International Journal of Computer Assisted Radiology and Surgery (2018) 13:1697–1706
https://doi.org/10.1007/s11548-018-1852-1

ORIGINAL ARTICLE

Fully automatic segmentation of paraspinal muscles from 3D torso CT images via multi-scale iterative random forest classifications

Naoki Kamiya[1] · Jing Li[2] · Masanori Kume[3] · Hiroshi Fujita[3] · Dinggang Shen[4] · Guoyan Zheng[2]

Received: 12 January 2018 / Accepted: 27 August 2018 / Published online: 1 September 2018
© CARS 2018

Abstract

Purpose To develop and validate a fully automatic method for segmentation of paraspinal muscles from 3D torso CT images.

Methods We propose a novel learning-based method to address this challenging problem. Multi-scale iterative random forest classifications with multi-source information are employed in this study to speed up the segmentation and to improve the accuracy. Here, multi-source images include the original torso CT images and later also the iteratively estimated and refined probability maps of the paraspinal muscles. We validated our method on 20 torso CT data with associated manual segmentation. We randomly partitioned the 20 CT data into two evenly distributed groups and took one group as the training data and the other group as the test data.

Results The proposed method achieved a mean Dice coefficient of 93.0%. It took on average 46.5 s to segment a 3D torso CT image with the size ranging from $512 \times 512 \times 802$ voxels to $512 \times 512 \times 1031$ voxels.

Conclusions Our fully automatic, learning-based method can accurately segment paraspinal muscles from 3D torso CT images. It generates segmentation results that are better than those achieved by the state-of-the-art methods.

Keywords Paraspinal muscles · CT · Segmentation · Random forest

Introduction

The paraspinal muscles play an important role in trunk movement and spinal stability. Several studies [2,3,6,8] have demonstrated an association between imaging parameters of the paraspinal muscles such as cross-sectional area (CSA) size, shape, density, and volume, and spinal degeneration and low back pain (LBP). The measurements of these imaging parameters in clinical practice, however, are not reliable enough as they are usually measured in a 2D axial CT image, which can be chosen differently from hospital to hospital. Although measuring the paraspinal muscles in 3D holds the potential to improve the accuracy, it has not become common as it requires expertise- and time-intensive manual segmentation. The integration of more automated procedures for the reliable 3D segmentation of paraspinal muscles may reduce the label-intensiveness associated with manual methods and provide reliability and reproducibility of the acquired imaging parameters with respect to segmentation bias and temporal drift, especially for multicenter, longitudinal studies.

Figure 1 (left) shows the entirety of the paraspinal muscles, which run along almost the entire spine. There is a pair of the muscles on both sides of the body. Automatic 3D segmentation of paraspinal muscles from CT images is challenging due to the size of the data, the large variability of muscle shape and appearance, and the close contact of paraspinal muscles with the surrounding muscles which appear with almost the same intensities as shown in Fig. 1(right).

Despite the fact that there are significant progresses made in automatic segmentation of muscles from MR images [5,11,13,15,17–19,22,28,30], only a few methods have been

Naoki Kamiya and Jing Li contributed equally to this paper.

✉ Dinggang Shen
dgshen@med.unc.edu

✉ Guoyan Zheng
guoyan.zheng@istb.unibe.ch

1 School of Information Science and Technology, Aichi Prefectural University, Nagakute, Japan

2 Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland

3 Department of Electrical, Electronic and Computer Engineering, Faculty of Engineering, Gifu University, Gifu, Japan

4 Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

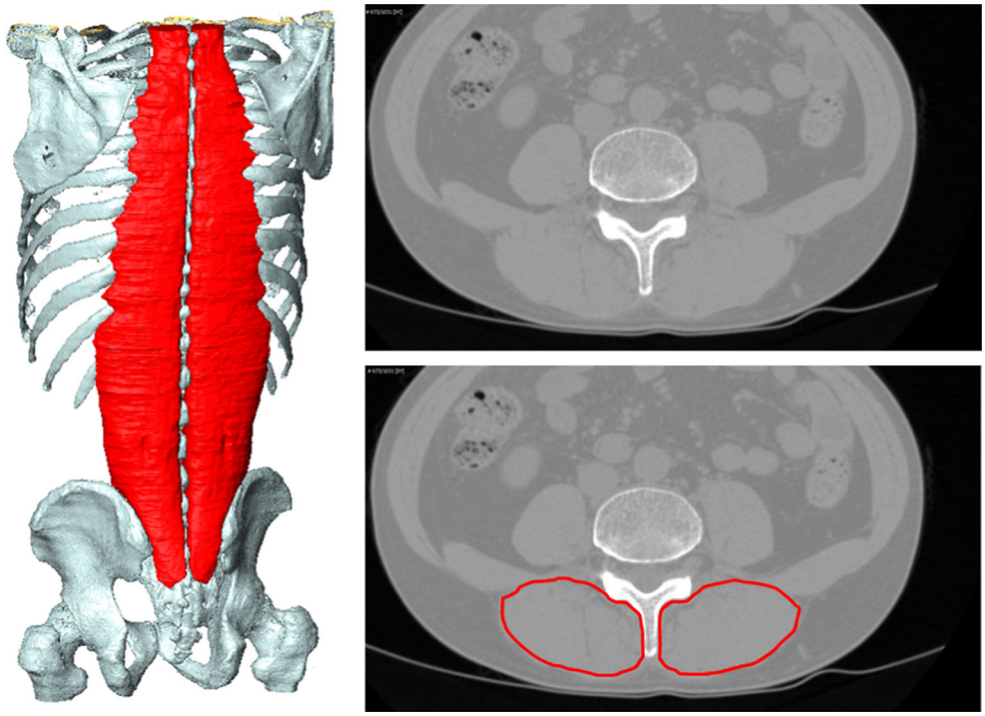

Fig. 1 Left: the paraspinal muscles (red) and bone (gray) in the 3D image. The paraspinal muscles are quite large and run along almost the complete spine. Right: the paraspinal muscles seen in an axial slice (top) and the expert manual segmentation (red contours in the bottom image)

