Segmentation of Skeleton and Organs in Whole-Body CT Images via Iterative Trilateration

Marie Bieth*, Loic Peter, Stephan G. Nekolla, Matthias Eiber, Georg Langs, Markus Schwaiger and Bjoern Menze

***Abstract*—Whole body oncological screening using CT images requires a good anatomical localisation of organs and of the skeleton. While a number of algorithms for multi-organ localisation have been presented, developing algorithms for a dense anatomical annotation of the whole skeleton, however, has not been addressed until now. Only methods for specialised applications, e.g., in spine imaging, have been previously described. In this work, we propose an approach for localising and annotating different parts of the human skeleton in CT images.**

We introduce novel *anatomical trilateration* features and employ them within iterative scale-adaptive random forests in a hierarchical fashion to annotate the whole skeleton. The anatomical trilateration features provide high-level long-range context information that complements the classical local context-based features used in most image segmentation approaches. They rely on anatomical landmarks derived from the previous element of the cascade to express positions relative to reference points. Following a hierarchical approach, large anatomical structures are segmented first, before identifying substructures. We develop this method for bone annotation but also illustrate its performance, although not specifically optimised for it, for multi-organ annotation.

Our method achieves average Dice Scores of 77.4 to 85.6 for bone annotation on three different datasets. It can also segment different organs with sufficient performance for oncological applications, e.g. for PET/CT analysis, and its computation time allows for its use in clinical practice.

***Index Terms*—Segmentation, Medical Imaging**

I. Introduction

DENSE skeleton annotation is necessary for a variety of clinical and research applications, in particular in orthopaedics or oncology. Planning orthopaedic interventions often requires the dense segmentation of bones and muscles in CT and MRI, for example in hip surgery or for interventions on the spine. Nearly all of these tasks only deal with a limited field of view. In oncology, the diagnosis of patients with primary tumours or secondary metastases of the bone requires the analysis and mapping of bone lesions, for example in whole body PET/CT scans, often several times during treatment. For heavily metastasised patients with dozens to hundreds of individual lesions, this is a very time consuming task if the annotation is performed manually, and diagnostic information is often only reported in a very qualitative fashion [1], [2]. Recently, an effort has been made towards quantitative analysis, but methods remain semi-automatic [3], [4] and only provide global statistics. Hence, a dense annotation of the skeleton and its substructures could ease the automation of the process, providing an anatomical reference frame for localising structures of interest, e.g., or for re-identifying previously detected lesions in follow-up scans. It could, in a later step, help in providing local rather than global lesion statistics across a population.

Segmenting the whole skeleton, as necessary in oncological applications, is a much more difficult task compared to segmenting well defined and constrained parts e.g., for orthopaedic applications, due to a wider range of anatomies, fields of view and patient position variations and a larger number of anatomical structures. To the best of our knowledge, the only study that approached the task of annotating the whole human skeleton, rather than narrow subregions of it, is the one that we presented in [5].

In the present paper, we extend this preliminary work and offer a robust registration-free method that can segment the whole skeleton to the accuracy needed for whole body oncological staging, also offering means for segmenting other structures of interest such as organs at an accuracy that meets clinical requirements for PET/CT analysis. It relies on a cascaded scale-adaptive random forest using *trilateration* features that express relative positions using landmarks in the skeleton that get updated in the cascade as a reference. This is coupled with a coarse-to-fine hierarchical refinement of labels.

A. State of the art

Different semi-automatic and automatic methods exist that perform skeleton annotation for the orthopaedic domain in MR or CT images. For example, methods exist for the spine [6]–[11], the knee [12], the ribs [13] and the hip region [14]. Methods also exist for multi-organ segmentation in CT [15]–[19].

Widely used methods for multiple structure segmentation include (but are not limited to) atlas based methods, deformable model based methods, random forest based methods,

M. Bieth and B. Menze are with the Department of Computer Science at the Technical University of Munich, Germany (e-mail: marie.bieth@tum.de)

L. Peter is with the chair for Computer Aided Medical Procedures at the Technical University of Munich, Germany and the Translational Imaging Group, University College London, UK

M. Bieth, S.G. Nekolla, M. Eiber and M. Schwaiger are with the Clinic for Nuclear Medicine, Klinikum Rechts der Isar at the Technical University of Munich, Germany

M. Eiber is with the Department of Molecular and Medical Pharmacology, David Greffen School of Medicine at UCLA, Los Angeles, USA

G. Langs is with the Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Austria

Published as M. Bieth, L. Peter, S. G. Nekolla, M. Eiber, G. Langs, M. Schwaiger, B. Menze, "Segmentation of Skeleton and Organs in Whole-Body CT Images via Iterative Trilateration," in IEEE Transactions on Medical Imaging, DOI: 10.1109/TMI.2017.2720261

convolutional neural network based methods, and graph based methods. Some hybrid methods combine ideas from several approaches. In the following, we will focus on how these methods try to leverage the local or global context of the image to improve the segmentation accuracy.

Segmentation by registration of (possibly multiple) atlases, which has proven successful for multi-organ segmentation [15], intrinsically uses context information by imposing constraints on the transformations. More context information can be incorporated by performing registration or atlas computation [16] at different scales. Context can also be explicitly used at label fusion time by modelling dependencies between voxels or labels [13], or taking a global decision, for example based on contours [20], instead of applying a voting scheme independently at each voxel. Similarly, in deformable model methods, context is implicitly considered by the constraints imposed on the deformation, and can additionally be explicitly considered by accounting for spatial [8] or hierarchical [18], [19] relations between different objects.

Others have used learning algorithms such as random forests (RFs) [21] that rely on decision trees and bagging. In these approaches, context information can be incorporated in the features such as with Haar-like [22] and geodesic features [23] or by using landmarks [12]. In combination with Haar-like features, a RF variant that learns the scale to which context is beneficial to the segmentation has been described [24] and a version where decisions are based on *all* features [25] showed good results for multi-organ segmentation. Context can also be used in the forest construction itself. For example, some approaches have incorporated global image similarities into the forest construction [26], [27] or predicted label and distance at the same time in a multi-task fashion [28], [29]. Cascaded systems are an alternative implicitly taking context into account either going from a global to a local scale [30] or using long range context information by providing the output of the forest (or an intermediate output [31]) to another forest for further training in an *auto-context* fashion, as initiated in [32]. The output of a classifier can also be processed before being used in the next one: in [9], the probability maps are regularised before being used as input for the next forest.

Deep convolutional neural network [33] approaches are related to the aforementioned cascaded system. By applying cascaded filters and pooling to the image, deeper feature maps contain information from a wider range of voxels in the image. The loss function can also be modified to explicitly take into account e.g., topological information [34].

Finally, local spatial constraints are often considered through conditional random fields (CRF) [35] and level set approaches, that try to minimise similar energies. For CRFs, with *classical* sparse binary terms, the α-expansion algorithm [36] is often used for energy minimisation. With dense graphs, a mean field approximation is prefered [37]. Other minimisation strategies are used for level set methods. Context information can be incorporated in the unary as well as the binary energy terms. In particular, any of the aforementioned methods can be used to generate unary terms. Additionally, appearance, shape and location can be explicitly considered in the unary term [38] (and jointly optimised [39]), constraints with respect to relative positions can be enforced using the binary terms [40], and multiphase or temporal data can be leveraged by using four-dimensional graphs [38]. These methods can also be used in combination with superpixels/supervoxels [41]–[43].

