N° d'ordre NNT : 2021LYSES026

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé (CREATIS)

Ecole Doctorale N° 488
Sciences Ingénierie Santé (SIS)

Spécialité de doctorat : Traitement d'images médicales
Discipline : Sciences et Technologies de l'information et de la communication

Soutenue publiquement le 15/10/2021, par :
Hoai-Thu Nguyen

Contributions aux approches multi-atlas et d'apprentissage profond pour la segmentation des muscles : application à l'étude longitudinale multi-paramétrique quantitative en IRM

Contributions to multi-atlas and deep learning approaches for muscle segmentation in multi-parametric quantitative MRI longitudinal studies

---

Devant le jury composé de :

Petitjean, Caroline	Professeur	Université de Rouen	Présidente
Garreau, Mireille	Professeur	Université de Rennes 1	Rapporteure
Edouard, Pascal	Professeur	Université Saint-Etienne	Examinateur
Viallon, Magalie	Docteur	CHU Saint-Etienne	Examinatrice
Wang, Hongzhi	Docteur	IBM	Examinateur
Croisille, Pierre	Professeur	CHU Saint-Etienne	Directeur de thèse
Grenier, Thomas	Maître de conférences	INSA de Lyon	Co-directeur de thèse

"I am the wisest man alive, for I know one thing, and that is that I know nothing."

Plato

Acknowledgements

À Thomas, dont la suggestion de stage a orienté ma future carrière dans une nouvelle direction. Ces cinq années de travail avec toi m'ont apporté le plus de joie, et j'ai beaucoup appris tant scientitifiquement qu'humainement. Tu es un mentor et un ami qui m'a toujours donné une immense confiance et un soutien qui m'ont permis d'arriver jusque là.

À Pierre, qui a dirigé et surveillé ce projet, merci pour l'opportunité de travailler sur ce projet unique et pour avoir apporté le côté interdisciplinaire dans ce travail, qui a donné vie à nos algorithmes.

À Magalie, dont le flot incessant d'idées a toujours apporté de l'enthousiasme à ce travail.

À Sylvain, Rémi et Malick pour avoir travaillé avec moi pendant ce projet et pour m'avoir aidé à terminer ce travail.

À Sarah, Yunyun, Fei, Anchen et PA pour avoir survécu à mes sarcasmes et pour m'avoir patiemment écouté me plaindre quand rien ne se passait comme je le voulais.

À tous les membres du laboratoire CREATIS pour avoir créé un environnement de travail agréable, merci pour toutes les discussions enrichissante et pour votre aide, qu'elle soit scientifique, technique ou administrative.

To my family for their unwavering confidence in me. To my parents, who have always been blindly supporting all of my decisions, who have left me space to learn and grow by myself, who have to remind themselves constantly that their girl is not stubborn, just willful. To my brother, Tien Anh, who have been there for me, both emotionally and physically, through the lowest point of my life. I can never ask for a better brother and a better friend. Thank you for listening to me so well and do not forget, I am always right!

To Hien, my best friend and my soulmate for the last 13 years. In Vietnam, in France, or with 10 000km distance, you have always been the core of my support system. Crazy young girls or crazy old ladies, I know we will always be crazy together.

To all of my friends, most live far away, for always been there when I need a good talk. Thank you for listening to me ranting about things you do not understand and do not have any interest in. Thank you for going drinking with me when some of you do not even drink. Thank you for giving me the view of a diverse society, of all the good things we could do in life outside my small world full of mathematical formulas and thousands of lines of code.

Finally, to all the people who were, at some point, part of my life, for each teaching me a lesson about myself and about life. Without you, I would not be the person I am today.

I would like to express my deepest gratitude to Caroline Petitjean and Mireille Garreau for agreeing to participate in the evaluation of my thesis work as reporters. It is an honor to be able to benefit from your experience and your expertise. I would also like to thank all the other members of the jury for their time and for sharing their insights.

UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Résumé

École Doctorale 488 Sciences Ingénierie Santé

Laboratoire CREATIS

Contributions aux approches multi-atlas et d'apprentissage profond pour la segmentation des muscles : application à l'étude longitudinale multi-paramétrique quantitative en IRM

par Hoai-Thu NGUYEN

Parmi les techniques d'imagerie, l'imagerie par résonance magnétique (IRM) est l'une des techniques d'imagerie les plus versatiles, capable de produire des données morphologiques mais également quantitatives et fonctionnelles. De par son caractère non-irradiant, elle permet en particulier les études temporelles et longitudinales. L'exploitation optimale de toutes ces potentialités requiert néanmoins la construction et l'optimisation de séquences de processus automatiques de traitements et d'analyses des images. L'étage de segmentation des images est une étape fondamentale garantissant la précision de l'exploitation des données, en permettant la séparation des structures anatomiques en différentes régions. La production massive des données image disponibles et la multiplication et complexité des structures anatomiques à contourer, rend caduque une segmentation manuelle en raison de la durée requise et la démotivation au regard de la tâche. Dès lors, il est nécessaire de mettre en place des méthodes de segmentation automatique adaptées.

Ce travail de thèse s'est nourri des problématiques du projet de recherche MUST dont l'objectif a été d'aider à comprendre l'effet de l'ultra-endurance sur l'organisme et son impact au niveau notamment musculaire. L'équipe de chercheurs du laboratoire CREATIS a choisi de mener cette étude sur l'Ultramarathon de Montagne longtemps considérée comme l'épreuve la plus extrême au monde : le Tor des Géants. Cette étude a notamment utilisé une IRM mobile dotées de techniques d'imagerie avancées pour étudier les variations longitudinales survenant lors d'un effort supra-physiologique. Afin d'étudier l'évolution de l'inflammation et de paramètres fonctionnels musculaires au niveau des quadriceps, notre thèse étudie et améliore les méthodes de segmentation automatique basées sur des approches supervisées. Notre objectif est de fournir une méthode cliniquement applicable permettant de segmenter les chefs musculaires des quadriceps aussi précisément que possible en longitudinal, et nécessitant le moins possible de segmentations manuelles pour la phase d'apprentissage.

A cet effet, nous explorons et appliquons dans un premier temps les approches multi-atlas pour la segmentation des quadriceps. Nos contributions permettent d'obtenir des segmentations de qualité sur une grande partie de la base de données. Cependant, afin de disposer d'une approche plus rapide et plus robuste, nous avons secondairement orienté nos travaux vers les approches de deep learning. Dans ce contexte méthodologique, nos deux contributions principales sont i) la proposition d'une étape de correction des segmentations basée sur un apprentissage machine, et ii) la proposition de stratégies d'augmentation de données pour optimiser l'apprentissage de réseaux de type U-Net, notamment une stratégie basée sur la ressemblance morphologique qui est évaluée grâce à une mesure originale. Cette mesure de la morphologie s'est aussi révélée très efficace pour sélectionner les atlas pour l'approche de segmentation multi-atlas. Enfin, nous montrons que les approches proposées se généralisent à d'autres problématiques de segmentation musculaires en IRM et permettent aussi des études statistiques longitudinales et localisées.

Notre travail met en évidence que, même si la quantité de données annotées est essentielle dans l'apprentissage supervisé, nous devons également prêter attention à la diversité morphologique de notre base de données, ce qui permet de réduire le temps de calcul et d'augmenter la précision ainsi que la robustesse des méthodes de segmentation.

UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Abstract

École Doctorale 488 Sciences Ingénierie Santé

Laboratoire CREATIS

Contributions to multi-atlas and deep learning approaches for muscle segmentation in multi-parametric quantitative MRI longitudinal studies

by Hoai-Thu NGUYEN

Among imaging techniques, magnetic resonance imaging (MRI) is one of the most versatile, capable of producing not only morphological but also quantitative and functional data. Due to its non-irradiative nature, it allows in particular temporal and longitudinal studies. The optimal exploitation of all these potentialities nevertheless requires constructing and optimizing an automatic framework for image processing and analysis. The image segmentation is a fundamental step providing the separation of anatomical structures into individual regions. The massive production of the available image data and the multiplication and complexity of the anatomical structures to be by-passed makes manual segmentation obsolete due to the time required and the lack of motivation concerning the task. It is, therefore, necessary to implement suitable automatic segmentation methods.

This thesis work was based on the questions of the MUST research project, whose objective was to help understand the effect of ultra-endurance on the body and its impact on the muscular level in particular. The researchers at the CREATIS laboratory chose to conduct this study on the mountain ultra-marathon, long considered the most extreme event in the world: the Tor des Géants. This study notably used a mobile MRI equipped with advanced imaging techniques to study the longitudinal variations during a supra-physiological effort. In order to study the evolution of inflammation and muscle functional parameters at the quadriceps level, our thesis studies and improves automatic segmentation methods based on supervised approaches. Our objective is to provide a clinically applicable method to segment the quadriceps muscle heads as precisely as possible in longitudinal and requiring as little manual segmentation as possible for the learning phase.

With this objective, we first explore and apply multi-atlas approaches for quadriceps segmentation. Our contributions allow us to obtain quality segmentations on a large part of our dataset. However, in order to have a faster and more robust approach, we then oriented our work towards deep learning approaches. In this methodological context, our two main contributions are i) the proposal of a segmentation correction step based on machine learning and ii) the proposal of data augmentation strategies to optimize the learning of U-Net type networks, in particular, a strategy based on morphological resemblance evaluated thanks to an original measurement. This morphology measurement also proved to be very efficient for selecting atlases for the multi-atlas segmentation approach. Finally, we show that the proposed approaches can be generalized to other muscle segmentation problems in MRI and allow longitudinal and localized statistical studies.

Our work shows that, even if the quantity of annotated data is crucial in supervised learning, we must also pay attention to our database's morphological diversity, which eventually reduces computation time and increases the precision and robustness of the segmentation methods.

Contents

List of Figures

List of Tables

Introduction

This introduction is written in french as requested by the doctoral school EDSIS. It serves as a summary of the dissertation and does not include any information that would not be detailed later.

L'imagerie médicale permet, par l'image, l'observation des caractéristiques internes d'un corps à des fins d'analyse clinique et d'interventions médicales.

Ce domaine connaît un développement rapide qui se traduit par une amélioration de la qualité des images ainsi que de la quantité des caractéristiques observées. De plus, sa démocratisation conduit à une large disponibilité des données d'images médicales et l'observation de quasi toutes les pathologies. Cependant, il reste primordial d'être capable d'extraire l'information utile pour l'analyse médicale ciblée. Devant cet afflux de données, ces traitements permettant l'extraction se doivent d'être les plus automatiques possibles, robustes et en accord avec les besoins des médecins afin d'améliorer leur efficacité sur l'analyse médicale.

Dans le cadre de ce travail de thèse à CREATIS, nous contribuons à l'étude de l'évolution fonctionnelle des muscles squelettiques. Cette étude s'appuie sur les données issues du projet MUST qui mène une étude longitudinale sans précédent sur l'effort supra-physiologique des athlètes de l'ultra-marathon de montagne en utilisant notamment des techniques avancées d'imagerie par résonance magnétique (IRM).

L'ultra-marathon connaît une popularité croissante depuis quelques années et de nombreuses manifestations sont organisées chaque année. La course d'ultra-endurance en montagne sur plusieurs jours (MUM) est le format le plus intense de ce sport. Elle met les athlètes dans de nombreuses conditions extrêmes et les pousse à leurs limites. Cependant, les effets de ces conditions sur le corps humain restent pour la plupart inexplorés. Les développements récents et innovants en matière d'IRM quantitative permettent une exploration approfondie et non invasive des altérations fonctionnelles des muscles squelettiques. Pour le projet de thèse, l'un des objectifs est de permettre de quantifier la principale réponse inflammatoire des muscles dans les conditions extrêmes de MUM qui, chez certains sujets, correspond étroitement à celle observée chez les patients en unité de soins intensifs après un polytraumatisme et/ou un infarctus du myocarde. Pour les muscles squelettiques, l'imagerie des cuisses réalisées dans le cadre de MUST est parfaitement adaptée à cette étude.

Ainsi, il va falloir traiter automatiquement ces IRM de jambes afin d'extraire des caractéristiques de l'image permettant d'étudier les réponses inflammatoires de chaque chef musculaire dans le temps. Parmi les traitements d'images à appliquer, celui qui est le plus critique ici est la segmentation d'images. Il s'agit de réaliser la délimitation de chaque chef musculaire dans les images IRM 3D. Ce travail, souvent réalisé à la main, est très long et demande aux radiologues experts attention et minutie éprouvantes.

Notre but sera donc de proposer une approche de segmentation automatique des quadriceps. Elle se devra d'être efficace sur l'ensemble de la base de données à analyser et reproductible dans le temps pour permettre le suivi des coureurs. Dans la suite de ce manuscrit, nous verrons aussi que les approches supervisées sont les seules alternatives permettant une segmentation suffisamment précise pour l'ensemble des coureurs. Nous en développerons deux : une méthode de segmentation basée sur du recalage multi-atlas et une méthode de segmentation utilisant l'apprentissage profond. Or, ces approches demandent un grand nombre d'images annotées manuellement (ou atlas) afin d'*apprendre*, automatiquement, la tâche à accomplir.

