This is a pre-print of an article published in **Sympoium on Graphics Patterns and Images**. The final authenticated version is available online at: https://doi.org/10.1109/SIBGRAPI.2019.00019

Fast and smart segmentation of paraspinal muscles in magnetic resonance imaging with CleverSeg

Jonathan S. Ramos[+1], Mirela T. Cazzolato[+], Bruno S. Faiçal[+], Oscar A. C. Linares[+], Marcello H. Nogueira-Barbosa[⋆], Caetano Traina Jr.[+] and Agma J. M. Traina[+]

[+]Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP).
[⋆]Ribeirão Preto Medical School (FMRP), University of São Paulo (USP).

Abstract

Magnetic Resonance Imaging (MRI) is a non-invasive technique, which has been employed to detect and diagnose many spine pathologies. In a Computer-Aided Diagnosis (CAD) context, the segmentation of the paraspinal musculature from MRI may support measurement, quantification, and analysis of muscle-related pathologies. Current semi-automatic segmentation techniques require too much time from the physicians to annotate all slices in the exams. In this work, we focus on minimizing the time spent on manual annotation as well as on the overall segmentation processing time. We use the mean absolute error between slices in order to minimize the number of annotated slices in each exam. Moreover, we optimize the manual annotation time by estimating the inside annotation based on the outside annotation, while the competitors demand the annotation of inside and outside annotation (seeds). The experimental evaluation shows that our proposed approach is able to speed up the manual annotation process in up to 50% by annotating only a few representative slices, without loss of accuracy. By annotating only the outside region, the process can be further speed up by another 50%, reducing the total time to only 25% of the previously required. Thus, the total time spent on manual annotation is reduced by up to 75%, and, since the human interaction is greatly diminished, allows a more productive and less tiresome activity. Despite that, our proposed CleverSeg method presented accuracy similar or better than the competitors, while managing a faster processing time.

Key-words: *3D vertebrae reconstruction; magnetic resonance imaging; Balanced Growth.*

1 Introduction

Back pain is one of the most common complaints worldwide. In general, it is related to spinal disease and can cause a significant loss of function and compromise the quality of life. Surgical spinal treatments have grown with the population aging, which require accurate diagnoses to avoid complications [1]. Magnetic Resonance Imaging (MRI) exams provide meaningful information to the detection and diagnosis of many spine pathologies and, at the same time, it is not harmful to the patient (do not use ionizing radiation) [2, 3].

The segmentation of the paraspinal musculature in the context of Computer-Aided Diagnosis (CAD) may allow a faster and more objective analysis of the muscle condition, supporting in the characterization and quantification of back muscle-related problems [4]. Many works in the literature have shown the disadvantage of dealing only with discrete image slices (2D), which can generate a loss of relevant information for precise measurements and diagnosis [5]. Accordingly, 3D segmentation approaches may assist in better visualization and analysis of the muscle structures in a reliable way.

Integrating automated procedures for reliable segmentation of selected muscles may reduce the labor-intensiveness associated with manual methods [6]. Manually segmenting many slices of a single 3D exam is also time-consuming. On the other hand, computer methods can now reduce inaccuracies occurred or aggregated due to subjective judgments, inter and intra-subjective variability [7, 8].

Fast and accurate segmentation plays a significant role and may assist the medical specialist in surgical planning and evaluation of suitable treatments [7, 9, 10]. One of the advantages of semi-automatic segmentation is the use of the specialists' knowledge, gained over the years, to improve the results of computer methods. Performing semi-automatic segmentation assists the physicians and specialists, leads to time savings, and reduces interpretation errors [11].

Semi-automatic segmentation serves as an essential tool for many tasks, whereas clinicians, as well as scientists, would strongly benefit from automated segmentation methods [12, 13]. Examples of such tasks include the extraction of semantic and agnostic features, the application of machine learning algorithms for the classification of anomalies, and Content-Based Image Retrieval (CBIR) techniques to obtain semantically similar historical data [14–17]. The muscle segmentation can be meaningful when combined with interactive tools, allowing the training and education of new radiologists [18], as students can learn how to segment muscles correctly, and to detect spine pathologies [19]. Visual tools may also help physicians and professors to evaluate and determine whether a student is ready to proceed to further tasks related to a specific pathology. In practice, the visualization of 3D human structures can also be used for the simulation of surgical procedures [20].

In [21], the authors describe the difficulties and challenges related to the problem of segmenting anatomical structures from MRI exams. They include the presence of ambiguous structure boundaries, the resemblance in the structures, insufficient contrast, low spatial resolution, intensity in-homogeneity, and image dimensionality of 3D exams.

In this work, we propose CleverSeg, a method that takes advantage of the motto "growth in unity is strength" to achieve a better-delineated segmentation. We manage the iterations to reduce the processing time. We propose the annotation of only a few slices

