[Brief Article]

First trial and evaluation of anatomical structure segmentations in 3D CT images based only on deep learning

Xiangrong Zhou[†], Takaaki Ito[†], Ryosuke Takayama[†], Song Wang[††], Takeshi Hara[†], and Hiroshi Fujita[†]

[†]Department of Intelligent Image Information, Graduate School of Medicine, Gifu University,
1-1 Yanagido, Gifu 501-1194, Japan
[††]Department of Computer Science and Engineering, University of South Carolina,
315 Main St., Columbia, SC 29208, USA
(Received on July 8, 2016. In final form on August 4, 2016)

Abstract : We propose a novel approach for semantic CT image segmentation based only on a fully convolutional network (FCN), which accomplishes an end-to-end, voxel-wise multiple-class prediction to map each voxel in a CT image directly to an anatomical label. The proposed method simplifies the segmentation of the anatomical structures (including multiple organs) in a CT image (generally in 3D) to majority voting for the semantic segmentation of multiple 2D slices drawn from three orthogonal viewpoints with redundancy. An FCN consisting of "convolution" and "de-convolution" parts is trained and re-used for the 2D semantic image segmentation of different slices of CT scans. We applied the proposed method to segment a wide range of anatomical structures that consisted of 19 types of targets in the human torso. A database consisting of 240 3D CT scans and a humanly annotated ground truth was used for training (230 cases) and testing (the remaining 10 cases). The results showed that the target regions for the entire set of CT test scans were segmented with acceptable accuracies (89% voxels were labeled correctly) against the human annotations. This performance was comparable to other recently reported state-of-the-art results. Compared to previous segmentation methods that have to be guided by human expertise, this data-driven approach showed better efficiency, generality, and flexibility.
Keywords : CT image, anatomical structure segmentation, 2D semantic segmentation, fully convolutional network (FCN), deep learning

1. INTRODUCTION

Fully automatic image segmentation is a fundamental step of computer-based image analysis in 3D CT scans by mapping the physical image signal to a useful abstraction [1]. Unfortunately, the current research on the image segmentation of CT images is still limited to the major organs (e.g., lungs, liver), and shows unsatisfactory computational efficiency, accuracy, and robustness, especially for abnormal CT cases. Therefore, the image-segmentation issue has become a bottleneck for further computer-based medical image analysis and image interpretation.

Conventional approaches to CT image segmentation usually try to transfer human knowledge directly to a processing pipeline, including numerous hand-crafted signal processing algorithms and image features. Although many mathematical models have recently been introduced in image segmentation [2-9], they still attempt to emulate limited human rules or operations in segmenting CT images. In order to further improve the accuracy and robustness of image segmentation, we need to be able to handle a large variety of ambiguous image appearances, shapes, and relationships of anatomical structures. It is difficult to achieve this goal by defining and considering human knowledge and rules explicitly. Instead, a data-drive approach using big image data—such as a deep convolutional neural network (CNN) —is expected to be better for solving this segmentation problem.

Recently, several studies were reported that applied deep CNNs to medical image analysis. Many of these used deep CNNs for lesion detection or classification [10-12]. A few of these embedded patch-based deep CNNs into conventional organ-segmentation processes to reduce the false positives (FPs) in the segmentation results or to predict the likelihoods of the image patches [13-15]. However, the anatomical segmentation of CT images over a wide region of the human body is still challenging because of the image appearance similarities between different structures, as well as the difficulty of ensuring global spatial consistency in the labeling of patches in different CT cases.

This paper proposes a novel segmentation approach based on deep CNNs that naturally imitate the thought processes of radiologists during CT image interpretation for image segmentation. Our approach models CT image segmentation in a way that can best be described as "multiple 2D proposals with a 3D integration." This is very similar to the way that a radiologist interprets a CT scan as many 2D sections, and then reconstructs the 3D anatomical structure as a mental image. Unlike previous work on medical image segmentation that labels each voxel/pixel by a classification based on its neighborhood information (i.e., either an image patch or a "super-pixel"), our work uses rich information from the entire 2D section to directly predict complex structures (multiple labels on images). Furthermore, the proposed approach is based on a fully convolutional network (FCN) [16] without using any conventional image-processing algorithms such as smoothing, filtering, and level-set methods. In addition, the proposed approach uses one simple network to segment multiple organs simultaneously. It is adaptive to 3D or 2D images over an arbitrary CT scan range (e.g., body, chest, abdomen), and CT segmentation with this capability has not previously been reported.