Although different strategies have been used to leverage context information, most of the methods mentioned in the previous paragraphs present drawbacks. Atlas and model based methods are applicable only when the variation among subjects is low. Otherwise, some cases may differ too much from the atlas/model to be correctly recovered. Topological variations of the structures to be segmented, in our problem, for example, a variation in the number of ribs, may lead to failure in atlas-driven methods. The use of multiple atlases can alleviate this problem and improve the performance, but usually leads to an increased computational burden with registration required at test time to every atlas. Deep learning methods require a large amount of training data, which is often impractical for medical applications, and are, in particular for three-dimensional images, limited by memory constraints. RFs consider voxels as independent, which can be alleviated by choosing explicitly context-oriented features or combining them with a CRF. Moreover, none of the aforementioned methods can fully take advantage of the very structured relations between different parts of the skeleton.

B. Contributions

In the following, we propose a whole body annotation approach that overcomes several of the current limitations. In particular, it leverages the very structured aspect of our problem, needs only a limited amount of training data and presents a good performance to computation time ratio. More specifically:

- We address for the first time the task of whole skeleton annotation.
- We introduce new anatomical trilateration features that efficiently incorporate long-range context information (sec. II-D).
- We propose a cascaded random forest approach where landmarks are updated between each element of the cascade (sec. II-E).
- We present an evaluation of our approach and demonstrate that it achieves high performance on three different datasets and compares favourably with autocontext and scale-adaptive random forest (sec. III).

II. Methods

A. Overview

In our approach, classification is performed jointly with localisation by a cascaded random forest (illustrated in Fig. 1, sec. II-E) with anatomical trilateration features: from the probability map obtained from a random forest, we compute not only the voxel labels, but also the centroid of each structure to label. These centroids are then given as supplementary input to the next random forest of the cascade and used as landmarks to compute our novel anatomical trilateration features (illustrated in Fig. 2, sec. II-D).

Algorithm 1 Overview of our method for iterative annotation of skeleton parts

```
1:  function PREDICT(V, LabelTree)
2:      landmarks = [ ]
3:      for h=0...H-1 do                                              ▷ Hierarchical approach
4:          for k=1...K^{L_h} do
5:              localCentroids = [ ]
6:              for iterations i=1...I^{L_h} do                       ▷ Cascaded trilateration
7:                  mask = voxels to label
8:                  probabilityMap = RF(V,mask,G_k^{L_h},landmarks,localCentroids)   ▷ Scale adaptive random forest
9:                  segm = argmax(probabilityMap)
10:                 localCentroids = computeCentroids(segm)           ▷ Local centroids updating
11:             end for
12:             landmarks.append(localCentroids)   ▷ Storing local centroids as landmarks for further computations
13:         end for
14:     end for
15:     skeleton=mergeResults()
16:     return skeleton
17: end function
```

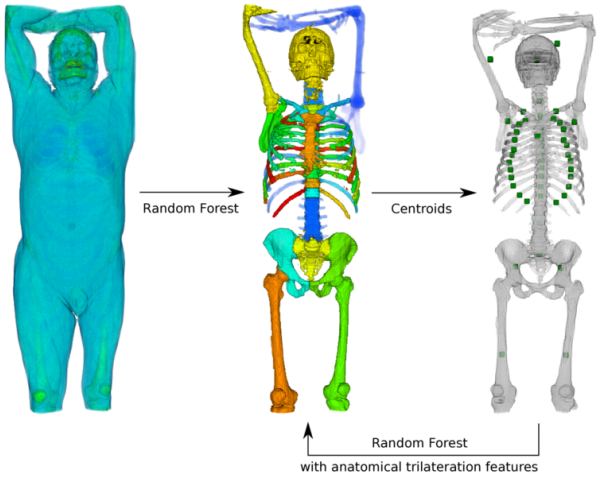

Fig. 1. Details of the cascaded RF (lines 6 and 10 in Algorithm 1). The landmarks obtained from the output of one iteration are used to compute the anatomical trilateration features in the next iteration.

The cascade described here constitutes a *super-classifier* that we use in a hierarchical coarse-to-fine fashion (sec. II-F). The labels are ordered in a hierarchical manner which is described by a label tree (Fig. 3). First, coarse groups of labels are segmented, and are thereafter refined by other super-classifiers in further classification procedures until all individual labels are segmented.

Details are provided in the following sections and an overview of the testing procedure is provided in Algorithm 1.

B. Notations

In all the following, K is the number of labels, $\mathbf{v}(x, y, z)$ is a voxel, $\mathbf{v_{sym}}$ is the symmetric of $\mathbf{v}$ with respect to the mid-sagittal plane, $\mathbf{l}(x_l, y_l, z_l)$ is a landmark, k_v is the label of voxel $\mathbf{v}$ and $a, b, c \in \mathbb{R}$. In a tree, for a node n, $Children(n)$ refers to the children of n.

C. Scale Adaptive Random Forests

We chose random forests as the atomic inference element of our iterative algorithm (line 8 of Algorithm 1) for their high performance to computation time ratio. More specifically, we use an implementation of scale adaptive random forest (saRF) [24] to perform a probabilistic segmentation of the structures to label. At training time, this particular version of RF samples the features sequentially in a fine-to-coarse fashion instead of sampling features uniformly for each node as done in the classical RF algorithm [21]. This can be seen as a guided sampling that learns the scale of the problem without user input.

At test time, a probabilistic segmentation $P^0(k_v|\mathbf{v})$ is obtained. It can be discretised by choosing for each voxel the label with the highest probability:

$$\hat{k}_v^0 = \arg\max_{k_v}(P^0(k_v|\mathbf{v})) \quad (1)$$

D. Features Description

Within the saRF, we use two kinds of features: Haar-like features for considering local intensity context and trilateration features for long-range anatomical context. At training, the Haar-like features are sampled in scale adaptive fashion whilst the trilateration features are sampled in the conventional uniform way. All features are described in details in the following sections.

1) Haar-like features: We use Haar-like features [22] as low-level context features. They are computed as an arithmetic operation between the intensity average of two boxes defined in the image domain, possibly in different modalities. Each feature $\mathcal{F}$ can be described by seven parameters $\mathcal{F} = (\mathbf{b_1}, \mathbf{b_2}, \mathbf{o_1}, \mathbf{o_2}, m_1, m_2, \sigma)$ where $\mathbf{b_1}, \mathbf{b_2}$ are three-dimensional vectors describing box sizes, $\mathbf{o_1}, \mathbf{o_2}$ are offsets describing the centre of the boxes compared to the current voxel $\mathbf{v}$, m_1, m_2 are the modalities in which the respective box value has to be computed and σ is an arithmetic operation ($\sigma \in$

