Segmentation of the Fascia Lata in Magnetic Resonance Images of the Thigh

Comparison of an Unsupervised Technique with a U-Net in 2D and Patch-wise 3D

Lis J. Louise P.[1], Klaus Engelke[1,2], Oliver Chaudry[1,2]

[1]Institute of Medical Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2]Department of Medicine 3, Friedrich-Alexander-Universität Erlangen-Nürnberg and University Hospital Erlangen
oliver.chaudry@imp.uni-erlangen.de

Abstract. To quantify muscle properties in the thigh, the segmentation of the fascia lata is crucial. For this purpose, the U-Net architecture was implemented and compared for 2D images and patched 3D image stacks in magnetic resonance images (MRI). The training data consisted of T_1 MRI data sets from elderly men. To test the performance of the models, they were applied on other data sets of different age groups and gender. The U-Net approaches were superior to an unsupervised semiautomatic method and reduced post-processing time.

1 Introduction

Quantification of adipose tissue (AT) in skeletal muscle is of growing interest to understand mechanisms of muscle weakness during ageing and in diseases such as sarcopenia and cachexia. The mid-thigh is the preferred anatomical location for such measurements. It was recently shown that intermuscular adipose tissue (IMAT), the combination of AT among muscles and the agglomeration of larger adipocytes within muscles is a very sensitive parameter to monitor exercise effects, the most widely used intervention to prevent muscle weakness with increasing age [1]. The assessment of IMAT requires an accurate segmentation of the fascia lata (FL), an envelope of fibrous tissue separating subcutaneous adipose tissue from muscles and IMAT (Fig. 1). Several unsupervised semiautomatic algorithms for FL segmentation of MRI images of the thigh have been developed, however, due to poor contrast of the FL in MRI images, corrective operator interactions are frequently required [2]. Segmentation of 3D MRI data sets of the thigh consisting of 20 to 30 slices requires a large effort as studies in the field typically consist of hundreds of patient visits. Therefore, the aim of this study was to use a convolutional neural network to reduce the overall segmentation and analysis time. The network should be applicable to populations of men and women of different ages. It was not the aim to fully automate

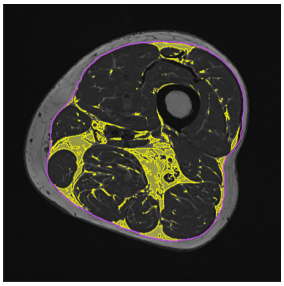


Fig. 1. Typical T_1 MR image of the thigh, 3D segmented fascia lata visible as thin line of fibrous tissue around the muscles (magenta), segmented IMAT within the fascia (yellow).

the process, but to achieve a segmentation accuracy comparable to our current gold standard [2], which includes a supervision of the semiautomatic process and necessary manual corrections by a medical expert.

2 Methods

We used four different MRI datasets (D1-D4) from three different longitudinal exercise intervention trials in subjects with and without sarcopenia. D1 was taken from the FROST study [3] consisting of 43 elderly men (age $\geq$ 72 years, BMI 24.5±1.9 kg/m^2), including two visits and 70 scans in total. From the FRANSO study [4], 17 young men (age $\geq$ 25 years, BMI 23.4±1.9 kg/m^2) were taken to form D2 and 17 elderly men (age $\geq$ 72 years, BMI 27.3±2.2 kg/m^2) to form D3. D4 consisted of 16 elderly women (age $\geq$ 71 years, BMI 24.9±1.4 kg/m^2), from the FORMOSA study [5]. All datasets comprised a wide range of muscle fat content.

2.1 Data acqusition

Whole MR image acquisition was performed using a 3T scanner (MAGNETOM Skyra$^{\text{fit}}$, Siemens Healthineers AG, Erlangen, Germany) and an 18-channel body receive array coil. The images were aquired by a T_1 weighted Turbo Spin Echo sequence: TR: 844 ms, TE: 14 ms, voxel size: 0.5×0.5×3.0 mm^3 (no slice gap), matrix size: 512×512 in 28 slices, bias field was corrected by the N4ITK algorithm [6].

2.2 Deep learning architecture and training

The FL segmentation was performed by the U-Net deep neural network. The implementation followed the original publication [7], except for the padding per convolution, which did not change the input dimensions. To overcome the class imbalance between the thin FL contour around the muscles and the background, the region bordered by the FL volume was segmented instead of the contour. In order to analyze performance, we compared an implementation in 2D with a patched 3D U-Net. The 2D model received the full image resolution while for the 3D approach the images were downsampled to 256×256 pixels, without

downsampling in through plane direction. The input consisted of patches of 4 slices each, both approaches were trained only on D1. Dropout regularization was applied at the end of every convolutional block. The training also included data augmentation, adding random spatial transformations (like shear, rotation, shifts and flips) to data samples. Subsequently, the model was tested on D2, D3 and D4.

