ICIC Express Letters
Part B: Applications
Volume **10**, Number **7**, July **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **651–656**

OBJECT DETECTION OF COCHLEA IN MICRO-COMPUTED TOMOGRAPHY USING FASTER REGION CONVOLUTIONAL NEURAL NETWORK

Jongwun Choi, Myeongjun Han and Namkeun Kim*

Department of Mechanical Engineering
Incheon National University
119, Academy-ro, Yeonsu-gu, Incheon 22012, Korea
whddnssnla@naver.com; gksaudwns57@gmail.com; *Corresponding author: nkim@inu.ac.kr

Received December 2018; accepted March 2019

Abstract. *In order to develop a personalized medical device for a human auditory periphery such as middle-ear prosthesis or to analyze cochlear structure using finite element model, it is important to obtain patient's cochlear geometry. In this study, the cochlear geometry was obtained from each person's computed tomography (CT) images. In addition, the data augmentation was performed to prevent overfitting. Furthermore, the faster region convolutional neural network (Faster R-CNN) was used to detect the cochlear region from the μCT images more effectively and automatically. Results showed that Faster R-CNN method could obtain the cochlear regions from the μCT images with 82.11% precision and 93.13% sensitivity.*
Keywords: Cochlea, Object detection, Faster region convolutional neural network

1. **Introduction.** Human's ear consists of three parts, which are outer ear, middle ear and inner ear. Among these components, the cochlea is the most significant organ to hear a sound because sound frequency can be distinguished in the cochlea. Therefore, there have been many studies for the human cochlea [1-4]. However, due to its location which is difficult for researcher to access as well as its complex geometry, it was hard to fully understand the hearing mechanism occurring in the cochlea. Therefore, obtaining the real geometry is a significant process for better understanding of the mechanism. Unfortunately, the process to obtain the region of interest (ROI) of cochlea from the micro computed tomography (μCT) images is considered to be too boring and time-consuming works.

To construct a three-dimensional geometry of the cochlea from μCT images, the interested object should be segmented from each image and then stacked. In this process, the intensity difference between bone and the other components of a cochlea makes the segmentation of the bone easily. There are several methods to segment the ROI. The most famous method is Gauss segmentation algorithm [5]. The key concept of this method is *shrink wrapping* of the ROI. In other words, the contour lines drawn manually by researchers were shrunk to the surface of the nearest bone to the contour line. However, the caveat of this method is that the processes should be mostly performed *manually* by researchers. The other method is an automated or semi-automated contouring procedures. For example, simple threshold method [6], snakes method [7,8], watershed method [9,10], and morphological approach method [11] are using the automated or semi-automated contouring procedures. In these methods, the contour lines can be automatically drawn, and shrunk or expanded to the surface of the nearest ROI within the boundary designated by researcher. However, these methods also need researchers' manual endeavors for some procedures such as designating the ROI.

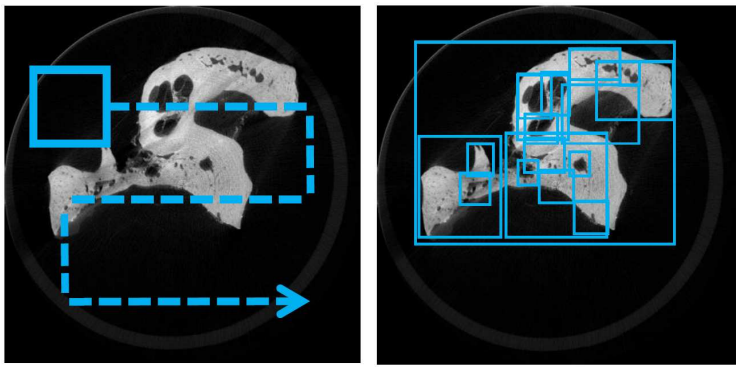
DOI: 10.24507/icicelb.10.07.651

In order to develop an automatic procedure for generating ROIs, a machine should obtain the information about the location, shape, size, etc., of the bone. In the case of human, this information can be obtained heuristically. In this study, however, we aim to train the machine taking the information through the faster region convolutional neural network (Faster R-CNN), namely, *deep learning*. Eventually, the trained machine will be able to automatically segment the interested ROIs from μCT images.

