GENTEL : GENerating Training data Efficiently for Learning to segment medical images

Rajat Prince Thakur[1,3] Sergi Pujades Rocamora[2] Lavika Goel[6]
Rolf Pohmann[4] Jürgen Machann[5] Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
[3] Birla Institute for Technology and Science, Pilani, India
[4] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[5] Eberhard Karls University of Tübingen, Tübingen, Germany
[6] Malaviya National Institute of Technology (NIT), Jaipur, India

Résumé

La segmentation précise d'images à résonnance magnétiques (IRM) est cruciale pour de nombreuses applications cliniques. Cependant, une segmentation manuelle visant une précision au niveau du pixel est une tâche longue et fastidieuse. Dans cet article, nous proposons une méthode simple pour améliorer l'efficacité de la segmentation d'images. Nous proposons de transformer la tâche d'annotation d'une image en une tâche de choix binaire. D'abord, nous utilisons plusieurs algorithmes classiques de traitement d'image pour générer plusieurs candidats de masques de segmentation. Ensuite, l'utilisat.eur.rice, au lieu de segmenter les pixels des images, décide si une segmentation est acceptable ou non. Cette méthode nous permet d'obtenir efficacement un grand nombre de segmentations de haute qualité avec une intervention humaine limitée. Avec les images et leurs segmentations sélectionnées, nous entrainons un réseau de neurones de l'état de l'art qui prédit les segmentations à partir des images d'entrée. Nous le validons sur un autre jeu de données IRM, acquis avec un protocole different, et qui contient des segmentations. Nous montrons que le réseau entrainé 1) est capable de segmenter automatiquement des cas où aucune des méthodes classiques n'a obtenu un résultat de haute qualité, 2) est capable de segmenter un autre jeu de données IRM, acquis avec un protocole different et jamais vu lors de l'entrainement, et 3) permet de détecter des annotations erronées dans ce jeu de données. Quantitativement, le réseau entrainé obtient de très bons résultats : Score DICE - moyenne 0,98 et médiane 0,99 - et distance de Hausdorff (en pixels) - moyenne 4,7, médiane 2,0.

Mots Clef

Segmentation IRM, méthode semi-automatique, données limitées

Abstract

Accurately segmenting MRI images is crucial for many clinical applications. However, manually segmenting images with accurate pixel precision is a tedious and time consuming task. In this paper we present a simple, yet effective method to improve the efficiency of the image segmentation process. We propose to transform the image annotation task into a binary choice task. We start by using classical image processing algorithms with different parameter values to generate multiple, different segmentation masks for each input MRI image. Then, the user, instead of segmenting the pixels of the images, she/he only needs to decide if a segmentation is acceptable or not. This method allows us to efficiently obtain high quality segmentations with minor human intervention. With the selected segmentations we train a state-of-the-art neural network model. For the evaluation, we use a second MRI dataset (1.5T Dataset), acquired with a different protocol and containing annotations. We show that the trained network i) is capable to automatically segment cases where none of the classical methods obtained a high quality result ii) generalizes to the second MRI dataset, which was acquired with a different protocol and never seen at training time ; and iii) allows to detect missannotations in this second dataset. Quantitatively, the trained network obtains very good results : DICE score - mean 0.98, median 0.99- and Hausdorff distance (in pixels) - mean 4.7, median 2.0-.

Keywords

MRI Segmentation, Semi-automatic method, scarce data.

1 Introduction

Accurately segmenting clinical images is crucial for many clinical applications, as it allows to focus on a specific part of the human body, such as bone, organs, or different tissues in the human body. However, accurately segmenting

clinical images is very challenging and the problem has attracted many research efforts.

Segmentation methods can mainly be clustered into automatic and semi-automatic methods, the latter requiring some human intervention in the segmentation process. In the automatic methods, we can also find two main categories, the ones based on empiric mathematical computations or models and the ones based on machine learning [1] [2]. Examples of the first ones are algorithms based on image histograms [12] or image features, such as contours [3]. These typically include some post-processing to refine the obtained results. In the second ones, starting with a cohort of segmentation examples, an algorithm is trained to generalize to new unseen data.

