and Visual Quality Comparison

In our experiment, we use JPEG codec with different quality factor (QF) to achieve different target total bit rate. And then the corresponding network for different value of QF can be trained. Here, we compare the performance of the proposed scheme with the traditional method in datasets of SET5 and

SET14. According to the MD coding quality assessment, we compare not only the rate central distortion performance when two descriptions can be received correctly, but also the rate side distortion performance when only one description can be received.

The objective quality comparisons have been shown in Table 1 and Table 2, which tabulate the pairs of central and side distortion results in terms of both PSNR and SSIM with respect to various QF for the test image sequences of SET5 and SET14, respectively. As seen from Table 1, for side distortion, the proposed method achieves on average 1.435–2.449dB gains in PSNR and 0.029–0.135 gains in SSIM compared against the basic method, and for central distortion, it achieves on average 1.023–1.981dB and 0.011–0.174 gains in PSNR and SSIM, respectively. Meanwhile, the performance on SET14 of our MDCNN is shown

Table 1 Rate distortions performance of PSNR (dB)/SSIM for SET5

SET5 QF	Proposed Scheme		Basic Scheme	
	Central	Side	Central	Side
5	26.715/0.767	25.969/0.741	24.734/0.655	24.533/0.670
25	31.429/0.872	30.509/0.860	30.251/0.843	28.816/0.830
50	34.060/0.920	32.060/0.901	32.497/0.893	29.827/0.866
75	36.130/0.941	32.983/0.921	35.107/0.931	30.534/0.892
80	37.260/0.951	33.228/0.926	36.163/0.942	30.712/0.898

Table 2 Rate distortions performance of PSNR (dB)/SSIM for SET14

SET14 QF	Proposed Scheme		Basic Scheme	
	Central	Side	Central	Side
5	25.010/0.672	24.357/0.647	23.651/0.598	23.257/0.600
25	29.303/0.835	27.776/0.796	28.577/0.811	26.538/0.769
50	31.606/0.883	28.830/0.837	30.944/0.874	27.363/0.811
75	33.811/0.923	29.724/0.870	33.849/0.922	27.944/0.840
80	35.025/0.936	30.012/0.878	34.978/0.935	28.085/0.847

Fig. 2 Objective quality comparisons.

Fig. 3 Visual quality comparisons for butterfly at total rate 0.5 bpp.

in Table 2, which demonstrates that the proposed scheme can gain around 1.101–1.780dB and 0.026–0.062 in PSNR and SSIM in side decoder, and gain around 0.662–1.360dB and 0.001–0.087 in central decoder. As a result, for side distortion, the proposed method outperforms the basic method at the same QF. For central distortion, in many cases, our method is better than the basic method. Although our PSNR is slightly inferior to the basic method on QF = 75 for SET14, our SSIM is better than the baseline.

Figure 2 shows the central and side distortion performance of the proposed scheme against the traditional method for the image Butterfly, Pepper, Lena and Barbara from the test datasets. It can be seen that the side and central PSNR outperform the conventional method for the given same or a very close total rate. Furthermore, the visual quality comparisons at the total rate of 0.5 bits/pixel (bpp) for Butterfly are provided in Fig. 3. We can see in the proposed method the blocking artifacts have been removed and more textures can be preserved.

4. Conclusion

In this paper, we propose a novel CNNs based MD image coding scheme, which consists of three sub-networks to improve the reconstruction quality of the side and central decoders. Experimental results demonstrate that the proposed scheme achieves better performance in both single and central channels than the conventional method, and it is compatible with the existing image coding standards. Our work indicates the effectiveness of deep neural networks in MD coding, which will inspire other researchers to design better

architectures to improve the performance of MD coding.

References

- [1] V.K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol.18, no.5, pp.74–93, Sept. 2001.
- [2] C. Lin, Y. Zhao, and C. Zhu, "Two-stage diversity-based multiple description image coding," *IEEE Signal Process. Lett.*, vol.15, no.1, pp.837–840, Aug. 2008.
- [3] H. Bai, C. Zhu, and Y. Zhao, "Optimized multiple description lattice vector quantization for wavelet image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol.17, no.7, pp.912–917, July 2007.
- [4] Y. Wang, M.T. Orchard, and A.R. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," *Proc. 1st Workshop Multimedia Signal Process.*, pp.419–424, June 1997.
- [5] S.-H. Chang, P.C. Cosman, and L.B. Milstein, "Performance analysis of n -channel symmetric FEC-based multiple description coding for OFDM networks," *IEEE Trans. Image Process.*, vol.20, no.4, pp.1061–1076, April 2011.
- [6] X. Zhang and X. Wu, "Standard-compliant multiple description image coding by spatial multiplexing and constrained least-squares restoration," *Proc. MMSP*, pp.349–354, Oct. 2008.
- [7] T. Tillo and G. Olmo, "Data-dependent pre- and postprocessing multiple description coding of images," *IEEE Trans. Image Process.*, vol.16, no.5, pp.1269–1280, May 2007.
- [8] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An end-to-end compression framework based on convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.
- [9] G. Toderici, D. Vincent, N. Johnston, S.J. Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," *arXiv: 1608.05148*, 2016.
- [10] C. Dong, C.C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.38, no.2, pp.295–307, Feb. 2016.
- [11] L. Zhao, J. Liang, H. Bai, Wang, Anhong, and Y. Zhao, "Convolutional neural network-based depth image artifact removal," *Proc. IEEE Int. Conf. Image Process.*, 2017.
- [12] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *Proc. IEEE Int. Conf. Comput. Vis.*, pp.416–423, 2001.
- [13] R. Timofte, V. De, and L.V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," *Proc. IEEE Int. Conf. Comput. Vis.*, pp.1920–1927, 2013.
- [14] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image superresolution via sparse representation," *IEEE Trans. on Image Process.*, vol.19, no.11, pp.2861–2873, 2010.
- [15] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proc. Int. Conf. Mach. Learn.*, pp.807–814, 2010.
- [16] M. Mathieu, C. Couprie, and Y. LeCun., "Deep multi-scale video prediction beyond mean square error," *arXiv: 1511.05440*, 2015.
- [17] M. Abadi, A. Agarwal, P. Barham, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.
- [18] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.
- [19] M. Bevilacqua, A. Roumy, C. Guillemot, and M.L. Alberi-Morel, "Lowcomplexity single-image super-resolution based on nonnegative neighbor embedding," *Proc. British Mach. Vis. Conf.*, pp.135.1–135.10, 2012.
- [20] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," *Proc. Int. Conf. Curves Surfaces*, vol.6920, pp.711–730, 2012.