introduced before to address the problem of automatic segmentation of muscles from CT data [9,10,12,20,20,29]. The published CT muscle segmentation methods can be classified into two categories: 2D methods and 3D methods. The methods in the former category usually work on 2D cross-sectional images taken at specific skeletal landmarks instead of 3D scans. For example, Wei et al. [29] presented a 2D atlas-based method for segmenting paraspinal muscles from 2D axial CT images. Another 2D method was introduced in [20], where a finite element method (FEM)-based deformable model was developed to incorporate a priori shape information via a statistical deformation model (SDM) within the template-based segmentation framework for automatic segmentation of skeletal muscle. Recently, Kume et al. [12] have investigated deep convolutional neural networks (CNN)-based approaches for automatic segmentation of paraspinal muscles at the level of the twelfth thoracic vertebrae in torso CT images. An average Dice coefficient of 86.3% was reported. In contrast, the methods in the latter category work directly on 3D scans. Along this line, Kamiya et al. proposed an rule-based expert system for the segmentation of the psoas major [9] and rectus abdominis [10] muscles from CT images, where the shape of the muscles was approximated by a simple quadratic function. An average Jaccard Similarity Coefficient (JSC) of 0.841 was reported in [10]. Inoue et al. [7] introduced a method to segment psoas major muscle using higher-order shape prior and reported an average JSC of 76.5%.

In this paper, we propose a novel learning-based method to address the challenging problem of fully automatic segmentation of paraspinal muscles from 3D torso CT images. In comparison with previous work, our contribution is as follows:

- To speed up the segmentation and to improve accuracy, we propose a novel multi-scale iterative random forest (RF) classification method for fully automatic segmentation of paraspinal muscles from CT images.
- Inspired by the auto-context model [21,25], we propose to employ features derived from multi-source information, including the original torso CT images and later also the iteratively estimated and refined probability maps of the paraspinal muscles.
- We conduct experiments to evaluate the performance of the present method and to compare the accuracy of the present method with a deep learning-based method.

The paper is organized as follows. In the next section, we will describe the method. "Experimental design and results" section will present the experimental results, followed by discussions and conclusions in "Discussions and conclusions" section.

Materials and method

We formulate the segmentation of paraspinal muscles as a two-class classification problem. To solve such a classifica-

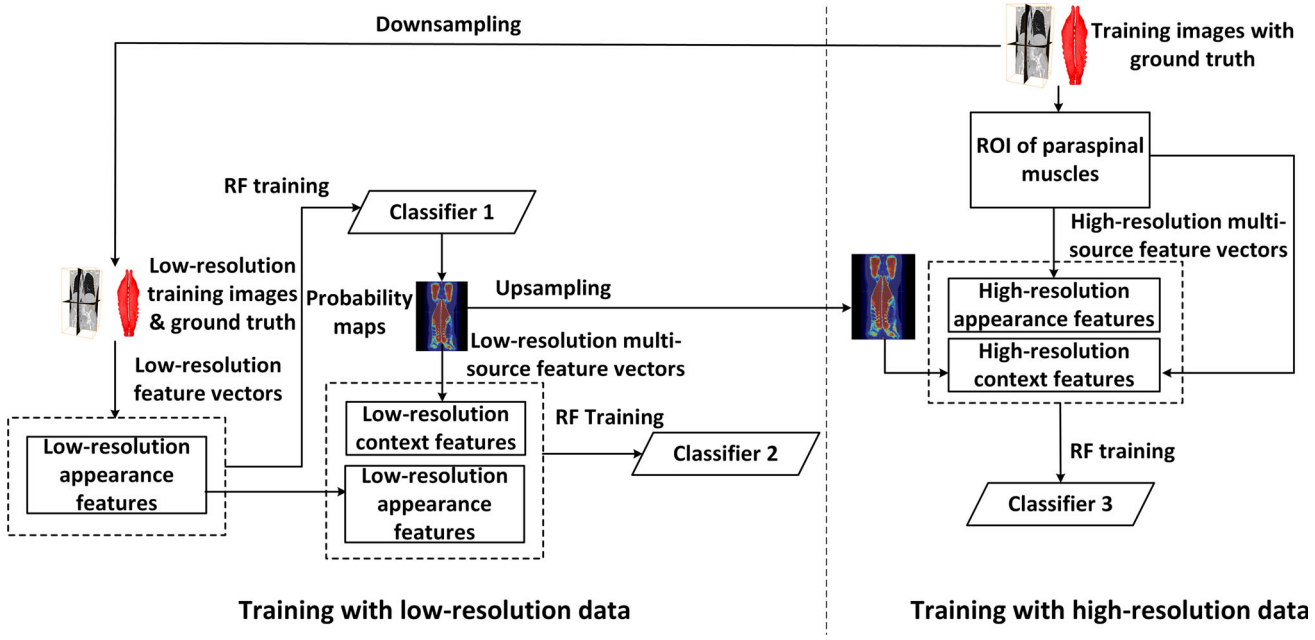

Fig. 2 A schematic illustration of how the training procedure works. The appearance features extracted from down-sampled CT images are used to train "Classifier 1," and then, both appearance features and the context features from probability maps are used to train the subsequent classifiers. We also employed a multi-scale strategy to speed up the training in high resolution

tion problem, we propose to employ random forests [1] and auto-context model [25], and conduct the classification in multiple scales.