Many machine learning algorithms exist [5], and in the recent years, neural networks [4] have been the ones obtaining the best performance. As neural networks have a large number of parameters that need to be solved for, one key component for them to achieve a good performance is the amount of available training data. This has guided researches addressing computer vision tasks, such as object recognition, to create large annotated datasets [6] [7] [8], which have unveiled the potential of neural networks. However, in some domains, such as the medical domain, the amount of available data for a given task may be limited. Although MICCAI Challenges, among others, are contributing to the creation of better and richer cohorts of segmented images, these cohorts still remain relatively small. For example, the Infant Brain MRI Segmentation [16] has 10 subjects, the Large Scale Vertebrae Segmentation Challenge [17] has 120 subjects, the Automatic Structure Segmentation for Radiotherapy Planning Challenge [18] provides 50 cases, the Liver Cancer Segmentation Challenge [19] has 50 cases and the Kidney Tumor Segmentation Challenge [20] provides 210 subjects. This relatively low number of cases is mainly due to two facts. First, the amount of raw data may be limited, as the access to clinical subjects presents several constraints. Second, the annotation of the raw data often presents an important challenge. In this paper we focus on how to efficiently obtain accurate segmentations from raw data.

Two key insights drive our work. The first is that *rater time* is scarce and expensive. While this might not apply for general computer vision problems, as for instance, anybody could segment a cat on an RGB image, segmenting clinical images requires medical skills. However, clinical personnel are heavily demanded and their time is precious. Our second key insight is that for a rater, it is fairly easy to distinguish if a segmentation is good or bad. Thus the binary task consisting of deciding if a segmentation is acceptable or not becomes very fast. Based on these insights, we present a simple, yet effective method to improve the efficiency of the image segmentation process, by transforming the pixel annotation task into a binary choice task.

We start by using classical image processing algorithms with several different parameter values to generate multiple, different segmentation masks for each input image. Then, the rater, instead of segmenting the pixels of the images, she/he only needs to decide if a segmentation is acceptable or not. With this methodology, we annotate a fairly large dataset with minimal rater intervention.
To evaluate our method, we train a neural network with the annotated data and illustrate the results, both qualitatively and quantitatively. Moreover, we show that the trained network i) is capable to automatically segment cases where none of the classical methods obtained a high quality result ii) generalizes to a second MRI dataset, acquired with a different protocol, never seen at training time and iii) allows to detect missannotations in this second dataset. The trained network is made available at `http://gentel.is.tue.mpg.de/`.
The proposed methodology is generic. We demonstrate it by addressing the specific problem of body segmentation in MRI images. We focus on MRI images as, in contrast to CT-Scans, in which standardized Hounsfield units are used, the absolute values in MRI images do not correspond to specific tissues. This makes their segmentation particularly challenging.

2 Task and Datasets

Segmentation task. In this paper we are interested in the detection of the human body, i.e. which parts belong to the human and which don't, in 2D MRI images. Our input images are 2D MRI images and the output is a binary image with pixels corresponding to the body being one, and the rest zero. In this work we used two datasets, one acquired with a 1.5T scanner and another with a 3T scanner. We name them 1.5T and 3T datasets in the rest of the paper.

1.5T Dataset - 6 subjects. The 1.5T dataset was acquired with a 1.5 T scanner (Magnetom Sonata, Siemens Healthcare). Subjects were lying in a prone position with extended arms reaching up. From feet to hands, axial T1-weighted fast-spin-echo images were recorded. The measurement parameters were : 256x192 matrix, 10-mm-thick sections, 10-mm gap between sections, adjusted field of view depending on the extension of the subject (450-530mm) and five sections per acquisition. Each MRI consists of around 110 slices, which varies slightly depending upon the height of the subject. Subjects gave prior written consent. The 1.5T dataset was segmented using an automatic method [14] which was specifically designed for data acquired with the used protocol.

