Multiorgan structures detection using deep convolutional neural networks

Jorge Onieva Onieva[a], Germán González Serrano[a], Thomas P. Young[a], George R. Washko[b], María Jesús Ledesma Carbayo[c], and Raúl San José Estépar[a]

[a]Applied Chest Imaging Laboratory, Dept. of Radiology, Brigham and Women's Hospital, 1249 Boylston St, Boston, MA USA
[b]Division of Pulmonary and Critical Care, Dept. of Medicine, Brigham and Women's Hospital, 72 Francis St, Boston, MA USA
[c]Biomedical Image Technologies Laboratory (BIT), ETSI Telecomunicación, Universidad Politécnica de Madrid and CIBER-BBN, Madrid, Spain

ABSTRACT

Many automatic image analysis algorithms in medical imaging require a good initialization to work properly. A similar problem occurs in many imaging-based clinical workflows, which depend on anatomical landmarks. The localization of anatomic structures based on a defined context provides with a solution to that problem, which turns out to be more challenging in medical imaging where labeled images are difficult to obtain. We propose a two-stage process to detect and regress 2D bounding boxes of predefined anatomical structures based on a 2D surrounding context. First, we use a deep convolutional neural network (DCNN) architecture to detect the optimal slice where an anatomical structure is present, based on relevant landmark features. After this detection, we employ a similar architecture to perform a 2D regression with the aim of proposing a bounding box where the structure is encompassed. We trained and tested our system for 57 anatomical structures defined in axial, sagittal and coronal planes with a dataset of 504 labeled Computed Tomography (CT) scans. We compared our method with a well-known object detection algorithm (Viola Jones) and with the inter-rater error for two human experts. Despite the relatively small number of scans and the exhaustive number of structures analyzed, our method obtained promising and consistent results, which proves our architecture very generalizable to other anatomical structures.

Keywords: organ detector, convolutional neural network, deep learning, computed tomography

1. INTRODUCTION

Machine learning has been extensively used in the field of medical imaging in the last years. Computer-aided diagnosis (CAD) systems and medical image analysis tools have been more and more widely adopted as the development of new algorithms overcome some of the challenges present in this area.[1]

One of these challenges, of vital importance for many of these algorithms, is the localization of an anatomical structure in a 2D slice inside of a 3D image volume. This is commonly needed for algorithm initialization, quick image retrieval or computation of biomarkers within the region of interest, among other tasks. A similar problem occurs in many clinical workflows supported by imaging, which are based on the detection of a particular anatomical landmark. Having in mind that an anatomic structure is typically visible in several slices, it is often needed to choose not only a slice where the structure is visible but also the most relevant one for the problem to solve based on a certain anatomical context.

In the last years, deep convolutional neural networks (DCNN) have proven to be very effective in image classification and object detection in traditional computer vision challenges like ImageNet.[2] However, the state

Further author information: (Correspondence: Jorge Onieva and Raúl San José Estépar)
Jorge Onieva: E-mail: jonieva@bwh.harvard.edu
Raúl San José Estépar: E-mail: rsanjose@bwh.harvard.edu

Medical Imaging 2018: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE Vol. 10574, 1057428 · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293761

Plane	Structures
Axial (n=25)	Ascending aorta, carina, left coronary artery, left/right diaphragm, left/right humerus, left/right kidney, heart, left/right pectoralis, left/right _/anterior/posterior chest wall, liver, pulmonary artery, spleen, sternum, transversal aorta, left/right scapula
Sagittal (n=17)	Ascending aorta, left/right anterior/posterior chest wall, left/right diaphragm, heart, left/right hilum, left ventricle, liver, pulmonary artery, spine, sternum, trachea, transversal aorta
Coronal (n=15)	Ascending/descending aorta, carina, left/right chest wall, left/right diaphragm, heart, left ventricle, liver, pulmonary artery, spine, spleen, trachea, left subclavian artery

Table 1. List of analyzed structures in axial, sagittal and coronal planes

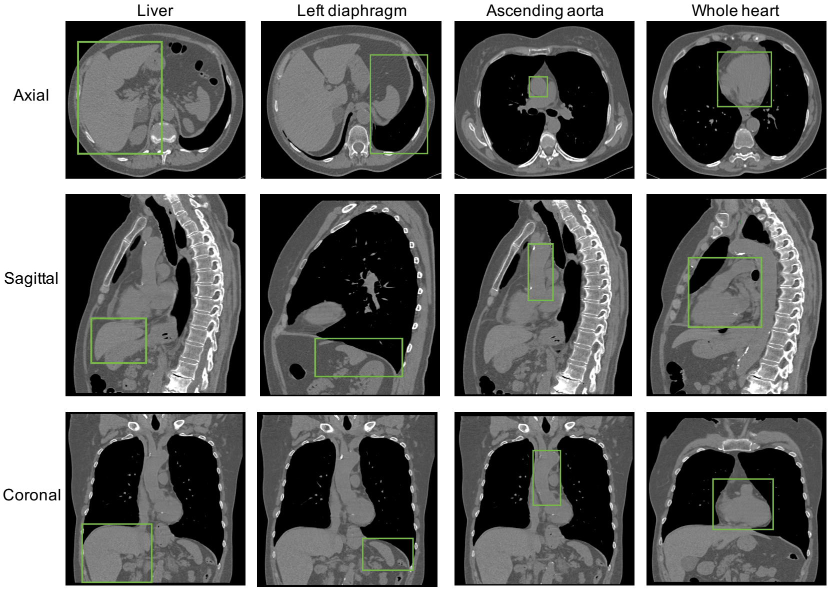


Figure 1. Sample structures (manual annotations)

of the art networks that perform well in this type of tasks, like Faster RCNN[3] or YoLo,[4] make use of very deep architectures with millions of parameters that require large datasets to be trained. This is a major hurdle in medical imaging, where labeled data is scarce.