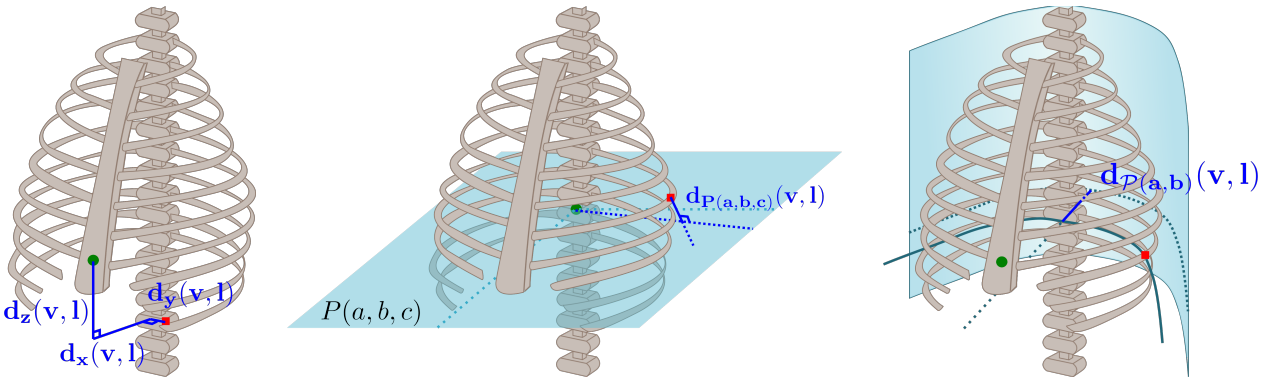

Fig. 2. Illustration of the anatomical trilateration features. In all images, the red square is the voxel to classify, and the green circle is the landmark. Left: signed distance features. Middle: planar distance feature. Right: parabolic distance features.

{sum, difference, sign of difference, absolute difference}). These features can be computed very efficiently by using integral volumes.

We also use symmetric Haar-like features that are similar to conventional Haar-like features, but for which the offset o_2 is computed with respect to voxel $\mathbf{v_{sym}}$. This is an extension of the symmetric features used in [44] and is particularly suited to the problem at hand because the human skeleton presents an approximate symmetry with respect to the mid-sagittal plane.

2) Anatomical trilateration features: Trilateration features represent the relative position of voxels or geometrical structures with respect to a reference point. Landmarks are needed for the computation of anatomical trilateration features. They can be manually annotated or obtained from previous computations of the cascade as described in section II-E. Trilateration features are illustrated in Fig. 2

2-a) Distance features: the distance features are the Euclidean and signed distances to landmarks. By the trilateration principle, it is possible to place a point in space if its distances to four landmarks are known and compatible. If the x, y, and z signed distances are known, one landmark is enough. Intuitively, in our case, because of anatomical variability and possible inaccurate landmark localisation, more landmarks are necessary to semantically trilaterate the position of a point.

The signed distances are defined as follows:

$$\begin{cases} \text{x signed distance } d_x(\mathbf{v},\mathbf{l}) = x - x_l \\ \text{y signed distance } d_y(\mathbf{v},\mathbf{l}) = y - y_l \\ \text{z signed distance } d_z(\mathbf{v},\mathbf{l}) = z - z_l \end{cases} \quad (2)$$

2-b) Geometric features: the planar and parabolic features detect the presence of planes and parabolic cylinders in the data and their positions relative to landmarks. Such structures occur for example in the rib-cage. For each feature, a plane or parabolic cylinder is defined with respect to a landmark. The respective feature then reflects how *far* from the structure $\mathbf{v}$ is. They are defined as follows:

$$\begin{cases} \text{planar distance} & d_{P(a,b,c)}(\mathbf{v},\mathbf{l}) = a(x-x_l) + b(y-y_l) \\ & \quad +c(z-z_l) \\ \text{parabolic distance} & d_{\mathcal{P}(a,b)}(\mathbf{v},\mathbf{l}) = (x-x_l) + a(y-y_l-b)^2 \end{cases} \quad (3)$$

During training, a, b and c are sampled randomly at each node.

E. Cascaded anatomical trilateration

We use the saRF and features described above in a cascade where the landmarks needed to compute the anatomical trilateration features are obtained from each element of the cascade for the next one (lines 6 and 10, in blue, of Algorithm 1). Using the initial probabilistic segmentation $P^0(k_v|\mathbf{v})$ obtained from the initial RF, the centroid of each segmented structure can be computed. The centroids can then be used as a densely meshed ensemble of landmarks to iterate the classification with trilateration features. An updated probabilistic segmentation $P^1(k_v|\mathbf{v})$ is obtained.

This process can be iterated by computing refined centroids from $P^i(k_v|\mathbf{v})$, $i \in \mathbb{N}^*$ and using them as landmarks for a classification step that outputs $P^{i+1}(k_v|\mathbf{v})$. The cascade is depicted in Fig. 1. Because they are recomputed after each iteration, the centroids can be considered as self-updating landmarks. Through the anatomical trilateration features described in the previous section, centroids are used to represent context information in a more condensed way than in autocontext [32].

F. Hierarchical segmentation

Because our ground truth contains many labels (K between 51 and 88 depending on the field of view), we use a coarse-to-fine hierarchical segmentation approach, following a user-defined label tree with H levels (lines 3 and 12, in green, of Algorithm 1). A simplified example of a label tree with $K = 9$ labels and $H = 2$ segmentation levels is depicted in Fig. 3.

The tree defines super-structures G (i.e. groups of labels) that are segmented in the coarser levels before being further divided in the finer levels. In level L_h, K^{L_h} (super-)structures $G_1^{L_h}, \cdots, G_{K^{L_h}}^{L_h}$ are segmented. By definition, the root of the tree contains only one super-structure (i.e. $K^{L_0} = 1$ with all labels $G_1^{L_0} = \{1, 2, \cdots, K\}$, and the tree has $K^{L_H} = K$ leaves $G_1^{L_H} = \{1\}$, ..., $G_K^{L_H} = \{K\}$.

A cascaded classifier (as described in II-E) is attached to each internal node $G_k^{L_h}$ of the tree. It receives landmarks from all already segmented (super-)structures, runs for I^{L_h} iterations and classifies voxels labelled with $G_k^{L_h}$ in level L_h into labels of $Children(G_k^{L_h})$. Fig. 3 shows a simplified example with the classifiers in red and the flow of landmarks in green.

When available, a mask can be applied before starting the hierarchical process, to remove the background voxels. This speeds up the computation by reducing the number of voxels to classify.

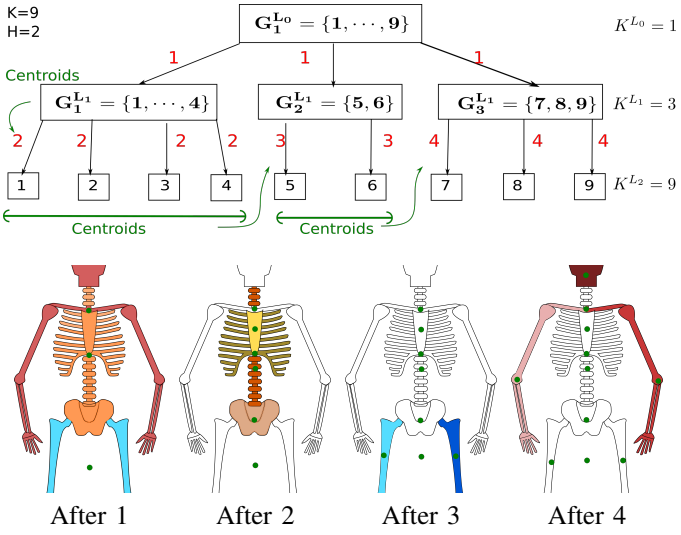

Fig. 3. Top: simplified hierarchical label tree example (9 labels and 2 levels of segmentation). The red digits indicate the ordering of the execution: all black arrows with the same number are processed simultaneously by the same cascaded RF. The green arrows indicate the subsequent flow of landmarks in the hierarchical tree (lines 3 and 12 in Algorithm 1). Bottom: results after the execution of the different steps in the example. Better viewed in color. Note that after each classifier, not only labels but also centroids are known.