3T Dataset - 33 subjects. The 3T dataset was captured with a 3.0T scanner (PRISMA Fit, Siemens Healthcare) with a bore diameter of 60 cm. The subjects were in supine position with arms lying on the chest. Axial T1-weighted turbo-spin-echo images were recorded from feet to head applying the following measurement parameters : 512x352 matrix, 10-mm-thick sections, 10-mm gap between sections, adjusted field of view depending on the extension of the subject. Five sections per acquisition were acquired within 27s, which is short enough to allow the subject to hold

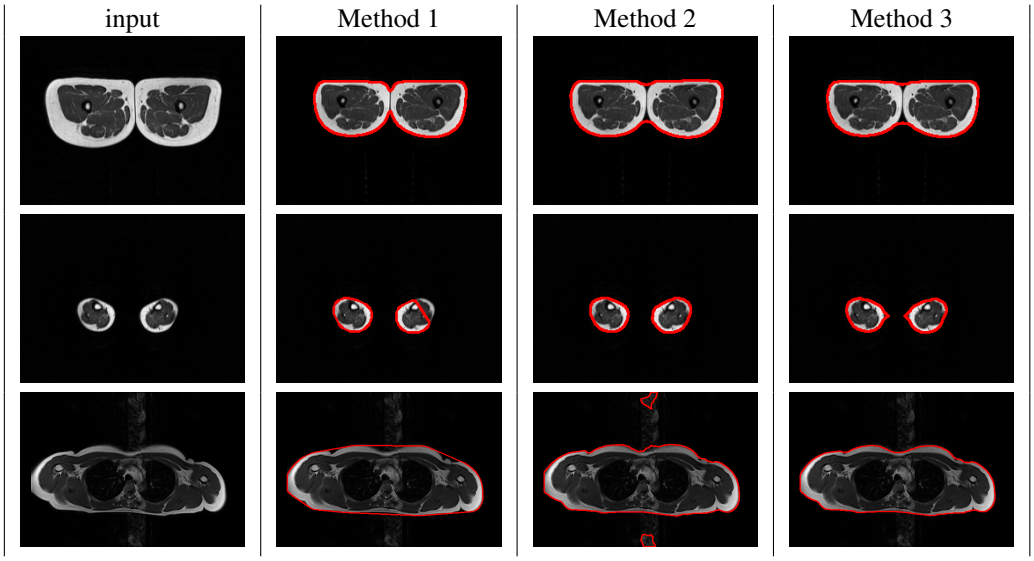


FIGURE 1 – Obtained segmentation candidates. The contours of the candidate segmentations are shown in red over the input image (contour width was enlarged for display purposes). Different methods are better at different images. The best variant for each method is shown.

his breath during the entire scan for slices in the region of the trunk. Each MRI consists of around 90 slices, which varies slightly depending upon the height of the subject. Subjects gave prior written consent.

The automatic segmentation method [14] completely failed on the 3T dataset, mainly due to two reasons : i) the method is not capable to deal with the arms being near the torso as it was designed to handle subjects with the arms reaching up ii) the MRI sequences had a different contrast (due to the TR parameters and the different machines 1.5T vs 3T). Thus, the 3T dataset has no segmentations.

3 Method

To avoid the cumbersome task of manual pixel annotation, we proceed in a two-step methodology to annotate the 3T MRI dataset. First we automatically generate candidate segmentations for each dataset image. Then the candidate segmentations are presented to a rater to decide if the segmentation is acceptable or not.

3.1 Creating segmentation candidates

To segment the body in the MRI images, we used several basic image processing operations to obtain candidate segmentation masks. We developed different methods, in order to handle different cases in the images such as a different number of body parts in the images. After the development of one method, it was run on the 3T dataset, and obvious error cases were used as guidance to create a new method and introduce parameter variants. In this manner we created 3 methods with variants, obtaining 18 candidate segmentations for each input image.

Method 1. We started by blurring the input image with a 5x5 Gaussian kernel, and performed Otsu's binarization [12]. We tried out other binarization methods also but, Otsu's method worked the best for the 3T dataset images. To fill in the remaining holes in the obtained binary image, we applied a floodfilling algorithm [13] starting from the top left corner of the image. We then extracted the contours from the floodfilled image. We discarded contours with a length shorter than 20 pixels and retained the largest 3 contours. We only consider the 3 largest contours as we don't expect to see any more contours in MRIs of subjects in the supine position. Smaller contours would basically correspond to fingers, that we do not consider, or noise. The remaining contours were convexified using a convex hull. The C largest contours were retained. *3 Variants :* We used $C = [1, 2, 3]$.