The networks were trained to optimize a loss function defined as a linear combination of the dice loss (DL) and the weighted cross entropy (WCE)

$$Loss(y,p) = \alpha WCE + (1-\alpha)DL \tag{1}$$

With DL defined as $1 - DSC$ and DSC being the dice similarity coefficient

$$DSC(y,p) = \frac{2\sum_i y_i p_i + s}{\sum_i y_i + \sum_i p_i + s} \tag{2}$$

With y being the true image and p the predicted segmentation map. The term s denotes a smoothing factor which was set to 0.0001. Analogously, WCE is defined as

$$WCE(y,p) = -(\beta \cdot ylog(p) + (1-y)log(1-p)) \tag{3}$$

with β empirically determined to be 0.4 in order to penalize false positive predictions. The total number of 2D samples was 1960 individual slices (28 slices per 3D volume), which were shuffled, normalized and split into training (70 %), validation (15 %) and test (15 %) sets. The models were implemented using tensorflow and trained on a Nvidia Geforce GTX 1060 6GB GPU, using the Adam optimizer. Due to GPU memory limitations, the batch size for the 2D model was restricted to 2. For the 3D model, the data set was split at patient level into 57 for training, 7 for validation and 6 patients for testing. For the 3D model the patch size was restricted to 4 slices with a batchsize of 1. The optimum values for α=0.2 and the learning rate of 2e-4 were found by a grid search in combination with a ten-fold cross validation. Training the 2D model took 2 hours for 12 epochs. The last 6 epochs were trained at a learning rate of 2e-5 to refine the results, initial weights came from pretraining the model on the segmentation of the cross sectional thigh area. Additional parameters were: dropout rate 0.5, augmentation: horizontal flip and nearest neighbor interpolation, with rotation range 20, width shift range 0.2 and height shift range 0.2.

2.3 Evaluation of the results

For the comparison of the segmentation results between the U-Net and the gold standard, the DSC and the Hausdorff distance (HD) of images were used. HD computes the maximum distance between the sets of non-zero pixels A and B

$$HD(A,B) = max(h(A,B), h(B,A)) \tag{4}$$

with h being the directed distance

$$h(A,B) = \max_{a \epsilon A} \min_{b \epsilon B} \|a - b\| \tag{5}$$

Table 1. Accuracy of the 2D U-Net and unsupervised method against gold standard.

Group	D1	D2	D3	D4
DSC unsup.	0.975±0.075	0.984±0.024	0.993±0.002	0.971±0.018
DSC U-Net 2D	0.996±0.002	0.991±0.008	0.995±0.002	0.985±0.008
HD [pixel] unsup.	15.02±16.76	14.18±14.38	7.344±4.557	23.49±14.83
HD [pixel] U-Net 2D	3.417±1.564	8.575±7.092	5.289±2.418	11.67±5.707
Number of samples	196	442	442	416

DSC was used in order to leave out the background, as this takes up approximately 80 % of the image. Additionally the HD was computed to quantify the offset between the boundaries of the prediction and the gold standard mask, because DSC is not sensitive enough against small differences at the borders of the masks. Before calculating the test accuracies, the prediction was thresholded, counting every pixel <0.1 as background.

3 Results

3.1 2D U-Net model

Final training accuracy for DSC was 0.9950, with a validation accuracy of 0.9955 (before thresholding). Tab. 1 shows the results of the 2D U-Net and the results of the unsupervised segmentation technique before the manual correction process. Fig. 2 shows the 2D U-Net results (after thresholding) and difference images in relation to the gold standard.

As expected, DSC of the 2D U-Net was highest in D1, the dataset used for training. In comparison to D1, DSC was similar for D2 and D3, with D3 being higher than D2. Lowest DSC was found for D4, the cohort of elderly women, which compared to D1 and D2 show a larger amount of IMAT and subcutaneous adipose tissue (SAT). Images of D4 also show more connective tissue, visible in the SAT. The young men of D3 with the least amount of AT, showed the best results. A similar trend was observed for the unsupervised segmentation technique. Numerically, the 2D U-Net results for DSC and HD were always superior.

3.2 3D U-Net model

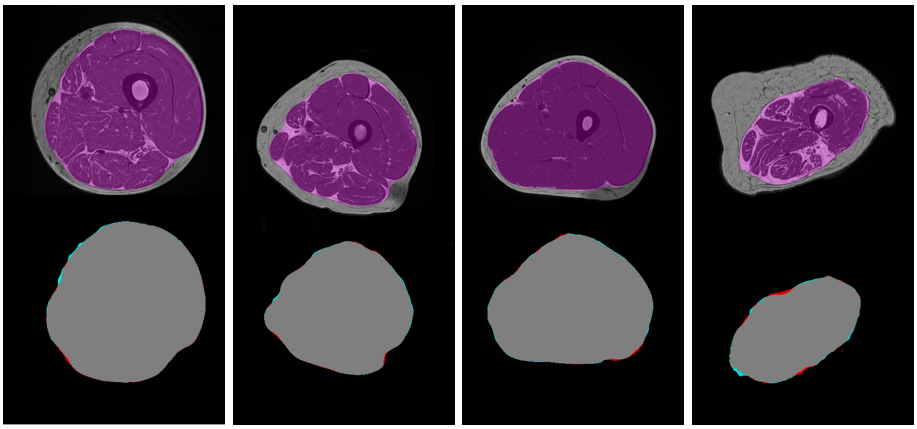
3D U-Net results are shown in Tab. 2. Accuracy is measured in the upsampled prediction masks, after applying a median filter to smoothen the edges. Results for the 3D model were similar to 2D results, but details around the fascia contour were less accurate in 3D, which can be seen in the HD values. In rare cases, in particular for very high levels of fat infiltration, the 2D U-Net model performed poorly and created holes or segmented too much of the SAT (Fig. 3 bottom row, left and right).