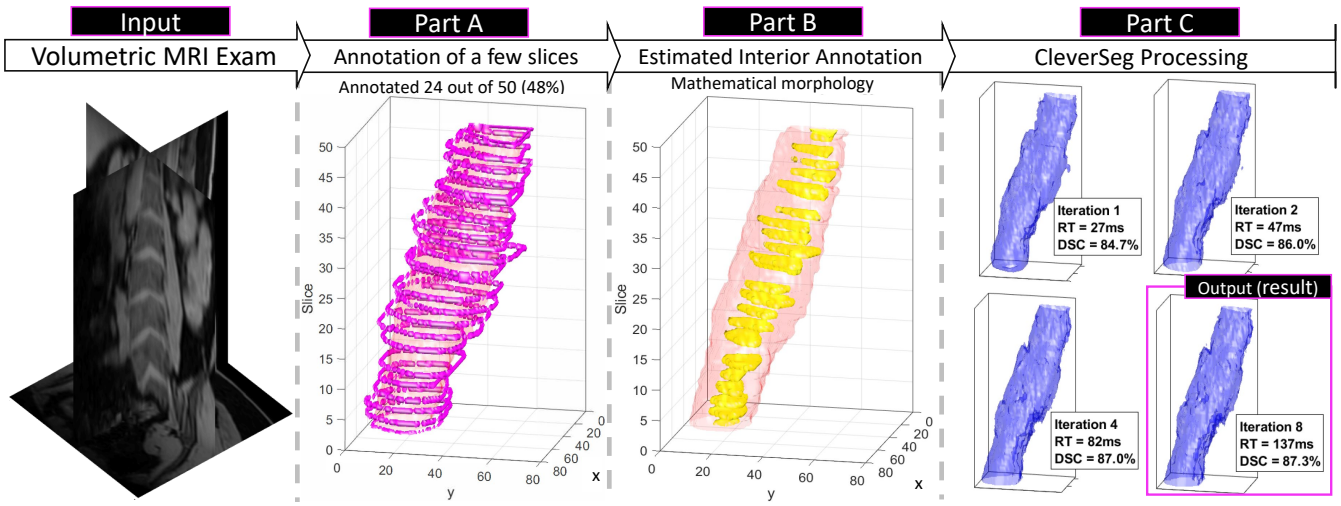


Figure 1: Proposed pipeline: ground-truth in red, exterior annotation in magenta, estimated interior annotation in yellow and CBG segmentation result in blue.

and the estimation of the interior annotation of each muscle. Moreover, CleverSeg uses a simple annotation (sloppy-like), which does not require detailed annotation, i.e., works with imprecise annotations. As a consequence, the time spent on manual annotation is greatly minimized, allowing a more productive and less tiresome activity.

The remainder of the paper is structured as follows. First, Section 2 presents the background and related works. Then, Section 3 details CleverSeg for the segmentation and reconstruction of paraspinal muscles in volumetric MRI. Next, Section 4 explores the materials and methods used in our experimental design. After that, Section 5 details the experimental design, results and discussion. Finally, Section 6 draws the conclusions.

2 Background and related works

There is an association between imaging parameters of the paraspinal muscles such as cross-sectional area size, shape, density, and volume with spinal degeneration and low back pain. As a consequence, measuring the paraspinal muscles in 3D is a crucial step in the analysis of the muscle conditions associated with low back pain [22]. Manually segmenting a large exam (several slices) is too time-consuming, thus automatic and semi-automatic approaches are highly attractive due to the reduction of labor-intensiveness associated with the manual approach [23].

Several fully automatic vertebrae segmentation methods are reported in the literature [5, 6, 21, 24]. Nonetheless, they take too much processing time or do not produce reasonably precise results, which may not suit clinical practice [5]. Semi-automatic approaches, in general, take considerably faster processing time and produce reasonably precise segmentation results. However, they often require too much time from the radiologists on manual annotation of a few or all the slices in the exam [23].

More recently, two novel semi-automatic approaches called Balanced Growth (BGrowth) [25] (for 2D images) and 3D Balance Growth (3DBGrowth) [23] (for 3D images) were proposed for the segmentation of vertebral bodies. BGrowth has presented promising results for the segmentation of crushed vertebral bodies in a single slice at a time, considering malignant (metastasis) and benign (osteoporosis) as well. 3DBGrowth works well for the segmentation of vertebral bodies in volumetric MRI scans. Briefly, both approaches balance the weights along the growing path of a region, so that small intensities transitions are better delineated.

Another semi-automatic segmentation method is GrowCut [26], which has been one of the most employed methods on the segmentation of medical imaging. GrowCut is based on cellular automata (analogous to a bacteria growth in biology) and works as a region-growing approach with an interactive labeling procedure [26]. Also, a faster but less accurate version of GrowCut, Fast GrowCut [27], has been widely used for segmentation of medical images in the 3D Slicer[2] open-source software [28]. 3D Slicer is a framework which provides a friendly Graphical User Interface (GUI) and allows interactive operations and visualization, which is especially helpful for semi-automatic segmentation approaches [27].

Currently, the results achieved by BGrowth and 3DBGrowth surpass the other methods from the literature, including GrowCut, presenting promising segmentation results, even with very simple/sloppy annotation (seed points). However, due to the balancing approach, 3DBGrowth may require more iterations for the segmentation of larger exams (more slices). Furthermore, 3DBGrowth was tested with only 18 slices (in average) on each exam, and a maximum number of iterations of 50 [23]. Besides, the slope coefficient defined by the authors may heavily rely on the physical spacing between slices.

3 The Proposed Method

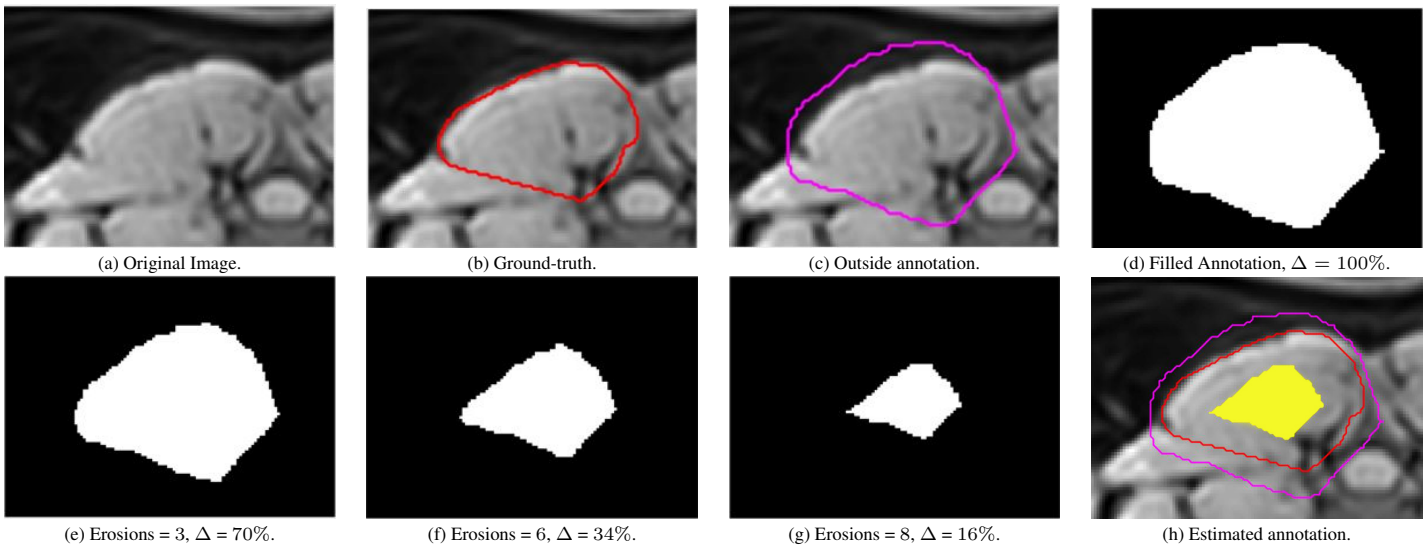
Placing (annotating) seeds appropriately in MRI data is a crucial initial step to produce accurate paraspinal muscles segmentation. Nevertheless, due to the 3D nature of MRI data and the complex structure of the human spine, it becomes a very difficult and tiresome task.

