DON'T MISS OUR SPECIAL SESSION AT ASTRO

Performing QA Remotely in the Age of COVID

October 26, 11:45 AM EST

At home. On site. **In sync.**

SunCHECK™ enables complete, collaborative **remote QA coverage** for COVID-19 & beyond.

Click to explore:

- Advantages of a centralized Patient & Machine QA solution
- How SunCHECK eased the transition to remote work for users worldwide
- Three new ways we're simplifying Platform adoption

Go to: **sunnuclear.com/getprepared**

SUN NUCLEAR corporation

Patient Safety Starts Here

BRR-Net: A tandem architectural CNN–RNN for automatic body region localization in CT images

Vibhu Agrawal, Jayaram Udupa[a)], Yubing Tong, and Drew Torigian
Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, USA

(Received 13 March 2020; revised 22 June 2020; accepted for publication 22 July 2020;
published xx xxxx xxxx)

Purpose: Automatic identification of consistently defined body regions in medical images is vital in many applications. In this paper, we describe a method to automatically demarcate the superior and inferior boundaries for neck, thorax, abdomen, and pelvis body regions in computed tomography (CT) images.

Methods: For any three-dimensional (3D) CT image *I*, following precise anatomic definitions, we denote the superior and inferior axial boundary slices of the neck, thorax, abdomen, and pelvis body regions by *NS(I)*, *NI(I)*, *TS(I)*, *TI(I)*, *AS(I)*, *AI(I)*, *PS(I)*, and *PI(I)*, respectively. Of these, by definition, *AI (I) = PS(I)*, and so the problem reduces to demarcating seven body region boundaries. Our method consists of a two-step approach. In the first step, a convolutional neural network (CNN) is trained to classify each axial slice in *I* into one of nine categories: the seven body region boundaries, plus legs (defined as all axial slices inferior to *PI(I)*), and the none-of-the-above category. This CNN uses a multichannel approach to exploit the interslice contrast, providing the neural network with additional visual context at the body region boundaries. In the second step, to improve the predictions for body region boundaries that are very subtle and that exhibit low contrast, a recurrent neural network (RNN) is trained on features extracted by CNN, limited to a flexible window about the predictions from the CNN.

Results: The method is evaluated on low-dose CT images from 442 patient scans, divided into training and testing sets with a ratio of 70:30. Using only the CNN, overall absolute localization error for *NS(I)*, *NI(I)*, *TS(I)*, *TI(I)*, *AS(I)*, *AI(I)*, and *PI(I)* expressed in terms of number of slices (nS) is (mean $\pm$ SD): 0.61 ± 0.58, 1.05 ± 1.13, 0.31 ± 0.46, 1.85 ± 1.96, 0.57 ± 2.44, 3.42 ± 3.16, and 0.50 ± 0.50, respectively. Using the RNN to refine the CNN's predictions for select classes improved the accuracy of *TI(I)* and *AI(I)* to: 1.35 ± 1.71 and 2.83 ± 2.75, respectively. This model outperforms the results achieved in our previous work by 2.4, 1.7, 3.1, 1.1, and 2 slices, respectively for *TS(I)*, *TI(I)*, *AS(I)*, *AI(I) = PS(I)*, and *PI(I)* classes with statistical significance. The model trained on low-dose CT images was also tested on diagnostic CT images for *NS(I)*, *NI(I),* and *TS(I)* classes; the resulting errors were: 1.48 ± 1.33, 2.56 ± 2.05, and 0.58 ± 0.71, respectively.

Conclusions: Standardized body region definitions are a prerequisite for effective implementation of quantitative radiology, but the literature is severely lacking in the precise identification of body regions. The method presented in this paper significantly outperforms earlier works by a large margin, and the deviations of our results from ground truth are comparable to variations observed in manual labeling by experts. The solution presented in this work is critical to the adoption and employment of the idea of standardized body regions, and clears the path for development of applications requiring accurate demarcations of body regions. The work is indispensable for automatic anatomy recognition, delineation, and contouring for radiation therapy planning, as it not only automates an essential part of the process, but also removes the dependency on experts for accurately demarcating body regions in a study. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14439]

Key words: automatic anatomy recognition, body region identification, computed tomography (CT), deep learning

1. INTRODUCTION

1.A. Background

An important step toward effective implementation of quantitative radiology is the recognition and delineation of objects in the human body. An essential part of this task is standardizing definitions for such objects, which may be organs, tissue regions, or well-defined body regions,[1,2] in order to develop generalizable body-wide methods, standardize clinical operations, and use the quantitative information meaningfully. With this ideology in mind, the standardization of body region boundaries is an equally important task, especially for objects that span body regions such as the thoracic spinal cord, the boundaries of which clearly depend on the definition of the thoracic region.

Our previous work on Automatic Anatomy Recognition (AAR)[2] to localize and/or delineate organs, tissue regions, and lymph-node zones uses standardized body region

definitions, but the body regions were located manually by specifying the superior and inferior axial slices for each body region. In this paper, we use the standardized body region definitions as described by Udupa et al.[2] and Wang et al.[3] and propose and substantiate a two-step approach for automatically locating the body region boundaries almost as accurately as knowledgeable human operators. Automation of this task is a vital step in minimizing the need of a radiological expert every time a framework like AAR is used. This work also helps in taking a step forward in the acceptance of standardized body region boundaries universally by providing an easy-to-use implementation for locating these boundaries in a volumetric image, the accuracy of which is remarkably close to human-level performance. The proposed methodology works regardless of whether the input volumetric computed tomography (CT) image spans the entire body or only a portion of the body.

1.B. Related works

Bai et al.[1] tackle our exact problem by using a system of virtual landmarks employing principal component analysis and recursive subdivision of objects, and subsequently using a neural network for mapping the virtual landmarks to boundary locations. Hussein et al.[4] propose a one-shot deep learning solution for automatically localizing the boundaries for the abdominal and thoracic regions by locating the superior and inferior boundaries of the abdomen and thorax, respectively, as a step in segmenting and quantifying intrathoracic adipose tissue in positron emission tomography/CT (PET/CT) images. In a more abstract sense of localization, Bai et al.[5] use an adaptive thresholding model to partition a PET/CT scan into three sections: above lungs, lungs, and below lungs. Criminisi et al.[6] use random forest regression for automatically detecting and localizing anatomical structures in CT images and predict bounding walls (and subsequently, the centroids) of various organs.

Operating on the slice level, Lee and Chung[7] segment each slice in a volumetric CT image into disconnected regions using a neural network and use the common information between adjacent slices and fuzzy rules based on spatial relationships for recognizing various organs in each slice. In a similar spirit as our work, Wang et al.[8] classify slices in volumetric CT images as either belonging/not belonging to the abdominal area using a convolutional neural network (CNN). They treat the problem as a binary classification problem considering each slice independently and using a one-dimensional median filter to smooth the results to remove spatial inconsistencies. de Vos et al.[9] use CNNs to localize anatomical structures in 3D images by detecting their presence in 2D image slices, using a single CNN to detect the presence of anatomical structures in axial, coronal, and sagittal slices from a 3D image and combining the outputs to create 3D bounding boxes.