One of the advantages of performing computer vision analysis in medical imaging versus natural scenes is the lower variability in the image content. Anatomic structures are pretty consistent in location, shape, and size for different image modalities. Based on that, we hypothesized that a simple architecture based on DCNNs is sufficient to perform both the slice and the bounding box detection of anatomical structures accurately, with the advantage of requiring less training data than more complex architectures.

In this paper, we present a DCNN architecture that addressed the detection of 2D structures in chest CT images. While previous work in this area focuses on a single or a few structures,[5,6] we analyzed 57 structures in different planes that have been considered of relevance for different diagnostic and quantitative tasks by medical experts, detailed in Table 1. We pose the problem in 2D instead of 3D since 2D detections have been proven sufficient in the search for image-based biomarkers in different clinical problems.[7,8] Besides, the inherent higher complexity of 3D architectures implies the need for more training data to achieve stable results.

Figure 1 shows some examples of the structures that have been analyzed. These bounding boxes were labeled manually by a trained expert. It is important to note that there is not a single solution to solve this problem since an anatomical structure can be visible in more than one slice. In order to decide which is the best slice to label a structure, we used anatomical criteria based on other nearby structures, either in the same plane or a different one. For example, note in the figure that the ascending aorta in the coronal plane is labeled on the same slice that the liver.

For those structures for which we have complete data for both methods (47 structures), we compared the results with Viola-Jones (VJ),[9] a 2D computer vision algorithm widely used in environments with limited computational resources.

2. METHODS

We developed a 2-stage approach based on two DCNNs. The first one determines which is the more likely slice within a 3D volume to define an anatomical structure of interest, while the second one locates the structure bounding box within the chosen slice. The architecture and hyper-parameters were both common for all the structures, but each one of the structures was trained independently from scratch. The system was tested in 57 anatomical structures of interest that we have defined along the principal planes of chest CT scans (axial, sagittal and coronal): 25 in the axial plane, 17 in the sagittal and 15 in coronal. Depending on the feasibility to detect the structure in the plane, some of them may be defined in the three planes or just in two of them. The structures were defined by an expert based on clinical criteria, and range from large organs (like lungs and heart) to smaller anatomical landmarks (like the hilum). Note how the size and shape of these structures can be dramatically different, which increases the problem complexity (a fixed size bounding box prediction will not be sufficient).

It is important to remark that while the VJ method required some customized parametrization for each one of the structures, we did not use any particular structure parametrization in DCNN.

2.1 Architecture

Our solution makes use of two DCNNs, one for the slice detection (hereinafter classification network) and another one for the bounding box regression (regression network). The detailed description of DCNNs are out of the scope of this article, and we refer the reader to[10] for a helpful introduction to the subject.

Our two DCNNs share a very similar architecture (Figure 2). We will describe first the classification network, and then we will just describe the differences with the regression network. The design resembles a typical