Method 2. In this case we did not blur the image and performed Otsu's binarization [12] directly on the input image followed by the flood-filling algorithm [13]. We extracted the connected components and removed the ones with an area smaller than 500 squared pixels. From these, we retained the 3 having the largest area. We then performed morphological closing on the remaining contours (using a circular structuring element of radius 10) and applied the flood-filling algorithm [13]. We then extracted the contours from the flood-filled image. We discarded contours with a length shorter than 20 pixels and retained the largest 3 contours as in Method 1. We did not convexify the contours and retained the C largest contours. *3 Variants :* We retained 3 values for $C = [1, 2, 3]$.

Method 3. We started by blurring the input image with a 5x5 Gaussian kernel, and performed Otsu's binarization [12]. To fill up the remaining holes, we applied a flood-filling operation [13]. We called this image img_{init}. We extracted the connected components and kept all of them.

After this we performed one IOP operation, which could be $IOP = 1$ or $IOP = 2$. In $IOP = 1$, we performed morphological closing with a circular structuring element of radius 18. In $IOP = 2$, we performed morphological closing with a circular structuring element of radius 21, and retained only the components having more than half the number of pixels before the morphological closing.
We called this image img_a. After this, we performed morphological closing on the image $max(img_{init} - img_a, 0)$ using the same kernel as above obtaining img_b. We then took the union of img_a and img_b and flood-filled the union image to obtain the final segmentation.
After the operation IOP we extracted the contours from the resulting image. The morphological operation did remove some of the minor details. However, this also provided good contours in noisy image regions where other methods created noisy contours. Performing morphological closing gave an alternative with overall smoother contours. We discarded contours with a length shorter than 20 pixels and retained the largest 3 contours as in Method 1 and 2. In one variant the remaining contours were convexified using a convex hull and in the other the contours were left untouched. The C largest contours were retained. *12 Variants :* We retained 3 values for $C = [1, 2, 3]$, 2 values for $IOP = [1, 2]$, and 2 values for doing or not doing the convexification $[True, False]$.

Generated Candidates. After this step, we had a total of 18 different algorithms generating potentially 18 different segmentation masks. Fig. 1 shows examples illustrating cases where one particular method worked or failed. In practice we observed that in some cases, several variants of the same method, or even different methods, produced the exact same result. For example, if the algorithm only detected one contour in the image, the variants retaining 2 and 3 contours obtained exactly the same result. However, given an MRI slice, we did not know which of the 18 algorithms would produce the most accurate segmentation. Thus we asked an expert rater to select acceptable segmentations.

3.2 Choosing acceptable segmentations

Although we had many candidate segmentations for each image, we did not ask the rater to identify the "best segmentation mask". Instead, we built a simple tool (see Sec 3.3), which would randomly select one of the candidate segmentation and display the contour(s) superimposed on the MRI image. The rater was then asked to either discard or keep the candidate segmentation, thus transforming the annotation task into a binary decision task. The rater was not able to refine the segmentation and was invited to discard it if errors were visible.
We had 18 candidate segmentations for each MRI image, which might seem a lot, but, for most of the cases several of the 18 possibilities were clear mismatches and were quickly identified. The rating process took around 12h, which accounts to an average of 5 seconds per image. Certain regions of the body required more careful examinations. After the process, 7146 segmentations were retained. Some input images had multiple segmentations, with slight subtle differences mostly located at the boundary of the segmentation. Also some input images had no retained segmentation, as none of the methods provided a correct segmentation.

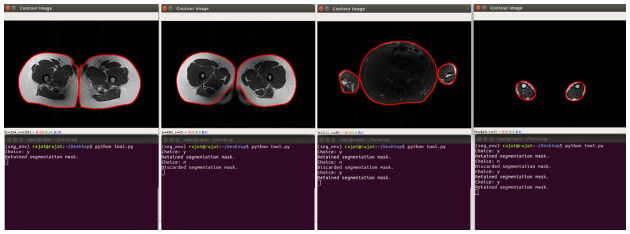


FIGURE 2 – Selection tool : the MRI image is displayed with a contour on it. The rater looks at the images sequentially and decides weather the segmentation should be retained or not.

