

Dissertation

Localising Anatomical Structures and Quantifying Tumour Burden in PET/CT Images using Machine Learning

Marie Bieth



Fakultät für Informatik
Technische Universität München



TECHNISCHE UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR INFORMATIK



Localising Anatomical Structures and Quantifying Tumour Burden in PET/CT Images using Machine Learning

Marie Bieth

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des Akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. M. Althoff
Prüfer der Dissertation: 1. Prof. Dr. B. Menze
2. Prof. Dr. M. Schwaiger

Die Dissertation wurde am 30.08.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 02.11.2017 angenommen.

ABSTRACT

With increasing life expectancy, cancer has become a more frequent disease. At the same time, progress in medicine has improved its outcome. In particular, advances in medical imaging techniques allow for better diagnosis and therapy response evaluation of cancer. However, the analysis of images is as important as their acquisition, and this analysis is often still performed manually by medical doctors. Manual analysis is very time consuming and impairs reproducibility. Moreover, most analysis methods are not quantitative. Machine learning can help fill the need for automatic and quantitative analysis methods.

This thesis presents contributions in the fields of automatic anatomical segmentation and quantitative analysis of medical images by using machine learning. Our work focusses on the analysis of positron emission tomography (PET) / computed tomography (CT) images for metastasised prostate cancer. The first contribution consists of new methods for automatic segmentation of anatomical structures in CT images. We present two methods for automatic localisation of bones and organs in CT images that rely on the use of context information within an iterative random forest framework. Our evaluation of the methods on real CT data showed that they exhibit high Dice scores and that the use of contextual information is key to their state of the art performance. The second contribution consists of new quantification indices for bone metastasis assessment in PET/CT images and a method to compute these quantities automatically with the possibility of manual corrections. An evaluation of the method on a metastasised prostate cancer patient cohort showed that the indices provide useful clinical information for therapy response evaluation. Moreover, we extended these indices to perform regional quantification and concluded that, in our patient cohort, lesions were not uniformly distributed in the skeleton and mixed responses to therapy were frequent. Finally, the computational efficiency of the methods presented in this work allows for their use in clinical practice, and they can be computed quickly even on modest hardware.

ZUSAMMENFASSUNG

Aufgrund der höheren Lebenserwartung tritt die Krebs immer häufiger auf. Gleichzeitig haben Fortschritte in der Medizin die Behandlungsmöglichkeiten der Krankheit verbessert. Insbesondere erlaubt die Fortentwicklung der medizinischen Bildgebung eine bessere Diagnose und Bewertung der Effektivität der Therapie. In diesem Bereich ist die Bildanalyse genauso wichtig wie die Erstellung der Bilder. Die Analyse wird aber oft immer noch manuell von Ärzten durchgeführt. Dieses Vorgehen ist zeitraubend und verschlechtert die Reproduzierbarkeit. Weiterhin sind die Analysemethoden meist nicht quantitativ. Machinelles Lernen kann dazu beitragen, den Bedarf für automatische und quantitative Analysemethoden abzudecken.

In dieser Doktorarbeit werden Beiträge zu den Bereichen der automatischen anatomischen Segmentierung und der quantitativen Analyse von medizinischen Bildern durch maschinelles Lernen beschrieben. Die vorliegende Doktorarbeit konzentriert sich auf die Analyse von Positronen Emissions Tomographie (PET) und Computertomographie (CT) Bildern für metastasierten Prostatakrebs. Der erste Teil der Arbeit stellt zwei neue Methoden zur automatischen Segmentierung von Knochen und Organen in CT Bildern vor. Die Methoden basieren auf der Auswertung von kontextuellen Informationen innerhalb von iterativem Random Forest Klassifikationsalgorithmus. Die detaillierte Auswertung anhand echter CT Bilder zeigte, dass die Methoden hohe Dice Scores erreichen und dass die Benutzung von kontextuellen Informationen wesentlich zu deren Leistung beisteuert. Der zweite Teil der Dissertation besteht aus neuen Indices für die Bewertung von Knochenmetastasen in PET/CT Bildern und einer Methode zur automatischen Berechnung dieser mit der Möglichkeit für manuelle Korrekturen. Die Auswertung der Methode anhand einer Kohorte von metastasierten Prostatakrebspatienten zeigte, dass die Indizes nützliche klinische Informationen zur Bewertung der Effektivität der Therapie liefern. Außerdem wurden die Indices zu einer regionalen Quantifizierungsmethode erweitert und es wurde gezeigt, dass Knochenläsion in der Patientenkohorte nicht gleichmäßig im Skelett verteilt waren, und dass gemischte Therapieeffektivität häufig auftrat. Die rechnerische Effizienz aller in dieser Dissertation vorgestellten Methoden ermöglicht ihre Benutzung in der klinischen Praxis und ihre schnelle Berechnung auch mit älterer Hardware.

ABRÉGÉ

Suite à l'augmentation de l'espérance de vie, le cancer est devenu une maladie de plus en plus fréquente. Parallèlement, les progrès en médecine ont amélioré sa prise en charge. En particulier, le perfectionnement des techniques d'imagerie médicale permet un diagnostic plus précis et une meilleure évaluation de l'efficacité des traitements. Pour cela, l'analyse des images est aussi importante que leur acquisition. Néanmoins l'analyse est encore souvent menée manuellement par les médecins, ce qui est coûteux en temps et diminue la reproductibilité. De plus, la plupart des méthodes d'analyse ne sont pas quantitatives. Les algorithmes d'apprentissage peuvent aider à combler le manque de méthodes d'analyse automatiques et quantitatives.

Dans cette thèse, nous présentons des contributions dans les domaines de la segmentation automatique et de l'analyse quantitative d'images médicales. Notre travail se concentre sur l'analyse d'images hybrides de tomographie par émission de positons (TEP) et de tomodensitométrie (TDM) de patients atteints de cancer de la prostate métastasé. La première contribution comprend de nouvelles méthodes pour la segmentation automatique de structures anatomiques dans des images TDM. Nous présentons deux méthodes pour la localisation automatique d'os et d'organes qui reposent sur l'utilisation d'informations contextuelles dans le cadre de random forests itératives. Notre évaluation de ces algorithmes à l'aide d'images de TDM a montré qu'elles produisent de hauts Dice Scores et que l'utilisation d'informations contextuelles concourt de façon essentielle à leur performance. La seconde contribution comprend de nouveaux indices pour l'appréciation de métastases osseuses dans des images TEP/TDM et une méthode pour leur calcul automatique avec la possibilité d'effectuer des corrections manuelles. L'évaluation de cette méthode à l'aide d'un panel de patients atteints de cancer de la prostate métastasé a montré que les indices fournissent des informations cliniques utiles pour évaluer l'efficacité des traitements. De plus, nous avons étendu ces indices pour permettre une quantification locale et conclu que, dans notre panel de patients, les lésions n'étaient pas réparties uniformément dans le squelette et que les réactions mitigées à la thérapie étaient fréquentes. Enfin, la rapidité de calcul des méthodes présentées ici permet leur utilisation en milieu clinique, même avec du matériel informatique modeste.

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisors Prof. Menze and Prof. Schwaiger for their support during the last three years. Furthermore, I would like to thank Dr. Matthias Eiber, Dr. Mona Mustapha and Dr. Karina Knorr for always finding time in their busy schedules to meet, discuss and help out with the medical aspect of my work and Dr. Stefan Nekolla for answering many of my questions. I would also like to thank Dr. Loic Peter for his time and his very valuable feedback on my work.

My thanks also go to my colleagues of the IBBM group, and in particular to Esther Alberts, Jana Lipkova and Markus Rempfler for our discussions on various topics, and to my office mates at the nuclear medicine clinic Dr. Jorge Cabello, Karl Kunze, Negar Omidvari, Giaime Rancan and Dr. Ian Somlai-Schwaiger for the moral support and the shared cakes. Thanks also to Esther, Chris and Jakob who took the time to comment on various parts of this thesis and made it clearer.

Finally, I would like to express my gratitude to my family and friends for their support and encouragement in all my decisions, and to Chris for always finding the right word to make me smile, in good as well as in bad days.

Je remercie du fond du coeur mes parents, mes grands-parents, mes frères et mes amis pour leur soutien et leurs encouragement dans toutes mes décisions.

TABLE OF CONTENTS

ABSTRACT	i
ZUSAMMENFASSUNG	iii
ABRÉGÉ	v
ACKNOWLEDGMENTS	vii
TABLE OF CONTENTS	viii
I Background information	1
1 Introduction	3
1.1 Contributions	3
1.2 Outline of the thesis	4
2 Medical Image Analysis	5
2.1 Background on cancer	5
2.1.1 Prostate Cancer	5
2.2 Introduction to medical imaging	6
2.2.1 X-ray computed tomography	7
2.2.2 Positron emission tomography	8
2.2.3 Planar scintigraphy	9
2.2.4 Hybrid PET/CT	10
2.3 Medical image analysis	10
2.3.1 CT analysis	11
2.3.2 PET analysis	12
2.3.3 Bone scintigraphy analysis	14
2.4 Conclusion	14
3 Image Segmentation	15
3.1 Introduction	15
3.1.1 Problem posing	15
3.1.2 Supervised learning	16
3.1.3 Performance evaluation	17
3.2 Feature representation	18
3.2.1 Definition	18
3.2.2 Feature examples	18
3.3 Random forests	20
3.3.1 Decision Trees	20
3.3.2 Injecting randomness	23
3.3.3 The ensemble model	24
3.4 Multiclass medical image segmentation	26
3.4.1 Random forest based methods	26
3.4.2 Atlas and model based methods	26
3.4.3 Deep learning methods	27
3.4.4 Conditional random field methods	28
3.5 Conclusion	29

II	Localisation and Quantification for Cancer Staging in PET/CT images	31
4	Segmentation of Skeleton and Organs in Whole-Body CT Images	33
4.1	Introduction	33
4.1.1	State of the art	34
4.1.2	Contributions	36
4.2	Methods	36
4.2.1	Overview	36
4.2.2	Notations	37
4.2.3	Scale Adaptive Random Forests	37
4.2.4	Features Description	38
4.2.5	Cascaded anatomical trilateration	40
4.2.6	Hierarchical segmentation	41
4.2.7	Regularisation	41
4.3	Experiments	42
4.3.1	Data sets	42
4.3.2	Setup	43
4.3.3	Computing time	45
4.3.4	Results	46
4.4	Discussion	54
4.5	Conclusion and outlook	56
5	From Large to Small Organ Segmentation in CT	57
5.1	Introduction	57
5.2	Methods	58
5.2.1	Vantage Point Forest	59
5.2.2	Initial labelling	59
5.2.3	Iterated Forest with regional context descriptors	60
5.2.4	Final shape voting	61
5.3	Experiments	62
5.4	Conclusion	64
6	Bone PET Indices for Bone Lesion Burden Staging in PET/CT	67
6.1	Motivation	67
6.2	Material and methods	68
6.2.1	Bone PET Index	68
6.2.2	Automatic computation method	69
6.2.3	Patient cohort	71
6.2.4	Data acquisition and analysis	72
6.2.5	Statistical analysis	73
6.3	Results	73
6.3.1	Technical validation	73
6.3.2	Quantification using BPI_{VOL} , SUV_{mean} and BPI_{SUV}	76
6.3.3	Correlation of BPI to clinical parameters	76
6.4	Discussion	78
6.5	Conclusion	81
7	Localised quantification of Bone Lesion Burden in PET/CT	83
7.1	Motivation	83
7.2	Material and methods	84
7.2.1	Localised Bone PET Index	84
7.2.2	Automatic computation method	84
7.3	Results	86
7.3.1	Individual cases	86

7.3.2	Analysis of the patient cohort	89
7.4	Discussion	90
7.5	Conclusion	92
III	Summary and outlook	93
8	Summary and outlook	95
8.1	Summary	95
8.2	Future work	96
8.2.1	Anatomical structure localisation in PET/CT images	96
8.2.2	Lesion segmentation in PET/CT images	96
8.2.3	Quantitative analysis of PET/CT images	97
IV	Appendices	99
A	Mathematical notations	101
A.1	Symbols	101
A.2	Number domains	101
A.3	Functions	101
B	List of publications	103
B.1	Journal publications	103
B.2	Conference proceedings	103
B.3	Conference abstracts	103
	LIST OF FIGURES	105
	LIST OF TABLES	107
	REFERENCES	108

Part I

Background information

Chapter 1

Introduction

With increasing life expectancy, cancer has become a more frequent disease. In 2012, female Americans had a 42.1%, male Americans a 37.6% risk of developing an invasive cancer during their life. However, due to rapid progress in medicine, both in diagnosis and treatment, the five-years survival rate has improved by 20% between 1975 and 2011 [20].

Since the discovery of X-rays in 1895, medical imaging techniques have improved and made it an essential tool for diagnosis, staging, and therapy response evaluation in diverse types of cancer. Nowadays, several techniques can produce a volumetric image of the body, and show even small structures with sizes below a millimetre. As more and more complex data is produced by imaging techniques, it is of the utmost importance that analysis techniques for these images improve concurrently. Because the amount of data produced has also increased drastically, to assist medical doctors in their assessment of the patient, new analysis methods should be as automatic as possible. Machine learning is therefore the tool of choice to develop such methods.

1.1 Contributions

In this thesis, I was interested in automatic localisation of anatomical structures and quantification of lesions in PET/CT images. I applied my methods to metastasised prostate cancer images. My contributions belong to two complementary fields:

- anatomical segmentation using machine learning: I present methods for segmenting bones (chapter 4) and organs (chapter 5) in CT images by using contextual information. The resulting segmentations can be used in combination with the lesion delineation to deduce the localisation of the lesions in the body.
- quantitative analysis: I present a new global quantitative analysis method for assessing bone lesions in PET/CT images (chapter 6). This method can be extended to localised quantification by using anatomical segmentation to obtain location information

(chapter 7). I applied both methods to a metastasised prostate cancer patient cohort, but they could also be used for other types of cancer.

1.2 Outline of the thesis

This thesis comprises three parts. The first part provides background information to the reader about concepts essential to the comprehension of this thesis. These are of medical and technical nature:

- *Chapter 2* introduces cancer in general and prostate cancer in particular, describes a number of medical imaging techniques and the related analysis methods.
- *Chapter 3* introduces the concept of image segmentation, and different methods for addressing it. Random Forests are described in detail.

The second part presents new methods for image segmentation and analysis:

- *Chapter 4* introduces a new, registration-free method for skeleton annotation in CT images that relies on hierarchical and iterative localisation of structures using context information.
- *Chapter 5* introduces a new method for organ segmentation in CT that focuses on small organs, does not require any deformable registration to be performed, and leverages semantic as well as image context information.
- *Chapter 6* introduces a new quantification method for bone metastases assessment in prostate cancer PET/CT images.
- *Chapter 7* extends the indices presented in chapter 6 to perform regional quantification of bone metastases in prostate cancer PET/CT images.

Finally, in the third part, *Chapter 8* concludes the thesis by discussing the work presented in the following and suggesting directions for future work.

Chapter 2

Medical Image Analysis

In this chapter, we present some medical information for readers that are unfamiliar with the medical field. In particular, we provide the necessary background on cancer, introduce several modalities used in medical imaging, and methods to analyse these images.

2.1 Background on cancer

Cancer is a group of diseases characterised by the uncontrolled proliferation of cells. Normal human cells are able to specialise into different functions, divide to provide new cells to the body when needed, and die when they become damaged or when they receive specific signals from the body. On the contrary, cancerous cells escape the proliferation control mechanisms, for example by not responding to death signals, and keep dividing, thus forming growths called tumours. Cancerous cells are also able to migrate to nearby tissues or through the blood or lymph systems. Malignant (i.e. cancerous) tumours may therefore spread to other parts of the body and lead to secondary metastatic tumours. Cancer can affect nearly any part of the human body; the most frequent cancers are lung, breast and prostate cancers. Cancer is the second cause of mortality in the USA and is responsible for more than 500 000 deaths per year [20].

Cancer should not be confused with benign tumours, where the cells proliferate without control but can not migrate to other tissues. These benign tumours can often be removed surgically and are therefore not life-threatening. A notable exception is the brain where even benign tumours can have severe consequences.

2.1.1 Prostate Cancer

Prostate cancer is the third most frequent type of cancer in the USA. 180 890 new cases were diagnosed in 2016 [20].

Anatomy The prostate is a small gland that belongs to the male reproductive system and produces a fluid that is part of the semen. When healthy, it measures a few centimetres, lies under the bladder and next to the rectum and surrounds the urethra.

Disease and diagnosis Prostate cancer is mostly asymptomatic, as long as it does not metastasise to other tissues. Therefore, in many European countries, a regular screening is proposed to older men in the form of Prostate-Specific-Antigen (PSA) blood tests. The PSA protein is produced by the prostate and present in small quantity in the blood of men with normal prostate function. In case of prostate cancer, its blood level often increases. In case of increased PSA values, a digital rectal examination, a biopsy (tissue sample extraction) and imaging can be performed to confirm the cancer diagnosis.

In advanced stages, prostate cancer can spread to other tissues, in particular bones and lymph nodes: it was reported in an autopsy study that from all patients with confirmed prostate cancer, 35 % were affected by distant metastases, of whom 90 % had bone metastases [18]. In prostate cancer, bone metastases are almost always osteoblastic (i.e. bone tissues are produced) and can cause complication such as pain and spinal cord compression. Medical imaging is used for diagnosing metastases.

Treatments Depending on the stage of the disease and the age and health of the patient, a simple active surveillance may be proposed, i.e. the tumour is closely monitored, but no treatment is applied. If the cancer is still localised in the prostate only, treatment possibilities include radical prostatectomy and radiotherapy or a combination. For metastasised cancer, treatment possibilities include hormone therapy, chemotherapy and immunotherapy. New treatments based on alpha and beta radiations are currently being developed.

2.2 Introduction to medical imaging

Medical imaging is the process by which an image of the interior of the (human) body is created. It is widely used for diagnosis and staging of pathologies, including primary and secondary tumours of prostate cancer, as mentioned in the previous section. Since the discovery of X-rays by Röntgen in 1895, several imaging modalities have been developed that show different types of information. In particular, modalities can show structural information (i.e. anatomical structures) or functional information (i.e. physiological activity).

In the following, we explain the principles of X-ray computed tomography (CT), positron emission tomography (PET) and bone scintigraphy. The goal is to provide a basic understanding to uninformed readers and not to offer an in-depth technical description of each modality. References are given where such descriptions can be found.

2.2.1 X-ray computed tomography

CT is a three-dimensional structural imaging modality.

Schematically, an X-ray CT scanner is composed of a rotating X-ray source and detectors. The source emits X-rays that traverse the object to be imaged before being measured by the detectors. X-rays are attenuated by the object, to an amount depending on the materials composing it. By rotating the source and detectors, all directions through the object are acquired during a scan. This allows for reconstruction of an image of the object that reflects the absorption coefficient of the different materials/tissues by inverting the Radon transform [85]. Typically, air exhibits values around -1000 Hounsfield units (HU), water around 0 HU and bone above 200 HU. An illustrative representation of a scanner and a slice from a thorax CT image are shown in Figure 2.1. A detailed description of imaging and reconstruction can be found in [48].

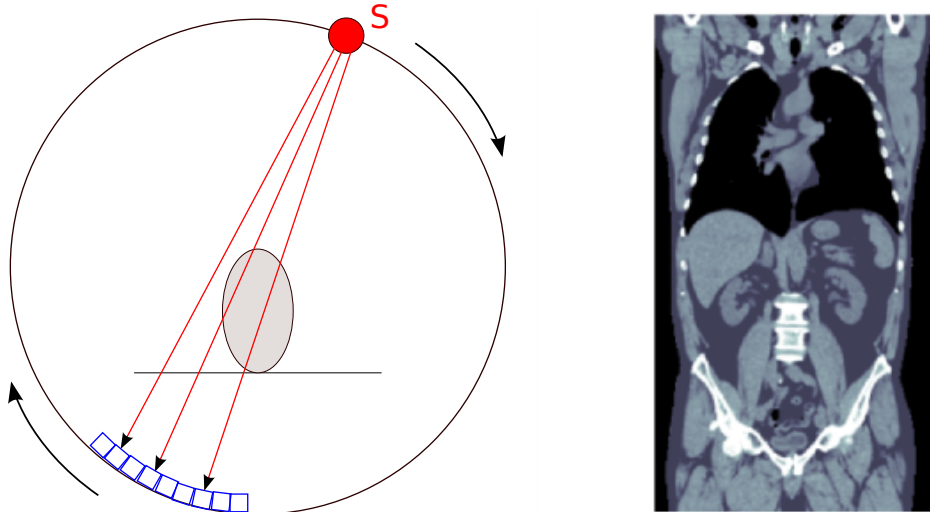


Figure 2.1 – Left: Schematic view of a CT scanner. The X-rays emitted from the rotating source S traverse the object before reaching the detectors. Right: Example of a slice from a thorax CT image. Bones have a larger absorption coefficient than other tissues and therefore appear brighter.

An abdomen-pelvis CT causes a radiation exposure of around 15 mSv [98], which is nearly four times the dose received in one year from the environment by a person living in Germany [117].

2.2.2 Positron emission tomography

PET is a three-dimensional functional imaging modality.

Physical principles PET is based on the decay of atom isotopes with a neutron deficit. These isotopes are unstable and decay by converting one of their protons into a positron e^+ and a neutron. Because of the proton loss, the chemical element changes with the decay. The emitted positron e^+ then annihilates with an electron e^- , producing two photons γ with energy 511 keV (Figure 2.2). Typical isotopes used for PET imaging are ^{11}C , ^{15}O and ^{18}F .

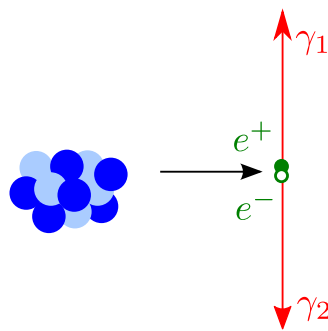


Figure 2.2 – Schematic representation of an atom of ^{11}C decaying. Protons are represented in dark blue, neutrons in light blue, positrons and electrons in green and photons in red. Adapted from [10].

Radiotracers A radiotracer for PET is a molecule that is labelled with positron emitting unstable isotopes, i.e. some atoms in the molecule have been replaced by positron emitting ones. Different radiotracers target different biological processes. For example, ^{18}F -Fluorodeoxyglucose (FDG) is an analogue of glucose and shows the glucose uptake of tissues whilst ^{68}Ga -PSMA-HBED-CC binds to the prostate specific membrane antigen (PSMA), a protein of the membrane of (normal and cancerous) prostate cells and can therefore be used to detect its presence.

PET Scan A PET scanner consists of photon detectors surrounding the object to image. Before a scan, a radiotracer is injected to the patient. During the scan, the detectors record the photons produced by the radiotracer decay. After acquisition, an image can be reconstructed that shows the spatial emission distribution of photons and therefore the concentration of the radiotracer in the body of the patient. Note that the scanner itself does not produce any radiation. A schematic representation of a PET scanner and an example of a PET image are shown in Figure 2.3. A detailed description of imaging and reconstruction can be found in [22].

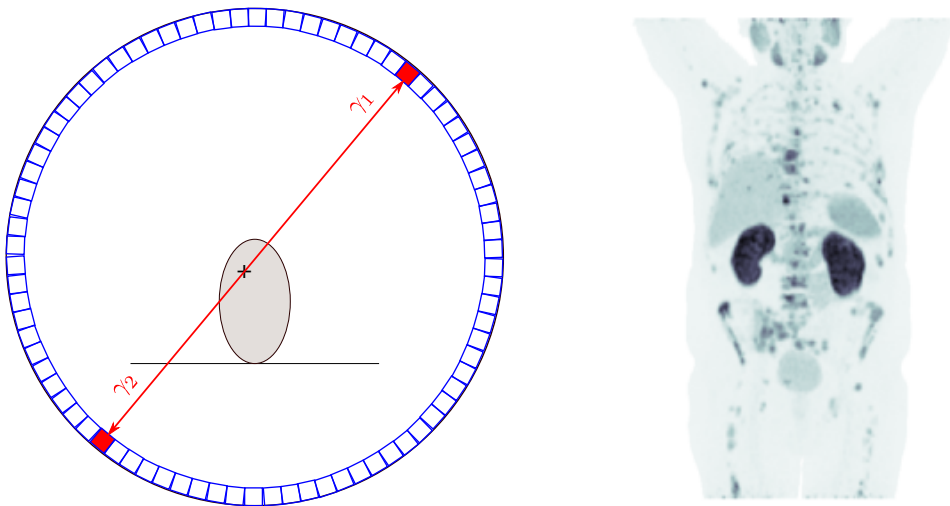


Figure 2.3 – Left: Schematic view of a PET scanner. The pair of photons emitted by one decay event in the object reaches the detectors. In practice, more complicated situation can arise (singles, triples, scattering, ...). Adapted from [10]. Right: Maximum intensity projection of a PSMA-PET image. The kidneys show a high uptake of the radiotracer.

2.2.3 Planar scintigraphy

Planar scintigraphy is a two-dimensional functional imaging modality.

Physical principles Like PET, scintigraphy relies on the radioactive decay of unstable atoms. For scintigraphy however, the unstable atomic nuclei with excess energy decay through gamma emission: the nucleus moves from a high energy state to a lower energy state by emitting a photon. The numbers of protons and neutrons in the nucleus do not change, so that the chemical element stays the same.

Scintigraphy acquisition Before acquiring a planar scintigraphy, a gamma-ray emitting radio tracer is injected to the patient. A gamma-camera [7] is then used to record the gamma-rays emitted by the radiotracer. Its detectors are arranged in a plane, so that only a planar image is obtained. This image is a projection of the decay events on the detector plane. As a result, overlapping structures can not be distinguished.

For bone scintigraphy, a typically ^{99m}Tc labelled radiotracer that targets osteoblasts (i.e. bone forming cells) is injected to the patient around three hours before the scan. The images show areas where the bone has a high turn-over rate. An anterior and a posterior projection are produced. An increased signal can amongst others indicate a fracture, a cancerous lesion or an infection. A schematic representation of a gamma-camera and an example of bone scintigraphy image are shown in Figure 2.4. A more detailed description can be found in [79, 81].

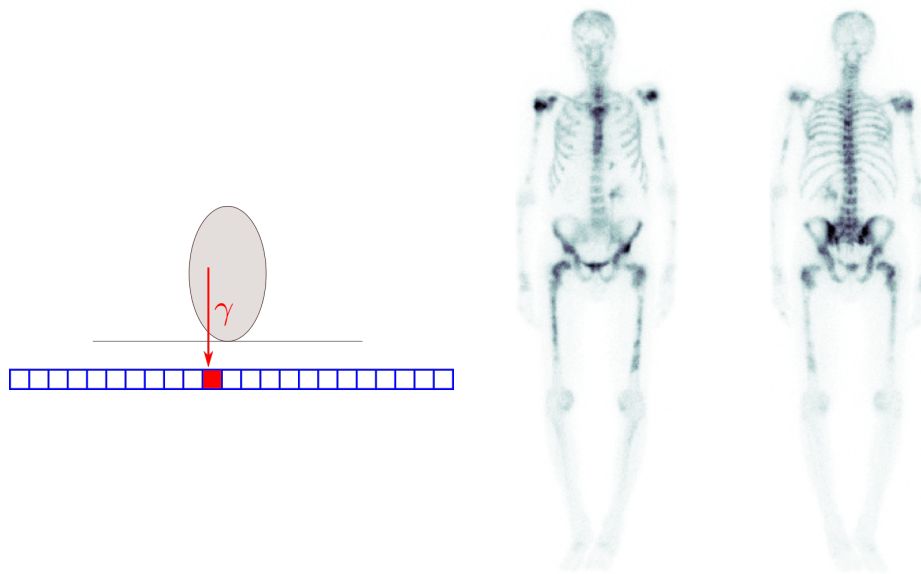


Figure 2.4 – Left: Schematic view of a gamma camera: a photon emitted by the radiotracer reaches a detector. Right: bone scintigraphy example: anterior and posterior projections. The dark areas have a high bone turn-over rate, which can be caused for example by arthrosis or bone lesions.