Nous veillerons à proposer des méthodes permettant de limiter ce nombre d'images annotées.

Pour réaliser cette étude, bien que capitale, la segmentation ne sera pas le seul traitement exploité ici. Il faudra notamment s'appuyer sur des prétraitements des images adaptés aux problématiques de l'IRM (correction de biais, recalage entre séquences, extraction de chacune des jambes). Il s'agira aussi de guider les experts médicaux dans la réalisation des annotations manuelles pour qu'elles soient justes suffisantes pour obtenir une segmentation manuelle de qualité en 3D. Puis, après la segmentation, il faudra appliquer plusieurs méthodes de post-traitements de manière séquentielle pour d'une part corriger automatiquement les petites erreurs de segmentation puis surtout extraire des données d'IRM des indices quantitatifs qui seront ensuite corrélés longitudinalement et avec d'autres marqueurs. Cependant, pour des raisons d'efficacité, ce manuscrit de thèse ne détaillera pas profondément ces approches. Pour obtenir ces éléments, le lecteur pourra lire nos contributions Nguyen et al. (2019b) et Nguyen et al. (2021a).

Dans cette thèse nous nous focalisons donc sur nos contributions aux méthodes de segmentation de muscles en IRM à partir d'un faible nombre d'atlas et permettant une analyse longitudinale de marqueurs issus de l'image et leur corrélation avec des biomarqueurs.

La thèse est organisée en 4 parties avec 11 chapitres.

La première partie dressera le contexte de cette thèse, en commençant par préciser comment l'IRM permet la quantification fonctionnelle et par rappeler l'importance de la segmentation d'images. Le second chapitre présentera le projet MUST dans son ensemble puis plus spécifiquement les données que nous traiterons.

La deuxième partie dresse les états de l'art. Pour le chapitre 3, il s'agit des critères de validation pour les méthodes de segmentation d'images, y compris les métriques de validation, la stratégie de validation croisée et l'évaluation du temps de calcul. Le chapitre 4 dresse l'état de l'art des méthodes de segmentation avec peu de données annotées. Enfin, les difficultés de la segmentation de muscles à partir des images IRM et les méthodes existantes pour la segmentation des quadriceps humains sont présentées dans Chapitre 5.

La troisième partie est dédiée à nos contributions à la segmentation des muscles quadriceps à partir d'images IRM. Le chapitre 6 donnera les prétraitements à appliquer spécifiquement aux images IRM et rappellera les donner dont on dispose. Le chapitre 7 présentera notre application et optimisation de la méthode de segmentation multi-atlas avec fusion d'étiquettes et apprentissage correctif sur la base de données MUST. Ensuite, le chapitre 8 présentera notre approche d'apprentissage profond faiblement supervisé et basé sur le réseau UNet comme alternative à la segmentation multi-atlas. Le chapitre 9 présentera notre proposition de descripteurs morphologiques permettant de mieux appréhender une base de données et ainsi de mieux définir les stratégies d'augmentation ou choix de données pour les méthodes de segmentation supervisées.

Enfin, la quatrième partie est consacrée aux applications et discussion de nos contributions. Le chapitre 10 montrera les résultats et améliorations obtenues avec nos approches sur l'ensemble des images MUST ainsi que leurs généralisations et limites de nos approches sur deux autres jeux de données nécessitant une segmentation de muscles. Quant au chapitre 11, il reviendra sur la problématique initiale

d'étude longitudinale de l'inflammation des muscles quadriceps sur la base de données MUST.

Ce manuscrit se termine par une conclusion générale et les perspectives envisageables issues de ces travaux de thèse. Chaque partie comporte un résumé rédigé en français, à la demande de l'école doctorale.

PART I

Context & Research project

Contents

Résumé

Cette partie présente le contexte de ce travail de thèse : l'IRM et le projet MUST

Le chapitre 1 donne une brève introduction à l'imagerie par résonance magnétique (IRM) qui sera la modalité d'imagerie utilisée dans cette thèse. Cette technique d'imagerie est non irradiante et non invasive, et a des applications potentielles dans l'étude des variations fonctionnelles. Pour explorer efficacement ces potentiels, il faut construire et optimiser une série de processus automatiques, tels que la segmentation des images, puis l'extraction des caractéristiques des images et enfin l'analyse ces caractéristiques extraites.

Le chapitre 2 présente le projet MUST, un projet de recherche important conduit par des membres du laboratoire CREATIS, qui vise à étudier l'effet de l'un des ultra-marathons de montagne les plus extrêmes au monde, le Tors des Géants, sur le corps des athlètes. Des prélèvements sanguins, des IRM et des examens par ultrasons ont été effectués sur les participants à différents moments de la course. Dans le cadre de ce projet de doctorat qui se focalise sur l'étude des muscles squelettiques, nous nous sommes intéressés aux images IRM du haut des jambes, en particulier les quadriceps, le groupe de muscles squelettiques le plus affecté par l'effort excentrique lors de la course de descente. Ce chapitre sera aussi le lieu de présentation des quadriceps et du lien entre cette étude MUST et des enjeux cliniques.

Introduction

This part presents the context of this thesis work: the MRI and the MUST project.

Chapter 1 gives a brief introduction of Magnetic Resonance Imaging (MRI) which will be the imaging modality used throughout this thesis. This imaging technique is non-irradiating and non-invasive and has potential applications in functional variation study. To effectively explore these potentials, a sequence of automatic processes, such as image segmentation, extraction of image features and features analysis, needs to be constructed and optimized.

Chapter 2 presents the MUST project, a major research project led by the members of CREATIS laboratory, which aims to study the effect of one of the most extreme mountain ultra-marathons in the world, Tors des Géants, on the corps of athletes. Blood sampling, MRI, and ultrasound examination were performed on the participants at multiple time points. In the context of this Ph.D. project, we are primarily interested in the upper legs' MRI data, specifically the quadriceps, the most affected skeletal muscle due to the eccentric effort during downhill running. This chapter will also present the quadriceps and how their studies can relate to clinical problems.

CHAPTER 1

Medical imaging for longitudinal functional variation studies

Medical imaging aims at providing morphological or parametric maps of internal characteristics of the human body to drive clinical decisions and monitor treatments. Among the most popular medical imaging methods, there are Magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), and computed tomography (CT). These techniques have been experiencing rapid development, which leads to improved quality and wide availability of medical image data. While the expertise of medical doctors is still irreplaceable, image analysis algorithms for medical applications are becoming more and more popular as a tool, helping them perform their tasks more efficiently and possibly with higher quality. Image analysis algorithms may provide automatic detection of anatomic structure or lesions in patients and, while focusing on selected structures, can also provide quantitative analysis at a specific time or at multiple time points during patient follow-up or during a longitudinal study design (e.g., functional evolution of organ or sub-regions structures). These applications motivate one of the most popular sub-domains in image analysis: image segmentation.

In this work, we mainly focus on MRI data analysis. MRI principles are shortly reminded in the next section (Sec. 1.1) with some muscle MR imaging specificities. Section 1.2 discusses the importance of image segmentation.

1.1 MRI and its ability for functional quantification

Magnetic Resonance Imaging (MRI) is a reliable, non-irradiating, and non-invasive imaging technique for tissue characterization and quantitative assessment of tissue integrity through its magnetic properties. An MRI pulse sequence is like an orchestral score describing a series of radiofrequency pulses, gradient manipulation, and signal measurements that result in a set of images with a particular appearance (Fig. 1.1). Each sequence has its own set of operator-selectable parameters that affect tissue contrast and spatial resolution. In general, the image pixel value depends on a host of intrinsic parameters, including the proton density, the T1 and T2 relaxation times, the field heterogeneity, and the physiological motion. The effects of these parameters can be suppressed or enhanced by the external magnetic fields and the operator-selectable parameters, such as repetition time (TR), echo time (TE), and flip

angle (Liang et al., 2000). Therefore, by altering these extrinsic parameters, one can get a quasi-infinitive number of different image contrast.

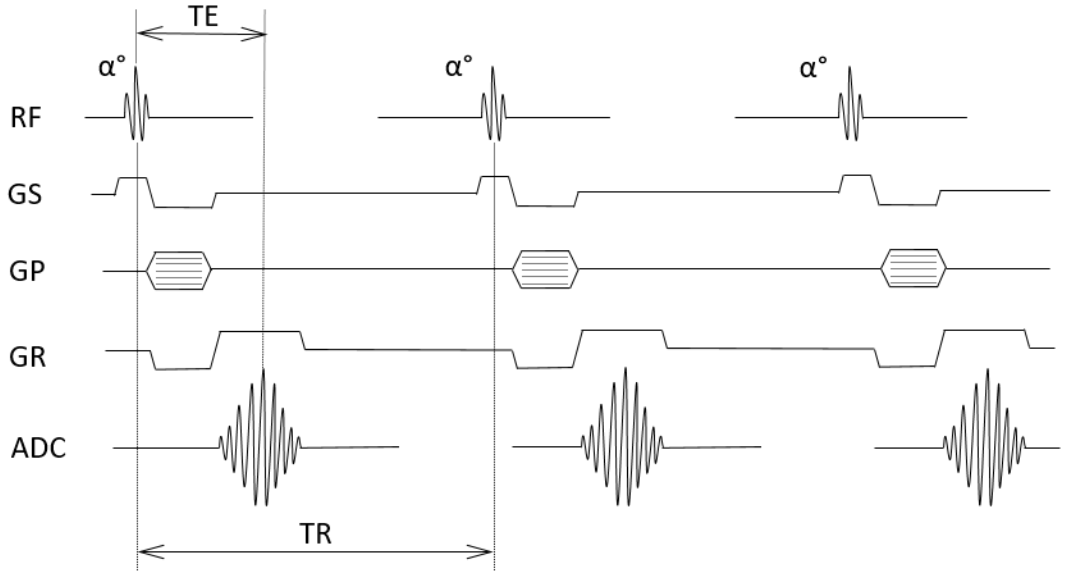

FIGURE 1.1: *A simple pulse sequence illustration. A pulse sequence is similar to an orchestral score, with time increasing from left to right, and each type of manipulation is displayed one above another. The first line illustrates the radiofrequency excitation pulses (RF) with flip angle α. GS, GP, and GR denote the magnetic field gradients in slice-select, phase-encoding, and frequency-encoding directions. ADC is the analog-to-digital converter that is turned on during the data acquisition period.*

Since different types of tissues have different T1 and T2, one can play on any of these parameters to maximize the contrast between tissues and weight the resulting MR signal in T1 and/or T2 by changing the combination of TR and TE (Fig. 1.2). T1 and T2 relaxation times and their corresponding rates R1 (=1/T1) and R2 (=1/T2) denote the characteristic time constants of the recovery back toward equilibrium of the z (longitudinal) and xy (transverse) components, respectively, of the nuclear magnetization. After being disturbed, the MR signal source, i.e., nuclear magnetization, does recover depending on physical laws that rely entirely on hydrogen nuclei interactions with the surrounding tissue environment to re-equilibrate (Haacke et al., 1999). R1 is called the spin-lattice relaxation rate. The "lattice" denotes the molecular environment surrounding a hydrogen nucleus and includes the remainder of the host molecule and other solute and solvent molecules. Spin–lattice relaxation occurs because of magnetic interactions between nuclear spin dipoles and the local, randomly fluctuating magnetic fields that exist on an atomic scale inside any medium. These originate mainly from neighboring magnetic nuclei, such as other hydrogen protons (e.g., within a water molecule, each hydrogen affects the neighbor) and are modulated by the motion of other surrounding dipoles in the lattice, which have components fluctuating with the same frequency as the resonance frequency (Conn, 2009).

Whereas T1 is sensitive to radiofrequency components of the local field, T2 is sensitive to low-frequency components. R2 (=1/T2) is called the spin-spin relaxation rate. In a time of 50 ms, water molecules diffuse distances of 20 cm to sample many different environments on the cellular level within the timescale of relaxation. A very rapid exchange may occur between bulk water, bound water, and interfacial

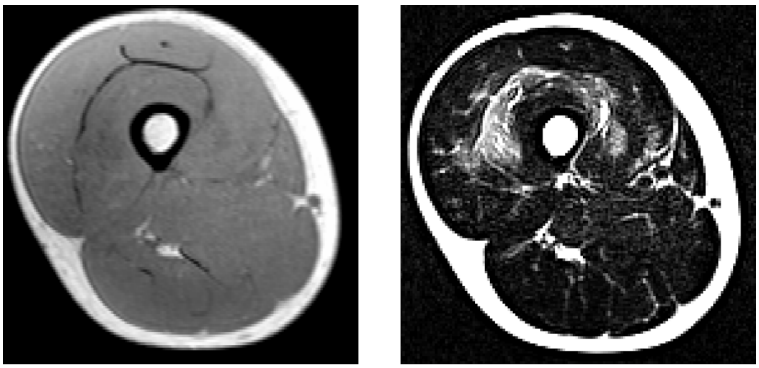

FIGURE 1.2: *Example of T1-weighted and T2-weighted axial images of human thigh obtained by using different combinations of TR and TE.*

water in biological systems in many situations, which is believed to be the origin for many increases in T1 or T2 in various pathologies, such as edematous changes following insults to tissue or in rapidly dividing cells with higher water fractions. Changes in tissue water and protein content, in general, will affect relaxation. Moreover, the existence of water in separate compartments that are only slowly exchanging gives rise to more complex behavior that is not described adequately by a single relaxation rate (Saab et al., 1999). The average water proton relaxation rate measured will depend on how effectively and at what rate these effects are spread through the rest of the water population.