3.3 Annotation Tool

We used a simple command line tool to annotate the 3T dataset. We used the MRI images and the corresponding segmentation masks obtained with the image processing based algorithms. The tool picks up a segmentation mask and generates contours from it. It then displays the contours superimposed on the MRI images as shown in Fig. 2. The user is then asked if the contours describe an accurate segmentation. The segmentation and the corresponding MRI image are retained if the user presses 'y'. They are discarded in all other cases. This is done for all of the segmentation masks one after the other.

4 Experiments

Training a segmentation network. With the annotated 3T dataset, we split it into train, test and validation sets to train our neural network based segmentation model. The annotated 3T dataset contains 5154 images in the training set, 847 in the validation set and 1145 in the test set. The 1.5T dataset contains 601 images, that were solely used as a second test set. We preprocess all images so that the min and max values in the MRI image are mapped to 0 and 255, as some of the MRI images in the dataset had low contrast. We used the pix2pix [9] network, consisting of a generator and a discriminator. The generator is a convolutional autoencoder and the discriminator is a convolutional encoder followed by a fully connected network. We trained the network for 200 epochs using the AdamOptimizer [15].

Quantitative Results. To evaluate our method, we used the segmentations of the 1.5T Dataset as ground truth and computed DICE and Hausdorff distance (HD) metrics for the trained network and the automatic methods presented in Sec. 3. Let us point out that for 14 images some of our methods detected a contour and the reference had none thus creating an undefined HD value. We observed that these

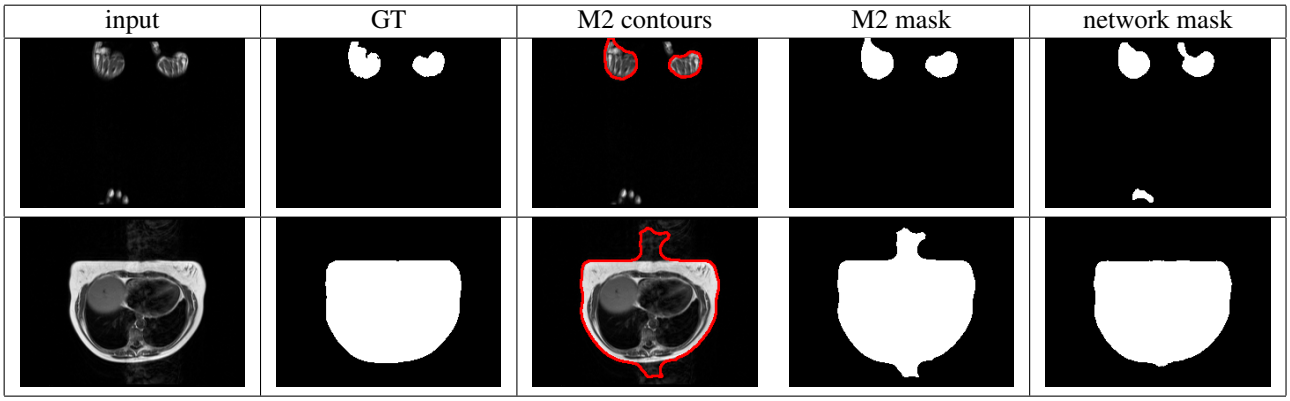


FIGURE 3 – Comparison of representative errors between the best classic algorithm M2(3) and the network. First row : M2 performs numerically better than the network (HD 8.0 vs 157.9). The network detects the fingers at the bottom of the image. Second row : the network performs numerically better than M2. M2 segments the heart beating artifact as body. The network is robust to the motion noise.