Multi-scale random forest classification with auto-context model

Our method is inspired by Qian et al. [21] and Tu and Bai [25]. It is a supervised learning method consisting of training and testing stages. In the training stage, we will train a sequence of classification forests, as shown in Fig. 2. In the first iteration, we extract only the appearance features from the CT images to train a classification forest ("Classifier 1" in Fig. 2). By applying the trained forests in the first iteration, each training subject will produce tissue probability maps for paraspinal muscles or background, respectively. In the subsequent iterations, the tissue probability maps obtained from the previous iteration will be used as additional source information for training, thus getting a subsequent classification forest (e.g., "Classifier 2" in Fig. 2). It was demonstrated in [21] that the context features could encode the spatial constraints into the classification, thus improving the quality of the estimated tissue probability maps.

Similarly, in the testing stage, given a target CT image, we can obtain the initial tissue probability maps by applying "Classifier 1" using only the appearance features, as shown in Fig. 3. In the subsequent iterations, along with the appearance features, the tissue probability maps obtained from the previous iteration are also fed into the subsequent classifier for refinement.

In theory, we can apply RF classification method directly to get 3D segmentation. In practice, however, due to the large size of the torso CT data (the size of the data ranges from $512 \times 512 \times 802$ voxels to $512 \times 512 \times 1031$ voxels), directly applying RF classification method will lead to long training and testing time. In this paper, we propose a multi-scale strategy to address this issue. We conduct both training and testing in multiple scales. More specifically, during training, we first train two classifiers ("Classifier 1" and "Classifier 2" as shown in Fig. 2) following the above procedure on down-sampled training images. For the high-resolution training images, instead of training a classifier from appearance features extracted from high-resolution data to get the initial tissue probability maps, we up-sample the probability maps obtained from classifiers in low resolution. We empirically found that the up-sampled probability maps from "Classifier 1" led to more accurate segmentation results. Furthermore, for each training data, we extract a region of interest of the paraspinal muscles by dilating the associated ground-truth segmentation and randomly sample training data points only from this region in order to train a classifier in high resolution ("Classifier 3" in Fig. 2). Similarly, during testing, we also up-sample the probability maps obtained from "Classifier 1" to provide an initial tissue probability maps in high resolution. We then up-sample and dilate (in this study, we dilate 10 voxels along each axis) the binary segmentation results obtained from the probability maps of "Classifier 2" by thresholding

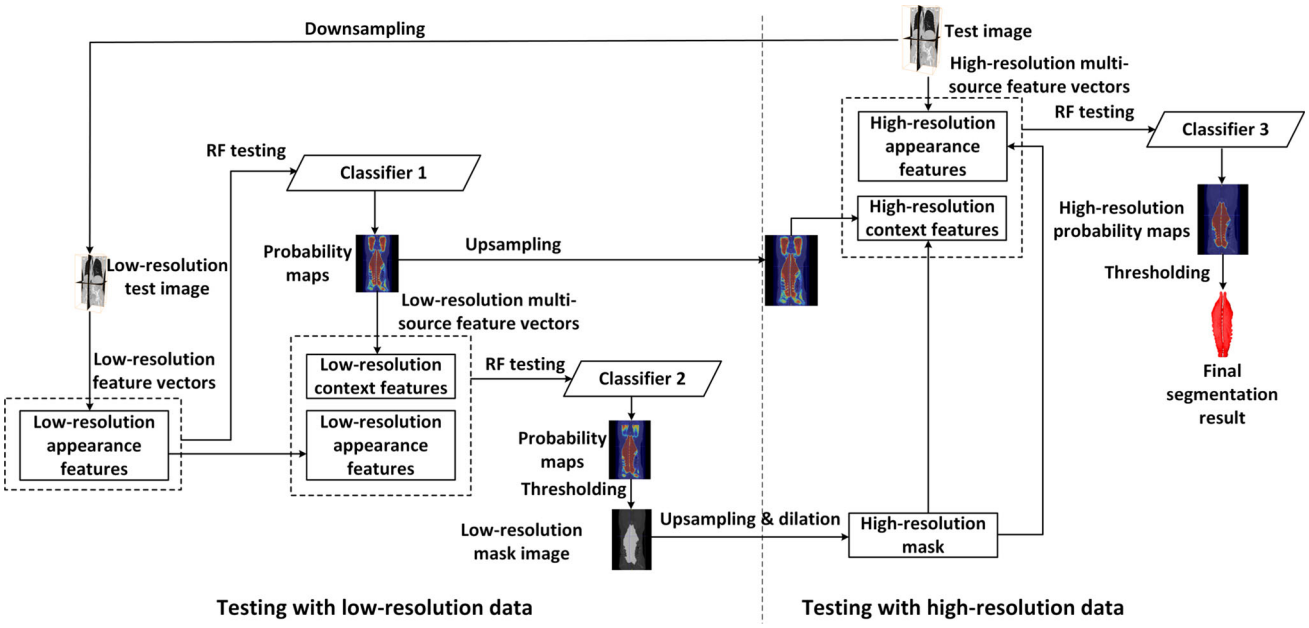

Fig. 3 A schematic illustration of how the testing procedure works. We used "Classifier 1" to get the tissue probability maps of the downsampled test image. Then, in the later iterations, the tissue probability maps obtained from previous iteration are also fed into the next classifier for refinement. Multi-scale strategy is used to speed up the testing

and morphological operations to provide a mask, which will then constrain the test region for "Classifier 3," i.e., we only apply "Classifier 3" to every voxel inside the masked region in order to compute the tissue probability maps in high resolution. Thresholding, followed by morphological operations to remove isolated small volumes and internal holes, is used to get the binary segmentation from the probability maps of "Classifier 3," which is then taken as the segmentation output of the present method.