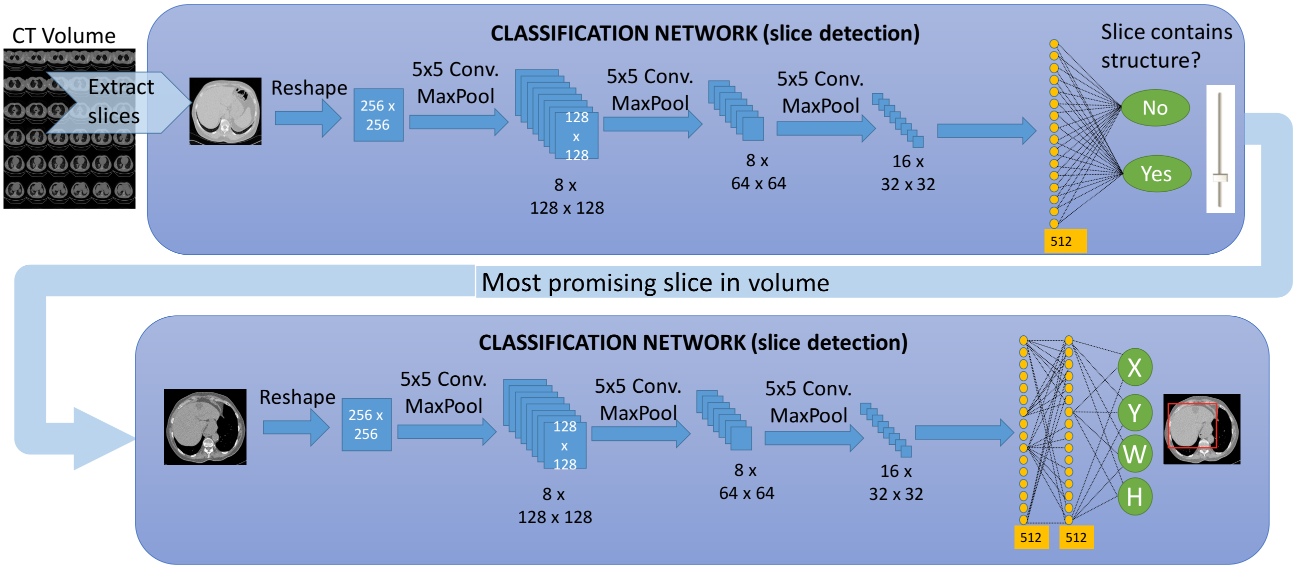


Figure 2. Classification and regression network architecture

classification architecture like the one described in[11] that has been proved to work in many general computer vision classification problems. Each network can receive as input images from any 2D plane in a 3D volume. In order to standardize this input, we reshape all the images to a size of 256x256 pixels, using linear interpolation. The reshaped input is processed by three convolutional layers of 8, 8 and 16 convolutional filters respectively, each one of them with a size of 5x5 pixels. Each layer is followed by a max pool layer of 2x2 that reduces the image dimensionality by a factor of 4 after each layer. All the layers (except the output of the network) make use of a rectified linear unit (ReLU) layer that discards all the negative activations. At the end of the convolutional layers, we use a simple 512 fully connected layer that will characterize the image features necessary to identify the anatomical landmarks. The output of the network is compound of just a size-2 vector whose second position contains a score (normalized with a Softmax activation) that indicates the probability of the input image to display the anatomical structure. The loss function used to train this network is the categorical cross entropy.

The bounding box detection is made with a regression network that outputs a vector (X, Y, W, H) with the top left corner coordinates, the width and the height of the region of interest (RoI). All the components are normalized to a [0.0-1.0] range. The output neurons employ a sigmoid activation function to produce the normalized result.

Both networks were trained using the Adam optimization implemented in the TensorFlow library (v0.10.0), with a learning rate, lr=0.001, that decays linearly with a 0.8 factor when there are no improvements in the loss function for the validation dataset. We did not use any data augmentation in the classification network. A slice was considered a positive sample when the distance between the slice and the labeled ground truth was less than 5 slices. Conversely, we did use data augmentation when training the regression network, consisting of random small rotations ($\pm 5°$), translations ($\pm 0.05 * originalSize$), and scaling ($\pm 0.05 * originalSize$). In total, we generated 20 images for every training sample.

2.2 Data preparation

We used a chest CT cohort extracted from the COPD Gene study[12] in order to evaluate our results. In total, we used 306 cases for the training and 198 cases for testing, with small variations depending on the study structure since some structures were not always visible in the chest CT scan.

An expert labeled a 2D bounding box for each one of the structures to study, which was used as ground truth for the training. Moreover, we asked another expert to repeat the labeling of 20 scans, to measure the inter-rate variability using the inter-class correlation. We also assessed the error between human raters to provide a reference.

Before training our models, we applied some preprocessing to the training datasets, both in the classification and the regression network. Every slice image was resized to a 256x256 pixels image. CT images were clamped in the range [-500,300] HU as all the structures were within that range. The intensities were later normalized to a [0.0-1.0] range, in order to prevent numerical instabilities during training.

Regarding the bounding box regression, all the coordinates are scaled to a [0.0-1.0] range, which allows the network to be invariant to the scale of the images. Also, setting the coordinates in this format simplifies the training, as the solution space is reduced to a [0.0-1.0] range for any structure.

2.3 Optimal slice selection

Some structures are large and visible in many slices, making it difficult to select the optimal view of the anatomical structure. The classification network may determine more than one slice with a maximum score of 1.0. For instance, if we think about axial plane where the whole heart is visible (one of the structures that we have defined), more than 25 slices may have the maximal response. In order to select the most likely slice, we selected the slices with maximum likelihood as defined by the DCNN output:

$$S = argmax_i P(x|s_i) \tag{1}$$

where si is the slice, x is the structure under detection and S denotes the set of slices that maximize the probability P. In case there was more than one maximum in S, we selected the subset with the largest cardinality and chose their middle slice as the candidate solution.

2.4 Method comparison

We compared the performance of our method with VJ.[9] For the slice detection stage, we used the absolute error between the reference slice labeled by an expert and the slice that the algorithm predicts as the best candidate. We also analyzed the correlation between the predicted and the correct slice for both methods. Differences in the mean between both approaches were assessed by means of the Wilcoxon test.[13] Bonferroni was used to correct for multiple comparisons.[14] The difference in correlations was analyzed using the Fisher r-to-z transformation.

The coordinates of the bounding box, (X, Y, W, H), were analyzed using a correlation analysis. X and Y represent the coordinates of the top-left corner of the bounding box, and W and H represent the width and the height of the structure respectively.