G. Regularisation

Similarly to [45], [46], we use a CRF as final step to ensure spatial consistency of the segmentation. The nodes $\mathcal{N}$ of the graphical modelare the voxels to classify. A 26-neighbourhood structure is used for the binary connections. The energy of the CRF is defined as follows:

$$E = -\sum_{v \in \mathcal{N}} log(P(k_v|\mathbf{v})) - \lambda \sum_{v_1 \sim v_2} B(k_{v_1}, k_{v_2}) \qquad (4)$$

where $P(k_v|\mathbf{v})$ is the probability map obtained from previous steps. B is a *compatibility* term computed as the logarithm of neighbouring frequencies in the training data. This way, the cost of assigning different labels to neighbouring voxels is higher if the association does not exist or is rare (e.g., skull and pelvis) than if the association is often found in the training data (e.g., vertebra L1 and vertebra L2). The binary costs are therefore *learnt*, so that our regularisation approach depends on only one hyperparameter $\lambda \in \mathbb{R}$ that sets the balance between fidelity to the probability map and spatial consistency. We use the α-expansion algorithm [36] implemented in the OpenGM library [47] to find the label configuration minimising the energy of the graph.

III. Experiments

A. Data sets

We conducted experiments on three datasets:

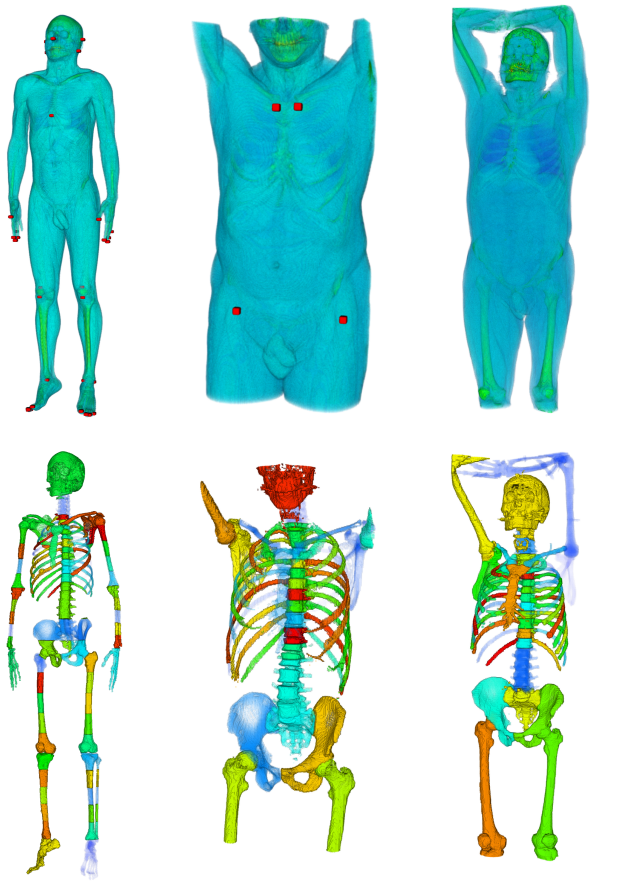

Fig. 4. Representative examples of the three datasets used in the experiments. The first line shows the original CT images with landmarks in red. The second line shows the ground truth bone annotation. Labels have been randomised for better visualisation.

(1) Healthy subjects dataset (HS): 20 whole body CT scans of healthy subjects.
(2) Prostate cancer dataset (PC): 30 thorax and trunk CT scans of prostate cancer patients.
(3) Multiple myeloma dataset (MM): 20 thorax and trunk CT scans of multiple myeloma patients.

Healthy Subjects Dataset: 20 non-contrast enhanced whole body CT images of healthy subjects from the *whole body morphometry project* (Mallinckrodt Institute of Radiology Washington University, School of Medicine, 2010), arms down, with a mean resolution of 1.3 mm $\times$ 1.3 mm $\times$ 1 mm were resampled to a mean resolution of 2.6 mm $\times$ 2.6 mm $\times$ 2 mm and a mean size of 256 $\times$ 256 $\times$ 896 voxels. The skeleton was annotated for 88 bone substructures. 57 landmarks at joints or tips of bones were also annotated.

Prostate Cancer Dataset: 30 contrast-enhanced CT images were extracted from PSMA-PET/CT images of prostate cancer patients, arms up. The field of view went from mid-thighs to skull. The images were resampled to an isotropic resolution of 2 mm and a mean size of 230 $\times$ 230 $\times$ 434 voxels. 51 bone structures were manually annotated. 4 landmarks at tips of bones were also annotated.

Multiple Myeloma Dataset: 20 non-contrast enhanced CT images of multiple myeloma patients, arms up, from the European VISCERAL project [48] were resampled to an isotropic resolution of 2 mm and a mean size of 168 $\times$ 216 $\times$ 657 voxels. The same bone structures as in Dataset 2 were manually annotated. Annotations for other structures (trachea, lungs, kidneys, psoas muscles, aorta, liver, spleen) are also available for this dataset.

The bone structures were chosen to be relevant for oncology analysis. Some structures, such as the hands in the healthy subjects dataset, group several bones because each of these bones is too small to be relevant alone for oncological mapping. Other structures are segments of bones, such as the femur segments in the healthy subjects dataset, because the whole bone is too long/big to obtain relevant local statistics in further analysis for oncological staging. Skeleton annotations of all datasets are depicted in Fig. 4.

B. Setup

1) Data preprocessing: In the preprocessing, images were windowed as follows: : regions below -150 HU were set to -150 HU, regions between -150 and -50 HU (approximate fat range) were set to -50 HU, and regions above 200 HU (bone range) were set to 200 HU. Moreover, a skeleton mask was generated for each patient by the approach described in [5]: the intensity of the different tissues was modelled by a Gaussian mixture, and a CRF was used to obtain the mask.

2) Parameters: Except for the transfer experiment (see III-D4), all experiments were run with a 2-fold cross validation. The random forests were trained with 100 trees. All other parameters are indicated in the supplementary material. In the planar distance features, b was set to 0 to consider only non-y-aligned planes. For all experiments, two hierarchical levels were used (the label groups are detailed in the supplementary material). The cascade II-E was run for 2 iterations in the first level and 5 iterations in the second. Except in experiment III-D5, only bone-voxels were classified. In experiment III-D5, all voxels in the body were classified. Unless stated otherwise, no initial landmarks were used. In the regularisation, λ was set to 1. We evaluated all experiments using the average Dice score over classes (DS).

C. Computing time

As a representative example, we recorded the testing time for 15 subjects of the prostate cancer Dataset on a Intel Xeon(R) CPU (3.20GHz$\times$4). With the settings described in the previous section, the average computing time for one subject was 488 seconds without CRF and 749 seconds with (Fig. 5). Each iteration of the second hierarchical level took around 65 seconds. Based on these numbers, each user can establish his own trade-off between speed and performance by choosing the number of iterations for each level and whether to use regularisation or not. Using only one iteration in the second hierarchical level brought the computing time down to an average of less than 4 minutes per subject.

