Abstract

Severity and progression of degenerative neuromuscular diseases can be sensitively captured by evaluating the fat infiltration of muscle tissue in T1-weighted MRI scans of human limbs. For computing the fat fraction, the original muscle needs to be first separated from other tissue. Five conceptionally different approaches were investigated and evaluated with respect to the segmentation of muscles of human thighs. Besides a rather basic thresholding approach, local (level set) as well as global (graph cut) energy-minimizing segmentation approaches with and without a shape prior energy term were examined. For experimental evaluations, a dataset containing 37 subjects was divided into four classes according to the degree of fat infiltration. Results show that the choice of the best method depends on the severity of fat infiltration. In severe cases, the best results were obtained with shape prior based graph cuts, whereas in marginal cases thresholding was sufficient. With the best approach, the worst-case error in fat fraction computation was always below 11 % and on average between 2 % for tissue showing no fat infiltrations and 6 % for heavily infiltrated tissue. The obtained Dice similarity coefficients, measuring the segmentation quality, were on average between 0.85 and 0.92. Although segmentation of heavily infiltrated muscle tissue is extremely difficult, an approach for reasonably segmenting these image data was identified. Especially the negative impact on the calculated fat fraction can be reduced significantly.

Keywords: Segmentation, Level set, Graph cut, Statistical shape model, Thighs, Muscle, T1-MRI
1. Introduction

Neuromuscular disease is a collective term that describes disorders in the motor function unit, such as hereditary or inflammatory myopathies and neuropathies, motor neuron disorders and neuromuscular junction diseases.

For assessment of neuromuscular diseases, established approaches exist, such as the Medical Research Council (MRC) sum score (Compston, 2010) and the Neuropathy Impairment Score (Dyck et al., 1997), relying on judgement by an expert or by patient’s impressions. A further commonly used method relies on hand-held or fixed myometry by force measurements, which is quite objective but becomes difficult and loses sensitivity in later stages of different neuromuscular diseases (Moxley, 1990).

Though very important for clinical practice and study outcomes, functional testing cannot reveal the underlying anatomical and morphological changes in muscles, which can routinely be visualized by muscle MRI. Driven by the demand for more objective disease markers, MRI has been increasingly utilized for the assessment of neuromuscular diseases. Recently, numerous methods have been proposed to extract markers from MRI scans for assessment of certain myopathies based on different visual scores (Lareau-Trudel et al., 2015; Mercuri et al., 2007; Morrow et al., 2016; Sookhoo et al., 2007; Tasca et al., 2012; Crawford et al., 2017; 2015; Elliott et al., 2015; Abbott et al., 2015; Karlsson et al., 2016).

Although relying on image data, most approaches still contain processing steps which are conducted manually. Lareau-Trudel et al. (2015) for example performed manual segmentation of muscle tissue in case of severely affected subjects while in most other literature, segmentation was performed fully manually (Morrow et al., 2016; Crawford et al., 2017; Elliott et al., 2015; Karlsson et al., 2016). The grading system proposed by Mercuri et al. (2007) is completely based on visual inspection of MRI scans.

This visual semi-quantitative analysis, however, is still subjective and time consuming which thereby provides a strong incentive for the development of fully-automated observer-independent methods for processing the MRI data (Crawford et al., 2017).
This work focuses on the so-called fat fraction (Morrow et al., 2016) computed on thigh MRIs, which proved to provide a high sensitivity for assessing disease progression of neuromuscular diseases. This measure is defined as the ratio between fat-infiltrated muscle tissue and overall muscle tissue and exhibits a single metric extracted from 3D MRI data for assessment, diagnosis and research. The fat-fraction in muscle tissue is not only relevant for assessing neuromuscular diseases, but is also used for investigations of other diseases such as low back pain (Hildebrandt et al., 2017; Teichtahl et al., 2015; Kjaer et al., 2007) and traumatic neck pain (Elliott et al., 2015; Karlsson et al., 2016; Abbott et al., 2015). It was suggested that the fat fraction should be calculated on the complete 3D muscle tissue instead of single 2D slices to gain additional information (Morrow et al., 2016; Willis et al., 2013) providing a further incentive for developing automated methods.

The first essential step in computing the fat fraction is to segment the original muscle tissue to primarily separate it from bone and subcutaneous fat tissue. However, a manual segmentation, especially if the complete 3D image is considered, is highly time consuming and is furthermore subject to significant interobserver variability (Mhuiris et al., 2016; Crawford et al., 2017) providing motivation for the development of automated segmentation methods.