T1 and T2 values are likely to evolve with time in pathologies. T1 and T2 maps are cartographying T1 and T2 directly in each pixel. Such quantitative maps are helpful to monitor the variation of these indexes in diseases before and after medical interventions or treatment since these parameters, once estimated, are independent of hardware changes and externalities to only reflect tissue changes. For example, elevated quantitative values in edematous areas in the heart are reported in acute ischemia, appearing bright in T2-weighted images. T1 and T2 values have also been shown to be influenced by treatments as pre-, post-, or remote ischemic conditioning (Thuny et al., 2012). T2* is the "effective" T2 resulting from inhomogeneities in the main magnetic field or susceptibility-induced field distortions produced in the tissue, related to the presence of chemical or paramagnetic substances such as fat or hemorrhage (Welsch et al., 2014). However, the first step in decoding the various effects of stress on the body using quantitative MRI (qMRI) markers is to develop systematic and comprehensive non-invasive exploration strategies to consistently extract the changes in each MR biomarker to identify which ones reflect the underlying physiological consequences.

When focusing on skeletal muscles, many studies witness that edematous areas appear bright in T2-weighted images because the T2 relaxation time becomes longer. Corresponding T1 and T2 maps show elevated quantitative values in matching areas (Ababneh et al., 2008; Ploutz-Snyder et al., 1997). From that perspective, MRI appears to be a unique imaging modality for extracting relevant anatomical and structural features of muscle tissue (Froeling et al., 2015; Maeo et al., 2017). Recently, quantitative imaging methods such as chemical shift-encoded MRI (Leporq et al., 2013, 2017) and MR relaxometry mapping (Patten et al., 2003; Tawara et al., 2011) have allowed users to understand chemical alterations noninvasively at the

imaged pixel size. Several postprocessing methods need to be sequentially applied to extract a quantitative index from MRI data. These methods include the segmentation of muscle heads on multiple large 3D images, the extraction of image features in each area of interest, and statistical analysis. Each of these steps is an area of research in its own right. Although various alternatives have been proposed to target each of these challenges (Froeling et al., 2015; Maeo et al., 2017), they are rarely applied to cohorts or longitudinal studies.

In the following section, we further focus on the crucial role of image segmentation.

1.2 Importance of image segmentation

Image segmentation is the process of dividing an image into different regions corresponding to different structures and assigning them unique labels. Segmentation has an enormous number of applications in many different fields, among which is medical imaging.

In many medical imaging applications, such as disease diagnosis, patient monitoring, and treatment planning, an accurate delineation is critical to isolate then study each anatomical component locally. As mentioned above, to fully explore the potential of qMRI in a longitudinal study, a robust segmentation method is crucial. The term *robust*, in the case of medical image segmentation, refers to a method that is i) as anatomically coherent as possible with a tolerable quantity of errors compared to a manual expert segmentation (dependent on the clinical application) and, ii) in case of errors, the quantity and the nature of errors should be predictable.

Image segmentation can be obtained by delineating the component borders manually or with computer-aided methods. Manual segmentation requires medical expertise of the anatomical regions in question, but it is very time-consuming and mentally exhausting. State-of-the-art 3D high-resolution isotropic imaging produces a huge number of images to be analyzed and makes manual segmentation in clinical uses irrelevant.

For simplicity, from this point onward, except when specified as manual, the term *image segmentation* in this dissertation is referred to computer-aided, also called automatic, image segmentation. Many works have been done to overcome medical image segmentation problems, yet there is still a need for more. Since there is no universal method, a method needs to be carefully chosen for each specific problem based on the imaging modality, the type of the body part, and the final clinical purpose. Modern approaches, such as deep learning methods, required an enormous number of annotated data to obtain an efficient model while, in our main case study (see chapter 2), the number of manual segmentations is limited.

CHAPTER 2

Longitudinal study case: MUST

Ultra-marathon has been growing rapidly in popularity in recent years, with numerous events being organized each year. Multi-day mountain ultra-endurance race (MUM) is the most intense format of this sport, which puts athletes in many extreme conditions and pushes them to their limits. However, the effects of such conditions on the human body remain mostly unexplored. Notably, microstructural and functional modifications and inflammation induced by these events at the skeletal muscles and myocardium level have never been explored using MRI.

This Ph.D. project is a part of a major research project of the laboratory CREATIS, the MUST study, which aimed to set an unprecedented longitudinal study on a supra- physiological effort of the mountain ultra-marathon athletes using advanced technology MRI techniques.

After reminding the interest of our study (Sec. 2.1), information about the MUM where the volunteers are recruited for our study is provided (Sec. 2.2). The data collection procedure is detailed in Section 2.3.

2.1 Motivation

The recent and innovative developments in quantitative Magnetic Resonance Imaging (qMRI) support a thorough and non-invasive exploration of multiple organs such as skeletal muscles, providing new perspectives to monitor functional alterations that typically occur during disease progression course and/or following any therapeutic interventions. The extreme conditions of MUM are known to lead to a sudden and significant inflammatory response of the body, including skeletal muscles, and as such, are considered as providing a unique experimental accelerated model of injury in humans, closely matching conditions such as those found in intensive care units (ICU) in patients following sudden events (polytrauma, myocardial infarction, ...). (Millet and Millet, 2012; Knechtle and Nikolaidis, 2018).

Therefore, this study was set to explore the capabilities of qMRI to demonstrate changes occurring during and after the MUM challenge on the body. It was also expected to provide new insights into the physiological mechanism of severe muscle damage and the recovery process, its dynamics, helping to identify new non-invasive biomarkers that could contribute to the monitoring of conditions leading to inflammation and skeletal muscle changes observed in clinical settings. There are indeed multiple scenarios met in clinical practice, leading to skeletal muscle changes,

such as disabilities secondary to stroke, cancer, or chronic obstructive pulmonary disease.

The following subsections present the studied skeletal muscles and damages that can be observed. Then, we present the MUM where the dataset was collected.

2.1.1 Skeletal muscle damage

Muscle injury is defined as the loss of muscle function caused by the physical disruption of muscle structures involved in producing or transmitting force (Tiidus, 2008). In this study, we concentrate on skeletal muscles and exercise-induced injury. The type of contractions required in an exercise plays an essential role in the amount of damage caused to the muscles. Some studies have concluded that eccentric contractions are particularly more injurious for skeletal muscles than isometric or concentric contractions (McCully and Faulkner, 1985; Pizza et al., 2002). Early events in eccentric contraction-induced injury include disruption of sarcomeres[1], damage to force-bearing cytoskeleton[2], loss of cell membrane integrity (Fig. 2.1). Loss of calcium homeostasis[3] may contribute to both the initial injury and the progression of the injury. The inflammatory process promotes muscle repair, regeneration, and growth, but it has been shown that the inflammatory cells may also exacerbate the injury (Tidball, 2005).

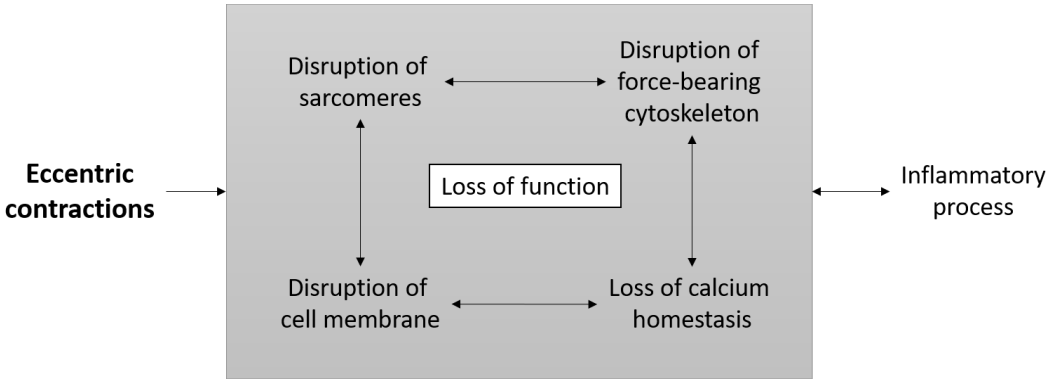

FIGURE 2.1: *Schematic of skeletal muscle injury following eccentric contractions. Early responses include disruption of sarcomeres, force-bearing cytoskeletal elements, and the cell membrane. Loss of calcium homeostasis may contribute to both the initial injury and the progression of the injury. The inflammatory process may also exacerbate the initial injury. Both the initial injury and later events contributing to its progression may cause impaired muscle force production.*

MRI has been used to observe muscular changes after eccentric exercises. The signal intensity of T2-weighted magnetic resonance (MR) image is dependent on the amount of water in the tissue and seems to be able to detect intracellular edema (Nurenberg et al., 1992). On the one hand, MRI was shown to be quite helpful in identifying local muscle damage during eccentric contractions (Nosaka and Clarkson, 1996; Takahashi et al., 1994). On the other hand, MRI is more sensitive than most conventional methods and can detect muscle damage many days after exercise

[1]Sarcomere is the basic unit of skeletal muscle and is composed of thick and thin bundles of proteins as filaments that slide past each other to create muscle contractions.

[2]Cytoskeleton is made up of protein filaments and motor proteins and helps eukaryotic cells maintain its shape and internal organization.

[3]Calcium homeostasis is the regulation of the extracellular concentration of calcium ions.

(Nosaka and Clarkson, 1996; Sayers et al., 1999; Sayers and Clarkson, 2001; Harrison et al., 2001).

In this longitudinal study, which involves MUM runners, the quadriceps are the most affected skeletal muscle due to eccentric effort during downhill running.

2.1.2 Quadriceps

The quadriceps femoris, also called quadriceps, is a large muscle group located on the front of the thigh which connects the hip and the knee joint. It consists of 4 muscles or also called 'heads': rectus femoris (RF), vastus lateralis (VL), vastus medialis (VM) and vastus intermedius (VI) (Fig. 2.2).

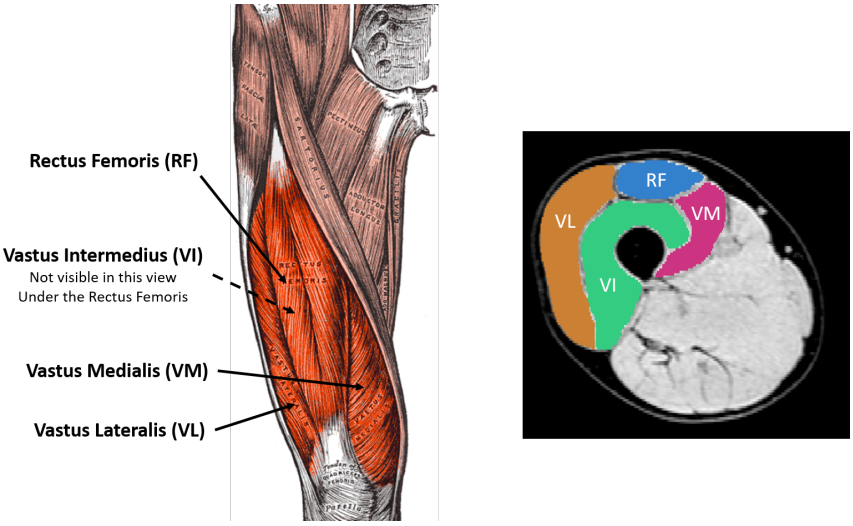

FIGURE 2.2: *The different muscles in the quadriceps. The vastus intermedius is not visible from the front view. On the left: Original image is extracted from Gray and Lewis (1918). On the right: T1W MRI of a right leg from MUST dataset.*

The quadriceps are the crucial extensor of the knee joint. It provides the human body with the abilities of walking, running, jumping, and squatting. Quadriceps damage will cause the loss or reduction of all these abilities and the decline in postural control.

2.2 The mountain ultra-marathon Tor des Géants

An ultra-marathon is an event that involves running for a distance longer than the traditional marathon length of 42.195 kilometers. There are many types of ultra-marathon events distinguished by the distance covered, the time limit, and the nature of the track. Mountain ultra-marathon (MUM), also called ultra-trail, is the most challenging form of this sport. It consists of running and hiking in mountain terrain for many days. Athletes face many severe obstacles, such as inclement weather, elevation change, sleep deprivation, or rugged terrain. Ultra-marathon events are organized globally every year, but only a few are ultra-trail events due to their high difficulty level and a relatively short history of being an organized sport.