None of the above works tackle the problem of body region localization as formulated in this paper, namely: *Given* a 3D image I comprising a stack of transaxial slices which represent a contiguous portion of the human body covering any of the four body regions — head and neck (H&N), thorax, abdomen, and pelvis — or any combination of them, *Objective 1* (O1): *to partition* the slices of I into the body regions to which they belong, *Objective 2* (O2): *label* each actual body region identified, and *Objective 3* (O3): *report* a message if the slices do not fully cover a body region at either end. For example, let I consist of 150 slices (numbered in the increasing order cranio-caudally) where in the superior and inferior boundaries of the H&N region are at slices 10 and 60, the superior boundary of the thorax is at slice 70, and slice 150 falls short of the inferior boundary of the thorax by 10 slices. Our system BRR-Net is designed to perform all three output actions O1-O3 and output the following: Superior boundary of H&N: slice 10; inferior boundary of H&N: slice 60; H&N body region: slices 10-60; superior boundary of thorax: slice 70; inferior boundary of thorax: not in I; thorax body region: incomplete. Bai et al.[1] focus on three body regions – thorax, abdomen, and pelvis, and do not consider H&N, although it is possible to extend their method to H&N. Hussein et al.[4] consider only the thorax and abdomen. More important than the smaller number of body regions considered by both works, their behavior becomes unpredictable (or not described/demonstrated in the papers) if I does not include the slices they aim to detect. The ability to perform O1-O3 has important consequences in practice where BRR-Net forms the front end of the application. We will provide three illustrations. (a) Standardized anatomy definition: In numerous applications, body-wide or body region-wide organ segmentation is required. Radiation therapy planning is an example. Unfortunately, none of the existing segmentation frameworks start off with a precise definition of the body region and the organs included in them, the AAR methodology[2,3,10] being an exception. Without such a definition, the segmentation problem and comparative evaluation of methods become ill-defined. For example, what is the definition of long objects such as esophagus, spinal cord, descending aorta, etc. that cross body regions? The AAR framework has demonstrated that standardized definitions have conceptual and algorithmic advantages. For example, by modeling object geographic relationship information with respect to the body region, objects can be quickly placed in an image based purely on prior knowledge. (b) Body composition analysis: Quantification of bodily tissues,[4,11] especially subcutaneous and visceral adipose components, has been shown to be useful as biomarkers in the study of various disease and treatment processes such as obstructive sleep apnea,[12] lung transplant surgery,[13] acute kidney injury in trauma,[14] etc. Without a precise definition of the body regions, standardized assessment of these tissues by body region becomes meaningless. (c) Disease quantification: It has been demonstrated recently[15] that body-wide, body region-wide, and organ-wide disease quantification can be performed via PET/CT images just following localization of these entities without having to perform their explicit delineation or the delineation of the disease sites. Again, without a precise definition and subsequent localization of the objects, this quantification has no

meaning. The outputs of BRR-Net are crucial in facilitating these practical applications.

1.C. Outline of the approach

We assume that the four body regions discussed in this paper — H&N, thorax, abdomen, and pelvis — are defined by their superior and inferior axial boundaries in the craniocaudal direction. Our goal is to classify each slice of the input image into one of eight categories — the eight region boundary slices plus another category which constitutes nonboundary slices. The process is divided into two stages, training and testing, each of which is further divided into three steps. In both the training and testing stages, the first step is the preprocessing step where each slice in a stack of axial slices is rescaled to 224 × 224 pixels and is combined with its neighboring slices to create compound five-channel images (Section 2.C). This provides additional appearance context to the model. The training stage has two more steps: training a deep CNN (Section 2.D) and training a recurrent neural network (RNN) (Section 2.H). The two networks work in a cascading manner with the RNN improving upon the predictions of the CNN for the body region boundaries which generally have high error rates. The RNN is trained on sequences of features extracted from an intermediate layer of the CNN, with the sequences defined by windows centered around the predictions from the CNN. Both the preprocessing step creating the five-channel images and the RNN exploit the inherently sequential nature of the axial stack of slices. In the testing stage, the first step is the same preprocessing step as in the training stage, and the other two steps are: drawing inference from the CNN, and improving its performance for the classes known to have high error rates by using the RNN. Our experiments involve 442 low-dose CT data sets from whole-body PET/CT scans and an additional 213 diagnostic CT scans of the H&N region, as described in Section 3. Our concluding remarks are summarized in Section 4.

2. MATERIALS AND METHODS

2.A. Data sets and notations

For this study, we use PET/CT scans from 442 patients obtained from the database of the Hospital of the University of Pennsylvania. Approval for data usage was obtained from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. Subjects include near-normal cases and patients with different types of disease conditions where all scans were obtained for clinical reasons only. Of the 442 scans, 262 were from head to pelvis, 39 from head to toe, and 17 from neck to toe. The mean voxel size for the low-dose CT images was $1.13 \times 1.13 \times 3.77 \text{mm}^3$ and the slice spacing varied from 2 to 5 mm: two images with slice spacing of 2 mm, 129 images with slice spacing of 3 mm, 1 image with slice spacing of 3.26 mm, 7 images with slice spacing of 3.27 mm, 266 images with slice spacing of 4 mm, 1 image with slice spacing of 4.25 mm, and 36 images with slice spacing of 5 mm. Apart from these 442 scans, diagnostic CT scans of the head and neck region were obtained for 213 patients for additional testing in a different scenario. The mean voxel size for these images was $1.10 \times 1.10 \times 2.05 \text{ mm}^3$ and the slice spacing varied from 1 to 3 mm: one image with slice spacing of 1 mm, 27 images with slice spacing of 1.5 mm, 158 images with slice spacing of 2 mm, 3 images with slice spacing of 2.5 mm, and 24 images with slice spacing of 3 mm.

For an image I in the data set, we denote the true superior and inferior boundaries of the neck, thorax, abdomen, and pelvis, respectively, as $NS(I)$, $NI(I)$, $TS(I)$, $TI(I)$, $AS(I)$, $AI(I)$, $PS(I)$, and $PI(I)$. For each volumetric image, these boundary locations were labeled manually under the guidance of a radiologist (Torigian) following our strict definition of the four body regions.[2,3]

2.B. Definition of body regions

For this study, we use the body region definitions described in our previous work.[2,3] We focus on four body regions: neck, thorax, abdomen, and pelvis, defining each body region by two boundary slices: the slice representing the superior boundary and the slice representing the inferior boundary. The definitions of these body region boundaries are tabulated in Table I, and the distinguishing features for each body region boundary are illustrated in Fig. 1. Note that, per our definitions, $AI(I) = PS(I)$.

The important point to note here is that some defining features have a high contrast with respect to their surrounding voxels (e.g., the apex of the lung for $TS(I)$ appears as two, or in some cases one, dark circular objects in the axial slice, with the slice immediately superior to this slice not having the dark circular objects, and the slice immediately inferior having two slightly larger dark circular objects), while some defining features are very subtle and exhibit very low contrast with respect to their surrounding voxels (e.g., the bifurcation of the abdominal aorta into the common iliac arteries as the distinguishing feature for $AI(I)$). This is an important observation which serves as motivation for the network design involving a two-step prediction process to improve only some of the predictions from the first step in the second step of our approach.

It is also important to note that there is often some degree of digital ambiguity present in the process of manual labeling of the ground truth for this study, arising largely due to the discrete nature of the axial stack, unlike the continuous nature of the human anatomy. For example, in the case of $AI(I)$, the slice where abdominal aorta bifurcates into the common iliac arteries may be marked differently by different experts, albeit within one to two slices of each other. This observation is important to keep in mind when analyzing and interpreting the results of this study.

2.C. Preprocessing

Each volumetric image is separated into its constituent slices. Each slice is resized from 512 × 512 to 224 × 224

TABLE I. Definition of body regions and their boundary locations.