1.1. Related Work

There is comparatively little literature on muscle segmentation of MRI scans of thighs showing pathological muscle tissue. Lareau-Trudel et al. (2015) applied a rather basic segmentation method, originally developed for the assessment of persons classified as obese (Positano et al., 2009), to patients with facioscapulohumeral muscular dystrophy. In a first step k-means clustering was utilized to discriminate between three tissue classes. Subsequently, the boundaries were determined by applying the active contours (snakes) approach (Kass et al., 1988). The proposed method failed in 20% of the slices. Especially when the fat fraction is high, snakes are unable to reliably determine the muscle’s boundary as the border between fat-infiltrated muscle and fat tissue cannot be effectively detected by edge-based active contours without relying on a shape model. Although a manual correction of 20% of the cases saves time
(compared to a completely manual segmentation), the approach is still subject to interobserver variability which is supposed to be particularly distinct for highly-affected subjects (Muiris et al., 2016).

Essafi et al. (2009a,b) investigated the problem of segmenting the medial gastocnemius muscle (of the calf) in T1-MRI images for healthy and diseased subjects is tackled. This problem definition is more difficult compared to (Lareau-Trudel et al., 2015) where muscle is considered as one single class. The goal of the proposed algorithm is to find landmark positions based on local texture features combined with shape knowledge. Geometrical information was inserted by means of diffusion wavelets. By applying a hierarchical diffusion operator, wavelet coefficients were gained for each individual training shape. Subsequently, the dimensionality of these coefficients was reduced by principle component analysis (Essafi et al., 2009a) or by means of the orthonorm method (Essafi et al., 2009b). Unfortunately, in both studies (Essafi et al., 2009b,a), the results were not separately assessed for healthy and pathological cases. Therefore, the performance of the approach for pathological cases is difficult to assess.

A mean landmark error of approximately 12 voxels (Essafi et al., 2009b) and a Dice similarity coefficient (DSC) of 0.55 (Essafi et al., 2009a) indicate that the method is not robust enough to process data showing highly affected muscle tissue. In further approaches, focus was on detecting the fascia lata (Kovacs et al., 2016; Yao et al., 2017) (tissue inside the fascia lata was labelled as muscle) which allows a rough localization of muscle tissue, but not a completely accurate segmentation.

Further segmentation approaches were developed and evaluated for healthy muscle tissue, or for tissue showing at least no distinct fat infiltration, leading to a completely different segmentation task. Orgiu et al. (2015); Positano et al. (2009) proposed a method based on active contours for segmenting MRI scans of human thighs. Baudin et al. (2012a,b) developed approaches for segmenting healthy muscle tissue in the thighs by means of random walks. A further approach relying on a PCA-based shape prior for separately segmenting single thigh muscles was proposed by Andrews et al. (2011). The considered data set contains both healthy and pathological cases, however only muscle deformation and no fat infiltration occurs. Gilles and Pai proposed an atlas-based approach for segmenting muscles in healthy thighs. Karlsson
et al. (2014) evaluated whole-body muscle segmentation method by applying a multi-atlas approach.

In summary, we identified a lack of a quantitative evaluation of segmentation approaches with respect to subjects showing fat infiltrations. Literature on segmenting thighs focus either on healthy subjects only or on subjects showing no distinct fat infiltration, or the evaluation is not performed separately.

Considering the proposed methodologies, most approaches are either atlas-based (Gilles and Pai, Karlsson et al., 2014) or incorporate a shape model (Andrews et al., 2011; Baudin et al., 2012a,b; Essafi et al., 2009a,b). One method only (Lareau-Trudel et al., 2015) performs segmentation by applying a level set approach without incorporating prior knowledge of the shape.

The utilization of state-of-the-art deep neural networks (Ronneberger et al., 2015) is currently inhibited by the small amount of available training data, especially considering severely affected patients.

