LETTER

Image Quality Improvement for Capsule Endoscopy Based on Compressed Sensing with K-SVD Dictionary Learning

Yuuki HARADA[†], *Nonmember*, **Daisuke KANEMOTO**[†a)], *Member*, **Takahiro INOUE**[††], **Osamu MAIDA**[†], *Nonmembers*, *and* **Tetsuya HIROSE**[†], *Member*

SUMMARY Reducing the power consumption of capsule endoscopy is essential for its further development. We introduce K-SVD dictionary learning to design a dictionary for sparse coding, and improve reconstruction accuracy of capsule endoscopic images captured using compressed sensing. At a compression ratio of 20%, the proposed method improves image quality by approximately 4.4 dB for the peak signal-to-noise ratio.

key words: *compressed sensing, capsule endoscopy, K-SVD, high quality*

1. Introduction

Capsule endoscopy (CE) [1] has recently attracted attention as a promising diagnostic method. CE acquires images of the gastrointestinal tract and transmits the images wirelessly, allowing observations just by swallowing a small capsule. Thus, CE is less invasive than conventional fiberscopic endoscopy. Nevertheless, CE faces problems due to the high power consumption of the radio-frequency transmitter, which should transmit multiple images wirelessly. A critical problem of CE is the short device recording time. If the battery energy of a CE device is depleted, the target organ or region may not be observed. Another problem is the large capsule size. Large power consumption makes the capsule larger, as most of the capsule volume is occupied by batteries. Current CE devices (e.g., PillCam SB 3; length: 26 mm, outer diameter: 11 mm) are larger than any pharmaceutical capsule and difficult to swallow, especially for patients with weak swallowing function. Furthermore, CE device retention has been reported in patients with intestinal stricture [2]. To reduce power consumption, the amount of data sent by the transmitter should be reduced, because 90% of the power is consumed during data transmission [3]. Thus, extensive research on image compression for CE has been conducted [4]. For instance, processing CE images with compressed sensing (CS) [5] is a promising option. However, it has not been explored thoroughly. In this study, we further investigated CS to improve its applicability for CE imaging.

Manuscript received April 20, 2021.
Manuscript revised July 20, 2021.
Manuscript publicized October 1, 2021.

[†]The authors are with the School of Engineering, Division of Electronic and Information Engineering, Osaka University, Suita-shi, 565-0871 Japan.

[††]The author is with the Office of Industry-Academia-Government Collaboration, Co-Creation Bureau, Osaka University, Suita-shi, 565-0871 Japan.

a) E-mail: dkanemoto@eei.eng.osaka-u.ac.jp

DOI: 10.1587/transfun.2021EAL2033

CS reconstructs a signal from only a small set of measurements under the sparsity constraint [6]–[8]. CS is suitable for CE because it enables compression through simple computations, unlike conventional image compression techniques [5]. The quality of the reconstructed image after CS mainly depends on the data sparsity, whose improvement is thus essential to obtain high-quality reconstructed CE images for diagnosis. As image data are not originally sparse, they should be converted into a sparse representation by sparse coding, which describes the original data into a sparse representation consisting of a sparse coefficient vector and a matrix, which is called the dictionary. In [5], a YUV image is converted into its sparse representation by using a fast Fourier transform (FFT) dictionary. Although this concept is simple and effective, a suitable sparse representation cannot be obtained in image channel Y and V. Consequently, the image reconstruction accuracy is deteriorated. We aim to improve the reconstruction accuracy by adapting a dictionary to CE images. To this end, we use K-SVD (singular value decomposition) dictionary learning [9], whose usefulness was confirmed through the study on compressed sensing for electroencephalography [10], to replace the FFT dictionary.

The remainder of this paper is organized as follows. Section 2 briefly describes CE with CS and its problems. Then, we introduce the proposed CS method for CE and justify the use of K-SVD dictionary learning. Section 3 presents experimental results, and Sect. 4 presents the conclusions.