2.2.4 Hybrid PET/CT

Structural and functional imaging modalities show different types of information that are often both necessary to establish a diagnosis. However, if both imaging procedures are performed on different scanners, the lack of alignment between modalities can considerably impair further steps.

To solve this problem, hybrid PET/CT scanners have been designed. On these scanners, both procedures can be acquired consecutively without moving the patient. PET and CT are therefore naturally aligned and the CT image is usually used in the PET reconstruction for scatter and attenuation correction. Misalignments due to patient movements and breathing can nonetheless not be excluded. The first prototype has been designed in the late 1990s [9] and the first commercial scanner appeared in 2001.

2.3 Medical image analysis

From a medical point of view, when looking at an image of a patient, it is essential to be able to say whether the patient is sick, but also *how* sick he is, as this can impact the therapy decision. In case of repeated imaging of the same patient, it is also important to be able to evaluate the evolution of the disease, and the possible response to therapy. Standard methods

for cancer staging exist, depending on the imaging modality. We present some of these for CT, PET and bone scintigraphy in the following.

2.3.1 CT analysis

Response evaluation criteria in solid tumours (RECIST) The RECIST method [33] has been designed for therapy response evaluation in anatomical images (CT, MR, or X-ray). It is based on the diameter values of target lesions. More precisely, lesions are first assessed for measurability. Amongst others, the following criterion are taken into account:

- the lesion must be accurately measured in at least one dimension,
- its longest diameter must be over a modality specific minimum value,
- if the lesion is in a lymph node, the lymph node must have a short axis longer than 15 mm,
- irradiated lesions are not measurable, unless progress has been shown,
- osteoblastic lesions are not measurable.

Up to five targets lesions are then selected amongst the largest lesions so that they are representative of the involved organs (with a limit of two target lesions per organ) and reproducibly measurable. The sum of longest diameters of all target lesions (shortest diameter for lymph nodes) is then considered to classify the disease evolution. The following categories are applied:

- complete response if all target lesions have disappeared,
- partial response if a decrease of at least 30% of the sum of diameters is observed,
- progressive disease if an increase of at least 20% and 5 mm of the sum of diameters is observed,
- stable disease if none of the other categories apply.

Limitations The RECIST method presents two major limitations. First, the measurement of lesions is unidimensional (longest diameter), which can not fully represent the evolution of a volume. Second, many lesions are considered non-measurable. This is in particular a problem for prostate cancer, as many patients present osteoblastic bone metastases that can not be assessed with this method.

2.3.2 PET analysis

Standardised uptake value (SUV) The SUV is a way to correct the activity read from the PET image for injected dose and patient weight. It is defined as follows at voxel \mathbf{v} :

$$\text{SUV}(\mathbf{v}) = \frac{\text{PET}(\mathbf{v}) \times \text{patient weight}}{\text{injected dose}} \quad (2.1)$$

When there is a long delay between injection and image acquisition, the injected dose is corrected for decay. The usual unit used for SUV is g/mL. It is also debated that, instead of correcting for patient weight, the SUV should correct for lean body mass or for body surface area.

The SUV is computed for each voxel of the PET image. However, for describing a lesion, usually only one number is used. The following are common SUV variants:

- SUV_{max} : maximum value in the lesion,
- SUV_{peak} : average SUV in a small sphere around the most intense voxel of the lesion,
- SUV_{mean} : average SUV in the lesion (the lesion being either manually drawn or defined as an isocontour of a percentage of SUV_{max}).

Note that, although it is the most widely used measure [109], SUV_{max} relies on only one voxel in the lesion and is therefore a very noisy measure.

Metabolic tumour volume (MTV) The metabolic tumour volume is defined as the overall lesion volume, segmented from the PET image:

$$\text{MTV} = \sum_{\mathbf{v} \in \text{lesions}} \text{voxel volume} \quad (2.2)$$

The segmentation can be done manually, or defined as an isocontour of a percentage of SUV_{max} , which leads to different numerical results. Note that the lesion contours defined from the PET image may not match those shown by anatomical imaging (CT, MR). MTV was shown to be outcome predictive for example in head and neck [63] and lung [67] cancer.

Total lesion glycolysis (TLG) The total lesion glycolysis is defined for ^{18}F -FDG PET and incorporates volume and uptake information:

$$\text{TLG} = \sum_{\text{lesion } l} \text{volume}(l) \times \text{SUV}_{\text{mean}}(l) \quad (2.3)$$

As for the MTV, different lesion segmentations are possible and give different numerical results. TLG was shown to be recurrence predictive in pancreas cancer [66] and outcome predictive in epithelial ovarian cancer [23] for example.

PET response criteria in solid tumours (PERCIST) The PERCIST method [109, 51] has been designed for therapy response assessment in ^{18}F -FDG PET, but can be used with other tracers as well. In a nutshell, target lesions are chosen, and their uptake is quantified. According to the changes between two images, the disease is classified as responsive to therapy, stable or progressive.

More precisely, the most intense lesions that fulfil the following criteria are first chosen as target lesions:

- up to two target lesions can be located in the same organ,
- the SUV_{peak} before treatment of all target lesions should be superior to 1.5 mean SUV in the liver plus 2 standard deviations of the SUV in the liver (if the liver is diseased, 2 blood pool activity plus 2 standard deviations of the activity in the mediastinum is used instead),
- for a precise measurement, target lesions should be at least 2 cm in diameters, but smaller lesions can be assessed as well.

Then, the SUV_{peak} of all target lesions are computed and summed to obtain a single index before and after treatment. The disease evolution is then classified according to the following criteria:

- complete metabolic response if all lesions have disappeared,
- partial response if the index decreased of at least 30%, the difference between the most intensive lesion before and after therapy (not necessarily the same lesion) is at least 0.8 g/mL, and there are no new lesions,
- progressive disease if the index increased of at least 30% and the difference between the most intensive lesion before and after therapy (not necessarily the same lesion) is at least 0.8 g/mL, or if there are new lesions,
- stable disease if none of the other categories correspond.

Note that the above definition means that a single new lesion is sufficient to classify the disease as progressive, even if all other lesions disappeared.

Limitations The above described MTV and TLG quantities are quantitative method analysis and have shown to be outcome predictive for a variety of cancer types. Nonetheless, they are not normalized for patient morphology, which may impede inter-patient comparison. By taking into account only up to 5 lesions, the PERCIST method is only semi-quantitative. Moreover, mixed responses with an important decrease in lesion uptake or in lesion volume but the appearance of even a single new lesion are classified as progressive disease, which is questionable in terms of clinical therapy monitoring.

2.3.3 Bone scintigraphy analysis

Bone Scan Index (BSI) The BSI [53] has been described as a quantitative method for bone tumour assessment in bone scintigraphy. It is an approximation of the percentage of the skeleton affected by lesions. More precisely, it is the average of the percentage lesion involvement of each bone weighted by the fractional mass of the bone in the skeleton. The bone to skeleton mass ratios are taken from [99]. Moreover, a semi-automatic computation method [104] has been developed and is commercialised in the EXINI Bone^{BSI} software (EXINI Diagnostics AB, Lund Sweden). The BSI has been shown to be predictive for survival in prostate cancer[93].

Limitations As a major limitation, the BSI is only a surrogate for the percental lesion involvement of the skeleton. This is due to the fact that the imaging technique is two-dimensional and that it is therefore not possible to correctly estimate the actual lesion mass. The use of standard weights for all patients is a further limitation that does not account for individual morphologies.

2.4 Conclusion

Prostate cancer is a very common type of cancer in men. The main tumour and secondary metastases can be imaged using different modalities, including PET, CT and bone scintigraphy. The available analysis methods for these modalities however present major drawbacks, and no method utilises the potential of hybrid PET/CT images. The clinical need for fully quantitative, multimodal, and, ideally, automatic analysis methods that allow for inter-patient comparison is therefore clear. As an automatic method needs to segment lesions and possibly other structures in the images, we give in the next chapter an introduction to the topic of medical image segmentation.

Chapter 3

Image Segmentation

In this chapter, we introduce technical concepts as background for the work presented in the following. First, we define the problem of image segmentation. We then introduce feature representation as a preprocessing step often used before addressing image segmentation directly. Finally, we describe methods for multi-class segmentation of medical images: because we used them in the rest of our work, we present random forests in details, but give only an overview of other methods. For explanations of mathematical notations, we refer the reader to Appendix A.

3.1 Introduction

In this section, we provide introductory information about the problem of image segmentation to uninformed readers to facilitate the comprehension of the rest of the thesis. We also introduce notations that are used throughout the thesis.

3.1.1 Problem posing

In our work, we consider digital images stored as collection of values on a regular grid. Each element of the grid is called a pixel for a two-dimensional image and a voxel for a three-dimensional image. More formally, we define a d -dimensional image I with n_c channels as a function from a d -dimensional grid $[[1, N_1]] \times \dots \times [[1, N_d]] \subset \mathbb{N}^d$ into \mathbb{R}^{n_c} that associates for each channel a value to each voxel \mathbf{v} in the grid:

$$\begin{aligned} I : [[1, N_1]] \times \dots \times [[1, N_d]] &\rightarrow \mathbb{R}^{n_c} \\ \mathbf{v} &\mapsto I(\mathbf{v}) \end{aligned} \tag{3.1}$$

(N_1, \dots, N_d) is the size of the image. A photograph taken with a common camera is a two-dimensional image ($d = 2$) and has $n_c = 3$ channels (red, green blue). In the following, images will typically be PET and CT images and have $d = 3$ dimensions and $n_c = 1$ channels. The image value at \mathbf{v} in channel c is denoted $I_c(\mathbf{v})$.

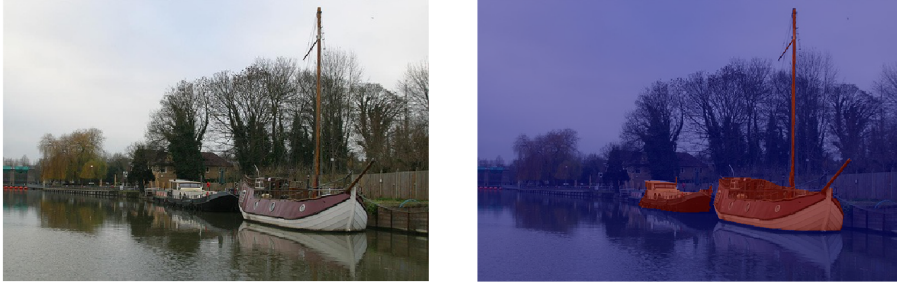


Figure 3.1 – Left: Image from the PASCAL VOC 2010 dataset [35]. Right: Overlaid true segmentation

We define a discrete segmentation S_I of image I (with finite number of labels) as an image of the same size whose values domain can without loss of generality be mapped to a subset $[[1, K]]$ of \mathbb{N} :

$$\begin{aligned} S_I : [[1, N_1]] \times \dots \times [[1, N_d]] &\rightarrow [[1, K]] \\ \mathbf{v} &\mapsto S_I(\mathbf{v}) \end{aligned} \quad (3.2)$$

where K is the maximum admissible label in I . This definition excludes the possibility of an infinite number of labels, as this case is not relevant for most medical imaging problems.

We assume that, for every image I , a *true* segmentation \hat{S}_I exists. The goal of image segmentation is to find \hat{S}_I . Note that for some problems, the definition of \hat{S}_I may be difficult. For medical problems in particular, \hat{S}_I might be unknown, as observing it would require an autopsy study. In that case, an expert's manual segmentation or a consensus between several experts is usually used in place of the *true* segmentation. An example of an image with its true segmentation from the PASCAL VOC 2010 dataset [35] is shown in Figure 3.1.

3.1.2 Supervised learning

Different learning frameworks exist depending on the nature of the training data (i.e. the data used to *train* a classifier). The unsupervised framework is defined by the absence of labelled training data. For example, clustering methods belong to this framework. The semi-supervised framework is defined by the presence of labelled and unlabelled training data. Typically, the data is very unbalanced, with only a small proportion of labelled data samples. This is in particular the case when a lot of data is available, but the ground truth acquisition process is very costly. The supervised learning is characterised by the availability of fully labelled training data.

In our work, we assume that labelled training data is available, i.e. for all training images, we know \hat{S}_I .

3.1.3 Performance evaluation

To evaluate the result of a segmentation process, we need to know the true segmentation and a measure of how *far* a segmentation S_I is from \hat{S}_I . We present here three performance measures used in our work. The three measures evaluate different properties of a segmentation. Which measure(s) should be used is problem-specific.

Two-label case ($K = 2$) Let us first consider the two-label case (i.e. $K = 2$).

Confusion matrix Let us consider label 1 as a positive label and label 2 as a negative label (for example in a tumour segmentation problem, v is part of a tumour or not). For a given segmentation S_I , we can then define the terms true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as follows:

$$\begin{array}{rcc} S_I(\mathbf{v}) = 2 & & S_I(\mathbf{v}) = 1 \\ \hat{S}_I(\mathbf{v}) = 2 & \text{True negative} & \text{False positive} \\ \hat{S}_I(\mathbf{v}) = 1 & \text{False negative} & \text{True positive} \end{array}$$

The confusion matrix is the table that contains the number of TP, TN, FP and FN.

True positive rate (TPR) This measure is sometimes called sensitivity or recall.

$$\text{TPR}(S_I, \hat{S}_I) = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \quad (3.3)$$

The TPR evaluates the proportion of voxels with true label 1 that have been identified. Note that a segmentation $S_I(\mathbf{v}) = 1, \forall \mathbf{v}$ would give a perfect TPR of 1, whilst being potentially very far from \hat{S}_I . The TPR is therefore often presented with the positive predictive value.

Positive predictive value (PPV) This measure is sometimes called precision.

$$\text{PPV}(S_I, \hat{S}_I) = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \quad (3.4)$$

The PPV evaluates the proportion of voxels with predicted label 1 that are correctly labelled.

Dice Score (DS) This measure is sometimes called F-score or F_1 -score. It is the harmonic mean of TPR and PPV

$$\text{DS}(S_I, \hat{S}_I) = \frac{\text{TPR} \times \text{PPV}}{\text{TPR} + \text{PPV}} = \frac{2 \sum \text{TP}}{2 \sum \text{TP} + \sum \text{FN} + \sum \text{FP}} \quad (3.5)$$

As the DS takes into account both the TPR and the PPV, it is often presented alone.

Multiple-label case ($K > 2$) Let us now consider the general case (i.e. $K > 2$). For the evaluation, we consider the multiple-label case as a collection of two-label problems: each label $k \in \llbracket 1, K \rrbracket$ is evaluated in a *one-against-all* fashion, where k is considered positive and all other labels are considered negative. Results can be presented individually for each label or averaged for groups of labels.

3.2 Feature representation

In this section, we present the concept of features representation and show feature examples.

3.2.1 Definition

Before computing the labelling, most methods perform feature extraction. Features hold information relevant to the problem at hand and can be extracted from the image itself or other sources (e.g., for medical problems, blood values, patient history, etc...). Information often relevant for image segmentation problems include: color, shape, texture, and location information. For some problems, feature properties such as translation or rotation invariance are desirable. Note that when using convolutional neural networks, features are directly learned by the network, and often, no handcrafted features are extracted.

In this work, we assume that features can be represented as images of the same size as the original image. To achieve this, the feature space has to be mapped to (a subset of) \mathbb{R} , which is only a mild restriction. For example, all categorical feature spaces with a finite number of categories can be mapped to an interval of the form $\llbracket 1, C \rrbracket \subset \mathbb{R}$, $C \in \mathbb{N}$. Each feature map \mathcal{F} can then be seen as a new channel of the original image.

3.2.2 Feature examples

In this section, we give examples of commonly used features.

Filter based features A filter Φ is a d -dimensional matrix of size $(\varphi_1, \dots, \varphi_d)$. When convolved with the image, it produces a feature map $\mathcal{F} = I * \phi$:

$$\mathcal{F}(\mathbf{v}) = \sum_{\mathbf{u} \in \llbracket 1, \varphi_1 \rrbracket \times \dots \times \llbracket 1, \varphi_d \rrbracket} I(\mathbf{v} - \mathbf{u})\Phi(\mathbf{u}) \quad (3.6)$$

Different ways of handling edges so that \mathcal{F} and I have the same size are possible: zeros padding, image wrapping, etc...

Depending on the content of Φ , different types of information can be extracted. Mean and Gaussian filters tend to decrease the noise level whilst blurring edges. On the contrary, Sobel filters enhance edges in a given direction. Examples with different filters are shown in Figure 3.2.



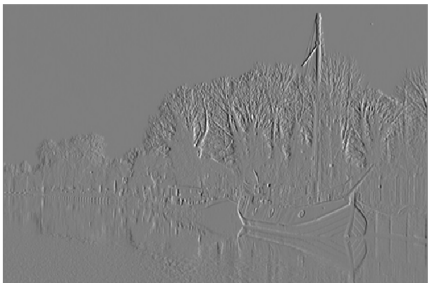
Filter	Parameters	Example
Mean	$\varphi = 3$ $\Phi(x, y) = 1/9$	
Gaussian	$\varphi = 16$ $\Phi(x, y) = \frac{1}{4\sqrt{2\pi}} \exp\left(-\frac{x^2 + y^2}{32}\right)$	
Sobel	$\begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix}$	

Figure 3.2 – Example of filters commonly used for feature extraction. All filters are square ($\varphi_1 = \varphi_2 = \varphi$) The last column shows the feature map obtained when applying the filter to the image in Figure 3.1.

Difference of Gaussians Differences of Gaussians enhance edges and are in particular used for detecting blob-like structures. They are computed as the difference of two Gaussian features (see Figure 3.2) with different standard deviations. The choice of standard deviations determines the scale at which the edges are enhanced. An example is shown in Figure 3.3.

Histogram of oriented gradients Histograms of gradients have been introduced in [28] for human detection and contain shape information. They are computed as follows: first the gradient is computed at each voxel of the image; then, the image is divided in cells, and a histogram

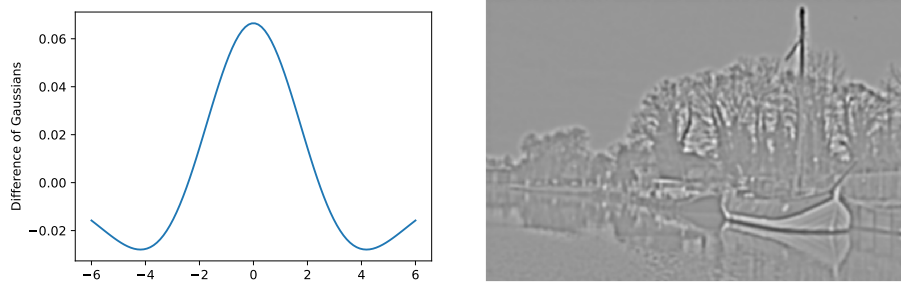


Figure 3.3 – Left: difference of two one dimensional Gaussians with mean 0 and standard deviations 2 and 3. Right: application of the difference of two-dimensional Gaussians with standard deviations 2 and 3 to the image in Figure 3.1.

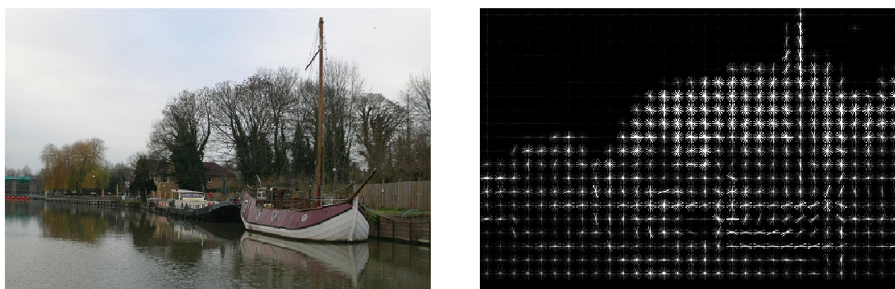


Figure 3.4 – Left: Image from the PASCAL VOC 2010 dataset [35]. Right: Representation of the histogram of gradient features

of the orientations of the gradients in the cell is computed for each cell. The gradient's magnitudes are also taken into account when building the histograms. The values of the bins of the histograms are used as features. An example is shown in Figure 3.4.

3.3 Random forests

Random forests are a classification method often used for image segmentation. They have first been introduced by Breiman in [16]. Because they are used extensively in the next chapters of this thesis, we give here a detailed presentation of random forests for image segmentation. Our description is based on the reference work by Criminisi and Shotton [25]. In the following, we assume that feature extraction has been performed, and that features are stored in image channels.

3.3.1 Decision Trees

A decision tree is a model for hierarchical decision making. The decision tree is the central element of the random forest. A tree is a connected and acyclic graph. A decision tree is a

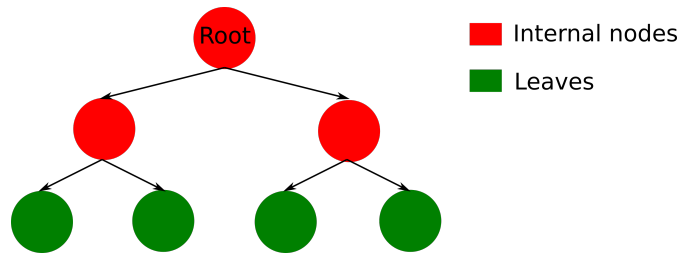


Figure 3.5 – Directed rooted tree: internal nodes have children while leaves do not.

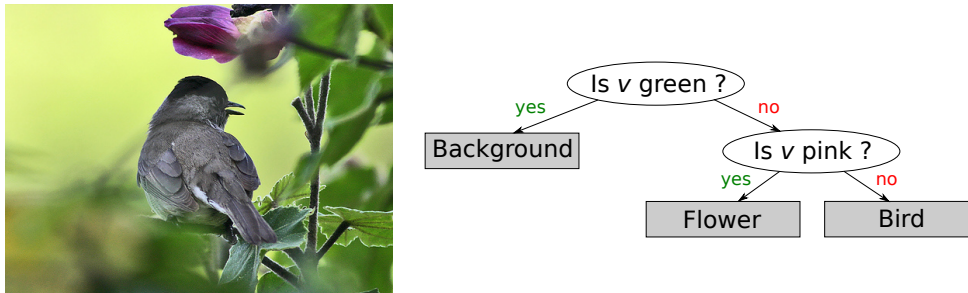


Figure 3.6 – Left: image from the ImageNet dataset [92]. Right: Example of decision tree for segmenting the image.

directed rooted tree. The root is the only node without incoming edge and all other nodes have exactly one incoming edge. Given a node n , the nearest node on the path from the root to n is the parent of n . All the other nodes connected to n are its children. Nodes without children are called leaves. Nodes with children are called internal nodes. Each internal node contains a split function, and each leaf a predictor. In all the following, we assume that decision trees are binary, i.e. all internal nodes have exactly two children. This does not restrict their discriminative power, as a node with more than two children can easily be transformed in an equivalent (deeper) binary tree. Figure 3.5 shows the different elements of a tree, and an example of binary decision tree is pictured in Figure 3.6.

Binary trees are first grown (or trained) using training data. When a new dataset has to be processed, this is done using the testing procedures. Both are detailed in the following paragraphs.

Training procedure The tree is grown in an offline recursive procedure using an optimisation function and a stopping criterion. Starting at the root, if the training data at the node doesn't fulfil the stopping criterion, the *best* split function for the training data at the node (according to the optimisation function) is set as the node split function. Training data is accordingly split and children are recursively grown. If a node fulfils the stopping criterion, it is classified as

a leaf and a predictor is chosen based on the training data that reached that leaf. Common stopping criteria are maximum node depth (i.e. only a fixed number of successive splits are allowed), minimum amount of data per internal node (i.e. datasets smaller than a fixed value are not split) and threshold on the optimisation function (i.e. if no good enough split can be achieved, the growth is stopped). The split function, optimisation function and predictor are discussed in the following paragraphs.

Testing procedure At test time, data samples travel from the root down to the leaves. When a sample enters an internal node, the binary split function determines to which child the sample is sent. This process is applied recursively until the sample reaches a leaf. The output for the sample is then obtained from the predictor contained in the leaf.

Split function The split functions contained in internal nodes determine the path followed by data samples in the tree. For image segmentation, these functions ξ are usually axis aligned weak learners, i.e. they consider only one channel and are of the form:

$$\begin{aligned} \xi : \mathbb{R}^{n_c} &\rightarrow \{0, 1\} \\ x &\rightarrow x_c < \delta \end{aligned} \quad (3.7)$$

with $c \in [1, n_c]$ and $\delta \in \mathbb{R}$.

In principle, any binary function can be used as the split function. Arbitrary hyperplanes in feature space [76], axis aligned hyperplanes in the canonical correlation projection space [86] and spheres [46] have shown good results for specific tasks. These are however less widely used than the axis aligned weak learners due to their intrinsic computational complexity and the more difficult optimisation at each internal node.

Optimisation function The optimisation function is an essential component of a decision tree, because it decides what is the best split function for each internal node. For segmentation tasks, the goal is to separate data of different classes and, ideally, obtain leaves that contain data of only one class. The optimisation function therefore measures how *pure* the datasets resulting from a split are. The most common optimisation functions are the information gain In_1 based on the entropy \mathcal{H} and the decrease in impurity In_2 based on the Gini purity measure \mathcal{G} . With E the training data at the given node, and E_1 and E_2 the two sets resulting from the split function that is evaluated, they are computed as:

$$\mathcal{H}(E) = - \sum_{l \in [1, K]} \frac{|S_l(\mathbf{v}) = k|}{|E|} \log\left(\frac{|S_l(\mathbf{v}) = k|}{|E|}\right) \quad (3.8)$$

$$In_1(E, E_1, E_2) = \mathcal{H}(E) - \frac{|E_1|}{|E|} \mathcal{H}(E_1) - \frac{|E_2|}{|E|} \mathcal{H}(E_2) \quad (3.9)$$

$$\mathcal{G}(E) = 1 - \sum_{l \in \{1, K\}} \frac{|S_l(\mathbf{v}) = k|^2}{|E|^2} \quad (3.10)$$

$$In_2(E, E_1, E_2) = \mathcal{G}(E) - \frac{|E_1|}{|E|} \mathcal{G}(E_1) - \frac{|E_2|}{|E|} \mathcal{G}(E_2) \quad (3.11)$$

where $|E|$ is the cardinality of E . Even though these are the most common optimisation functions, in principle, any function can be used.

Predictor For classification, the leaf predictors are usually taken as the class distribution of the training data that reached that leaf. In that case, the output of the tree is a probability map for each label. An alternative is to choose a single class by majority voting of the training data that reached the leaf. For pure leaves (i.e. that contain training data of only one class), this is the same. For mixed leaves however, the first option allows to encode uncertainty information.

Limitations The main limitation of decision trees is their pronounced tendency to overfitting, i.e to perform well on the training data but have poor generalisation power for new data. More precisely, for separable data¹ (i.e. there are no identical samples with different labels), a decision tree can always achieve perfect classification on the training data (provided that the stopping criterion does not stop the growth before it does). However, this does not guarantee perfect or even good performance on new data. In particular, if the training labels are noisy, or if the training data is not representative of the new data, performance on new data can be very poor. Overfitting is illustrated in Figure 3.7. Limiting the height of the tree alleviates this problem, but also limits its discriminative power. Another possibility is to inject randomness in the training procedure by bagging and/or randomised node optimisation.

3.3.2 Injecting randomness

By using randomness in the training procedure, a different tree may be obtained every time that a decision tree is trained. Randomness can be injected in the data or in the training procedure itself. More randomness produces more uncorrelated trees. The effect of randomness is shown in Figure 3.7.

Bagging This method is also called bootstrap aggregating. It consists of growing the tree with a bootstrap replica of the training data instead of the original training data. The replicas

1. Whether data is separable or not may depend on the features used to describe it.

are of fixed size, and are built by uniform sampling with replacement of the training data. This means that, when training several trees, each tree gets a different training data set.

Randomised node optimisation In traditional decision tree training, at each internal node, the weak learner that achieves the best split for the training data at that node is chosen amongst *all* weak learners. Instead, randomness can be injected by making available at each node only a fraction of all weak learners. The number of learners available is usually common to all nodes, and the learners are randomly chosen for each node. In the extreme case where only one learner is available at each node, the trees are called *extremely randomised* and are uncorrelated. When randomised node optimisation is used, two trees trained with the same training data can be different.