2. **Study Methods.** The object detection is the most important process in the computer-vision field using deep learning technique. Through the *detection process*, we can determine not only if the interested object exists in an image, but also the probability (represented by percentage) if the detected object in an image identifies with what we are interested. In other words, the object detection is the algorithm to find ROIs with the position as well as the probability for the object to be identified with what we need. When an input image is given, the object-detection-process is carried out by *localization* and *classification* in all the meaningful regions. For example, when an image with cat and dog is given as an input, drawing a bounding box on each of the cat and dog is called *localization*. On the other hand, the categorization of the cat image into the cat category and the dog image into the dog category, respectively, are called *classification*.

In this study, we used '*Faster R-CNN* (region-convolutional neural network)' for object detection process. The Faster R-CNN is a combination of *CNN* and *region proposal network*. The CNN is a network which obtains the feature map from an original image, and it has the spatial information of the image as well. This is a strength of the CNN in comparison with ANN which can obtain the feature map from an original image, but cannot have the spatial information [12].

On the other hand, 'region proposal' (see Figure 1) represents to suggest any region(s) from an original image to users. At the moment, the suggested region(s) can be an interested region, but or not. The *region proposal network* is a network which can do the 'region proposal' through the *training* (see Figure 2). Details for the training can be found in Ren et al. [13].



(a) (b)

FIGURE 1. The example of *region proposal* method: (a) sliding window detector approach, and (b) selective search algorithm

We investigated the feasibility of using Faster R-CNN for detection of the cochlea from μCT images. The Faster R-CNN was implemented by 'Faster R-CNN inception resnet v2' provided by the Tensorflow. It should be noted that 'Faster R-CNN inception resnet v2' has the advantages to detect an interested object from an original image (i.e., high mean average precision) in spite of the expensive calculation.

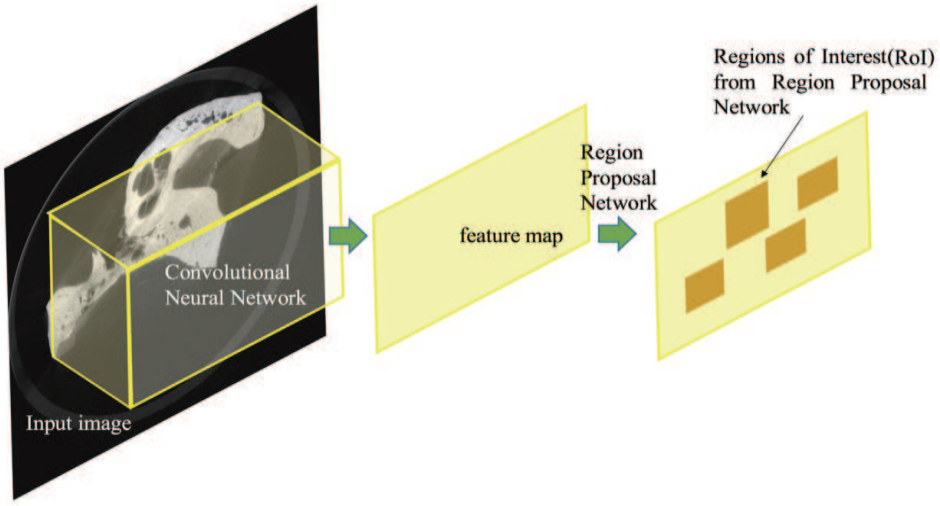


FIGURE 2. The schematic diagram of the region proposal network in the Faster R-CNN

3. Experimental Materials.

3.1. **Image datasets.** The used image datasets in this study were obtained from the micro-CT scanner developed by SCANCO Medical AG (www.scanco.ch). The visibility of CT is determined by the photon energy, the intensity of the X-ray and the integration time. This scanner has high maximum X-ray intensity, which means CT images have good signal-to-noise ratio and image clarity. In addition, the machine can perform high resolution scans up to 2,048 × 2,048 pixels per image. When the specimen (human temporal bone) is reduced to fit into the 21.5 mm diameter holder, we could obtain the best resolution of 10.5 μm. The default scan length was about 12 mm in the depth direction. These values in scan length could make approximately 1,140 slices at the 10.5 μm resolution [5]. We acquire 10 image datasets from the scan in this study.

