Effectiveness of Pseudo 3D Feature Learning for Spinal Segmentation by CNN with U-Net Architecture

Naofumi Shigeta, Mikoto Kamata, and Masayuki Kikuchi
School of Computer Science, Tokyo University of Technology, Tokyo, Japan
Email: {c0115170ee, c0114161c5}@edu.teu.ac.jp, kikuchi@stf.teu.ac.jp

***Abstract*—In the medical field, automatic extraction of spinal region from CT images has been desired. Among various methods for image segmentation, one of the convolutional neural network models called U-Net [1] has been shown to attain good performance with small data set size. Previous study by Kamata *et al.* [2] applied U-Net for spine segmentation task and achieved 82.7% accuracy for unlearned CT images. However, the method had difficulty in the precision of the 3D shape. This study attempted extraction of spine region with higher precision by adopting pseudo 3D feature learning for U-Net.**

***Index Terms*—medical image processing, spine segmentation, convolutional neural network, U-Net**

I. Introduction

Recently tomographic images of the human body by CT scan are indispensable for medical diagnosis. However, the extraction of extracting the necessary parts from the captured CT image is still hard manual work. Radiologists have to label the desired region from whole organization image for hundreds of slice data per one subject. In particular, since the spine has complicated shapes and structure connected to the ribs near the thoracic vertebra, etc., it is difficult to segment such region due to only CT value.

Several studies were presented at CSI 2014 on extraction of the spinal region by various automated methods. Huang *et al.* achieved 96% of the lumbar vertebrae using Ada Boost [3], [4], Wang *et al.* achieved 92.7% of the total vertebrae extraction using multi atlas segmentation [3], [5], and Korez *et al.* reported that the healthy vertebrae (thoracic and lumbar part of the spine) by using the Canny method and some contrivances achieved an accuracy of 94% [3], [6]. Vania and colleagues conducted experiments using Convolutional Neural Network (CNN) using redundant class labels, achieved accuracy of up to 94.3% by dice coefficient for extraction of all spines, whose performance was excellent compared by existing methods such as the Level-set method [7].

Manuscript received March 5, 2019; revised August 8, 2019.

There is U-Net as CNN model used for image extraction task. It is widely used in the medical image field since it can achieve high extraction accuracy with a small number of samples and operation is fast [1]. As a previous study using U-Net, Kamata *et al.* [2] achieved the accuracy of 82.7% for unlearned data using a model with unique improvement of U-Net. Also Vania *et al.* used U-Net for the task and achieved 96.0%. As a difficult aspect in spine segmentation, Kamata *et al.* showed that there are incorrectly extracted regions that cannot be resolved only by numerical CT values, such as some bones being extract thicker than usual when creating a 3D model from actually extracted spinal images.

II. Method

In order to improve accuracy of 3D shape extraction of the spine, which was a problem in the previous research, this study attempted to solve by introducing CNN to learn the 3D shape of the spine. Initially, we performed experiments using 3D U-Net on the computer with NVIDIA GeForce GTX 1080 Ti and 32GB RAM, as the result, the total data amounted 32GB per one spine image sample. The data was too large to execute. Therefore, this study adopted pseudo 3D convolution 2-channel learning which gives 3D spatial context information to the model and performs 2D convolution from different coordinate axis directions for memory resource reduction.

A. Dataset

Ten total healthy spinal samples (Case 1 - Case 10) included in "Dataset 15: Test set for CSI 2014 Vertebra Segmentation Challenge" plus one sample with compression fracture published on SpineWeb [3] were used for the experiment in total. Three of these healthy vertebrae (Case 1-Case 3) were used for learning, and the remaining eight (7 healthy spines and 1 compression fracture spine) were used as verification data.

B. Implementation 3D Shape Information

For the sake of learning 3D features, we combined following two methods. First, we gave differentiation maps (variation maps) of the 3D data by taking the difference between two neighboring slices for the spinal data according to formula (1):

doi: 10.18178/joig.7.3.107-111

$$D_z(x,y,z) = C(x,y,z-1) - C(x,y,z+1)$$
$$D_x(x,y,z) = C(x-1,y,z) - C(x+1,y,z) \quad (1)$$
$$D_y(x,y,z) = C(x,y-1,z) - C(x,y+1,z)$$

where $D_x(x,y,z)$, $D_y(z,y,z)$, $D_z(x,y,z)$ are the value of variation maps for each axes, $C(x,y,z)$ is pixel values of each CT slices.