MUM has been described as an outstanding model for investigating the physiological responses to extreme load and stress (Millet and Millet, 2012). MUM is a prolonged whole-body exercise with repeated eccentric contractions, which makes

it highly injurious. Many previous studies have established evidence of acute consequences of MUM, such as a decline in pulmonary function (Vernillo et al., 2015) and postural control (Degache et al., 2014), cardiac and neuromuscular fatigue (Maufrais et al., 2016; Saugy et al., 2013), muscle damage and inflammation (Saugy et al., 2013; Hoffman et al., 2012).

To reproduce a model of extreme human tissue stress, the MUST project studies the ultra-racers who participated in the most extreme mountain trail in Europe and one of the most extreme races in the world, the *Tor des Géants*. This is an annual event that takes place in Aosta Valley, Italy, in September, with the start and finishing line in Courmayeur. The route is around 330km with 24000m of total elevation gain (Fig. 2.3), which must be completed in 150 hours (6.25 days).

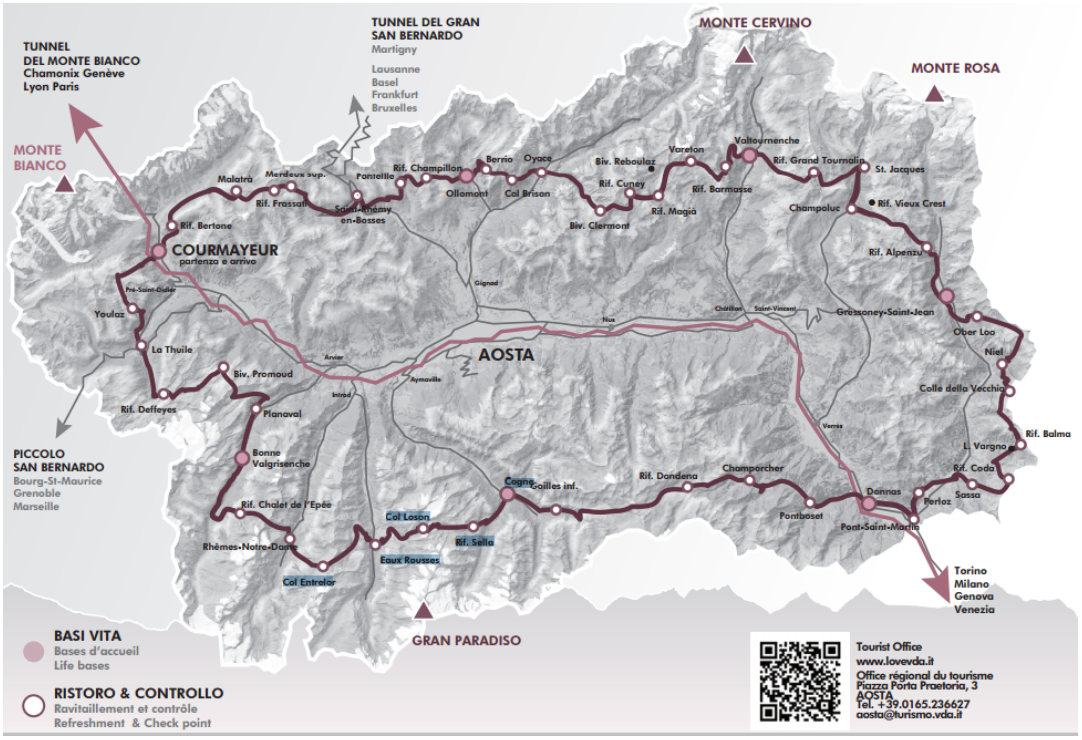

FIGURE 2.3: *The route of the Tor des Géants in the portable roadbook of the competition. The participants have to run in the mountain terrain around the Aosta Valley. They cross 34 municipalities, 25 over-2000-meters mountains, 30 alpine lakes, and 2 natural parks.*

Around 700 people can participate, with only approximately 60% of them who completed the race. In 2014, the best performance was 71 h 49 min among 740 starters and 446 finishers. There is no money prize at stake, so all racers are there to test their limits. Most of the racers are experienced ultra-marathon runners that have taken part in other ultra-marathon races. Well-trained racers like them usually have leg and hip musculature that is relatively resistant to injury (Tiidus, 2008). Participants of the race must also have the ability to adapt to encountered or expected problems, such as change of weather, physical problems, and sleep deprivation. They can only shower, sleep, or get medical care at the official refreshment points; otherwise, they are autonomous. The Tor des Géants is no doubt one of the most demanding races in the world.

2.3 MUST data collection

The MUST project followed the Tor de Géants 2014 that took place from September 7th to September 14th, 2014. This study was approved by the local ethical committee (Aosta Valley, Azienda USL 101/946), and the experimental plan was conducted in accordance with the Helsinki Declaration (2001). Subjects were recruited through mailing and public announcements to registered runners by race organizers. Exclusion criteria were smoking, substance abuse, regular intake of medications, medical or psychiatric illness, and any contraindication to MRI (e.g., claustrophobia, nonremovable metal devices) or abnormalities detected upon laboratory screening. As it was impossible to collect data from all 740 participants and as many runners are so highly motivated that they do not want to miss any moment of the race, a limited number of runners were recruited to be the subjects of the study. In the end, we had 51 runners that volunteered and provided written consent to participate in this study. Only 27 among them finished the race, which is coherent with the percentage of finishing runners of the race overall. Demographic information of the participants is listed in the table A.1

The experiment design was longitudinal, which involved following the runners at four time points during the race. Blood samples were taken at each time point, while, due to time restriction, MRI and ultrasound examinations were performed only at three time points. The MRI is principally focused on upper leg muscles, heart, and brain. In the context of this Ph.D. project, we are interested mostly in the MRI data of the upper legs, more specifically, the quadriceps muscles, and the biological data:

- The first point (pre-race: **Pre**) was at the start location. The data collection was performed within 4 days before the race and consisted of MRI acquisition and biological sampling.
- The second point was located halfway through the race (middle: **Mid**). Only biological sampling was performed.
- The third one was at the arrival of the race (arrival: **Post**): athletes who finished the race were transported by car to the laboratory and were evaluated (MRI and biological sampling) within one hour after finishing the race.
- The last point (recovery: **Post+3**) was 48-72 h after arrival time. Both MRI and biological sampling were acquired.

The flowchart of the study is shown in the figure 2.4. While the image feature extraction and analysis is also a part of this project, our main interest here is to search for a robust segmentation method that will allow us to study, with precision, each muscle head in the quadriceps locally.

2.3.1 MRI acquisitions

At Pre, Post, and Post+3, MRI acquisitions were performed on-site using a mobile 1.5 T MR scanner system (MAGNETOM Avanto, Siemens Healthcare, Erlangen, Germany installed in a truck from Alliance Medical, England). A standard coil configuration was used: a 4-channels body-array surface coil combined with 4 elements of the spine coil, resulting in an 8-channels coil in total. Three MRI acquisitions of the legs were sequentially performed:

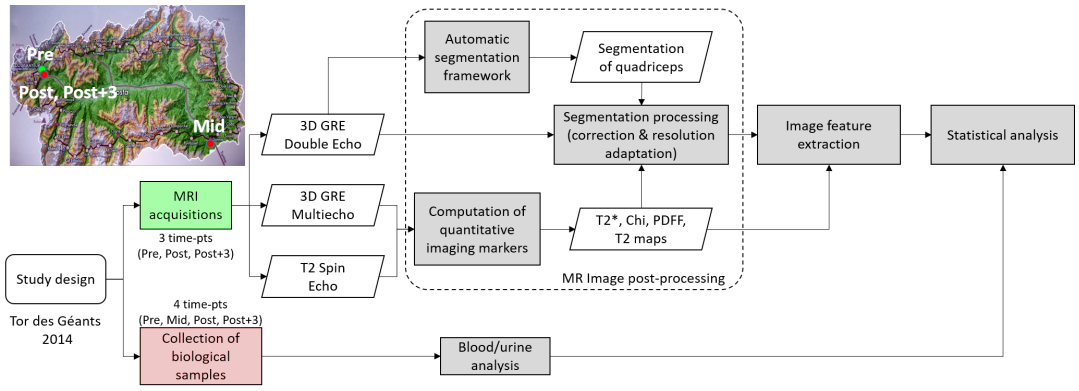

FIGURE 2.4: *Flowchart summarizing the main steps of the study of the functional variation in the quadriceps. MRI was performed at three time points: before (Pre), and for participants who finished the race, immediately after (Post), and 48-72 h after the race (Post+3). Meanwhile, biological sampling was performed at four time points (an additional sample was performed at half-race (Mid)).*

- **A three-dimensional (3D) isotropic gradient dual-echo sequence**: The coronal acquisition included the entire upper leg (from the tibial tuberosity to the anterior superior iliac spine) with a total scan time of 3 minutes. The voxel size was $0.781\,25 \times 0.781\,24 \times 1.3\,\text{mm}^3$, the number of slices was 176 resulting in a total coverage in the z-direction of 20.8 mm and an explored 3D volume of $437.5 \times 500 \times 208\,\text{mm}^3$, i.e. an in-plane field of view (FOV) of $437.5 \times 500\,\text{mm}^2$. The pixel-wise volume size is $560 \times 640 \times 176$. The reconstruction of the water and fat images from the acquired multi-echo data sets was performed inline using a Dixon approach (Leyendecker et al., 2010) enabling four 3D isotropic in-phase, out-of-phase, fat-only and water-only coronal images to be calculated on the MR scanner, hereafter denoted in-phase (IN), out-of-phase (OUT), water (W), and fat (F) images, respectively (Fig. 2.5).

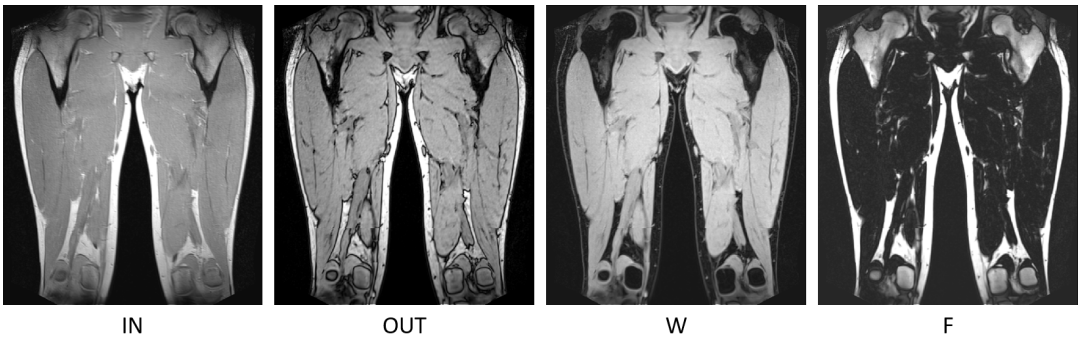

FIGURE 2.5: *Coronal view of the four T1-weighted images obtained with the Dixon approach: in-phase (IN), out-phase (OUT), water (W), and fat (F)*

- **A 3D spoiled gradient echo sequence (3D GRE)**: acquired in axial plane with the voxel size of $1.5625 \times 1.5625 \times 5\,\text{mm}^3$, and with a total coverage in the z-direction of 240 mm and 140 mm prior and after aliasing elimination in the slice direction, 3D volume of $400 \times 280 \times 140\,\text{mm}^3$ i.e. an in-plane FOV of $400 \times 280\,\text{mm}^2$. The pixel-wise volume size is $256 \times 160 \times 28$. Eight echoes were acquired in the transverse plane with a flyback readout gradient (first echo: 1.58 ms and echo spacing: 2.52 ms). TR and flip angle were adjusted

to minimize the T1-related bias. Phase and magnitude images were systematically reconstructed. Prescription of localization was performed using anatomic images from an isotropic 3D gradient-echo acquisition. For standardization purposes, the central partition in the z-direction was planned at a 15 cm distance from the upper part of the patella using the sagittal multiplanar reconstruction of the first acquired 3D gradient isotropic sequence.

Using this sequence, a set of quantitative indexes and maps can be calculated, including the magnetic susceptibility (χ) map, the T2* relaxation time (T2*) map, and the proton density fat fraction (PDFF) map (Leporq et al., 2017). These maps were computed using an in-house program in MATLAB (MATLAB, 2017) and the process encompassed two main steps: fat-water separation and magnetic susceptibility quantification (Fig. 2.6).