In order to have a global measurement of the accuracy of the full bounding box, we calculated the intersection over union (IoU) of the predicted bounding box with respect to the ground truth manually labeled. The IoU between two bounding boxes is defined as the coefficient between their area of overlap (area of the intersection) and the area of the smallest bounding box that would contain the two boxes (area of the union). Wilcoxon test was employed in the IoU evaluation to assess differences between our method and VJ.

We tested a total of 8765 case-structure combinations distributed in 198 test CT scans corresponding to subjects that were not used in training and 47 anatomic structures. Note that not all the structures were visible in every scan.

2.5 Inter-reader variability

The structures detections in our training and testing datasets were performed by one trained expert. We assessed the inter-reader variability for the slice detection in 20 cases that were randomly selected from our training dataset and in a blind fashion. The intraclass correlation coefficient between readers was excellent for 35 structures and good for 6. However, 12 structures (Left Anterior Chest Wall Sagittal, Left Diaphragm Axial, Left Hilum Sagittal, Left Posterior Chest Wall Sagittal, Left Ventricle Sagittal, Liver Axial, Liver Sagittal, Right Diaphragm Axial, Right Hilum Sagittal, Spine Sagittal, Spleen Axial, Whole Heart Sagittal) and 4 structures (Left Kidney Axial, Right Anterior Chest Wall Sagittal, Right Diaphragm Sagittal, Right Posterior Chest Wall Sagittal) structures yielded a fair and poor intraclass correlation coefficient respectively. This indicates that there is a reasonable inter-reader variability in the slice detection when performed by different humans and provides a level of consistency for our training database even if it was performed just by one reader. The mean absolute slice prediction error between readers was 10.15 ± 6.23 providing a lower bound for the mean detection error.

3. RESULTS

In the following sections, we describe the detection error analysis that we conducted for the slice prediction and the bounding box regression methods.

3.1 Slice prediction

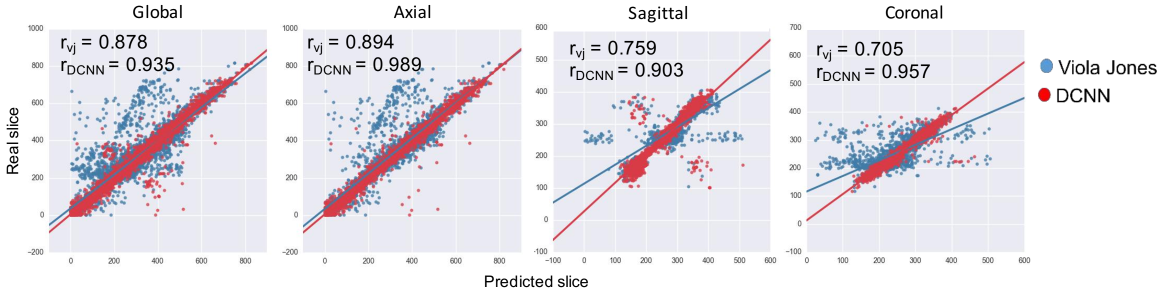


Figure 3. Correlations comparison VJ vs DCNN for the slice detection prediction. Our approach shows a higher correlation with less variance around the unit line. Performance in the sagittal and coronal planes is superior to VJ

The Pearson correlation of the slice prediction versus the reference value for all the structures was 0.878 and 0.935 for VJ and DCNN respectively. The correlation difference between VJ and DCNN was highly significant ($p < 10^{-8}$). The slope of the regression lines for the slice detection was 0.851 and 0.986 for VJ and DCNN respectively, which shows a better fit to the unity line for the DCNN solution. We further analyzed the correlations for the structures belonging to each plane (axial, sagittal or coronal), displayed in Figure 3. For all cases, DCNN correlations were significantly stronger than VJ ($p < 10^{-8}$). We noted that correlations were higher in the axial plane that in the other two planes for both VJ and DCNN. This result may be explained by the higher resolution of the axial plane in our CT scans, as well as the resampling of the images that we need to do for sagittal and coronal planes to fit into the fixed input size of the network. While axial images are reconstructed in a fixed 512x512 sampling grid, coronal and sagittal images dimensions depend on the selected slice thickness and scanning field of view.