This study used U-Net as the method of Kamata *et al.* (Fig. 1).

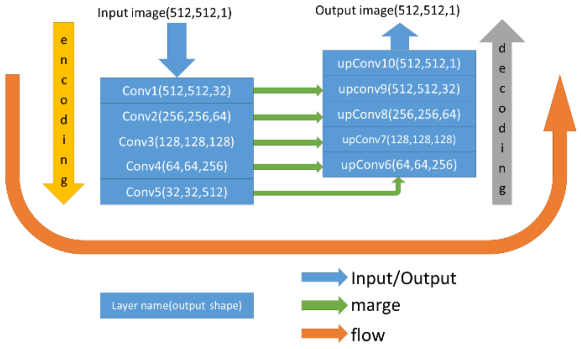

Figure 1. Conceptual diagram of U-Net architecture.

The difference between our method and the model used by Kamata *et al.* is that, as mentioned earlier, our model has two channels for each of input / output (pixel data of CT value and Variation-Map) and performs pseudo 3D learning. The model learns the given spinal data in the x, y, z axis directions respectively, and by combined them by equation (2):

$$N(x,y,z) = \frac{N_z(x,y) + N_y(x,z) + N_x(y,z)}{3} \quad (2)$$

where $N(x,y,z)$ is the value of output CT-slice image, $N_z(x,y)$ is output pixel data of axis z, $N_y(x,z)$ is output pixel data of axis y and $N_x(y,z)$ is output pixel data of axis x.

The output obtained by above prosecing represents one 3D spinal model. This is the content of pseudo 3D feature learning (Fig. 2). We compared three methods; Kamata's 2D method, our 1-Channel pseudo 3D learning model (our-method 1), and the 2-channel pseudo 3D learning model with Variation-Map (our-method 2).

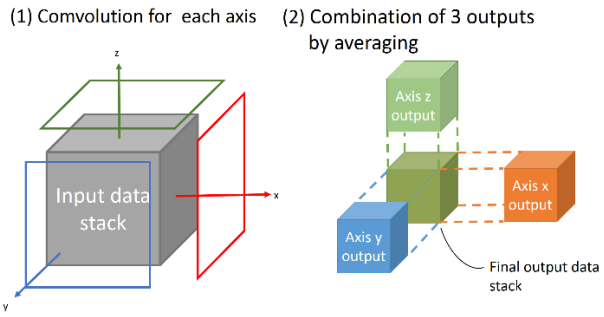

Figure 2. Pseudo 3D feature learning.

III. Results

A. Comparing Kamata's 2D Method and Our Methods

Table I shows the degree of accuracy by the correct label as a result of extracting the healthy spinal region from the unlearned data by the Kamata's method (2D U-Net), and two our methods with pseudo 3D learning. The definitions of Dice coefficient is given by formula (3):

$$\mathrm{DC} = \frac{2|V_c \cup V_o|}{|V_c| + |V_o|} \quad (3)$$

where DC is dice coefficient, V_c is the set of pixels of correct label data, and V_o is set of output pixels of each models.

Fig. 3-Fig. 5 show the spine extraction results for unlearned healthy spine data (Case No.10) by each method of Kamata and ours.

Table II shows the number of epochs and accuracy at the end of training in our-method 2. "Epochs" is number of epochs until finishing learning, "Dice Coef" is dice coefficient about learned data, and "Val Dice Coef" is coefficient about unlearned data. Fig. 5-Fig. 7 shows the transition of dice coefficient to number of epochs for each axes during learning of our-method 2. The blue line represents dice coefficient of learned case and the orange line shows "val_dice_coef" which is dice coefficient of unlearned case. These figures show that our-method 2 finally achieved nearly 0.980 for learned data and about 0.60 to 0.80 for the unlearned data at the end of learning.