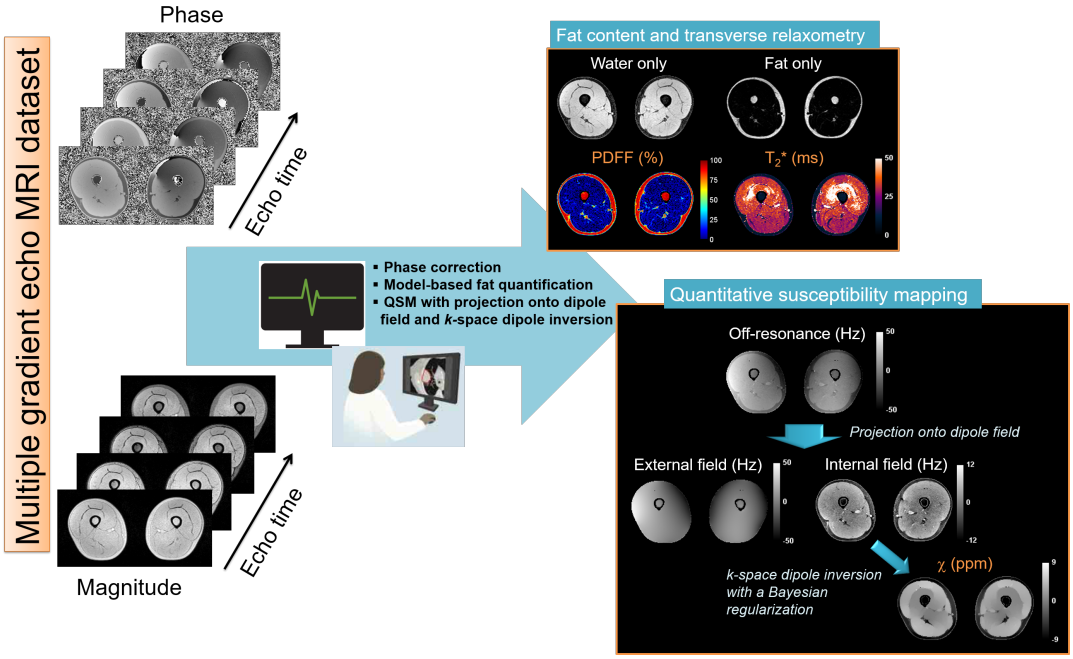

FIGURE 2.6: *Quantitative map reconstruction from 3D multi-echo GRE sequences. Phase images were unwrapped beforehand to compute the B_0 field inhomogeneities (ΔB_0) map and the ΔB_0-demodulated real part images from which fat-water separation was performed. The fat-water separation step provided parametric T2* and PDFF maps. From the ΔB_0 map, the external (B_{out}) and internal fields (B_{int}) were separated using the projection onto the dipole field. From B_{int}, performed with a single orientation Bayesian regularization including spatial priors derived from magnitude images for the boundary conditions, error and smoothness weighting to compute the susceptibility map (Viallon et al., 2019).*

- **A 2D multi-echos T2 weighted spin-echo sequence** with 16 echo times (TEs) ranging from 10 to 178 ms: the voxel size is of $1.25 \times 1.25 \times 10\,mm^3$, the FOV is of $400 \times 250\,mm^2$ with a z-direction coverage of 70 mm. The pixel-wise volume size is $320 \times 200 \times 7$. The central slice was also planned at the same location as previously described for the 3D GRE sequences. The T2 relaxation time (T2) maps were calculated using an in-house written program in MATLAB considering a mono-component T2 relaxation and thus implementing a single-parameter least square fit (Levenberg-Marquardt)(Gavin, 2013). The first non-stimulated echo was removed as recommended to account for the presence of stimulated echoes as indicated in (Azzabou et al., 2015; Kan et al., 2009) (Fig. 2.7).

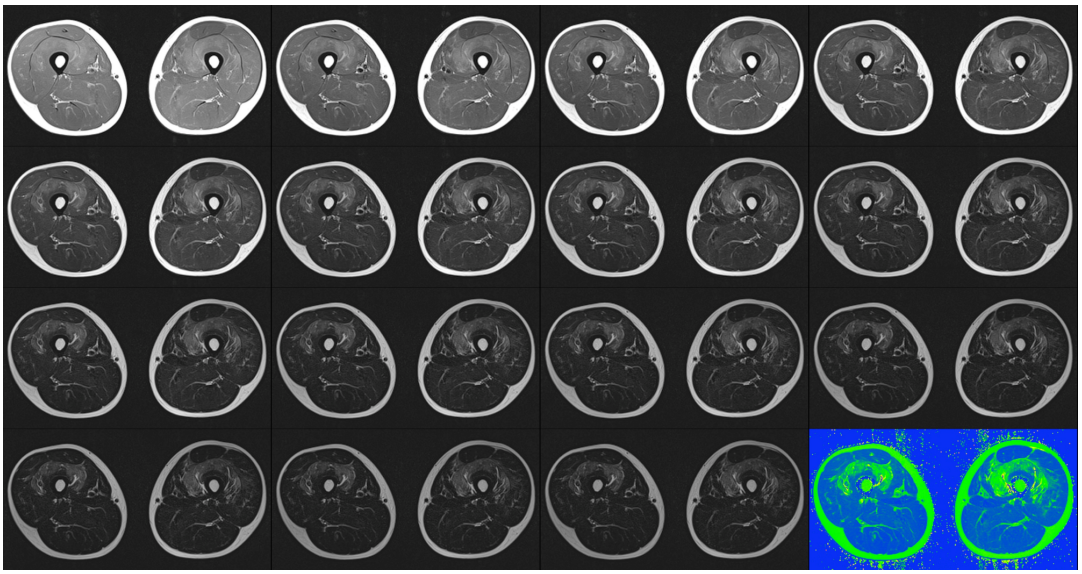

FIGURE 2.7: *Quantitative T2 color maps obtained after 2-parameter fit of the multiecho spin-echo (meSE) T2-weighted images acquired with different echo times (TEs) in one athlete after arrival, showing clearly increased T2 in the vastus intermedius (VI) regions.*

As we can see, each sequence had a different resolution and field of view (individually optimized in terms of SNR, coverage, and MR properties). The area explored by each technique is summarized in the figure 2.8a, and the main MR parameters are listed in the table A.2. An automatic segmentation based on shape matching (Gilles et al., 2016) was performed on the W images.

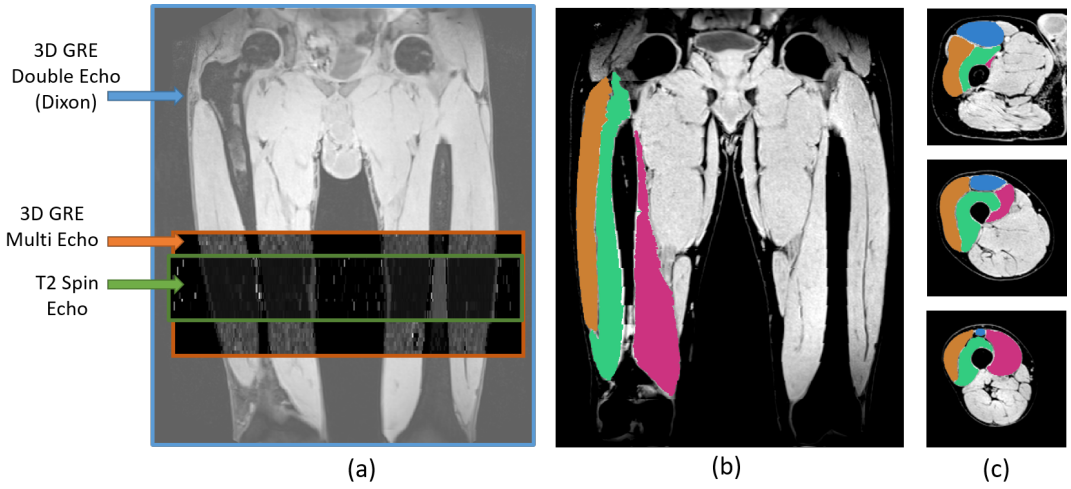

FIGURE 2.8: *MRI Input Sequences and result of Gilles et al.'s segmentation. 3D GRE and T2 Spin Echo sequences superposed on the 3D GRE Double Echo water image (a), the coronal (b) and axial (c) views of the manual segmentation the right leg displayed with the isotropic water image as background.*

2.3.2 Manual segmentations

An automatic segmentation method needs to be validated by applying to an image dataset whose segmentation has been done manually or verified by medical experts. Such a dataset for quadriceps muscles does not exist publicly. Our medical experts at CHU Saint-Étienne have been manually segmenting MRI images from

MUST dataset, more precisely, the images T1-Water (Figure 2.8b,c). Since our high-resolution T1 images cover the entire upper legs of the athletes, which take a lot of time to segment, only the right legs in 7 images at Pre are fully segmented.

Additionally, to validate the segmentation methods on left legs and other time-points, partial manual segmentations are available for 5 left legs and then 4 right legs at all of the 3 time points. The detail on the manual segmentation dataset is presented in the table 2.1. A total of 4 medical experts were involved in the manual segmentation of right leg volumes, while for left leg partial segmentation, each subject among the 5 segmented was segmented by two medical experts simultaneously. The manual segmentation and slice selection (for partial manual segmentation) will be described in Chapter 6.

	Right legs			Left legs
Subject	*Pre* (all slices)	*Post* (17 slices)	*Post+3* (17 slices)	*Pre* (50 slices)
ALB-2725	①	③	③	③ ④
ALF-4529	①	③	③	
ANG-2014	②			
ANS-3229	②			
ARS-4026	②	③	③	
BRG-1924				③ ④
CAL-4223	③			③ ④
MAV-526				③ ④
OUK-2927	④	③	③	
YAG-47				③ ④

TABLE 2.1: *Available manual segmentations per subject. Subjects are presented by their monograms. Each medical expert is represented by a number. Pre, Post, and Post+3 are the three MRI acquisition time points. Image slices are axial slices. There are, in total, 640 axial slices per image volume.*

2.3.3 Biological sampling and analysis

The usage of the biological data is not considered in this Ph.D. thesis but was intensively used in our work Nguyen et al. (2021b). Introduction on these biomarkers can be found in Appendice A.3.

Conclusion

While extreme ultra-endurance races are growing in popularity, their effects on skeletal muscles remain mostly unexplored, although some subjects display inflammatory symptoms matching those found in ICU patients after polytrauma or in patients with muscle disorders. The MUST longitudinal study was built to, among multiple objectives, explore physiological changes in mountain ultramarathon athletes' quadriceps using quantitative magnetic resonance imaging (qMRI) coupled with biological markers. This Ph.D. project resulted from an urgent demand for a robust automatic segmentation method for muscle in MR images, specifically quadriceps images. Automatic image segmentation is a rapidly growing domain, but few studies involved quadriceps segmentation, especially ones of professional athletes. Meanwhile, one of our main difficulties is the limited number of available manual segmentation. In the next part, we will present the state-of-the-art of medical image segmentation, and more specifically, methods that used a small number of manual annotations.

PART II

Segmentation Methods & Validation

Contents

Résumé

Dans cette partie, nous explorons les techniques de segmentation adaptées à notre problématique que nous commençons par décrire.

La segmentation de l'image consiste à créer une partition d'une image en groupes de pixels ou en régions non chevauchantes ayant des propriétés communes qui les différencient des autres groupes. Un label est ensuite attribué pour identifier les pixels d'une même région. Ces régions sont associées à différents objets, en particulier en imagerie médicale, à des pathologies ou des structures anatomiques. Ce processus de délimitation peut être effectué manuellement. Cependant, selon le nombre de structures à délimiter, la taille de l'image et la modalité d'imagerie, la complexité de l'annotation manuelle peut exiger un temps considérable d'expert pour cette tâche. D'autres facteurs externes, tels que la fatigue ophtalmologique et mentale due à la répétitivité et à l'exigence d'une forte concentration pour cette tâche, peuvent affecter la qualité de la segmentation et éventuellement le diagnostic voire le traitement du patient.

Un problème commun de la segmentation automatique des images médicales est le manque de données annotées manuellement. En fait, il existe un nombre limité de bases de données d'images médicales publiques, dont la plupart concernent soit le cerveau, soit le cœur. À notre connaissance, il n'existe pas de base de données publiques portant sur les cuisses humaines.

La segmentation des muscles à partir d'IRM a toujours été un défi en raison de l'absence de limites musculaires définies, de l'inhomogénéité de l'intensité du champ ou des artefacts d'acquisition (Prescott et al., 2011). Dans le cas des coureurs MUM qui ont participé au projet MUST, des athlètes professionnels ayant des muscles quadriceps très développés et une quantité minimale de graisse corporelle, la plupart du temps, la détermination des limites musculaires n'est pas évidente, même pour les experts médicaux. Ces problèmes rendent inopérantes les méthodes de segmentation automatiques standards telles que le seuillage ou la croissance de région.

Cette partie présente d'abord, au chapitre 3, les mesures d'évaluation quantitatives des méthodes de segmentation, puis elle donne un aperçu des méthodes de segmentation existante pour l'imagerie médicales (Chapitre 4) et puis des détails de deux méthodes que nous avons jugées adaptées à notre tâche spécifique : la segmentation par recalage multi-atlas et la segmentation par réseau de neurones UNet. Les deux approches seront ensuite appliquées à notre jeu de données puis seront la base de nos contributions (Partie III). La partie se termine avec un rapport de l'état de l'art sur la segmentation automatique des quadriceps humains et les difficultés spécifiques à notre jeu de données qu'il faudra prendre en compte pour l'interprétation des résultats de nos contributions (Chapitre 5).