Appearance features and context features

Considering the size of the data, we use the random Haar-like features as introduced in [27] for both appearance features and context features. Specifically, as shown in Fig. 4, for each voxel x, its Haar-like features are computed as the local mean intensity of any randomly displaced cubical region R_1 or as the mean intensity difference over any two randomly displaced cubical regions (R_1 and R_2) within the cubic image patch R around the voxel x in a source image I.

$$f(x, I) = \frac{1}{|R_1|} \sum_{p \in R_1} I(p) - b \frac{1}{|R_2|} \sum_{q \in R_2} I(q), b \in [0, 1] \tag{1}$$

where R is the patch centered at voxel x, I is any kind of source image, and the parameter $b \in [0, 1]$ indicates whether one or two cubical regions are used (as shown in Fig. 4, for $b = 0$ and $b = 1$).

To accelerate the feature extraction within each cubical region, we use the well-known integral image technique as introduced in [26]. Details about how to compute the integral image of a quantity can be found in [26]. The quantity can be the voxel intensity value or the estimated tissue probability value. Advantage of using integral image lies in the fact that once we obtain an integral image of the quantity over the complete CT volume, the sum of the quantity in any sub-volume or cubical region can be calculated quickly in constant time no matter how big the size of the cubical region is [26].

Data description

After local institution review board (IRB) approval, the present method was evaluated on torso CT data with associated manual segmentation of 20 subjects. CT images used in this study are non-contrast torso CT images taken at Light Speed Ultra 16 scanner (manufactured by GE) at Gifu University Hospital. We randomly partitioned the 20 subjects into two evenly distributed groups. We then took one group as the training data and the other group as the test data. Table 1 shows the demographic data of all 20 subjects used in our study.

All the CT data have an isotropic voxel resolution of 0.625 mm. The manual segmentation for each of data was created by Mr. Masanori Kume using a graph cut-based interactive method implemented in the common software platform called "PLUTO" (http://pluto.newves.org/trac) [16].

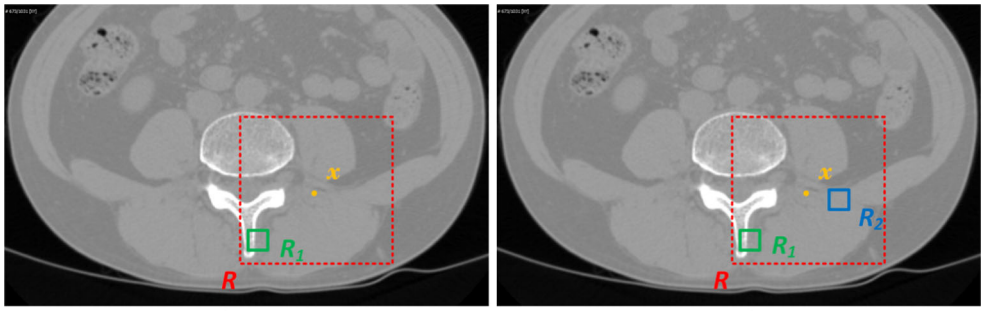

Fig. 4 A schematic illustration of how the Haar-like features as defined by Eq. (1) are computed for two different situations: when $b = 0$ (left) and when $b = 1$ (right)

Table 1 Demographic data of 20 subjects included in our study

Training group			Test group		
Training case	Gender	Age	Test case	Gender	Age
#01	M	49	#01	M	65
#02	F	84	#02	M	52
#03	M	71	#03	M	73
#04	M	40	#04	M	88
#05	M	68	#05	M	60
#06	F	74	#06	F	76
#07	M	53	#07	M	73
#08	M	68	#08	M	58
#09	M	70	#09	M	65
#10	F	61	#10	M	41
Average	NA	63.8 ± 13.1	Average	NA	65.1 ± 13.3

The obtained segmentation was then verified and corrected slice by slice by an anatomical specialist.

Implementation details

We trained and tested the random forest classifiers in two different scales. In order to train "Classifier 1" and "Classifier 2" in low resolution, we first down-sampled each training data into its one fourth of its original resolution along each axis. During training, we always sample evenly distributed data points from each training data, i.e., half of the data points sampled from the paraspinal muscle region and the other half from background. Specifically, in training "Classifier 1," we randomly sampled 20,000 points from each training data and compute 10,000 Haar-like features for each data point. The size of R was chosen to be 25 voxels. In training "Classifier 2," again we randomly sampled 20,000 data points from each training data. For each data point, we computed 10,000 multi-source Haar-like features with 5000 from the appearance and the other 5000 from the initial probability maps obtained from "Classifier 1." The size of R was chosen to be 45 voxels. "Classifier 3" was trained with data in the original resolution. We constrained the region to sample the data points for each training data to be within a ROI computed from the ground-truth segmentation. Again, we sampled 20,000 evenly distributed data points, and for each data point, we computed 10,000 multi-source features for each data point where 5000 features were computed from the training data and the other 5000 features from the up-sampled probability maps as shown in Fig. 2. The size of R for computing Haar-like features in high resolution was chosen to be 180 voxels.