In this work, we present CleverSeg, which focus on minimizing the specialist's effort to segment and reconstruct MRI exams built on 2D slices. To this end, we contribute in three main aspects, as the pipeline shown in Fig. 1. The three main parts of CleverSeg are detailed as follows:

A – *Annotation of a few slices*: we use the Mean Absolute Error (MAE) to verify slices that look alike and, therefore, are not required to be manually annotated. The error is calculated slice vs. slice and is not dependent on the physical spacing between slices.

B – *Estimation of the interior annotation*: given the initial outside annotation of each slice annotated in part A, we estimate the inside annotation using mathematical morphology. As a consequence, the time spent on manual annotation may be greatly reduced.

[2] https://www.slicer.org/



(a) Original Image. (b) Ground-truth. (c) Outside annotation. (d) Filled Annotation, $\Delta = 100\%$.

(e) Erosions = 3, Δ = 70%. (f) Erosions = 6, Δ = 34%. (g) Erosions = 8, Δ = 16%. (h) Estimated annotation.

Figure 2: Sample slice with ground-truth (red) and exterior annotation (magenta) and estimated interior annotation in yellow using a Δ threshold of 20%.

C – *Fast and effective semi-automatic segmentation method*: our proposed CleverSeg method works in a faster and smart way, using only a few iterations. CleverSeg efficiently propagates the annotated slices in parts A-B into non-annotated slices. Therefore, the processing time speeds up while keeping high accuracy.

In the next Subsections, we explore the three main aforementioned contributions. CleverSeg method is publicly available[3] as an extension for 3D Slicer [28].

3.1 Outside annotation

To choose the most representative slices to be annotated, we use the MAE to compute the difference between two paired variables. The MAE (M) is also applied to measure the difference between two 2D images S_1 and S_2, as shown in Eq. 1. Here, S_1 and S_2 represent two distinct slices of the volumetric MRI exam V. Both slices have size n, which corresponds to the total number of pixels from each slice. The i-th entries in S_1 and S_2 are represented by $s_{1,i}$ and $s_{2,i}$, respectively. Note that the closer $M(S_1, S_2)$ is to zero, the more similar the images are [29, 30].

$$M(S_1, S_2) = \frac{1}{n}\sum_{i=1}^{n} |s_{1,i} - s_{2,i}| \tag{1}$$

Considering a volumetric exam $V = \{S_1, S_2, .., S_i, .., S_n\}$, in which S_i represents the i-th slice in the exam, similar sequential slices can be avoided from the manual annotation process. The most representative slices may be selected in a bottom-up manner in three steps, as follows.

Step 1: define S_1 to be manually annotated and set S_1 as the initial slice for comparison with the next slice S_2.

Step 2: if the error $M(S_1, S_2)$ is lower then a threshold η, then compare S_1 with the next slice S_3. This process repeats until the i-th slice S_i, such that $M(S_1, S_i) \geq \eta$. Then, define S_i as the initial slice for comparison and go back to Step 1.

Step 3: repeat Steps 1 and 2 until the final slice S_n is reached. The last slice is always set to be manually annotated.

Note that, the first (S_1) and last (S_n) slices are always set to be manually annotated and the number of non-annotated slices between them may vary according to the threshold η. The value of the threshold η can be set by the user.

[3] https://github.com/JonathanRamos/SlicerCleverSeg.git

For analysis purposes, we define a Percentage of Annotated Slices (PAS), which is the number of annotated slices ($\#annotatedSlices$) divided by the total number of slices in the exam with muscle content ($\#slices$):

$$PAS = \frac{\#annotatedSlices}{\#slices}$$

3.2 Inside annotation estimation

Given the outside annotation of the i-th slice (S_i) as exemplified in Fig. 2c (in magenta), we estimate the inside annotation using morphological operations [31] in four steps:

Step I: represent the manually annotated outside region of S_i as a 2D binary mask k_i. Considering that the region in k_0 is always a closed boundary, fill this boundary with ones (white pixels), as exemplified in Fig. 2d.

Step II: set k_0 as the initial interior annotation and t_0 as the total number of white pixels in k_0 and apply an erosion operation over k_0 using a 5×5 square structuring element, which results into a new mask k_1. As a consequence, the number of white pixels in k_1 is reduced to t_1. The 5×5 square structuring element allows a smooth erosion without losing the main shape of the region.

Step III: apply the same erosion over k_1, which results in k_2 with t_2 white pixels. Repeat the erosion process until the i-th iteration, resulting in the mask k_i with t_i white pixels. The erosion process stops when the percentage of remaining white pixels in the eroded mask (k_i) associated to the initial mask k_0 is lower than a threshold Δ:

$$\Delta > \frac{t_i}{t_0} \tag{2}$$

Step IV: finally, use the k_i mask as the inside annotation. The whole erosion process is depicted in Fig. 2e to Fig. 2g, in which Δ was set to 20%, yielding the estimated inside annotation shown in Fig. 2h (in yellow).

3.3 Clever Segmentation – CleverSeg

In our proposed CleverSeg method (Algorithm 1), for the sake of simplicity, initially, we segment grayscale volumes into foreground and background. However, the algorithm works for more than two

regions, and the difference between voxels intensities may be easily adapted for color images.