2. METHODS

2.1 Overview

As shown in Fig.1, the input is a 3D CT case (the method can also handle a 2D case, which can be treated as a degenerate 3D case), and the output is a label map of the same size and dimension, in which the labels are a pre-defined set of anatomical structures. Our segmentation process is repeated to sample 2D sections from the CT case, pass them to FCN for 2D image segmentation, and stack the 2D labeled results back into 3D. Finally, the anatomical structure label at each voxel is decided based on majority voting at the voxel. The core part of our segmentation is an FCN that is used for the anatomical segmentation of the 2D sections. This FCN is trained based on a set of CT cases, with the human annotations as the ground truth. All of the processing steps of our CT image segmentation are integrated into an all-in-one network under a simple architecture with a global optimization.

2.2 3D-to-2D image decomposition and 2D-to-3D label stacking

In the proposed approach, we decompose a CT case (a 3D matrix, in general) into numerous sections (2D matrices) with different orientations, segment each 2D section, and finally, assemble the outputs of the segmentation (labeled 2D maps) back into 3D. Specifically, each voxel in a CT case (a 3D matrix) can lie on different 2D sections that pass through the voxel with different orientations. Our idea is to use the rich image information of the entire 2D section to predict the anatomical label of this voxel, and to increase the robustness and accuracy by redundantly labeling this voxel on multiple 2D sections with different orientations. In this work, we select all the 2D sections in three orthogonal directions (axial, sagittal, and coronal-body); this ensures that each voxel in a 3D case is located on three 2D CT sections.

After the 2D image segmentation, each voxel is redundantly annotated three times from these three 2D CT sections. The annotated results for each voxel should ideally be identical, but may be different in practice because of mislabeling during the 2D image segmentation. A label fusion by majority voting (selecting mode from three labels) is then introduced to improve the stability and accuracy of the final decision. Furthermore, *a prior* for each organ type (label) is estimated by calculating voxel appearance frequency of the organ region within total image based on training samples. In the case of no consensus between three labels during the majority voting process, our method simply selects the label with the biggest *prior* as the output.

2.3 FCN-based 2D image segmentation via convolution and de-convolution networks

We use an FCN for semantic segmentation in each 2D CT slice by labeling each pixel. Convolutional networks are constructed using a series of connected basic components (convolution, pooling, and activation functions) with translation invariance that depends only on the relative spatial coordinates. Each component acts as a nonlinear filter that operates (e.g., by matrix multiplication for convolution or maximum pooling) on the local input image, and the whole network computes a general nonlinear transformation from the input image. These features of the convolutional network provide the capability to adapt naturally to an input image of any size and any scan range of the human body, producing an output with the corresponding spatial dimensions.

Our convolutional network is based on the VGG16 net structure (16 layers of 3×3 convolution interleaved with maximum pooling plus 3 fully connected layers) [17], but with a change in the VGG16 architecture by replacing its fully connected layers (FC 6 and 7 in Fig.2) with convolutional layers (Conv 6 and 7 in Fig.2). Its final fully connected classifier layer (FC 8 in Fig.2) is then changed to a 1×1 convolution layer (Conv 8 in Fig.2) whose channel dimension is fixed at the number of labels (the total number of segmentation targets was 20 in this work, including the background). This network is further expanded by docking a de-convolution network (the right-hand side in Fig.2). Here,