TABLE I. The Degree of Coincidence between Segmented Halthy Spine Region and Correct Label Region for Each Unlearneed Spine Data

	Dice Coefficient		
Case No.	Kamata's Method (2D U-Net)	Our-method 1 (1-Channel learning)	Our-Method 2 (2-Channel Learning)
4	0.859	0.930	0.943
5	0.837	0.933	0.962
6	0.792	0.933	0.942
7	0.915	0.974	0.975
8	0.889	0.976	0.972
9	0.832	0.971	0.977
10	0.864	0.977	0.978
Average	0.855	0.956	0.964

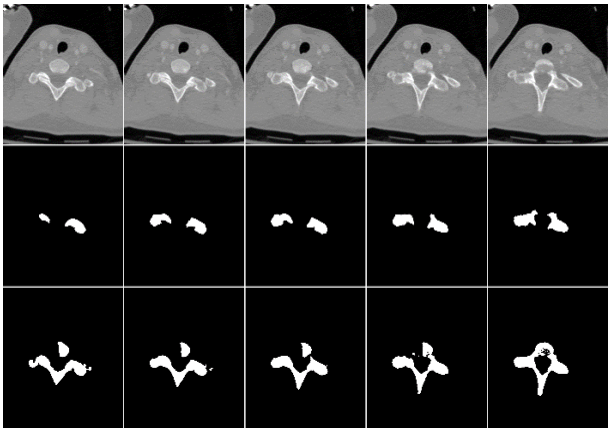

Figure 3. Result of spine segmentation by Kamata's method. This figure represents the original CT image (upper row), the correct label (middle row), and the extraction result (lower row), respectively.

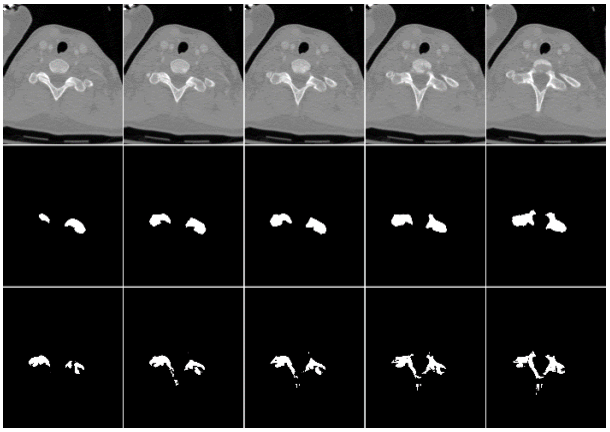

Figure 4. Result of spine segmentation by our-method1. This figure represents the original CT image (upper row), the correct label (middle row), and the extraction result (lower row), respectively.

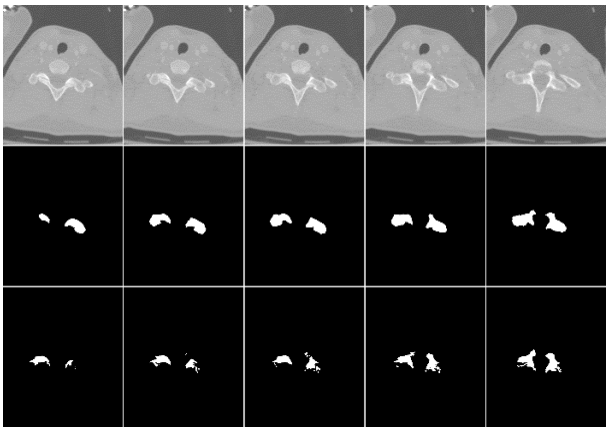

Figure 5. Result of spine segmentation by our-method1. This figure represents the original CT image (upper row), the correct label (middle row), and the extraction result (lower row), respectively.

TABLE II. EPOCS AND ACC WHEN FINISHED TRAINING

Axis	Epochs	Dice Coef	Val Dice Coef
z	117	0.975	0.816
x	132	0.979	0.602
y	307	0.983	0.620

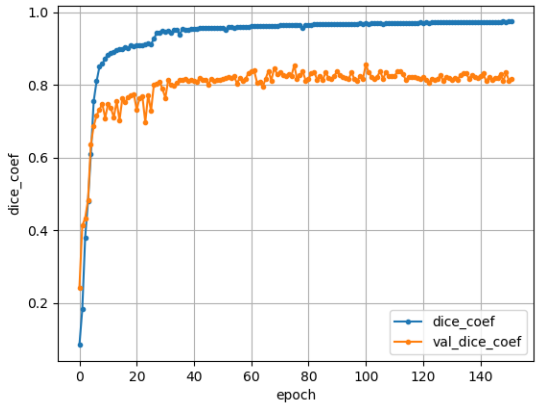

Figure 6. Learning curve for axis z. The vertical axis is dice coefficient, and horizontal axis is number of epochs.