3.2. **Labeled region of interest (ROI).** For both the training and test datasets, the labeled ROI containing each cochlea was manually selected from the images (See Figure 3). The number of ROIs obtained from each dataset was approximately limited from 240 to 440. The average number of ROIs of ten datasets was 286. And, the resolution of the ROIs is varied from 226 × 220 to 300 × 490, with an average of 308 × 342. The samples of μCT images and the labeled cochlear images are shown in Figure 3.

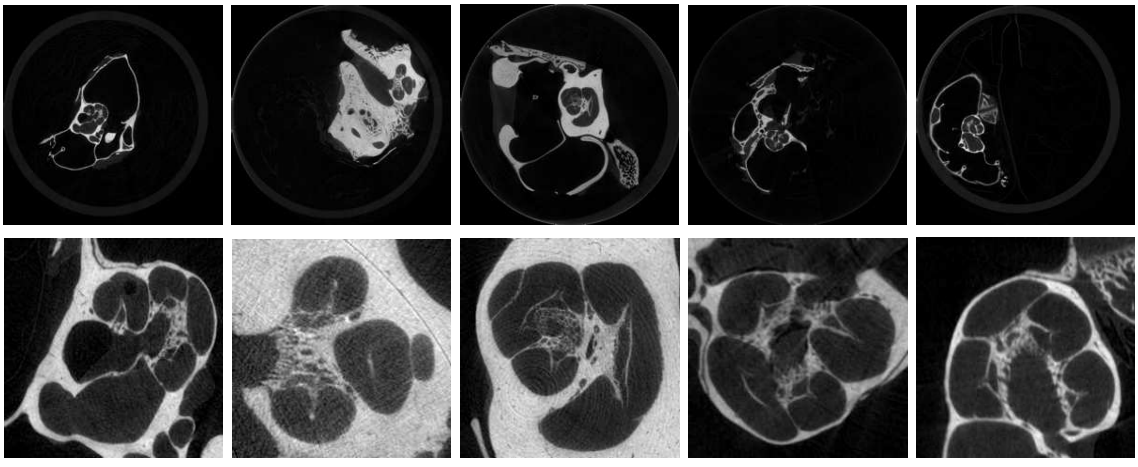


FIGURE 3. The first row represents the examples of μCT images whereas the second row shows the labeled ROI (cochlea) images.

3.3. **Data augmentation.** In order to avoid an overfitting in the machine-learning training, we performed data augmentation by 1) varying the brightness, 2) rotating [14], 3) upscaling/downscaling 4) adding noise [15], 5) distorting [16] and, 6) mirroring of images (i.e., flip the image horizontally). At first, we changed the brightness of the images. The brightness was varied by Equation (1):

$$Y = \frac{I_{\max}}{\phi} * \left(\frac{X * \theta}{I_{\max}} \right)^{\frac{1}{\gamma}} \tag{1}$$

where X and Y mean the input and output images, respectively. The $I_{\max}$ represents the maximum intensity. And, ϕ, θ, and γ describe the variables to control the brightness. In this study, the ϕ and θ were fixed as 1 whereas the γ was varied from 1.4 to 0.6 to describe the brightness and darkness. For example, the lower γ represents the darker image. The rotational angle was varied from -10 degrees to 10 degrees, and the scaling factor was changed from 95% to 105% with respect to the original image scale. Furthermore, in order to add the noise, we used the Gaussian Noise [15] with average and variation as 5 and 70, respectively. The distortion used in this study has the same meaning with '*elastic deformations*' [16].