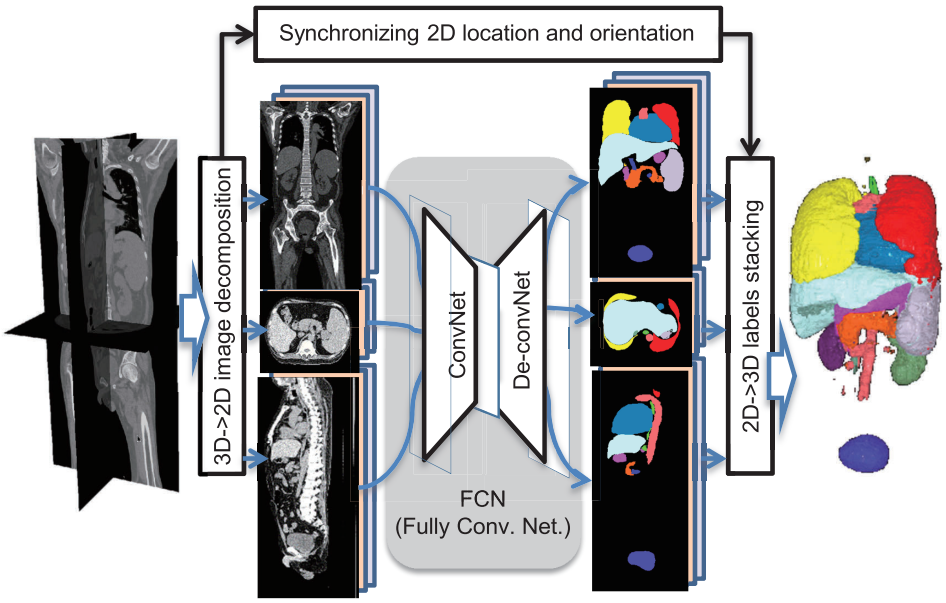

Fig.1 Pipeline of proposed anatomical structure segmentation for 3D CT scan. See Fig.2 for the details of the FCN part.

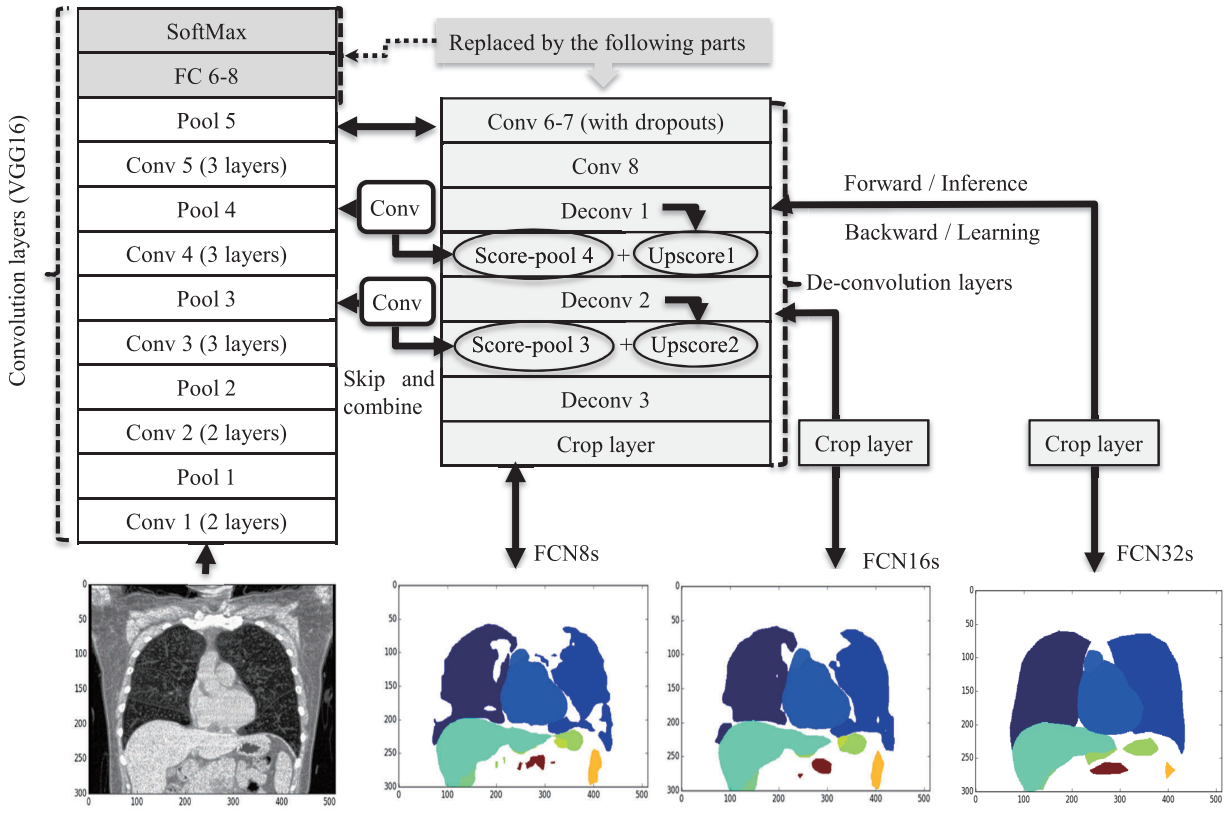

Fig.2 Semantic image segmentation of 2D CT slice using fully convolutional network (FCN) [16]. Conv : convolution, Deconv : deconvolution, and FC : fully connected.

we use the structures of the de-convolution [16], which are constructed using three de-convolution layers, each of which consists of up-sampling, convolution, and crop layers. As in the original idea [16], the intermediate results of the convolution network (the lower layers—Pools 3 and 4 —of VGG16 with higher image resolution) are skipped and combined sequentially into de-convolution layers. This skip structure passes the information that is lost in the lower convolution layers of VGG16 directly into the de-convolution process, which recovers the detailed contour sequentially under a higher image resolution.