Table 2. Accuracy of the 3D U-Net against gold standard.

Group	D1	D2	D3	D4
DSC (upsampled)	0.994±0.002	0.992±0.002	0.994±0.004	0.982±0.001
HD [pixel] (upsampled)	4.224±1.719	6.459±3.813	5.179±3.514	21.698±48.458
Number of patients	7	17	17	16

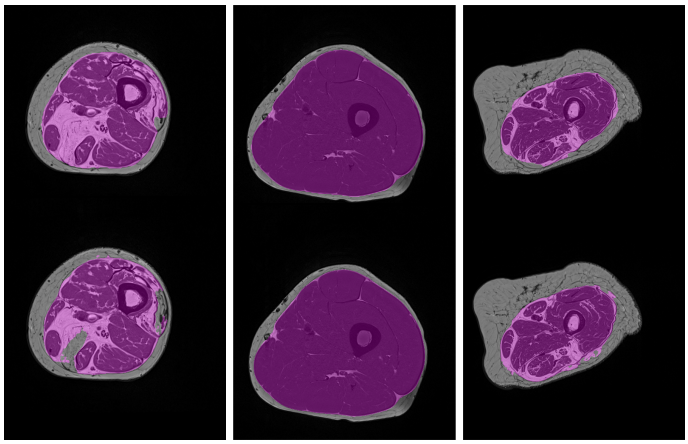
3.3 Post-processing time

Compared to the unsupervised segmentation, manual post-processing times were significantly reduced by the 2D U-Net model (Tab. 3.)



(a) D1 (b) D2 (c) D3 (d) D4

Fig. 2. Top row: MR image with overlay of the prediction mask (magenta). Bottom row: gold standard mask (red), prediction mask (cyan) and intersection of both masks (gray).



(a) D2 (b) D3 (c) D4

Fig. 3. Comparison of the 3D segmentation (upper row) with the 2D segmentation (bottom row) for the same patients in D2, D3 and D4.

Table 3. Manual post-processing time.

Group	D2	D3	D4
Unsupervised [min]	4.3±1.7	2.4±1.1	6.8±2.9
2D U-Net [min]	1.8±0.9	1.0±0.4	4.6±2.5

4 Discussion and Conclusion

Two different U-Net models for the segmentation of the FL were implemented. Although the models were only trained on images of elderly men, both approaches also delivered excellent results for young men. The models were able to predict the FL segmentation with higher accuracy than the unsupervised method, which halved the time for final manual corrections. The patched 3D approach overall showed sligthly lower accuracy in the test datasets but produced smoother, less convex segmentations. Results for D4 indicated that higher subcutaneous and intramuscular AT content caused poorer segmentation. Data from both genders and 3D information should be included in the training data to gain best results.

References

1. Grimm A, Nickel MD, Chaudry O, et al. Feasibility of dixon magnetic resonance imaging to quantify effects of physical training on muscle composition—A pilot study in young and healthy men. European Journal of Radiology. 2019;114:160–166.
2. Chaudry O, Friedberger A, Grimm A, et al. Segmentation of the fascia lata and reproducible quantification of intermuscular adipose tissue (IMAT) of the thigh. Magn Reson Mater Phy. 2020;.
3. Kemmler W, Kohl M, Jakob F, et al. Effects of high intensity dynamic resistance exercise and whey protein supplements on osteosarcopenia in older men with low bone and muscle mass. Final results of the randomized controlled FrOST study. Nutrients. 2020;12(8):2341.
4. Kemmler W, Grimm A, Bebeneck M, et al. Effects of combined whole-body electromyostimulation and protein supplementation on local and overall muscle/fat distribution in older men with sarcopenic obesity: the randomized Controlled Franconia Sarcopenic Obesity (FranSO). Calcif Tissue Int. 2018;103:266–277.
5. Kemmler W, Teschler M, Goisser S, et al. Prevalence of sarcopenia in Germany and the corresponding effect of osteoarthritis in females 70 years and older living in the community: results of the FORMoSA study. Clin Interv Aging. 2015;10:1565–1573.
6. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging. 2010;29:1310–1320.
7. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015;9351:234–241.