DeepSeg: Abdominal Organ Segmentation Using Deep Convolutional Neural Networks

Måns Larsson*, Yuhang Zhang* and Fredrik Kahl*†
*Chalmers University of Technology, Göteborg, Sweden
Email: {mans.larsson, zhangyu, fredrik.kahl}@chalmers.se
†Lund University, Lund, Sweden

***Abstract*—A fully automatic method for abdominal organ segmentation is presented. The method uses a robust initialization step based on a multi-atlas approach where the center of the organ is estimated together with a region of interest surrounding the center. As a second step a convolutional neural network performing pixelwise classification is applied. The convolutional neural network consists of several full 3D convolutional layers and takes two input features, which are designed to ensure both local and global consistency. Despite limited training data, our preliminary experimental results are on par with state-of-the-art approaches that have been developed over many years. More specifically the method is applied to the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault" in the free competition for organ segmentation in the abdomen. It achieved the best results for 3 out of the 13 organs with a total mean dice coefficient of 0.757 for all organs. Top score was achieved for the gallbladder, the aorta and the right adrenal gland.**

I. Introduction

Segmentation is a key problem in medical image analysis, and an automated method for organ segmentation can be crucial for numerous applications in medical research and clinical care such as computer aided diagnosis and surgery assistance. The high variability of the shape and position of abdominal organs makes segmentation a challenging task. Previous work done on segmentation of abdominal organs includes, among others, multi-atlas methods [1], patch-based methods [2], and methods based on probabilistic atlas [3], [4]. These techniques achieve great results for a lot of abdominal organs but might struggle with segmentation of organs where the anatomical variability is large.

Recently, deep convolutional neural networks have shown great performance and achieved state of the art results in many computer vision applications [5], [6]. This fact can be partly attributed to the constant increase in available computing power, most notably GPU computing solutions, and the availability of large annotated datasets. In the field of medical image analysis this development has led to an increase in methods based on deep convolutional neural networks, often with great results [7], [8]. This recent development serves as a motivation to utilize convolutional neural network for abdominal organ segmentation.

In this paper, a fully automatic method for segmentation of abdominal organs in contrast enhanced CT images is presented. The first part of the method serves as a coarse and robust localization of the target organ. This part is based on the feature-based multi-atlas approach presented in [9]. The second part utilizes a deep convolutional neural network to perform pixelwise classification. The input features used for the network are two 3D patches of different resolution centered around the voxel to be classified. The first input feature has fine resolution and is meant to provide the network with local information ensuring local precision while the second input feature has a coarse resolution and is meant to ensure global spatial consistency. This dual input approach is used to give the network enough information to make good predictions despite the high anatomical variability of some abdominal organs. As a final step a simple post-processing is done by removing all parts of the segmentation except the largest connected component.

The presented method is used in the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault" where it achieved state of the art results in the free competition for organ segmentation in the abdomen. To this date, our method gives the best results for 3 out of the 13 organs.

II. Proposed Solution

Our method segments each organ independently and can be divided into three parts:

1) Localization of region of interest using a multi-atlas approach.
2) Pixelwise binary classification using a convolutional neural network.
3) Postprocessing by thresholding and removing all positive samples except the largest connected component.

Each step will now be described in detail.

A. Localization of region of interest

This part of the method provides a robust initialization of the segmentation. The goal is to locate the center voxel of the organ in the target image. When this has been done a prediction mask is placed centered around the predicted organ center. The prediction mask later defines the region of interest where the convolutional neural network is initially applied. The use of an initialization method enables us to train more specialized networks that only need to differentiate between a certain organ and the background. This means that the classification task that the network needs to perform is simplified and smaller networks can be used.

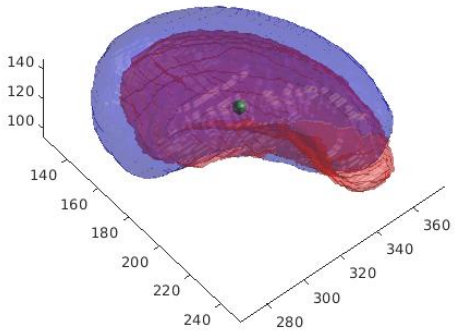

Fig. 1. Example of localization of region of interest for the Spleen. The green sphere is the estimated center point, the red mask describes the ground truth and the blue mask describes the estimated region of interest.