Introduction

Image segmentation consists of partitioning an image into non-overlapping groups of pixels or regions with common properties that differentiate them from the other groups. A label is then assigned to identify the pixels in a region. These regions are associated with different objects, specifically in medical imaging, anatomical structures. This delineation process can be performed manually; however, depending on the number of structures to be delimited, the image size, and the imaging modality, the complexity of the annotation may increase to the point of requiring a trained expert with several years of study and practice, as is the case for a radiologist. Other external factors, such as ophthalmological and mental fatigue due to the repetitiveness and the demand for a high concentration of the task, can affect the quality of the segmentation and possibly the diagnosis or treatment of the patient.

A general problem in medical image segmentation is the lack of annotated data. As a matter of fact, there is a limited number of public medical image datasets, with most of them either involving brains or hearts. To our knowledge, there is no sizeable public dataset for human thighs.

In this part, we first present, in Chapter 3, essential elements of the evaluation procedure of a segmentation method. Chapter 4 offers an overview of existing segmentation methods for medical imaging and details of methods that we judged suitable for our specific task. Finally, we present the difficulties of muscle segmentation from MRI data and the existing automatic segmentation methods for human quadriceps (Chapter 5).

CHAPTER 3

Segmentation method validation

The validation of a segmentation method is the primary step in understanding its strengths and weaknesses. In addition to the quality of the segmentation, which can be quantified by specific metrics (Sec. 3.1), the computational time (Sec. 3.3) is also a crucial factor in evaluating the usability in clinical practice of the method. Furthermore, in this chapter, we present the cross-validation procedure (Sec. 3.2) for generalization capability evaluation.

3.1 Metrics for segmentation evaluation

There are many metrics for assessing the quality of segmentation (Taha and Hanbury, 2015). These measures are often described for a segmentation into two classes of an image (or binary): an object and the background. In general, they are well generalized to the multi-class case where several objects and a background are segmented.

3.1.1 Sørensen–Dice coefficient

The DICE coefficient (Dice, 1945) is the most common measure comparing two binary sets R (reference) and T (test).

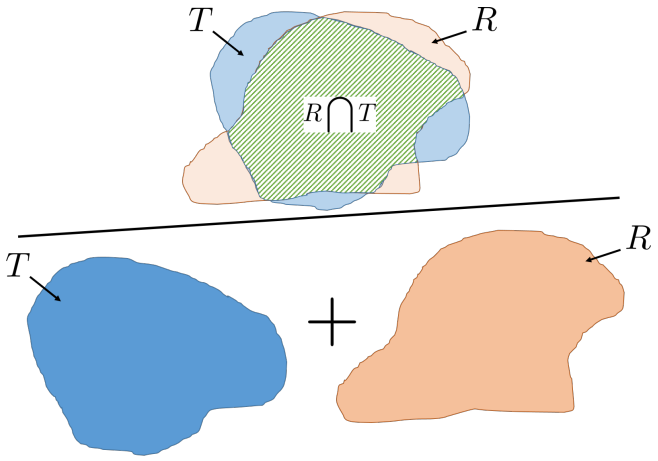

FIGURE 3.1: *Illustration of the DICE Similarity Coefficient.*

$$DICE(R,T) = \frac{2|R \bigcap T|}{|R| + |T|} = \frac{2|R \bigcap T|}{|R \bigcup T| + |R \bigcap T|} \tag{3.1}$$

or

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{3.2}$$

where:

- TP is *True Positive*: number of pixels where T and R both have the *object* value
- TN is *True Negative*: number of pixels where T and R both have the *background* value
- FP is *False Positive*: number of pixels where T has the *background* value and R has the *object* value
- FN is *False Negative*: number of pixels where T has the *object* value and R has the *background* value

DICE is equivalent to the F_1 score. The higher the DICE coefficient, the more similar the 2 segmentations.

In the multi-class case, the global DICE score, here denoted *DSC (Dice Similarity Coefficient)*, is the average value of individual DICE scores of all the classes. There is also a weighted version of DSC, here denoted DCSw:

$$DSCw = 2\frac{\sum_{c=1}^{N} |R_c \cap T_c|}{\sum_{c=1}^{N} |R_c| + |T_c|} \tag{3.3}$$

where N is the number of classes in the segmentation.

The main difference between DSC and DSCw is the impact of small regions on the overall score: DSC is more sensitive to small region error than DSCw that weights all regions DICE proportionally to their size.

3.1.2 Jaccard coefficient or IoU

The Jaccard coefficient (JC) or IoU (*Intersection over Union*) measures the ratio between the intersection and the union of two sets.

Its formula is as follows:

$$JC(R,T) = \frac{|R \bigcap T|}{|R \bigcup T|} \tag{3.4}$$

This metrics is very similar to the DICE coefficient but is formulated differently. The relationship between the Jaccard and DICE coefficients is:

$$JC(R,T) = \frac{DICE(R,T)}{2 - DICE(R,T)} \quad \text{and} \quad DICE(R,T) = \frac{2JC(R,T)}{1 + JC(R,T)} \tag{3.5}$$

The higher the Jaccard coefficient, the more similar the 2 segmentations.

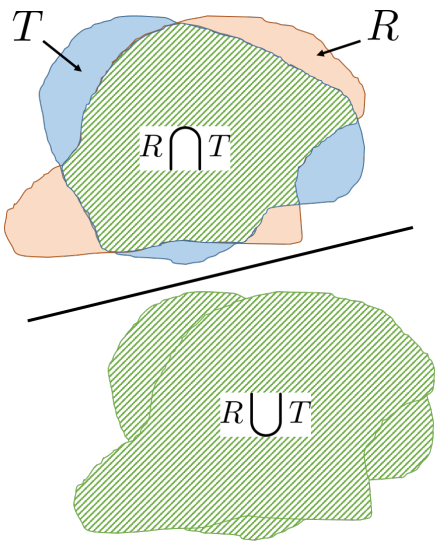

FIGURE 3.2: *Illustration of the Jaccard coefficient (also called IoU score)*

3.1.3 Hausdorff distance

The Hausdorff distance, denoted HD, is based on the distance between the segmentation surfaces (see figure 3.3).

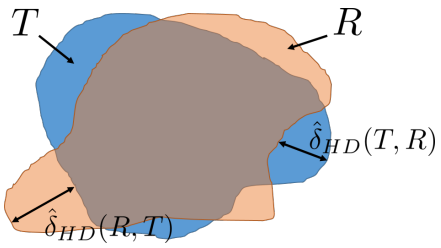

FIGURE 3.3: *Illustration of Hausdorff Distance.*

The Hausdorff pseudo-distance $\hat{\delta}_{HD}$, non-symmetrical, is defined as follows:

$$\hat{\delta}_{HD}(R,T) = \max_{r \in R} \min_{t \in T} ||r - t|| \tag{3.6}$$

and the Hausdorff distance (with the symmetry property) δ_{HD} is defined as:

$$\delta_{HD}(R,T) = \delta_{HD}(T,R) = \max\left(\hat{\delta}_{HD}(R,T), \hat{\delta}_{HD}(T,R)\right) \tag{3.7}$$

The smaller the HD, the smaller the maximum errors between the two segmentations. For multi-class, it is recommended to compute the HD for each class and then to keep the maximum of all these distances as the global HD, which is the way the global HD is computed in this thesis.

3.1.4 Mean absolute distance

The mean absolute distance, denoted MAD or d_m, between two sets R and T is defined as :

$$d_m(R,T) = MAD(R,T) = \frac{1}{2}\left(\bar{d}(R,T) + \bar{d}(T,R)\right) \tag{3.8}$$

with

$$\bar{d}(R,T) = \frac{1}{|R|}\sum_{r\in R}\min_{t\in T}\|r-t\| \tag{3.9}$$

The illustration of this metric is shown in figure 3.4.

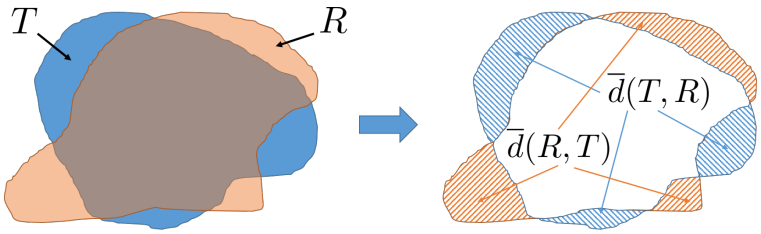

FIGURE 3.4: *Illustration of the Mean Absolute Distance. On each of the hatched regions, the minimum distance to the other region is accumulated for the calculation of the* $\bar{d}$

Meanwhile, it should be noted that this metrics is not equivalent to the error area ($|R \bigcup T - R \bigcap T|$) since, for each point, it is the minimum distance to the other set that is searched and then accumulated.

For multi-class cases, similar to HD, the global MAD is the maximum of all the individual MAD of classes.

3.1.5 Volume similarity

The volume similarity, denoted VS, between two sets R and T is defined as :

$$VS(R,T) = 2\frac{|T| - |R|}{|T| + |R|} \tag{3.10}$$

The VS metric ignores all the spatial information; therefore, while VS should correlate with DICE score, two segmentation with DICE score equal to 0 can have a perfect VS ($VS = 0$) if they have the same volume. A high VS is only in favor if the corresponding DICE score is also relatively high.

In the case of multi-class segmentation, the global VS is defined as:

$$VS(R,T) = \frac{2}{N}\sum_{c=1}^{N}\frac{||T_c| - |R_c||}{|T_c| + |R_c|} \tag{3.11}$$

where N is the number of classes in the segmentation.

3.2 Cross-validation

Cross-validation (Hastie et al., 2009) is a resampling procedure used to evaluate the performance of a given method on a data sample, specifically in the case of image

segmentation on a limited amount of annotated data. To estimate the generalization capability of a method, one needs to test that method on *unseen* data. By splitting the N available annotated data into k groups (*k-fold cross-validation*), we can test the method on each of these k groups using the other $k-1$ groups as reference or training data.

In the case of a very small N, which is our case ($N = 7$), we can use the *N-fold* strategy or also called *Leave-One-Out* (LOO) consisting of taking each subject as a test and all the others as references. The LOO strategy is employed in most of our experiments presented in this dissertation.

3.3 Computation time

In clinical practice, the computation time of a method is a non-negligible factor. For many applications, the computation time must be relatively short to be transferred to everyday practice. While working with methods involving machine learning, there are two types of computation time:

- **Training time**: The amount of time needed to train a certain model, i.e., the time it took for our computer to learn a given pattern.
- **Inference time**: The amount of time taken to apply the trained model on a test subject

The training time directly affects the duration of the research and development phase of a method, but once a proper model is established, the inference time is the more critical factor in the evaluation phase.

CHAPTER 4

Segmentation methods dealing with few annotations

There are many different approaches to the problem of image segmentation based on different perspectives, such as the similarity between objects and boundaries of regions of interest. Below is a non-exhaustive list of common segmentation methods frequently employed in medical image processing:

- **Thresholding**: The image pixels are classified by comparing their gray level (or intensity) to one or many threshold values. Many threshold-derived methods have been developed for medical imaging applications (Ng et al., 2008; Li et al., 2006; Khare and Tiwary, 2005).
- **Region growing**: From initial seed points selected manually or automatically, regions of interest are expanded to adjacent points depending on predefined criteria. These criteria could be pixel intensity grayscale texture, color and/or gradients (Grenier et al., 2006). This method is frequently used in tumor or abnormality segmentation (Haider et al., 2011; Deng et al., 2010; Siddique et al., 2006).
- **Deformable models**: Deformable models (Metaxas, 1996; Hegadi et al., 2010) are also called active contours as they deformed the objects' boundaries based on the shape of objects, smoothness of contours, internal forces, and external forces on the objects.
- **Atlas-based segmentation**: The segmentation is carried out by mapping a reference image that has verified segmentation on a new image that needs to be segmented. This is one of the most popular methods in medical image segmentation as it can handle from very large to very small anatomical changes (Rohlfing et al., 2005; BachCuadra et al., 2015; Wang et al., 2013).
- **Classification**: Supervised learning methods are trained using manual segmentations of training images to derive a classification model that can be applied to images of the same modality and the same body part. Each pixel in an image is treated as an individual characterized by its intensity, position, neighborhood, or the local features around it. Two of the most common classification techniques used in medical image segmentation are Decision Tree (Safavian and Landgrebe, 1991) and Artificial Neural Networks (Litjens et al., 2017).

- **Clustering**: While classification methods use training data and are considered as supervised learning, the clustering approach is unsupervised. It divides an image into different regions based on the statistics of the dataset. The two most common clustering methods used for medical image segmentation are K-Means (Lee et al., 2008; Muda and Salam, 2011) and Fuzzy C-Means (Li et al., 2008; Balafar et al., 2008).