3.3.3 The ensemble model

Ensemble learning is inspired by Condorcet's jury theorem which states that for a binary classification problem, if each jury member has a probability $p > 0.5$ of giving the right classification, the probability of the majority of voters being right is greater than p and tends to 1 when the number of jury members tends to infinity (see [89] for a detailed description of ensemble learning).

In other words, combining the answers from different classifiers can, when certain assumptions are fulfilled, lead to an improved performance over each single classifier. In our case, a random forest is a combination of several decision trees trained using randomness and usually outperforms single trees. The set of trees is designated as \mathcal{T} . In particular, using several uncorrelated trees helps avoiding overfitting and improves performance when the training labels are noisy.

Although this is not the only possibility, we choose to define the combined predictor as an average of the tree individual predictors, these being the class distributions of the training data in the leaves:

$$\forall k \in \llbracket 1, K \rrbracket, P(S_I(\mathbf{v}) = k | I(\mathbf{v})) = \frac{\sum_{T \in \mathcal{T}} P_T(S_I(\mathbf{v}) = k | I(\mathbf{v}))}{|\mathcal{T}|} \quad (3.12)$$

where P_T are the tree-individual predictors. When needed, a discretised segmentation can be obtained as:

$$S_I(\mathbf{v}) = \operatorname{argmax}_{k \in \llbracket 0, K \rrbracket} P(S_I(\mathbf{v}) = k | I(\mathbf{v})) \quad (3.13)$$

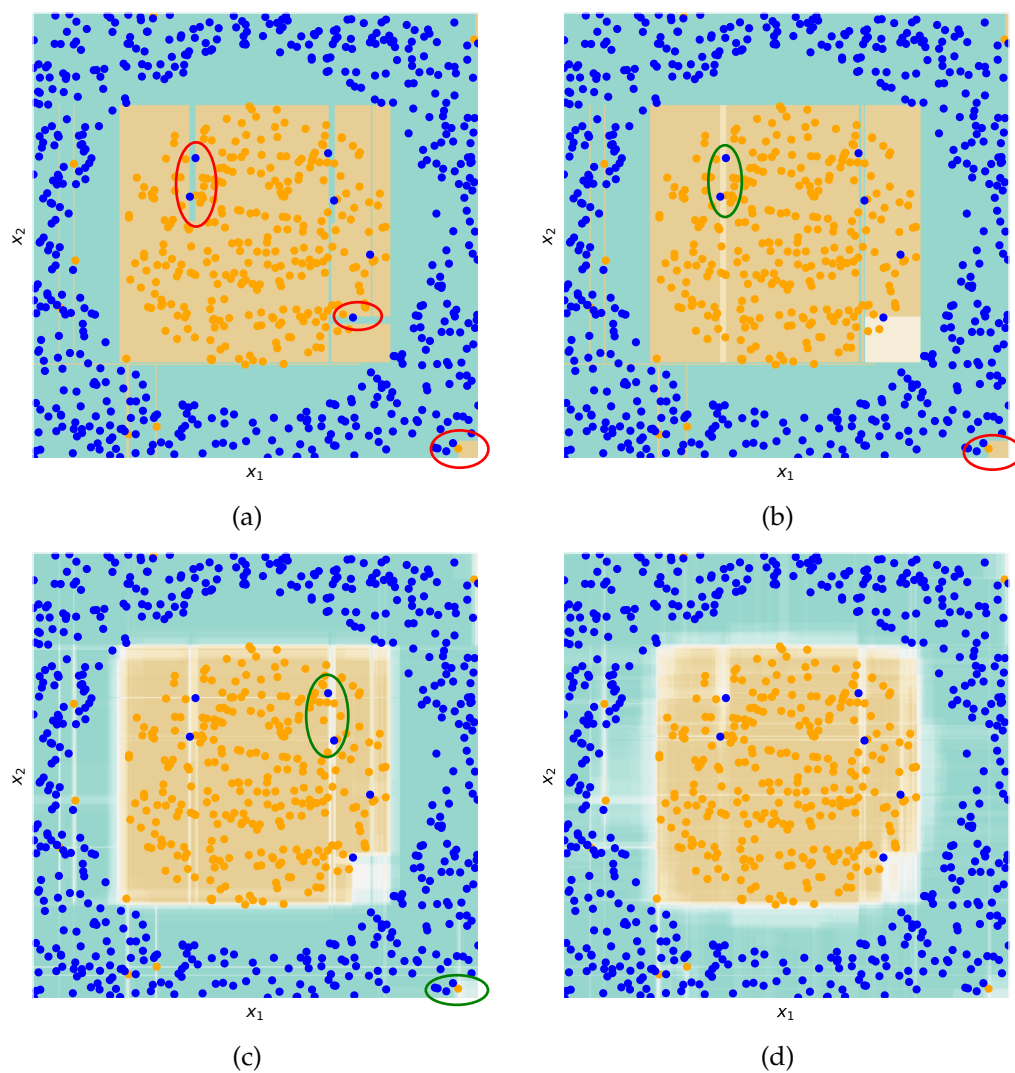


Figure 3.7 – Effect of injecting randomness in the training procedure. In this example, $n_c = 2$, $K = 2$. The dots are the training samples, their color reflect their label, and the background color shows the value obtained at testing. Overfitting examples are encircled in red, alleviated overfitting in green. (a) Single tree fully grown without randomisation. Overfitting is observed around mislabelled samples. (b) Single tree with height limited to 6, and grown without randomisation. Overfitting is still present, but less pronounced. (c) 50 trees grown using bagging. Smoother transitions are observed. (d) 50 extremely randomised trees trained with bagging. The resulting probability map is smoother, and the mislabeled training samples have only a minor effect on the result. The effect would be more pronounced with more features.

3.4 Multiclass medical image segmentation

We have provided in the previous section a detailed introduction to random forests for multi-class image segmentation. In this section, we briefly review existing methods commonly used for medical image segmentation. We focus on the supervised learning framework and in particular on variants of random forest based methods, atlas and model based methods, deep learning methods, and graph-based methods. Our review is not exhaustive.

3.4.1 Random forest based methods

In the previous section, we have described in detailed the original version of classification random forests. However, many variations have been used for segmenting natural [96] as well as medical images [42, 110]. Regression Forests, which are similar to Random Forests but predict a continuous output have been also used for sparse annotation of medical images [27, 24].

3.4.2 Atlas and model based methods

In atlas and model based methods, an atlas or model of the structure(s) to segment is obtained prior to segmentation. An image with the corresponding labelling is usually called an atlas, whilst a mesh representation of the surface of the object to segment is usually called a model.

At segmentation time, the atlas or model is geometrically deformed to match the image to segment. The quality of the match is measured by a similarity function. Depending on the class of deformations chosen for a particular task (translation, rigid, affine, free, etc...), numerous methods are available to compute the best deformation to match the image. Different similarity functions commonly result in different "best" deformations for the same image. It has also been shown that the choice of the atlas itself has an important influence on the accuracy of the results [4]. Thus, a common refinement of atlas-based methods consists in using several atlases. For example, the most similar atlas to a dataset can be used at test time, or all atlases can be used and the different segmentation obtained are subsequently fused [88]. A review of multi-atlas segmentation methods can be found in [52].

A major limitation of atlas and model based methods is that they are applicable only when data is relatively homogeneous: if the subject to segment is too different from the available atlas(es)/model(s), the method may fail. Moreover, these methods usually require the registration of each atlas or model to the image to segment, which is computationally very expensive.

Atlas-based methods have been used for example for segmenting brain structures [73] and abdominal organs [82].

3.4.3 Deep learning methods

Neural networks consist of neurons connected by weighted connections. Neurons are typically organised in layers and the weights are learned during the network training. Artificial neural networks trained by backpropagation have first been used by Lecun [65] for digit recognition, but the amount of hidden layers and the size of images to which they were applied were small due to the limited amount of computational power available.

Recently, this technique has however been used very successfully for image classification [62] and segmentation [72]. In this section, we give a basic description of convolutional neural networks (CNNs). For details about variants about these or different types of neural networks, we refer the reader to the considerable amount of literature on the topic.

A CNN typically consists of convolutional layers, pooling layers, non-linear layers, fully connected layers, and a loss layer. The output of internal layers consists of an image with several channels, called feature maps, by analogy with the feature extraction process. A convolutional layer consists of neurons that share weights so that the result can be expressed as a convolution of the image with a filter (similarly to feature computation in section 3.2.2). The filters typically have a small spatial extent, with a size of only a few voxels, but extend through all features maps present in the input of the layer. One feature map per filter is obtained. Pooling layers perform non-linear downsampling; for example, max-pooling layers perform a max operation with a given stride. Non-linear layers perform a non-linear operation at each element. The most common operations are the rectified linear unit operation ($ReLU(x) = \max(0, x)$), the saturating hyperbolic tangent and the sigmoid function. In fully connected layers, each element is connected to all elements of the previous layer. Finally, the loss layer computes the cost of the discrepancy between the network output and the true segmentation. The size and number of filters as well as the arrangement of layers are hyperparameters of the method.

CNNs are trained by backpropagation: filters are learnt by updating the weights of the connections at each iteration of the training following the gradient of the loss (i.e. the output of the loss layer) with respect to each weight. Data augmentation is usually used both to artificially increase the amount of training data and to “teach” the network wished invariances (translation, rotation, etc...).

Recently, Ronneberger [90] initiated the use of upsampling layers and skip connections to maintain spatial information in the network. His work was also one of the first successful

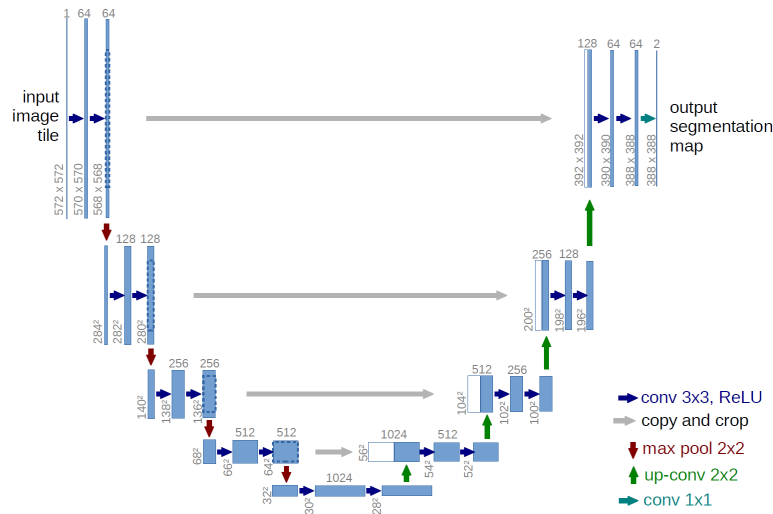


Figure 3.8 – Architecture of the U-net used in [90]. Copyright Springer, 2015.

applications of CNNs to biomedical image segmentation. A sketch of his network architecture is shown in Figure 3.8.

Deep neural networks have proven in the last year that they are very powerful for a variety of tasks. However, the training time, the quantity of training data needed, and their large number of hyperparameters are major limitations.

3.4.4 Conditional random field (CRF) methods

CRFs have been thoroughly described in [64]. In this overview, we will limit ourselves to CRFs with cliques of order up to two. A description of higher-order CRFs can be found in [56].

In CRF-based methods, the image is modelled as a graph where each voxel is a vertex and pairwise interactions are represented by edges. Feature values and labels are considered as random variables and Hammersley and Clifford [45] have shown that the joint distribution of S_I conditioned on I (with features as channels) is proportional to the exponential of the negative of an energy term E :

$$\log(P(S_I|I)) \propto -E \quad (3.14)$$

$$E = \sum_{\mathbf{v} \in I} U(\mathbf{v}, S_I(\mathbf{v})) + \lambda \sum_{\mathbf{v}_1 \sim \mathbf{v}_2} B(\mathbf{v}_1, \mathbf{v}_2, S(\mathbf{v}_1), S(\mathbf{v}_2))$$

U is commonly called the unary cost (depending on one voxel only), and B the binary cost. \sim denotes the presence of an edge between two voxels. The parameter λ balances the importance of both types of costs. Finding the maximum a posteriori estimate of the labelling is equivalent to finding the labelling minimising the energy term. Commonly, the natural grid

structure of the image is used to define edges (four or eight neighbours for each voxel in two dimensions, eight or twenty-six in three dimensions), yielding a sparse graph with few edges. Numerous methods exist to solve the maximisation. They have been reviewed and compared in [56]. If more than two labels are being segmented ($K > 2$), most methods are approximate and only reach a local maximum. Recently, a method has also been developed to approximately solve the minimisation problem when the graph is complete (i.e. all possible edges exist) under constraints on the costs [61].

CRFs are often used as post-processing for other methods. In that case, the unary costs are based on the output of the other method and the binary costs on voxel similarity. This can for example enforce spatial consistency after methods that handle each voxel independently.

One of the earliest uses of CRFs for image segmentation was for semantic segmentation of photographs [97].

3.5 Conclusion

We have presented in this section the problem of image segmentation and diverse methods commonly used to solve it for the particular case of medical imaging. In the following two chapters, we show for two practical cases how contextual information can be leveraged to improve the quality of the obtained segmentation.

Part II

Localisation and Quantification for Cancer Staging in PET/CT images

Chapter 4

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

In this chapter, we present work on skeleton segmentation in CT images. This work has been originally published in *Segmentation of skeleton and organs in whole-body CT images via iterative trilateration*, M. Bieth, L. Peter, S.G. Nekolla, M. Eiber, G. Langs, M. Schwaiger, B. Menze, IEEE Transactions on Medical Imaging, vol. 36, no. 11, pp. 2276-2286 [14]. © 2017 IEEE¹. This chapter is based on this publication with minor modifications.

4.1 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 [33, 109]. Recently, an effort has been made towards quantitative analysis, but methods remain semi-automatic [34] 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

1. In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of the university's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

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 [12].

Here, 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.

4.1.1 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 [95, 50, 57, 87, 40, 42], the knee [110], the ribs [112] and the hip region [31]. Methods also exist for multi-organ segmentation in CT [54, 111, 102, 80, 21].

Widely used methods for multiple structure segmentation include (but are not limited to) atlas based methods, deformable model based methods, random forest based methods, convolutional neural network based methods, and graph based methods. Some hybrid methods combine ideas from several approaches. We have given an overview of these method in section 3.4. Here, we will focus on how they can 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 [54], intrinsically uses context information by imposing constraints on the transformations. More context information can be incorporated by performing registration or atlas computation [111] at different scales. Context can also be explicitly used at label fusion time by modelling dependencies between voxels or labels [112], or taking a global decision, for example based on contours [108], 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 [57] or hierarchical [80, 21] relations between different objects.

Others have used learning algorithms such as random forests [16] that rely on decision trees and bagging. In these approaches, context information can be incorporated in the features such as with Haar-like [107] and geodesic features [59] or by using landmarks [110]. In combination with Haar-like features, a RF variant that learns the scale to which context is beneficial to the segmentation has been described [83] and a version where decisions are based on *all* features [47] 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 [60, 71] or predicted label and distance at the same time in a multi-task fashion [41, 37]. Cascaded systems are an alternative implicitly taking context into account either going from a global to a local scale [38] or using long range context information by providing the output of the forest (or an intermediate output [78]) to another forest for further training in an *auto-context* fashion, as initiated in [103]. The output of a classifier can also be processed before being used in the next one: in [87], the probability maps are regularised before being used as input for the next forest.

Deep convolutional neural network [65] 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 [8].

Finally, local spatial constraints are often considered through conditional random fields [64] and level set approaches, that try to minimise similar energies. For CRFs, with *classical* sparse binary terms, the α -expansion algorithm [15] is often used for energy minimisation. With dense graphs, a mean field approximation is preferred [61]. 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 [70] (and jointly optimised [94]), constraints with respect to relative positions can be enforced using the binary terms [58], and multiphase or temporal data can be leveraged by using four-dimensional graphs [70]. These methods can also be used in combination with superpixels/supervoxels [36, 100, 116].

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.

4.1.2 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 (subsection 4.2.4).
- We propose a cascaded random forest approach where landmarks are updated between each element of the cascade (subsection 4.2.5).
- 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 (section 4.3).

4.2 Methods

4.2.1 Overview

In our approach, classification was performed jointly with localisation by a cascaded random forest (illustrated in Figure 4.1, subsection 4.2.5) with anatomical trilateration features: from the probability map obtained from a random forest, we computed not only the voxel labels, but also the centroid of each structure to label. These centroids were 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 Figure 4.3, subsection 4.2.4).

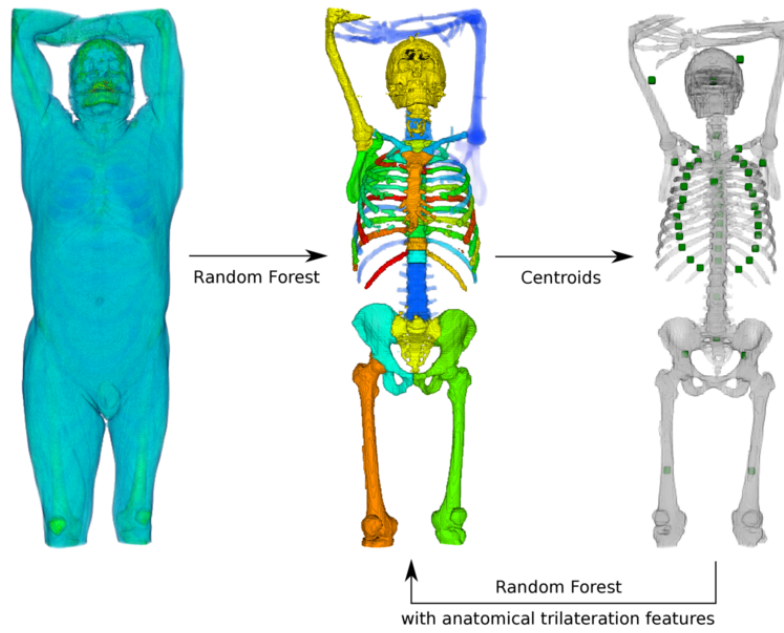


Figure 4.1 – Details of the cascaded RF (lines 6 and 10 in Algorithm 1). The landmarks obtained from the output of one iteration are used to compute the anatomical trilateration features in the next iteration.

The cascade described here constitutes a *super-classifier* that we used in a hierarchical coarse-to-fine fashion (subsection 4.2.6). The labels were ordered in a hierarchical manner which was described by a label tree (Figure 4.4). First, coarse groups of labels were segmented, and were thereafter refined by other super-classifiers in further classification procedures until all individual labels were segmented.

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

4.2.2 Notations

In all the following, K is the number of labels, k is a label, $\mathbf{v}(x, y, z)$ 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, S is a segmentation, and $a, b, c \in \mathbb{R}$. In a tree, for a node n , $Children(n)$ refers to the children of n .

4.2.3 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 used an implementation of scale adaptive random forest (saRF) [83] to perform a probabilistic segmentation of the structures to label. At training time, this particular version of RF samples