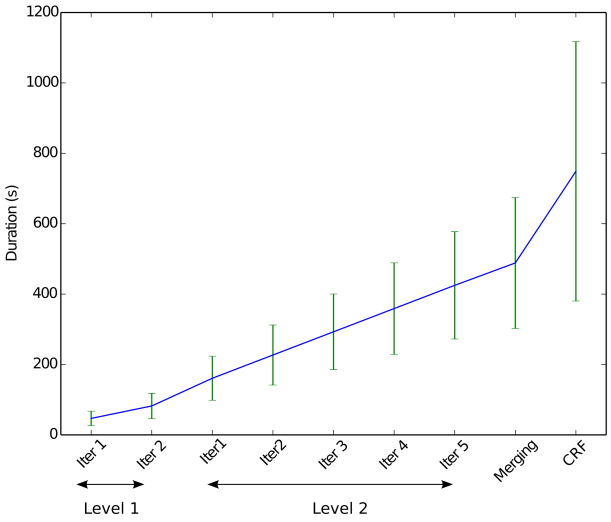

Fig. 5. Average computing time for the different steps of the algorithm measured using 15 subjects of the PC Dataset. The blue curve shows the average computing time and the green bars the range.

TABLE I
MEAN DS OVER SUBJECTS FOR SKELETON ANNOTATION FOR DIFFERENT METHODS

	Method	Ours	saRF	Autocontext
Whole	HS	**84.2** ($\pm$ 6.5)	80.6 ($\pm$ 6.8)	73.8 ($\pm$ 9.3)
	PC	**81.6** ($\pm$ 9.5)	74.7 ($\pm$ 8.9)	67.5 ($\pm$ 12.9)
	MM	**74.8** ($\pm$ 11.5)	68.9 ($\pm$ 9.6)	61.9 ($\pm$ 13.8)
Ribs	HS	**82.0** ($\pm$ 11.5)	76.2 ($\pm$ 11.9)	65.6 ($\pm$ 16.4)
	PC	**78.8** ($\pm$ 12.5)	70.6 ($\pm$ 12.4)	63.6 ($\pm$ 16.3)
	MM	**71.2** ($\pm$ 15.6)	66.7 ($\pm$ 13.2)	56.4 ($\pm$ 17.7)

D. Results

1) Comparison to other methods: In a first experiment, we compared our method to two other methods: autocontext [32] and saRF [24]. Autocontext was run for two iterations as this gave the best results and with a tree depth reduced to 15 to avoid overfitting. For a fair comparison, no regularisation was applied to any of the methods. Weighted DS are shown in table I for the whole skeleton and for the ribs.

Our method significantly outperformed saRF and autocontext for the task at hand, both considering the whole body and the ribs only.

2) Relevant components of our method: In a second experiment, we explored the influence of the different components in our method and of the regularisation. Overall DS of 85.6, 83.8 and 77.4 were obtained for the three datasets respectively.

Results with hierarchical model but no trilateration features (HM) on one hand and with trilateration features but no hierarchical model (TF) on the other hand are shown in table II. For the TF approach, the cascade was run for 5 iterations. It showed that the trilateration features contributed more than the hierarchical model to the increase in DS compared to the saRF method. It also demonstrated that the hierarchical model improved the DS by up to 2 points, depending on the

dataset, but only when combined with trilateration features. Without the trilateration features, the results with and without the hierarchical model were similar. This was likely because the trilateration features benefit from the supplementary centroids generated by the hierarchical model whilst the Haar-like features do not. The importance of the different features is shown in the supplementary material.

TABLE II
MEAN DS OVER SUBJECTS FOR SKELETON ANNOTATION FOR VARIATIONS ON OUR METHOD

	Method	Ours (HM+TF)	HM	TF
Whole	HS	84.2 (± 6.5)	81.3 (± 6.7)	**84.4** (± 6.6)
	PC	**81.6** (± 9.5)	74.8 (± 8.9)	79.6 (± 7.5)
	MM	**74.8** (± 11.5)	68.6 (± 9.9)	73.5 (± 9.1)
Ribs	HS	**82.0** (± 11.5)	76.9 (± 11.7)	80.7 (± 12.8)
	PC	**78.8** (± 12.5)	70.5 (± 12.3)	78.0 (± 10.9)
	MM	**71.2** (± 15.6)	62.3 (± 13.7)	69.7 (± 13.5)

Results for different iterations and with regularisation are shown in Fig. 6, 7 and 8. A table with DS for all iterations can be found in the supplementary material. The cascaded approach and its adaptive landmarks improved the segmentation scores. The largest difference was observed in the second iteration, that was the first one using trilateration features. In further iterations, landmarks were refined, which resulted in better dense segmentations. In particular in the rib region, the densely meshed landmarks helped distinguishing ambiguous regions. For all datasets, the regularisation step also improved the accuracy by better following anatomical borders.

The examples in Fig. 9 show that most errors occurred either at the interfaces between two labels or in the ribs.

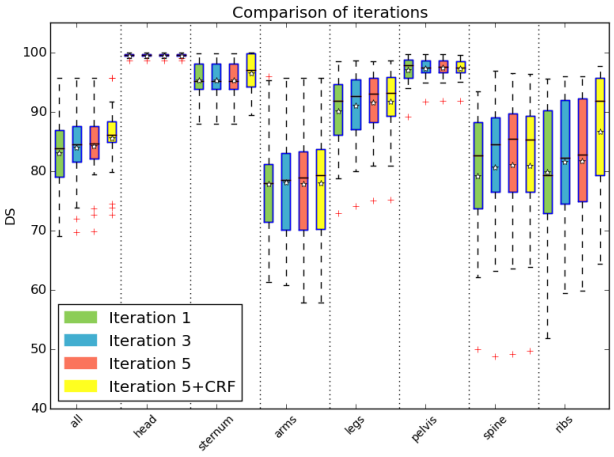

Fig. 6. DS for different groups of parts of the skeleton in the HS dataset, without initial landmarks.

3) Initial landmarks: Our method can be used without initial landmarks as done in the two previous experiments. However, if landmarks are available for the data at hand, these can be incorporated into the first classifier of the cascade in our method. Using the healthy subjects and the prostate cancer

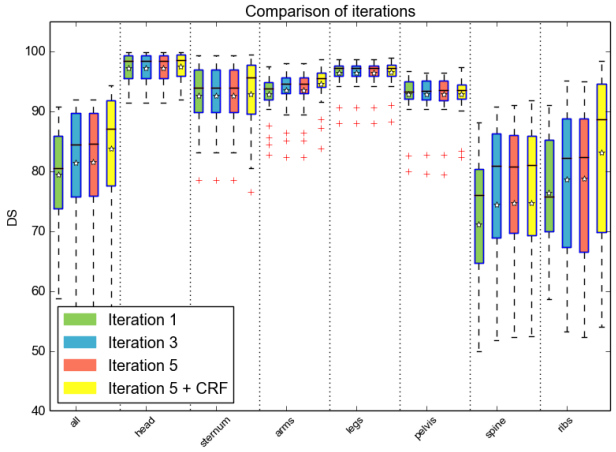

Fig. 7. DS for different groups of parts of the skeleton in the PC dataset, without initial landmarks.