Considering a volume V and a matrix L of corresponding annotations/seeds, both with dimensions $M \times N \times Z$, representing the number of rows, columns and slices, respectively. The total number of voxels in V is represented by n. The entries of L may have values 0 (unlabelled), 1 (background or outside annotation) or 2 (foreground or inside annotation).

The method is divided in three steps:

Step A: a weighted matrix W is initialized, so that every entry $\forall w_i \in W$ is filled with ones for corresponding labelled entries $l_i \in L$ (inside/outside annotation) and zeros otherwise as in Eq. 3 (lines 1 to 2). The maximum voxel intensity is calculated/represented as m (line 3).

$$w_i = \begin{cases} 1.0 & \text{if } l_i \neq 0, \\ 0.0 & \text{otherwise.} \end{cases} \quad (3)$$

Step B: every voxel $v_i \in V, i = \{1, 2, 3, ..., n\}$ and each of its 26 neighbors $v_i^j \in V, j = \{1, 2, 3, ..., 26\}$ is analyzed (line 4) as follows. The absolute difference between the voxel intensity (v_i) and its neighbor's (v_i^j) is calculated and subtracted by the maximum voxel intensity (m). The result is represented as h (line 5). A strength s is calculated (line 6) normalizing h by m and multiplying with the current voxel strength w_i.

Step C: if the difference between s and the neighbor's current strength w_i^n (line 7) is greater than a threshold (θ), then, the strength w_n^i is averaged with the new strength s (line 8) and its label l_i^j is updated (line 9).

Input: Image V and annotation/labels matrix L.
Output: Segmented image (grown regions in L).

```
/* Initialization                                          */
1 if l_i ≠ 0, ∀l_i ∈ L then
2     w_i ← 1.0
3 m ← max v_i, ∀v_i ∈ V, i = {1, 2, 3, ..., n}
  /* For every voxel and its 26
     corresponding neighbors                               */
4 for ∀v_i ∈ V and ∀v_i^j ∈ V, j = {1, 2, 3, ..., 26} do
5     h ← m − |v_i − v_i^j|
6     s ← w_i × h/m
7     if (s − w_i^j) > θ then
8         w_i^j ← mean(s , w_i^j)
9         l_i^j ← l_i
```

Algorithm 1: CleverSeg method overview.

Steps A and B repeat for a maximum number of iterations or until the algorithm converges. The threshold θ avoids iterations that do not contribute to a better segmentation. For example, it avoids balancing (averaging) values that only change the third or fourth decimal place.

4 Materials and methods

In order to evaluate the performance of our proposed CleverSeg method, due to space limitations, we present only a meaningful dataset of lumbar muscles. Next, we compare our method with the state-of-the-art semi-automatic segmentation techniques, such as Balanced Growth and GrowCut, considering default parameters settings for each method. Then, we consider the measures of Jaccard Coeficcient, Dice Score, Hausdorff's Distance and F-measure to analyze the segmentation results. To further validate the results, we employ statistical testing. Table 1 shows a summary of acronyms used throughout this work and Section 4.5 explores the computational setup. We highlight that no deep-learning approaches were applied due to the limited number of available exams.

Table 1: Summary of symbols/acronyms used in this work.

Acron.	Description	Acron.	Description
DSC	Dice Score Coeff.	3DBGrowth	3D Balanced Growth
JAC	JAccard Coeff.	CleverSeg	Clever Segmentation
HD	Hausdorff's Distance	FM	F-Measure
RT	Running Time (processing time)		
PAS	Percentage of manually Annotated Slices		
ANT	ANnotation Time (time spent in manual annotation)		
G	Ground-truth		
S	Segmentation yielded by a semi-automatic method		

4.1 Image Dataset

In the work [12], the authors present a manually segmented dataset, called *MyoSegmenTUM spine*. This dataset contains segmented lumbar muscle groups and vertebral bodies, from MRI scans of 54 healthy volunteers. Each exam contains the muscles erector spinae left and right as well as psoas left and right muscles with corresponding manual segmentation. Summing up, there are $54 \times 4 = 216$ muscles. The exams have an average resolution of $334 \times 334 \times 67.4 \pm 135 \times 135 \times 5.02$ voxels and a spatial resolution of $1 \times 1 \times 3.6 \pm 0 \times 0 \times 0.5$ mm. In order to assure the best conditions to all segmentation algorithms, the grayscale intensities of the exams are normalized into 256 gray levels (8 bits/pixel):

$$V = 255 \times \frac{V - \min v_i}{\max v_i - \min v_i}, \forall v_i \in V, i = \{1, 2, 3, ..., n\}$$

in which V represents the data volume of the exam, with resulting entries v_i (voxel intensities) within [0, 255]. The volumes are not normalized to isotropic resolution to avoid adding noise to the image and manual segmentations. We employed *MyoSegmenTUM spine* in the validation of our method, as it provides the ground truth to support testing and refining computer vision methods and is fully available at [12].

4.2 Segmentation algorithms and parameters settings

In order to evaluate our proposed method for the segmentation of spinal muscles, we compared it with 3DBGrowth [23] and GrowCut [26]. Since Fast GrowCut presents a lower accuracy than Growcut [27], it was not considered in the analysis. The maximum number of iterations was set to 500 for all algorithms considered in the analysis. However, in general, the algorithms take less than 500 iterations to converge.

For CleverSeg, the threshold θ, which controls the approximate roundness' of the averaged weights during the balancing of region expansion, was set to 1% and avoids averaging values on the third decimal place. In order to allow a simple or sloppy-like annotation, the Δ threshold was set to a small value, 20%. Thus, imprecise external annotation do not compromise the estimated internal annotation.

For the η threshold, we considered an initial value of 0, increasing by 1.5, up until 10.5, which sums up to 8 thresholds.

4.3 Comparison measures

We considered four well-known measures to compare the resulting segmentation yielded by a segmentation method (S) and the ground-truth (G) as follows.

