Feature Pyramid Network for Liver Segmentation from Whole Body CT Images of Nonhuman Primates with Infectious Diseases

Editors: Under Review for MIDL 2020

Abstract
With the advent of deep learning, convolutional neural networks (CNNs) have evolved as an effective method for the automated segmentation of different tissues in medical image analysis. In certain infectious diseases, the liver is one of the more highly affected organs, where an accurate liver segmentation method may play a significant role to improve the diagnosis, quantification, and follow-up. Although several segmentation algorithms have been proposed for liver or liver-tumor segmentation in computed tomography (CT) of human subjects, none of them have been investigated for nonhuman primates (NHPs), where the livers have a wide range in size and morphology. In addition, different NHPs have heterogeneous immune responses to the infections, which eventually appear with a diverse radiodensity distribution in the CT imaging. In this study, we employed one of the state-of-the-art algorithms, feature pyramid network (FPN) for automatic liver segmentation in whole-body CT images of NHPs. The efficacy of the method was evaluated on 82 scans of 37 animals, including pre and post-exposure to different viruses such as Ebola, Marburg, and Lassa. Using 10-fold cross-validation, we obtained an average 93% Dice score, and 5.6% relative absolute volume difference for the segmented livers. While more investigation is required, the preliminary results of our study demonstrated the efficacy of FPN for liver segmentation in infectious disease imaging research.

Keywords: Feature pyramid network, infectious diseases, automated segmentation, liver, CT, whole body.

1. Introduction
Accurate volumetric segmentation of livers from computed tomography (CT) images is crucial for dose estimation, follow-up assessment, and evaluation of therapeutic response. The liver is a spatially complex, three-dimensional (3D) organ that usually shows similar density to the nearby organs on non-contrast CT scans. Manual delineation of liver is considered as the gold-standard, however the process is time-consuming and suffers from inter- and intra-observer variability. Specifically, for a large dataset, the manual process is quite unmanageable in a short period of time. Therefore, an automated segmentation method is highly desirable to accurately segment the liver and allows faster processing of large data sets, while removing inter- and intra-observer biases and variation.

Among recent methods, a number of 2D and 3D convolutional neural network (CNN) based automated liver and liver-tumor segmentation methods were reported in the Liver Tumor Segmentation (LiTS) challenge (Christ) with a primary focus on liver tumor segmentation. Han et al. (Han, 2017) proposed a combination of two UNet-like models in a 2.5 D framework, where the first model was used only for coarse liver segmentation and the
second model was trained with the segmented liver region to segment both liver and tumors in a fine-tuned manner. Li et al. (Li et al., 2018b) developed another UNet based model in a hybrid densely connected fashion (H-DenseUNet) for liver and liver-tumor segmentation. The H-DenseUNet consisting of a 2D DenseUNet and a 3D counterpart aggregates the volumetric contexts in an auto-context mode (Tu and Bai, 2009). Jin et al. (Jin et al., 2018) used a 3D hybrid residual attention-aware segmentation method RA-UNet for liver and liver-tumor segmentation, which has a basic 3D UNet structure with low-label and high-label feature fusion to extract the contextual information. On the other hand, Jiang et al. (Jiang et al., 2019) proposed a 3D fully connected network structure, composed of multiple attention hybrid connection blocks (AHCBlocks) densely connected with both long and short skip connections and soft self-attention modules for liver-tumor segmentation.

Although the above methods were effective in human livers and liver-lesion segmentation from contrast-enhanced CT scans, their performances haven’t been evaluated yet for animal models. For a high consequence infectious disease study in a biosafety level 4 (BSL-4) environment, animal models are used to study viral pathogenesis as well as to test medical countermeasures. The automated liver segmentation evaluated in this study using
FPN for Liver Segmentation of nonhuman primates

Table 1: Diversity in the dataset for automated liver segmentation.

<table>
<thead>
<tr>
<th>Virus Types</th>
<th># of Animals (total 37)</th>
<th># of scans (total 82)</th>
<th>Range of Animal Weights (kg)</th>
<th>Range of Liver Volumes (ml$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rhesus</td>
<td>Cynomolgus</td>
<td>Pre</td>
<td>Post</td>
</tr>
<tr>
<td>EBOV</td>
<td>25</td>
<td>0</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>LASV</td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>MARV</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

the whole-body CT scans of these animal models is challenging due to the diversity of the subjects that includes different species with a wide range of animal ages/sizes, and change of imaging field of view for different scanners. In addition, the infected livers appear with heterogeneous radiodensity distribution, changes in size and shape, and unusual deformation due in part to the presence of subcutaneous/intra-abdominal fat or applying anesthesia at the time of scanning that favors intestinal gas, and motion blur due to breathing artifact. Example images shown in Figure 1 demonstrate the challenges associated with this segmentation task.