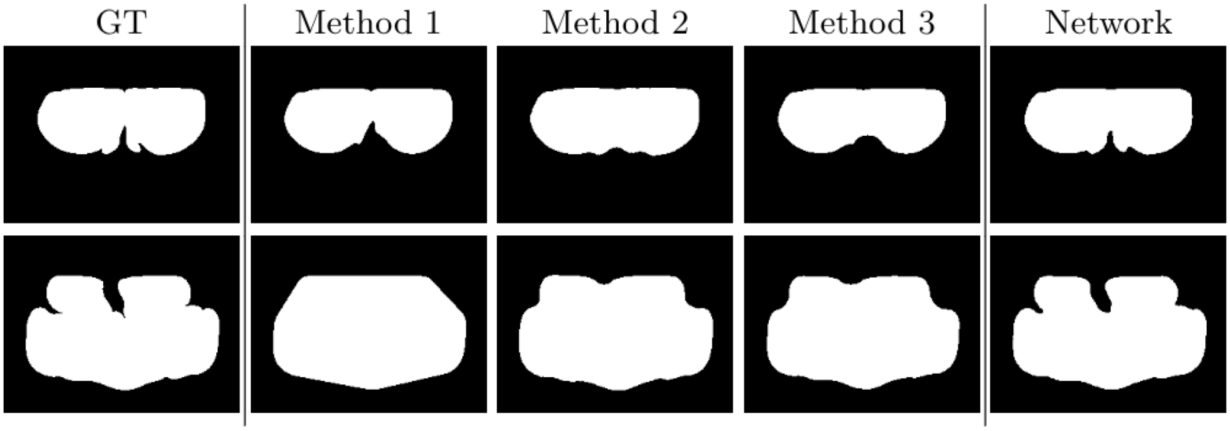


FIGURE 4 – From left to right : Ground truth mask, segmentation obtained with the automatic methods (best variant), the segmentation mask obtained with the trained network. While none of the automatic methods obtained a good segmentation, the trained network predicted a very good mask.

	DICE score		**Hausdorff dist. (pixels)**	
Method	mean	median	mean	median
M1 (3)	0.974	0.988	6.74	2.17
M2 (3)	**0.985**	**0.993**	6.30	2.22
M3 (1 3 0)	0.968	0.980	10.20	6.0
Net	**0.985**	0.989	**4.69**	**2.0**

TABLE 1 – DICE (highest best) and HD (lowest best) results of the methods on the 1.5 Dataset. The best performing variant for M1, M2 and M3 is retained. Best value is highlighted in bold.

errors were consistent across the methods so we visually inspected these 14 images. Two cases arose : 10 very noisy images and 4 errors in the ground truth annotations (see Fig. 6 first four rows). We decided to exclude these 14 images from the 1.5T dataset only for numerical evaluation. We used the remaining 601 images for evaluation. Table 1 shows the obtained results. The network outperforms all methods in terms of Hausdorff distance. Only the automatic method 2, variant 3 is marginally better (+0.004) in terms of median DICE scores.

Qualitative Results. We observed that the algorithm M2(3) had comparable, even slightly better, quantitative DICE results compared to the network. However, when visually inspecting the worse performing cases for each method, we observed that the networks' errors were qualitatively more acceptable in terms of body segmentation. Fig. 3 shows two representative cases. In one case the network segments what seems to be fingers in the bottom of the image and thus the HD is very high (157.9). We found several numeric high errors due to these phantom fingers. On the contrary, M2 (3) was not robust to heart beating motion artifacts, and these were segmented as part of the body. The network was robust to these artifacts. We further show in Fig. 4 that the network produced visually better segmentation masks than the rest of the other methods.

Let us also highlight, that the network was able to gene-

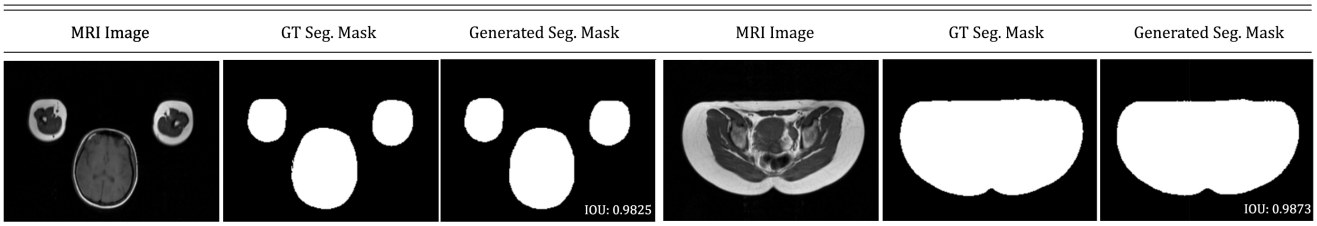


FIGURE 5 – Example of segmentations produced by the trained network on the 1.5T dataset. (First triplet) The network had never seen arms near the head, as all subjects in the training set had their hands on their chest. (Right triplet) The network had never seen prone subjects, as all subjects in the training set were lying in supine position. The trained model was able to obtain valid segmentations in these cases.