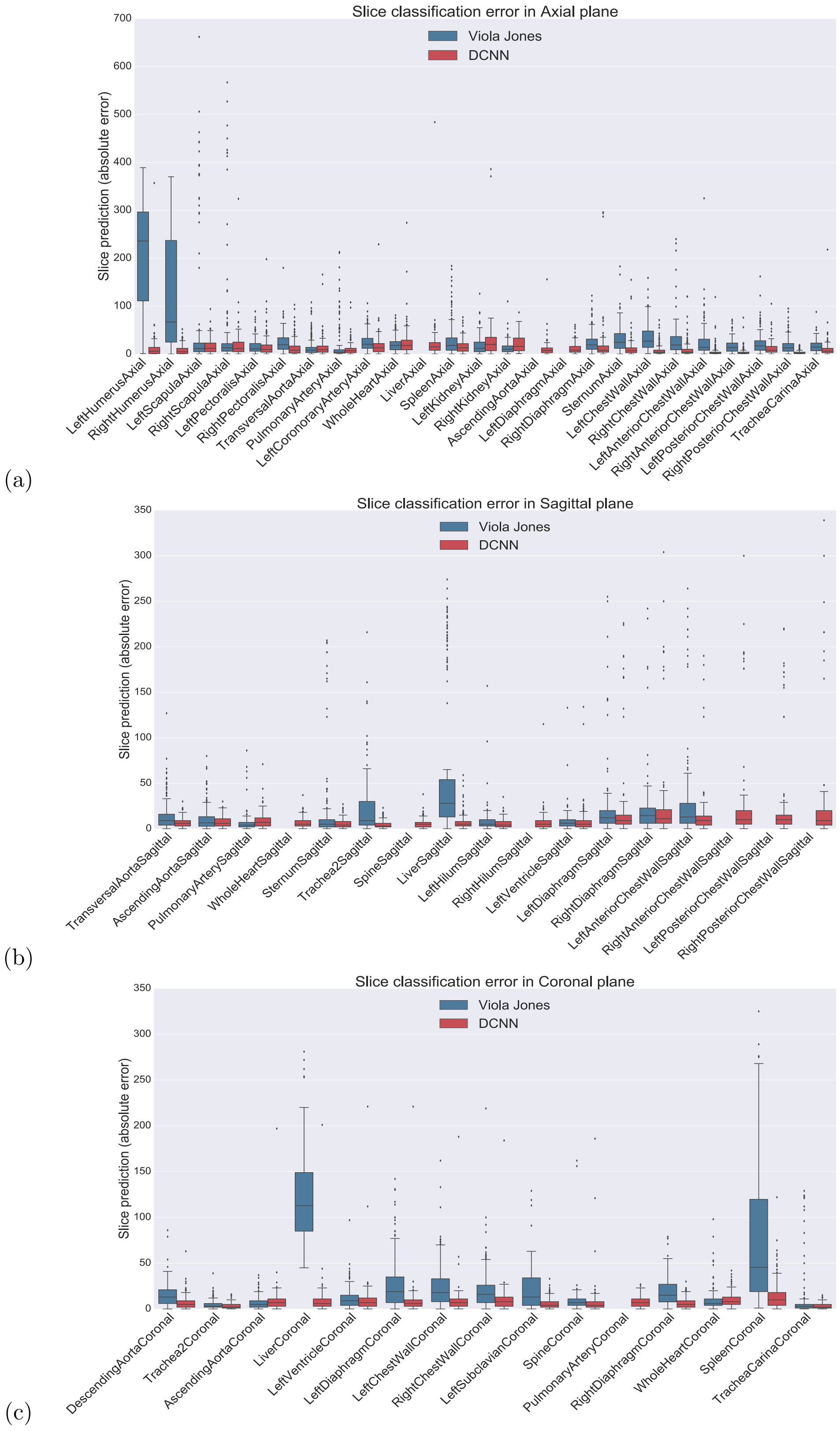
The absolute error between the slice prediction and the reference value was computed for each structure. 28 out of 47 structures reached statistically significant difference between our method and VJ after Bonferroni correction ($p < \frac{0.05}{47} = 0.001$). Figure 4 shows the absolute slice distance error for every structure in the axial, sagittal and coronal planes. For all cases (except the Pulmonary Artery in the sagittal plane), our approach has a mean error smaller than VJ.

3.2 Bounding box regression

Our bounding box analysis was performed in those case-structure pairs where the error in the slice prediction for VJ and DCNN was less than 10 slices to perform an unbiased evaluation of this stage. A total of 3085 case-structure combinations were studied. One of the structures (the liver in the coronal plane) did not meet these criteria.

The global correlation for each bounding box coordinate is shown in Figure 5. For all cases, DCNN correlations were statistically higher than VJ ($p < 10^{-8}$). The top-left coordinates, X and Y, show a better correlation that width (W) and height (H) consistently for both methods. This might indicate that the location of the bounding box is more precise than the sizing.

The difference in IoU between our method and VJ was statistically significant in 25 out of the 47 anatomical structures ($p \leq 0.001$). Overall, DCNN presented a higher IoU mean value in each one of the structures analyzed, which indicates the clear superiority of the method for the structures detection. Figure 6 shows the IoU error for each one of the structures in the axial, sagittal and coronal planes respectively.



(a)

(b)

(c)

Figure 4. Absolute slice error comparison between VJ and DCNN in the axial (a), sagittal (b) and coronal (c) planes