Body region	Boundaries	Description	Definition
Neck	NS	Neck superior axial boundary location	Superior-most aspect of the mandible
	NI	Neck inferior axial boundary location	Level of bifurcation of the superior vena cava into left and right brachiocephalic veins
Thorax	TS	Thoracic superior axial boundary location	15 mm superior to the apex of the lungs
	TI	Thoracic inferior axial boundary location	5 mm inferior to the base of the lungs
Abdomen	AS	Abdominal superior axial boundary location	Superior-most aspect of the liver
	AI	Abdominal inferior axial boundary location	Level of bifurcation of the abdominal aorta into common iliac arteries
Pelvis	PS	Pelvic superior axial boundary location	Inferior boundary of the abdominal region
	PI	Pelvic inferior axial boundary location	Inferior-most aspect of the ischial tuberosities

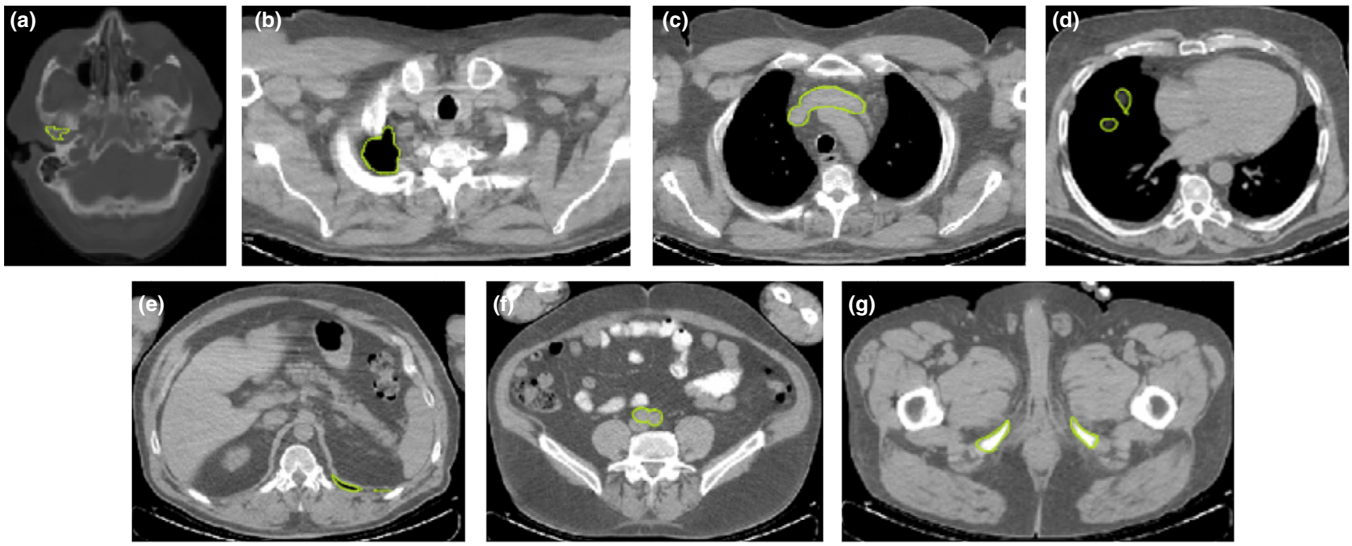

FIG. 1. Distinguishing features (marked in green) for each body region boundary. Note that not all of these slices correspond to boundaries. The slices shown depict: (a) The superior-most aspect of the mandible; (b) The apex of the lung; the TS(I) slice exists 15 mm above this slice; (c) The level of bifurcation of the superior vena cava into left and right brachiocephalic veins; (d) The superior-most aspect of the liver; (e) The base of the lungs; the TI(I) slice is located 5 mm below this slice; (f) The level of bifurcation of the abdominal aorta into common iliac arteries; (g) The inferior-most aspect of the ischial tuberosities of the pelvis.

pixels. For each slice, a five-channel compound image is created with the slice in question as the third channel. The first and second channels are the two slices immediately inferior to the slice in question, and the fourth and fifth channels are the two slices immediately superior to the slice in question. This is done to provide BRR-Net with additional visual context which is crucial to the classification task. This aids in accurately classifying body region boundaries by exploiting the contrast among the channels due to the nature of the definition of the boundaries. For example, for locating the apex of the lung, the axial slice immediately superior to the slice with the apex of the lung will not contain any dark circular objects (lungs) while the axial slice immediately inferior to the slice with the apex of the lung will contain slightly larger dark circular objects (lungs); in the case of locating the slice with the inferior-most aspect of the ischial tuberosities of the pelvis, the slice immediately inferior to this slice will not contain any bright bony objects of the pelvis while the slice immediately superior will have a larger representation of the bright ischial tuberosities than the slice with the inferior-most aspect.

Each five-channel image is then zero-center normalized by subtracting the entire data set's mean image from the five-channel images. Thus, each slice is now represented by a five-channel image. For cases where two slices are not available either before or after the slice, zero padding is used.

2.D. Convolutional neural networks

Convolutional neural networks[16] employ convolution operations on the input with weights and introduce nonlinearity. They are suited for processing spatial information through the spatial arrangement of convolution operations, local connectivity, parameter sharing, and pooling operations. CNNs have shown great promise in processing spatial data, especially in tasks like image classification and image segmentation.

For the task of classification (the goal of our application), many network architectures have been proposed and tested in literature. A few most prominent ones include VGG-16,[17] AlexNet,[16] and GoogLeNet.[18] In this paper, we use the GoogLeNet architecture owing to its efficiency, both in terms of number of operations needed for a single inference and the space required for the storage of the parameters.[19] The GoogLeNet architecture is modified to accept five-channel images as input and classify the input image into one of the nine categories — seven categories corresponding to the seven body region boundaries (recall that $AI(I) = PS(I)$), an eighth category called *Legs* which comprises of all the slices below $PI(I)$, and a ninth category denoting the remaining nonboundary locations. The GoogLeNet architecture is shown in Fig. 2, where each "Inception Block" is a network of layers shown in Fig. 3.

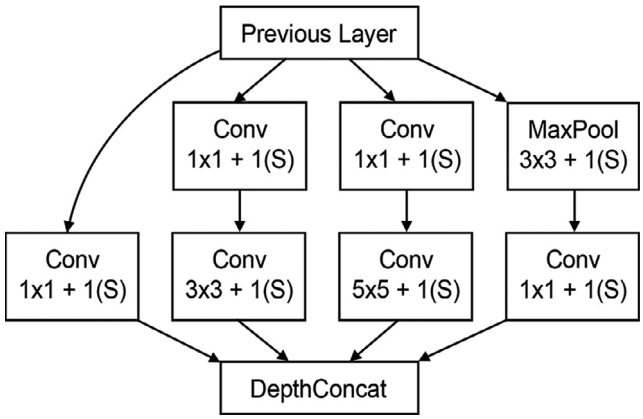

FIG. 3. An inception block with dimension reductions.[18]

2.E. Training the convolutional neural network

The data set containing 442 volumetric images is partitioned into the training + validation data set (containing 70% of the images) and the testing data set (containing 30% of the images). As mentioned earlier, for each volumetric image, all slices below the $PI(I)$ slice are labeled as *Legs*; all other slices that do not belong to any of the seven body region boundary classes or the Legs class are termed *none-of-the-above (NOTA)*. The Legs class is defined in order to reduce the variance in the NOTA class. From the training + validation data set, the seven body region boundary images are over-sampled to twice the number of samples in each class, and the Legs and NOTA classes are heavily under-sampled to create a relatively balanced data set. As our data ensemble is relatively small and over-sampling is employed, data augmentation was performed on the training + validation data set. Each image is:

a Randomly horizontally flipped with a probability of 50%.
b Rotated by x degrees where x is a random number between −15 and 15.
c Translated along the horizontal axis by y pixels where y is a random integer between −15 and 15.
d Translated along the vertical axis by z pixels where z is a random integer between −15 and 15.