Evaluation metrics

Assuming the automatically segmented set of voxels as AS and the manually defined ground truth as GT, we used both volume overlap metrics and distance-based metrics to evaluate the present method.

Volume overlap metrics

We computed following volume overlap metrics:

– *Dice Coefficients (DC)* It quantifies the match of two sets by normalizing the size of their intersection over the average of their sizes and is defined as follows:

$$\mathrm{DC} = \frac{2|\mathrm{AS} \cap \mathrm{GT}|}{|\mathrm{AS}| + |\mathrm{GT}|} \tag{2}$$

where the operator $|\cdot|$ returns the number of voxels contained in a region.

– *Jaccard Similarity Coefficients (JSC)* It is defined as the number of common voxels of the automatically segmented and ground-truth regions over their union:

$$\mathrm{JSC} = \frac{|\mathrm{AS} \cap \mathrm{GT}|}{|\mathrm{AS} \cup \mathrm{GT}|} \tag{3}$$

– *Precision (PR)* It is defined as the fraction of all automatically segmented voxels that are correct:

$$\mathrm{PR} = \frac{|\mathrm{AS} \cap \mathrm{GT}|}{|\mathrm{AS}|} \tag{4}$$

– *Recall (RC)* It is defined as the fraction of all ground-truth voxels that have been corrected segmented by an automatic method:

$$\mathrm{PR} = \frac{|\mathrm{AS} \cap \mathrm{GT}|}{|\mathrm{GT}|} \tag{5}$$

Distance-based metrics

Before we present the definitions of different distance-based metrics, we first define a distance measure for a voxel x from a set of voxels A as:

$$d(x, A) = \min_{y \in A} d(x, y) \tag{6}$$

where $d(x, y)$ is the Euclidean distance of the voxels incorporating the real spatial resolution of the volume data.

We further define the directed Hausdorff measure from a point set A to a point set B as the maximum distance, for all points in A, to the closest point in B. Mathematically, this is given as:

$$\overrightarrow{d}_H(A, B) = \max_{x \in A}(\min_{y \in B}(d(x, y))) \tag{7}$$

The directed percent Hausdorff measure, for a percentile r, is the r^{th} percentile distance over all distances from points in A to their closest point in B. For example, the directed 95% Hausdorff distance is the point in A with the distance to its closest point in B is greater or equal to exactly 95% of the other points in A. Mathematically, denoting the r^{th} percentile as K_r, this is given as:

$$\overrightarrow{d}_{H,r}(A, B) = K_r(\min_{y \in B} d(x, y)), \forall x \in A \tag{8}$$

With these definitions, we can define a number of distance-based metrics to quantify the dissimilarity of the automatic segmentation from the ground truth:

– *Average Surface Distance (ASD)* It is defined as the average of all the distances from points on the boundary of AS (we denote them as B_{AS}) to the boundary of GT (B_{GT}):

$$\mathrm{ASD} = \frac{1}{|B_{\mathrm{AS}}|} \sum_{x \in B_{\mathrm{AS}}} d(x, B_{\mathrm{GT}}) \tag{9}$$

– *Average Symmetric Surface Distance (ASSD)* It is defined as the average of all the distances from points on the boundary B_{AS} to the boundary B_{GT} and from points on B_{GT} to B_{AS}:

$$\mathrm{ASSD} = \frac{1}{|B_{\mathrm{AS}}| + |B_{\mathrm{GT}}|} \times \left(\sum_{x \in B_{\mathrm{AS}}} d(x, B_{\mathrm{GT}}) + \sum_{y \in B_{\mathrm{GT}}} d(y, B_{\mathrm{AS}}) \right) \tag{10}$$

– *Modified Hausdorff Distance (MHD)* It is defined as the undirected 95 percentile Hausdorff measure [4]:

$$\mathrm{MHD} = \frac{\overrightarrow{d}_{H,95}(\mathrm{AS,GT}) + \overrightarrow{d}_{H,95}(\mathrm{GT,AS})}{2} \tag{11}$$

Experimental design and results

Experimental design

We conducted two different studies in order to evaluate the efficacy of the present method. For the first study, the segmented result of each test data obtained by the present method was compared with the associated manual segmentation. For the second study, due to the large size of input data, we implemented a 2D fully convolutional network (FCN) [14] based on the network structure of VGG 16 [24]. In this FCN, the fully connected layer in VGG 16 is replaced by a convolutional layer, which is then followed by a 1×1 convolutional layer to generate segmentation in a down-sampled resolution. In order to get the segmentation in full resolution, up-sampling is done via deconvolutions [23].

Table 2 Segmentation results of the 10 test torso CT data

Case	JSC (%)	DC (%)	RC (%)	PR (%)	ASD (mm)	ASSD (mm)	MHD (mm)
#1	84.4	91.5	93.3	89.8	0.99	1.02	3.35
#2	87.0	93.1	97.5	89.0	0.74	0.83	2.98
#3	79.0	88.3	96.4	81.4	1.22	1.25	4.0
#4	85.1	92.0	97.3	87.2	0.89	0.98	3.19
#5	91.7	95.7	97.7	93.8	0.51	0.57	1.88
#6	88.8	94.0	95.0	93.1	0.64	0.73	2.58
#7	87.4	93.3	95.9	90.9	0.73	0.78	2.65
#8	87.9	93.6	98.0	89.5	0.79	0.85	2.80
#9	89.2	94.3	95.7	92.9	0.70	0.78	2.65
#10	89.8	94.6	97.5	91.9	0.72	0.75	2.50
Average	87.0 ± 3.5	93.0 ± 2.1	96.4 ± 1.5	89.9 ± 3.6	0.79 ± 0.20	0.85 ± 0.19	2.85 ± 0.56