The location of the organ center in the target image is done using a feature-based multi-atlas approach. Each atlas image is registered to the target using the method described in [9]. This registration is performed individually for each organ and atlas image. The transformations estimated are then used to transform each organ center point from an atlas image to the target image. The median of these transformed center points is then used as the center point for the region of interest in the target image. The reason for using the median, and not for instance the mean operator, is that it provides a robust estimate of the center point, that is, it is not affected by a few, spurious outliers.

The prediction mask is estimated using the ground truth segmentations of the atlas images. Let the ground truth segmentations be represented by a binary image of the same dimension as the atlas image $G^{(l)}$, where l is the image id, and $G^{(l)}_{ijk} = 1$ if and only if voxel with index i, j, k in image with id l is foreground (or organ). Further, define $D^{(l)}$ as the binary image formed by dilating $G^{(l)}$ by a cube of size $25 \times 25 \times 25$ voxels and translating it so that the center of the organ is located at the center of the image. The prediction mask P is then defined as the binary image where each element P_{ijk} is

$$P_{ijk} = \begin{cases} 1 & \text{if} \quad \frac{1}{N} \sum_{l=1}^{N} D^{(l)}_{ijk} \geq \delta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where N is the number of atlas images and δ is a threshold set to $\delta = 0.5$ for the majority of the organs.

Finally, the region of interest R is defined as the prediction mask centered around the estimated center point. An example of a localization of region of interest is show in Figure 1.

B. Voxel classification using a convolutional neural network

The convolutional neural network is applied using a sliding window approach. For each voxel to be segmented two cubes of different resolutions centered around said voxel are extracted and used as input features to the network. The network in return outputs a probability, denoted p_{ijk}, of the voxel being organ.

To speed up the process the network is not applied to every voxel in the area that is being segmented, denoted S. Instead, it takes steps of three in each dimension over S. The probabilities output by the network are then interpolated to every voxel in S. Lastly, all voxels in S that has been assigned an interpolated probability neither close to zero nor close to one will be classified by the network once more. The idea behind this approach is that for easily classified regions the network is only applied to a grid of the voxels while for regions where classification might be harder, such as the boundaries of organs, the network classifies every voxel explicitly.

To reduce the dependency on the quality of the initial region of interest where the convolutional neural network is applied, a region growing algorithm is used. Call the set of voxels that should be segmented S. Further, call the set of voxels already classified by D and the set of voxels with an assigned probability larger than 0.5 by O. The region growing algorithm can then be described by Algorithm 1.

Initialize
- S as the region of interest R
- D as $\varnothing$
- O as $\varnothing$.

while $S \neq \varnothing$ **do**
- Classify voxels in S
- Set $D = D \cup S$, and O as the set of voxels with an assigned probability larger than 0.5
- Let O^+ be the set O dilated by a cube of size $12 \times 12 \times 4$ voxels
- Set $S = O^+ \setminus D$

end

Algorithm 1: Region growing algorithm for convolutional neural network classification.

The usage of the region growing algorithm means that even though the initial region of interest only covers part of the organ, a successful segmentation is still possible.

Convolutional neural network setup: The convolutional neural network used performs pixelwise binary classification. The input features for the network are two image cubes, one with a fine resolution similar to the original CT image and the other with a coarse resolution. The fine resolution input feature is meant to provide the network with local information ensuring local precision while the coarse resolution input feature is meant to ensure global spatial consistency. These inputs are then processed separately by two sets of 3D-convolution and max-pooling layers. Afterwards the aggregated image features of both these image patches are merged and processed by two consecutive fully connected network layers. Finally, a two-way softmax operation is applied calculating the