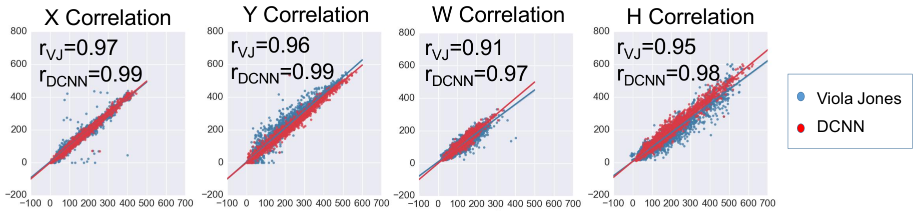
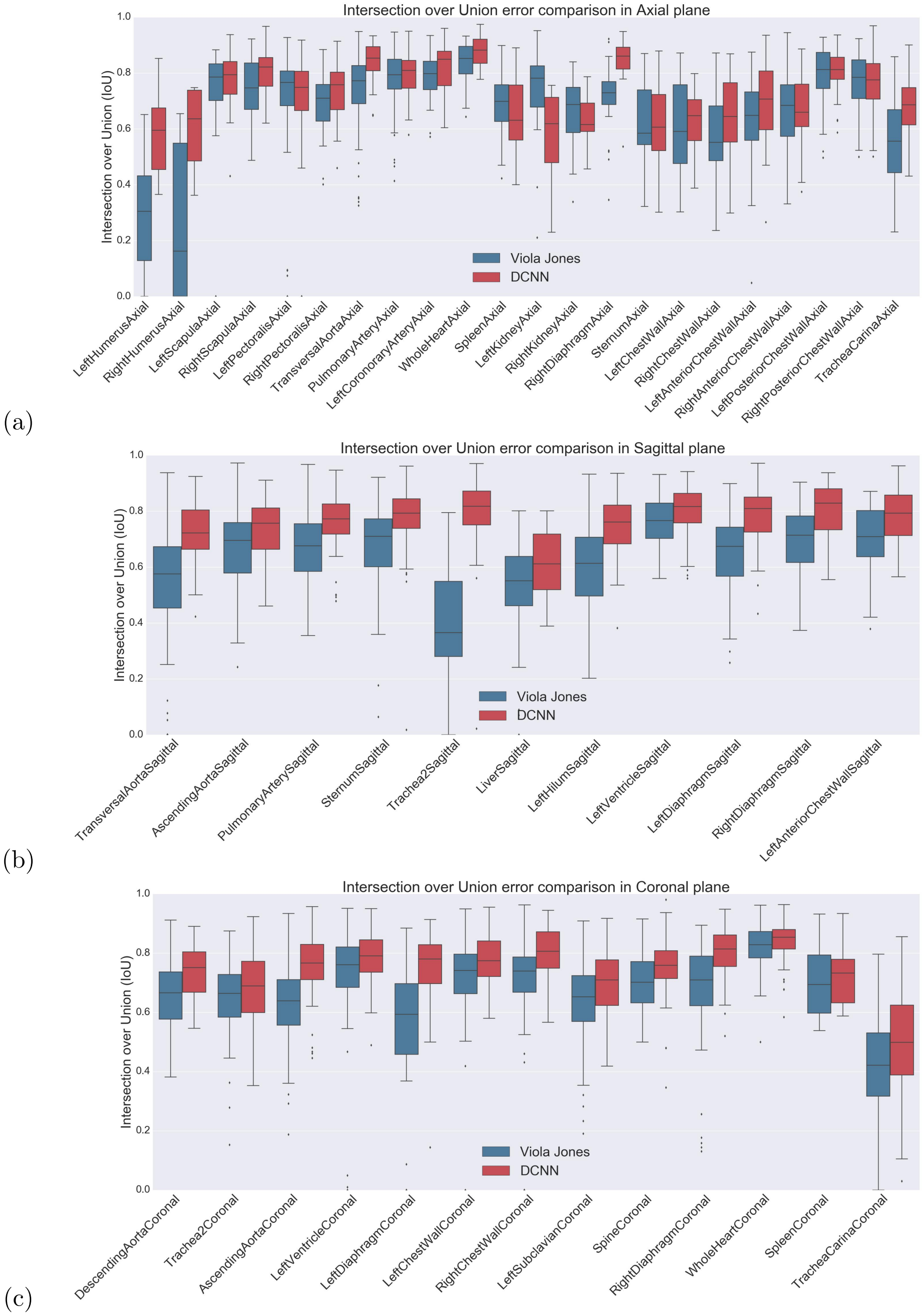


Figure 5. Correlation for each one of the bounding box coordinates: top-left corner X, top-left corner Y, Width and Height



(a)

(b)

(c)

Figure 6. IoU comparison between VJ and DCNN in the axial (a), sagittal (b) and coronal (c) planes.

3.3 Full cases examples

We selected two patients from our testing dataset to illustrate the results of our method. To that aim, we selected two cases that were in the upper and lower performance percentiles. In order to choose a case with a good performance, we first selected the cases with a mean absolute slice detection error below the 10th percentile in the test dataset. Then, we chose the case with highest mean IoU for all the structures. The process is analogous for the case with a lower performance, using the 90th percentile and the lowest mean IoU. For each case, we selected the three structures with a higher/lower IoU for each plane.

The results for the cases with a higher/lower performance are displayed in Figures 7 and 8 respectively. For each structure, we show the automatic prediction obtained by our method (left side, in red), versus the manually labeled figure (right side, in green). It is important to observe that even the structures with a higher quantitative error show very reasonable results in a qualitative sense. This stability of the results, even for the worst performing cases, may be in part explained by the spatial consistency of the structures across multiple slices. Although our method does not always choose the same slice as defined in the reference standard, the selected slide is consistent within the proposed structure definition. In addition to that, the detection of the bounding box seems to be quite robust once the slice is selected as reflected in our numerical results in Figure 6.