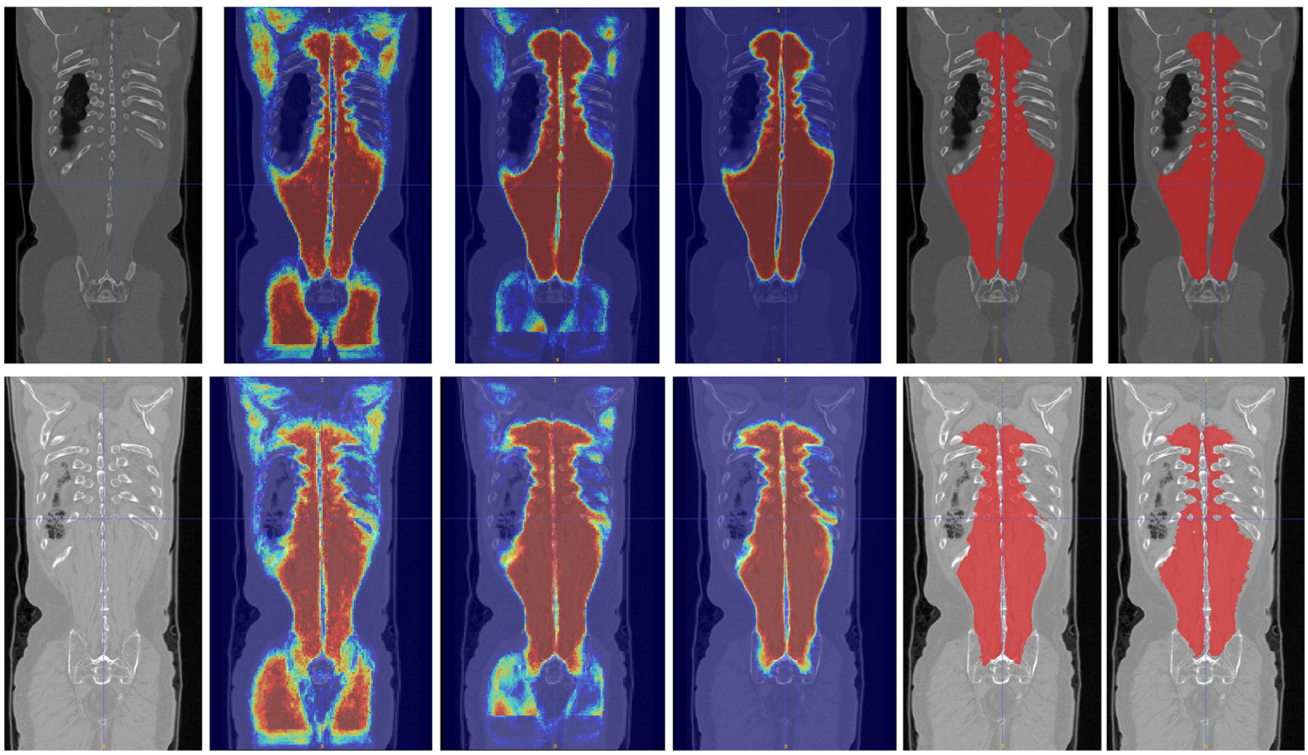

Fig. 5 Segmentation of the best (top, test case 05) and the worst (bottom, test case 03) cases. From left to right, the input image, the probability map from "Classifier 1," the probability map from "Classifier 2," the probability map from "Classifier 3," the final segmentation result, and the ground truth segmentation

Results

Quantitative segmentation results of the 10 test data is shown in Table 2. Our approach achieved a mean DC of $93.0 \pm 2.1\%$, a mean JSC of $87.0 \pm 3.5\%$, a mean RC of $96.4 \pm 1.5\%$, a mean PR of $89.9 \pm 3.6\%$, a mean ASD of 0.79 ± 0.20 mm, a mean ASSD of 0.85 ± 0.19 mm and a mean MHD of 2.85 ± 0.56 mm. Figure 5 shows the segmentation procedures for the best case (top row) and the worst case (bottom row). Qualitatively, it can be found that without incorporating context features, the probability maps (the second column) from "Classifier 1" show high values in relatively large portion of false positive regions. After integrating context features, the area of false positive regions is reduced as reflected by the probability maps (the third column) from "Classifier 2" but not completely removed. By incorporating the up-sampled context features with the constrained region of interest in the high-resolution image space, "Classifier 3" generates probability maps (the forth column) that have significantly

Table 3 Comparison of the results obtained by a 2D FCN and our method

Methods	JSC	DC	RC	PR
2D FCN	81.7 ± 3.2	89.9 ± 2.0	92.8 ± 5.0	87.5 ± 4.3
our method	87.0 ± 3.5	93.0 ± 2.1	96.4 ± 1.5	89.9 ± 3.6

reduced false positive regions, demonstrating the efficacy of the present method.

Implemented on a machine with a 3.5GHz Intel(R) i7 CPU with 12 cores and 64 GB RAM, it took on average 46.5 s to segment a torso CT image with the size ranging from $512 \times 512 \times 802$ voxels to $512 \times 512 \times 1031$ voxels. In contrast, without using the proposed multi-scale strategy, we have to test each voxel in a given 3D scan, which leads to an average test time of 205.0 s.