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...KLh do
5:       localCentroids = [ ]
6:       for iterations i=1...iLh do ▷ Cascaded trilateration
7:         mask = voxels to label
8:         probabilityMap = RF(V, ▷ Scale adaptive random forest
                               mask,
                               GkLh,
                               landmarks,
                               localCentroids)
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

```

the features sequentially in a fine-to-coarse fashion instead of sampling features uniformly for each node as done in the classical RF algorithm [16]. This can be seen as a guided sampling that learns the scale of the problem without user input.

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

$$S^0(\mathbf{v}) = \operatorname{argmax}_k(P^0(S(\mathbf{v}) = k|\mathbf{v})) \quad (4.1)$$

4.2.4 Features Description

Within the saRF, we used 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 were sampled in scale adaptive fashion whilst the trilateration features were sampled in the conventional uniform way. All features are described in details in the following paragraphs.

Haar-like features We used Haar-like features [107] as low-level context features. They are computed as an arithmetic operation between the intensity average of two boxes defined in the

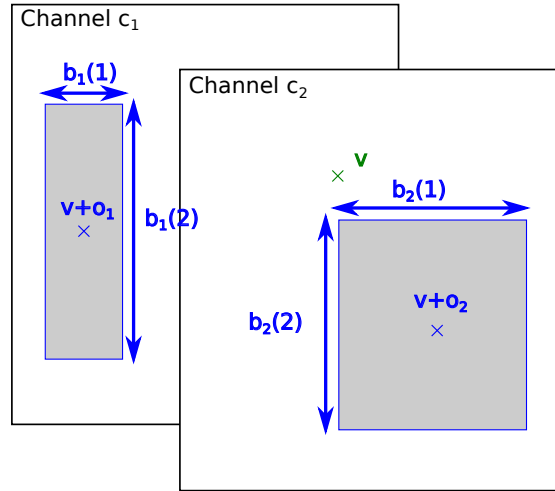


Figure 4.2 – Illustration of Haar-like features

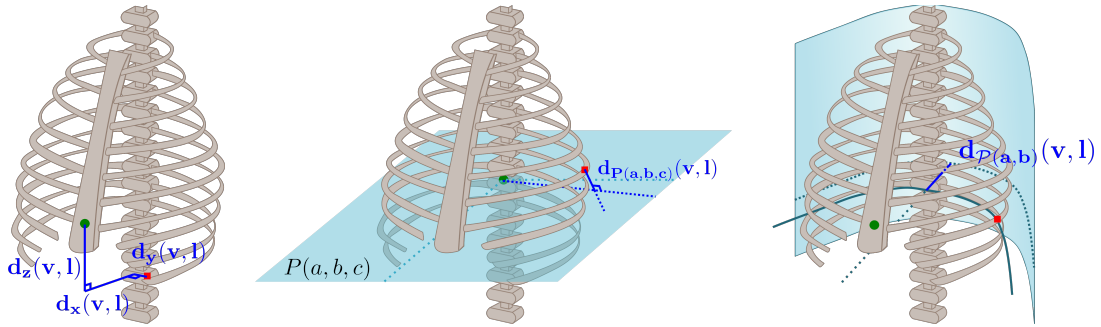


Figure 4.3 – Illustration of the anatomical trilateration features. In all images, the red square is the voxel to classify, and the green circle is the landmark. Left: signed distance features. Middle: planar distance feature. Right: parabolic distance features.

image domain, possibly in different channels. Each feature \mathcal{F} can be described by seven parameters $\mathcal{F} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{o}_1, \mathbf{o}_2, c_1, c_2, \sigma)$ 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} , c_1, c_2 are the channels in which the respective box value has to be computed and σ is an arithmetic operation ($\sigma \in \{ \text{sum, difference, sign of difference, absolute difference} \}$). An illustration of the features is shown in Figure 4.2. These features can be computed very efficiently by using integral volumes.

We also used symmetric Haar-like features that are similar to conventional Haar-like features, but for which the offset \mathbf{o}_2 is computed with respect to voxel \mathbf{v}_{sym} . This is an extension of the symmetric features used in [39] and is particularly suited to the problem at hand because the human skeleton presents an approximate symmetry with respect to the mid-sagittal plane.

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 subsection 4.2.5. Trilateration features are illustrated in Figure 4.3

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 were necessary to semantically trilaterate the position of a point.

The signed distances are defined as follows:

$$\begin{cases} x \text{ signed distance } d_x(\mathbf{v}, \mathbf{l}) = x - x_l \\ y \text{ signed distance } d_y(\mathbf{v}, \mathbf{l}) = y - y_l \\ z \text{ signed distance } d_z(\mathbf{v}, \mathbf{l}) = z - z_l \end{cases} \quad (4.2)$$

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:

$$\begin{cases} \text{planar distance } d_{P(a,b,c)}(\mathbf{v}, \mathbf{l}) = a(x - x_l) + b(y - y_l) + c(z - z_l) \\ \text{parabolic distance } d_{P(a,b)}(\mathbf{v}, \mathbf{l}) = (x - x_l) + a(y - y_l - b)^2 \end{cases} \quad (4.3)$$

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

4.2.5 Cascaded anatomical trilateration

We used the saRF and features described above in a cascade where the landmarks needed to compute the anatomical trilateration features were 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 could be computed. The centroids could then be used as a densely meshed ensemble of landmarks to iterate the classification with trilateration features. An updated probabilistic segmentation $P^1(k_v|\mathbf{v})$ was obtained.

This process was 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 output $P^{i+1}(k_v|\mathbf{v})$. The cascade is depicted in Figure 4.1. Because they were recomputed after each iteration, the centroids can be considered as self-updating landmarks. Through the anatomical trilateration features described in the previous section, centroids were used to represent context information in a more condensed way than in autocontext [103].

4.2.6 Hierarchical segmentation

Because our ground truth contained many labels (K between 51 and 88 depending on the field of view), we used 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 Figure 4.4.

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

A cascaded classifier (as described in subsection 4.2.5) was attached to each internal node $G_k^{L_h}$ of the tree. It received landmarks from all already segmented (super-)structures, ran for i^{L_h} iterations and classified voxels labelled with $G_k^{L_h}$ in level L_h into labels of $Children(G_k^{L_h})$. Figure 4.4 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.

4.2.7 Regularisation

Similarly to [101, 75], we used a CRF as final step to ensure spatial consistency of the segmentation. The nodes \mathcal{N} of the graphical model were the voxels to classify. A 26-neighbourhood structure was used for the binary connections. The energy of the CRF was defined as follows:

$$E = - \sum_{\mathbf{v} \in \mathcal{N}} \log(P(S(\mathbf{v})|\mathbf{v})) - \lambda \sum_{\mathbf{v}_1 \sim \mathbf{v}_2} B(S(\mathbf{v}_1), S(\mathbf{v}_2)) \quad (4.4)$$

where $P(S(\mathbf{v})|\mathbf{v})$ 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

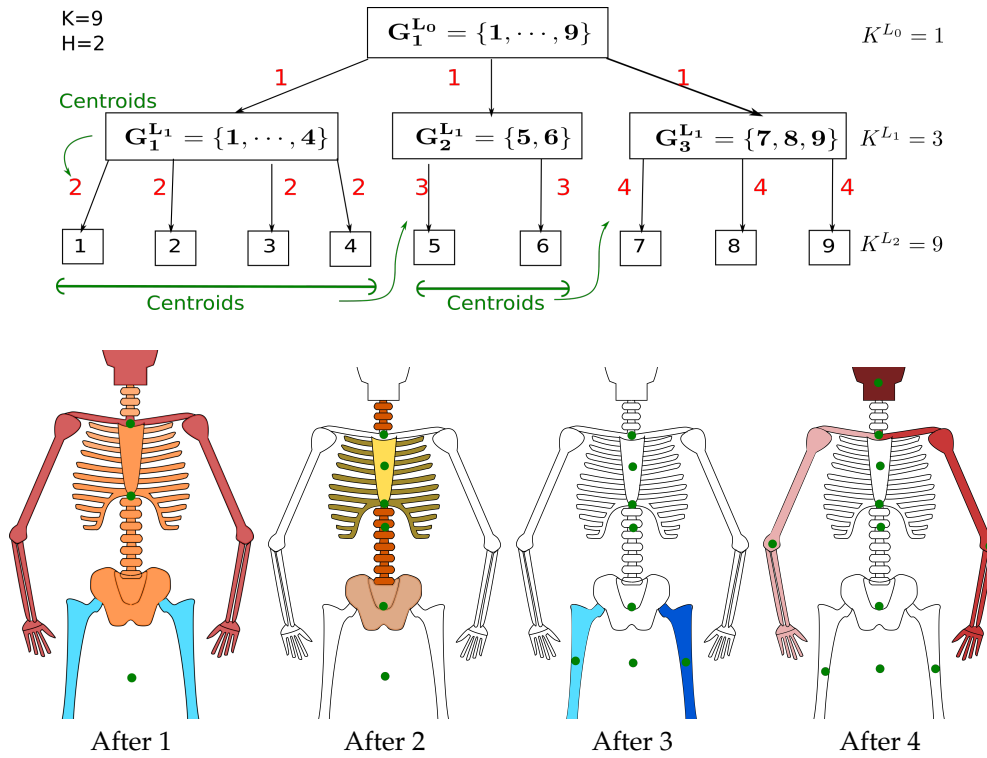


Figure 4.4 – Top: simplified hierarchical label tree example (9 labels and 2 levels of segmentation). The red digits indicate the ordering of the execution: all black arrows with the same number are processed simultaneously by the same cascaded RF. The green arrows indicate the subsequent flow of landmarks 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.

of assigning different labels to neighbouring voxels was higher if the association did not exist or was rare (e.g., skull and pelvis) than if the association was often found in the training data (e.g., vertebra L1 and vertebra L2). The binary costs were therefore *learnt*, so that our regularisation approach depended on only one hyperparameter $\lambda \in \mathbb{R}$ that set the balance between fidelity to the probability map and spatial consistency. We used the α -expansion algorithm [15] implemented in the OpenGM library [6] to find the label configuration minimising the energy of the graph.

4.3 Experiments

4.3.1 Data sets

We conducted experiments on three datasets:

- (1) Healthy subjects dataset (HS): twenty whole body CT scans of healthy subjects.

- (2) Prostate cancer dataset (PC): thirty thorax and trunk CT scans of prostate cancer patients.
- (3) Multiple myeloma dataset (MM): twenty thorax and trunk CT scans of multiple myeloma patients.

Healthy Subjects Dataset: twenty 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 $1.3 \text{ mm} \times 1.3 \text{ mm} \times 1 \text{ mm}$ were resampled to a mean resolution of $2.6 \text{ mm} \times 2.6 \text{ mm} \times 2 \text{ mm}$ 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: thirty 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: twenty non-contrast enhanced CT images of multiple myeloma patients, arms up, from the European VISCERAL project [44] 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 the PC dataset were manually annotated. Annotations for other structures (trachea, lungs, kidneys, psoas muscles, aorta, liver, spleen) were 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 Figure 4.5.

4.3.2 Setup

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 [12]: the intensity

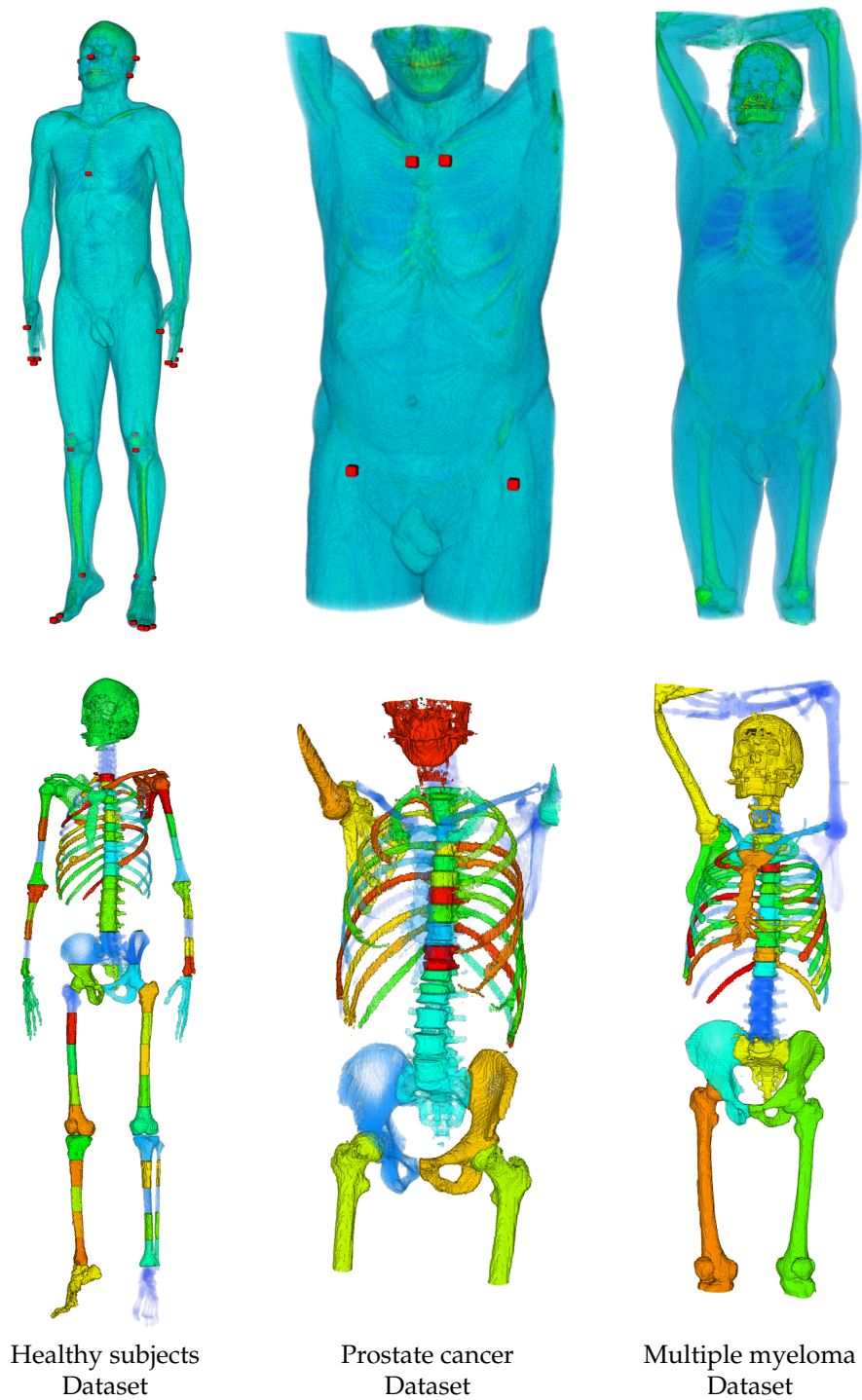


Figure 4.5 – Representative examples of the three datasets used in the experiments. The first line shows the original CT images with landmarks in red. The second line shows the ground truth bone annotation. Labels have been randomised for better visualisation.

of the different tissues was modelled by a Gaussian mixture, and a CRF was used to obtain the mask.

Parameters For all experiments, a custom-implementation of random forests was used. All forests were trained with 100 trees. For each node, 150 Haar-like features (as well as 50 distance and 50 geometric features when included) were sampled. For each sampled feature, 10 thresholds (chosen in a grid search fashion) were tested, and the Gini index was used as splitting criterion to determine the best feature and threshold for each node. Bagging was employed with a rate of 30% and a maximum number of 150 000 training samples per tree (effectively lowering the bagging rate) to limit the computational burden. For Haar-like features, the maximum offset was 100 voxels, and the maximum box size 50 voxels. 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 cascade (subsection 4.2.5) was run for 2 iterations in the first level and 5 iterations in the second. Except in the organ segmentation experiment, only bone-voxels were classified. In the organ segmentation experiment, 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. Except for the transfer experiment, all experiments were run with a 2-fold cross validation.

For our method, the hierarchical model was trained sequentially, each level in turn. For each group G , a cascade was trained. It consisted of two classifiers: one with only landmarks available before G was segmented (used for the first testing iteration) and one with the centroids of the elements of G as additional landmarks (used for all subsequent testing iterations). For all experiments, level 0 contained one label group with all K labels and level 2 K groups with 1 label. For skeleton segmentation, level 1 contained 8 groups: head, sternum, both arms, both legs, pelvis, rib cage (ribs and spine). For organ segmentation, level 1 contained 2 groups: background and organs.

4.3.3 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 (Figure 4.6). 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

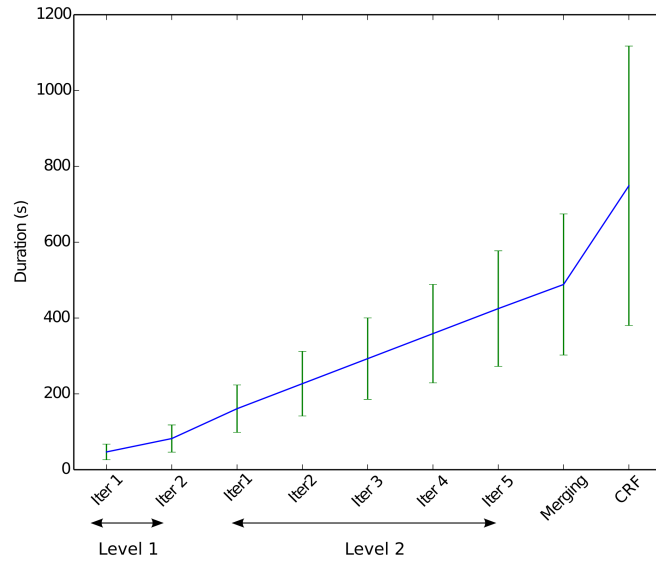


Figure 4.6 – Average computing time for the different steps of the algorithm measured using 15 subjects of the PC Dataset. The blue curve shows the average computing time and the green bars the range.

Table 4.1 – Mean DS over subjects for skeleton annotation for different methods

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

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.

4.3.4 Results

Comparison to other methods In a first experiment, we compared our method to two other methods: autocontext [103] and saRF [83]. 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 4.1 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.

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 Table 4.2. 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.

Table 4.2 – Mean DS over subjects for skeleton annotation for variations on our method

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

Results for different iterations and with regularisation are shown in Figure 4.7, Figure 4.8, Figure 4.9 and Table 4.3. 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 Figure 4.10 and Figure 4.11 show that most errors occurred either at the interfaces between two labels or in the ribs.

Table 4.3 – Mean DS over subjects for skeleton annotation for different iterations

	It.	1	2	3	4	5	5+CRF
Whole	HS	83.0	83.8	84.0	84.1	84.2	85.6
		± 6.6	± 6.5	± 6.5	± 6.5	± 6.5	± 6.1
	PC	79.4	81.2	81.4	81.6	81.6	83.8
		± 8.2	± 9.4	± 9.5	± 9.5	± 9.5	± 9.7
	MM	70.4	73.8	74.4	74.5	74.8	77.4
		± 10.4	± 11.3	± 11.4	± 11.4	± 11.5	± 12.2
Ribs	HS	79.9	81.2	81.5	81.7	81.8	85.6
		± 12.1	± 11.8	± 11.7	± 11.6	± 11.6	± 11.1
	PC	76.5	78.4	78.7	78.7	78.8	83.2
		± 10.2	± 12.3	± 12.5	± 12.6	± 12.5	± 13.8
	MM	63.8	69.9	70.7	71.0	71.2	76.2
		± 14.0	± 15.3	± 15.5	± 15.6	± 15.6	± 17.4

Feature importance To complement the analysis of the performance without the trilateration features in the previous section, we show here the importance of the different type of features. The importance was defined by Breiman *et al.* [17] as the average over all internal nodes of the forest that use the feature of the Gini information gain (computed on the training data) weighted by the probability of reaching that node. We show the importance of features for the prostate cancer dataset in table Table 4.4. The importance was averaged over all groups of a level in the hierarchical structure. The importance was similar for the two other datasets.

In the first iteration of the first level, because no landmarks were available, the trilateration features could not be computed, and only the Haar-like features were used. In the other forests, the trilateration features represented between 55% and 68% of the importance. Note that, as expected, they were used more in the second iteration of each level, for which *local* landmarks were available.

Table 4.4 – Normalized feature importance for the prostate cancer dataset

Level	Iteration	Haar	Distance	Geometric
1	1	1.00	0.00	0.00
1	2	0.32	0.45	0.22
2	1	0.45	0.37	0.18
2	2	0.34	0.44	0.21

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

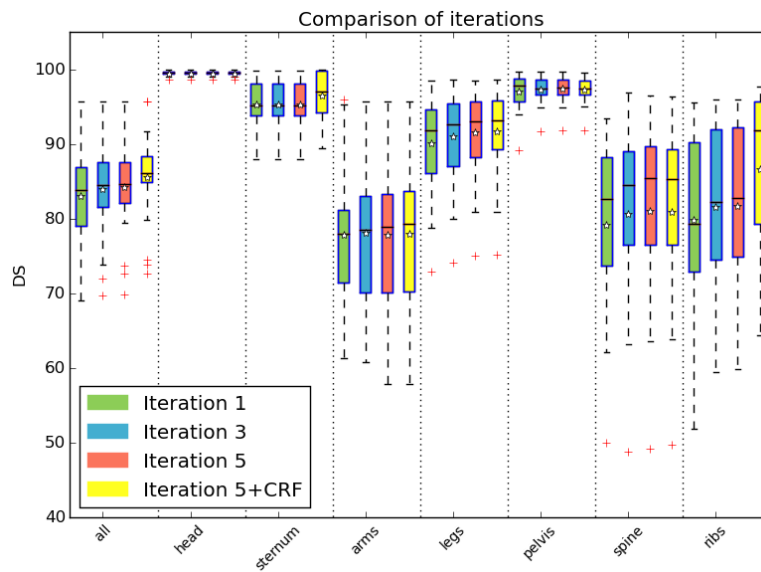


Figure 4.7 – DS for different groups of parts of the skeleton in the HS dataset, without initial landmarks.

the prostate cancer 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 Table 4.5. 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.

Table 4.5 – Mean DS over subjects for skeleton annotation with initial landmarks

	It.	1	2	3	4	5	5+CRF
Whole	HS	84.8	85.4	85.6	85.7	85.8	87.2
		± 5.7	± 5.6	± 5.6	± 5.7	± 5.7	± 5.6
	PC	80.2	81.7	81.8	81.8	81.8	84.2
		± 8.0	± 8.7	± 8.8	± 8.8	± 8.8	± 9.0
Ribs	HS	80.0	81.2	81.5	81.6	81.7	86.5
		± 12.1	± 11.4	± 11.2	± 11.2	± 11.2	± 11.2
	PC	77.2	79.2	79.4	79.4	79.4	84.5
		± 10.8	± 11.6	± 11.8	± 11.8	± 11.8	± 12.8

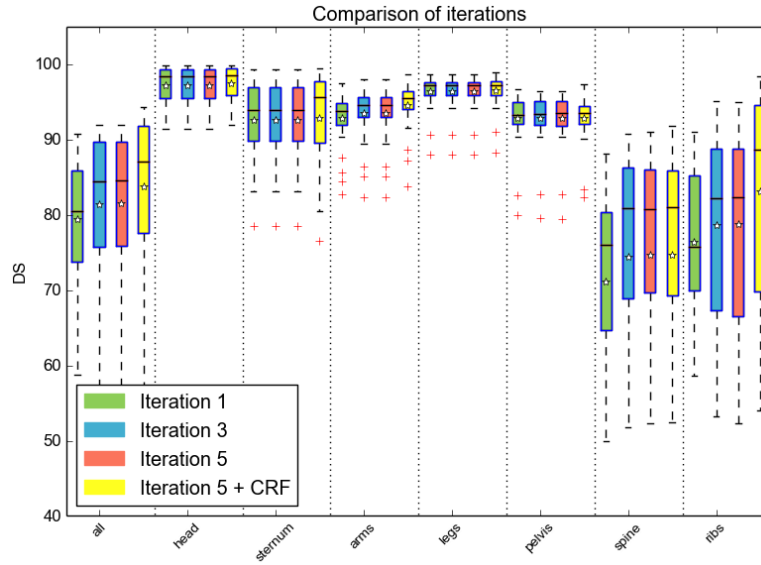


Figure 4.8 – DS for different groups of parts of the skeleton in the PC dataset, without initial landmarks.

Table 4.6 – 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)

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 the cascade. To obtain a fair comparison between methods, no regularisation was used. Results are presented in Table 4.6. 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.

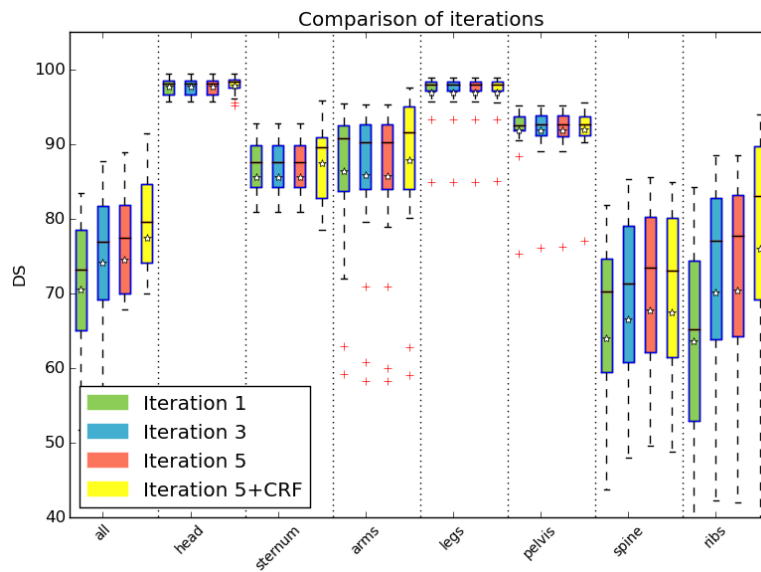


Figure 4.9 – DS for different groups of parts of the skeleton in the MM dataset, without initial landmarks.

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.

We used the multiple myeloma dataset, because annotations were available for various non-bony structures from the VISCERAL benchmark [43]. In this experiment, the first 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 Figure 4.12. Detailed results for individual organs are shown in Table 4.7. For the lungs and the liver, our results were close to the best ones obtained in the Visceral Benchmark 2015 [43]. 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 [54] is a multi-atlas method that has to perform a registration to each atlas for each structure to segment. From the information given in [54], 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

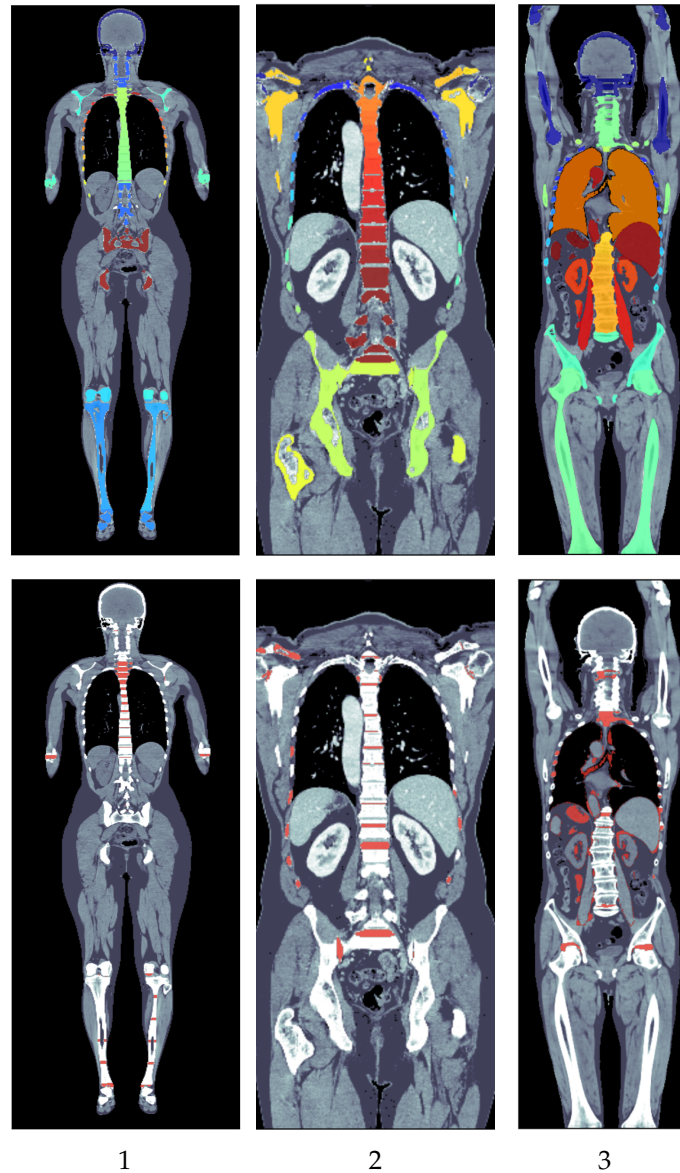


Figure 4.10 – First row: segmentation resulting from our method overlaid on the windowed CT image for an example of each of the three datasets. Second row: errors in the segmentations are shown in red. Most errors occurred at the interfaces between two labels or in the ribs for the bones, and in close zones of similar intensities for the organs (e.g. the trachea in column 3). For column 3, results of bones and organs segmentation are presented on the same image but have been computed separately.

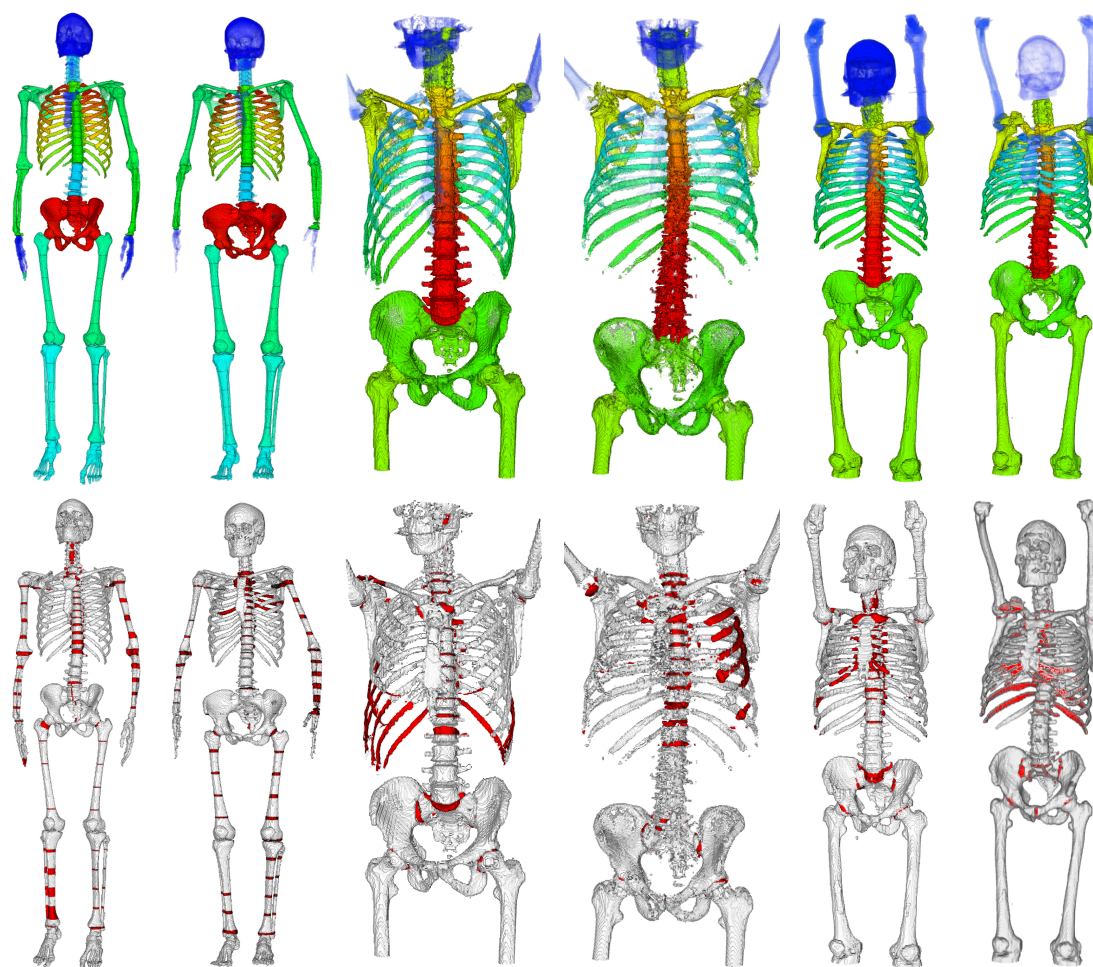


Figure 4.11 – First row: segmentation resulting from our method in 3D view for two examples of each of the three datasets. Second row: errors in the segmentations are shown in red.

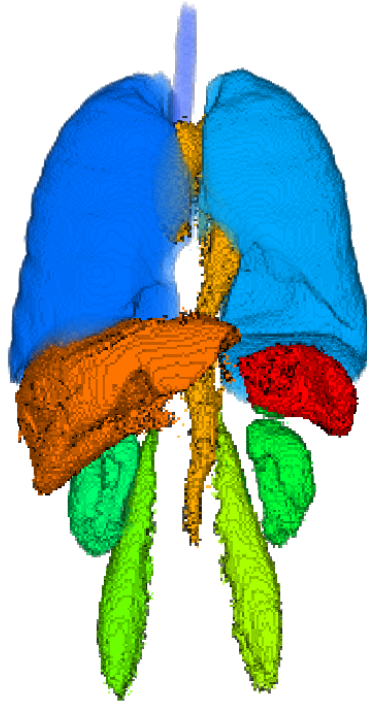


Figure 4.12 – Organ segmentation for one subject from the MM Dataset using our method, localising structures annotated in the VISCERAL Benchmark.

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 Figure 4.10 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.

4.4 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. Even in the ribs, which are the most challenging part to annotate due to similar appearance 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.

Table 4.7 – Mean DS over subjects for organ annotation

	Our method	Visceral [43]
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

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, 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. Our method can therefore be combined with an automatic landmark annotation method (e.g., the method described in [30]) 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 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. 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.

4.5 Conclusion and outlook

We have developed a method for skeleton and organs annotation in CT images that outperforms saRF and Autocontext 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.* [69] and incorporating shape models after each iteration as in [87] or as postprocessing of the final result are promising research directions.

Chapter 5

From Large to Small Organ Segmentation in CT using Regional Context

In this chapter, we present a method for segmenting small organs in CT image. This work has been originally published in *From large to small organ segmentation in CT using regional context*, M.Bieth, E. Alberts, M. Schwaiger, B. Menze in the proceedings of the international workshop on Machine Learning in Medical Imaging (MLMI), 2017, Springer [11]. This chapter is based on that publication with minor modifications.

5.1 Introduction

Precise organs segmentation is necessary in diverse medical applications, including diagnostics, computer-aided interventions and radiotherapy planning. Thus, the problem is well-studied and multiple methods exist that produce good segmentations for larger organs such as the lungs and the liver. For these organs, state of the art methods reach Dice Scores of over 0.9. However, for smaller organs with a higher variability in location or shape (e.g. pancreas, glands), most existing segmentation methods do not yield good results. Locating these organs is nonetheless crucial for example for radiotherapy planning to avoid irradiation with dramatic consequences. It would also be useful to automatically locate normal glandular uptake when segmenting lesions in PET images or lesions in these organs. In this chapter, we are therefore interested in fully automatic small organ segmentation. To be usable in clinical practice, the method should have a short computation time and a good sensitivity.

The problem of multi-organs segmentation has been addressed with different approaches. Multi-atlas registration methods followed by label fusion such as [29] generally produce better results than patch or voxelwise labelling methods. However, they usually have a higher computation time because each atlas has to be non-linearly registered to the test image. Deep learning has also been successfully used for segmenting larger organs [49] such as kidneys and liver as well as the pancreas [91], but requires large amounts of training data, which is often

difficult to obtain. Lately, more time efficient approaches such as Regression Forests [26] for organ localization, Atlas Forests [115] and Vantage Point Forests [46] for organ segmentation have shown good results. In forest-based methods, prior knowledge can be built in the features and high performance can be achieved with smaller training sets and relatively short computation times. Taking into account local as well as long-range context in the features has been shown to improve the performance. Haar-like [107] and BRIEF features [19] describe the local context based on intensities. Longer-range context can be provided for example by the output of a previous classifier, in an autocontext fashion [87, 103], by the use of distances to landmarks [12] or by the use of shape features [68]. The importance of semantic context has also been explored in [80].

Whilst local context is enough to segment larger organs such as the lungs, liver or kidneys, it doesn't allow for segmentation of smaller structures of interest for radiotherapy planning that are sometimes only scarcely visible in CT. In this work, we introduce a novel approach for small organ segmentation that makes use not only of local but also of regional context through features that encode semantic knowledge on nearby anatomy similarly to [12] and organ shapes as in [68]. Moreover, our method does not require *any* deformable registration to be performed (in contrast to [46] and [115]). By using fast Vantage Point Forests [46] for inference, it has a computation time that is significantly lower than multi-atlas methods and scales well to large data sets. We implement it in an iterative procedure where small organ labellings are iteratively refined by gradually incorporating better context information in the classification process. A final shape voting step ensures spatial consistency. In the following, we present our approach (section 5.2), evaluate it on the Visceral Challenge 2015 dataset [43] (section 5.3) and offer conclusions (section 5.4).

5.2 Methods

In this section, we detail the different components of our method for small organs segmentation. We encode regional context in the form of anatomical context and shape features. These are used within an iterative procedure where, after an initial labelling of all organs using local context only, the segmentation of small organs is refined using regional context. Finally, the segmentations are regularised by shape voting.