4. CONCLUSIONS

We have proposed a structure detection method based on a two-stage DCNN that has demonstrated good performance for a large number of structures, and it is competitive with human raters. After an exhaustive comparison between VJ and our method, we can conclude that DCNN presents a consistent improvement over VJ for most of the anatomical structures analyzed. While the improvement in the slice detection was statistically significant in approximately half of the defined structures, the bounding box object detection was superior in DCNN, where all the structures performed better on average with no exception.

It is also proven that the proposed method is very scalable. Conversely to VJ, we did not use any structure-specific parameterization or data preprocessing in any case, which shows the potential for the generalization of our solution for its use in different structures with no customization required.

When we compare our method to the performance achieved by humans, the mean absolute slice prediction error between readers is 10.15±6.23 in comparison to 31.61±29.27 for DCNN. This implies that there is still room for improvement for the method to be within the inter-reader error. Nevertheless, our visual results suggest that the performance is consistent in average for those structures even for those cases where our method performed the upper error percentile.

One limitation of our method is that each structure is trained independently, so our approach does not take advantage of the spatial relationship between structures a strong prior in medical imaging detection tasks. Our future work will make use of a multiclass network that could identify more than one structure at the same time, which allows learning some anatomical information and potentially reducing the computation time.

ACKNOWLEDGMENTS

This work has been funded by NIH NHLBI grants R01-HL116931 and R21HL140422. The Titan Xp used for this research was donated by the NVIDIA Corporation.

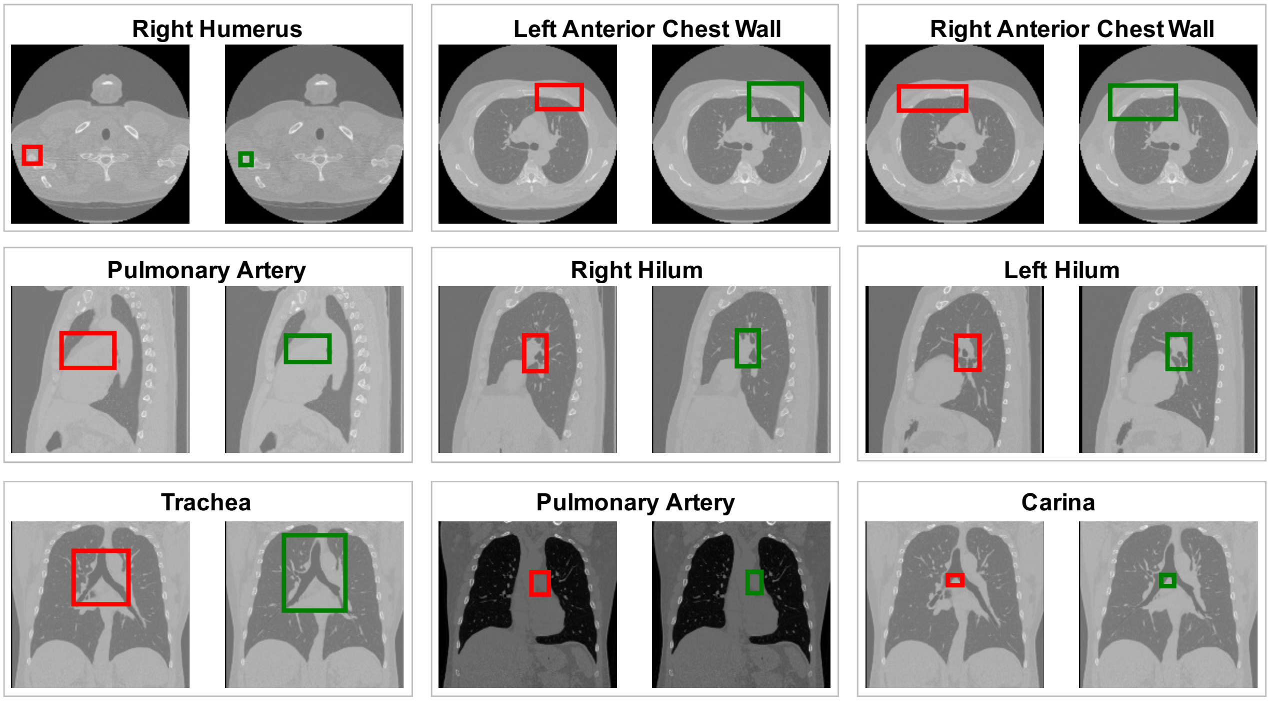


Figure 7. Structures for the lower percentile error case. Automatic detection (red) versus manually labeled (green) structures. The structures with the highest IoU were selected for each plane.