1.2. Contribution

In this study, we systematically implemented, extended and evaluated several conceptionally different segmentation approaches which were inspired by methods in literature (Essafi et al. 2009b; Lareau-Trudel et al. 2015). Evaluation was performed in combination with variably affected (fat-infiltrated) T1-MRI scans (Fig. 1) which were partitioned into four categories reaching from healthy to severely affected muscle tissue. For segmentation, we compared a basic clustering-based technique with a statistical level set (featuring a local optimization method) and a graph cut approach (featuring a global optimization method). To introduce knowledge of the shape, which was supposed to be important especially in severe cases (Andrews et al., 2011; Baudin et al., 2012a,b; Essafi et al. 2009a,b), we added statistical shape models into the energy formulation. We focused on segmentation approaches in combination with implicitly parametrized shape models because especially in the case of severely affected patient’s data, a reliable manual annotation based on anatomic keypoints (Essafi et al., 2009a,b) for training would be extremely difficult. Level set (also motivated by the work of Lareau-Trudel et al. (2015)) and graph cut approaches were compared in or-
Figure 1: For evaluation purposes, the data set was partitioned into four categories according to the prevalent fat infiltration. Whereas the ‘healthy’ (a) and the ‘easy’ (b) samples do not show any fat infiltration, samples of the ‘moderate’ category (c) show moderate, local and the ‘hard’ category (d) exhibits infiltrations in significant muscle areas.

...der to assess the impact of global compared to local energy minimization requiring for an appropriate initialization. Evaluation was performed based on the obtained segmentation performance (DSC) as well as on the accuracy of the obtained fat fraction exhibiting a (finally even more relevant) domain specific measure.

2. Material & Methods

2.1. Image Material

The complete data set consists of 37 whole-body T1-weighted MRI-scans. It was partitioned into four categories (corresponding to levels of difficulties, Fig. [1] Table[1]). Figure 2 shows an example slice for each of the investigated patients, separately for each category. The images were acquired on a 1.5 Tesla Phillips device between 2012 and 2016 at the University Hospital Aachen. Echo time (17 ms), bandwidth (64 kHz) and echo train length (6) were fixed for all patients. Relaxation time varied between 721 ms and 901 ms. In z-direction the sampling interval was fixed to 7 mm. In x-y-direction the sampling rate varies between 0.931 mm and 1.18 mm.

After concatenation of the whole body scans (Sect. 2.2), a ground truth segmentation was created for all scans. Under supervision of a medical expert (co-author Madlaine Müller), the complete data set was segmented manually. The segmentation started at the most distal slice, where the kneecap could be recognized. Subsequently,
Figure 2: This graphic shows an example (central) slice for each of the investigated MRI scans. As fat infiltration is not uniformly distributed, single 2D slices might seem wrongly categorized as for categorization, the whole 3D scan was considered.
Table 1: Overview of the available MRI data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th># Patients (f/m)</th>
<th>Age</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>'healthy'</td>
<td>15 (8/7)</td>
<td>35.53 ± 12.62</td>
<td>no diagnosed myopathy</td>
</tr>
<tr>
<td>'easy'</td>
<td>6 (6/0)</td>
<td>43.16 ± 13.09</td>
<td>myopathy, no fat infiltration visible</td>
</tr>
<tr>
<td>'moderate'</td>
<td>7 (3/4)</td>
<td>54.00 ± 21.27</td>
<td>myopathy, fat infiltration in small areas</td>
</tr>
<tr>
<td>'hard'</td>
<td>9 (5/4)</td>
<td>57.44 ± 16.21</td>
<td>myopathy, large affected areas</td>
</tr>
<tr>
<td>overall</td>
<td>37 (22/15)</td>
<td>45.59 ± 18.20</td>
<td>complete data set</td>
</tr>
</tbody>
</table>

every fourth slice was labelled to increase efficiency. The segmentation was stopped at
the greater trochanter of the femur indicating the beginning of the hip.

In order to fit the shape model (for level set segmentation with shape prior), the
3D shapes were normalized to a coordinate system with dimensions $251 \times 251 \times 10$.
The reduction of slices in z-direction is motivated by the ground truth segmentation,
which was only available for every fourth slice. Note that in z-direction linear inter-
polation was used, whereas in x-y directions a nearest neighbour method was applied
in order to avoid an inherent enlargement of the segmented shape by low-pass filtering. Nevertheless, segmentation and evaluation were performed on the original image
resolutions.

During this study, we considered one thigh (only the right) per patient only to espe-
cially decrease the manual annotation effort while keeping variability high. This strat-
egy is applied as it is obvious that inter-patient variability is higher than intra-patient
variability.

2.2. Preprocessing

First, single 3D subimages (typically two or three) had to be combined, partly
exhibiting different resolutions and coordinate systems to obtain the complete thigh
volume. Therefore, we utilized cubic B-spline interpolation in combination with his-
togram matching in order to correct for intensity inhomogeneities in colliding and over-
lapping slices.