In the following, we describe our base classifier, the Vantage Point Forest (VPF), how the initial labelling is performed, and how we further use it to refine the labellings by employing context-richer descriptors in subsequent iterations. We then describe how to regularise the

labelling using shape voting. In all the following, I is the current image and $\mathbf{v} = (x, y, z)$ is the current voxel. Note that no pre-alignment of the images is performed, and all the voxels in the image volume enter the first classifier.

5.2.1 Vantage Point Forest

In our work, we chose to use a clustering approach that is able to consider all features of a sample simultaneously and is therefore less prone to overfitting than a classical Random Forest [16] that considers only one feature at each node. As a base-classifier, we used the VPF. It is an algorithm for approximate nearest neighbour search whose atomic element, the Vantage Point Tree, was first described by Yianilos [114]. Instead of using axis-aligned splits, each tree of the VPF describes a partition of the data space using hyperspheres centered on training data samples. These center-samples are randomly selected during training, and each tree is grown up to a fixed leaf-size. After training, each internal node therefore contains a center-sample and a radius describing a hypersphere and each leaf contains a set of training samples. At test time, each sample is pushed through the trees by determining at each node whether it is located inside or outside the hypersphere and recursively searching either the left or the right subtree until a leaf node is reached. The training samples contained in that leaf node are approximate nearest neighbours of the test sample. Heinrich *et al.* [46] showed that using a linear nearest neighbour search over the union of the sets returned by all the trees improves the segmentation results for large organs. We therefore followed this approach in our work. The distribution of classes of the nearest neighbours was then used as the output of the classifier.

Note that, similarly to extremely randomized trees, in VPF, node splitting is not optimized. Moreover, training labels are only used at test time to calculate the class distribution of the nearest neighbours set. When using binary features, the training and searching of VPF is very efficient, even compared to classical Random Forests.

5.2.2 Initial labelling

We defined the initial labelling as a multi-class segmentation problem. We used a VPF with BRIEF features as weak descriptors to obtain a tentative labelling of large and small organs. A performance close to the state of the art was reached for the lungs, liver, spleen and kidneys. For small and more variable structures however, incorrect segmentations were obtained, because BRIEF features were not able to describe small organs precisely enough.

BRIEF features BRIEF features [19] encode local intensity differences and are computed on the smoothed image \tilde{I} . For \mathbf{v} , the n th BRIEF feature $\mathcal{F}_n^{\text{BRIEF}}$ was :

$$\mathcal{F}_n^{\text{BRIEF}}(\mathbf{v}) = \text{sign}(\tilde{I}(\mathbf{v} + \mathbf{o}_{n,1}) - \tilde{I}(\mathbf{v} + \mathbf{o}_{n,2})) \quad (5.1)$$

where $\mathbf{o}_{n,1}$ and $\mathbf{o}_{n,2}$ are randomly chosen offsets. By imposing, for a fixed proportion of the features, $\mathbf{o}_{n,2} = 0$, we could ensure that not only the relations between neighbouring structures are described, but also their relation to the current voxel.

5.2.3 Iterated Forest with regional context descriptors

Even though the initial dense segmentation of small organs using VPF with weak descriptors was sub-optimal, the probability maps obtained contained valuable information. In particular, we computed an approximate location $\mathbf{v}_o = (x_o, y_o, z_o)$ of each object o to segment using its probability map \mathcal{P}_o computed from the initial classifier as an average location weighted by \mathcal{P}_o :

$$\mathbf{v}_o = (x_o, y_o, z_o) = \sum_{\mathbf{v} \in I} (x, y, z) \times \mathcal{P}_o(\mathbf{v}) / \sum_{\mathbf{v} \in I} \mathcal{P}_o(\mathbf{v}) \quad (5.2)$$

We then defined the small organ segmentation problem as a two class problem and restricted it to a box around \mathbf{v}_o . The size of that box was computed as the maximum size of the object to segment amongst the training subject plus a security margin.

We propose anatomical context features and shape features to describe the regional context. These descriptors provide richer information than BRIEF features whilst still benefiting from a low computation time. We iteratively performed new labellings as a two-class problem for each object of interest separately, using BRIEF features, anatomical context features and shape features. The latter are described in the next paragraphs and all features are illustrated in Figure 5.1. \mathbf{v}_o was recomputed after each iteration.

Anatomical context features We introduce anatomical context features to describe spatial relationships between neighbouring structures. They are related to the *autocontext* method of Tu [103]. Tu used the probability maps obtained from a classifier as inputs for the next classifier in an iterative process. Here, we considered the labellings of structures segmented in the previous iterations. For \mathbf{v} , a segmentation S_I obtained from the previous classifier and a reference label k , the n th anatomical context feature $\mathcal{F}_n^{\text{anat}}$ was:

$$\mathcal{F}_n^{\text{anat}}(\mathbf{v}) = (S_I(\mathbf{v} + \mathbf{o}_n) == k) \quad (5.3)$$

where \mathbf{o}_n was a randomly chosen offset.

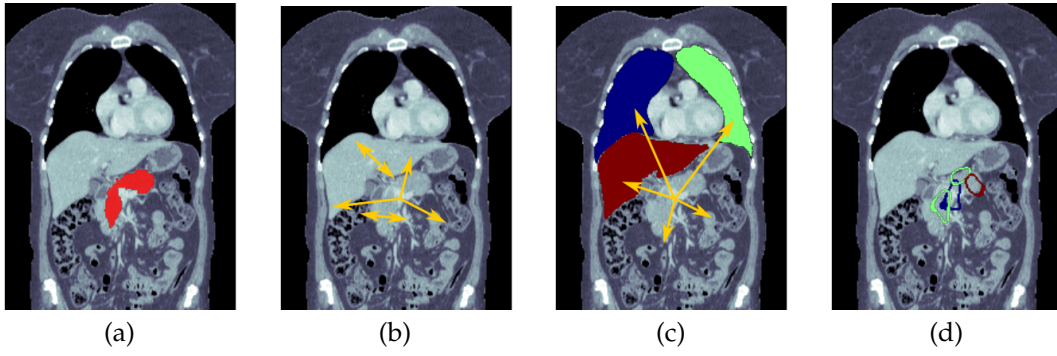


Figure 5.1 – Illustration of different types of features: (a) Ground truth for pancreas segmentation (b) BRIEF features: sign of intensity differences between or with nearby voxels are computed (c) Anatomical context features: the initial labelling at nearby voxels is considered (d) Shape features: training shapes at the approximate location of the object of interest are considered. Here, only contours of the shapes are shown.

Shape features We introduce shape features as an emulation of regional shape atlases. For each training subject s , a window around the object of interest with label k was selected, yielding a cropped image I_c^s and the corresponding segmentation S_c^s . I_c^s was then aligned to I by translation $T_{\rightarrow(x_o, y_o, z_o)}$ and affine transformation $\text{Aff}_{I_c^s \rightarrow I}$. The transformation was then applied to S_c^s to obtain the shape features $\mathcal{F}_s^{\text{trans}}$ and $\mathcal{F}_s^{\text{aff}}$:

$$\begin{cases} \mathcal{F}_s^{\text{trans}}(\mathbf{v}) &= (T_{\rightarrow(x_o, y_o, z_o)}(S_c^s)(\mathbf{v}) == k) \\ \mathcal{F}_s^{\text{aff}}(\mathbf{v}) &= (\text{Aff}_{I_c^s \rightarrow I}(S_c^s)(\mathbf{v}) == k) \end{cases} \quad (5.4)$$

At a local level, affine transformations are complex enough to meaningfully express deformations whilst retaining a lower computational time than deformable registration methods. Because we registered a cropped region and not the full image, we chose to restrict the shape features to translations and affine transformations. Our shape features are similar to local multi-atlas features. However, by operating at a local scale and with affine transformations only, they are computationally more efficient.

5.2.4 Final shape voting (SV)

The procedure described above output a probability map for each of the organs to segment. It could be discretised by choosing for each voxel the label with the highest probability. However, despite the shape features used in the classification, this approach doesn't ensure an optimal spatial consistency of the resulting segmentation. We propose instead to allow each training structure to vote on the probability map. Each training structure was affinely registered to the corresponding probability map. In each voxel, the number of votes was counted, leading to our final labelling.

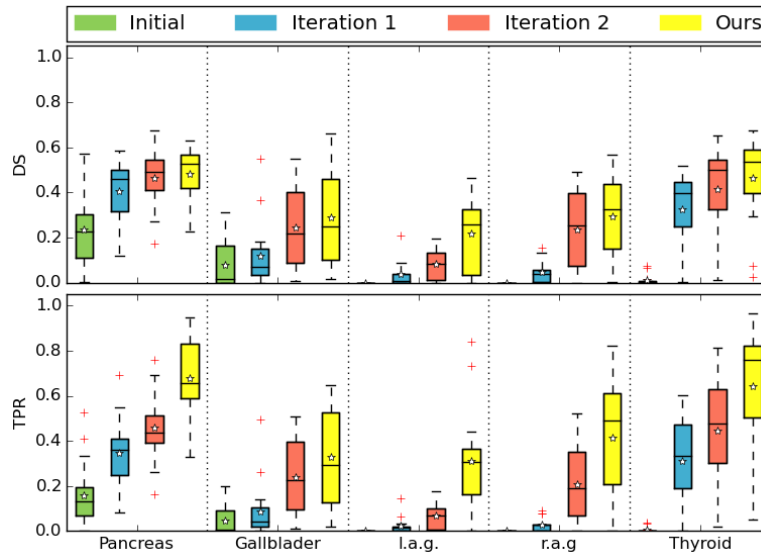


Figure 5.2 – Evaluation of DS and TPR after the initial labelling, the first iteration, the second iteration and the shape voting.

5.3 Experiments

We performed segmentation of the pancreas, the gallbladder, the left and right adrenal glands and the thyroid gland in CT images.

Datasets We evaluated our method on the twenty training images available from the Visceral Anatomy 3 dataset [43] for non contrast-enhanced CT. In total, annotations for twenty structures were available, but we chose to concentrate on the small and variable ones, with which other approaches had had difficulties. All data was resampled to a 2mm isotropic resolution. When the ground truth was not available for a particular structure in a subject, the subject was removed of the score calculation for that structure only.

Parameters Evaluation was performed in a leave-one-out fashion. For each iteration of our method, 15 trees were grown. At test time, the linear nearest neighbour search following the tree search selected 21 nearest neighbors. For BRIEF features, the image was smoothed by a Gaussian with a standard deviation of 3 voxels and the offsets were sampled for one half from a Gaussian distribution with a standard deviation of 20 and for the other half from a Gaussian distribution with a standard deviation of 40. For a third of the BRIEF features, $\sigma_{n,2}$ was set to 0. For anatomical context features, the offsets were sampled from a Gaussian distribution with a standard deviation of 20. All image registrations were performed using SimpleITK [74], with mean square difference as metric and gradient descent optimisation.

For the initial labelling, 640 BRIEF features were used. For refining the labelling, we performed two iterations as described in subsection 5.2.3. For the first one, 160 BRIEF features were used. For the second one, the liver, lungs, spleen and kidneys were considered as reference organs for the anatomical context features. 128 BRIEF features, 64 anatomical features for each reference organ (384 in total) and 38 shape features (each repeated three times, 108 in total) were used.

We compared our method to VPF (15 trees, parameters as in [46]), VPF followed by shape voting (VPF+SV), scale-adaptive Random Forest (saRF) [83] (99 trees) and the best multi-atlas method [29] for small organs of the Visceral Anatomy Challenge at ISBI 2015.

Table 5.1 – Comparison of the average DS and TPR obtained with different methods. Note that the scores for [29] were obtained on the testing and not the training set of the Visceral dataset.

DS	Pancreas	Gallbladder	Thyroid	L. adrenal g.	R. adrenal g.
Ours	0.481	0.288	0.463	0.220	0.294
VPF	0.234	0.081	0.0	0.0	0.010
VPF+SV	0.447	0.243	0.0	0.0	0.228
saRF	0.246	0.012	0.294	0.08	0.018
Multi-atlas [29]	(0.408)	(0.276)	(0.549)	(0.373)	(0.355)
TPR	Pancreas	Gallbladder	Thyroid	L. adrenal g.	R. adrenal g.
Ours	0.681	0.328	0.643	0.311	0.417
VPF	0.158	0.047	0.0	0.0	0.005
VPF+SV	0.559	0.233	0.0	0.0	0.238
saRF	0.553	0.010	0.598	0.0	0.013

Results We evaluated our method using the True Positive Rate and the Dice Score. As both for radiotherapy planning and false positive removal in PET, sensitivity is more important than specificity, the TPR is our main guideline.

In Figure 5.2, we show the influence of each element of our method on its overall performance. It demonstrates that the anatomical context and shape features included in Iteration 2 significantly increased both DS and TPR. In particular, after iteration 1, the adrenal glands were still not found, but the inclusion of anatomical context and shape features allowed to locate them for many subjects. The final shape voting had a limited influence on the DS, but caused a substantial increase in TPR.

Table 5.1 compares the DS and TPR obtained with our method to those obtained by the best method of the Visceral challenge 2015, which is a multi-atlas method, and other forest-based methods. For the pancreas and the gallbladder, our method obtained a higher DS than

the multi-atlas method (0.073 and 0.012 more respectively). As the scores for the multi-atlas method were computed on the Visceral challenge testing set, they can not be directly compared to ours, but still give a good orientation as to the respective performance of both methods. Our method also reached a higher DS and TPR than the forest-based methods for all organs, even when using VPF in combination with shape voting.

Note that, for our method, no deformable registrations need to be performed, which makes it computationally efficient. Typically, feature computation for each iteration for one subject was around one minute (for all organs), and the search of each tree was a matter of seconds.

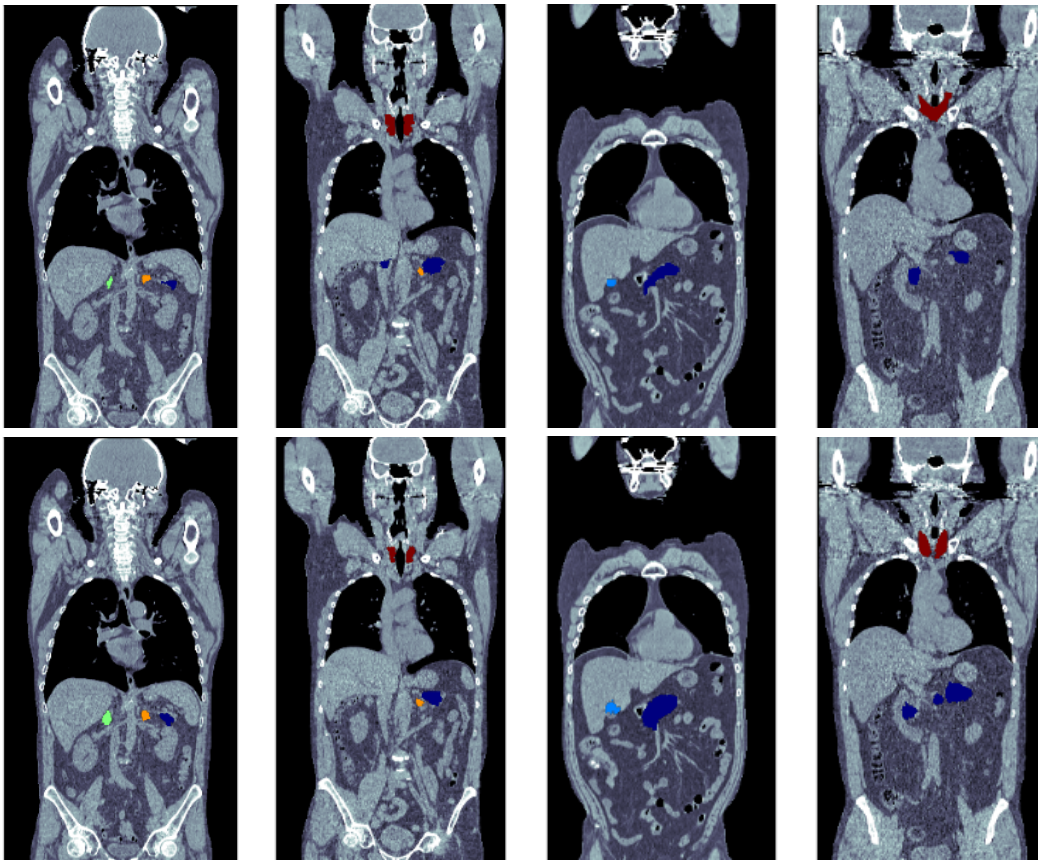


Figure 5.3 – Illustration of the labelling obtained for two subjects. The first line shows the ground truth and the second line the segmentation obtained with our method. The adrenal glands are shown in green and orange, the pancreas in blue, the gallbladder in light blue, and the thyroid in red.

5.4 Conclusion

We have presented a novel iterative method for small organ segmentation in CT and introduced the anatomical context and shape features for describing regional context. Through the

shape features and the final shape voting, our method has similarities with multi-atlas methods. By using only affine registrations however, it is computationally efficient and still outperformed the winning multi-atlas method for pancreas and gallbladder segmentation on a similar dataset. It also outperformed other forest-based methods for all small organs.

Chapter 6

Bone PET Indices: Multimodal Quantitative Indices for Bone Lesion Burden Staging in PET/CT

In this chapter, we present our work on quantitative assessment of bone metastases in PSMA-PET/CT images. This work has been originally published in: M. Bieth, M. Krönke, R. Tauber, M. Dahlbender, M. Retz, S.G. Nekolla, B. Menze, T. Maurer, M. Eiber, M. Schwaiger. *Exploring New Multimodal Quantitative Imaging Indices for the Assessment of Osseous Tumour Burden in Prostate Cancer using ^{68}Ga -PSMA-PET/CT*. J Nucl Med. 2017, vol. 58, no. 10, pp. 1632-1637 [13]. © by the Society of Nuclear Medicine and Molecular Imaging, Inc. This chapter is based on that publication with modifications.

6.1 Motivation

In case of recurrence, prostate cancer often spreads to other structures, in particular the bones. An accurate staging of the disease and its response to therapy are of utter importance for deciding to continue, change or abandon treatment. In clinical routine, patient performance status, blood parameters and imaging are common elements used to assess the risk, life expectancy and treatment response of individual patients.

In clinical practice, bone scintigraphy and ^{68}Ga -PSMA-PET/CT are the most widely used methods for prostate cancer staging and re-staging. However, bone scintigraphy is a two-dimensional modality which lacks detailed anatomical information, has suboptimal specificity and does not show lymph node and visceral metastases. It is therefore increasingly replaced by ^{68}Ga -PSMA-PET/CT, that has been shown to offer high detection rates and superb specificity for prostate cancer staging [2, 3]. Additionally, all existing CT and PET/CT analysis methods that we described in section 2.3 present drawbacks for the assessment of prostate cancer therapy response. In particular, RECIST considers osteoblastic bone metastases as non measurable and both RECIST and PERCIST are not fully quantitative. Moreover, a first tentative at quantitative assessment of ^{18}F -Fluoride-PET/CT in prostate cancer has been done by Etchebere *et al.*

[34] but the approach is unimodal and neglects the information contained in the CT image, impeding inter-patient comparison. Therefore, a comprehensive quantitative imaging biomarker measuring complete tumour load allowing for inter-patient comparison is an unmet clinical need.

In this work, our aim was to define intrinsically multimodal quantitative imaging indices incorporating both anatomical information from a CT image and functional information from a PET image acquired in the same session. In a first step towards full body quantification, we focused on bone tumour load in the definition of the indices. We also developed a method to compute them automatically with possible manual corrections, so that the indices can be easily used in clinical practice. We have applied this method to a cohort of prostate cancer patients with bone metastases that underwent ^{68}Ga -PSMA-PET/CT for staging or re-staging and compared the results of our multimodal imaging indices to the standards of BSI, serum PSA, and clinical expert reading using PERCIST.

6.2 Material and methods

6.2.1 Bone PET Index (BPI)

BSI is an approximation of the percentage of skeletal mass affected by tumor calculated on a bone scintigram [53, 55]. Because bone scintigraphy is intrinsically two-dimensional and lacks detailed anatomical information, a standard weighting of bones is incorporated in the calculation. Inspired by this definition, we define two new multimodal imaging indices for PET/CT: BPI_{VOL} is the percental bone volume (including bone marrow) affected by tumour. BPI_{SUV} additionally considers the target expression measured by average of the standardised uptake value SUV_{mean} . In both indices, the anatomical information is extracted from the CT image while the functional information is extracted from the PET image, making them intrinsically multimodal. The formulas are as follows:

$$\text{BPI}_{\text{VOL}} = 100 \times \frac{\text{Bone metastases volume}}{\text{Skeleton volume}} \quad (\text{no unit}) \quad (6.1)$$

$$\text{SUV}_{\text{mean}} = \frac{\sum_{v \in \text{lesions}} \text{SUV}(v)}{\sum_{v \in \text{lesions}} 1} \quad (\text{g/mL}) \quad (6.2)$$

$$\text{BPI}_{\text{SUV}} = \text{BPI}_{\text{VOL}} \times \text{SUV}_{\text{mean}}/100 \quad (\text{g/mL}) \quad (6.3)$$

Contrary to the calculation of BSI, no standard weighting of the bones is needed because patient-specific anatomical information from the CT image is used instead. In PET/CT, depending on the type of cancer, regularly only the trunk and not the whole body is imaged. Therefore, to achieve a standardized calculation of the BPI, arms and legs as well as part of the head were excluded from the computation: only the image slices between the bottom of the ischium (easily recognized on CT) and the caudal edge of the sub-lingual gland (easily recognized due to glandular uptake in PET) were taken into account. Of note, in the computation of BSI by EXINI bone^{BSI}, the forearms and lower legs are excluded as well.

6.2.2 Automatic computation method

To compute BPI_{VOL} , SUV_{mean} and BPI_{SUV} , a precise segmentation of the skeleton in CT and of bone metastases in PET are necessary. This can be done manually with appropriate software, but is time-consuming. We propose instead an automated method with possible manual corrections. The method has been incorporated in a package programmed in Python.

Preprocessing The tool read images in DICOM format. PET and CT were affinely registered using information contained in the DICOM headers. This was possible because both images were acquired on the same scanner during the same session. The bed was automatically removed from the CT by simple morphological operations.

Bone segmentation on CT On CT, the skeleton can be segmented by using its density in Hounsfield Units, which is higher than that of soft-tissue and air. We used the first two steps of the method of Kang *et al.* [55], i.e. global and local thresholding, followed by morphological operations. These are detailed in the next paragraphs.

Global thresholding First, a low and a high threshold were computed by fitting a mixture of two Gaussian distributions $\mathcal{G}_1(m_1, \sigma_1)$ and $\mathcal{G}_2(m_2, \sigma_2)$ with respective means m_1 and m_2 and respective standard deviations σ_1 and σ_2 to the histogram of CT intensities (excluding the background). Without loss of generality, we assumed $m_1 < m_2$. After fitting, the low and high thresholds LT and HT were computed as:

$$\begin{aligned} LT &= \min(160, m_2 + 1.7\sigma_2) \\ HT &= LT + 400 \end{aligned} \tag{6.4}$$

All pixels that had an intensity greater than HT were considered as bone. All pixels that had an intensity smaller than LT were considered as not bone. All pixels that had an intensity between LT and HT were considered as undetermined and were labelled in the next step. To avoid

labelling endoprosthesis as bones, all voxels with an intensity above 2000 HU were excluded from the bone mask.

Local thresholding In a second step, the pixels that could not be labelled by global thresholding were considered. For each pixel, the local mean m_{loc} and local standard deviation σ_{loc} were computed within a 26-pixels neighbourhood. The local threshold for the current voxel was set to $m_{loc} - 0.8\sigma_{loc}$.

Morphological operations To further correct the segmentation and include the bone marrow, morphological operations were applied. All connected components with a size inferior to forty-five pixels were removed. The bone mask was dilated by one pixel, hole filling was applied to each slice and the bone mask was eroded by one pixel.

Manual corrections In case of heavy calcification or in case of artefact generating medical objects (e.g. a pacemaker), manual corrections were possible: corrections could be applied either with a brush, or by removing in one click the whole "bone" from a slice.