The results of the second study are shown in Table 3. In comparison with the 2D FCN method, our method demonstrated better performance. More specifically, the 2D FCN method achieved a mean DC of $89.9 \pm 2.0\%$, a mean JSC of $81.7 \pm 3.2\%$, a mean RC of $92.8 \pm 5.0\%$ and a mean PR of $87.5 \pm 4.3\%$. In contrast, our method achieved a mean DC of $93.0 \pm 2.1\%$, a mean JSC of $87.0 \pm 3.5\%$, a mean RC of $96.4 \pm 1.5\%$, and a mean PR of $89.9 \pm 3.6\%$.

Discussions and conclusions

Manual and automated segmentation of individual muscles in CT images has been recognized as a challenging task, given the high variability of shapes between muscles and subjects and the discontinuity or lack of visible boundaries between the target muscles and surrounding muscles. In this paper, we proposed a novel learning-based method for automatic segmentation of paraspinal muscles from 3D torso CT images and conducted a validation study to confirm the efficacy of the proposed method.

The results achieved by our method are better than those reported in previous work. For example, based on deep learning techniques, Kume et al. reported a mean DC of 86.3%, while our method achieved a mean DC of 93.0%. Using higher-order shape prior, Inoue et al. [7] reported an average JSC of 76.5% in segmenting psoas major muscles which is lower than what our method achieved. The reason why our method achieved better results than others is probably due to the integration of the multi-source information in a multi-scale learning-based framework. As shown in Fig. 5, the integration of multi-source information and the adoption of the multi-scale strategy progressively refine the probability maps obtained in different stages, leading to an accurate segmentation at the final stage. To get a fair comparison, we implemented a 2D FCN method. Our experimental results showed that the results achieved by our method were better than those achieved by the 2D FCN method.

The present method is not only accurate but also fast, largely due to the proposed multi-scale strategy. It is known that for random forest classification, the test time is proportional to the number of voxels in the test data. The initial segmentation obtained from "Classifier 2" at low resolution allows us to define a mask to constrain the test at high resolution to a smaller region of interest. This can not only improve the learning efficacy, as we concentrate on a smaller region than the complete image space, but also lead to faster algorithm as we will test on less number of voxels. Our experimental results demonstrate that our algorithm is four times faster than the one without using the multi-scale strategy.

It is worth to compare the method introduced in [21] with the present method. First, both methods are based on random forest classification with auto-context model [25]. Second, both studies confirm the effectiveness of incorporating context features for refined segmentation, despite the fact that the method introduced in [21] is applied to multi-parametric prostate MR images while the present method is evaluated on torso CT data. The differences between these two methods, however, are also apparent. More specifically, due to the purpose of the study reported in [21], which aims to localize prostate cancer from in vivo MR images, the resolution of their data is relatively low, leading to small data dimension along the out of plane direction. For example, the highest resolution of the multi-parametric MR images used in [21] is $0.3125 \times 0.3125 \times 3\,\text{mm}^3$. Additionally, their data were cropped around the prostate, which is a relatively small organ, in order to localize the prostate cancer from the cropped MR images. This is the reason why they can repeatedly apply the random forest classification with auto-context model in the original data space to get refined results. In contrast, the resolutions of our data are high in all three axes, leading to large data dimensions. Additionally, as we shown in Fig. 1, the paraspinal muscles are quite large, running along almost the complete spine. Furthermore, we did not purposely crop our torso CT data around the paraspinal muscles, which complicated the learning task for our problem. This has been demonstrated in the second and third columns of Fig. 5, where false positive predictions appear above and below the paraspinal muscles. By combining information extracted from the outputs of two classifiers that are trained in low resolution, we focus the third classifier on learning important multi-source features in a constrained region instead of the whole volume. As demonstrated in the fourth column of Fig. 5, such a strategy significantly reduced the false positive prediction, leading to refined segmentation.

There are limitations in our study. First, the dataset used in our study is relatively small. We are expecting to enlarge the dataset to include torso CT data of over 50 subjects, but the main challenge is to get the ground-truth annotations.

Second, all the CT data used in this study were acquired with the same scanner from Gifu University Hospital. It would be interesting to apply our trained model to CT images from other scanners in order to test the inter-scanner robustness. Considering the fact that unlike MR image values, CT values are correlated with tissue attenuation coefficients, we hypothesize that we can directly apply our trained model to CT data acquired from other scanners. Such a hypothesis needs to be verified in our future work. Last but not least, the present method was evaluated on CT data collected with a standard clinical protocol. Whether it will work or not on heterogeneous data acquired in clinical routine needs to be further checked in the future.

In summary, we proposed a novel learning-based method to address the challenging problem of automatic segmentation of paraspinal muscles from 3D torso CT images. Our method is based on multi-scale iterative random forest classifications with multi-source information. The experimental results demonstrated the efficacy of our proposed approach.

Acknowledgements This work was supported in part by a JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Multidisciplinary Computational Anatomy, #26108005 and #17H05301), JAPAN.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was obtained from all individuals included in the study.