To suppress background noise, first the Canny-edge detector ($\sigma = 1.8$, $t_{low} = 0.05$,
$t_{high} = 0.1$) was applied to determine the regions-of-interest (i.e. the two thighs refer
to the contours surrounding the largest area). All other pixels were set to an intensity
value corresponding to fat tissue. Thereby, also non-relevant structures such as fingers were effectively suppressed.

For compensation of inhomogeneities within the regions-of-interest, intensity correction as proposed by [Tustison et al. (2010)] was applied. This correction is not only relevant for computing the fat fraction but also to facilitate intensity-based segmentation.

2.3. Segmentation Approaches

For segmenting MRI scans of human thighs, we identified five effective methods:

The first method (GMM) is a basic segmentation approach relying on fitting a Gaussian mixture model followed by thresholding. Although it must be assumed that this model is not appropriate for highly-affected tissue, it serves as a baseline for comparison, especially for non or only slightly affected tissue as well as for initialization purposes (a similar strategy was applied by [Lareau-Trudel et al. (2015)]). Apart from thresholding, we investigated two different level set approaches. The first approach utilized a conventional statistical image energy formulation (LS) whereas the second additionally relied on a statistical shape model (LS-SP). Furthermore, a graph cut approach was investigated (GC) and was also incorporated with a statistical shape model (GC-SP).

2.3.1. Gaussian Mixture Model (GMM)

A Gaussian mixture model is fitted to the data in order to identify clusters of three different classes: muscle, fat and bone/vessels. Initial cluster centres are fixed to the minimum grey value ($s_{\text{min}}$), maximum grey value ($s_{\text{max}}$) and finally a value in between ($s_{\text{min}} + \frac{s_{\text{max}} - s_{\text{min}}}{6}$). The value in-between is chosen to be nearer to the lower value to consider the smaller intensity distance between muscle tissue and background than between muscle tissue and fat. Note that the background is extracted in advance and thus excluded from the centre computation. Two thresholds are obtained by taking the mean value between adjacent cluster centres. A binary opening followed by a binary closing operation removes thin structures and smooths the segmented area (3D circular structuring element (seven voxels), one iteration).
2.3.2. Level Set (LS)

We make use of a statistical level set formulation (Cremers et al., 2006b) because of its high flexibility compared to other typical formulations requiring for edges (Caselles et al., 1997) or restricting the distribution of intensity values (Chan and Vese, 2001) which definitely do not hold true in our application scenario.

Prior knowledge of pixel intensities is exploited by estimating the probability densities for intensities $s$ to belong to class ’Muscle’ ($p_m(s)$) or class ’No-Muscle’ ($p_f(s)$). These probabilities are estimated by calculating the min-max-normalized histograms of the ’Muscle’ ($p_m(s) \sim h_m(s)/\text{volume(m)}$) and the ’No-Muscle’ class ($p_f(s) \sim h_f(s)/\text{volume(f)}$) relative to the respective volume. For estimating $p_m$ and $p_f$, only data of the same fat infiltration category is considered. The level set function is denoted by $\phi(x)$ and a corresponding heavyside step function is denoted by $H_\phi(x)$ (for better readability $H_\phi$).

The energy function (adapted from (Cremers et al., 2006b)) is as follows
\[ E_{LS}(\phi) = -\lambda_i \int H_\phi \log p_m(s) + (1 - H_\phi) \log p_f(s) dx + \lambda_s E_{smooth}(\phi), \]

where $E_{smooth}$ describes a smoothness term, which is based on the gradient of $H_\phi$ (Cremers et al., 2006b). The outer surface of the initialisation of $\phi(x)$ is set to a dilated background area (3D circular structuring element (seven voxels), two dilations) to improve the fit of the initialization. Dark areas (e.g. vessels) extracted by the GMM approach and the bone are removed from the initialisation.