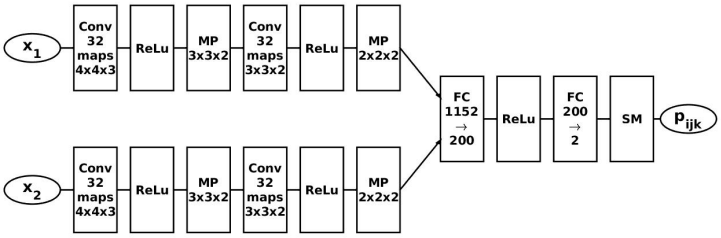

Fig. 2. Structure of the convolutional neural network used, both type and size of each layer is shown. The following abbreviations are being used: Conv: Convolutional layer, ReLu: Rectified Linear Unit, MP: Max Pooling, FC: Fully Connected and SM: Soft Max. Both inputs are cubes containing $27x27x12$ voxels and are centered around the voxel being classified. Input $\mathbf{x_1}$ has as high resolution with voxels of size $1 \times 1 \times 3$ mm^3, while input $\mathbf{x_2}$ is downsampled by a factor of five in each dimension.

probabilities for foreground or background classification. Between each layer a rectified linear unit is added as an activation function. A schematic of the convolutional neural network is shown in Figure 2.

Implementation and Training: For the implementation of the convolutional neural network the framework Torch7 was used [10]. For each convolutional neural network the training and validation set were extracted from the region of interest calculated as described previously and the area around the part of the image describing the organ. This was done for each image in the training set. For the majority of the organs a balanced training set was used, meaning that there was an equal amount of foreground and background samples in the training set. However, since some of the organs are quite small this leads to a relatively small training set. Several methods, listed below, were used to deal with this problem.

1) For organs present in pairs, kidneys and adrenal glands, training samples from both the left and the right organ were used. Note that this does not pose a problem during inference since the initialization part of the method will separate the organs. Hence, the network will not need to differentiate between for example the left and the right kidney.
2) Expansion of the training set by adding slightly distorted CT images, transforming them using a random affine transformation similar to the identity transformation. The transformation T was randomized as

$$T = \begin{pmatrix} 1+\delta_{11} & \delta_{12} & \delta_{13} & 0 \\ \delta_{21} & 1+\delta_{22} & \delta_{23} & 0 \\ \delta_{31} & \delta_{32} & 1+\delta_{33} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where δ_{ij} are independently and uniformly randomized numbers between -0.25 and 0.25 for $i = 1, 2, 3$ and $j = 1, 2, 3$.
3) Including a greater number of background samples than foreground samples in the training set. This leads to a larger but unbalanced training set.

The choice of what methods to use were empirically decided individually for each organ. The evaluation used for the decision was how well the network performed on the validation set.

The networks were trained in mini batches using stochastic gradient descent with Nesterov's momentum [11] and weight decay. The training parameters were set to: batch size 100, learning rate $5 \cdot 10^{-3}$, momentum weight 0.9, weight decay 10^{-5}. The error function used was negative log likelihood. When an unbalanced training set was used the loss was multiplied by a factor k for foreground samples where k is the ratio between background and foreground samples. To avoid overfitting the layers of the network were restricted using dropout during training [12]. The networks were trained for ten epochs or more, the network obtaining the highest validation score were finally picked to be used for the segmentation.

C. Postprocessing

As a final step the probabilities from the convolutional neural network are thresholded with a value of 0.5 creating a binary image. For this binary image everything but the largest connected component is set to zero producing the final segmentation.

III. Experimental Result

This method were tested by submitting an entry to the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault" in the free competition for organ segmentation in the abdomen [13]. In this challenge, there are 30 CT images coupled with manual segmentations of the following organs: (1) spleen, (2) right kidney, (3) left kidney, (4) gallbladder, (5) esophagus, (6) liver, (7) stomach, (8) aorta, (9) inferior vena cava, (10) portal vein and splenic vein, (11) pancreas, (12) right adrenal gland, (13) left adrenal gland. The numbers in the parentheses will from now on be referred to as organ id. These 30 images and segmentations are available for method development and validation. Out of the 30 images 20 were used for training and 10 were used for testing.