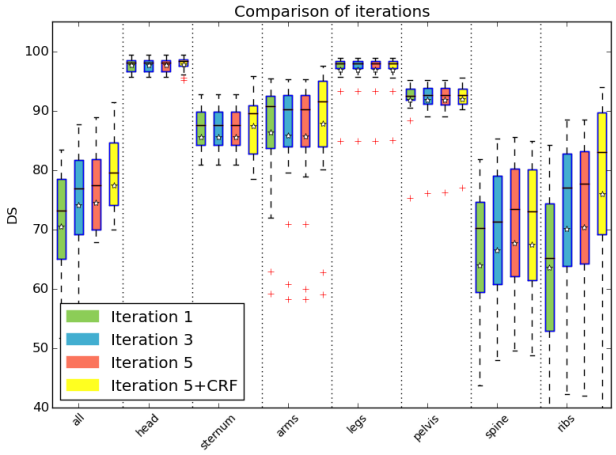

Fig. 8. DS for different groups of parts of the skeleton in the MM dataset, without initial landmarks.

datasets, for which we had landmarks annotations available, we tested the influence of these initial landmarks on the final segmentation. Overall improvements in DS of 1.6 and 0.4 were observed for the healthy subjects and the prostate cancer datasets respectively. Detailed results for all iterations can be found in the supplementary material. For the prostate cancer dataset, a *plateau* was reached at the third iterations and results did not significantly change any more. This was most likely due to the fact that, for the prostate cancer dataset, only 4 initial landmarks were available and these were not as accurate as the 57 initial landmarks of the healthy subjects dataset.

4) Transfer experiment: To show the stability of the method, we performed a transfer experiment, training on the multiple myeloma dataset and predicting on the prostate cancer dataset. This test is challenging because both datasets consist of different modalities (contrast-enhanced CT vs non contrast-enhanced CT) and have slightly different fields of view. No cross-validation was done, since training and testing datasets were different. Our method was used with the parameters described above, but only one iteration in the first level of

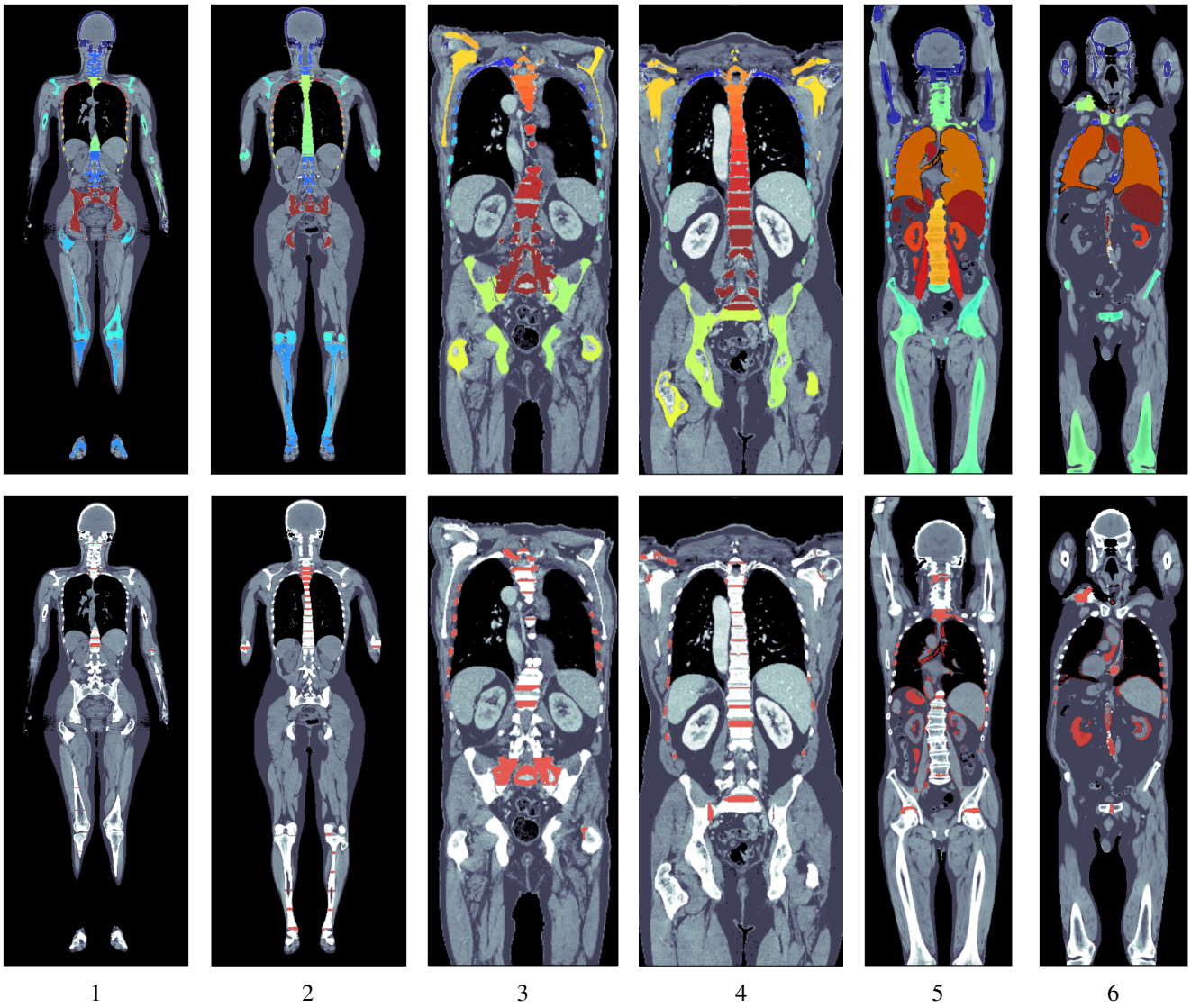

Fig. 9. First row: segmentation resulting from our method overlayed on the windowed CT image for two examples of each of the three datasets. Second row: errors in the segmentations are shown in red. Most errors occured at the interfaces between two labels or in the ribs for the bones, and in close zones of similar intensities for the organs (e.g. the trachea in columns 5 and 6). Column 6 shows a failure case for the left kydney. For columns 5 and 6, results of bones and organs segmentation are presented on the same image but have been computed separately in experiments III-D1 and III-D5. More images in 3D view can be found in the supplementary material.

the cascade. To obtain a fair comparison between methods, no regularisation was used. Results are presented in table III. Our method outperformed saRF and autocontext and obtained a DS of 64.4 for the whole body. This was as expected lower than when training and testing on the same type of dataset but shows that our method was relatively robust to changes in modality (contrast-enhanced vs non contrast-enhanced) and imaging parameters, especially minor changes in field of view. This was likely due to the use of the anatomical trilateration features which do not depend on intensities in the image.

5) Organ segmentation: In a final experiment, we explored whether organ segmentation can be performed with our method. Note that this is a more difficult task, because, in contrast to skeleton annotation where a bone mask can be easily computed, we did not use an organ mask and therefore also needed to separate the background from the structures of interest. In this experiment, a label was predicted for all voxels inside the body, and "background" was used as an additional label for voxels that did not belong to any of the structures being segmented.