In this study, we employed a 3D version of the feature pyramid networks (FPN) (Lin et al., 2017a) to segment liver in animal models. FPN is chosen primarily for its inherent multi-resolution feature extraction capability. To the best of our knowledge, this is the first study that uses CNN-based method to segment liver in nonhuman primates (NHPs) with a history of infectious diseases.

2. Methodology

2.1. Dataset

A total of 82 CT images were acquired from 37 nonhuman primates (rhesus (Macaca mulatta) and cynomolgus macaques (Macaca fascicularis)), ranging in weight from 3kg to 14 kg. Pre and post-exposure scans at different time intervals for different studies were used for this project. These animals were exposed to either Ebola (EBOV), Marburg (MARV) or Lassa (LASV) viruses. The data set is summarized in Table 1. Sixty-four non-contrast and eighteen contrast-enhanced scans were obtained from using either a Philips Gemini 16-slice CT scanner or a Philips Precedence 16-slice CT scanner (Philips Healthcare, Cleveland, OH, USA) in helical scan mode. Whole-body scans covered 1) head to pelvis or 2) head to knee. The scanner parameters were set at 140 kVp, 250 or 300 mAs/slice, 1-mm thickness with 0.5-mm increments. The manual liver outlining was performed by a previously trained imaging fellow and monitored by a CT body radiologist. If breathing motion was observed (even if subjects were anesthetized and intubated), the contour was drawn halfway through the extents of the breathing artifact. The gallbladder was excluded if it was clearly visible.

2.2. Preprocessing

In the pre-processing step, first, all the CT images were resampled to 1mm$^3$ isotropic voxels using cubic interpolation. Second, the density values were thresholded to include the range [-40, 160] that excluded irrelevant organs and objects. Finally, the thresholded image
intensities were rescaled between 0 to 1. An anisotropic diffusion filter (Perona and Malik, 1990) was applied to each image for noise reduction and boundary preservation.

2.3. CNN architecture

We employed feature pyramid network, FPN (Lin et al., 2017a), which has also been adopted in many object-recognition tasks (Zhao et al., 2019; Li et al., 2018a). Recognizing objects at vastly different scales is a fundamental challenge in computer vision. A standard solution to this scale-variant issue is to use feature pyramids that enables a model to detect objects across a large range of scales by scanning the model over both positions and pyramid levels. The FPN consists of two pyramidal pathways; bottom-up and top-down. The top-down architecture with lateral connections builds the high-level semantic feature maps at all scales. The principle advantage of FPN is that it produces a multi-scale feature representation in which all levels are semantically strong, including the high-resolution levels (Abdulla, 2017). For the bottom-up feature encoder, we have chosen to use ResNet50 (He et al., 2017).

2.4. Training

The network was trained using input patches of size $64 \times 64 \times 64$ voxels. The input patches were extracted randomly from both liver and non-liver areas with equal numbers. Patches were called liver or non-liver patches based on the center voxel. The corresponding manual segmentation mask was blurred with a 3D Gaussian filter with sigma equal to $1 \times 1 \times 1 \text{mm}^3$. We kept the mini-batch size set to 90. The parameter updates were performed using the Adam optimizer with a learning rate of $5 \times 10^{-5}$ and decay rate of $5 \times 10^{-5}$. The model was trained for 50 epochs, whereas the input patches were reshuffled before each epoch.

2.5. Loss function

Segmentation of any organ is often seen as a voxel-wise classification task, where high class-imbalance (sample ratio of different classes) has a significantly negative impact on the performance. Considering the variation of class-ratios (foreground/background) among the subjects and hard-to-detect boundaries with the nearby organs, we chose focal (Lin et al., 2017b) as the loss function.

2.6. Post-processing

The output of the CNN is a probability map image at 1mm$^3$ resolution. This probability map image was resampled to the original image size. Similar to the preprocessing step, a Gaussian filter was used to smooth the resampled probability map image. Later, an empirically chosen threshold of value 0.5 was used to suppress the low probability values and set higher probability values as the liver. Finally, 3D connected component-based object processing was used to reduce the false positives.

2.7. Performance comparison

In order to compare the efficacy of the proposed method, we applied a 3D version of UNet (Ronneberger et al., 2015) with the same input patch-size, batch-size, loss-function, opti-
FPN for Liver Segmentation of nonhuman primates

Figure 2: Example images of liver segmentation; column-wise from left to right, original CT images, the probability map provided by the FPN, liver mask by threshold of the probability map, and the manually outlined liver mask. Row-wise from top to bottom: diffusely hypodense liver from an infected animal, unusually shaped liver, and the liver with lower in-plane resolution due to a change in field of view as shown in Figure 1.a. Note: All the images are cropped and zoomed for better visualization.