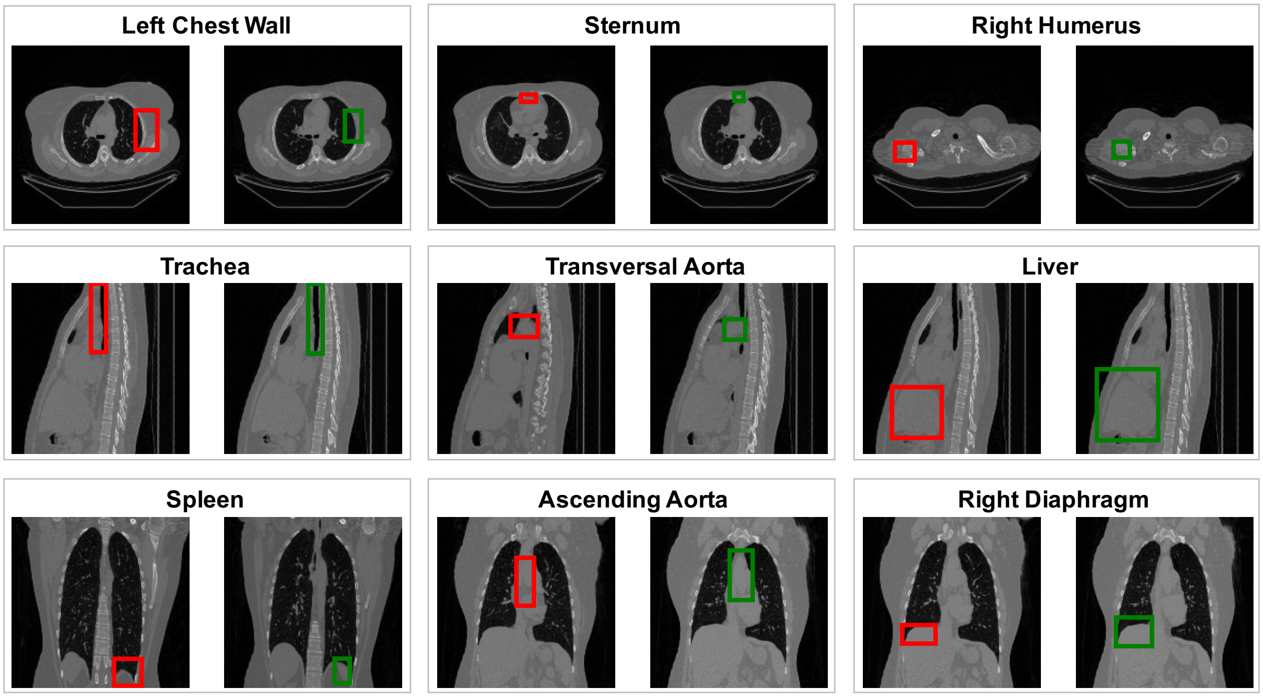


Figure 8. Structure for the upper percentile error case. Automatic detection (red) versus manually labeled (green) structures. The structures with the lowest IoU were selected for each plane.

REFERENCES

[1] Duncan, J. and Ayache, N., “Medical image analysis: progress over two decades and the challenges ahead,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 85–106 (jan 2000).

[2] Krizhevsky, A., Sulskever, I., and Hinton, G. E., “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information and Processing Systems (NIPS)* , 1–9 (2012).

[3] Ren, S., He, K., Girshick, R., and Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *ArXiv 2015* , 1–10 (2015).

[4] Redmon, J. and Farhadi, A., “Yolo9000: Better, Faster. Stronger,” *Arxiv* (dec 2017).

[5] Lu, X., Xu, D., and Liu, D., “Robust 3D organ localization with dual learning architectures and fusion,” in [*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*], (2016).

[6] de Vos, B., Wolterink, J., de Jong, P., Leiner, T., Viergever, M., and Isgum, I., “ConvNet-Based Localization of Anatomical Structures in 3D Medical Images,” *IEEE Transactions on Medical Imaging* , 1–1 (2017).

[7] Wells, J. M., Washko, G. R., Han, M. K., Abbas, N., Nath, H., Mamary, a. J., Regan, E., Bailey, W. C., Martinez, F. J., Westfall, E., Beaty, T. H., Curran-Everett, D., Curtis, J. L., Hokanson, J. E., Lynch, D. a., Make, B. J., Crapo, J. D., Silverman, E. K., Bowler, R. P., and Dransfield, M. T., “Pulmonary Arterial Enlargement and Acute Exacerbations of COPD,” *New England Journal of Medicine* **367**(10), 913–921 (2012).

[8] Ross, J. C., Kindlmann, G. L., Okajima, Y., Hatabu, H., Díaz, A. a., Silverman, E. K., Washko, G. R., Dy, J., and San José Estépar, R., “Pulmonary lobe segmentation based on ridge surface sampling and shape model fitting,” *Medical physics* **40**, 121903 (dec 2013).

[9] Viola, P. and Jones, M., “Rapid object detection using a boosted cascade of simple features,” *Computer Vision and Pattern Recognition (CVPR)* **1**, I—-511—-I—-518 (2001).

[10] LeCun, Y., Bengio, Y., and Hinton, G., “Deep learning,” *Nature* **521**(7553), 436–444 (2015).

[11] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE* **86**(11), 2278–2323 (1998).

[12] Regan, E. a., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. a., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D., “Genetic epidemiology of COPD (COPDGene) study design,” *Copd* **7**(1), 32–43 (2010).

[13] Siegel, S., “Nonparametric statistics for the behavioral sciences,” (1956).

[14] Bonferroni, C. E., [*Teoria statistica delle classi e calcolo delle probabilita*], Libreria internazionale Seeber (1936).