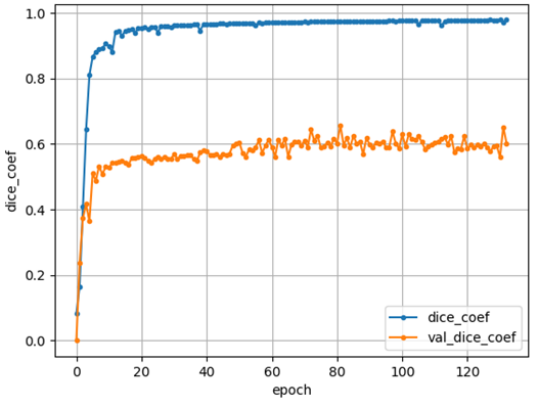

Figure 7. Learning curve for axis x. The vertical axis is dice coefficient, and horizontal axis is number of epochs.

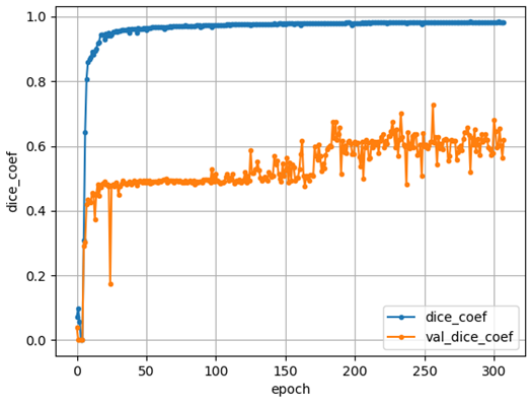

Figure 8. Learning curve for axis z. The vertical axis is dice coefficient, and horizontal axis is number of epochs.

B. Evaluation by Three-Dimensional Visualization

Fig. 9 shows the results of 3D reconstruction of healthy spine from the correct label and outputs of Kamata's method, our-method 1, and our-method 2. Table III shows dice coefficient of extraction result for compression fracture sample by each method. Kamata's method achieved 0.805, however ours two methods were 0.739 (our-method 1), and 0.574 (our-method 2).

Fig. 10 shows the results of 3D reconstruction of compression fracture sample from the correct label and outputs of Kamata's method, our-method 1, and our-

method 2. Looking at the figure, Kamata's model shows over-extraction at the upper cervical vertebrae, and missing parts are seen in the lower part. On the other hand, our two methods show missing parts at the upper and lower ends of the 3D model.

TABLE III. DICE COFFICIENT OF COMPRESSION FRACTURE SAMPLE

method	Dice Coefficient
Kamata's 2D method	0.805
Our-method 1	0.739
Our-method 2	0.574

Figure 9. 3D reconstructed images of spinal CT from healthy spine (Case10).

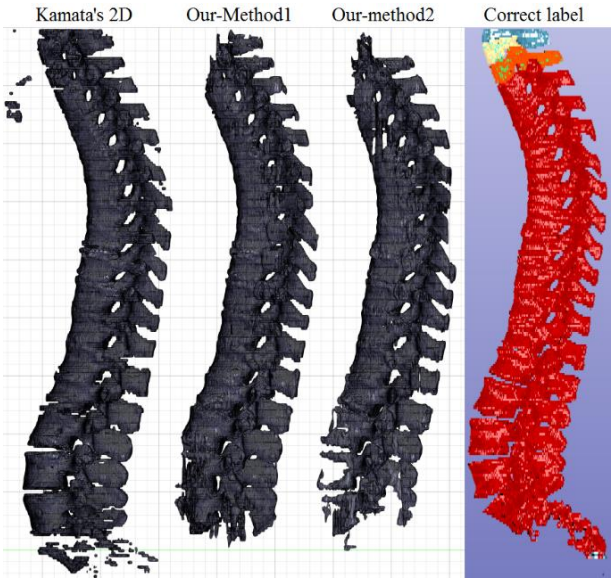

Figure 10. 3D reconstructed images of spinal CT from compression fracture sample.