The training data set was resampled every 10 epochs to have a healthy representation of the Legs and NOTA classes, and to reaugment the data set for more effective training.

2.F. Improving training for the convolutional neural network

During inference for any volumetric image I, multiple consecutive slices in I may have high probabilities for any body region boundary because of their similarity of appearance. These are the slices that are in the vicinity, both superior and inferior, of the ground truth slice for the body region boundary under consideration. In order to (a) localize these high-probability predictions and (b) take into account marginal labeling errors (to the order of one slice), the trained network is fine-tuned using a subset of the entire training data set. This subset contains only the slices which are in the vicinity of the ground truth slice. For each body region boundary label, five images on either side, inferior and superior, are considered, resulting in 11 images per body region boundary label. Of these ten neighboring images (each image being a five-channel compound image, consisting of five axial slices), the two images immediately adjacent to the ground truth slice, one superior and one inferior, are also marked as the ground truth for that label, effectively extending the ground truth for each body region boundary label to three slices for this stage of refined training.

2.G. Recurrent neural networks (RNNs)

The performance of the above CNN model is not uniformly accurate for all boundary slices. The goal for designing an RNN is to improve the predictions for the classes which did not perform well, that is, for those classes which had relatively large errors in the predictions from the CNN.

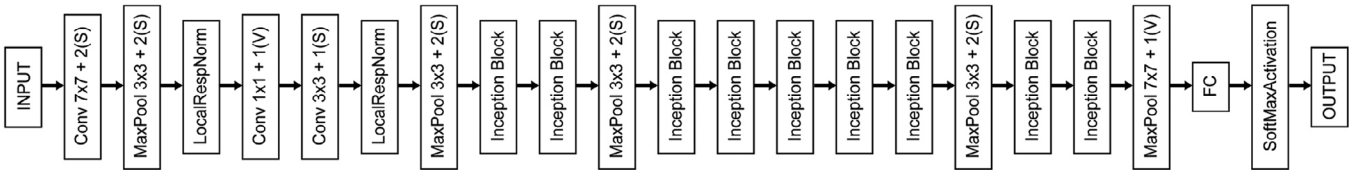

FIG. 2. The GoogLeNet architecture (Szegedy et al., 2015).

In order to exploit the inherently sequential nature of a stack of axial slices and the contiguity of appearance, RNNs were employed on the features extracted from the CNN. Recurrent neural networks are types of artificial neural networks which are used to process sequential information. These types of networks have connections between their nodes to form a sequence of connected neural networks. RNNs have proved to work exceedingly well in processing temporal sequences like audio and video data.[20] However, in practice, traditional RNNs are often difficult to train because during training, the gradients tend to either vanish or, in rare cases, explode, making gradient-based optimization difficult. Long short-term memory (LSTM)[21] units tackle this problem by using a modification of the simple RNN architecture in which the hidden state is allowed to be updated, be reset, or be propagated without modification using learned gates.[22]

If an axial stack of slices is considered analogous to a video, and each individual slice in the axial stack is identified with a frame in a video, an RNN seems like a natural choice of architecture for its processing. We use an architecture comprised of bi-directional LSTM cells, drop-out layers, and fully connected layers for our task. Thus, we implement a two-step process to solve the problem. The CNN model gives the predictions for each body region boundary, and we use the RNN model to improve upon the predictions that are not very accurate.

The input for the RNN is the 1024-dimensional feature vector from the last max pooling layer of the GoogLeNet architecture (Figs. 2, 3). The RNN model is trained separately for each of the classes that did not perform well. The network architecture is shown in Fig. 4.

2.H. Data and labels for RNN

For each class that did not perform well, a window (contiguous sequence) of 31 slices is defined centered around the ground truth slice in each axial stack. In each window, all slices before the ground truth label are labeled 0 and all other slices are labeled 1. Then, to simulate the error produced during inference by the CNN, each window is offset by a random number generated by a normal distribution with the mean and standard deviation equal to the mean and standard deviation, respectively, of the errors for each of the underperforming classes from the CNN predictions on the validation set. For each slice in this window, a 1024-dimensional feature representation is extracted from the last max-pooling layer of the CNN. At the end, for each axial stack (image I), we have a 31 x 1024-dimensional vector which is treated as a sequence of length 31, together with a 31-dimensional binary-valued vector which holds the labels, either 0 or 1, corresponding to each feature vector in the window. Such sequences are generated for each image I in the data set, and the set of sequences along with their binary-valued label vectors are used to train the RNN.

At inference time, a sequence of 31 1024-dimensional vectors corresponding to a 31-slice window centered around the slice whose classification is predicted by the CNN is given as input to the RNN. A corresponding sequence of length 31 consisting of 0s and 1s is predicted by the RNN, and the element after the one where the sequence changes from 0 to 1 is marked as the predicted slice location. In case there are <15 slices on either side (inferior or superior) of the slice being predicted by the CNN, the window is truncated to include all the slices from that side, and a sequence of <31 slices (thereby making the window size *flexible*) is used as input, and the corresponding predicted sequence is processed in the same manner as described before.

2.I. The final BRR-Net architecture

We call the tandem CNN–RNN architecture formed as described above BRR-Net. The complete architecture is displayed in Fig. 5.

At the time of inference, each slice in the given axial stack I is scaled to 224 × 224 pixels and is converted into a compound five-channel image as described in Section 2.C. Each slice is then sequentially fed into the CNN as the input image, and the probability of that image belonging to one of the body region boundary classes is obtained. The slice that is taken as the predicted slice at the CNN stage should satisfy two conditions: It should have the highest probability for the predicted class among all classes and it should be in the right sequential order of the slice numbers with respect to other body region boundaries. The latter check is not implemented for all body region boundary classes, and is used only for those classes that pose a risk of being wrongly predicted by a large number of slices (enough to erroneously cross other body region boundary locations in either direction) due to their defining features

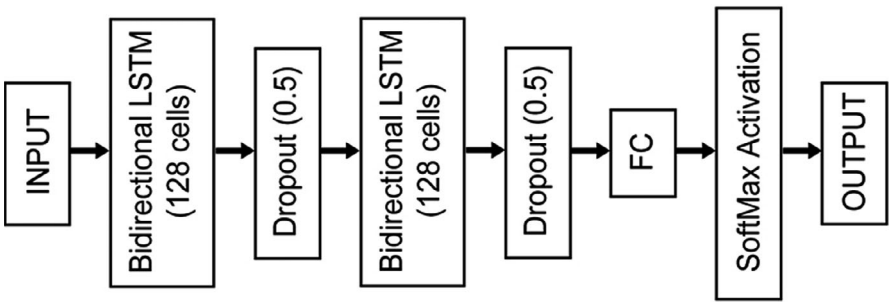

FIG. 4. Network architecture of the recurrent neural network used.

