Automatic detection and evaluation of spine from CT images using deep learning

Yusuke Mita[1], Toru Kato[1], Shoto Sekimura[2], Hiroki Takahashi[1], Akio Doi[1], Taro Mawatari[3], and Taku Sugawara[4]

[1]Iwate Prefectural University, Japan
[2]Research Institute of System Planning, Inc., Japan
[3]Hamanomachi Hospital, Japan
[4]Akita Cerebrospinal and Cardiovascular Center, Japan
(Tel: 81-019-694-2000, Fax: 81-019-694-2001)

[1]{g231q029, katoul2011, t-hiroki, doia}@s.iwate-pu.ac.jp, [2]sekimura@isp.co.jp, [3]mawatari@gmail.com, [4]sugawara-taku@akita-noken.jp

Abstract: The main disorders of the lumbar spine include spondylolysis, spondylolisthesis, and fractures. Spinal fusion is used to treat these diseases, and careful preoperative planning is important. Generally, in preoperative planning for spinal fusion, it is necessary to extract bone regions from CT images and classify the lumbar spine into L2 to L5 vertebrae. However, accurately extracting and classifying the surface shape of the lumbar vertebrae is cumbersome and time-consuming given the complex morphology of the vertebrae. In addition, it is necessary to prepare a large amount of learning data in order to improve the extraction accuracy of deep learning. In the present study, we attempt to improve the accuracy of automatic spine extraction with less learning data by adding an adversarial network and boundary information to automatic segmentation using a fully convolutional network and evaluate the results.

Keywords: Deep Learning, Medical CT Image, Fully Convolutional Neural Networks, Segmentation.

1 INTRODUCTION

The main spinal disorders include spondylolysis, spondylolisthesis, vertebral body fracture, and scoliosis. Spinal fusion is used to treat these diseases and requires thorough preoperative planning. A recently developed tailor-made implant stabilizes the upper and lower vertebrae by fixing the laminae with fit-and-lock platesand rods. This technique requires accurate measurement of each lumbar vertebra, especially when the plate is attached to the lumbar spine (Fig. 1). There are five lumbar vertebrae, which are referred to, from the top, as the first through fifth lumbar vertebra (L1-L5). However, accurate extraction is difficult because the lumbar vertebrae overlap in a complicated manner. These are important problems for diagnosing spinal disorders.

In the present study, we classify the lumbar vertebrae of CT images from the spine segment into L2 to L5 using a fully convolutional network (FCN) [2], which is a convolutional neural network (CNN). In the FCN, we attempted to improve the extraction accuracy by changing the contour weight using bone contour information. Furthermore, adversarial training [3] was introduced in order to prevent over-learning.

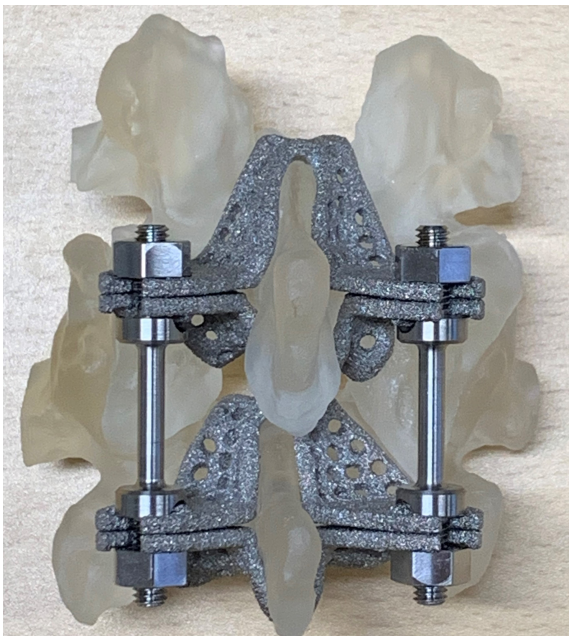

Fig. 1. Tailor-made spinal fixation implants [1]

2 RELATED RESEARCH

2.1 Spine instance segmentation

Kato et al. developed an instance segmentation method that classifies L3, L4, L5, and S1 using deep learning from spinal CT images [4]. In this method, two lumbar spine cases were successfully classified. However, it was clarified that it was not possible to cope with cases that were not learned (cases with curved bones or cases with different luminance distributions). In addition, since four lumbar vertebrae are extracted simultaneously, it is necessary to reduce the image size with an FCN.

2.2 Deep learning using a fully convolutional network

Fully convolutional networks have shown high accuracy in the field of segmentation in recent years. The FCN is a network model in which a fully connected layer is excluded from a CNN, and all layers are convolutional layers. The FCN is often used for semantic segmentation because the position information of the object on the image is not lost by eliminating the fully connected layer and there is no restriction on the size of the input image.