TABLE III
MEAN DS OVER SUBJECTS FOR SKELETON ANNOTATION IN TRANSFER EXPERIMENT

Method	Ours	saRF	Autocontext
Whole	**64.4** (± 11.0)	57.9 (± 8.2)	54.0 (± 13.9)
Ribs	**59.8** (± 14.9)	50.0 (± 10.1)	48.9 (± 17.9)

We used the multiple myeloma dataset, because annotation was available for various non-bony structures from the VISCERAL benchmark [49]. In this experiment, the first

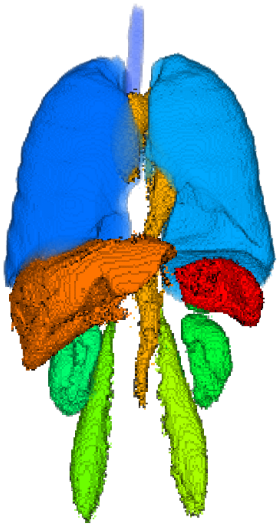

Fig. 10. Organ segmentation for one subject from the MM Dataset using our method, localising structures annotated in the VISCERAL Benchmark.

hierarchical level was used to separate the background from the structures of interest (lungs, liver, spleen, kidneys, aorta, trachea, psoas muscles), and the second to label the organs. Centroids of the bones were provided as initial landmarks to the classifier. No regularisation was done. An example is depicted in Fig. 10. Detailed results for individual organs are shown in table IV. For the lungs and the liver, our results were close to the best ones obtained in the Visceral Benchmark 2015 [49]. For other organs, our method resulted in slightly lower DS than the one obtained by the best benchmark participant. Note however that the winning method [15] is a multi-atlas method that has to perform a registration to each atlas for each structure to segment. From the information given in [15], one registration takes approximately 110-210s (with refinement). With twenty atlases, the time needed for segmenting ten organs is therefore close to five hours for one subject. The computation time of multi-atlas methods also grows linearly with the number of training subjects whilst the computation time of our method is approximately constant with respect to the quantity of training data.

The examples in Fig. 9 show that the algorithm has problems distinguishing between organs with similar intensities that are close to each other. It can be seen in the confusion between trachea and lungs as well as between stomach and spleen. Nonetheless, the segmentation we obtained were good enough to be used to perform lesion localisation in cancer staging using PET/CT image data.

IV. Discussion

The experiments showed that our method could achieve bone annotation in contrast-enhanced and non-contrast enhanced CT with high DS and outperformed saRF and Autocontext for this task (see sec. III-D1). Even in the ribs, which are the most challenging part to annotate due to similar appearance

TABLE IV
Mean DS over subjects for organ annotation

	Our method	Visceral [49]
Lungs	96.9	97.4
Liver	89.7	92.3
Spleen	80.1	87.4
Kidneys	72.9	92.5
Aorta	63.7	84.7
Trachea	74.1	93.1
Psoas muscles	74.8	85.4

and high variability amongst subjects, overall DS of over 78 were obtained for the healthy subjects and the prostate cancer dataset. For the multiple myeloma dataset, the DS for ribs was slightly lower, likely due to the larger variability in fields of view.

The experiments also showed that, in particular when initial landmarks were missing, repeated iterations within the cascaded approach with adaptive landmarks improved the accuracy of the final annotations (see sec. III-D2), and that the trilateration features were an essential component of our method. The final regularisation ensured spatial smoothness of the segmentation and helped disambiguating similar regions, which was particularly important for the rib cage. This was still valid if initial landmarks were available (see sec. III-D3). Our method can therefore be combined with an automatic landmark annotation method (e.g., the method described in [50]) to obtain a fully automatic annotation pipeline with an improved final result.

Our method is strongly based on the assumption that the positions of structures relatively to one another are approximatively constant among instances. While this assumption is not fulfilled for example in natural images segmentation, it holds for anatomical annotation in 3D medical image scans of the trunk or whole body, because all patients of a dataset are usually scanned with the same protocol, and in particular in the same position. The transfer experiment (see sec. III-D4) nonetheless showed that our method was relatively robust to changes in imaging parameters, and could handle minor changes in fields of view such as the ones present between the multiple myeloma and the prostate cancer datasets.

In our approach, we used Euclidean centroids as landmarks for our trilateration features due to their low computation time. For non-convex structures however, in particular the ribs, the centroid is not located within the structure. Whilst this does not impede the computation of the trilateration features, other ways of representing location, such as centre lines, should be explored in future work.

Using the full method with 5 iterations in the second level and regularisation took on average 13 minutes per patient. However, our method also has the advantage of allowing each user to choose his own trade-off between time and performance by computing or manually adding initial landmarks or not, choosing for each level of the cascade the number of iterations wished and using regularisation or not. Using the full method brought a gain of up to 7.0 overall DS and up to 12.6 DS in the ribs for an average cost of 9 minutes per

patient. The user's individual trade-off therefore has to depend on the performance needed for the given application.

When considering applications in oncology staging, it is also interesting to note that our method was able to segment diverse organs and muscles to an accuracy that is under the state of the art but is sufficient for oncology applications (see sec. III-D5). In PET/CT, for example, the detection of false positive regions associated with specific organs can easily be accomplished with the current accuracy. Coupled with the low computation time of our method, it makes it usable in clinical practice for example for lesion mapping at different time points in cancer patients with a large number of lesions.

V. Conclusion and outlook

We have developed a method for skeleton and organs annotation in CT images that outperforms saRF and Auto-context for skeleton annotation. Our method is based on a cascaded RF combined with a hierarchical approach with adaptive landmarks and a final CRF. It relies on anatomical trilateration features that we introduced here and can annotate the skeleton and different organs to an accuracy that will enable lesion mapping and remapping for oncological staging and handle the difficult task of generalisation between different CT acquisition protocols. The user can choose his individual application oriented trade-off between computation time and performance by adapting the length of the cascade and the use of regularisation. The reasonable computation time allows for a clinical application of the method.

Since not only the position of structures relative to one another, but also the shape of individual structures is approximately constant among subjects, using shape aware features like the one developed by Li *et al.* [51] and incorporating shape models after each iteration as in [9] or as postprocessing of the final result are promising research directions.

Acknowledgment

This work was partially funded by the German ministry for education and research (Bundesministerium für Bildung und Forschung) under Grant Agreement No. 01IS12057, by the European Union Seventh Framework Program (FP7) under Grant Agreement No. 294582 ERC MUMI and by the Austrian Science Fund (FWF I2714-B31).

References

[1] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney *et al.*, "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," *European journal of cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[2] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors," *Journal of nuclear medicine*, vol. 50, no. Suppl 1, pp. 122S–150S, 2009.

[3] E. C. Etchebehere, J. C. Araujo, P. S. Fox, N. M. Swanston, H. A. Macapinlac, and E. M. Rohren, "Prognostic factors in patients treated with 223Ra: the role of skeletal tumor burden on baseline 18F-fluoride PET/CT in predicting overall survival," *Journal of Nuclear Medicine*, vol. 56, no. 8, pp. 1177–1184, 2015.

[4] M. Bieth, M. Krönke, R. Tauber, M. Dahlbender, M. Retz, S. G. Nekolla, B. Menze, T. Maurer, M. Eiber, and M. Schwaiger, "Exploring new multimodal quantitative imaging indices for the assessment of osseous tumour burden in prostate cancer using 68Ga-PSMA-PET/CT," *Journal of Nuclear Medicine*, 2017.