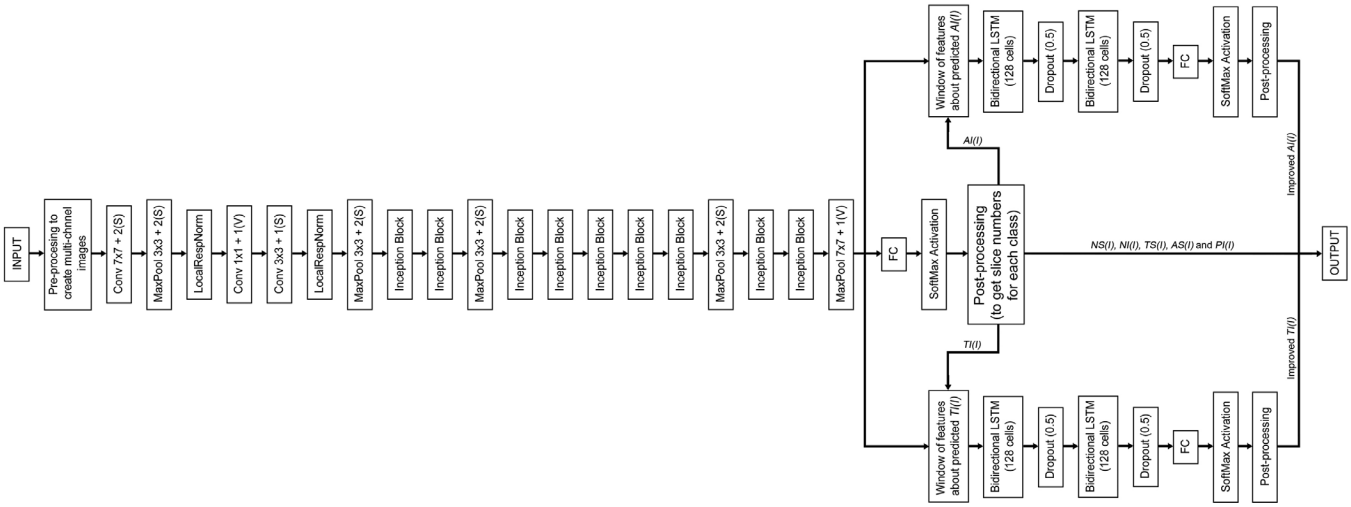

FIG. 5. Architecture of BRR-Net.

resembling anatomical structures elsewhere in the body. To clarify, consider an example: The defining feature for *AS(I)* — the superior-most aspect of the liver — sometimes resembles the superior-most aspect of the skull. This is because while moving in the caudo-cranial direction, the superior-most aspects of both the liver and the skull appear to be high-density masses on low-density backgrounds, with the high-density masses in both cases exhibiting a convex circular shape. Infrequently, this causes the slice containing the superior-most aspect of the skull to be predicted as the slice with the highest probability for the *AS(I)* class. In order to mitigate this tendency, the prediction for *AS(I)* is always compared against the predictions for *NS(I)* and/or *NI(I)*, neither of which demonstrate a tendency to be wrongly predicted by a large margin (i.e., even when they are inaccurate, they are always in close vicinity of the actual *NS(I)* and *NI(I)* slices, respectively. Please refer Tables IX and X). If the predicted slice number for *AS*(*I*) lies superior to the predicted slice of either of the body region boundaries for neck, then that prediction is ignored and the slice with the second highest probability is marked as the prediction, and so on.

As mentioned in the example, only those classes that are predicted with high accuracy and which do not demonstrate any tendency for being outliers are used as "anchors" for checking for correctness of anatomical order. In our study, only *AS(I)* was found to require such a check, but this methodology may be extended for any class that exhibits such properties, provided that the classes used for checking for the correctness of anatomical order ("anchors") are known to not be outlier-prone.

For classes that are known to underperform with the CNN, the predicted slice is fed into the RNN by creating sequence of feature vectors about that slice, each vector having been obtained from an intermediate layer of the CNN as described in Section 2.H, and the refined prediction is obtained. Thus, the final output of BRR-Net for a given axial stack of slices is obtained as the slice numbers corresponding to each of the classes representing the seven body region boundaries.

3. EXPERIMENTS, RESULTS, AND DISCUSSION

3.A. Computational considerations

The training and testing of the two components of BRR-Net were performed as follows. The computing platform had an i7-5930K CPU clocked at 3.50 GHz, 16 GB of RAM, and a NVIDIA Titan Xp GPU with 12 GB of memory. We used MATLAB for training the CNN and python + keras with TensorFlow backend for training the RNN. The MATLAB model was converted to a TensorFlow-compatible model using ONNX. One epoch of training the CNN took about 150 s and one epoch of training the RNN took about 3 s (excluding the feature extraction from CNN). The learning rate was adjusted using learning rate schedulers. For the CNN training, the learning rate was reduced by a factor of 10 every 10 epochs. The batch size of 32 was kept constant while training. While training the RNNs, the learning rate was reduced by a factor of 10 every 20 epochs, and the batch size used was 256 and was kept constant throughout the training. Prediction time for a single study varies from 5 to 10 s depending on the number of axial slices in the image.

3.B. Ground truth and its precision

Before examining the results presented henceforth, it is important to examine the variations in human performance in the task of labeling the body region boundaries. In a previous work[1] (Bai et al., 2019), the *TS*(*I*), *TI*(*I*), *AS*(*I*), *AI*(*I*), and *PI*(*I*) classes were labeled for 180 volumetric images by an expert. For this study, the images were labeled again, by a different expert. The mean and standard deviations for the absolute difference in the labels by the two experts are tabulated in Table II. It is clear that the variation in slice location for the class $AI(I) = PS(I)$ is much larger than that for other classes and is quite considerable with a mean of ~4 slices. This is largely due to the ambiguous nature of the defining feature of the *AI*(*I*) class. For all other classes, the mean variation is less than one slice.

3.C. Prediction results

The results from the initial CNN are tabulated in Tables III–V. In Table III, each cell represents the percentage of total number of cases that were within (≤) the absolute error mentioned in the first column which is expressed in terms of number of slices. Table IV shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in terms of number of slices and mm. Table V shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in mm for images with slice spacing equal to 3, 4, and 5 mm separately.

The results from the fine-tuned CNN (henceforth referred to as just "CNN") are tabulated in Tables VI–VIII. In Table VI, each cell represents the percentage of total number of cases that were within (≤) the absolute error mentioned in the first column which is expressed in terms of number of slices. Table VII shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in terms of number of slices and mm. Table VIII shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in mm for images with slice spacing equal to 3, 4, and 5 mm separately.

As is evident, the model performs exceptionally well except for classes *AI*(*I*) and *TI*(*I*). This is predominantly due to the nature of the defining features of these two classes. For the case of *AI*(*I*) the defining feature, the level of bifurcation of the superior vena cava into the left and right brachiocephalic veins, is sometimes difficult to locate accurately. For the case of *TI*(*I*), the defining feature, the base of the lungs, often appears as a mere sliver of black (low image intensity), making it difficult to observe in some cases.

As explained in Section 2.G, the RNN model is employed only for those classes which have relatively higher errors. Classes *AI*(*I*) and *TI*(*I*) are empirically chosen on the basis of Tables VI and VII to employ the RNN. The RNN was trained on these classes as described in Section 2.H, and the results from the CNN, followed by the RNN for them are tabulated in Tables IX–XI. In Table IX, each cell represents the percentage of total number of cases that were within the absolute error mentioned in the first column. Table X shows mean and standard deviation of the prediction errors with respect to the ground truth expressed in terms of number of slices and mm. Notable improvements can be seen for both *AI*(*I*) and *TI*(*I*) classes, which have benefitted from the sequential processing of their neighboring slices. Table XI shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in mm for images with slice spacing equal to 3, 4, and 5 mm separately.

TABLE II. Mean and standard deviation (SD) of absolute difference expressed in number of slices (first row) and mm (second row) in the labels assigned by two different experts.