FCN training. The proposed network is trained with numerous CT cases of humanly annotated anatomical structures. All of the 2D CT sections (corresponding to the label maps) along the three body orientations are shuffled, and used to train the FCN. The training process repeats feed-forward computation and back-propagation to minimize the loss function, which is defined as the sum of the pixel-wise losses between the network prediction and the label map annotated by the human experts. The gradients of the loss are propagated from the end to the start of the network, and the method of stochastic gradient descent with momentum is used to refine the parameters of each layer.

The FCN is trained sequentially by adding de-convolution layers. To begin with, a coarse prediction (by a 32-pixel stride) is trained for the modified VGG16 network with one de-convolution layer (called FCN32s). A finer training is then added after adding one further de-convolution layer at the end of the network. This is done by using skips that combine the final prediction layer with a lower layer with a finer stride in the modified VGG16 network. This fine-training is repeated with the growth of the network layers to build FCN16s and FCN8s, which are trained from the predictions of 16 and 8 strides on the CT images, respectively.

2D CT segmentation using trained FCN. The density resolution of the CT images is reduced from 12 to 8 bits using linear interpolation. The trained FCN is then applied to each 2D section independently, and each pixel is labeled automatically. The labels from each 2D section are then projected back to their original 3D locations for the final vote-based labeling, as described above.

3. EXPERIMENT AND RESULTS

Our experiment used a CT image database that was produced and shared by a research project entitled “Computational Anatomy [18]”. This database included 640 3D volumetric CT scans from 200 patients at Tokushima University Hospital. The spatial resolution of these CT images was distributed from 0.625 to 1.148 mm on transverse plane with slice thickness of 1 mm. However, any normalization of the spatial resolution to an isotropic value was not done in both training and test stages. The anatomical ground truth (a maximum of 19 labels that include Heart, right/left Lung, Aorta, Esophagus, Liver, Gallbladder, Stomach and Duodenum (lumen and contents), Spleen, left/right Kidney, Inferior Vein Cava, region of Portal Vein, Splenic Vein, and Superior Mesenteric Vein, Pancrease, Uterus, Prostate, and Bladder in 240 CT scans was also distributed with the database [19]. Our experimental study used all of the 240 ground-truth CT scans, comprising 89 torso, 17

chest, 114 abdomen, and 20 abdomen-with-pelvis scans. Furthermore, our research work was conducted with the approval of the Institutional Review Boards at Gifu and Tokushima Universities.

We picked 10 CT scans at random as the test samples, using the remaining 230 CT scans for training. As previously mentioned, we took 2D sections along the axial, sagittal, and coronal body directions. For the training samples, we obtained a dataset of 84,823 2D images with different sizes (width : 512 pixels ; height : 80-1141 pixels). We trained a single FCN based on the ground-truth labels of the 19 target regions. Stochastic gradient descent (SGD) with momentum was used for the optimization. A learning rate of 10^{-4}, momentum of 0.9, and weight decay of 2^{-4} were used as the training parameters. All the 2D images were used directly as the inputs for FCN training, without any patch sampling. We did not balance the target classes by weighting the training loss, although the sample numbers and occupied regions of each target class were significantly different.

We tested the proposed FCN network (Fig.1) using 10 CT cases that were not used in the FCN training. An example of the segmentation result for a 3D CT case covering the human torso is shown in Fig.3. The accuracy of the segmentation was evaluated per organ type and per image. First, we measured the intersection over union (IU) (also known as the Jaccard similarity coefficient) between the segmentation result and the ground truth. The mean IU values in each organ type are listed in Table 1 for both training and test data.

Because each CT case may contain different anatomical structures—with the information about these unknown before the segmentation—we performed a comprehensive evaluation of multiple segmentation results for all the images in the test dataset by considering the variance of the organ number and volume. Four measures (voxel accuracy, mean voxel accuracy, IU, and frequency weighted IU) that are commonly used in

Table 1 Accuracy evaluations in terms of mean value of IUs per target type between segmentation and ground truth in 230 training and 10 test CT scans after voting in 3D.