2. Capsule Endoscopy with Compressed Sensing

2.1 Dictionary Selection

The sparsity of the representation coefficient vector depends on the dictionary. In [5], RGB images are converted into YUV images and sparse coded using an FFT dictionary aiming to obtain a sparse representation. In the YUV color space, the RGB image information is converted into three channels, namely a structural channel Y, a blue chroma channel U (=B-Y), and a red chroma channel V (=R-Y). The blue chroma channel U can then be represented sparsely because endoscopic images contain very few blue components. However, channel Y and V cannot be represented sparsely enough in the FFT basis, leading to critical image degradation after reconstruction. Consequently, to compress the data without compromising the accuracy of the diagnosis, we must reduce

the compression ratio (CR). Thus, to achieve a high CR, a more appropriate sparse dictionary should be devised to replace the FFT dictionary. Either an existing dictionary or a new dictionary that fits a dataset of samples can be used. In magnetic resonance imaging (MRI), an existing dictionary has been successfully used, representing the best CS application in the medical field [11]. Sparse coding for MRI is based on a wavelet dictionary and achieves excellent performance, because MR images resemble line drawing in that their intensity varies only at tissue boundaries. In contrast, the diverse features of CE images hinder the generation of sparse representations using existing dictionaries. Thus, a dictionary should be specifically designed for CE by adapting its contents to the CE image characteristics.

Dictionary learning can be used to design a dictionary. Let $\mathbf{X} \in \mathbb{R}^{N\times P}$ be a set of training data matrices, where each column of training data matrix $\{\boldsymbol{x}_i\}_{i=1}^{P}$ represents small image patches segmented from the training image. $\mathbf{X}$ can be represented by the product of an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{N\times T}(N < T)$ and a sparse representation coefficient matrix $\mathbf{S} \in \mathbb{R}^{T\times P}$ that contains P sparse representations of small image patches, $\{\boldsymbol{s}_i\}_{i=1}^{P}$, obtaining $\mathbf{X} = \mathbf{DS}$. Dictionary learning aims to find dictionary matrix $\mathbf{D}$ based on training data. K-SVD dictionary learning computes the dictionary matrix by determining $\mathbf{D}$ and $\mathbf{S}$ that minimize residual Frobenius norm $\|\mathbf{X} - \mathbf{DS}\|_F^2$, as expressed in (1).

$$\min_{\mathbf{D},\mathbf{X}}\{\|\mathbf{X} - \mathbf{DS}\|_F^2\} \text{ subject to } \forall i, \|\boldsymbol{s}_i\|_0 \le K \tag{1}$$

where K is the upper limit on the number of non-zero elements contained in $\boldsymbol{s}_i$. If K is small enough, dictionary $\mathbf{D}$ can probably represent signals similar to the training data sparsely.

2.2 Proposed Method

In the proposed method, compression and reconstruction are conducted as follows. First, CE images are represented in the YUV color space and divided into patches. Small patches reduce the reconstruction complexity, whereas large patches often improve the reconstruction accuracy. In [5], the CE images were divided into 8×8-pixel patches to balance the tradeoff between computation time and reconstruction accuracy. However, we performed simulations using 16×16-pixel patches, which provide higher reconstruction accuracy. The reconstruction time is became less important than reconstruction accuracy because reconstruction is performed by an external workstation whose computing power has been improved since the [5] was published. Then, the patches are compressed by multiplying Gaussian distribution sampling matrix $\mathbf{A} \in \mathbb{R}^{N\times M}$, with $M/N \times 100(\%)$ being the CR. In this study, we compressed the CE images according to the CRs listed in Table 1. For example, a CR of 30% for an RGB image represents a compression at 10% for channels U and V, and 70% for channel Y, instead of a uniform compression at 30% for all YUV channels. The reason for varying the CR is that each channel of YUV images has different visual properties and degrees of compressibility. Channels U and V, which represent chroma, have less visual impact than channel Y. Moreover, in CS, the high sparsity of channel U facilitates reconstruction. During reconstruction, compressed data are sparse coded using the K-SVD dictionary and reconstructed by orthogonal matching pursuit [12]. Finally, the YUV image is converted back into an RGB image. In this paper, the proposed method is called K-SVD-CSCE, and that proposed in [5] is called FFT-CSCE.