For the distortion, we had to generate a grid. A grid has the same size with an input image and grid's nodes have its own x and y position. Then, we changed the grid's node by adding randomly generated displacements, Δx and Δy. The displacements, Δx and Δy, are generated through the following processes: 1) make the matrix which has the same size with input images, 2) assign the random number between -1 to 1 to every element of matrix, 3) to preserve shape information, apply the 'Gaussian filter' with the Gaussian standard deviation, σ which is equal to 2 times of the height of the input image (i.e., the length of the input image in the y axis), and 4) multiply the scaling factor, α, equal to 0.08 times of the height of the input image. Then, we mapped the input image to the changed grid, which became the *distorted* images. By applying 13 different transformations, we could obtain 13 times more image datasets than the original images. In this study, we used 10 μCT datasets. After data augmentation, number of images becomes 37,000. The 80 percentages of the images were used for machine-learning training, and the remained images were used for machine-learning test. The samples of the data augmentation are shown in Figure 4.

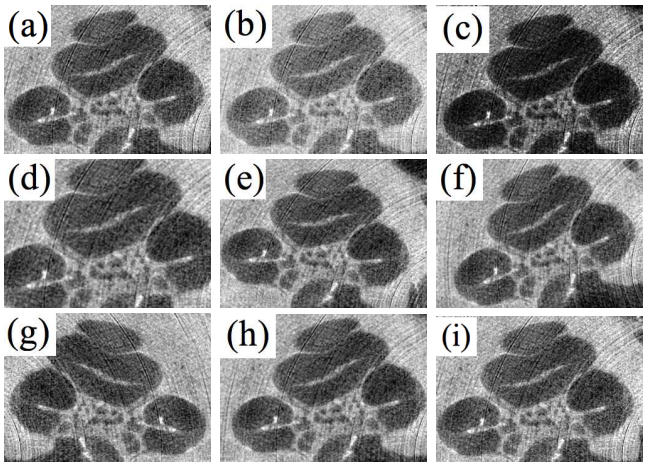


FIGURE 4. The examples of data augmentation: (a) original images, (b) brightness ($\gamma = 1.4$), (c) darkness ($\gamma = 0.6$), (d) upscaling, (e) rotation (10 degrees), (f) rotation (-10 degrees), (g) mirroring, (h) distortion, and (i) adding noise

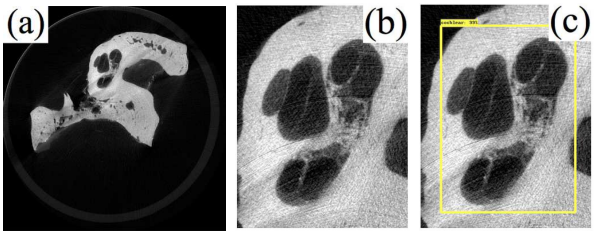


FIGURE 5. The results of object detection: (a) μCT images, (b) ROI (selected by researcher before machine-learning training), and (c) detected image (automatically detected by computer after machine-learning training; yellow square)

4. **Results and Discussion.** Figures 5(a)-5(c) show the original μCT image, the selected cochlear ROI by researcher before machine-learning training, and the automatically detected cochlear ROI by Faster R-CNN, respectively. We quantitatively evaluate the detection accuracy using 'positive predictive value (= precision)' and 'sensitivity'. Those two variables are defined as follows, respectively:

$$Precision = \frac{\text{number of } \textit{True Positives}}{\text{number of } \textit{Ture Positives} + \text{number of } \textit{False Positives}} \tag{2}$$

$$Sensitivity = \frac{\text{number of } \textit{True Positives}}{\text{number of } \textit{Ture Positives} + \text{number of } \textit{False Negatives}} \tag{3}$$

where *True Positives, False Positives*, *False Negatives* mean the cochlear region classified as cochlea, the non-cochlear region classified as cochlea, and the cochlear region classified as non-cochlea, respectively [15]. For example, the *precision* describes the probability if the cochlea predicted by the machine is a real cochlea. On the other hand, the *sensitivity* means the probability if the machine finds the cochlear image among all the real cochlear images in the original images. In other words, if we obtain 100% sensitivity, the machine can predict an image as the cochlea whenever the image has a real cochlea. If we have 80% sensitivity, 8 images would be predicted to be a cochlea out of total 10 images having a real cochlea. We could obtain the detected cochlear ROI (Figure 5(c)) with 82.11% precision as well as with 93.13% sensitivity. Without the Faster R-CNN process, the cochlear images were taken from the original CT images through very tiresome processes, which is drawing contours manually along each ROI. Therefore, it took several days to obtain the cochlear images. However, in the current study, we can detect the cochlear images within 30 minutes using the Faster R-CNN. It was proceeded by a Macbook Pro equipped with 2.9 GHz intel core i7 processor and 16 GB RAM.