The FCN network structure of the proposed method adopts the Encoder-Decoder type and is shown in Fig. 2. The network consists of five layers. The Encoder part uses feature extraction by convolution. The convolution filter is a 5×5×5 filter, and the activation function uses a parametric rectifier linear unit (PReLU) [5]. Down-sampling uses a stride-2 2×2×2 filter convolution so that the number of feature map channels is doubled. In the Decoder part, upsampling is performed to return the feature map to its original size and perform segmentation. For the convolution filter, we used a 13×13×13 global convolutional network (GCN) [6] that can apply a large-scale kernel filter with a small number of parameters and reduce the amount of computation. The activation function is a PReLU, as in the Encoder part. Upsampling uses a stride-2 2×2×2 filter transposed convolution so that the number of feature map channels is halved. In order to prevent the disappearance and divergence of the gradient, residual learning is performed by adding the first and last feature maps in each layer [7]. In order to solve the problem of local feature loss during downsampling, a feature map is transmitted from the encoder to the decoder [8]. Finally, the feature map output from the network is converted into the probability that a voxel belongs to each class using a softmax function.

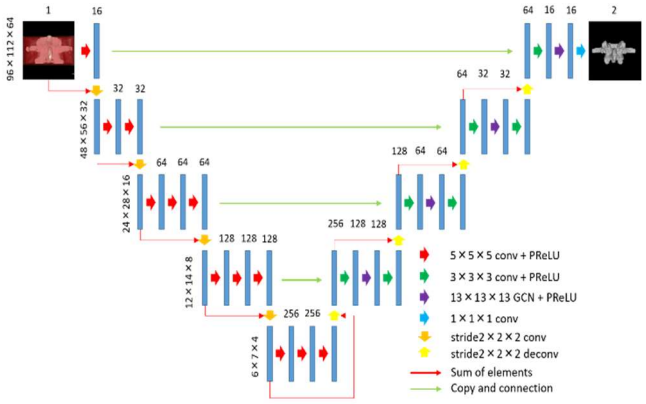

Fig. 2. FCN network

2.3 Prevention of over-fitting with Adversarial Training

In order to prevent over-fitting, the loss value obtained by adversarial training is added to the loss value. Adversarial training is a technique to improve the generalization performance of a model by learning for an adversarial example. An adversarial example is the original image plus the hostile perturbation calculated to maximize the loss value. The overall loss function is expressed as follows:

$$L_{seg} = L_{st} + L_{at} \cdots (1)$$

$$L_{st} = -\sum_{d,h,w}\sum_{c\in C} y_n^{(d,h,w,c)}\, log(S(x_n)^{(d,h,w,c)}) \cdots (2)$$

$$L_{at} = -\sum_{d,h,w}\sum_{c\in C} y_n^{(d,h,w,c)}\, log(S(x_n + r)^{(d,h,w,c)}) \cdots (3)$$

where L_{seg} is calculated as the sum of L_{st} and L_{at}, where L_{st} is the sum of the label data of each pixel and the cross-entropy of the segmentation result, and L_{at} is the loss related to adversarial training. In addition, y_n is the pixel value of the label data, and 0 or 1 is entered depending on the class to which the pixel belongs. As will be described later, if the target pixel is a bone contour, $y_n > 1$ is used to enhance learning with bone contour information. Moreover, x_n is the original data (CT image), d, h, and w indicate the voxel position (depth, height, and width), c is the number of channels, and $S(x_n)$ is the probability of belonging to a certain class of FCN. The perturbation added to the original CT image is output by the FCN, and the sum of cross-entropy is calculated based on the result. Here, r is the perturbation obtained by adversarial training.

3 METHOD

3.1 Isotropic voxelization

In order to improve the accuracy of image classification, a data set of isotropic 3D images is used for training. Tri-linear interpolation was used to estimate pixel values where there are no pixels, which is necessary for isotropic image processing.

Isotropic image processing unifies the actual size in each direction (number of voxels × voxel width) across all datasets. A reference value is set for the distance in millimeters set around a specific affected area of the CT image, and the isotropic voxel width and the number of voxels were adjusted to satisfy this standard (Fig. 3).

The actual size was determined as shown in Fig. 4, and the voxel width was unified at 1.0 mm. Each data set was arranged so that the center of gravity of the vertebral body was at the center of the image. In order to reduce unnecessary areas, the number of voxels was set to X: 96.0, Y: 112.0, and Z: 64.0.