Table 1 CRs per channel in YUV image.

CR of overall image (%)	10	20	30	40	50
CR of channel Y (%)	10	40	70	100	100
CR of channel U (%)	10	10	10	10	10
CR of channel V (%)	10	10	10	10	40

3. Experimental Results

We conducted experiments using MATLAB on 109 images of 448×448 pixels. We used the images which correspond to small intestine recorded by double-balloon endoscopy and are derived from an anonymous patient with Crohn's disease. From the images, 99 were used for dictionary learning, obtaining 77,616 patches. The remaining 10 images were used for a compression and reconstruction test. We obtained the reconstruction accuracy for each image and analyzed the average results. The number of columns of the K-SVD dictionary was set to 1024, and the number of iterations of K-SVD dictionary learning was set to 50.

We used two measures to evaluate reconstruction accuracy: peak signal-to-noise ratio (PSNR) and mean structural similarity (MSSIM) [13]. The reason for using both is to reduce the influence that each measure has on the results. The PSNR is given by

$$PSNR(L,\hat{L}) = 10\log_{10}\left(\frac{255^2}{MSE(L,\hat{L})}\right), \tag{2}$$

$$MSE(L,\hat{L}) = \frac{\sum_{w=1}^{W}\sum_{h=1}^{H}\left(L_{w,h} - \hat{L}_{w,h}\right)^2}{W \times H}. \tag{3}$$

The PSNR simply compares $H \times W$ images L and $\hat{L}$ pixel by pixel. Its value ranges from 0 to ∞ and PSNR of ∞ indicates that images L and $\hat{L}$ are equal. This is a mathematically intuitive measure, but it fails to suitably distinguish between large inaccuracies confined to a small area of the image, versus small inaccuracies over a large area.

The structural similarity compares pairs of images regarding luminance, contrast, and structure on 8×8-pixel patches. The MSSIM averages the structural similarity values across patches contained in an image to be evaluated. The MSSIM is given by

$$SSIM(l,\hat{l}) = \frac{(2\mu_l\mu_{\hat{l}} + c_1)(2\sigma_{l\hat{l}} + c_2)}{(\mu_l^2 + \mu_{\hat{l}}^2 + c_1)(\sigma_l^2 + \sigma_{\hat{l}}^2 + c_2)}, \tag{4}$$

$$MSSIM(L,\hat{L}) = \frac{1}{B}\sum_{b=1}^{B} SSIM(l_b,\hat{l}_b), \tag{5}$$

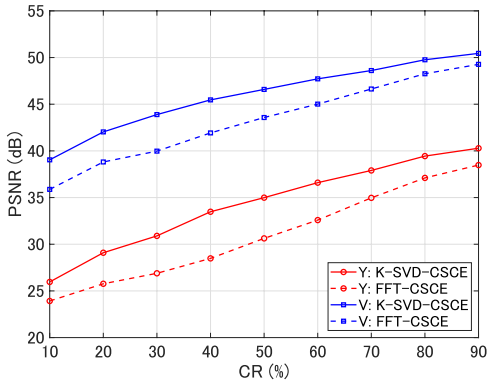

Fig. 1 PSNR of YUV images reconstructed by FFT-CSCE and K-SVD-CSCE.

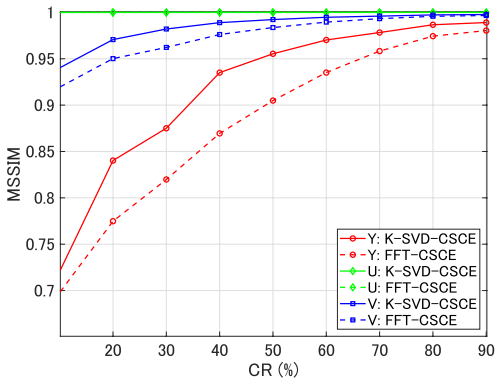

Fig. 2 MSSIM of YUV images reconstructed by FFT-CSCE and K-SVD-CSCE.