In addition to these images training data from the VISCERAL challenge was also used for training. The VISCERAL training data consists of 20 unenhanced

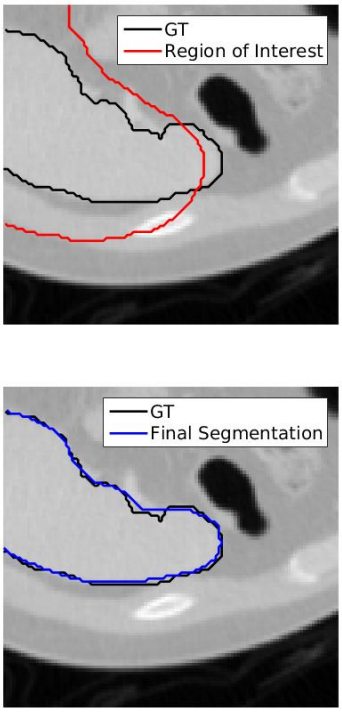

Fig. 3. Example of the resulting segmentation of the spleen for on CT slice. In both images the edge of the ground truth is marked in black. In the left image the edge of the initial region of interest is marked in red and in the right image the edge of the final segmentation is marked in blue. Note that even though the initial region of interest did not contain the entire organ the final result still does, this is due to region growing. This segmentation was one of the most successful in the validation set and achieved a DICE coefficient of 0.967.

whole body CT images and 20 contrast enhanced CT images over the abdomen and thorax. In these images organ with organ ids 1, 2, 3, 4, 6, 8, 11, 12 and 13 were manually segmented. The unenhanced whole body CT images were excluded from the training set for organs with organ id 1, 2, 6, 8, 10, 11 and 12 since they differed too much from the enhanced CT images. All images were resampled to the same resolution of 1 mm $\times$ 1 mm $\times$ 3mm. For the right kidney, a network trained on a training set formed by samples from both the right and the left kidney was used. For the stomach, the data set was expanded with distorted CT images and for the left adrenal gland an unbalanced data set was used with twice as many background samples as foreground samples.

For the test set of the MICCAI challenge the CT images are available for download. The competitors then apply their methods, segmenting the organs present in the CT images. The segmentation files are then submitted to a test server that calculates the DICE coefficient for each organ and posts the result to the publicly available leaderboard. The final results are given in Table I with the currently two best competitors:

- **IMI** - algorithm name: *IMI_deeds_SSC_jointCL* submitted by Mattias Heinrich at the Insitute of Medical Informatics, Lübeck, Germany.
- **CLS** - algorithm name: *CLSIMPLEJLF_organwise* submitted by Zhoubing Xu at the Vanderbilt University, Nashville, TN, USA.
- **other best** - this column contains results from other competitors, the score is only shown if they are the highest for that organ.

TABLE I
FINAL RESULTS MEASURED IN DICE METRIC FOR ORGAN SEGMENTATION IN CT IMAGES. OUR APPROACH GIVES THE BEST RESULTS FOR 3 OUT OF THE 13 ORGANS. HERE '-' MEANS THAT ONE OF THE SPECIFIED METHODS ACHIEVED BEST RESULT.

Organ	*IMI*	*CLS*	*other best*	*Our*
Spleen	0.919	0.911	**0.964**	0.930
Right Kidney	**0.901**	0.893	-	0.866
Left Kidney	0.914	0.901	**0.917**	0.911
Gallbladder	0.604	0.375	-	**0.624**
Esophagus	**0.692**	0.607	-	0.662
Liver	**0.948**	0.940	-	0.946
Stomach	**0.805**	0.704	-	0.775
Aorta	0.857	0.811	-	**0.860**
Inferior Vena Cava	**0.828**	0.760	-	0.776
Portal Vein and Splenic Vein	0.754	0.649	**0.756**	0.567
Pancreas	**0.740**	0.643	-	0.602
Right Adrenal Gland	0.615	0.557	-	**0.631**
Left Adrenal Gland	**0.623**	0.582	-	0.583
Average	**0.790**	0.723	-	0.757

IV. DISCUSSION

In Table II a comparison of the validation score and the test score of our method is presented. As can be seen from the table there is a large difference between validation and testing scores for some organ. This means that our networks do not generalize well to the test set for these organs which might be an indication of overfitting and that the input features and structure of our network is not ideal to learn high order information that generalize to all other CT images. However, since the validation data has not been used for the actual training, only for the decision on when to stop the training, these differences might not be only due to overfitting. Instead it might be due to the existence of anatomical variations in the test set that differs too much from anything seen in the training and validation images for the network to perform well. A specific example of where our method performed badly on the test data, for an organ with good validation result, is shown in Figure 4. Here, the network has classified most of the right kidney correctly. However, it has also classified a lot of surrounding organs or tissue as right kidney as well.