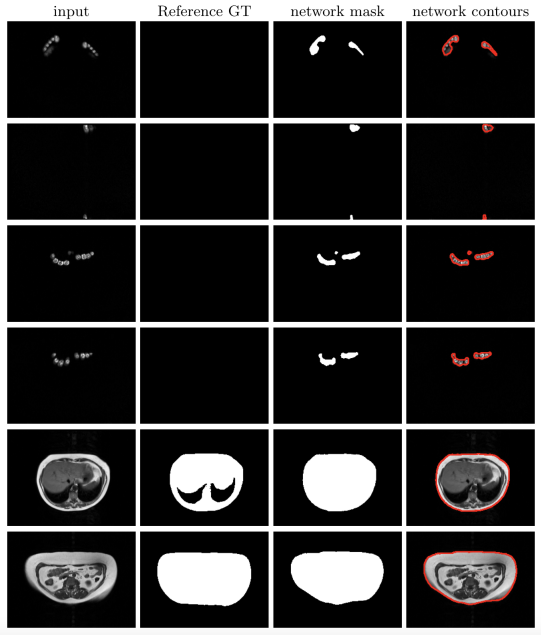


FIGURE 6 – Our method allowed to detect several wrong reference segmentations. From left to right : input image, reference segmentation, segmentation mask and network segmentation contours.

ralize to the 1.5T dataset, which had two significant differences from the 3T dataset. For instance i) subjects had their arms near the head in the 1.5T dataset, while in the 3T dataset all subjects had their hands on their chest ii) subjects where in a prone position in the 1.5T dataset, while in the 3T dataset, the subjects were in a supine position. Fig. 5 illustrates the results for two representative examples where the trained model provides better segmentation masks than all classical segmentation methods.

5 Application

Next we show an application where we use the proposed method to extract a 3D point-cloud from the individually segmented MRI slices of a subject and fit a body model (SMPL [21]) to it. Fitting a body model on the point cloud generated with the MRI segmentation has several interesting applications, such as the registration of the MRI volumes through the body model parametrization, or the study of the deformations on the body due to gravity and the platform experienced during the MRI scan.

We start with the full body MRI images, and we use the proposed method to extract the segmentations masks. Then, with the MRI metadata (pixel-space and slice distance), we transform the segmentations into a 3D point-cloud. We then register the SMPL [21] body model to it by optimizing the point to mesh distance between the point-cloud and the SMPL body mesh. We initialized the SMPL mesh with the shape of the subject - obtained with a 3D scanner in a standing position - and optimized the SMPL pose parameters to best match the segmented point-cloud. We illustrate the segmented 3D point-clouds and the corresponding SMPL mesh fits in Fig. 7.

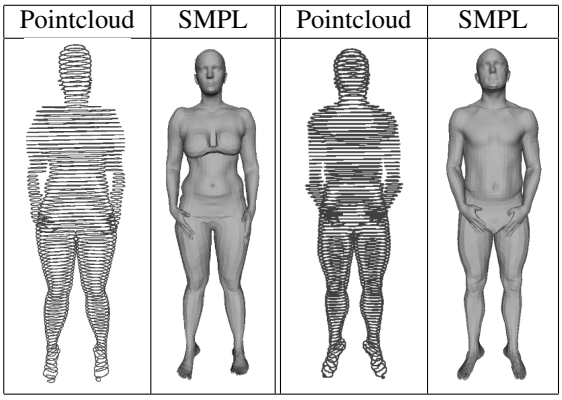


FIGURE 7 – Pairs of obtained point-clouds and SMPL body fit. We initialized the SMPL mesh with the shape of the subject - obtained with a 3D scanner in a standing position - and optimized the SMPL pose parameters to best match the segmented point-clouds.