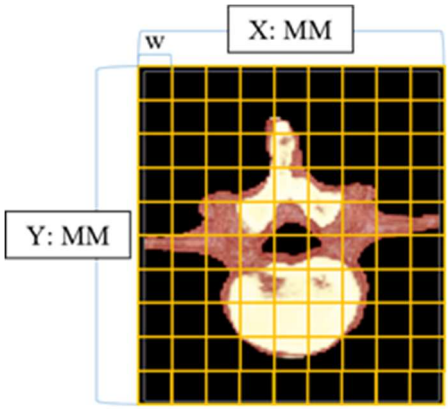

Fig. 3. Conceptual diagram of isometricization of an image of actual size. MM = actual size (mm), w = voxel width, n = number of voxels; MM = w × n is equalized so that all data sets match.

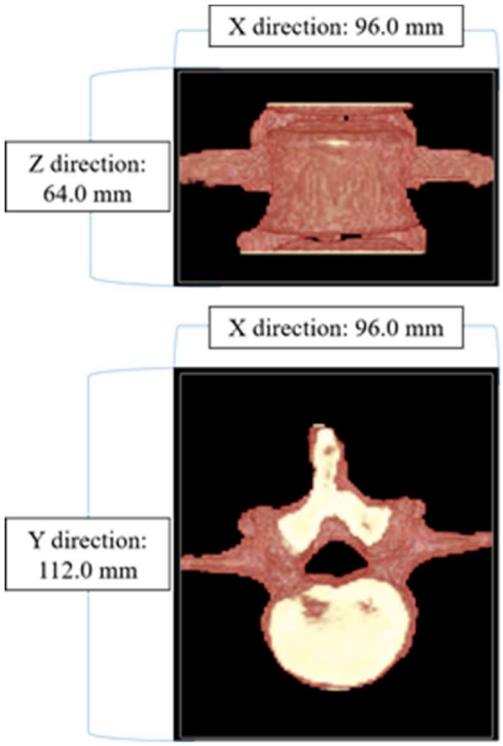

Fig. 4. Cutout reference value for vertebral

3.2. Learning enhancement of contour using edge images

Lumbar vertebrae were classified by an FCN incorporating bone contour information. The contour information is used to change the value of y_n when calculating the loss function. The calculation used to determine the loss for each pixel is expressed by the following equation:

$$loss = -y_n^{(d,h,w,c)} \log S(x_n)^{(d,h,w,c)} \cdots (4)$$

The loss function uses cross-entropy and calculates the similarity between label data and the segmentation results for all pixels. Here, x is the probability from 0 to 1 that the output from the FCN belongs to a class of voxels. Moreover, $-\log x$ approaches ∞ as x approaches 0 and approaches 0 as x approaches 1 (Fig. 5).

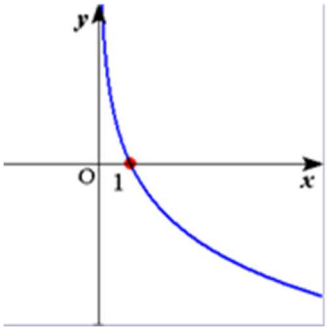

Fig. 5. Probability transition for the value of x

Here, y is the pixel value of the label data. For example, if y is pattern divided into 0, 1, and 1.1, it is expressed by the following formula. The greater y_n is, the greater the impact on the loss value is. Therefore, in order to reduce the loss value, it is necessary to learn that the value of x more closely approaches 1 for pixels with large y_n than for pixels with small y_n. Therefore, pixels with large y_n can be learned more efficiently.

Based on the above properties, a Laplacian filter was applied to the label data to create a boundary image with a background of 0 and a boundary of 1. The label data and the boundary image were compared on a pixel-by-pixel basis. If both the label data and the boundary image were pixel 1, then learning was performed by changing pixel y of the label data to 1.1. As a result, learning with enhanced contours is possible, and the accuracy of bone contour extraction is improved.

3.3. Creating training data

We created a data set that extracted L2, L3, L4, and L5 vertebral bodies from 11 spinal CT images. The data set was 44 cases (11 × 4), as shown in Fig. 4, and nine cases were used for training. In the training data creation, in creating the L2 data set, when the L2 of the original image is clipped, the boundary between the upper L1 and L3 is given, and the label data is only that for L2. In the second experiment, the surface deviation of the coordinate data measured by 3D scanner after shaving off the meat around the bone of the same part as the DICOM data taken from the abdominal multi-slice CT prepared separately from Experiment 1 were visualized using GOM Inspect and compared.

4 RESULTS

Based on the results of learning with the proposed method without distinguishing L2 through L5, we conducted two experiments. In the first experiment, segmentation was performed for eight cases that were not used for learning, and the extraction accuracy was evaluated using Intersection over Union (IoU).

4.1. Segmentation results and IoU evaluation

Figure 6 shows an example of the results of segmenting eight cases based on the learned data. Table 1 shows the average IoU for all cases.