2.3.3. Shape-Prior Level Set (LS-SP)

A statistical shape prior is incorporated into the level set formulation, which is based on (Cremers et al., 2006a). No registration of the image to the training set is needed, because shapes are aligned intrinsically. The energy function of Eq. (1) is expanded by an additive term $E_{SP}$, which includes the probability of a shape $P(\phi)$:
\[ E(\phi) = E_{LS}(\phi) + \lambda_{sp} E_{SP}(\phi) = E_{LS}(\phi) - \lambda_{sp} \log P(\phi) \]

In this formulation, the shape prior is weighted by the factor $\lambda_{sp}$. The shape probability is determined by a Gaussian kernel density estimation.
available and a distance function between shapes $d(\phi_i, \phi_j)$ is defined, $P(\phi)$ is calculated as
\[
P(\phi) = \frac{1}{N} \sum_{i=1}^{N} \exp \left( -\frac{d^2(\phi_i, \phi_j)}{2\sigma^2} \right).
\] (3)

Note that the standard deviation $\sigma$ is determined automatically (Cremers et al., 2006a).

The shape distance $d^2(\phi_i, \phi_j)$ is defined to be the non-overlapping areas of the shapes $\phi$ and $\phi_i$. Before calculating $d^2(\phi_i, \phi_j)$, the shapes have to be aligned based on geometric features to obtain invariance to natural variability as far as possible. Due to the consistent setting during MRI examination (e.g. without strong rotations), in our case only the mean is taken into account, increasing computational efficiency and providing a translation invariant shape prior. The transformation $T(x, \phi)$ can be expressed by
\[
T(x, \phi) = x + \mu_{\phi} \quad \text{with} \quad \mu_{\phi} = \int H_{\phi}(x)dx.
\] (4)

The transformed shape $\phi(T(x))$ is used to compute the distance function
\[
d^2(\phi_i, \phi_j) = \int \left( H_{\phi_i}(T(x, \phi)) - H_{\phi_j}(x) \right)^2 dx.
\] (5)

For further details, the reader is referred to (Cremers et al., 2006a).

2.3.4. Graph Cut (GC)

The image term of the energy function for graph cut segmentation is equal to the level set formulation (Eq. (1)). The smoothness term is adopted from (Boykov and Funka-Lea, 2006; Quispe and Petitjean, 2015), where the object boundary is attracted by edges. Before calculating the smoothness, the image is low-pass filtered with a Gaussian filter ($\sigma = 1$) for noise suppression. In order to reduce the computation time, a six-connected neighbourhood is considered for the 3D-images.

2.3.5. Shape-Prior Graph Cut (GC-SP)

The insertion of shape information to the graph cut method differs distinctively from the level set approach. In this case, only 2D-images are considered. The corresponding training shapes are obtained by selecting for each ground truth segmentation in the training set the slice with the closest relative z-coordinate (which can be interpreted as nearest-neighbour interpolation). Subsequently, the training shapes
are pre-aligned, such that the mean of the bones are congruent. A probabilistic atlas map \( p'_m \) is created by estimating the coordinate-wise probability of muscle tissue. This probability is estimated based on the occurrences in the prevalent training data by computing the coordinate-wise fraction between positive ('Muscle') samples and overall samples. With \( f_n \) denoting a binary variable indicating the class membership of node \( n \) and \( E_{GC}(f_n) \) denoting the previous energy function without shape prior, the energy formulation changes to

\[
E(f_n) = E_{GC}(f_n) + \lambda E_{SP}(f_n)
\]

with \( E_{SP}(f_n) = \begin{cases} 
-\log(p'_m(n)) & \text{if } f_n = 1 \ ('\text{Muscle}') \\
-\log(1 - p'_m(n)) & \text{if } f_n = 0 \ ('\text{No-Muscle}').
\end{cases} \) \label{eq:7}

Note that if \( p'_m(n) \) is smaller than 0.5, the formulation of Eq. \(\text{(7)}\) favours a classification to the 'No-Muscle'-class. This behaviour is desirable in certain situations, since it can prevent large negative areas from belonging to the muscle group. Therefore, the probability map \( p'_m \) is rescaled by a piecewise-linear function \( g_{p_n}(p) \) to move the 'neutral probability' from originally 0.5 to \( p_n \):

\[
g_{p_n}(p) = \begin{cases} 
0.5 \frac{p}{p_n} & \text{if } p \leq p_n \\
0.5 \frac{p + 0.5 - p_n}{1 - p_n} & \text{if } p > p_n
\end{cases} \]

\label{eq:8}

2.4. Evaluation Details

For assessing the segmentation outcomes, our evaluation was performed based on two different metrics consisting of the DSC as well as a second domain specific metric, referred to as maximum fat fraction error (MFFE).