References

1. Beriman L (2001) Random forests. Mach Learn 45:5–32
2. Bresnahan L, Smith J, Ogden A, Quinn S, Cybulski G, Simonian N, Natarajan R, Fessler R, Fessler R (2017) Assessment of paraspinal muscle cross-sectional area after lumbar decompression: minimally invasive versus open approaches. Clin Spine Surg 30(3):E162–E168
3. Cooper R, Clair Forbes W, Jayson M (1992) Radiographic demonstration of paraspinal muscle wasting in patients with chronic low back pain. Rheumatology 31(6):389–394
4. Dubuisson M, Jain A (1994) A modified hausdorff distance for object matching. In: Proceedings of international conference on pattern recognition (ICPR). pp 566–568
5. Engstrom C, Fripp J, Jurcak V, Walker D, Salvado O, Crozier S (2011) Segmentation of the quadratus lumborum muscle using statistical shape modeling. J Magn Reson Imaging 33:1422–1429
6. Hides J, Stokes M, Saide M, Jull G, Cooper D (1994) Evidence of lumbar multifidus muscle wasting ipsilateral to symptoms in patients with acute/subacute low back pain. Spine 19(2):165–172
7. Inoue T, Kitamura Y, Li Y, Ito W, Ishikawa H (2015) Psoas major muscle segmentation using higher-order shape prior. In: Proceedings of MICCAI-MCV workshop. pp 116–124
8. Kalichman L, Carmeli E, Been E (2017) The association between imaging parameters of the paraspinal muscles, spinal degeneration, and low back pain. Biomed Res Int 2017:14
9. Kamiya N, Zhou X, Chen H, Hara T, Hoshi H, Yokoyama R, Kanematsu M, Fujita H (2009) Automated recognition of the psoas major muscles on X-ray CT images. In: Proceedings of IEEE-EMBC 2009. pp 3557–3560
10. Kamiya N, Zhou X, Chen H, Muramatsu C, Hara T, Yokoyama R, Kanematsu M, Hoshi H, Fujita H (2011) Automated segmentation of recuts abdominis muscle using shape model in X-ray CT images. In: Proceedings of IEEE-EMBC 2011. pp 7993–7996
11. Karlsson A, Rosander J, Romu T, Tallberg J, Groenqvist A, Borga M, Dahlqvist Leinhard O (2015) Automatic and quantitative assessment of regional muscle volume by multi-atlas segmentation using whole-body water-fat mri. J Magn Reson Imaging 41(6):1558–1569
12. Kume M, Kamiya N, Zhou X, Kato H, Chen H, Muramatsu C, Hara T, Miyoshi T, Matsuo M, Fujita H (2017) Automated recognition of the erector spinae muscle based on deep CNN at the level of the twelfth thoracic vertebrae in torso CT images. In: Proceedings of the 36th JAMIT annual meeting
13. Le Troter A, Foure A, Guye M, Confort-Gouny S, Mattei J, Gondin J, Salort-Campana E, Bendahan D (2016) Volume measurements of individual muscles in human quadriceps femoris using atlas-based segmentation approaches. Magn Reson Mater Phys Biol Med (MAGMA) 29(2):245–257
14. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of 2015 IEEE conference on computer vision and pattern recognition (CVPR 2015). pp 3431–3440
15. Makrogiannis S, Serai S, Fishbein K, Schreiber C, Ferrucci L, Spencer R (2012) Automated quantification of muscle and fat in the thigh from water-, fat-, and nonsuppressed mr images. J Magn Reson Imaging 35(5):1153–1161
16. Nimura Y, Deguchi D, Kitasaka T, Mori K, Suenaga Y (2008) Pluto: a common platform for computer-aided diagnosis. Med Imaging Technol 26(3):187–191
17. Ogier A, Sdika M, Foure A, Le Troter A, Bendahan D (2017) Individual muscle segmentation in MR images: A 3D propagation through 2D non-linear registration approaches. In: Proceedings of IEEE-EMBC 2017. pp 317–320
18. Orgiu S, Lafortuna C, Rastelli F, Cadioli M, Falini A, Rizzo G (2016) Automatic muscle and fat segmentation in the thigh from t1-weighted MRI. J Magn Reson Imaging 43(3):601–610
19. Ozdemir F, Karani N, Fuernstahl P, Goksel O (2017) Interactive segmentation in MRI for orthopedic surgery planning: bone tissue. Int J Comput Assist Radiol Surg 12(6):1031–1039
20. Popuri K, Cobzas D, Esfandiari N, Baracos V, Jaegersand M (2016) Body composition assessment in axial CT images using FEM-based automatic segmentation of skeletal muscle. IEEE Trans Med Imaging 35(2):512–520
21. Qian C, Wang L, Gao Y, Yousuf A, Yang X, Oto A, Shen D (2016) In vivo MRI based prostate cancer localization with random forests and auto-context model. Comput Med Imaging Graph 52:44–57
22. Sdika M, Tonson A, Le Fur Y, Cozzone P, Bendahan D (2016) Multi-atlas-based fully automatic segmentation of individual muscles in rat leg. Magn Reson Mater Phys Biol Med (MAGMA) 29(2):223–235
23. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651
24. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
25. Tu Z, Bai X (2010) Auto-context and its application to high-level vision tasks and 3D brain image segmentation. IEEE Trans Pattern Anal Mach Intell 32:1744–1757

26. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of 2001 CVPR conference. IEEE pp 511–518
27. Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vis 57:137–154
28. Wang C, Teboul O, Michel F, Essafi S, Paragios N (2010) 3D knowledge-based segmentation using pose-invariant higher-order graphs. In: Proceedings of MICCAI 2010. vol Part 3. pp 189–196
29. Wei Y, Xu B, Tao X, Qu J (2015) Paraspinal muscle segmentation in CT images using a single atlas. In: Proceedings of IEEE international conference on progress in informatics and computing (IPC). pp 211–215
30. Yang Y, Chong M, Tay L, Yew S, Yeo A, Tan C (2016) Automated assessment of thigh composition using machine learning for Dixon magnetic resonance images. Magn Reson Mater Phys Biol Med (MAGMA) 29(5):723–731