	TS	TI	AS	AI = PS	PI
Number of slices	0.1 ± 0.7	0.4 ± 1.0	0.1 ± 0.4	4.1 ± 5.3	0.7 ± 0.6
mm	0.6 ± 2.6	1.3 ± 3.5	0.6 ± 1.7	15.5 ± 20.8	2.8 ± 2.4

TABLE III. Results after training the modified GoogLeNet on all classes.

Abs. Err.	NS	NI	TS	TI	AS	AI = PS	PI
1	81.37	61.90	97.74	49.62	93.23	30.08	100
2	96.08	83.81	98.50	65.41	94.74	46.62	100
3	100	92.38	99.25	75.19	96.99	59.40	100
4	100	95.24	100	84.96	97.74	68.42	100
5	100	98.10	100	87.97	97.74	77.44	100

TABLE VI. Results after training the convolutional neural network on all classes.

Abs. Err.	NS	NI	TS	TI	AS	AI = PS	PI
1	95.10	80.95	100	53.38	96.24	30.83	100
2	100	93.33	100	78.95	97.74	48.12	100
3	100	95.24	100	84.96	98.50	61.65	100
4	100	97.14	100	91.73	98.50	74.44	100
5	100	99.05	100	94.74	98.50	80.45	100

TABLE IV. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via the modified GoogLeNet, expressed in number of slices (first row) and mm (second row).

	NS	NI	TS	TI	AS	AI = PS	PI
Number of slices	0.9 ± 0.8	1.5 ± 1.4	0.4 ± 0.6	2.4 ± 2.5	0.6 ± 2.6	3.8 ± 3.5	0.4 ± 0.5
mm	3.2 ± 3.1	5.6 ± 4.9	1.5 ± 2.3	8.4 ± 8.0	2.2 ± 8.1	13.6 ± 11.7	1.7 ± 1.9

TABLE V. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via the modified GoogLeNet, expressed in mm, for images with different slice spacings.

Slice spacing	NS	NI	TS	TI	AS	AI = PS	PI
3 mm	2.9 ± 2.3	5.5 ± 4.8	1.4 ± 2.3	9.0 ± 8.5	3.9 ± 13.9	17.1 ± 13.7	1.5 ± 1.5
4 mm	3.3 ± 3.2	5.8 ± 5.0	1.5 ± 2.3	8.4 ± 8.0	1.3 ± 3.6	12.2 ± 10.8	1.6 ± 2.0
5 mm	5.0 ± 7.1	2.5 ± 2.9	2.5 ± 2.9	5.0 ± 4.1	7.5 ± 9.6	13.8 ± 9.5	3.8 ± 2.5

TABLE VII. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via CNN, expressed in number of slices (first row) and mm (second row).

	NS	NI	TS	TI	AS	AI = PS	PI
Number of slices	0.6 ± 0.6	1.1 ± 1.1	0.3 ± 0.5	1.8 ± 2.0	0.6 ± 2.4	3.4 ± 3.2	0.5 ± 0.5
mm	2.2 ± 2.1	3.8 ± 3.9	1.2 ± 1.8	6.8 ± 6.8	2.0 ± 7.5	12.7 ± 11.7	1.9 ± 1.9

TABLE VIII. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via CNN, expressed in mm, for images with different slice spacings.

Slice spacing	NS	NI	TS	TI	AS	AI = PS	PI
3 mm	1.9 ± 1.9	4.3 ± 4.6	0.8 ± 1.4	7.6 ± 8.1	3.5 ± 13.6	12.1 ± 10.4	2.1 ± 1.7
4 mm	2.4 ± 2.2	3.5 ± 3.2	1.3 ± 1.9	6.4 ± 6.3	1.2 ± 1.9	12.8 ± 12.3	1.8 ± 2.0
5 mm	2.5 ± 2.9	3.8 ± 7.5	1.3 ± 2.5	7.5 ± 2.9	7.5 ± 6.5	16.3 ± 9.5	2.5 ± 2.9

TABLE IX. Results after training the convolutional neural network on all classes followed by RNN (BRR-Net) for $AI(I)$ and $TI(I)$.

Abs. Err.	NS	NI	TS	TI	AS	AI = PS	PI
1	95.10	80.95	100	70.68	96.24	35.34	100
2	100	93.33	100	86.47	97.74	57.14	100
3	100	95.24	100	92.48	98.50	69.92	100
4	100	97.14	100	93.98	98.50	81.20	100
5	100	99.05	100	96.99	98.50	89.47	100

Of course, the accuracy of this algorithm depends entirely on how precisely the ground truth has been labeled, but the results in Table X establish BRR-Net's ability to closely mimic an expert's labels. An important point to note here, perhaps, would be that during the labeling of the data, the expert must maintain consistency in their methodology for labeling ambiguous cases and that visual ambiguities must be handled in the same manner for all subjects in the data set. This ensures homogeneity in the labeled slices, which is critical to achieving high accuracy in our classification task.

3.D. Display examples

Samples of results, both good (within one slice of the ground truth) and not-as-good (more than one slice from the ground truth and labeled as "poor") are shown in Fig. 6. "Poor" examples for classes $TS(I)$ and $PI(I)$ were not available as the error was within one slice for all test cases for these classes.

Even though our models were trained on low-dose CT images, tests were carried out to see if they can be used on diagnostic CT images as well in an "as-is" condition. As described in Section 2.A, 213 diagnostic CT images of the neck region with an average voxel size of $1.10 \times 1.10 \times 2.05$ mm^3 were obtained and labeled for three classes: $NS(I)$, $NI(I)$, and $TS(I)$. The results from the final two-step BRR-Net model are tabulated in Tables XII–XIV. Table XIV shows the mean and standard deviation of the prediction errors with respect to the ground truth expressed in mm for images with slice spacing equal to 1.5, 2.5, and 3 mm separately. Samples of results for diagnostic CT images are shown in Fig. 7.

TABLE X. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via BRR-Net, expressed in number of slices (first row) and mm (second row).

	NS	NI	TS	TI	AS	AI = PS	PI
Number of slices	0.6 ± 0.6	1.1 ± 1.1	0.3 ± 0.5	1.4 ± 1.7	0.6 ± 2.4	2.8 ± 2.7	0.5 ± 0.5
mm	2.2 ± 2.1	3.8 ± 3.9	1.2 ± 1.8	5.4 ± 5.6	2.0 ± 7.5	10.8 ± 9.9	1.9 ± 1.9

TABLE XI. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes via BRR-Net, expressed in mm, for images with different slice spacings.