5. **Conclusion.** We applied Faster R-CNN to object-detection process, which is selecting of the ROIs from μCT images. To prevent overfitting caused by limited input data, data augmentations were performed by varying the brightness, rotating, upscaling/downscaling, adding noise, distorting, and mirroring images. We could obtain the automatically selected cochlear ROI with 82.11% precision and 93.13% sensitivity. For better performance, we will consider additional data augmentation with more new datasets. Furthermore, whereas the current study is limited to detect the area in which we are interested, we will obtain the contour along the interested area using *morphological filtering* [11]. The obtained contours will be used to reconstruct the 3D cochlear structure, and a *super-resolution generative adversarial network* will be helpful to increase the resolution of the reconstructed 3D cochlear structure.

Acknowledgment. This research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (NRF-2017M3A9E2065287).

REFERENCES

[1] N. Kim, K. Homma and S. Puria, Inertial bone conduction: Symmetric and anti-symmetric components, *J. Assoc. Res. Otolaryngol.*, vol.12, pp.261-279, 2011.

[2] N. Kim, C. R. Steele and S. Puria, Superior-semicircular-canal dehiscence: Effects of location, shape, and size on sound conduction, *Hear. Res.*, vol.301, pp.72-84, 2013.

[3] X. Wang, L. Wang, J. Zhou and Y. Hu, Finite element modelling of human auditory periphery including a feed-forward amplification of the cochlea, *Comput. Methods Biomech. Biomed. Eng.*, vol.17, pp.1096-1107, 2014.

[4] W. Nogueira, D. Schurzig, R. Penniger, A. Büchner and W. Würfel, Validation of a cochlear implant patient specific model of the voltage distribution in a clinical setting, *Front. Bioeng. Biotechnol.*, vol.4, p.84, 2016.

[5] J. H. Sim, S. Puria and C. R. Steele, Calculation of inertial properties of the malleus-incus complex from micro-CT imaging, *Journal of Mechanics of Materials and Structures*, vol.2, pp.1515-1524, 2006.

[6] T. Mclnerney and D. Terzopoulos, Deformable models in medical image analysis, *IEEE Medical Image Analysis*, vol.1, no.2, pp.91-108, 1996.

[7] C. Xu and J. L. Prince, Gradient vector flow: A new external force for snakes, *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition*, pp.66-71, 1997.

[8] M. Kass, A. Witkin and D. Terzopoulos, Snake: Active contour models, *International Journal of Computer Vision*, pp.321-331, 1988.

[9] L. Vincent and V. Solle, Watershed in digital spaces: An efficient algorithm based on immersion simulation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.13, no.6, pp.583-598, 1991.

[10] J. Sijbers and P. Scheunders, Watershed-based segmentation of 3D MR data for volume quantization, *Magnetic Resonance Imaging*, vol.15, pp.679-688, 1997.

[11] P. Marquez-Neila, L. Baumela and L. Alvarez, A morphological approach to curvature-based evolution of curves and surfaces, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.36, no.1, pp.2-17, 2014.

[12] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Proc. of the 25th International Conference on Neural Information Processing Systems*, pp.1097-1105, 2012.

[13] S. Ren, K. He, R. Grishick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Proc. of the 28th International Conference on Neural Information Processing Systems*, pp.91-99, 2015.

[14] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata and H. Fujita, Classification of teeth in cone-beam CT using deep convolutional neural network, *Comput. Biol. Med.*, vol.80, pp.24-29, 2017.

[15] S. E. Umbaugh, *Digital Image Processing and Analysis – Human and Computer Vision Applications*, 2nd Edition, CVIPtools, 2016.

[16] P. Y. Simard, D. Steinkraus and J. C. Platt, Best practices for convolutional neural networks applied to visual document analysis, *Proc. of the 7th International Conference on Document Analysis and Recognition*, pp.958-963, 2003.