With the increasing complexity of medical imaging problems, the interest of the research community has shifted from atlas-based or deformable models to methods of the machine learning family (Clarke et al., 1995; Pham et al., 2000; Sharma and Aggarwal, 2010). The swift emergence of deep learning, a branch of machine learning, has been gaining much attention, if not almost all of the attention (Qayyum et al., 2018; Cai et al., 2020) for its efficiency and potential in solving not only image segmentation problems but also many other medical image analysis problems (Litjens et al., 2017).

As mentioned above, for our particular muscle segmentation problem with few annotations, non-supervised methods such as threshold, deformable model, or clustering are not suitable. In this work, our focus is on atlas-based segmentation and deep learning methods.

4.1 Atlas-based segmentation

Atlas-based segmentation (Rohlfing et al., 2005; BachCuadra et al., 2015) is a widely used method in medical image analysis that involves using an *atlas* to help separate the anatomic structures in a medical image. An *atlas* is a *labeled reference*, which here we considered as two images: a reference MRI image and a label image that contains the segmentation information of the reference image done by medical experts. The reference image is registered to a new image that needs to be segmented; the transformation found by the registration process is then applied to the label image of the reference to obtain a new label image that contains the segmentation information of the image considered. The segmentation problem now turns into a registration problem.

4.1.1 Registration

In image processing, registration is a process of finding a correct transformation that can bring an image into spatial correspondence with another image. A registration problem between 2 images (F: fixed image, M: moving image) can be formulated as an optimization problem that maximize a similarity measure $\mathcal{S}$ with a spatial transformation T^* of type $\mathcal{T}$:

$$T^* = \arg\max_{T \in \mathcal{T}} \mathcal{S}(F, M; T) \tag{4.1}$$

A registration system is defined based on the main criteria: the primitives used, the nature of the transformation, the similarity criterion, and the optimization method. These criteria are not independent of each other and depend on the type of images, the imaging modalities, and the specific registration problem (2D/3D, mono/multi-modal, intra/inter-patient ...).

4.1.1.1 The primitives used

There are two types of primitives: extrinsic and intrinsic. The extrinsic primitives are the stereotactic framework, the markers or the calibration of the acquisition systems. Intrinsic primitives (the content of the image) are the most used. A method can be based on one of the characteristics extracted from the image (points of reference, anatomical structures, local descriptors, ...) or the intensities of the pixels.

4.1.1.2 The nature of transformation

The transformations used in the registration are divided into four categories. A rigid transformation contains only translations and rotations. An affine transformation includes, in addition to rigid transformation, scaling, homothety, and shear mapping. It retains the parallelism between the lines. A projective transformation has more degrees of freedom than an affine transformation as it does not preserve parallelism but only the projective structure. The three previous transformation types are all linear. On the other hand, a transformation can also be non-linear.

The nature of the transformation $\mathcal{T}$ has to be defined according to the problem. In our case, a linear transformation would not be able to find the small local changes in the muscles, so we aimed for a deformable registration method. There are two types of models to represent such transformation: parametric and non-parametric. A preliminary test on two popular approaches that represent these two types: B-spline deformation (Rueckert et al., 1999) for parametric model and Demons (Thirion, 1998) for non-parametric model has yielded a result in favor of the B-spline deformation method on which we will concentrate from now on.

The application of B-splines in non-rigid image registration of MR images was introduced by Rueckert *et al.* (Rueckert et al., 1999). The global deformations of images needed to be modeled first by a rigid or affine transformation, while the local changes would be modeled by a free-form deformation (FFD) based on B-splines.

Let the image volume be denoted as $\Omega = \{(x, y, z) \mid 0 \leq x < X, 0 \leq y < Y, 0 \leq z < Z\}$. The basic idea of FFD is to manipulate a mesh Φ of $n_x \times n_y \times n_z$ control points $\phi_{i,j,k}$ of Ω. the FFD can be written as the 3-D tensor product of the familiar 1-D cubic B-splines (Rueckert et al., 1999):

$$T_{local}(x, y, z) = \sum_{l=0}^{3} \sum_{m=0}^{3} \sum_{n=0}^{3} B_l(u) B_m(v) B_n(w) \phi_{i+l, j+m, k+n} \tag{4.2}$$

where $i = \lfloor x/n_x \rfloor$, $j = \lfloor y/n_y \rfloor$, $k = \lfloor z/n_z \rfloor$, $u = x/n_x - \lfloor x/n_x \rfloor$, $v = y/n_y \lfloor y/n_y \rfloor$, $w = z/n_z - \lfloor z/n_z \rfloor$ and B_l represents the lth basis function of B-spline:

$$B_0(u) = (1-u)^3/6$$
$$B_1(u) = (3u^3 - 6u^2 + 4)/6$$
$$B_2(u) = (-3u^3 + 3u^2 + 3u + 1)/6$$
$$B_3(u) = u^3/6$$

The B-splines are locally controlled, so they are very computationally efficient even for a large number of control points. Thus, it is possible to use this transformation in a multi-resolution approach where the local transformation is calculated

at different levels defined with different resolutions of the control mesh. The final transformation will be the sum of the transformation computed at all levels.

4.1.1.3 Similarity criterion

The criterion of similarity depends mainly on the primitives used. If the method is based on features extracted from the image, the similarity criterion will be the distance between the corresponding primitives (points, lines, or pieces), which will be minimized. In intensity-based registration, the most common criterion is based on the quadratic error for images acquired with the same modality and on correlation ratios or measures from information theory (and mainly: mutual information) for images from different modalities (Maes et al., 2015). We use the measure of mutual information in the following, even if the modalities of the images are the same because we have observed that it allows us to obtain satisfactory registration for our different projects.

The mutual information can be defined with the following formula:

$$MI(A,B) = H(A) - H(B) - H(A,B) \tag{4.3}$$

where $H(A)$ and $H(B)$ are marginal entropies of images A and B respectively and $H(A,B)$ is their joint entropy. The entropies are defined as:

$$H(A) = -\int p_A(a) \log p_A(a) da \tag{4.4}$$

$$H(B) = -\int p_B(b) \log p_B(b) db \tag{4.5}$$

$$H(A,B) = -\int p_{AB}(a,b) \log p_{AB}(a,b) da db \tag{4.6}$$

where p_A, p_B and p_{AB} are respectively marginal probability density functions for A and B and their joint probability density function. The larger the MI, the more similar the two images.

4.1.1.4 The optimization method

The optimization method depends mainly on the choices of the previous criteria. For methods based on geometric primitives (landmarks, local descriptors, etc.), the optimal transformation can be found with the least squares algorithm. In the case of intensity-based methods, gradient descent methods are very often used.

4.1.2 Mono-atlas segmentation

The figure 4.1 summarizes the principle of mono-atlas segmentation. An atlas A will be the grouping of the intensity image and the manual segmentation $A = (A_F, A_S)$.

The optimized transformations are then applied to the manual segmentation using, for the interpolation step, a method that does not create new labels like the k nearest neighbors. With a single atlas, relevant multi-organ segmentation can be obtained, as in the case of bones in X-ray imaging (Moreau et al., 2016).

Nevertheless, segmentation by atlas registration is very dependent on the registration process: segmentation errors will mainly come from errors related to the fact that the registration of the atlas image on the target image has partially or entirely

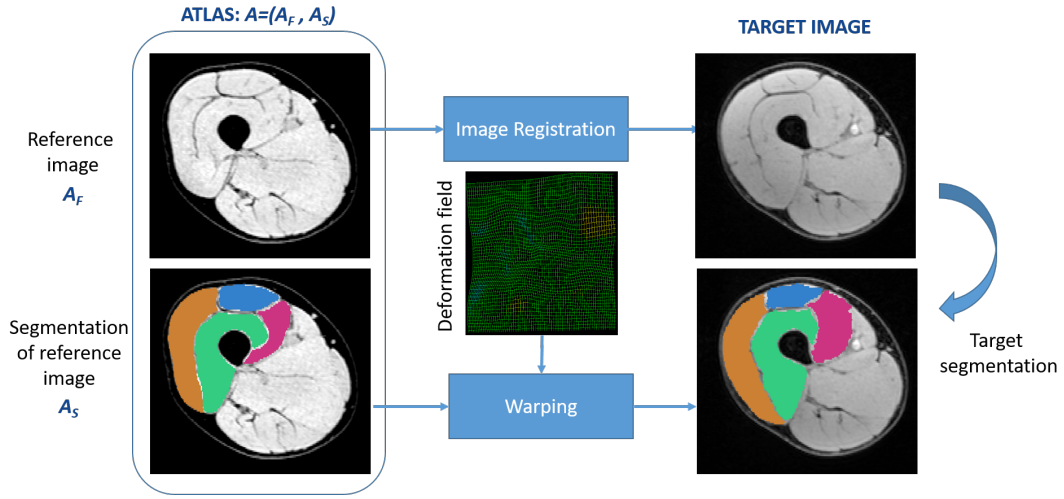

FIGURE 4.1: *Principle of segmentation by image registration. The atlas image (left) is registered on the image to be segmented (right); the resulting deformation field is applied to the atlas segmentation to obtain a segmentation of the target image.*

failed. In order to improve the capability of this method, one approach consists in using several atlases to represent more anatomical diversity and intensity variability.

4.1.3 Multi-atlas segmentation

Multi-atlas segmentation (MAS) is one of the most recent segmentation methods introduced for medical imaging applications (Rohlfing et al., 2004). The two main strategies to take into consideration different atlases are: either to find and use only the atlas that is most representative of the image to be segmented or to merge the segmentations obtained with the different atlases.

In the case that involves merging segmentations produced with different atlases, for each atlas, we apply the procedure as in the mono-atlas method to acquire a segmentation of the target image (referred as an *candidate segmentation*). After this step, we have a number of candidate segmentations equal to the number of atlases. These segmentations should be merged to produce one final segmentation.

Label fusion is the core step in MAS which can define the accuracy of the method. The simplest fusion method is *major voting*, which chooses, for each pixel, the most frequent label suggested by the candidate segmentations. This method is likely to have low accuracy when the target image is very different from the atlases as it ignores the image intensity information in voting.

An extension of major voting is *weighted voting*. Global or local weights can be distributed to favorite the candidate segmentation derived from the training images the most similar to the test image. In general, global weighting (Artaechevarria et al., 2008) is not adapted to the spatial anatomical variations in medical image registration. There are many label fusion methods based on local or semi-local weighted voting that make use of local intensity (local cross-correlation (Artaechevarria et al., 2009), local mutual information (Nie and Shen, 2013), local intensity difference (Išgum et al., 2009)), local features (Kasiri et al., 2014) and many other different metrics (Ramus et al., 2010; Tamez-Peña et al., 2012; Depa et al., 2010). However, optimal weight metrics remain unclear.

In most cases, weights are distributed independently for each atlas, not considering errors produced by correlated atlases. Wang et al. (2013) has developed the *joint label fusion* method that took into consideration the structural correlation between the atlases to minimize expected error. We have adopted this approach of Wang et al. that includes a posterior correction algorithm (Wang and Yushkevich, 2013) based on an automatic learning process. This approach has been used in particular to segment regions of the brain and leg muscles of dogs in MRI scans.

4.1.4 Multi-atlas segmentation with joint label fusion and corrective learning

This approach involves two consecutive steps: joint label fusion (JLF) Wang et al. (2013) and an *optional* step of correction by automatic learning (CL) Wang et al. (2011). These two steps are described below and are illustrated in Figure 4.2.

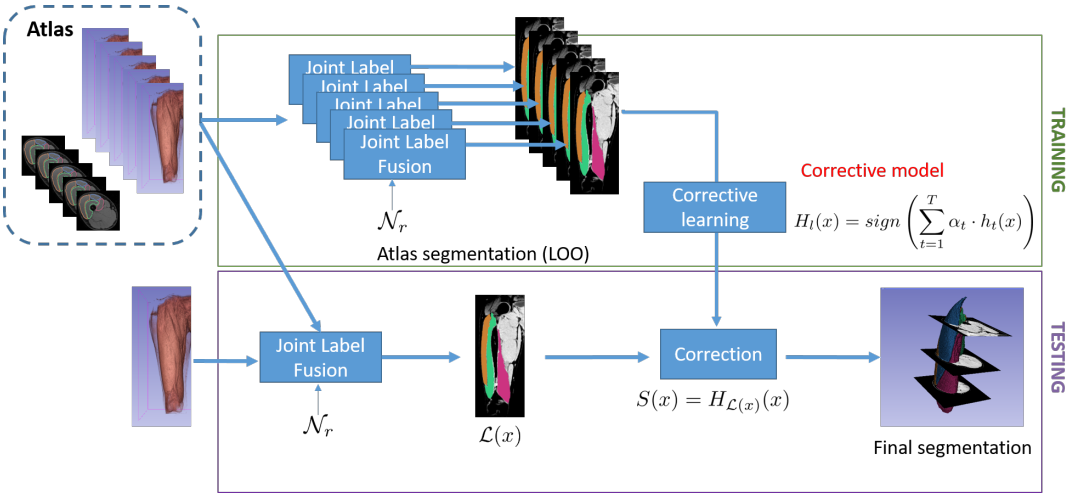

FIGURE 4.2: *Principle of the multi-atlas segmentation method with JLF+CL Wang and Yushkevich (2013). This approach has two parts: the joint label fusion (JLF) then a corrective learning (CL) part. For the latter, a training step is necessary.*

4.1.4.1 Joint label fusion

To remedy the problem of label fusion in muli-atlas segmentation, Wang et al. (2013) proposed a strategy derived from weighted voting where the problem is the optimization of weights to minimize the error made on the segmentation of the target image. However, since this error is unknown, Wang et al. (2013) propose to estimate it using the similarity of intensities in the proximate neighborhood of pixels, which addresses the problem of noisy segmentation results.