Jaccard Coefficient (JAC): calculates the intersection of the manual (G) and semi-automatic (S) segmentation, and divides it

by the union of them as in Eq. 4. This indicates the similarity between the segmentations, in which 0 indicates no similarity and, the closer JAC is to 1, the more alike the segmentations are [32].

$$JAC(G,S) = \frac{|G \cap S|}{|G \cup S|} \tag{4}$$

Dice Score Coefficient (DSC): measures, in voxels, the spatial overlap of several segmentations of the same object, i.e, quantifies the overlap degree between two segmented objects, as in Eq. 5. A DSC close to 0 indicates very low overlap, while a DSC closer to 1 indicates a higher overlap [33, 34].

$$DSC(G,S) = \frac{2 \times |G \cap S|}{|G| + |S|} \tag{5}$$

Hausdorff's Distance (HD): indicates how far away (in voxels) G and S are, as in Eq. 6. A HD of 0 indicates comparable segmentations [35].

$$\begin{aligned} HD(G,S) &= \max\{mm_1, mm_2\} \\ mm_1 &= \max_{g_i \in G} \left(\min_{s_i \in S} \{d(g_i, s_i)\} \right) \\ mm_2 &= \max_{s_i \in S} \left(\min_{g_i \in G} \{d(s_i, g_i)\} \right) \end{aligned} \tag{6}$$

in which d denotes the Euclidean distance [36]:

$$d(g_i, s_i) = \sqrt{(s_i^x - g_i^x)^2 + (s_i^y - g_i^y)^2 + (s_i^z - g_i^z)^2}$$

F-measure (FM): calculates the harmonic mean between precision (P) and recall (R) as in Eq. 7.

$$FM = 2 \times \frac{P \times R}{P + R} \ , \tag{7}$$

in which P and R are defined, respectively, as: $P = TP/(TP + FP)$, $R = TP/(TP + FN)$, considering that True Positive (TP) represents the number of voxels correctly segmented as part of the foreground (G), False Positive (FP) represents the number of voxels miss-segmented as belonging to G and False Negative (FN) represents the number of voxels incorrectly segmented as part of the background.

4.4 Statistical tests

According to [25], if the data present several similar values, the Kolmogorov-Smirnov [37] is the most suitable normality test. Then, in order to analyze if there are significant statistical differences, the Wilcoxon [38] test may be employed. In this test, the null hypothesis is that data from two paired sample groups were selected from populations having the same distribution, against the opposite alternative [23].

4.5 Computational setup

Every experiment used a 2.40GHz Intel(R) Core(TM) i7 CPU and 8GB RAM machine, using Matlab(R) version 2018a. To assure the same conditions for all segmentation methods, no pre or post-processing techniques were applied.

5 Experiments, results and discussion

In our experimental design, three main parts are analyzed, as depicted in Fig. 3. First, we analyze the segmentation of each muscle, considering that all slices are annotated on the outside and we estimate the inside annotation. Then, we vary the number of slices annotated, based on the error (η) between slices. Finally, we statistically evaluate the results.

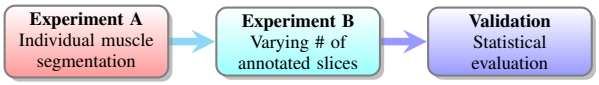


Figure 3: Experimental design.

5.1 Muscle segmentation

We performed the segmentation of each muscle considering the manual annotation of all slices on each MRI exam. Fig.4 shows the average results for all muscles. Note that, CleverSeg presented better results than GrowCut for the measures DSC, FM, JAC and RT, while keeping a similar HD. CleverSeg presented better results for the measures FM, JAC and RT than 3DBGrowth, while keeping comparable DSC and HD.

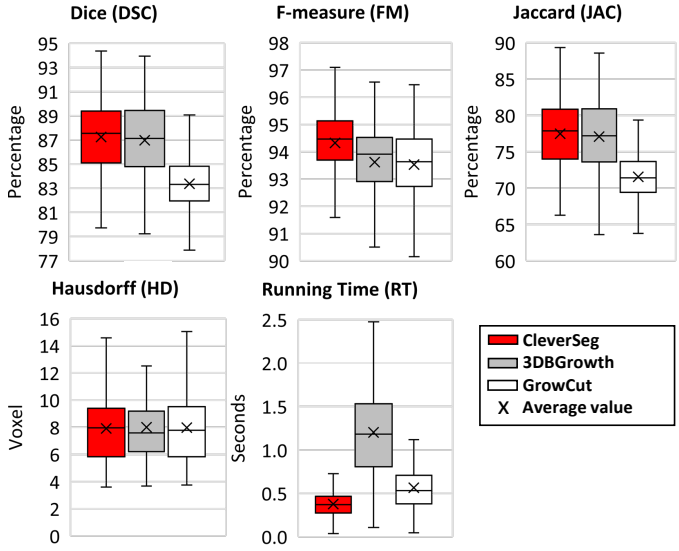
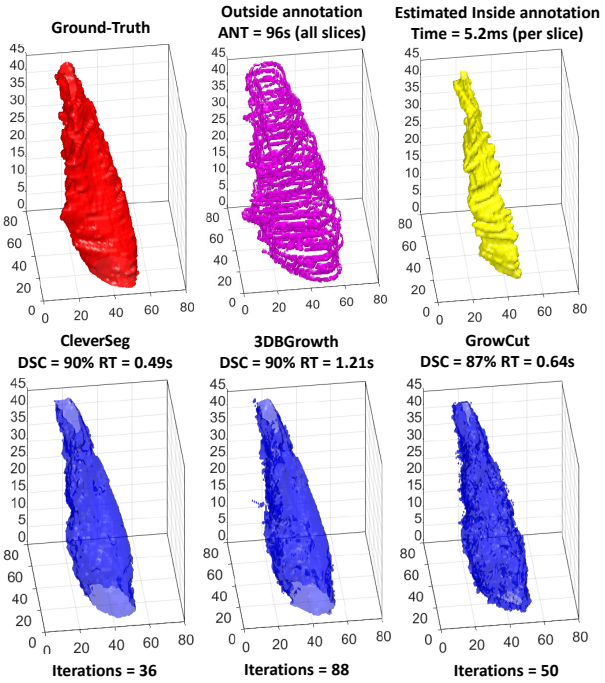


Figure 4: Comparison between the segmentation approaches.