Target name	Mean value of IUs	
	Training samples (230)	Test samples (10)
Right Lung	0.92	0.87
Left Lung	0.91	0.88
Heart	0.87	0.87
Aorta	0.72	0.63
Esophagus	0.18	0.27
Liver	0.91	0.91
Gallbladder	0.58	0.48
Stomach and Duodenum (2nd pos.)	0.48	0.43
Stomach and Duodenum Lumen	0.59	0.61
Contents inside of Stomach and Duodenum	0.21	0.10
Spleen	0.85	0.86
Right Kidney	0.85	0.86
Left Kidney	0.85	0.84
Inferior Vena Cava	0.56	0.51
Portal Vein, Splenic Vein, and Superior Mesenteric Vein	0.32	0.03
Pancreas	0.48	0.45
Uterus	0.23	0.09
Prostate	0.48	0.35
Bladder	0.67	0.72

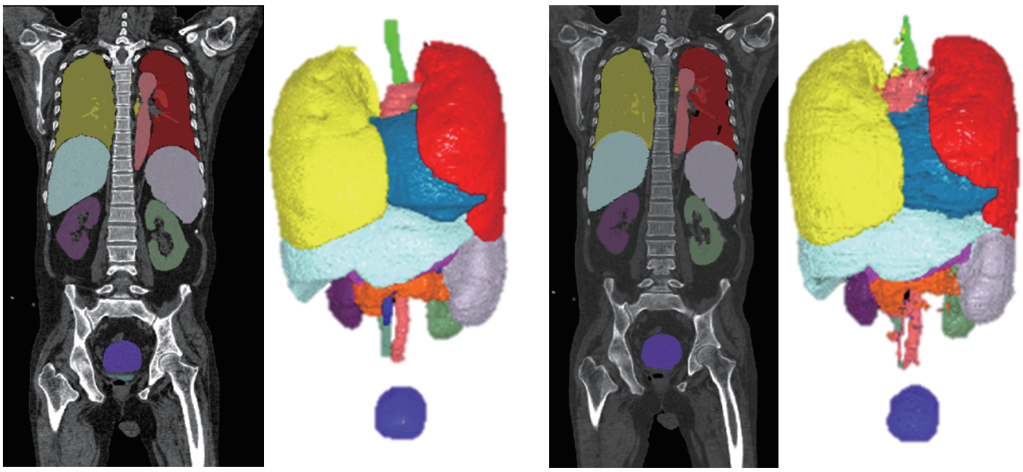

Fig.3 An example of segmentation in a 3D CT case. Left : corresponding ground truth, Right : segmented regions labeled with different colors for one 2D CT slice and 3D visualization based on surface-rendering method.

semantic segmentation and scene parsing were employed for the evaluations [16]. The evaluation results for the voxel accuracy, mean voxel accuracy, IU, and frequency weighted IU were 89%, 66%, 59%, and 84%, respectively, when averaged over all the segmentation results of the test dataset. These results show that 89% of the voxels within the anatomical structures (constructed using multiple target regions) were labeled correctly, with an coincidence (which is the same as the Jaccard similarity coefficient) of 59% for the test dataset. After normalizing the voxel accuracy and IU using the target numbers and volumes for different CT cases, these two values changed to 66 % and 84%, respectively.

4. DISCUSSION

We found that the target organs were recognized and extracted correctly in all the test CT images, except for oversights of the portal vein, splenic vein, and superior mesenteric vein in two CT cases. Because our segmentation targets covered a wide range of shapes, volumes, and sizes, either with or without contrast enhancement, and at different locations in the human body, these experimental results demonstrated the potential capability of our approach to recognize whole anatomical structures appearing in CT images. The IUs of the organs with larger volumes (e.g., liver : 91%, heart : 87%) were comparable to the accuracies reported from the previous state-of-the-art methods [2-9]. For some smaller organs (e.g., gallbladder) or line structures (e.g., portal vein, splenic vein, and superior mesenteric vein) that have not been reported in previous work, our segmentation did not show particularly high IUs, but this performance was deemed reasonable because the IU tends to be lower for those organs with smaller volumes. The physical CT image resolution is the major cause of this limited performance, rather than the segmentation method. Our evaluation showed that the average segmentation accuracy of all the targets over all the test CT images was approximately 84% in terms of the frequency weighted IUs. From this, we see that our approach can recognize and extract all types of major organs simultaneously, achieving a reasonable accuracy according to the organ volume in the CT images.