Slice spacing	NS	NI	TS	TI	AS	AI = PS	PI
3 mm	1.9 ± 1.9	4.3 ± 4.6	0.8 ± 1.4	6.5 ± 7.6	3.5 ± 13.6	12.0 ± 9.8	2.1 ± 1.7
4 mm	2.4 ± 2.2	3.5 ± 3.2	1.3 ± 1.9	4.9 ± 4.6	1.2 ± 1.9	10.3 ± 10.1	1.8 ± 2.0
5 mm	2.5 ± 2.9	3.8 ± 7.5	1.3 ± 2.5	6.3 ± 2.5	7.5 ± 6.5	8.8 ± 6.3	2.5 ± 2.9

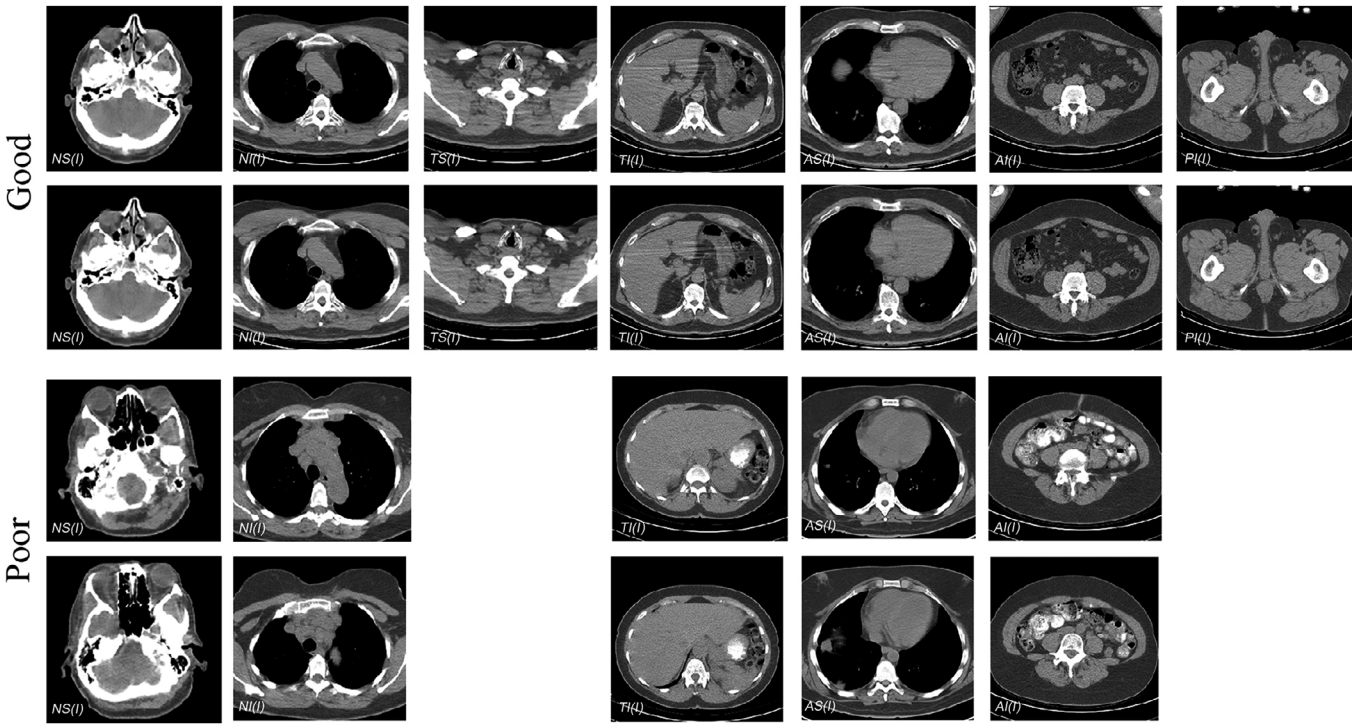

FIG. 6. Display examples for low-dose computed tomography images. Sample true (Row 1) and predicted (Row 2) slices for a "good" case for each class (error 1 slice) and a "poor" case (Rows 3 — true and 4 — predicted) for each class (error > 1 slice).

TABLE XII. Results from the final model on diagnostic computed tomography images.

Abs. Err.	NS	NI	TS
1	60.56	35.89	93.43
2	83.10	57.42	97.65
3	92.96	73.68	99.53
4	96.24	81.82	100
5	99.06	89.95	100
6	99.53	93.78	100
7	99.53	97.61	100
8	99.53	99.04	100
9	100	99.52	100
10	100	100	100

TABLE XIII. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes *TS*(*I*), *NS*(*I*), and *NI*(*I*) for diagnostic CT images via BRR-Net, expressed in number of slices (first row) and mm (second row).

	NS	NI	TS
Number of slices	1.5 ± 1.3	2.6 ± 2.1	0.6 ± 0.7
mm	2.9 ± 2.5	5.1 ± 3.9	1.2 ± 1.4

One interesting point to note here is that the performance on diagnostic CT seems worse in terms of error expressed in number of slices than that on the low-dose CT images. However, the mean interslice separation in the diagnostic CT images is almost half of that of low-dose CT images, meaning that even though the error is higher in number of absolute slices, it is quite comparable to the performance on low-dose CT images in absolute distance from the ground truth. This is quite remarkable considering the fact that we did not retrain BRR-Net on diagnostic CT images.

TABLE XIV. Mean and standard deviation (SD) of the error from the ground truth in the prediction of the classes *TS*(*I*), *NS*(*I*), and *NI*(*I*) for diagnostic computed tomography images via BRR-Net, expressed in mm, for images with different slice spacings.

Slice spacing	NS	NI	TS
1.5 mm	2.6 ± 2.9	5.8 ± 4.0	1.5 ± 1.4
2 mm	3.1 ± 2.4	5.0 ± 3.9	1.0 ± 1.2
3 mm	1.9 ± 1.7	4.6 ± 3.8	1.6 ± 2.2

3.E. Comparison with methods from the literature

As mentioned in Section 1.B, the only other works that tackled the problem addressed in this paper in a much-limited sense are Bai et al.[1] and Hussein et al.[4] Bai et al. focused on localizing *TS*(*I*), *TI*(*I*), *AS*(*I*), *AI*(*I*) = *PS*(*I*), and *PI*(*I*) in low-dose CT of PET/CT acquisitions as well. Using a different concept of virtual landmarks and neural networks for prediction, they report errors (mean ± SD, in number of slices) of, respectively, 2.7 ± 1.8, 3.0 ± 3.0, 3.7 ± 1.9, 3.9 ± 3.2, and 2.5 ± 1.8 for these five body region boundaries. BRR-Net results (Table X) are better for all classes with statistical