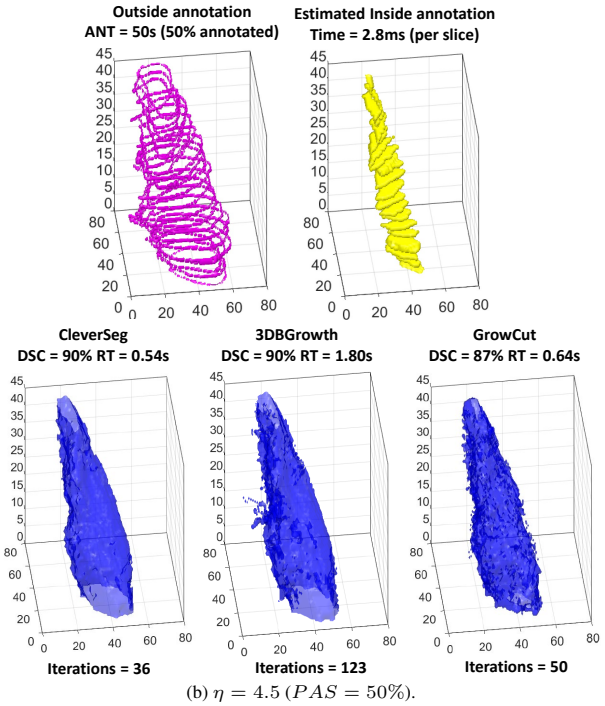
In average, CleverSeg presented a running time of 377ms, while 3DBGrowth took 1202ms and GrowCut 565ms. The number of iterations were 32 ± 6.5, 99 ± 22.6 and 49 ± 13.2 for CleverSeg, 3DBGrowth and GrowCut, respectively. In the original work [23], 3DBGrowth was tested with 18 slices (in average) and a maximum of 50 iterations. In our experimental dataset, in average, there are 67 slices for each exam.

3DBGrowth requires more iterations to converge as the number of slices increases, consequently, increasing the processing time. On the other hand, CleverSeg presented results comparable or better than the competitors, while managing a faster processing time (RT). To better illustrate this, Fig. 5a shows the segmentation results for a single muscle. Note that, CleverSeg presented the fastest running time and a comparable or better DSC than the competitors. For this example, the interior annotation took 96s (ANT) while the estimation of the interior annotation was considerable faster (5.2ms). In average, the interior estimation for all muscles took 4.5 ± 3.6 ms. If the inside annotation was to be manually given, the time spent on manual annotation (ANT) would possible double.

Analyzing the segmentation results in Fig. 5a, GrowCut presented spiculated borders with a lower DSC, while CleverSeg and 3DBGrowth presented smooth borders and equal DSC. However, CleverSeg had the lowest number of iterations, and presented the fastest processing time (RT). To further validate CleverSeg and, at the same time, reduce the annotation time (ANT), on the next experiment we vary the number of slices with exterior annotation for each muscle.

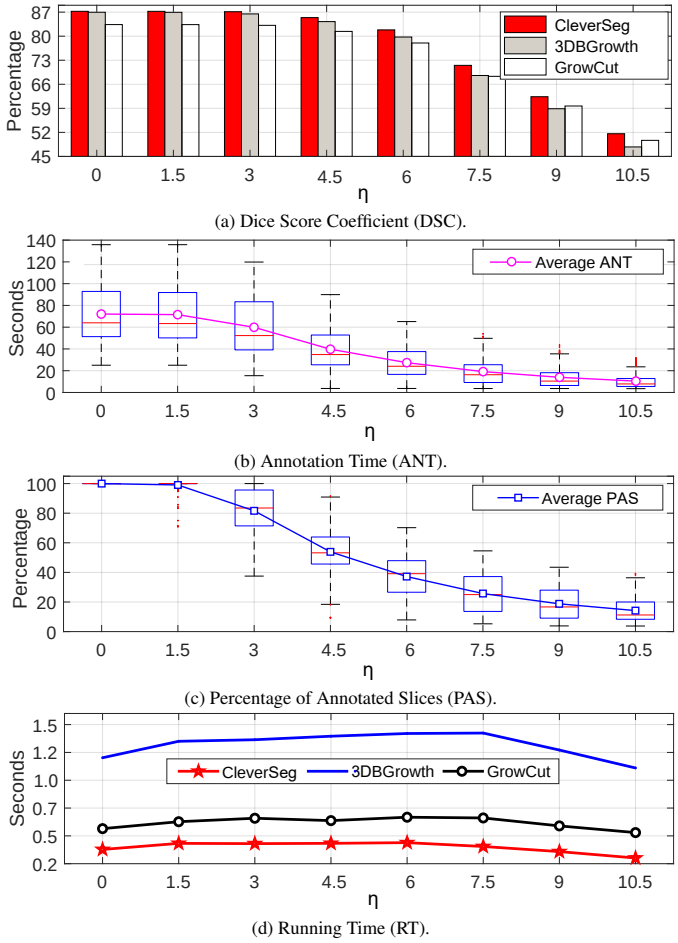


(a) All slices annotated ($PAS = 100\%$).



(b) $\eta = 4.5$ ($PAS = 50\%$).

Figure 5: Segmentation results for a single muscle: exam 52, psoas left.



(a) Dice Score Coefficient (DSC).

(b) Annotation Time (ANT).

(c) Percentage of Annotated Slices (PAS).

(d) Running Time (RT).

Figure 6: Results for varying the mean absolute error (η) threshold.

Table 2: Wilcoxon test results: ✓ means that CleverSeg was significantly better and × means that no significant difference is observed.