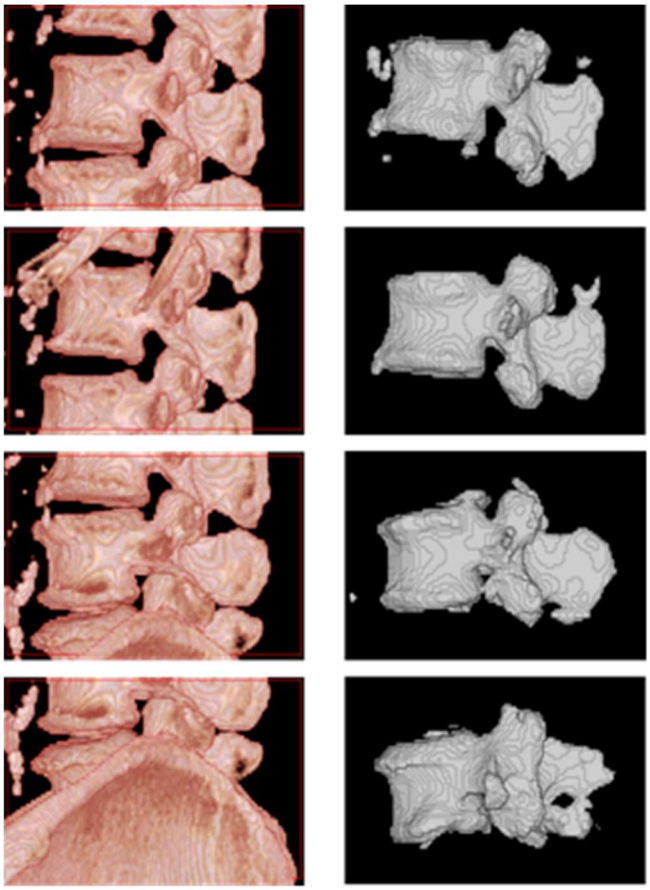

Fig. 6. Example of experimental results (from the top: L2, L3, L4, and L5)

Table 1. Average Intersection over Union of the evaluation results

Extraction site	Baseline	Using edge images
L2	0.846	0.904
L3	0.835	0.889
L4	0.790	0.840
L5	0.789	0.705

4.2. Visualization evaluation of surface deviation

Using GOM Inspect, the segmentation results are compared with CT images and 3D scan data, and the surface deviations are visualized in color. The color of the surface deviation shows that it shifts to the outside from the original data as it changes from green to red, and it shows that it shifts to the inside as it changes to blue.

4.3. Discussion

When the test images of two cases were evaluated in Experiment 1, the contours of L2 to L4 could be extracted, and the results were less affected by other parts than without contour enhancement. However, the extraction results for L5 were not stable in any case. This may be due to the fact that there is a large individual difference in the boundary with S1 (sacrum), which is the lower part of L5. In addition, the shape is different from L2 to L4, and the number of cases is smaller than that for L2 to L4. Therefore, the shape may be influenced by the contours learned from L2 to L4.

In addition, when the degree of coincidence with the CT image was compared in Experiment 2, there was a large difference between the upper and lower vertebral bodies and the intervertebral disc and the lumbar vertebrae for the proposed method, as compared to the conventional method. The erosion of the outer surface is reduced. In comparison with 3D scan data, the results varied depending on the number of cases. This is because the data used for learning was a CT image.

5 CONCLUSION

In the present study, 36 cases for L2, L3, L4, and L5 single vertebral body data sets were carefully selected from spinal CT images and were used for training. In addition, the surface deviation between the scan data and the CT image prepared separately and the results of segmentation using an FCN was visualized and evaluated.

Here, for L2 to L4, there were good cases and deformed cases, and L5 was not stable as compared to L2 to L4. The cause may be that there is a large individual difference in the boundary with a lower S1 and that the training is not sufficient. Finally, L2 to L5 also have individual differences in bending that should be taken into account.

REFERENCES

[1] Taku S (2017), "Spine fusion without hurting bones" created by a specialist (in Japanese). Forbes JAPAN October 2017 issue (https://forbesjapan.com/articles/detail/18162/2/1/1)

[2] Jonathan L, Evan S, Trevor D (2014), Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038

[3] Ian J. Goodfellow, Jonathon S, Christian S (2015), Explaining and Harnessing Adversarial Examples. arXiv:1412.6572

[4] Toru K, Shoto S, Akio D, Taro M, Sadafumi I (2017), Development of automatic bone extraction tool from CT images using deep learning (in Japanese). arXiv:1412.6572

[5] Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2015), Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv:1502.01852

[6] Chao P, Xiangyu Z, Gang Y, Guiming L, Jian S (2017), Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Network. arXiv:1703.02719

[7] Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2015), Deep Residual Learning for Image Recognition. arXiv:1512.03385

[8] Olaf R, Philipp F, Thomas B (2015), U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597