6 Conclusion

We proposed a simple and effective method to transform the manual pixel annotation task into an easy binary image choice task. By using several basic methods we automatically computed candidate segmentations. These were labeled as correct or wrong by a rater. We demonstrated that

this process provided rich annotations which can be leveraged by a state of the art neural network. We showed that the trained network i) generalized well to another MRI dataset (1.5T), acquired with a different protocol and never seen at training time ii) was capable of automatically segmenting cases where none of the basic methods produced a correct result and iii) allowed to find segmentation errors in the 1.5T dataset. The trained network is made available at `http://gentel.is.tue.mpg.de/`. We also present an application where the segmented images are converted into coherent point-clouds and a body model is fit. Future work will address the segmentation of other anatomic structures, such as bones and the subcutaneous adipose tissue.

7 Disclosure

Michael J. Black has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While he is a part time employee of Amazon and has financial interests in Amazon and Meshcapade GmbH, his research was performed solely at, and funded solely by, MPI.

8 Acknowledgments

We would like to thank Andrea Keller and Tsvetelina Alexiadis for the 3T MRI data acquisition. This research was partially supported by the French National Agency (ANR) under the JCJC project SEMBA.

Références

[1] Bezdek, James C., L. O. Hall, and L.P Clarke. "Review of MR image segmentation techniques using pattern recognition." Medical physics 20, no. 4 (1993) : 1033-1048.

[2] Sivakumar, P., and S. Meenakshi. "A review on image segmentation techniques." International Journal of Advanced Research in Computer Engineering Technology (IJARCET) 5, no. 3 (2016) : 641-647.

[3] Kass, Michael, Andrew Witkin, and Demetri Terzopoulos. "Snakes : Active contour models." International journal of computer vision 1, no. 4 (1988) : 321-331.

[4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015) : 436.

[5] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[6] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet : A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.

[7] Lin TY. et al. (2014) Microsoft COCO : Common Objects in Context. In : Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham

[8] Kuznetsova, Alina Rom, Hassan Alldrin, Neil Uijlings, Jasper Krasin, Ivan Pont-Tuset, Jordi Kamali, Shahab Popov, Stefan Malloci, Matteo Duerig, Tom Ferrari, Vittorio. (2018). The Open Images Dataset V4 : Unified image classification, object detection, and visual relationship detection at scale.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017.

[10] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble supervision alone," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2018, pp. 236–244.

[11] Drapikowski, Pawe and Domagała, Zuzanna. (2014). Semi-Automatic Segmentation of Ct/Mri Images Based on Active Contour Method for 3D Reconstruction of Abdominal Aortic Aneurysms. Image Processing Communications. 19.

[12] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9, no. 1 (1979) : 62-66. 10.1109/TSMC.1979.4310076

[13] Flood fill algorithm. `https://docs.opencv.org/2.4/modules/imgproc/doc/miscellaneous_transformations.html#floodfill`

[14] Würslin C, Machann J, Rempp H, Claussen C, Yang B, Schick F. "Topography mapping of whole body adipose tissue using A fully automated and standardized procedure." Journal of Magnetic Resonance Imaging. 2010 Jan 23;31 :430-439

[15] Kingma, Diederik P and Ba, Jimmy Lei. Adam : A method for stochastic optimization. arXiv preprint arXiv :1412.6980, 2014.

[16] Infant Brain MRI Segmentation Challenge, `http://iseg2017.web.unc.edu/`

[17] Large Scale Vertebrae Segmentation Challenge, `https://verse2019.grand-challenge.org/`

[18] Automatic Structure Segmentation for Radiotherapy Planning Challenge, `https://structseg2019.grand-challenge.org/`

[19] Liver Cancer Segmentation, `https://paip2019.grand-challenge.org/`

[20] Kidney Tumor Segmentation Challenge (KiTS 2019), `https://kits19.grand-challenge.org/`

[21] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J. : SMPL : A skinned multi-person linear model. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 34(6), 248 :1–248 :16 (2015).