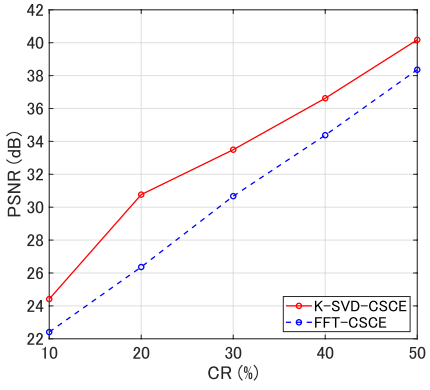

Fig. 3 PSNR of RGB images reconstructed by FFT-CSCE and K-SVD-CSCE.

where l_b and $\hat{l}_b$ represent the b-th patch of images L and $\hat{L}$, respectively, μ is the average pixel value in a patch, σ_l is the standard deviation of the pixel values, $\sigma_{l\hat{l}}$ is the covariance between images l and $\hat{l}$, and c_1 and c_2 are constants to stabilize the structural similarity value. The MSSIM value ranges from 0 to 1 and MSSIM of 1 indicates that images L and $\hat{L}$ are equal. The MSSIM complements the PSNR but is insensitive to changes in chroma.

First, we compressed channels Y, U, and V at CRs from

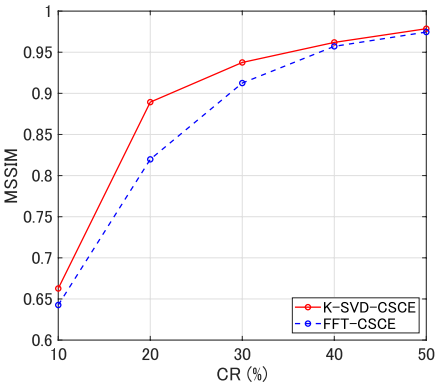

Fig. 4 MSSIM of RGB images reconstructed by FFT-CSCE and K-SVD-CSCE.

Fig. 5 RGB image reconstructed by FFT-CSCE at CR of 20%.

10 to 90% in 10% increments using FFT-CSCE and K-SVD-CSCE, and compared the reconstruction accuracies of the two methods. The PSNR and MSSIM values per channel are shown in Figs. 1 and 2, respectively. In Fig. 1, the PSNR values for channel U are not displayed because their large values reduce the readability of the graph. These figures show that the proposed K-SVD-CSCE achieves improvements for both channels Y and V regardless of the CR, for both PSNR and MSSIM. In particular, 40% compression for channel Y exhibits the greatest improvement in reconstruction accuracy, with a PSNR value of approximately 5 dB and MSSIM value of 0.066. Regarding channel U, no difference in PSNR and MSSIM was observed between the results of K-SVD-CSCE and FFT-CSCE. Thus, the K-SVD dictionary does not improve U-channel reconstruction. Moreover, the K-SVD dictionary is computationally more expensive than the FFT dictionary because it expands the size of the matrix to be computed during sparse coding. Thus, we used the K-SVD dictionary for reconstruction of channels Y and V and the FFT dictionary for reconstruction of channel U.

Second, we reconstructed RGB color images, obtaining the PSNR and MSSIM results shown in Figs. 3 and 4. The proposed K-SVD-CSCE provides higher PSNR values than

Fig. 6 RGB image reconstructed by KSVD-CSCE at CR of 20%.

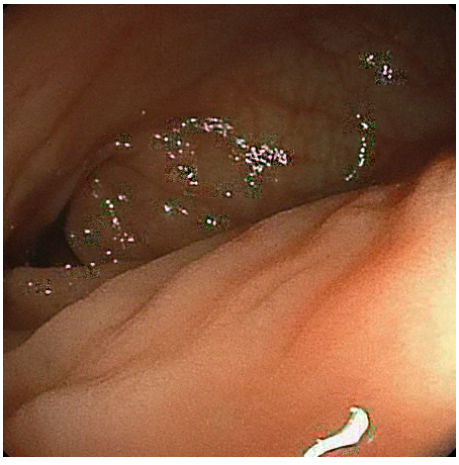

Fig. 7 RGB image reconstructed by FFT-CSCE at CR of 30%.