The fat fraction, representing the final disease marker, is determined by choosing a threshold, which separates fat and muscle tissue within the segmented muscle volume. For sake of general validity, we did not focus on a specific threshold selection approach (in fact this step is often conducted manually), but we rely on a metric computing the upper bound of the fat fraction error.
Based on the normalised histogram $h(s)$ of the segmented muscle volume $s$, consisting of intensity values, the fat fraction $FF$ can be computed by

$$FF_\gamma(s) = \int_{\gamma}^{s_{\text{max}}} h(s) ds$$

after selecting a threshold $\gamma$. Instead of selecting a certain threshold (manually or automatically), we are interested in the worst-case error (Borovskikh, 1979) considering the fat fraction between the ground truth volume ($s_{\text{GT}}$) and an automated segmentation output $s_A$ which is obtained by

$$MFFE(s_A) = \max_\gamma |FF_\gamma(s_A) - FF_\gamma(s_{\text{GT}})| .$$

For parameter optimization, a grid search was implemented in order to determine the best combination. To avoid any bias, for all experiments a k-fold cross-validation was executed with $k = \min\{10, \# \text{ patients per category}\}$. Parameter evaluation was performed individually for the four categories (Table 1).

The number of iterations of the level set approach was set to 250 and the range of the weighting factors was chosen accordingly as follows: $\lambda_s \in [0.2, 0.5, 1.0]$, $\lambda_i \in [0.15, 0.25, 0.5, 0.75, 1]$ and $\lambda_{sp} \in [0.3, 0.5, 1.0]$. Due to the high computing effort, $\lambda_{sp}$ was optimized employing cross-validation only for the 'hard' category and the optimum value was utilized for the other data sets.

Considering the graph cut approach, the evaluated parameters consisted of curvature weight $\lambda_s \in [0.001, 0.002, 0.05, 0.1, 0.2, 0.5]$, low-pass filtering weight $\sigma \in [1, 2]$ (Gaussian Kernel), shape prior weight $\lambda_{sp} \in [0.1, 0.2, 0.5, 0.7, 1]$ and neutral probability $p_n \in [0.2, 0.3, 0.4, 0.5]$. Due to the distinctly higher efficiency of the graph cut approach, $\lambda_{sp}$ was optimized employing cross-validation for all data sets.

3. Results

For evaluation, we individually assessed the segmentation performance (i.e. the DSCs) as well as the impact on the final fat fraction (MFFE). Figure 3 shows the obtained DSCs for the five segmentation settings. Considering the categories without visible fat infiltration (‘easy’, ’moderate’), the obtained DSCs were between 0.88 and
0.95. The average DSCs, for the different segmentation methods, were between 0.90 and 0.93. The 'moderate' category exhibited similar average performances (between 0.89 and 0.91) with few outliers showing rates below 0.90. Considering the 'hard' data set, outcomes between 0.34 and 0.91 were obtained strongly depending on the segmentation method. The best measures were achieved in case of the GC-SP method (0.85 on average) followed by LS-SP (0.83) and LS (0.82). With GC and GMM mean DSCs of 0.67 and 0.72 were obtained.

Figure 4 shows the effect of the performed segmentation on the error during calculating the fat fraction. For the classes 'moderate', 'easy' and 'Healthy' we generally noticed small errors with mean errors always below 0.07. In the case of the 'hard' category, the mean error was between 0.06 and 0.33. The best outcomes were achieved with the same methods as for the evaluation of DSCs.

To assess whether differences in performance are statistically significant, Wilcoxon signed rank tests were performed. Due to the high and consistent performances in the case of the categories 'healthy', 'easy' and 'moderate', we focussed here on the 'hard' category. Comparing the outcomes of the different approaches, p-values are provided.
in Fig. 5. Significance could be shown between GC-SP, providing the best rates, and GMM, LS-SP and GC in case of the DSCs. The level of significance was generally higher for the MFFEs (Fig. 6(b)) where GC-SP statistically significantly outperformed all other techniques (i.e. the p-value is always less than 0.05).

For visual assessment, example segmentations for one patient of the 'hard' data set are shown in Fig. 5.

4. Discussion

For this experimental study, a set of 37 3D-T1-weighted MRI scans of the thigh was divided into four groups based on the visual impression of the fat infiltration in order to facilitate detailed performance interpretation. A local as well as a global statistical energy minimizing method were implemented and combined with statistical shape models.