3. Results

We performed subject-wise 10-fold cross-validation using images from 37 animals. In other words, all pre and post scans of the test subjects/animals were left out from the training. Animal images were randomized before splitting into 10 folds. The example images of the segmentation outputs are shown in Figure 2 for challenging cases mentioned in Figure 1. Figure 3 shows example images for a comparison between UNet and FPN. As seen in the CT images, when the livers appear with an iso-dense distribution with the surrounding tissues like muscles, kidney, stomach, and the partial volume effect of the gallbladder, the segmentation algorithms mostly suffer in finding a clear boundary between the tissues. In
Figure 3: State-of-the-art comparison. From left to right, original CT image, segmented liver by UNet, segmented liver by FPN, and the manual liver mask. Example images show where both UNet and FPN have quite similar output (top), and FPN outperformed UNet (marked by yellow arrow in middle and bottom rows). Note: All the images are cropped and zoomed for better visualization.

Figure 3 as example, we observed that the proposed FPN is better than UNet in these regions.

3.1. Quantitative evaluation

To validate the effectiveness and robustness of our approach, we used four metrics of the segmented liver including Dice similarity, sensitivity, average surface distance (ASD), and relative absolute volume difference (RAVD). For the last two evaluation metrics, the smaller the value is, the better the segmentation.

In order to understand the effect of virus infections on the liver, we separately evaluated the segmentation performance of the pre and post scans for both algorithms. The average scores of these metrics can be found in Table 2. More detail statistics on the above scores can be found in Figure 4.
Table 2: Summary of the quantitative scores using UNet and FPN.

<table>
<thead>
<tr>
<th>Scans</th>
<th>Dice Similarity (%)</th>
<th>Sensitivity (%)</th>
<th>ASD (mm)</th>
<th>RAVD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UNet</td>
<td>FPN</td>
<td>UNet</td>
<td>FPN</td>
</tr>
<tr>
<td>Pre</td>
<td>91.68</td>
<td>94.46</td>
<td>89.39</td>
<td>94.19</td>
</tr>
<tr>
<td>Post</td>
<td>91.50</td>
<td>90.42</td>
<td>89.85</td>
<td>90.01</td>
</tr>
<tr>
<td>Overall</td>
<td>91.60</td>
<td>92.69</td>
<td>89.59</td>
<td>92.35</td>
</tr>
</tbody>
</table>

Figure 4: State-of-the-art comparison using evaluated scores; (a) overall Dice score and sensitivity, (b) Dice, (c) sensitivity for pre/post scans, and (d) RAVD.

4. Discussion

In automated liver segmentation, one of the most challenging aspects is to detect a clear boundary between liver and the nearby organs such as gallbladder, muscles, kidney, and
stomach. Example images in Figures 2 and 3 suggest that FPN is capable of handling the diversity of our dataset. Figure 3 shows that FPN has better boundary detection capability than UNet. In this study of automated NHP liver segmentation, we obtained 93% overall Dice score which is on par with the reported methods in the LiTS challenge (Christ).

In Table 2, we notice the overall scores for UNet and FPN by our chosen metrics are comparable however, a closer look of those metrics in Figure 4 indicates that the scores are compromised by the few outliers. Both algorithms show better performance on the pre-scan images than the post-infection scans. Specifically, for the post-infection scans, the Dice scores are lower with more sparse outliers (Figure 4. b). Note that after exposure to the viruses, different animals have varying host responses to the viruses, causing variability in liver shape, size and density. While some animals show severe changes in liver characteristics post exposure, this number is under-represented in our dataset. From the subject-wise outputs, we noticed that most of these outliers came from those under-represented scans. Well-represented training data might improve those scans’ outputs.

5. Conclusion

In this work, we investigated the efficacy of a CNN-based method for automated liver segmentation in nonhuman primates using whole-body CT scans. Published research has focused on liver tumor segmentation in human subjects, whereas for animal models the liver segmentation task is limited likely due to the challenging nature of the effort. To the best of our knowledge, this is the first work that proposes a CNN-based automated segmentation method for liver in NHPs with active viral infections. Our proposed liver segmentation method provides accurate results in this challenging task. Future work will include data augmentation with spatial deformations (Ciçek et al., 2016), which may help to represent the scans from heavily infected animals well in the training. We also plan to apply cascaded CNNs (Reza et al., 2019) in auto-context mode (Tu and Bai, 2009) which have been shown effective in many segmentation tasks.

References