Lesion segmentation on PET ^{68}Ga -PSMA usually does not exhibit unspecific uptake in the bones and bone marrow. Therefore, regions of the skeleton with increased uptake can be considered as bone metastases. Thus, the lesions were segmented by using a SUV-threshold on PET and restricting the result to the skeleton segmented from the CT image (Figure 6.1). This use of anatomical information avoided manual removal of normal uptake sites (e.g. bladder, kidneys) as proposed in [34].

Even though CT and PET were acquired consecutively on the same scanner, e.g. breathing can cause misalignment in the region of the ribs. As the liver and spleen show high physiological uptake of ^{68}Ga -PSMA, projection of the ribs in CT on liver and spleen in PET can potentially generate false positives. For now, such false positives had to be manually corrected.

SUV-threshold choice The final lesion segmentation depends on the SUV threshold. Rather than choosing it arbitrarily, we propose a method to calculate it using a negative training patients cohort. For the cancer-negative patients, the value of BPI_{VOL} is the percentage of the skeleton that is falsely segmented as lesions. Since the training cohort was chosen to be negative for cancer, BPI_{VOL} should have been 0 for these patients. However, because of the noise in the image, in some voxels, the intensity exceeded the normal background uptake. Depending

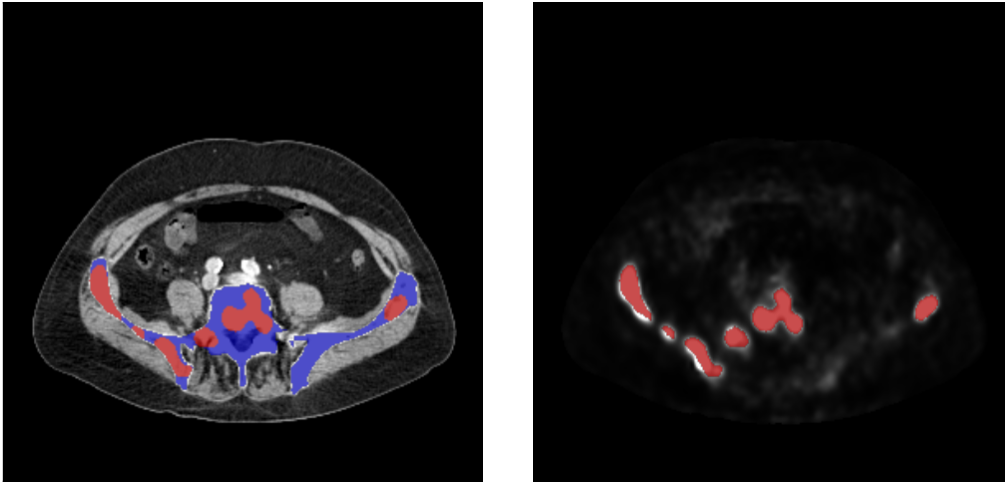


Figure 6.1 – PET and CT image of one patient. The blue overlay shows the bone mask computed by the tool and the red overlay shows the tumour mask computed with a SUV-threshold of 3.

on the chosen SUV-threshold, the thresholding of the PET image then resulted in a non-empty set of bone lesions, which lead to BPI_{VOL} having a strictly positive value.

We defined a user-specific percentage of tolerated false positives FP_{max} . For each patient of the cohort, BPI_{VOL} was then computed using the method described above with a wide range of thresholds $[0, t_{max}]$. With a SUV-threshold of 0 g/mL, BPI_{VOL} was 100, and with a SUV-threshold of 4 g/mL, for cancer-negative patients, it was very close to 0. The recommended threshold t_{rec} for each patient was defined as :

$$t_{rec} = \min_t(t \in [0, t_{max}] | BPI_{VOL} < FP_{max}) \quad (6.5)$$

The SUV-threshold for the testing cohort could then be chosen as the maximum of all t_{rec} obtained for the training cohort.

6.2.3 Patient cohort

Data of forty-five patients with metastatic castration resistant prostate cancer undergoing a Radium-223-dichloride therapy (Xofigo, Bayer Healthcare, Leverkusen, Germany) were analysed retrospectively. They received therapy at a dose of 50 kBq/kg per therapy cycle in monthly intervals with up to 6 cycles. Mean age of the patients was 71 (± 8 years). All patients had bone metastases, but none of the patients showed organ or relevant (>3 cm) lymph node metastases. Fifteen patients that underwent a ^{68}Ga -PSMA-PET/CT at our institution for prostate cancer staging or re-staging and were regarded as negative for bone metastases by an

Test	Before therapy	After three cycles	After six cycles
Training patients			
⁶⁸ Ga-PSMA-PET/CT	15	N/A	N/A
Prostate cancer patients			
⁶⁸ Ga-PSMA-PET/CT	45	32	20
PSA serum value	43	33	21
BSI	31	21	18

Table 6.1 – Available data for the total of 60 patients.

experienced nuclear medicine physician were also retrospectively randomly selected to serve as training subjects.

The institutional review board of the Technical University Munich approved the retrospective analysis (permit 5665/13) and all subjects signed a written informed consent for anonymised evaluation and publication of their data.

All patients underwent ⁶⁸Ga-PSMA-PET/CT within four weeks prior to initiation of Radium-223-dichloride therapy. Thirty-one patients also underwent bone scintigraphy. Thirty-two patients underwent additional ⁶⁸Ga-PSMA-PET/CT and twenty-two of them bone scintigraphy three to six months after the first scan (equivalent to after three or six cycles of Radium-223-dichloride). Table 6.1 shows a summary of the data available for all patients.

6.2.4 Data acquisition and analysis

⁶⁸Ga-PSMA-PET/CT was obtained approximately 62 minutes (range: 45-99) after injection of mean 131 MBq (± 69 MBq, range: 52-239 MBq) ⁶⁸Ga-labelled HBED-CC. A diagnostic CT scan was performed in the portal venous phase after intravenous injection of contrast agent (Imeron 300). Immediately after the CT, the PET scan was acquired with 6-8 bed positions (3-5 minutes per bed position). PET was reconstructed using ordered-subset expectation-maximisation with point spread function and time-of-flight information (3 iterations, 21 subsets) and corrected for normalisation, attenuation, scatter, randoms and decay. The transaxial pixel size was 4.07 mm for PET and 1.52 mm for CT and the slice thickness was 5 mm for both. ^{99m}Tc-HDP whole-body bone scintigraphy was performed in planar imaging mode with an acquisition time of 1 minute / 10 cm body height. Activity was body weight-adjusted (9 MBq/kg) and injected 3 hours before imaging.

6.2.5 Statistical analysis

To validate our new tool and the introduced BPI, we performed a reproducibility analysis. Ten datasets randomly selected from our prostate cancer patient cohort prior to application of Radium-223-dichloride were analysed by two trained observers applying manual corrections independently. The reproducibility threshold was then defined as the maximum absolute difference observed between both observers for each index. For BSI, Anand et al. [5] defined the reproducibility threshold as 0.30.

For response assessment, BPI was compared to BSI and PERCIST by an experienced reader as well as PSA. BSI was computed from the bone scintigraphy images using the commercially available software EXINI bone^{BSI}. Response by PERCIST was evaluated by an experienced physician using recently published criteria [109, 51], and criteria were adapted for ⁶⁸Ga-PSMA similarly to a recent work [113]. In brief, SUV_{peak} value was measured in one to five target lesions and the appearance of new lesions was investigated. As PERCIST is not quantitative, but only classifies the patient in progressive metabolic, stable metabolic disease or partial metabolic response, we also defined these categories for BPI and BSI using the respective reproducibility threshold: a change of magnitude smaller than the reproducibility threshold was considered as stable metabolic disease, an increase in value larger than the reproducibility threshold was considered as progressive disease and a decrease in value larger than the reproducibility threshold was considered as partial metabolic response. Moreover, two separate analyses based on PERCIST criteria were performed: for metastatic status based on all types of target lesions (including potential new lymph node and visceral metastases, as prescribed by the criteria) and for metastatic status based on bone involvement only (to allow for direct comparison with BPI). For comparing quantitative methods (i.e. BPI, BSI and PSA), we used the Pearson coefficient r . For all tests, a p -value smaller than 0.05 was considered significant.

6.3 Results

6.3.1 Technical validation

Bone segmentation After manual correction, the mean bone volume of the forty-five treated patients was $4,184 \text{ cm}^3$ ($\pm 503 \text{ cm}^3$, range: $3,327\text{-}5,739 \text{ cm}^3$). For 32 patients with two sequential ⁶⁸Ga-PSMA-PET/CT, the mean difference in computed bone volume between two scans was 66 cm^3 ($\pm 61 \text{ cm}^3$) with a maximum of 270 cm^3 . This is an average difference of less than 2%. The small discrepancy can be explained by slightly different positions of the patient in the scanner. The absolute values obtained were in the expected range with a reference human

skeleton having an estimated total volume (including bone marrow, averaged for both sexes) of $7,700 \text{ cm}^3$ [99] and 50.3 % of the total mass of the skeleton being included in our segmentation [106].

Manual corrections After bones had been segmented and manually corrected, false positives (e.g. in the rib cage) were corrected by an expert reader. On fifty-four scans from the patient cohort, an average of 3.8 cm^3 false positives per patient had to be manually corrected. This represented an average difference in BPI_{VOL} of 0.0008 per patient.

Selection of SUV-threshold for lesion segmentation The SUV-threshold lesion segmentation was determined using the fifteen training patients. For each of them, we computed the threshold that resulted in a BPI_{VOL} of 0.1 and 1 (equal to 0.1% and 1% of false positive voxels) respectively. Corresponding SUV-thresholds for all training patients were in the range of 1.15 to 1.95 g/mL (mean: 1.42 g/mL) for a false positive threshold of 1 and of 1.7 to 2.65 g/mL (mean: 2.06 g/mL) for 0.1, respectively. Figure 6.2 shows the different BPI_{VOL} values obtained for different thresholds for one negative training patient.

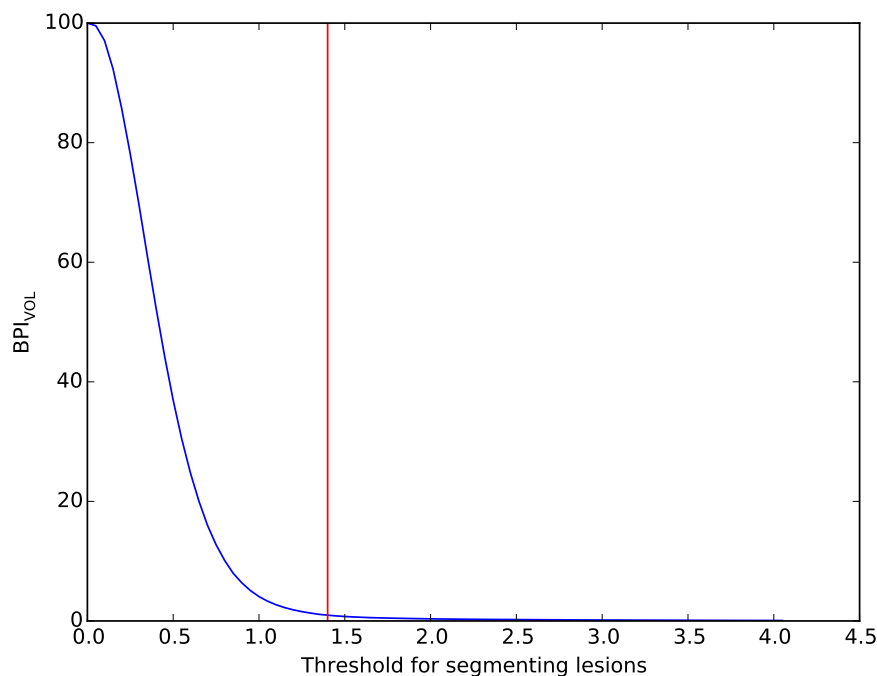


Figure 6.2 – BPI_{VOL} values obtained with different thresholds for one cancer-negative training patient. The red line shows that a threshold of 1.4 g/mL gave a BPI_{VOL} of 1.

SUV-threshold influence on lesion segmentation Based on these results, thresholds of 1.5 g/mL (average value obtained for the scenario of 1% false positive results) and 3 g/mL (conservative approach ensuring a maximum of 0.1% false positive lesions in all patients) were used for the initial analysis of the baseline ^{68}Ga -PSMA-PET/CT of all forty-five patients. There was a strong correlation between the BPI_{SUV} values obtained with these two thresholds ($r=0.99$; $p<0.001$, Figure 6.3). The values of BPI_{VOL} and SUV_{mean} computed with the two thresholds showed a similarly high correlation ($r=0.95$ for both; $p<0.001$). Due to the high correlation between both values, we chose a threshold of 3 g/mL to ensure a high specificity of the BPI, with less of 0.1 of BPI_{VOL} being related to false positives. All following results were computed using a SUV-threshold of 3 g/mL.

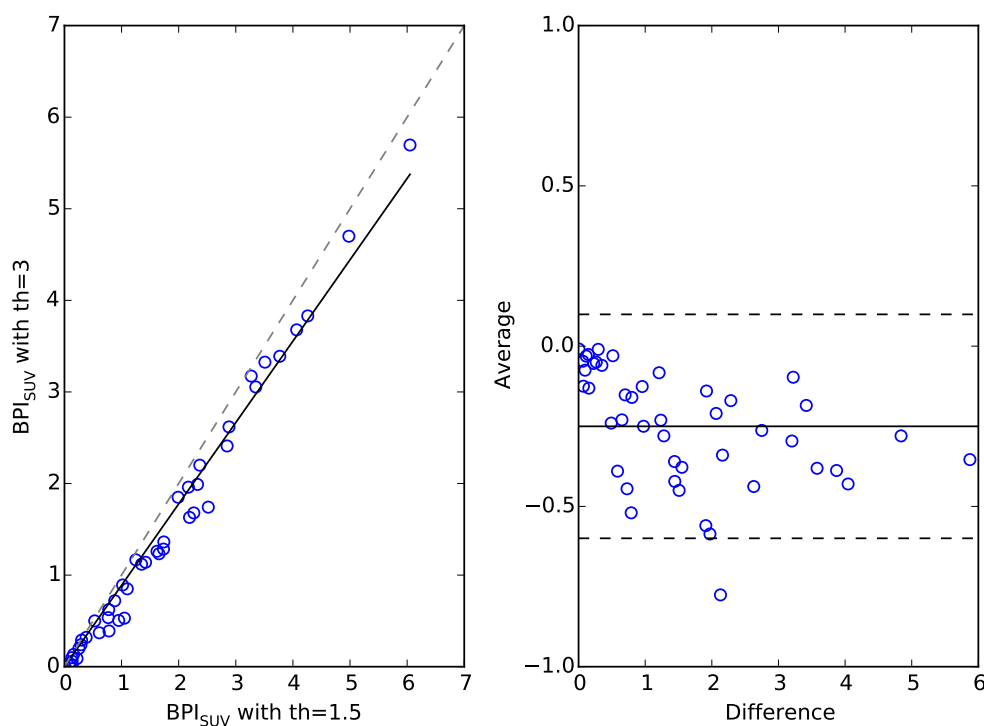


Figure 6.3 – The difference between the two BPI_{SUV} obtained with cut-off values 1.5 g/mL and 3 g/mL is shown by the Bland-Altman plot on which the differences between two BPI_{SUV} are plotted against their average. They show a mean difference of $-0.25 \text{ BPI}_{\text{SUV}}$ (95% confidence intervals, $+0.1$ and $-0.6 \text{ BPI}_{\text{SUV}}$), indicating systematically lower BPI_{SUV} values for a cutoff value of 3, as expected.

Reproducibility Comparison from two independent observers using 10 randomly selected datasets showed a nearly perfect correlation ($r=0.999$; $p<0.001$ for both). The maximum observed percentage differences between both observers were 3.5% for BPI_{SUV} and 2.2% for BPI_{VOL} . The maximum absolute difference was 0.055 g/mL for BPI_{SUV} and 0.37 for BPI_{VOL} . We defined

0.06 g/mL and 0.4 as reproducibility thresholds for the respective index. Note that a wide range of disease was present in the analysed patients (range of BPI_{SUV} : 0.09-3.39 g/mL, BPI_{VOL} : 1.53-38.05).

6.3.2 Quantification using BPI_{VOL} , SUV_{mean} and BPI_{SUV}

BPI_{VOL} , SUV_{mean} and BPI_{SUV} before and after therapy The average values of BPI_{VOL} , SUV_{mean} and BPI_{SUV} before therapy were 19.5, 8.3 g/mL and 1.59 g/mL respectively. After therapy, the average values were 26.0, 7.7 g/mL and 1.99 g/mL. This represents changes of +33%, -7% and +25% respectively.

Correlation between BPI_{SUV} and BPI_{VOL} BPI_{VOL} and BPI_{SUV} for all forty-five patients before therapy were strongly correlated ($r=0.89$, $p<0.001$, Figure 6.4). The percentage changes of BPI_{VOL} and BPI_{SUV} during therapy were very strongly correlated ($r=0.97$, $p<0.001$, Figure 6.5).

It is notable that the two introduced indices BPI_{VOL} and BPI_{SUV} are in principle highly correlated although they are not completely equivalent, since BPI_{SUV} also takes into account the level of expression of PSMA. The percentage changes of BPI_{VOL} and BPI_{SUV} during therapy were highly correlated despite one outlier for which they changed in opposite direction (Figure 6.5). Interestingly, while the average values of BPI_{VOL} and BPI_{SUV} increased during therapy, the average value of SUV_{mean} decreased. This shows that BPI_{VOL} , SUV_{mean} and BPI_{SUV} provide different information and that their predictive properties have to be explored in a prospective study with a large patient cohort.

6.3.3 Correlation of BPI to clinical parameters

At baseline, BPI_{VOL} and BPI_{SUV} showed a moderate and significant correlation with BSI ($r=0.76$ and 0.74 respectively, $p<0.001$, Figure 6.6). There was a tendency to a stronger correlation with PSA-value for BPI_{VOL} and BPI_{SUV} ($r=0.57$ and 0.54 respectively, $p<0.01$) than for BSI ($r=0.49$, $p<0.01$). A moderate correlation between change of BPI_{VOL} and BPI_{SUV} and percentage change of PSA-value after treatment was observed ($r=0.70$; $p<0.01$). There was no significant correlation of change in BSI with percentage change in PSA-value ($r=0.24$; $p=0.32$).

When compared to PERCIST for the whole body (Table 6.2), BPI_{VOL} , BPI_{SUV} and BSI showed agreement for 65.6% (21/32), 68.7% (22/32) and 57.9% (11/19) of patients and opposite results for 25.0% (8/32), 15.6% (5/32) and 21.1% (4/19) respectively. When compared to PERCIST applied to bones only (Table 6.3), BPI_{VOL} , BPI_{SUV} and BSI showed agreement for 62.5% (20/32),

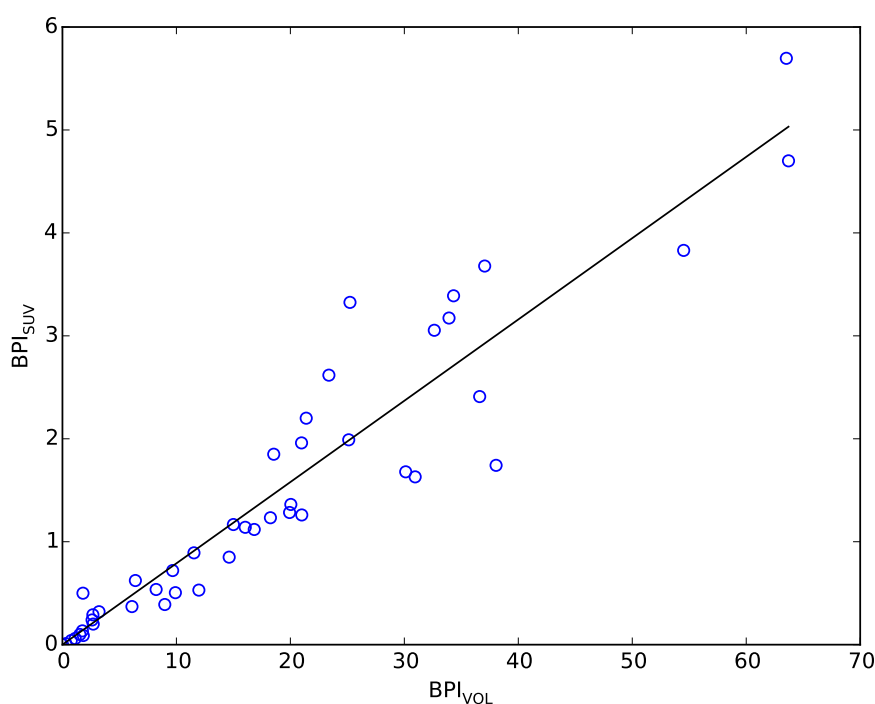


Figure 6.4 – BPI_{VOL} and BPI_{SUV} values. A significant correlation was observed. Images from all forty-five patients with bone metastases were used.

Table 6.2 – Comparison of comprehensive PERCIST to BPI and BSI classifications. A total of thirty-two patients for BPI and nineteen patients for BSI were classified.

		BPI_{VOL}			BPI_{SUV}			BSI		
		Prog.	Stab.	Resp.	Prog.	Stab.	Resp.	Prog.	Stab.	Resp.
Percist	Prog.	17	2	6	17	3	5	8	4	4
	Stab.	1	0	0	1	0	0	0	1	0
	Resp.	2	0	4	0	1	5	0	0	2

68.7% (22/32) and 63.2% (12/19) of patients and opposite results for 12.5% (4/32), 6.2% (2/32) and 10.5% (2/19) respectively.

When comparing both parameters to objective imaging response evaluation for the whole body using PERCIST, BPI_{SUV} showed a higher agreement than BSI and also BPI_{VOL} . It was not unexpected as BPI_{SUV} takes into account also intensity values, as does PERCIST. When comparing to objective imaging response evaluation using PERCIST on bones only, the number of patients showing opposite results was much lower, as expected. For BPI_{VOL} and BPI_{SUV} , two patients were classified as responsive to therapy while PERCIST classified them as progressive. Both patients experienced a considerable decrease of PSMA-expression under therapy, whilst new small bone lesions appeared (example in Figure 6.7). As by definition these patients were

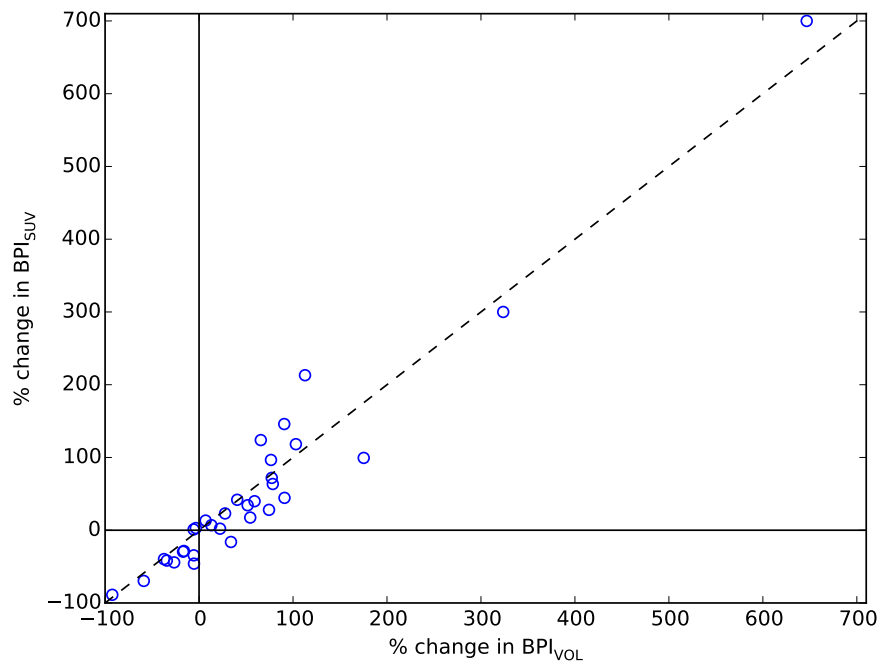


Figure 6.5 – Change in BPI_{SUV} and BPI_{VOL} during therapy for thirty-two patients. A high and significant linear correlation was observed.

Table 6.3 – Comparison of bone-PERCIST to BPI and BSI classifications. A total of thirty-two patients for BPI and nineteen patients for BSI were classified. Expert reader took only bone tumour into account to establish the PERCIST-classification.

		BPI_{VOL}			BPI_{SUV}			BSI		
		Prog.	Stab.	Resp.	Prog.	Stab.	Resp.	Prog.	Stab.	Resp.
Percist	Prog.	15	2	2	15	2	2	7	2	2
	Stab.	3	0	3	3	1	2	1	3	2
	Resp.	2	0	5	0	1	6	0	0	2

classified as progressive by PERCIST, the clinical significance of this mixed response by imaging remains unclear.

6.4 Discussion

In this study, we have described the bone PET indices as new quantitative multimodal imaging indices for the assessment of bone metastases in PET/CT using a novel automatic computation method. To the best of our knowledge, this is the first report of a multimodal imaging index taking into account both anatomical information from CT and functional information from PET. Our data indicated that the BPI is robust and reproducible. A small amount of manual correction was still necessary, especially due to misalignments, calcifications and endoprosthesis.

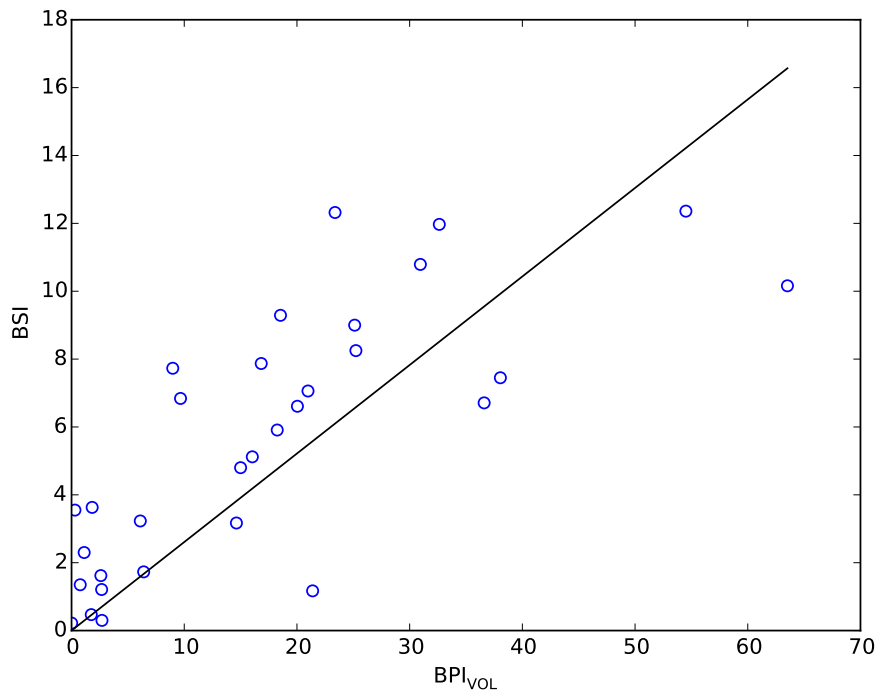


Figure 6.6 – BPI_{VOL} and BSI values for thirty-one patients before the beginning of the therapy. A moderate correlation was observed.

The computation with corrections only took a few minutes, whilst a fully manual segmentation would have taken several hours even to an experienced reader. We have shown that BPI has potential for quantitative response assessment. This is documented by a high correlation of BPI with BSI and PSA in metastatic prostate cancer patients, and reasonable prediction of tumour response compared to PERCIST despite the fact that neither of them (due to known limitations) can serve as standard of reference for BPI and they only allow a first estimation of the potential clinical usability. For further assessment, future clinical studies using more reliable endpoints (survival, radiographic progression-free-survival, skeletal adverse events) have to be conducted.

Compared to imaging biomarkers TLF_{10} and FTV_{10} introduced by Etchebere *et al.* [34], BPI_{VOL} and BPI_{SUV} pursue an intrinsically multimodal approach. For TLF_{10} and FTV_{10} , no correction is possible with regard to the patient size whereas for BPI_{VOL} and BPI_{SUV} , the skeleton volume is included, thus allowing for inter-patient comparison.

We first demonstrated that the bone volume using the newly introduced tool was reproducible between different scans of the same patient. After bones had been corrected, the average false positive correction by an expert reader represented a negligible difference in BPI_{VOL} . The

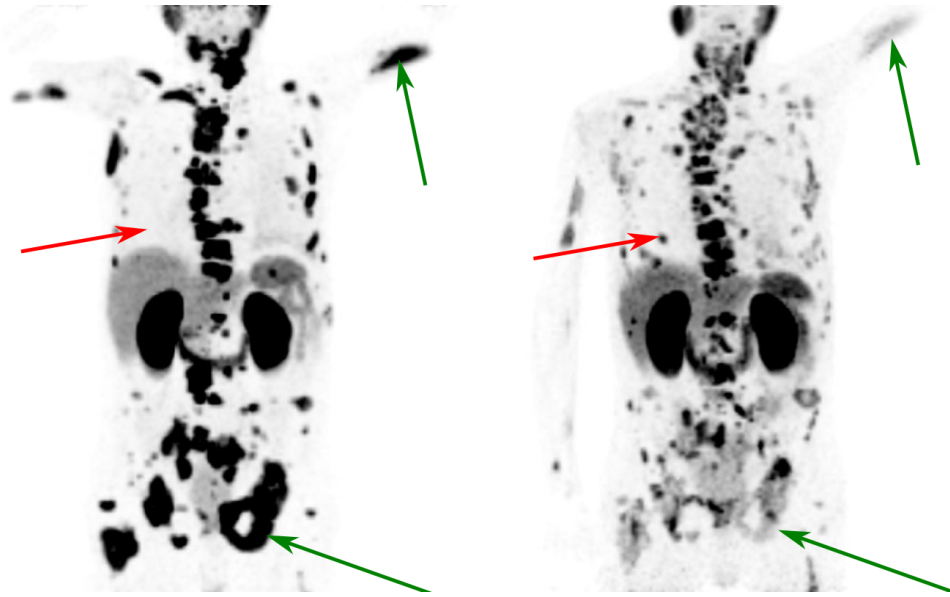


Figure 6.7 – Maximum intensity projection of ^{68}Ga -PSMA-PET images of the same patient before (left) and during (right) Radium-223-dichloride therapy. Green arrows show examples of lesions that responded to therapy. Red arrows show small new lesions in the ribs. Despite the decrease in tumour load, the disease was classified as progressive when following PERCIST criteria.

false positive correction was done with only a few clicks and the full image segmentation with manual corrections only took a few minutes, which makes it usable in clinical practice.

Both imaging indices BPI_{VOL} and BPI_{SUV} were highly reproducible with a maximum inter-observer difference of 3.5%, facilitating their use in clinical practice. Even though the limited influence of different SUV-thresholds for computation underlined the robustness of our clinical analyses, we chose a conservative approach (SUV-cut-off of 3 g/mL) to ensure that less than 0.1 of BPI_{VOL} was due to false positives.

Notably, some clear outliers in the comparison between BPI and BSI (Figure 6.6) were observed. It has to be respected that bone scintigraphy and ^{68}Ga -PSMA-PET/CT image two different biological processes: bone scintigraphy displays the reactive changes of the tumour on the skeleton [77], while ^{68}Ga -PSMA-PET directly shows the intensity of PSMA-expression on viable tumour cells. Thus, no absolute equivalence of BSI and BPI can be expected. Other explaining factors are the higher sensitivity of ^{68}Ga -PSMA-PET/CT for bone metastases [84], the delay after which bone scintigraphy usually shows changes as well as the *flare* phenomenon [77, 105].

The fact that BPI agreed better to PERCIST applied to bones than PERCIST applied to the whole body underlines the limitation of the current definition of BPI as quantitative PET-imaging index for assessment of bone disease and urges the need for further expansion to soft-tissue tumour burden.

Our study is limited in being retrospective. It has to be noted that data for change of BPI after initiation of Radium-223-dichloride was based on a mixed collective of ^{68}Ga -PSMA-PET/CT scans after three or six cycles. Furthermore, the influence of previous lines of treatment was not assessed, which could potentially impact signals derived from both ^{68}Ga -PSMA-PET and bone scan [32, 1]. These confounding factors have to be investigated in future studies encompassing larger patient cohorts. Moreover, in the evaluation of BPI as new quantitative imaging biomarker, clinical endpoints have to be taken into account. For the patient collective presented here, this data is being collected. Another limitation is the fact that no respiratory gating was used, which potentially would minimize the need for manual correction.

6.5 Conclusion

We have introduced BPI_{SUV} and BPI_{VOL} as new multimodal quantitative imaging indices for PET/CT, representing a robust tool for quantitative assessment of osseous tumour burden. We have shown that their automatic computation (with possible manual corrections) was feasible and highly reproducible on a cohort of metastatic prostate cancer patients. Finally, our results demonstrated that BPI_{VOL} and BPI_{SUV} provide clinically meaningful information when correlated to PERCIST, BSI and PSA-value. However, their value in predicting patient outcome has to be explored in future studies.

Chapter 7

Localised quantification of Bone Lesion Burden in PET/CT

In this chapter, we present a localised quantification method for bone tumour assessment in PSMA-PET/CT images.

7.1 Motivation