[5] M. Bieth, R. Donner, G. Langs, M. Schwaiger, and B. Menze, "Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015, pp. 84.1–84.10.

[6] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, and C. Schnörr, "Spine detection and labeling using a parts-based graphical model," in *Information Processing in Medical Imaging*. Springer, 2007, pp. 122–133.

[7] S.-H. Huang, Y.-H. Chu, S.-H. Lai, and C. L. Novak, "Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI," *IEEE Transactions on Medical Imaging*, vol. 28, no. 10, pp. 1595–1605, 2009.

[8] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in CT images," *Medical Image Analysis*, vol. 13, no. 3, pp. 471–482, 2009.

[9] D. Richmond, D. Kainmueller, B. Glocker, C. Rother, and G. Myers, "Uncertainty-driven forest predictors for vertebra localization and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 653–660.

[10] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 590–598.

[11] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine CT via dense classification from sparse annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2013, pp. 262–270.

[12] Q. Wang, D. Wu, L. Lu, M. Liu, K. L. Boyer, and S. K. Zhou, "Semantic context forests for learning-based knee cartilage segmentation in 3D MR images," in *Medical Computer Vision. Large Data in Medical Imaging*. Springer, 2014, pp. 105–115.

[13] D. Wu, D. Liu, Z. Puskas, C. Lu, A. Wimmer, C. Tietjen, G. Soza, and S. K. Zhou, "A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 980–987.

[14] J. Ehrhardt, H. Handels, T. Malina, B. Strathmann, W. Plötz, and S. J. Pöppl, "Atlas-based segmentation of bone structures to support the virtual planning of hip operations," *International Journal of Medical Informatics*, vol. 64, no. 2, pp. 439–447, 2001.

[15] F. Kahl, J. Alvén, O. Enqvist, F. Fejne, J. Ulén, J. Fredriksson, M. Landgren, and V. Larsson, "Good features for reliable registration in multi-atlas segmentation," *Proceedings of the VISCERAL Challenge at ISBI*, vol. 1390, pp. 12–17, 2015.

[16] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.

[17] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal, and D. Rueckert, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Medical Image Analysis*, vol. 23, no. 1, pp. 92–104, 2015.

[18] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors," *Medical Image Analysis*, vol. 26, no. 1, pp. 1–18, 2015.

[19] J. J. Cerrolaza, R. M. Summers, and M. G. Linguraru, "Soft multi-organ shape models via generalized PCA: A general framework," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016, pp. 219–228.

[20] C. Wachinger, K. Fritscher, G. Sharp, and P. Golland, "Contour-driven atlas-based segmentation," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2492–2505, 2015.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2001, pp. I–511.

[23] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi, "Geof: Geodesic forests for learning coupled predictors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 65–72.

[24] L. Peter, O. Pauly, P. Chatelain, D. Mateus, and N. Navab, "Scale-adaptive forest training via an efficient feature sampling scheme," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 637–644.

[25] M. P. Heinrich and M. Blendowski, "Multi-organ segmentation using vantage point forests and binary context features," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016, pp. 598–606.

[26] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi, "Neighbourhood approximation using randomized forests," *Medical image analysis*, vol. 17, no. 7, pp. 790–804, 2013.

[27] H. Lombaert, D. Zikic, A. Criminisi, and N. Ayache, "Laplacian forests: semantic image segmentation by guided bagging," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 496–504.

[28] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *European Conference on Computer Vision*. Springer, 2012, pp. 870–881.

[29] Y. Gao, Y. Shao, J. Lian, A. Z. Wang, R. C. Chen, and D. Shen, "Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests," *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1532–1543, 2016.

[30] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Medical image analysis*, vol. 23, no. 1, pp. 70–83, 2015.

[31] A. Montillo, J. Tu, J. Shotton, J. Winn, J. Iglesias, D. Metaxas, and A. Criminisi, "Entanglement and differentiable information gain maximization," in *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013, pp. 273–293.

[32] Z. Tu, "Auto-context and its application to high-level vision tasks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[33] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Winter 1989.

[34] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 460–468.

[35] J. Lafferty, A. McCallum, F. Pereira *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol. 1, 2001, pp. 282–289.

[36] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[37] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," *arXiv preprint arXiv:1210.5644*, 2012.

[38] M. G. Linguraru, J. A. Pura, V. Pamulapati, and R. M. Summers, "Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT," *Medical image analysis*, vol. 16, no. 4, pp. 904–914, 2012.

[39] A. Saito, S. Nawano, and A. Shimizu, "Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs," *Medical image analysis*, vol. 28, pp. 46–65, 2016.

[40] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck, and S. K. Zhou, "Automatic multi-organ segmentation using learning-based segmentation and level set optimization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2011, pp. 338–345.

[41] B. Fulkerson, A. Vedaldi, S. Soatto *et al.*, "Class segmentation and object localization with superpixel neighborhoods." in *International Conference on Computer Vision (ICCV)*, vol. 9. Citeseer, 2009, pp. 670–677.

[42] Z. Tian, L. Liu, Z. Zhang, and B. Fei, "Superpixel-based segmentation for 3D prostate MR images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 3, pp. 791–801, 2016.

[43] V. Zografos, A. Valentinitsch, M. Rempfler, F. Tombari, and B. Menze, "Hierarchical multi-organ segmentation without registration in 3D abdominal ct images," in *MCV workshop (MICCAI 2015)*, 2015.

[44] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images," *NeuroImage*, vol. 57, no. 2, pp. 378–390, 2011.

[45] S. Tomoshige, E. Oost, A. Shimizu, H. Watanabe, and S. Nawano, "A conditional statistical shape model with integrated error estimation of the conditions; application to liver segmentation in non-contrast CT images," *Medical image analysis*, vol. 18, no. 1, pp. 130–143, 2014.

[46] R. Meier, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes, "A hybrid model for multimodal brain tumor segmentation," *Multimodal Brain Tumor Segmentation*, vol. 31, 2013.

[47] B. Andres, T. Beier, and J. Kappes, "OpenGM: A C++ library for discrete graphical models," *ArXiv e-prints*, 2012.

[48] K. Gruenberg, M. A. Weber, O. A. Jiménez del Toro, O. Goksel, B. Menze, H. Müller, G. Langs, I. Eggel, M. Holzer, G. Kontokotsios, M. Krenn, R. Schaer, A. A. Taha, M. Winterstein, and A. Hanbury, "Visceral-visual concept extraction challenge in radiology: Segmentation challenge: overview, insights and preliminary results," in *European Congress of Radiology (ECR) 2015*, Vienna, Austria, 2015.

[49] O. Göksel, O. A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, "Overview of the VISCERAL challenge at ISBI 2015," in *Proceedings of the VISCERAL Anatomy Grand Challenge at the 2015 IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2015.

[50] R. Donner, B. H. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization," *Medical Image Analysis*, vol. 17, no. 8, pp. 1304–1314, 2013.

[51] Y. Li, C. Pahng Ho, N. Chahal, R. Senior, and M.-X. Tang, "Myocardial segmentation of contrast echocardiograms using random forests guided by shape model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016, pp. 158–165.