Considering the easier-to-segment categories ('healthy', 'easy' and 'moderate'), we noticed that all investigated segmentation approaches exhibit similar performances independent of the performance measure (DSC or MFFE). Although the LS-SP ap-
Figure 5: Example segmentations for different slices of one patient of the 'hard' data set compared to the ground truth segmentation (GT). Especially GMM and GC completely failed to segment severely affected tissue. The LS method misses details by segmenting smooth structures due to a high smoothness term. Shape prior-based approaches suffer especially in case of fat infiltrations in boundary regions.

The approach was not completely optimized for these data sets (see Sect. 2.4), we did not notice distinct decreases in performance compared to the other methods.

Completely different outcomes were obtained in the case of the difficult-to-segment class ('hard'). Here we notice that the DSCs as well as the MFFEs, strongly depend on the applied segmentation approach. The methods incorporating statistical shape models (LS-SP, GC-SP) show distinctly better outcomes than the version without shape prior knowledge. This effect is particularly pronounced for the graph cut methods: Without shape prior searching for the global minimum energy (GC) is inferior compared to searching for a local minimum (LS). This is supposed to be due to cases showing 'non-affected' tissue in border areas and distinct fat infiltration in central regions (Fig. 5, second row). Reasonably initialized level set-based approaches remain at the real edge, although the global minimum solution could correspond to a completely different segmentation.

However, this positive effect of the level set approach is not maintained after introducing the shape model. Adding this model, graph cut techniques are inhibited to produce implausible shapes (Fig. 5) which, in combination with finding the global optimum, even outperforms the shape prior based level set method.

Considering the MFFEs, we noticed that by introducing shape models, the tradi-
Figure 6: P-values for the 'hard' category obtained by Wilcoxon signed rank tests: The null hypothesis is that the data in two distribution (DSCs or MFFEs of two different segmentation approaches) comes from populations with equal medians.

Although generally assessing the shape model-based approach as highly effective considering two metrics, it should also be outlined that a proper segmentation of especially fine structures still often fails (Fig. 5). Due to the small area of these structures, however, the effect of a mis-segmentation of these areas is small. In the case of GC-SP, we assume that this effect is due to the relatively straightforward shape model which is not capable of modelling fine details. Considering the LS-SP method, we suspect that, if larger training data (which is rather small compared to the high degree of variability) would be available, this behaviour could be improved. In spite of the relatively small effect on the final measures, we suppose that the final scores can thereby be improved even further.

Interestingly, the finally relevant (and domain specific) MFFE is improved more distinctly (Fig. 4 vs. Fig. 3) which is also confirmed by decreasing p-values (Fig. 6). We expect that this is due to the fact that the statistical shape prior prevents completely degenerated segmentations which show potential to strong errors and thereby weak MFFEs and DSCs. In case of a rather too smooth contour on the other hand, we expect
that, although the DSCs are not perfect, errors can be balanced out leading to low MFFEs.

An inherent restriction (which also motivated us to introduce the domain-relevant MFFE measure) is given by the rather low image resolution. Although segmentation methods are able to deal with low resolution images, segmentation accuracy is limited by the resolution of the image data. Additionally, voxels in border regions showing a gray value in between muscle and fat are highly difficult to classify even for human observers which define the ground truth for training and for evaluation. This problem definitely constitutes a challenge for further research in this as well as in related fields dealing with segmentation of MRI scans. It becomes even more evident if considering the segmentation of smaller muscle structures such as lower legs or spinal muscles. For such data, it might be beneficial to apply super-resolution techniques in advance to further processing to increase the resolution even before acquiring the manual ground truth. This could enhance the final segmentation quality and also the reliability of e.g. the DSC measure if comparing the segmented masks with the ground truth annotations.

5. Conclusion

In conclusion, we notice that the graph cut approach incorporating shape knowledge showed an attractive performance, especially for the hard-to-segment category, but also for the easy-to-segment image data, representing the best approach overall. Especially due to the high fat fraction accuracy (i.e. low MFFEs), we suggest to utilize the proposed combination of a segmentation approach including a statistical shape model for fully-automated segmentation of muscle MRI data sets of patients with neuromuscular diseases.

Future work will investigate the applicability of fully-convolutional neural networks in combination with domain specific data augmentation as well as a combination of neural networks with statistical shape models. Moreover, future work will assess the progression of fat infiltration and muscle atrophy during the disease course of specific neuromuscular disorders.
Acknowledgement

This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

Declaration of interest

Conflicts of interest: none