As mentioned in previous chapters, ^{68}Ga -PSMA-PET/CT is increasingly used for diagnosis and staging of prostate cancer, but, thus far, the only fully quantitative analysis method that uses both modalities is the one that we presented in chapter 6. It utilises new quantitative indices for bone lesion assessment in PSMA-PET/CT images. These indices overcome several limitations of other existing image analysis methods: they are quantitative, take all lesions into account, and allow for inter-patient comparison. They also give a different approach when there is a mixed response to therapy than the one used by PERCIST: as all lesions are used to compute the indices, these reflect the global tumour burden. While PERCIST designates new lesions as a clear sign of progress, new lesions only lead to a (possibly small) increase in BPI_{VOL} and BPI_{SUV} . Figure 6.7 shows an example of patient where new lesions appear but BPI_{VOL} and BPI_{SUV} decrease during therapy. Similarly, it is possible that in case of a heterogeneous response, tumour growth and tumour shrinkage in different lesions *compensate* each other, leading to deceptively stable BPI_{VOL} and BPI_{SUV} .

Global quantification can therefore lead to unclear results. For patients with high tumour load however, following individual lesions can be difficult, as lesions can merge due to progress or split when responding to therapy. To avoid both pitfalls, we propose a new method for *localised* quantification of bone lesions in PSMA-PET/CT images. Instead of quantifying the image globally, we propose to delineate anatomically meaningful regions and assess each of these individually. The situation described above with different tumour responses compensating each other is much less likely to happen in a small homogeneous region than in a whole-body or thorax-abdomen image.

In the following, we define new localised quantitative image indices and describe a method to compute them automatically (section 7.2), present a proof-of-concept analysis on a prostate cancer patient cohort (section 7.3), discuss our results (section 7.4) and offer conclusions (section 7.5).

7.2 Material and methods

7.2.1 Localised Bone PET Index (LBPI)

In chapter 6, BPI_{VOL} was defined as the percental bone volume of the skeleton affected by tumour whilst BPI_{SUV} additionally took SUV_{mean} into account. We propose to apply these definitions to parts of the skeleton in order to define the analogous localised indices $LBPI_{VOL}$, $LSUV_{mean}$ and $LBPI_{SUV}$:

$$LBPI_{VOL}(\text{part}) = 100 \times \frac{|\text{lesions} \cap \text{part}|}{|\text{part}|} \quad (\text{no unit}) \quad (7.1)$$

$$LSUV_{mean}(\text{part}) = \frac{\sum_{v \in \text{lesions} \cap \text{part}} SUV(v)}{|\text{part}|} \quad (\text{g/mL}) \quad (7.2)$$

$$LBPI_{SUV}(\text{part}) = LBPI_{VOL}(\text{part}) \times LSUV_{mean}(\text{part})/100 \quad (\text{g/mL}) \quad (7.3)$$

These definitions can be applied to any part of the skeleton for which a segmentation is available. If the part is the whole skeleton, these indices are the same as BPI_{VOL} , SUV_{mean} and BPI_{SUV} . In section 7.3, we evaluate them for anatomically meaningful parts of the skeleton.

7.2.2 Automatic computation method

For computation of $LBPI_{VOL}$, $LSUV_{mean}$ and $LBPI_{SUV}$ for a region of the skeleton, a precise segmentation of that part of the skeleton in CT and of lesions in the corresponding area of the PET image are necessary. This could be done manually, but would be very time-consuming. We instead adapt methods described in chapter 4 and chapter 6 to compute the indices automatically.

Bone segmentation in CT For bone segmentation, we used the dual thresholding method described in section 6.2.2.

Part localisation in the skeleton For localising parts in the skeleton, we used the method described in section 4.2 and segmented the following ten parts: skull, sternum, both arms, both legs, pelvis, spine, right-side and left-side ribs. We used no initial landmarks, only one hierarchical level and three iterations of saRF before regularisation with the CRF.

Lesion segmentation in PET For segmenting lesions, we used the thresholding method described in section 6.2.2. We showed in section 6.3.1 that manual correction of false positives caused for example by projection of liver and spleen in PET on the ribs in CT had only a negligible influence on the index values. Thus, we did not perform false positive manual correction. Moreover, we also showed in section 6.3.1 that, within reasonable bounds, the exact value of the SUV threshold used for segmenting lesions had little influence on the analysis. We therefore chose the same threshold as in chapter 6 and used 3 g/mL throughout this chapter. An example of part localisation and lesion segmentation is shown in Figure 7.1.

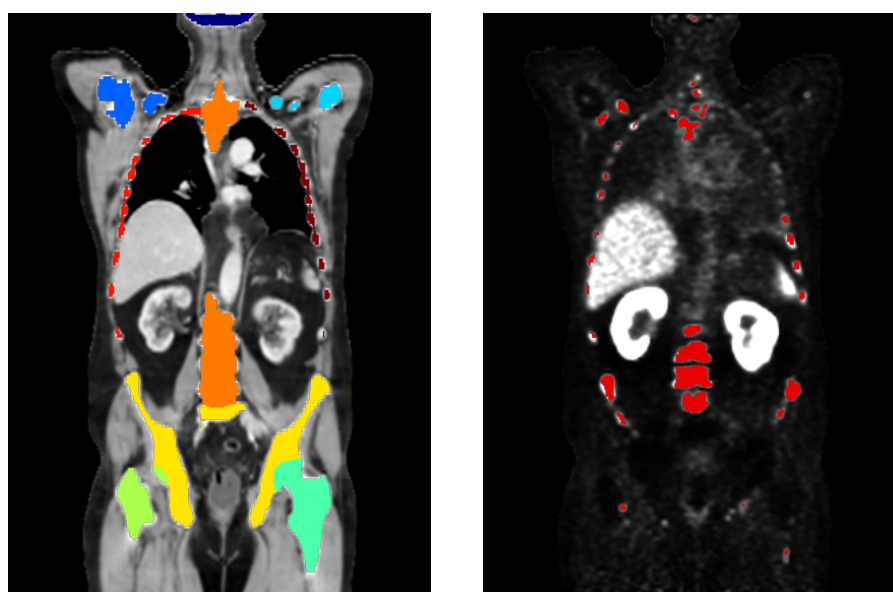


Figure 7.1 – Left: example of CT image overlaid with skeleton segmentation and part localisation. Right: example of PET image overlaid with lesion segmentation.

Patient cohort We used the same patient cohort as in chapter 6 for analysing the indices. The cohort includes forty-five patients with metastatic castration resistant prostate cancer undergoing a Radium-223-dichloride therapy (Xofigo, Bayer Healthcare, Leverkusen, Germany). A detailed description of the cohort can be found in subsection 6.2.3.

All patients underwent ^{68}Ga -PSMA-PET/CT within four weeks prior to initiation of Radium-223-dichloride therapy and further scans every three months until interruption of the treatment. The available data is summarised in Table 7.1.

Data acquisition A detailed description of the acquisition and reconstruction protocols for ^{68}Ga -PSMA-PET/CT can be found in subsection 6.2.4. When possible, patients were scanned with their arms up. However, when pain or other impairments made this impossible, they were

Test	Before therapy	After 3 cycles	After 6 cycles
^{68}Ga -PSMA-PET/CT	45	32	20

Table 7.1 – Available data for the total of forty-five prostate cancer patients.

scanned with their arms down. In that case, manual corrections were performed on the part localisation to account for position differences. Moreover, for some acquisitions, the full head was scanned instead of scanning only up to the eyes. In our patient cohort, both arms and head positions varied between different scans of the same patient. To allow for a fair comparison, we excluded these of our analysis.

Data analysis In chapter 6, we found the reproducibility threshold of BPI_{VOL} and BPI_{SUV} to be 0.4 and 0.06 g/mL respectively. As the exact reproducibility threshold for LBPI_{VOL} and LBPI_{SUV} may depend on the analysed part of the skeleton, we chose to use the same thresholds as BPI_{VOL} and BPI_{SUV} for LBPI_{VOL} and LBPI_{SUV} . We also defined an index to be stable if it changed with a magnitude less than its reproducibility threshold, and to increase or decrease otherwise.

$\text{LBPI}_{\text{VOL}}(\text{part})$ can be interpreted as the probability of a voxel to belong to a lesion when drawn randomly from that part. To compare distributions in different parts, we used the paired Wilcoxon test and considered p-values below 0.01 as significant.

7.3 Results

We present here results of the analysis of our patient cohort using LBPI_{VOL} and LBPI_{SUV} .

7.3.1 Individual cases

For some patients, all lesions responded to therapy in the same way. However, for many patients, the response was heterogeneous, i.e. some lesions progressed whilst other regressed during therapy. Figure 7.2 shows a case where both LBPI_{VOL} and LBPI_{SUV} evolved in the same way during therapy in all studied parts of the skeleton: over six months, the indices increased in all parts.

Figure 7.3 shows a case where LBPI_{VOL} and LBPI_{SUV} evolved in different directions in different parts of the skeleton: for the spine and the ribs, as well as for the whole skeleton, both LBPI_{VOL} and LBPI_{SUV} decreased during therapy. In the legs however, they increased substantially, being multiplied by more than four.

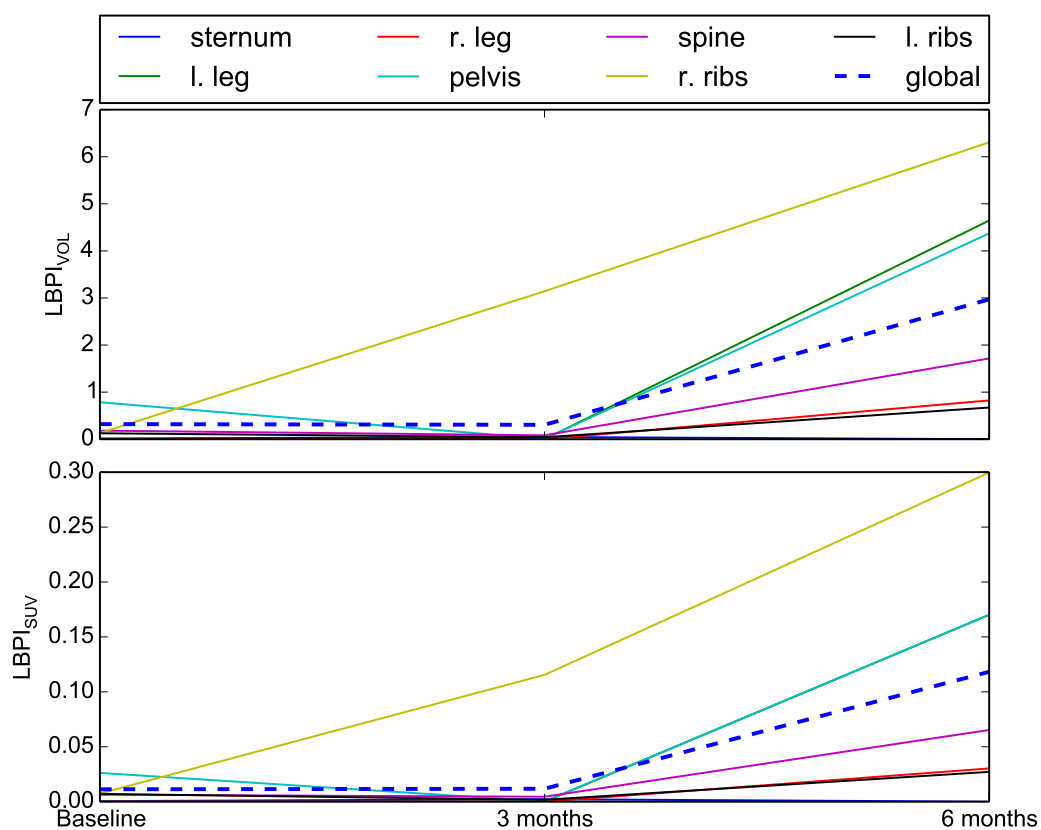


Figure 7.2 – Clinical case 1: evolution of $LBPI_{VOL}$ and $LBPI_{SUV}$ during therapy for seven parts of the skeleton and the whole skeleton (dotted line).

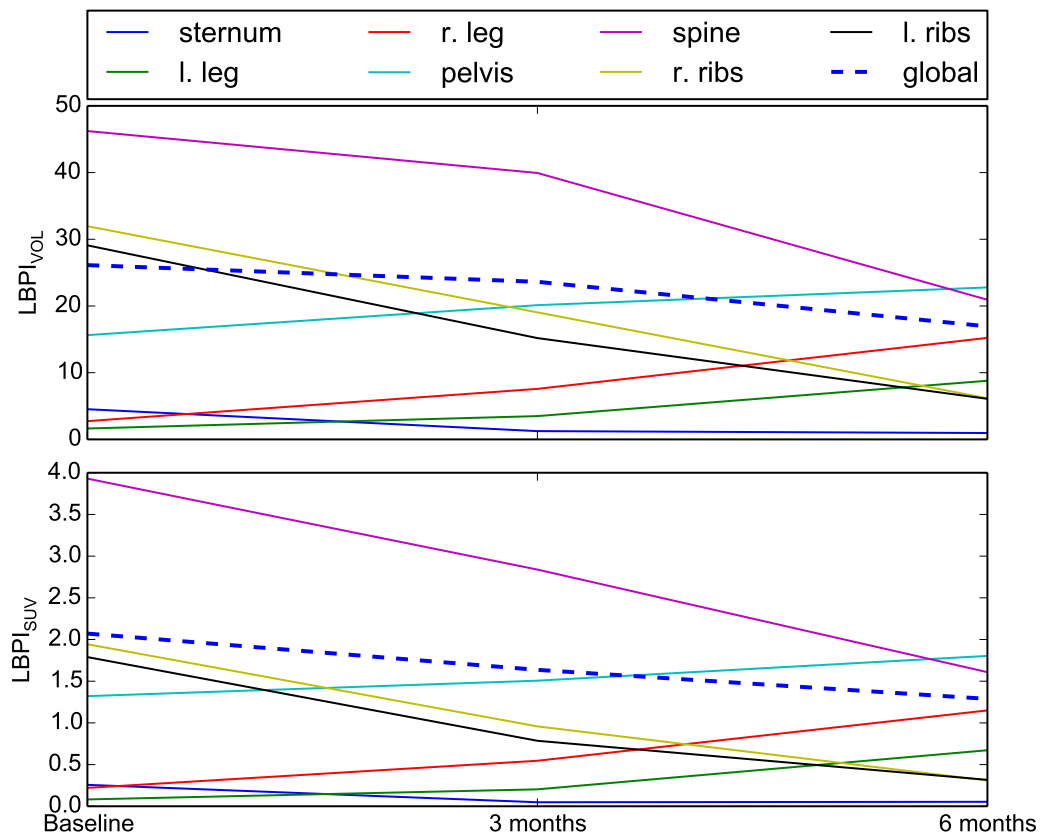


Figure 7.3 – Clinical case 2: evolution of $LBPI_{VOL}$ and $LBPI_{SUV}$ during therapy for seven parts of the skeleton and the whole skeleton (dotted line).

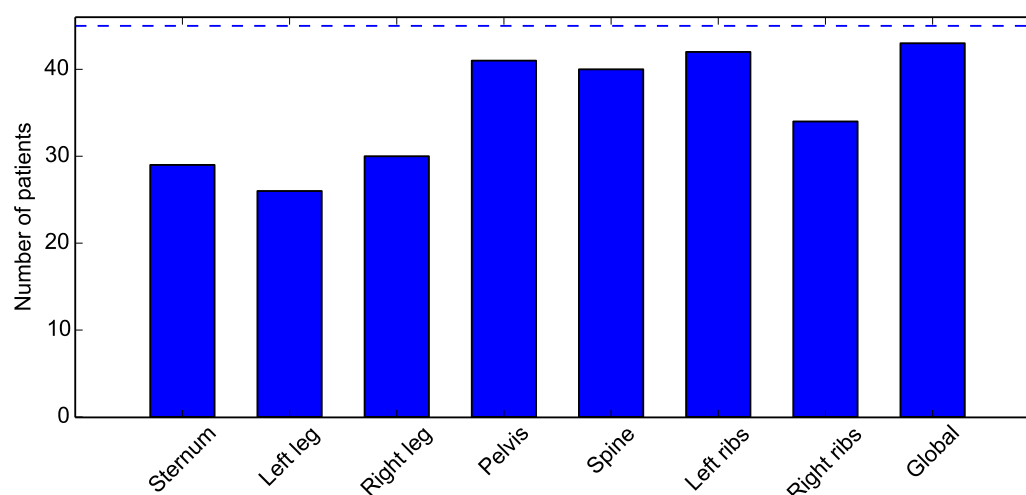


Figure 7.4 – Number of patients affected by lesions in different parts of the skeleton before therapy. The dashed line shows the total number of patients.

7.3.2 Analysis of the patient cohort

At baseline At the patient cohort level, the localised indices allow for analysis of lesions location. In particular, we define a part to be affected by lesion if $LBPI_{VOL}$ in that part is greater than its reproducibility threshold 0.4. Figure 7.4 shows for each part of the skeleton the number of patients affected by lesions in that part. In our patient cohort, lesions were not distributed evenly in the skeleton: the least affected frequently parts were the legs (26 and 30 patients) and the sternum (29 patients).

The frequency of lesions in a part of the skeleton does not take into account the difference in lesion extents and part sizes. However, $LBPI_{VOL}$ is the probability of a randomly drawn voxel to belong to a lesion, knowing in which part of the skeleton it is located, and accounts for size differences. Figure 7.5 shows the distribution of $LBPI_{VOL}$ for different parts of the skeleton over the patient cohort. It demonstrates that not all distributions were identical. For example, the paired Wilcoxon test showed the distributions of $LBPI_{VOL}$ in the left leg and the pelvis, as well as in the right leg and the whole skeleton to be significantly different ($p < 0.001$ for both). On the contrary, the distributions of $LBPI_{VOL}$ for the pelvis and the spine were very similar.

Evolution during therapy For twenty-two patients, a ^{68}Ga -PSMA-PET/CT was available before and after six months of therapy. Figure 7.6 shows for each part of the skeleton the evolution of $LBPI_{VOL}$ and $LBPI_{SUV}$. In particular, for all parts of the skeleton, more than half of the patients experienced an increasing or stable $LBPI_{VOL}$. Moreover, by considering a binomial distribution,

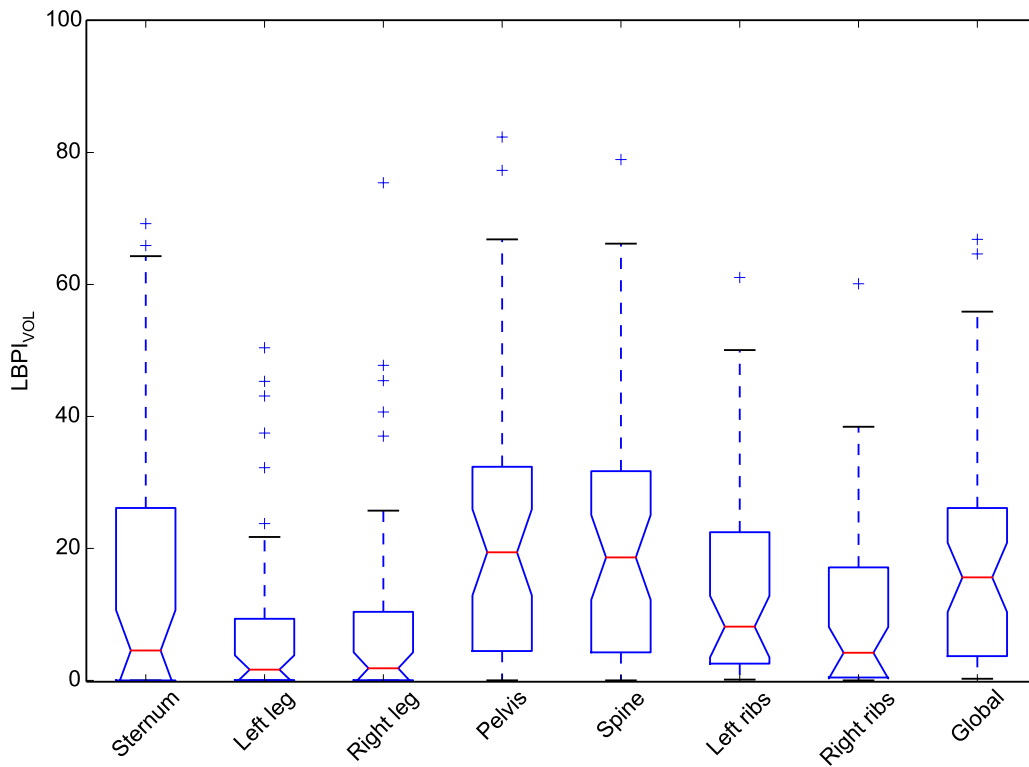


Figure 7.5 – Distribution of $LBPI_{VOL}$ at baseline for different parts of the skeleton.

both for $LBPI_{VOL}$ and $LBPI_{SUV}$, there was no significant difference between parts of the skeleton in the probability of experiencing an increase of the index.

Finally, nineteen of the twenty-two patients experienced a heterogeneous response, i.e. there were at least two parts of the skeleton where $LBPI_{VOL}$ changed in different directions.

7.4 Discussion

In this study, we have described new localised bone PET indices and an automatic computation method. Additionally, we have shown that they can be computed on a metastasised prostate cancer patient cohort. The computation required only a few minutes per patient. To the best of our knowledge, our approach is the first one examining localised quantification of ^{68}Ga -PSMA-PET/CT images. Our evaluation of the indices was a proof of concept and does not constitute a full validation.

First, we have shown that some parts of the skeleton were more frequently affected by lesions than others. This was in particular the case for the pelvis and spine. When taking into account extent of lesions and size of the different parts by studying the distributions of $LBPI_{VOL}$ in these parts, the probability of lesions in the legs still proved significantly lower than

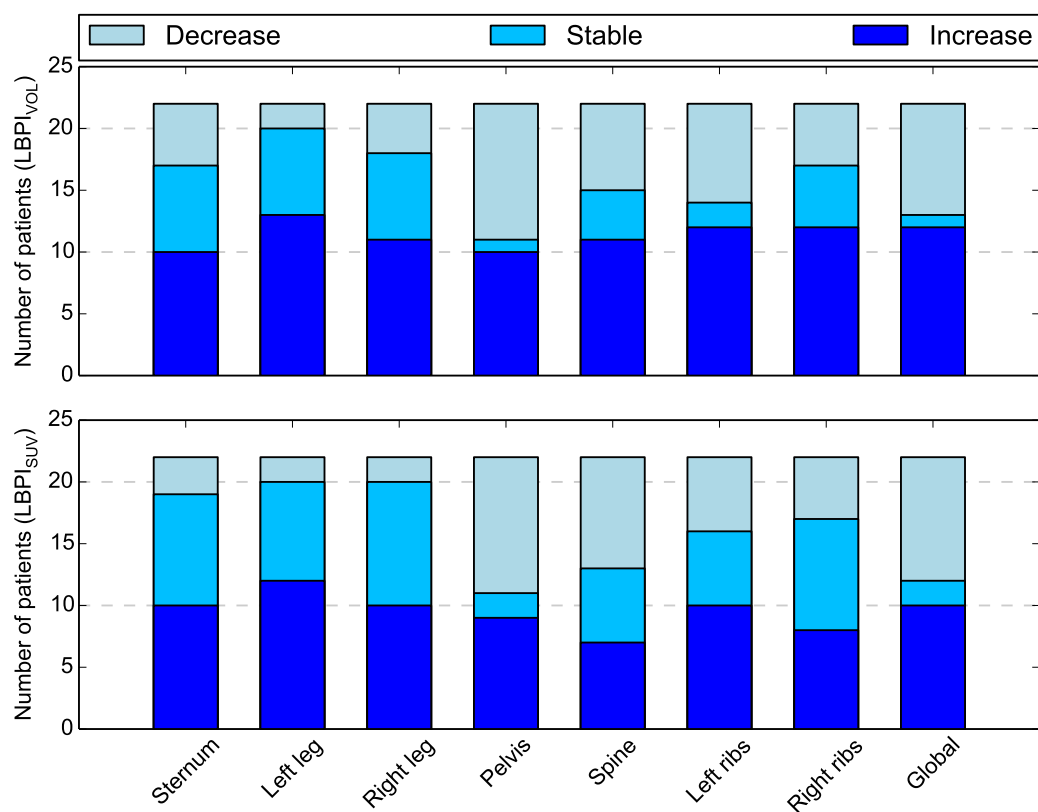


Figure 7.6 – Evolution of $LBPI_{VOL}$ and $LBPI_{SUV}$ in different parts of the skeleton (top: $LBPI_{VOL}$, bottom: $LBPI_{SUV}$).

in the pelvis, spine, and in the whole skeleton. We can therefore conclude that lesions were not uniformly distributed in the skeleton. This confirms an autopsy study [18] that had already described a non-uniform distribution of lesions in the skeleton with more frequent lesions in the spine than long bones and skull. Its results were however not conclusive, because some skeleton parts had not been systematically examined.

Moreover, we have shown that lesions do not always respond uniformly to therapy, i.e. in the same patient, some lesions can respond well to therapy whilst others do not. Heterogeneous response to therapy was very widespread in our patient cohort. However, we could show no systematic therapy response difference between parts of the skeleton and no significant relation between the indices at baseline and their change during therapy. The frequent occurrence of heterogeneous response to therapy shown by our indices should be further explored to see whether it can be related to the biological properties of individual lesions.

As $LBPI_{VOL}$ and $LBPI_{SUV}$ are used to perform *local* quantification, every single one of them can not be expected to correlate well with global biomarkers such as PSA serum value or to be predictive for overall survival. $LBPI_{VOL}$ and $LBPI_{SUV}$ could however be used to decide whether to apply (when possible) a local therapy such as radiotherapy to single lesions to complement an ongoing systematic therapy. To validate the use of our indices in this context, a prospective study should be conducted.

7.5 Conclusion

We have introduced $LBPI_{VOL}$ and $LBPI_{SUV}$ as extensions of BPI_{VOL} and BPI_{SUV} for localised quantification of ^{68}Ga -PSMA-PET/CT images and provided an automatic computation method. We have shown that lesions were not uniformly distributed in the skeleton on a metastasised prostate cancer patient cohort. Moreover, we have demonstrated that heterogeneous responses to therapy with individual lesions of the same patient responding differently to therapy were frequent. Further, prospective, studies should be conducted to explore the clinical meaning of these findings.

Part III

Summary and outlook

Chapter 8

Summary and outlook

In this chapter, we summarize the contributions presented in the rest of the thesis and propose directions for future work.

8.1 Summary

With the improvement of medical imaging acquisition methods, medical imaging has become a central tool in cancer diagnosis and staging. For improving the disease outcome however, image analysis is as important as image acquisition. Moreover, because more and more data is acquired daily, it is important that only minimal, or ideally no manual intervention is required to analyse the images. Machine learning provides automatic methods to learn from data that can be applied to medical image analysis.

In this thesis, we have introduced new approaches based on machine learning for the automatic analysis of PET/CT images for cancer assessment.

In chapter 4 and chapter 5, we have described methods for the automatic segmentation of anatomical structure in CT images. Both methods rely on iterative use of contextual information in a random forest framework. The first method aims at segmenting bones and bone segments, but can also be used to locate large organs. It is computationally efficient and reaches high Dice scores, even for fine structures such as the ribs. The second method has been designed for locating small organs with high variation among subjects such as the pancreas and adrenal glands. Our method does not require any deformable registration to be performed, but still reaches state-of-the-art performance for several organs. By combining both approaches, most anatomical structures can be located in a CT image of the human body.

Furthermore, we have introduced multimodal indices, BPI_{VOL} and BPI_{SUV} , for assessment of bone lesions in PET/CT images in chapter 6. Contrary to existing approaches, they are quantitative, take into account the full tumour load, and allow for inter-patient comparison. We have also provided an automatic method with possibility for manual corrections for calculating these

indices for ^{68}Ga -PSMA-PET/CT images. We proved our indices to be technically robust and a comparison to several biomarkers showed that they provide clinically meaningful information.

Finally, we have combined our methods for localisation in CT and quantification in PET/CT to develop new localised quantification indices, LBPI_{VOL} and LBPI_{SUV} , that can be computed automatically, in chapter 7. These indices offer new insights into lesion location and heterogeneous responses to therapy, that can lead to a deceiving stability of global indices. By applying LBPI_{VOL} and LBPI_{SUV} to a prostate cancer patient cohort, we were able to show that lesions were not distributed uniformly in the skeleton, and that mixed responses to therapy are frequent.

8.2 Future work

Our work has opened new directions for research, that we present here.

8.2.1 Anatomical structure localisation in PET/CT images

In chapter 4 and chapter 5, we have shown that contextual information is essential to the good performance of our segmentation methods. Further ways of incorporating such information should be explored. In particular, the distances that we used to convey the location of other structures could be replaced by richer information such as center lines. We also expect that incorporating shape information would improve the performance of the methods. Models of each part could be included, or in a more parametric way, curvatures of surfaces could be used.

Additionally, our methods have been designed to segment CT images, discarding the PET modality of PET/CT images completely. This modality contains valuable information that could be incorporated in the segmentation methods. Utilising PET data however implies dealing with the problems intrinsic to PET/CT such as misalignments due to respiratory movement, and halos around high uptake areas in PET.

8.2.2 Lesion segmentation in PET/CT images

In this thesis, we have focused on bone lesions. As visceral and lymph node secondary lesions are also frequent in many types of cancer, future work should be directed towards segmenting these.

Moreover, we have worked with ^{68}Ga -PSMA-PET/CT images. ^{68}Ga -PSMA is a very specific tracer for prostate cancer, which facilitates lesion segmentation. For other cancer types however, only tracers with high non-specific uptake are available. This is for example the case

with melanoma, for which ^{18}F -FDG is often used. For such images, new methods should be developed to differentiate true lesions from non-specific uptake.

8.2.3 Quantitative analysis of PET/CT images

Quantitative biomarkers are of prime importance for patient management. These biomarkers can be for example concentrations of proteins in the blood, or can be calculated from images like the BSI. Therapy decisions are based on these biomarkers. In chapter 6, we have developed new indices to quantify the bone tumour load and applied them to ^{68}Ga -PSMA-PET/CT images. Prospective studies of these indices should be conducted to relate them to patient survival. They should also be extended to take into account non-osseous lesions and evaluated for other types of cancer.

In chapter 7 we have shown that heterogeneous responses to therapy are frequent. Global methods such as PERCIST, BPI_{VOL} and BPI_{SUV} can not detect such mixed responses. Localised quantification should therefore be performed along global quantification to recognise heterogeneous responses. The clinical meaning of mixed responses should be explored in further studies. The question of possible (micro)biological causes to different therapy responses of lesions in a single patients should also be investigated.

Finally, exploring the indices that we have developed with very large patient cohorts (hundreds to thousands of patients) including patients at different stages of the disease could lead to new insights into the disease evolution and improved patient management.

Part IV

Appendices

Appendix A

Mathematical notations

Here, we detail the mathematical notations used in the rest of the thesis.

A.1 Symbols

\in	element of
\forall	for all
\subset	subset of
\cap	intersection
$ E $	number of elements in E
$*$	convolution operator
α	proportional to

Table A.1 – Meaning of different mathematical symbols

A.2 Number domains

\mathbb{R} denotes the domain of real numbers. \mathbb{N} denotes the domain of positive integer numbers. Intervals in \mathbb{N} are denoted as $[[,]]$ i.e. $[[a, b]] = [a, b] \cap \mathbb{N}$.

$A_1 \times \dots \times A_k$ denotes the Cartesian product of $A_1 \dots A_k$, i.e. an element x of $A_1 \times \dots \times A_k$ is a vector $x = (a_1, \dots, a_k)$ where $a_1 \in A_1, \dots, a_k \in A_k$. By extension, A^k denotes the cartesian product $A \times \dots \times A$ where A is present k times in the equation.

A.3 Functions

A function is a relation defined over an input and an output domain, that associates with each element of the input domain an single element of the output domain. A function f is denoted as follows:

$$\begin{aligned} f : A &\rightarrow B \\ x &\mapsto f(x) \end{aligned} \tag{A.1}$$

where A is the input domain, B the output domain, also called codomain, and $f(x)$ is the value associated with x , for each element x in A . Note that A and B are part of the definition of f .

A simple example is the square function:

$$\begin{aligned} f: \mathbb{R} &\rightarrow \mathbb{R} \\ x &\mapsto x^2 \end{aligned} \tag{A.2}$$

Appendix B

List of publications

B.1 Journal publications

- **M. Bieth**, M. Krönke, R. Tauber, M. Dahlbender, M. Retz, S.G. Nekolla, B. Menze, T. Maurer, M. Eiber, M. Schwaiger. Exploring New Multimodal Quantitative Imaging Indices for the Assessment of Osseous Tumour Burden in Prostate Cancer using ^{68}Ga -PSMA-PET/CT. *Journal of Nuclear Medicine*, 2017.
- **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. *IEEE Transactions on Medical Imaging*, 2017.

B.2 Conference proceedings

- **M. Bieth**, R. Donner, G. Langs, M. Schwaiger, B. Menze. Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images. *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. (oral presentation)
- **M. Bieth**, E. Alberts, M. Schwaiger, B. Menze. From Large to Small Organ Segmentation in CT using Regional Context. *Proceedings of the international Workshop on Machine Learning in Medical Imaging (MLMI) at MICCAI*, 2017.