Fig. 8 RGB image reconstructed by K-SVD-CSCE at CR of 30%.

FFT-CSCE regardless of the CR for an RGB image. In contrast, at CRs of 40% and 50% for an RGB image, there is almost no difference in the MSSIM between K-SVD-CSCE and FFT-CSCE. This occurs because the structural channel Y is uncompressed when the CRs for an RGB image are 40 and 50%, as listed in Table 1, and the improvement only appears in chroma; MSSIM is insensitive to chroma defects. Unlike MSSIM, PSNR shows the improvement in chroma achieved by the proposed method. The best improvement of about 4.4 dB PSNR and 0.07 MSSIM is observed at a CR of 20% for an RGB image. This is because, as shown in Table 1, when the whole RGB image is compressed to 20%, channel Y is compressed to 40% with the greatest improvement in image quality as in Figs. 1 and 2. Figures 5 to 8 show reconstructed RGB images at CR of 20% and 30%, respectively. Most of the noisy patches around the center in Figs. 5 and 7 are removed in Figs. 6 and 8, and the noise is barely noticeable especially in Fig. 8.

4. Conclusions

We propose a reconstruction method for CE images using CS. The proposed method improves the image quality by 4.4 dB for PSNR and 0.07 for MSSIM at a CR of 20% for an RGB image and achieves a high CR while mitigating reconstruction degradation compared with an existing method. Thus, the proposed method may be suitable for further compressing images to reduce power consumption in CE devices.

Acknowledgments

This work was supported by JKA and its promotion funds from AUTO RACE.

References

[1] K.D. Robertson and R. Singh, "Capsule endoscopy," StatPearls [Internet], Aug. 2020, https://www.ncbi.nlm.nih.gov/books/NBK482306/

[2] S.F. Pasha, M. Pennazio, E. Rondonotti, D. Wolf, M.R. Buras, J.G. Albert, S.A. Cohen, J. Cotter, G. D'Haens, R. Eliakim, D.T. Rubin, and J.A. Leighton, "Capsule retention in Crohn's disease: A meta-analysis," Inflamm Bowel Dis., vol.1, no.26, pp.33–42, Jan. 2020.

[3] Z. Abdelkrima, A. Ashwag, and E. Majdi, "Low power design of wireless endoscopy compression/communication architecture," Journal of Electrical Systems and Information Technology, vol.5, no.1, pp.35–47, May 2018.

[4] M.W. Alam, M.M. Hasan, S.K. Mohammed, F. Deeba, and K.A. Wahid, "Are current advances of compression algorithms for capsule endoscopy enough? A technical review," IEEE Rev. Biomed. Eng., vol.10, pp.26–43, 2017.

[5] J. Wu and Y. Li, "Low-complexity video compression for capsule endoscope based on compressed sensing theory," 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, Minnesota, USA, pp.3727–3730, Sept. 2009.

[6] E.J. Candes and M.B. Wakin, "An introduction to compressive sampling," IEEE Signal Process. Mag., vol.25, no.2, pp.21–30, March 2008.

[7] J. Romberg, "Imaging via compressive sampling," IEEE Signal Process. Mag., vol.25, no.2, pp.14–20, March 2008.

[8] Y. Tsaig and E.J. Candes, "Extensions of compressive sensing," Signal Process., vol.86, no.3, pp.549–571, March 2006.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Process., vol.54, no.11, pp.4311–4322, Nov. 2006.

[10] K. Nagai, D. Kanemoto, and M. Ohki, "Applying K-SVD dictionary learning for EEG compressed sensing framework with outlier detection and independent component analysis," IEICE Trans. Fundamentals, vol.E104-A, no.9, pp.1375–1378, Sept. 2021.
[11] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly, "Compressed sensing MRI," IEEE Signal Process. Mag., vol.25, no.2, pp.72–82, March 2008.
[12] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," Proc. 27th Asilomar Conference on Signals, Systems and Computers, pp.40–44, Nov. 1993.
[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, April 2004.