IV. DISCUSSION

Based on the experimental results, the spine region can be extracted by U-Net with the accuracy of 96.4% on an average for healthy bone samples by applying pseudo 3D convolution learning and variation map. This is a better result than the existing method using U-Net. Pseudo 3D convolution learning achieved 60% to 80% for unlearned data in each axis. However, it was confirmed that by integrating them by averaging, it finally improved to about 96%. This is probably because above operation practically creates an effect of ensemble learning.

From these results, the learning of cubic shape features showed certain effectiveness in improving accuracy of spine region automatic extraction. However, because there is a difference between the extraction result and the correct label in the extraction of the healthy vertebrae, we conclude that further examination is necessary for the learning method.

It seems that the networks of proposed pseudo 3D methods learned by healthy spinal samples failed to segment correctly for diseased samples. We think that this can be improved by learning of compression fracture samples.

V. CONCLUSION

We tried automatic extraction of spinal area from CT images as an attempt to apply CNN to medical image processing.

In order to solve the problem of the shape reproducibility that was shown in the previous study by Kamata *et al.*, we used a U-Net with an improvement to learn 3D features based on the Kamata's method. As a result of learning three vertebrae using the model, our-method 1 (only pseudo 3D convolution learning) achieved the extraction accuracy of 95.6%, and our-method 2 (add variation map) achieved 96.4% for the unlearned data. By comparing our-method 1 and our-method 2, over-extraction was observed in our-method 1 for the region near the cervical vertebrae. Conversely, our-method 1 archieved accuracy of 73.9% from compression fracture sample, but our-method 2 achieved 57.4%. These results show that the learning of 3D features has a certain effect on improving extraction accuracy; however, it seems that generalization ability is not sufficient with these methods. We think that further research is necessary for the 3D shape learning method.

ACKNOWLEDGMENT

The authors wish to thank Drs. Kunihiko Fukushima, Isao Hayashi, and Hayaru Shouno for helpful discussion at the early stage of this study. This study was supported in part by JSPS KAKENHI Grant Number JP17K00251.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox. (May 2015). UNet: Convolutional networks for biomedical image segmentation, computerized medical imaging and graphics. arXiv. [Online]. Available: https://arxiv.org/pdf/1505.04597.pdf

[2] M. Kamata, K. Fukushima, H. Shouno, I. Hayashi, and M. Kikuchi, "Automatic detection of spine in CT image by U-Net," in *Proc. NCSP*, March 2018, pp. 608-610.

[3] J. Yao, *et al.* (April 2016). A multi-center milestone study of clinical vertebral CT segmentation, computerized medical imaging and graphics. [Online]. 49. pp. 16-28. Available:

https://www.sciencedirect.com/science/article/pii/S0895611115001937
[4] J. Huang, F. Jian, and H. Wu. (May 2013). An improved level set method for vertebra CT image segmentation, Biomedical engineering online. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701568/
[5] Y. Wang, J. Yao, H. R. Roth, J. E. Burns, and R. M. Summers, "Multi-Atlas segmentation with joint label fusion of osteoporotic vertebral compression fractures on CT," in *Proc. 3rd Workshop & Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 2015, ch. 7, pp. 74-84.
[6] R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "Interpolation-Based shape-constrained deformable model approach for segmentation of vertebrae from CT spine images," in *Proc. 2nd Workshops on Computational Methods and Clinical Applications for Spine Imaging*, Boston, 2014, pp. 235-240.
[7] M. Vania, D. Mureja, and D. Lee. (November 2017). Automatic segmentation of spine using convolutional neural networks via redundant generation of class labels, Arxiv. [Online]. Available: https://arxiv.org/pdf/1712.01640v1.pdf

Naofumi Shigeta is an undergraduate student at School of Computer Science, Tokyo University of Technology, Japan. His research interest is a medical image processing using convolutional neural network.

Mikoto Kamata received B. degree in Computer Science in 2018 from Tokyo University of Technology, Japan. Her research interest was a medical image processing using convolutional neural network.

Masayuki Kikuchi is a senior assistant professor at School of Computer Science, Tokyo University of Technology, Japan. He received his B. Eng. degree in 1994 from Waseda University, Japan, and M. Eng. degree in 1996 and Ph.D. degree in Engineering in 1999 from Osaka University, Japan. He was a research associate at Osaka University from October 1996 to March 1999, and a research associate at University of Tsukuba from April 1999 to March 2003. His current research interests are modeling of neural network of the visual system, psychophysical experiment on visual perception, and analysis of brain activity during various cognitive tasks.