5. CONCLUSION

We proposed a novel approach for the automatic segmentation of anatomical structures (multiple organs) in CT images, based on a deep fully convolutional network. This approach was applied to segment 19 types of targets in 3D CT cases, demonstrating highly promising results. Our work is the first to tackle anatomical segmentation (with a maximum of 19 targets) on scale-free CT scans (both 2D and 3D images) through a single deep neural network. The proposed approach could also be extended as a general solution for more complex anatomical structure segmentation in other image modalities that remain fundamental problems in medical physics (e.g., MR and PET imaging).

ACKNOWLEDGMENTS

The authors would like to thank all the members of the Fujita Laboratory in the Graduate School of Medicine, Gifu University for their collaborations. We would like to thank all the members of the Computational Anatomy [18] research project, especially Dr. Ueno of Tokushima University, for providing the CT image database. This research was supported in part by a Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 26108005), and in part by a Grant-in-Aid for Scientific Research (C26330134), MEXT, Japan.

References

[1] Pham DL, Xu C, Prince JL : Current methods in medical image segmentation, Biomed. Eng., 2, 315-333, 2000.
[2] Heimann T, and Meinzer HP : Statistical shape models for 3D medical image segmentation : A review, Med. Image Anal., 13(4), 543-563, 2009.
[3] Xu Y, Xu C, Kuang X, et al. : 3D-SIFT-Flow for atlas-based CT liver image segmentation, Med. Phys. 43(5), 2229-2241, 2016.
[4] Lay N, Birkbeck N, Zhang J, et al. : Rapid multi-organ segmentation using context integration and discriminative models, Proc. IPMI, 7917, 450-462, 2013.
[5] Shimizu A, Ohno R, Ikegami T, et al. : Segmentation of multiple organs in non-contrast 3D abdominal CT images, Int. J. Compt. Assisted Radiol. Surg., 2, 135-142, 2007.
[6] Wolz R, Chu C, Misawa K, et al. : Automated abdominal multi-organ segmentation with subject-specific atlas generation, IEEE Trans. Med. Imaging, 32(9), 1723-1730, 2013.
[7] Okada T, Linguraru MG, Hori M et al. : Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors, Med. Image Anal., 26(1), 1-18, 2015.
[8] Bagci U, Udupa JK, Mendhiratta N, et al. : Joint segmentation of anatomical and functional images : Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images, Med. Image Anal., 17(8), 929-945, 2013.
[9] Sun K, Udupa JK, Odhner D, et al. : Automatic thoracic anatomy segmentation on CT images using hierarchical fuzzy models and registration, Med. Phys., 43(4), 1882-1896, 2016.
[10] Shin HC, Roth HR, Gao M, et al. : Deep convolutional neural networks for computer-aided detection : CNN architectures, dataset characteristics and transfer learning, IEEE Tran. Med. Imaging, 35(5), 1285-1298, 2016.
[11] Teramoto A, Fujita H, Yamamuro O, et al. : Automated detection of pulmonary nodules in PET/CT images : Ensemble false-positive reduction using a convolutional neural network technique, Med. Phys., 43(6), 2821-2827, 2016.
[12] Nappi JJ, Hironaka T, Regge D, et al. : Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography, Proc. SPIE Medical Imaging 2016 : Computer-Aided Diagnosis, 9785, 97852B-1-97852B-8, 2016.
[13] Brebisson AD, and Montana G : Deep neural networks for anatomical brain segmentation, Proc. CVPR workshops, 20-28, 2015.
[14] Roth HR, Farag A, Lu L, et al. : Deep convolutional

networks for pancreas segmentation in CT imaging, Proc. SPIE Medical Imaging 2016 : Image Processing, 9413, 94131G-1-94131G-8, 2015.
[15] Cha KH, Hadjiiski L, Samala RK, et al. : Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets, Med. Phys., 43(4), 1882-1896, 2016.
[16] Long J, Shelhamer E, Darrell T : Fully convolutional networks for semantic segmentation, Proc. CVPR, 3431-3440, 2015.
[17] Simonyan K and Zisserman A : Very deep convolutional networks for large-scale image recognition, Proc. ICLR, arXiv : 1409.1556, 2015.
[18] http : //www.comp-anatomy.org/wiki/
[19] Watanabe H, Shimizu A, Ueno J, et al. : Semi-automated organ segmentation using 3-dimensional medical imagery through sparse representation, Trans. of Jap. Society for Med. and Bio. Eng., 51(5), 300-312, 2013.