CleverSeg against	Meas.	η 0.0	1.5	3.0	4.5	6.0	7.5	9.0	10.5
3DBGrowth	DSC	×	×	×	✓	✓	✓	✓	✓
	JAC	×	×	×	✓	✓	✓	✓	✓
	HD	×	×	✓	✓	✓	✓	✓	✓
	RT	✓	✓	✓	✓	✓	✓	✓	✓
GrowCut	DSC	✓	✓	✓	✓	✓	✓	✓	✓
	JAC	✓	✓	✓	✓	✓	✓	✓	✓
	HD	×	×	×	×	×	×	×	×
	RT	✓	✓	✓	✓	✓	✓	✓	✓

5.2 Varying the number of annotated slices

For this experiment, we use the error η between slices as a threshold in order to find which slices should be annotated. The average results for all muscles are summarized in Fig. 6, and analyzed as follows. The DSC starts dropping at $\eta = 4.5$ (Fig. 6a), while ANT fastly drops at $\eta = 1.5$ (Fig. 6b) along with the PAS slices (Fig. 6c). For the running time, CleverSeg was the fastest method for all thresholds (Fig. 6d).

According to the results reported, $\eta = 4.5$ presented the best trade-off between DSC and ANT. For this threshold, the PAS drops to approximately 50% and the ANT drops from 72 to 40 seconds (almost 2x faster), losing just a tiny bit of DSC (from 87%, 87% and 83% to 85%, 84% and 81% for CleverSeg, 3DBGrowth and GrowCut, respectively). Fig. 5b illustrates this, in which, compared to Fig. 5a, ANT drops almost to a half, while DSC drops 1% for GrowCut. To further validate the results presented herein, in the next section we perform statistical testing.

5.3 Statistical evaluation

As the data for all measures presented several similar values, the Kolmogorov-Smirnov [37] test was applied at the 5% significance level. The null hypothesis was rejected for all measures, which indicates the data do not follow a normal distribution. Therefore, the Wilcoxon [38] test was employed at the 5% significance level.

The Wilcoxon test results are reported in Table 2. Note that, CleverSeg presented significantly better running time (RT) than 3DBGrowth and GrowCut. Compared to 3DBGrowth, CleverSeg presented better results for DSC, JAC and HD from $\eta = 4.5$ to 10.5. In general, CleverSeg presented comparable or significantly better results than the competitors, while always achieving a faster processing time.

6 Conclusion

The semi-automatic segmentation of muscles in larger volumetric MRI exams is a challenging task. In general, too much time is spent on manual annotations of each slice of the exam, both inside and outside the object of interest (muscles in this work). For this reason, allowing a fast and accurate segmentation of slices is crucial in order to obtain a proper 3D reconstruction of the muscle. Aimed at overcoming this issue, we used the mean absolute error to remove not needed slices from the manual annotation process. Furthermore, we estimated the inside annotation based on the outside annotation, not requiring manual inside annotations, while the competitors demand the annotation of both inside and outside seeds.

The experimental results showed that, on average, only 50% of the slices required outside annotations. Moreover, the time spent on overall annotations is 50% faster by using only the outside annotation and quickly estimating the interior annotation with our approach. As a final remark, we highlight that CleverSeg presented better or similar results than 3DBGrowth and GrowCut while managing a statistically significant lower processing time. As a future work, we intend to exploit the segmentation of temporal images sequences.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and grant No.: 0487/17083480, by the São Paulo Research Foundation (FAPESP, grants No. 2016/17078-0, 2017/23780-2, 2018/06228-7, 2018/24414-2), and the National Council for Scientific and Technological Development (CNPq).

References

[1] M. G. Fehlings, L. Tetreault, A. Nater, T. Choma, J. Harrop, T. Mroz, C. Santaguida, and J. S. Smith, "The aging of the global populationthe changing epidemiology of disease and spinal disorders," *Neurosurgery*, vol. 77, no. 1, pp. 1–5, 2015.

[2] M. Rak and K. D. Tönnies, "On computerized methods for spine analysis in MRI: a systematic review," *Int. J. Computed Assisted Radiology and Surgery*, vol. 11, no. 8, pp. 1445–1465, Aug 2016.

[3] Y. X. J. Wang, F. R. Santiago, M. Deng, and M. H. Nogueira-Barbosa, "Identifying osteoporotic vertebral endplate and cortex fractures," *Quant Imaging Medical Surgery (QIMS)*, vol. 7, no. 5, pp. 555–591, Oct 2017.

[4] E. Klupp, B. Cervantes, S. Schlaeger, S. Inhuber, F. Kreuzpointer, A. Schwirtz, A. Rohrmeier, M. Dieckmeyer, D. M. Hedderich, M. N. Diefenbach, F. Freitag, E. J. Rummeny, C. Zimmer, J. S. Kirschke, D. C. Karampinos, and T. Baum, "Paraspinal muscle DTI metrics predict muscle strength," *J. of Magnetic Res. Imaging (JMRI)*, pp. 1–8, 2019.

[5] G. Hille, S. Saalfeld, S. Serowy, and K. Tnnies, "Vertebral body segmentation in wide range clinical routine spine MRI data," *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 93 – 99, 2018.

[6] C. M. Engstrom, J. Fripp, V. Jurcak, D. G. Walker, O. Salvado, and S. Crozier, "Segmentation of the quadratus lumborum muscle using statistical shape modeling," *J. of Magnetic Resonance Imaging (JMRI)*, vol. 33, no. 6, pp. 1422–1429, 2011.

[7] Z. Dženan, V. Aleš, E. Jan, H. Daniel, N. Christopher, and K. Andreas, "Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images," *Computer Graphics Forum*, vol. 33, no. 6, pp. 190–204, 2014.

[8] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb 2016.

[9] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic X-Ray images," in *SIBGRAPI*, Oct 2018, pp. 400–407.

[10] H. Oliveira and J. dos Santos, "Deep transfer learning for segmentation of anatomical structures in chest radiographs," in *SIBGRAPI*, Oct 2018, pp. 204–211.

[11] J. Egger, C. Nimsky, and X. Chen, "Vertebral body segmentation with GrowCut: Initial experience, workflow and practical application," *SAGE Open Med*, vol. 5, pp. 1–5, 2017.

[12] E. Burian, A. Rohrmeier, S. Schlaeger, M. Dieckmeyer, M. N. Diefenbach, J. Syväri, E. Klupp, D. Weidlich, C. Zimmer, E. J. Rummeny, D. C. Karampinos, J. S. Kirschke, and T. Baum, "Lumbar muscle and vertebral bodies segmentation of chemical shift encoding-based water-fat MRI: the reference database myosegmentum spine," *BMC Musculoskeletal Disorders*, vol. 20, no. 1, p. 152, Apr 2019.