- E. Alberts, G. Tetteh, S. Trebeschi, **M. Bieth**, A. Valentinitich, B. Wiestler, C. Zimmer, B. Menze. Multi-modal image classification using low-dimensional texture features for genomic brain tumor recognition. *Proceedings of the MICCAI Workshop on Image Genetics (MICGen)*, 2017.

B.3 Conference abstracts

- I. Somlai-Schwaiger, **M. Bieth**, S. Reder, S.G. Nekolla, S.I. Ziegler, M. Schwaiger. Proof of Principle of a Custom-Made γ -Ray Coincidence Counter for Analysis of Cardiac PET

- Tracer Kinetics on Isolated Perfused Rat Hearts. *European Molecular Imaging Meeting*, 2016.
- M. Krönke, **M. Bieth**, R. Tauber, M. Retz, B. Menze, J. Gschwend, T. Maurer, M. Schwaiger, M. Eiber. PSMA PET Index-Introduction of a new imaging quantitative biomarker for assessment of osseous tumor burden in patients with metastatic castration-resistant prostate cancer. *Annual Congress of the European Association of Nuclear Medicine*, 2016.
 - **M. Bieth**, M. Krönke, T. Maurer, R. Tauber, M. Dahlbender, M. Retz, J. Gschwend, S.G. Nekolla, B. Menze, M. Eiber, M. Schwaiger. Introducing PSMA-Bone-PET-Index for quantitative assessment of osseous tumor burden in prostate cancer. *Annual European Association of Urology Congress*, 2017.
 - **M. Bieth**, M. Krönke, R. Tauber, M. Dahlbender, M. Retz, S.G. Nekolla, B. Menze, T. Maurer, M. Eiber, M. Schwaiger. BPI: multimodal quantitative imaging indices for assessment of osseous tumour burden in Ga-68-PSMA-PET/CT. *Gemeinsame Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaften für Nuklearmedizin*, 2017. (oral presentation)

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
2.1 CT principles and example	7
2.2 ^{11}C decay through positron emission	8
2.3 PET principles and example	9
2.4 Bone scintigraphy principles and example	10
3.1 Image segmentation example	16
3.2 Filter based feature examples	19
3.3 Histogram of oriented gradients	20
3.4 Histogram of oriented gradients	20
3.5 Directed rooted tree	21
3.6 Decision tree	21
3.7 Effect of randomness in the training procedure	25
3.8 Architecture of the U-net	28
4.1 Details of the cascaded random forest	37
4.2 Illustration of Haar-like features	39
4.3 Illustration of the anatomical trilateration features	39
4.4 Hierarchical model	42
4.5 Examples from the three datasets	44
4.6 Average computing time	46
4.7 Results for the HS dataset	49
4.8 Results for the PC dataset	50
4.9 Results for the MM dataset	51
4.10 2D view of skeleton and organ annotation	52
4.11 3D view of skeleton annotation	53
4.12 Organ segmentation example	54
5.1 Illustration of different types of features	61
5.2 DS and TPR at different steps	62
5.3 Organ segmentation examples	64
6.1 Bone and lesion segmentation example	71

6.2	BPI _{VOL} for different thresholds	74
6.3	BPI _{SUV} with different thresholds	75
6.4	Correlation of BPI _{VOL} and BPI _{SUV}	77
6.5	Change in BPI _{SUV} and BPI _{VOL} during therapy	78
6.6	Correlation of BPI _{VOL} and BSI	79
6.7	Unclear disease evolution example	80
7.1	Example of bone and lesion segmentation	85
7.2	Clinical case 1	87
7.3	Clinical case 2	88
7.4	Lesions location	89
7.5	LBPI _{VOL} distribution	90
7.6	Evolution of LBPI _{VOL} and LBPI _{SUV}	91

LIST OF TABLES

<u>Table</u>	<u>page</u>
4.1 Mean DS for different methods	46
4.2 Mean DS for variations on our method	47
4.3 Mean DS for different iterations	48
4.4 Normalized feature importance for the prostate cancer dataset	48
4.5 Mean DS with initial landmarks	49
4.6 Mean DS in transfer experiment	50
4.7 Mean DS over subjects for organ annotation	55
5.1 Comparison with other methods	63
6.1 Data for 60 patients	72
6.2 Comparison of PERCIST to BPI and BSI	77
6.3 Comparison of bone-PERCIST to BPI and BSI	78
7.1 Data for forty-five patients	86
A.1 Meaning of different mathematical symbols	101

REFERENCES

- [1] A. Afshar-Oromieh, E. Avtzi, F. Giesel, et al. The diagnostic value of PET/CT imaging with the ^{68}Ga -labelled PSMA ligand HBED-CC in the diagnosis of recurrent prostate cancer. *European journal of nuclear medicine and molecular imaging*, 42(2):197–209, 2015.
- [2] A. Afshar-Oromieh, A. Malcher, M. Eder, et al. PET imaging with a ^{68}Ga gallium-labelled PSMA ligand for the diagnosis of prostate cancer: biodistribution in humans and first evaluation of tumour lesions. *European journal of nuclear medicine and molecular imaging*, 40(4):486–495, 2013.
- [3] A. Afshar-Oromieh, C. Zechmann, A. Malcher, et al. Comparison of PET imaging with a ^{68}Ga -labelled PSMA ligand and ^{18}F -choline-based PET/CT for the diagnosis of recurrent prostate cancer. *European journal of nuclear medicine and molecular imaging*, 41(1):11–20, 2014.
- [4] P. Aljabar, R. Heckemann, A. Hammers, et al. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.
- [5] A. Anand, M. Morris, R. Kaboteh, et al. Analytic validation of the automated bone scan index as an imaging biomarker to standardize quantitative changes in bone scans of patients with metastatic prostate cancer. *Journal of Nuclear Medicine*, 57(1):41–45, 2016.
- [6] B. Andres, T. Beier, and J. Kappes. OpenGM: A C++ library for discrete graphical models. *ArXiv e-prints*, 2012.
- [7] H. Anger. Scintillation camera. *Review of scientific instruments*, 29(1):27–33, 1958.
- [8] A. BenTaieb and G. Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 460–468. Springer, 2016.
- [9] T. Beyer, D. Townsend, T. Brun, et al. A combined PET/CT scanner for clinical oncology. *Journal of nuclear medicine*, 41(8):1369–1379, 2000.
- [10] M. Bieth. *Master thesis: Kinetic analysis and inter-subject registration of brain PET images*. 2013.
- [11] M. Bieth, E. Alberts, M. Schwaiger, et al. From large to small organ segmentation in CT using regional context. In *International Workshop on Machine Learning in Medical Imaging*, pages 1–9. Springer, 2017.
- [12] M. Bieth, R. Donner, G. Langs, et al. Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 84.1–84.10, 2015.
- [13] M. Bieth, M. Krönke, R. Tauber, et al. Exploring new multimodal quantitative imaging indices for the assessment of osseous tumour burden in prostate cancer using ^{68}Ga -PSMA-PET/CT. *Journal of Nuclear Medicine*, pages jnumed–116, 2017.
- [14] M. Bieth, L. Peter, S. Nekolla, et al. Segmentation of skeleton and organs in whole-body CT images via iterative trilateration. *IEEE Transactions on Medical Imaging*, 36(11):2276–2286, 2017.
- [15] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.
- [16] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [17] L. Breiman, J. Friedman, C. Stone, et al. *Classification and regression trees*. CRC press, 1984.

- [18] L. Bubendorf, A. Schöpfer, U. Wagner, et al. Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients. *Human pathology*, 31(5):578–583, 2000.
- [19] M. Calonder, V. Lepetit, M. Ozuysal, et al. BRIEF: Computing a Local Binary Descriptor Very Fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, 2012.
- [20] A. cancer society. Cancer facts and figures 2016. *Atlanta. American Cancer Society*, 2016.
- [21] J. Cerrolaza, R. Summers, and M. Linguraru. Soft multi-organ shape models via generalized PCA: A general framework. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 219–228. Springer, 2016.
- [22] S. Cherry and M. Dahlbom. PET: physics, instrumentation, and scanners. In *PET*, pages 1–117. Springer, 2006.
- [23] H. Chung, H. Kwon, K. Kang, et al. Prognostic value of preoperative metabolic tumor volume and total lesion glycolysis in patients with epithelial ovarian cancer. *Annals of surgical oncology*, 19(6):1966–1972, 2012.
- [24] A. Criminisi, D. Robertson, E. Konukoglu, et al. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical image analysis*, 17(8):1293–1303, 2013.
- [25] A. Criminisi and J. Shotton. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.
- [26] A. Criminisi, J. Shotton, and S. Bucciarelli. Decision forests with long-range spatial context for organ localization in CT volumes. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–80. Cite-seer, 2009.
- [27] A. Criminisi, J. Shotton, D. Robertson, et al. Regression forests for efficient anatomy detection and localization in CT studies. In *Proceedings of the International MICCAI Workshop on Medical Computer Vision*, pages 106–117. Springer, 2010.
- [28] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.
- [29] O. del Toro and H. Müller. Hierarchical multi-structure segmentation guided by anatomical correlations. In *Proceedings of the VISCERAL Challenge at ISBI*, pages 32–36. Citeseer, 2014.
- [30] R. Donner, B. Menze, H. Bischof, et al. Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical Image Analysis*, 17(8):1304–1314, 2013.
- [31] J. Ehrhardt, H. Handels, T. Malina, et al. Atlas-based segmentation of bone structures to support the virtual planning of hip operations. *International Journal of Medical Informatics*, 64(2):439–447, 2001.
- [32] M. Eiber, T. Maurer, M. Souvatzoglou, et al. Evaluation of hybrid 68Ga-PSMA ligand PET/CT in 248 patients with biochemical recurrence after radical prostatectomy. *Journal of nuclear medicine*, 56(5):668–674, 2015.
- [33] E. Eisenhauer, P. Therasse, J. Bogaerts, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

- [34] E. Etchebehere, J. Araujo, P. Fox, et al. 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*, 56(8):1177–1184, 2015.
- [35] M. Everingham, L. Van Gool, C. Williams, et al. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [36] B. Fulkerson, A. Vedaldi, S. Soatto, et al. Class segmentation and object localization with superpixel neighborhoods. In *Proceedings of the International Conference on Computer Vision (ICCV)*, volume 9, pages 670–677. Citeseer, 2009.
- [37] Y. Gao, Y. Shao, J. Lian, et al. Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests. *IEEE transactions on medical imaging*, 35(6):1532–1543, 2016.
- [38] R. Gauriau, R. Cuingnet, D. L., et al. Multi-organ localization with cascaded global-to-local regression and shape prior. *Medical image analysis*, 23(1):70–83, 2015.
- [39] E. Geremia, O. Clatz, B. Menze, et al. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.
- [40] B. Glocker, J. Feulner, A. Criminisi, et al. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 590–598. Springer, 2012.
- [41] B. Glocker, O. Pauly, E. Konukoglu, et al. Joint classification-regression forests for spatially structured multi-object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 870–881. Springer, 2012.
- [42] B. Glocker, D. Zikic, E. Konukoglu, et al. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 262–270. Springer, 2013.
- [43] O. Göksel, O. Jiménez-del Toro, A. Foncubierta-Rodríguez, et al. 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.
- [44] K. Gruenberg, M. Weber, O. Jiménez del Toro, et al. Visceral-visual concept extraction challenge in radiology: Segmentation challenge: overview, insights and preliminary results. In *European Congress of Radiology (ECR) 2015*, Vienna, Austria, 2015.
- [45] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.
- [46] M. Heinrich and M. Blendsowski. Multi-organ segmentation using vantage point forests and binary context features. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 598–606. Springer, 2016.
- [47] M. Heinrich and M. Blendsowski. Multi-organ segmentation using vantage point forests and binary context features. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 598–606. Springer, 2016.
- [48] G. Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media, 2009.
- [49] P. Hu, F. Wu, J. Peng, et al. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–13, 2016.

- [50] S. Huang, Y. Chu, S. Lai, et al. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE Transactions on Medical Imaging*, 28(10):1595–1605, 2009.
- [51] J. Hyun, M. Lodge, and R. Wahl. Practical PERCIST: A simplified guide to PET response criteria in solid tumors 1.0. *Radiology*, 2016.
- [52] J. Iglesias and M. Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.
- [53] M. Imbriaco, S. Larson, H. Yeung, et al. A new parameter for measuring metastatic bone involvement by prostate cancer: the bone scan index. *Clinical Cancer Research*, 4(7):1765–1772, 1998.
- [54] F. Kahl, J. Alvé, O. Enqvist, et al. Good features for reliable registration in multi-atlas segmentation. *Proceedings of the VISCERAL Challenge at ISBI*, 1390:12–17, 2015.
- [55] Y. Kang, K. Engelke, and W. Kalender. A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. *IEEE transactions on medical imaging*, 22(5):586–598, 2003.
- [56] J. Kappes, B. Andres, F. Hamprecht, et al. A comparative study of modern inference techniques for discrete energy minimization problems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1335. IEEE, 2013.
- [57] T. Klinder, J. Ostermann, M. Ehm, et al. Automated model-based vertebra detection, identification, and segmentation in CT images. *Medical Image Analysis*, 13(3):471–482, 2009.
- [58] T. Kohlberger, M. Sofka, J. Zhang, et al. Automatic multi-organ segmentation using learning-based segmentation and level set optimization. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 338–345. Springer, 2011.
- [59] P. Kotschieder, P. Kohli, J. Shotton, et al. Geof: Geodesic forests for learning coupled predictors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 65–72, 2013.
- [60] E. Konukoglu, B. Glocker, D. Zikic, et al. Neighbourhood approximation using randomized forests. *Medical image analysis*, 17(7):790–804, 2013.
- [61] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, 2012.
- [62] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [63] T. La, E. Filion, B. Turnbull, et al. Metabolic tumor volume predicts for recurrence and death in head-and-neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, 74(5):1335–1341, 2009.
- [64] 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)*, volume 1, pages 282–289, 2001.
- [65] Y. LeCun, B. Boser, J. Denker, et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

- [66] J. Lee, C. Kang, H. Choi, et al. Prognostic value of metabolic tumor volume and total lesion glycolysis on preoperative 18F-FDG PET/CT in patients with pancreatic cancer. *Journal of Nuclear Medicine*, 55(6):898–904, 2014.
- [67] P. Lee, D. Weerasuriya, P. Lavori, et al. Metabolic tumor burden predicts for disease progression and death in lung cancer. *International Journal of Radiation Oncology, Biology, Physics*, 69(2):328–333, 2007.
- [68] Y. Li, C. Ho, N. Chahal, et al. Myocardial segmentation of contrast echocardiograms using random forests guided by shape model. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 158–165. Springer, 2016.
- [69] Y. Li, C. Pahng Ho, N. Chahal, et al. Myocardial segmentation of contrast echocardiograms using random forests guided by shape model. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 158–165. Springer, 2016.
- [70] M. Linguraru, J. Pura, V. Pamulapati, et al. Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT. *Medical image analysis*, 16(4):904–914, 2012.
- [71] H. Lombaert, D. Zikic, A. Criminisi, et al. Laplacian forests: semantic image segmentation by guided bagging. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 496–504. Springer, 2014.
- [72] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.
- [73] J. Lötjönen, R. Wolz, J. Koikkalainen, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*, 49(3):2352–2365, 2010.
- [74] B. Lowekamp, D. Chen, L. Ibáñez, et al. The design of SimpleITK. *Frontiers in neuroinformatics*, 7:45, 2013.
- [75] R. Meier, S. Bauer, J. Slotboom, et al. A hybrid model for multimodal brain tumor segmentation. *Multimodal Brain Tumor Segmentation*, 31, 2013.
- [76] B. Menze, B. Kelm, D. Splitthoff, et al. On oblique random forests. *Machine Learning and Knowledge Discovery in Databases*, pages 453–469, 2011.
- [77] C. Messiou, G. Cook, and N. deSouza. Imaging metastatic bone disease from carcinoma of the prostate. *British journal of cancer*, 101(8):1225–1232, 2009.
- [78] A. Montillo, J. Tu, J. Shotton, et al. Entanglement and differentiable information gain maximization. In *Decision Forests for Computer Vision and Medical Image Analysis*, pages 273–293. Springer, 2013.
- [79] V. Müller, M. de Wit, K. Bohuslavizki, et al. Bone scintigraphy in clinical routine. *Radiology and Oncology*, 35(1), 2001.
- [80] T. Okada, M. Linguraru, M. Hori, et al. Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors. *Medical Image Analysis*, 26(1):1–18, 2015.
- [81] A. Oppelt. *Imaging systems for medical diagnostics*. Publicis MCD, 2005.
- [82] H. Park, P. Bland, and C. Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *IEEE Transactions on medical imaging*, 22(4):483–492, 2003.

- [83] L. Peter, O. Pauly, P. Chatelain, et al. Scale-adaptive forest training via an efficient feature sampling scheme. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 637–644. Springer, 2015.
- [84] T. Pyka, S. Okamoto, M. Dahlbender, et al. Comparison of bone scintigraphy and ⁶⁸Ga-PSMA PET for skeletal staging in prostate cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–8, 2016.
- [85] J. Radon. Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5, 2005.
- [86] T. Rainforth and F. Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.
- [87] D. Richmond, D. Kainmueller, B. Glocker, et al. Uncertainty-driven forest predictors for vertebra localization and segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 653–660. Springer, 2015.
- [88] T. Rohlfing, R. Brandt, R. Menzel, et al. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004.
- [89] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, 2010.
- [90] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.
- [91] H. Roth, L. Lu, A. Farag, et al. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 556–564. Springer, 2015.
- [92] O. Russakovsky, J. Deng, H. Su, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [93] P. Sabbatini, S. Larson, A. Kremer, et al. Prognostic significance of extent of disease in bone in patients with androgen-independent prostate cancer. *Journal of clinical oncology*, 17(3):948–948, 1999.
- [94] 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*, 28:46–65, 2016.
- [95] S. Schmidt, J. Kappes, M. Bergtholdt, et al. Spine detection and labeling using a parts-based graphical model. In *Information Processing in Medical Imaging*, pages 122–133. Springer, 2007.
- [96] F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–10, 2008.
- [97] J. Shotton, J. Winn, C. Rother, et al. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–15. Springer, 2006.
- [98] R. Smith-Bindman, J. Lipson, R. Marcus, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of internal medicine*, 169(22):2078–2086, 2009.

- [99] W. Snyder, M. Cook, E. Nasset, et al. Report of the task group on reference man, international commission on radiological protection no. 23, 1975.
- [100] Z. Tian, L. Liu, Z. Zhang, et al. Superpixel-based segmentation for 3D prostate MR images. *IEEE Transactions on Medical Imaging*, 35(3):791–801, 2016.
- [101] S. Tomoshige, E. Oost, A. Shimizu, et al. A conditional statistical shape model with integrated error estimation of the conditions; application to liver segmentation in non-contrast CT images. *Medical image analysis*, 18(1):130–143, 2014.
- [102] T. Tong, R. Wolz, Z. Wang, et al. Discriminative dictionary learning for abdominal multi-organ segmentation. *Medical Image Analysis*, 23(1):92–104, 2015.
- [103] Z. Tu. Auto-context and its application to high-level vision tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.
- [104] D. Ulmert, R. Kaboteh, J. Fox, et al. A novel automated platform for quantifying the extent of skeletal tumour involvement in prostate cancer patients using the bone scan index. *European urology*, 62(1):78–84, 2012.
- [105] D. Ulmert, L. Solnes, and D. Thorek. Contemporary approaches for imaging skeletal metastasis. *Bone research*, 3:15024, 2015.
- [106] J. Valentin. Basic anatomical and physiological data for use in radiological protection: reference values: International commission on radiological protection no. 89. *Annals of the ICRP*, 32(3):1–277, 2002.
- [107] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511. IEEE, 2001.
- [108] C. Wachinger, K. Fritscher, G. Sharp, et al. Contour-driven atlas-based segmentation. *IEEE Transactions on Medical Imaging*, 34(12):2492–2505, 2015.
- [109] R. Wahl, H. Jacene, Y. Kasamon, et al. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *Journal of nuclear medicine*, 50(Suppl 1):122S–150S, 2009.
- [110] Q. Wang, D. Wu, L. Lu, et al. Semantic context forests for learning-based knee cartilage segmentation in 3D MR images. In *Medical Computer Vision. Large Data in Medical Imaging*, pages 105–115. Springer, 2014.
- [111] R. Wolz, C. Chu, K. Misawa, et al. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Transactions on Medical Imaging*, 32(9):1723–1730, 2013.
- [112] D. Wu, D. Liu, Z. Puskas, et al. A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 980–987. IEEE, 2012.
- [113] M. Yadav, S. Ballal, M. Tripathi, et al. ¹⁷⁷Lu-DKFZ-PSMA-617 therapy in metastatic castration resistant prostate cancer: safety, efficacy, and quality of life assessment. *European journal of nuclear medicine and molecular imaging*, 44(1):81–91, 2017.
- [114] P. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *SODA*, volume 93, pages 311–21, 1993.

- [115] D. Zikic, B. Glocker, and A. Criminisi. Atlas encoding by randomized forests for efficient label propagation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 66–73. Springer, 2013.
- [116] V. Zografos, A. Valentinitich, M. Rempfler, et al. Hierarchical multi-organ segmentation without registration in 3D abdominal ct images. In *Proceedings of the International MICCAI Workshop on Medical Computer Vision*, 2015.
- [117] Natürliche Strahlenbelastung. http://www.bfs.de/DE/themen/ion/umwelt/natuerliche-strahlenbelastung/natuerliche-strahlenbelastung_node.html. accessed: 10/10/2016.