The ideal solution to this problem would be to include more images in the training set. This however, requires more manually segmented CT images which are not always easy to acquire. Other approaches to solve this problem would be to train a network on several organs, and then fine tune the network weights for each specific organ. This could enable the network to learn higher order features that differentiates well between all organs in the CT image, not only between the organs located closest to the organ that is currently being segmented.

TABLE II
COMPARISON OF VALIDATION SCORE AND TEST SCORE MEASURED IN DICE METRIC FOR ORGAN SEGMENTATION IN CT IMAGES.

Organ	*Validation*	*Test*
Spleen	0.944	0.930
Right Kidney	0.940	0.866
Left Kidney	0.928	0.911
Gallbladder	0.744	0.624
Esophagus	0.724	0.662
Liver	0.947	0.946
Stomach	0.823	0.775
Aorta	0.892	0.860
Inferior Vena Cava	0.823	0.776
Portal Vein and Splenic Vein	0.632	0.567
Pancreas	0.689	0.602
Right Adrenal Gland	0.600	0.631
Left Adrenal Gland	0.580	0.583
Average	0.790	0.757

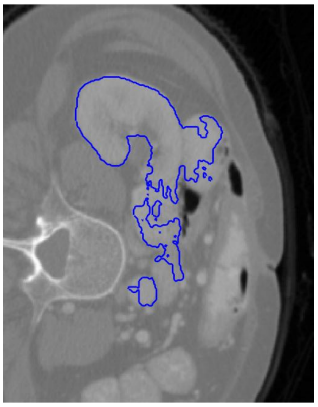

Fig. 4. Example of the resulting segmentation of the right kidney for a CT slice from the test set. The final segmentation is marked in blue. This segmentation was one of the examples where the method performed badly.

V. CONCLUSION

In this paper, a method for abdominal organ segmentation was presented. The method uses a robust initialization algorithm based on a multi-atlas approach for finding a region of interest where the organ to be segmented is located. As a second step a convolutional neural network is applied performing pixelwise classification, the network uses two sets of input features to ensure both global and local consistency. The method was evaluated by submitting an entry to the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault" in the free competition for organ segmentation in the abdomen. This entry achieved on par with state-of-the-art for a majority of the organs to be segmented with a mean dice coefficient of 0.757. Future work includes redesign of the network structure and input features used, this to improve the performance of the network and how well it generalizes. Also, a more sophisticated method for postprocessing can be used, incorporating shape information.

REFERENCES

[1] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 9, pp. 1723–1730, 2013.

[2] Z. Wang, K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert, "Geodesic patch-based segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, ser. Lecture Notes in Computer Science, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer International Publishing, 2014, vol. 8673, pp. 666–673. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10404-1_83

[3] H. Park, P. Bland, and C. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 4, pp. 483–492, April 2003.

[4] C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, and K. Mori, "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3d abdominal ct images," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, ser. Lecture Notes in Computer Science, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer Berlin Heidelberg, 2013, vol. 8150, pp. 165–172. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40763-5_21

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR (to appear)*, Nov. 2015.

[7] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *MICCAI*, vol. 2, 2013, pp. 411–418.

[8] H. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. Turkbey, and R. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, vol. 9349, pp. 556–564. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24553-9_68

[9] F. Kahl, J. Alvn, O. Enqvist, F. Fejne, J. Uln, J. Fredriksson, M. Landgren, and V. Larsson, "Good features for reliable registration in multi-atlas segmentation," in *Proceedings of the VISCERAL Anatomy3 Segmentation Challenge co-located with IEEE International Symposium on Biomedical Imaging (ISBI 2015)*, 2015, pp. 12–17.

[10] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[11] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th international conference on machine learning (ICML-13)*, 2013, pp. 1139–1147.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[13] Z. Xu, "Multi-atlas labeling beyond the cranial vault - workshop and challenge," 2016, [Online; accessed 18-January-2016]. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/217752