[13] T. Huber, G. Alber, S. Bette, T. Boeckh-Behrens, J. Gempt, F. Ringel, E. Alberts, C. Zimmer, and J. S. Bauer, "Reliability of semi-automated segmentations in glioblastoma," *Clinical Neuroradiology*, vol. 27, no. 2, pp. 153–161, Jun 2017.

[14] J. R. F. Junior, M. Koenigkam-Santos, F. E. G. Cipriano, A. T. Fabro, and P. M. de Azevedo-Marques, "Radiomics-based features for pattern recognition of lung cancer histopathology and metastases," *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 23 – 30, 2018.

[15] P. Casti, A. Mencattini, M. H. Nogueira-Barbosa, L. Frighetto-Pereira, P. M. Azevedo-Marques, E. Martinelli, and C. Di Natale, "Cooperative strategy for a dynamic ensemble of classification models in clinical applications: the case of MRI vertebral compression fractures," *IJCARS*, vol. 12, no. 11, pp. 1971–1983, Nov 2017.

[16] Z. Xue, L. R. Long, S. Antani, and G. R. Thoma, "Spine X-ray image retrieval using partial vertebral boundaries," in *Symposium on Computer-Based Medical Systems (CBMS)*, June 2011, pp. 1–6.

[17] L. Bergamasco, K. Lima, C. Rochitte, and F. d. L. d. S. Nunes, "3d medical objects retrieval approach using spharms descriptor and network flow as similarity measure," in *SIBGRAPI*, Oct 2018, pp. 329–336.

[18] D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," *IJCARS*, vol. 13, no. 8, pp. 1211–1219, 2018.

[19] P. Stefan, S. Habert, A. Winkler, M. Lazarovici, J. Fürmetz, U. Eck, and N. Navab, "A radiation-free mixed-reality training environment and assessment concept for C-arm-based surgery," *IJCARS*, vol. 13, no. 9, pp. 1335–1344, Sep 2018.

[20] P. Banerjee, M. Hu, R. Kannan, and S. Krishnaswamy, "A semi-automated approach to improve the efficiency of medical imaging segmentation for haptic rendering," *Journal of Digital Imaging (JDI)*, vol. 30, no. 4, pp. 519–27, Aug 2017.

[21] R. Korez, B. Likar, F. Pernuš, and T. Vrtovec, "Model-based segmentation of vertebral bodies from MR images with 3D CNNs," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer Int. Publishing, 2016, pp. 433–441.

[22] N. Kamiya, J. Li, M. Kume, H. Fujita, D. Shen, and G. Zheng, "Fully automatic segmentation of paraspinal muscles from

3D torso CT images via multi-scale iterative random forest classifications," *IJCARS*, vol. 13, no. 11, pp. 1697–1706, Nov 2018.

[23] J. S. Ramos, M. T. Cazzolato, B. S. Faiçal, M. H. Nogueira-Barbosa, C. Traina Jr., and A. J. M. Traina, "3DBGrowth: volumetric vertebrae segmentation and reconstruction in magnetic resonance imaging," *Computer-Based Medical Systems (CBMS)*, pp. 1–6, June 2019.

[24] B. Gaonkar, Y. Xia, D. S. Villaroman, A. Ko, M. Attiah, J. S. Beckett, and L. Macyszyn, "Multi-parameter ensemble learning for automated vertebral body segmentation in heterogeneously acquired clinical MR images," *Journal of Translational Engineering in Health and Medicine (JTEHM)*, vol. 5, pp. 1–12, 2017.

[25] J. S. Ramos, C. Y. V. Watanabe, M. H. Nogueira-Barbosa, and A. J. M. Traina, "BGrowth: an efficient approach for the segmentation of vertebral compression fractures in magnetic resonance imaging," *Symposium on Applied Computing (SAC)*, pp. 1–8, April 2019.

[26] V. Vezhnevets and V. Konouchine, "GrowCut - interactive multi-label N-D image segmentation by cellular automata," *International Conference on Computer Graphics and Vision (GraphiCon)*, vol. 1, Nov 2005.

[27] L. Zhu, I. Kolesov, Y. Gao, R. Kikinis, and A. Tannenbaum, "An effective interactive medical image segmentation method using Fast GrowCut," in *Medical Image Computing and Computed-Assisted Intervention (MICCAI)*, vol. 17, 2014.

[28] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, *3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support*. New York, NY: Springer New York, 2014, pp. 277–289.

[29] W. S. Silva, D. L. Jasbick, R. E. Wilson, P. M. Azevedo-Marques, A. J. M. Traina, L. F. D. Santos, A. E. S. Jorge, D. de Oliveira, and M. V. N. Bedo, "A two-phase learning approach for the segmentation of dermatological wounds," *Computer-Based Medical Systems (CBMS)*, pp. 1–6, June 2019.

[30] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," *IOP Conference Series: Materials Science and Engineering*, vol. 324, p. 012049, mar 2018.

[31] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

[32] P. D. Barbieri, G. V. Pedrosa, A. J. M. Traina, and M. H. Nogueira-Barbosa, "Vertebral body segmentation of spine MR images using Superpixels," in *Symposium on Computer-Based Medical Systems (CBMS)*, C. T. Junior, P. P. Rodrigues, B. Kane, P. M. de Azevedo Marques, and A. J. M. Traina, Eds. São Carlos and Ribeirão Preto, BR: Conference Publishing Services (CPS), June 2015, pp. 44–49.

[33] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.

[34] T. Sørensen, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*, ser. Biologiske skrifter. I kommission hos E. Munksgaard, 1948.

[35] T. Marošević, "The hausdorff distance between some sets of points," *Mathematical Communications*, vol. 23, no. 2, Jan 2018.

[36] M. M. Deza and E. Deza, *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009.

[37] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *J. of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[38] F. Wilcoxon, S. Katti, and R. Wilcox, "Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test," *Selected Tables in Math. Statistics*, vol. 1, pp. 171–259, 1970.