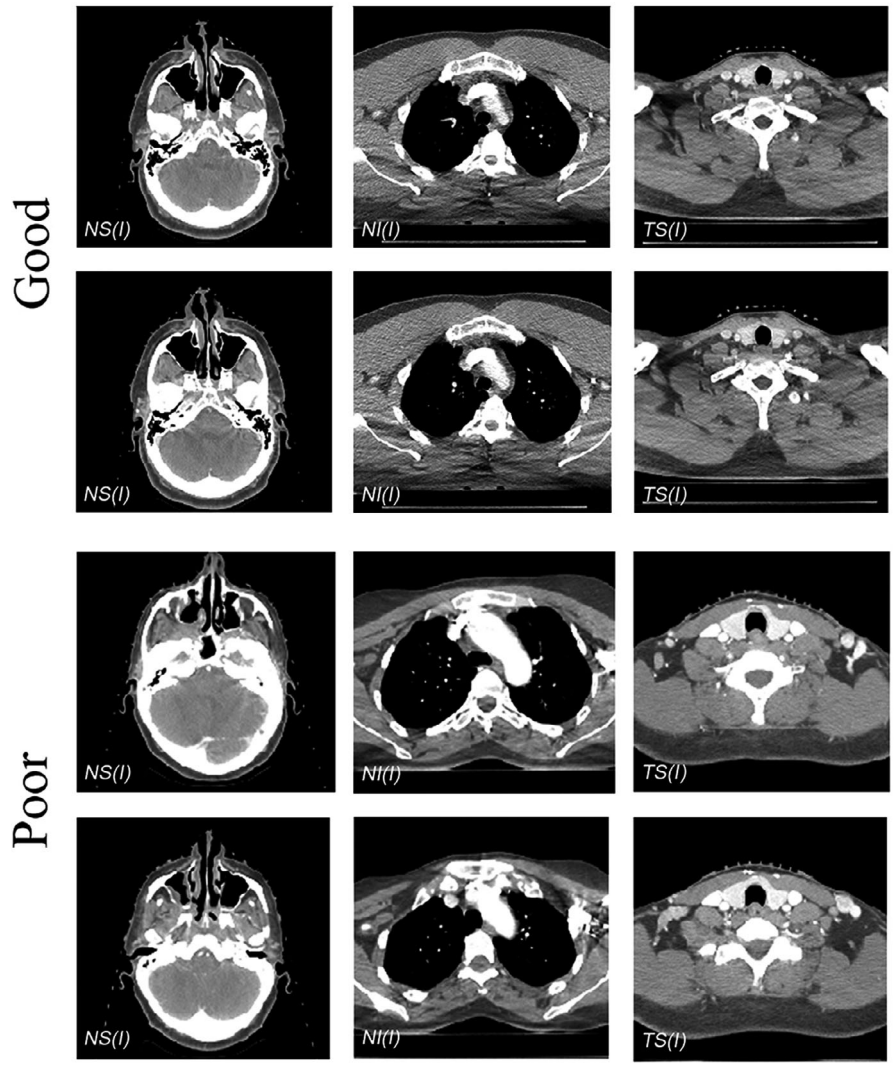

FIG. 7. Display examples for diagnostic CT images. Sample true (Row 1) and predicted (Row 2) slices for a "good" case for each class (error 1 slice) and a "poor" case (Rows 3 — true and 4 — predicted) for each class (error > 1 slice).

significance ($P < 0.05$) and by 2.4, 1.7, 3.1, 1.1, and 2 slices, respectively. Notably, BRR-Net achieves tighter predictions with considerably lower standard deviation (except for *AS* where it is comparable), suggesting that it is overall more robust. Hussein et al. report an error of ~47 mm, which is considerably higher than our error, in localizing the thoracic region boundaries following the same body region definitions as we have used (Table I). It is to be noted that (a) their study's scope, data set, and application were different from ours and (b) the nature of their algorithm required only a rough initial estimation and precise localization was not necessary in their study.

4. CONCLUDING REMARKS

Automatically localizing body regions in medical images by locating their superior and inferior axial boundaries is an important step toward the acceptance and subsequent application of standardized body region definitions. It is also increasingly important in developing applications based on this ideology, especially the systems designed for anatomical regions which depend on the precise boundaries. In this work, we have described a novel technique to automatically locate the axial body region boundaries of four body regions — neck, thorax, abdomen, and pelvis — within one slice for a majority of the locations, and within at most three slices overall. If additional body regions are defined or if existing body region definitions are modified in the future, this model can be retrained to learn the representations of their defining features. The models trained on low-dose CT images also work remarkably well on diagnostic CT images. Due to the nature of the algorithm, this approach can be generalized to other imaging modalities, as well, such as MRI, although in this work we have demonstrated its performance only on low-dose CT and diagnostic CT images.

In a production-mode set up, a BRR-Net-type system becomes indispensable for the reasons we explained in Section 1.B with the application example of RT planning. The AAR methodology fully exploits the facility of precisely defining body regions (and the objects contained in them) to

reduce population variations in the geographic layout of objects and thus to improve the accuracy of object delineation, and quantification. A question that remains for AAR-type methods is what level of accuracy is required or error is tolerable in locating body region boundaries by methods such as BRR-Net. Given that expert localization has its own variability, will computerized localization be more accurate than manual methods for AAR and its applications? We are investigating such questions in the context of using BRR-Net as the front end of such applications.

CONFLICT OF INTEREST

Drs. Udupa and Torigian are cofounders of Quantitative Radiology Solutions, LLC, whose goal is to develop automated software solutions for quantitative analysis in radiological practice.

[a)]Author to whom correspondence should be addressed. Electronic mail: jay@pennmedicine.upenn.edu.

REFERENCES

1. Bai P, Udupa JK, Tong Y, Xie S, Torigian DA. Body region localization in whole-body low-dose CT images of PET/CT scans using virtual landmarks. *Med Phys*. 2019;46:1286–1299.
2. Udupa JK, Odhner D, Zhao L, et al. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. *Med Image Anal*. 2014;18:752–771.
3. Wang H, Udupa JK, Odhner D, Tong Y, Zhao L, Torigian DA. Automatic anatomy recognition in whole-body PET/CT images. *Med Phys*. 2016;43:613–629.
4. Hussein S, Green A, Watane A, et al. Automatic segmentation and quantification of white and brown adipose tissues from PET/CT scans. *IEEE Trans Med Imaging*. 2017;36:734–744.
5. Bai P, Udupa JK, Tong Y, Xie S, Torigian DA. Automatic thoracic body region localization.In: Medical Imaging 2017: Computer-Aided Diagnosis. 2017;10134:101343X. https://doi.org/10.1117/12.2254862
6. Criminisi A, Robertson D, Konukoglu E, et al. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Med Image Anal*. 2013;17:1293–1303.
7. Lee CC, Chung PC. Recognizing abdominal organs in CT images using contextual neural network and fuzzy rules. In: *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2000;3:1745–1748. https://doi.org/10.1109/IEMBS.2000.900421
8. Wang Y, Qiu Y, Thai T, Moore K, Liu H, Zheng B. A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images. *Comput Methods Progr Biomed*. 2017;144:97–104.
9. de Vos BD, Wolterink JM, de Jong PA, Leiner T, Viergever MA, Išgum I. ConvNet-based localization of anatomical structures in 3-D medical images. *IEEE Trans Med Imaging*. 2017;36:1470–1481.
10. Wu X, Udupa JK, Tong Y, et al. AAR-RT - a system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med Image Anal*. 2019;54:45–62.
11. Liu T, Udupa JK, Miao Q, Torigian DA. Quantification of body-torso-wide tissue composition on low-dose CT images via automatic anatomy recognition. *Med Phys*. 2019;46:1272–1285.
12. Tong Y, Udupa JK, Sin S, Liu Z, Wileyto PE, Torigian DA, Arens R. MR image analytics to characterize upper airway structure in obese children with obstructive sleep apnea syndrome. *PLOS ONE*. 2016;11:e0159327.
13. Anderson MR, Udupa JK, Edwin EA, et al. Adipose tissue quantification and primary graft dysfunction after lung transplantation: the lung transplant body composition study. *J Heart Lung Transplant*. 2019;38:1246–1256.
14. Shashaty MGS, Kalkan E, Bellamy SL, et al. Computed tomography-defined abdominal adiposity is associated with acute kidney injury in critically ill trauma patients. *Crit Care Med*. 2014;42:1619–1628.
15. Tong Y, Udupa JK, Odhner D, Wu C, Schuster SJ, Torigian DA. Disease quantification in PET/CT images without explicit object delineation. *Med Image Anal*. 2019;51:169–183.
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*. 2017;60:84–90. https://doi.org/10.1145/3065386
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014; [published online September 4; 2014. arXiv preprint arXiv:1409.1556.
18. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015:1–9. https://doi.org/10.1109/CVPR.2015.7298594
19. Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. 2016; [published online May 24. 2016. arXiv preprint arXiv:1605.07678.
20. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning.2015; [published online May 29; 2015. arXiv preprint arXiv:1506.00019.
21. Hochreiter S, Schmidhuber J. Long short-term memory. *Neur Comput*. 1997;9:1735–1780.
22. Donahue J, Hendricks LA, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans Pattern Anal Mach Intell*. 2015;39:2625–2634.