This method, which merges labels using both spatial domain and grayscale similarity, is named *Joint Label Fusion* and abbreviated to JLF. The JLF algorithm produces, for each label and at each pixel, a probability that will allow a consensus vote to be taken. In the end, the algorithm determines for each pixel the label that received the highest probability. The main difference between JLF and other label fusion methods is that it considers the correlation between atlases when calculating weight maps. Thus, in the extreme case where an atlas is included twice, this approach will lead to the same result as if it had been included only once.

Let T_F be an image to be segmented and $A^1 = (A_F^1, A_S^1), \ldots, A^n = (A_F^n, A_S^n)$ the n mono-atlas segmentation results of T_F based on n available atlases, A_F^i is the i^{th}

warped atlas image registered on T_F and A_S^i the corresponding candidate segmentation.

For each pixel $\mathbf{x}$, it is possible to model the segmentation error for the label l (where $l \in [1, L]$ and L is the number of labels) by:

$$\delta_l^i(\mathbf{x}) = \mathcal{I}[T_S(\mathbf{x}) = l] - \mathcal{I}[A_S^i(\mathbf{x}) = l] \tag{4.7}$$

where $\mathcal{I}[.]$ represents the indicator function (which is 1 if the condition is true, 0 otherwise) and thus $\delta^i(x)$ can take only three values $\delta^i(x) \in \{-1; 0; 1\}$ and $T_S(\mathbf{x})$ represents the unknown segmentation to be obtained. The distribution of this error for n atlas can be written:

$$q_l^i(\mathbf{x}) = p(|\delta_l^i(\mathbf{x})| = 1 \mid T_F, A_F^1, \ldots, A_F^n) \tag{4.8}$$

To produce the consensus segmentation $\overline{S}$, the weighted voting strategy is used. In the binary case, this weighting is formulated as :

$$\overline{S}(\mathbf{x}) = \sum_{i=1}^{n} w^i(\mathbf{x}) A_S^i(\mathbf{x}) \quad \text{with} \quad \sum_{i=1}^{n} w^i(\mathbf{x}) = 1. \tag{4.9}$$

The goal is to determine the weights w^i that minimize the mean error between the proposed segmentation $\overline{S}$ and the reference segmentation T_S:

$$\begin{aligned} & E_{\delta^1(\mathbf{x}),\ldots,\delta^n(\mathbf{x})} \left[\left(T_S(\mathbf{x}) - \overline{S}(\mathbf{x})\right)^2 \,\middle|\, T_F, A_F^1, ..., A_F^n \right] \\ & = E_{\delta^1(\mathbf{x}),\ldots,\delta^n(\mathbf{x})} \left[\left(\sum_{i=1}^{n} w^i(\mathbf{x}) \delta^i(\mathbf{x}) \right)^2 \,\middle|\, T_F, A_F^1, ..., A_F^n \right] \\ & = \sum_{i=1}^{n} w^i(\mathbf{x}) \sum_{j=1}^{n} w^j(\mathbf{x}) E_{\delta^i(\mathbf{x}),\delta^j(\mathbf{x})} \left[\delta^i(\mathbf{x}) \delta^j(\mathbf{x}) \,\middle|\, T_F, A_F^1, ..., A_F^n \right] \\ & = \mathbf{w}_\mathbf{x}^T \mathbf{M}_\mathbf{x} \mathbf{w}_\mathbf{x} \end{aligned} \tag{4.10}$$

where $\mathbf{w}_\mathbf{x}$ is the vector $\left[w^1(\mathbf{x}); \ldots; w^n(\mathbf{x})\right]$ and $\mathbf{M}_\mathbf{x}$ is the matching matrix of atlases i and j.

From the known $\mathbf{M}_\mathbf{x}$, the optimal weights $\mathbf{w}_\mathbf{x}^*$ are determined by the minimization :

$$\mathbf{w}_\mathbf{x}^* = \underset{\mathbf{w}_\mathbf{x}}{\operatorname{argmin}}\, \mathbf{w}_\mathbf{x}^T \mathbf{M}_\mathbf{x} \mathbf{w}_\mathbf{x} + \alpha \|\mathbf{w}_\mathbf{x}\|_2 \tag{4.11}$$

where α is a coefficient of the regularization term that constrains the $\mathbf{w}$ dynamic, whose value is generally set at 0.1 (Wang and Yushkevich, 2013).

Wang proposed to compute the elements of $\mathbf{M}_\mathbf{x}$ using the local similarity of the warped atlas images i and j with the image to be segmented T_F. That means:

$$\begin{aligned} \mathbf{M}_\mathbf{x}(i,j) &= E_{\delta^i(\mathbf{x}),\delta^j(\mathbf{x})} \left[\delta^i(\mathbf{x}) \delta^j(\mathbf{x}) \,\middle|\, T_F, A_F^1, ..., A_F^n \right] \\ &= p\left(\delta^i(\mathbf{x}) \delta^j(\mathbf{x}) = 1 \,\middle|\, T_F, A_F^1, ..., A_F^n \right) \end{aligned} \tag{4.12}$$

can be formulated with A_F^i and A_F^j and be assumed independent to the other atlases:

$$\mathbf{M}_\mathbf{x}(i,j) = p\left(\delta^i(\mathbf{x})\delta^j(\mathbf{x}) = 1 \,\middle|\, T_F, A_F^i, A_F^j\right) \tag{4.13}$$

Now assuming that pixels far from $\mathbf{x}$ have no influence on this probability, then the $\mathbf{M}_\mathbf{x}$ element (i,j) can be expressed only on $\mathcal{N}(\mathbf{x})$, the neighborhood. We have then:

$$\begin{aligned}\mathbf{M}_\mathbf{x}(i,j) &= p\left(\delta^i(\mathbf{x})\delta^j(\mathbf{x}) = 1 \,\middle|\, \left\{T_F, A_F^i, A_F^j | \mathbf{y} \in \mathcal{N}(\mathbf{x})\right\}\right) \\ &\propto \left[\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} |T_F(\mathbf{y}) - A_F^i(\mathbf{y})|.|T_F(\mathbf{y}) - A_F^j(\mathbf{y})|\right]^\beta\end{aligned} \tag{4.14}$$

with β, a parameter of the model which will be classically set to 2, larger β will put more weight on the most similar patch.

In the multi-class case, the average error is generalized as:

$$\begin{aligned}&E_{\delta^1(\mathbf{x}),\ldots,\delta^n(\mathbf{x})}\left[\left(\mathcal{I}[T_S(\mathbf{x}) = l] - \sum_{i=1}^n w^i(\mathbf{x})\mathcal{I}[A_S^i(\mathbf{x}) = l]\right)^2 \middle| T_F, A_F^1, \ldots, A_F^n\right] \\ &= \mathbf{w}_\mathbf{x}^t \mathbf{M}_\mathbf{x} \mathbf{w}_\mathbf{x}\end{aligned} \tag{4.15}$$

and $\mathbf{M}_\mathbf{x}$ becomes:

$$\mathbf{M}_\mathbf{x}(i,j) \sim \left[\langle |A_F^i(\mathcal{N}(\mathbf{x})) - T_F(\mathcal{N}(\mathbf{x}))|, |A_F^j(\mathcal{N}(\mathbf{x})) - T_F(\mathcal{N}(\mathbf{x}))| \rangle\right]^\beta \tag{4.16}$$

with $\langle .,. \rangle$ the scalar product and $|A_F^i(\mathcal{N}(\mathbf{x})) - T_F(\mathcal{N}(\mathbf{x}))|$ the vector of the absolute deviations, on the patch $\mathcal{N}$ centered in $\mathbf{x}$, of the intensities between the image to be segmented T_F and the warped atlas images A_F^i and A_F^j.

The context around $\mathbf{x}$ in the image to be segmented may not correspond perfectly to those in the resized atlases. In order to improve the estimation of $\mathbf{M}_\mathbf{x}$, several patches $\mathcal{N}(\mathbf{x}+\epsilon)$ centered at different positions ϵ around $\mathbf{x}$ will be tested. The one that minimizes the error between $A_F^i(\mathcal{N}(\mathbf{x}+\epsilon))$ and $T_F(\mathcal{N}(\mathbf{x}))$ will be kept for the calculation of $\mathbf{M}$ and $A_S^i(\mathcal{N}(\mathbf{x}+\epsilon))$ for the weighted voting. The largest value of ϵ for testing is defined by the radius of the search neighborhood that we note $\mathcal{N}_r$.

To make the algorithm more robust in the case of images with different intensity dynamics, which is common in MRI, the patches are normalized before being compared. The size of the patches $\mathcal{N}$ and of the research neighborhood $\mathcal{N}_r$ depend on the size of the structures to be segmented. In our case, we set the size of the image patches $\mathcal{N}$ to 5x5x5 pixels and the size of the search neighborhood $\mathcal{N}_r$ to 8x8x8 pixels.

The probability of having the label l at the position $\mathbf{x}$ for the image to be segmented T_F is:

$$p(l|\mathbf{x}, T_F) = \sum_{i=1}^n \mathbf{w}_\mathbf{x}^i \mathcal{I}(A_S^i(\mathbf{x}) = l) \tag{4.17}$$

To obtain a segmentation, all that remains to be done is to determine the most likely label for each $\mathbf{x}$.

4.1.4.2 Corrective learning

The algorithm of *Corrective Learning* (CL) aims to detect and then correct segmentation errors systematically committed by an automatic segmentation algorithm. Unlike random errors, which are due to noise or random anatomical variations, systematic errors are predictable given a set of conditions (e.g., shape, location, organ, and intensity in the image). The source varies: small variations in anatomical definitions, discontinuities in manual segmentations, or biases between a priori knowledge included in automatic methods and the data to be segmented.

A correction approach, based on machine learning, is proposed in Wang et al. (2011). Applied to MRI imaging, a 20% to 70% decrease in the number of poorly segmented pixels is observed on four automatic segmentation algorithms, including JLF. In the following, this correction algorithm for JLF is explained, starting with the learning phase.

Each atlas will be segmented by JLF as if it were an image to be segmented. However, the JLF will only rely on the other available atlases to perform this segmentation. Since the expert segmentation is available for this atlas under test, it will be possible to determine the errors made by JLF. Applying this for all atlases (Leave-One-Out protocol) makes it possible to build a database to learn systematic errors made by JLF and establish a correction model for each label.

To consider only the areas that potentially need to be corrected, a working region is created for each label. This region is associated with the label of interest morphologically dilated with a structural element of radius r_d.

For each pixel $\mathbf{x}$ in this working region, a feature vector $\mathcal{F}(\mathbf{x})$ is extracted, which covers the relative spatial position, characteristics of appearance (intensity value) and segmentation context in a neighborhood $\mathcal{X}$ of size N_f. The relative spatial position is the relative coordinate of $\mathbf{x}$ to the barycenter of the working region. The appearance and contextual characteristics are also reduced by each of the spatial components in order to increase the spatial correlation. Wang proposed a neighborhood of size $5 \times 5 \times 5$ and thus obtains a feature vector $\mathcal{F}(\mathbf{x})$ in $\mathbf{x}$ of $3 + 125 + 125 + 3 \times 125 = 1003$ dimensions.

All pixels in this region are used for training, and the training dataset will be constructed by the pairs $\mathcal{F}(\mathbf{x}, \mathbf{x}_l)$ where $\mathbf{x}_l$ corresponds to the label of the pixel $\mathbf{x}$. The learning algorithm used by Wang is the AdaBoost binary classification algorithm (Freund and Schapire, 1996), a precise description of which is given in Zhou (2012) and explained in the appendix C. It will produce, for each label and each pixel, a correction model.

When correcting a new image, each pixel, previously labeled by the JLF, will be tested using the same working region definition and characteristics. The pixel will be reassigned to the label whose correction model has given the highest confidence.

4.1.4.3 Perspectives

An assessment of the adaptation capability of this method on our dataset will be presented in Chapter 7. Briefly, the method gives highly accurate segmentations with an average DSC of **0.918** on 7 subjects with manual segmentation. However, since our image volume is much larger than images in the brain dataset, for which this method was initially built, the computational time is exceptionally high, with inference time around 50 hours for an image. Moreover, the large morphological variation in our

dataset also causes an unbalance in the results with the lowest DSC only at 0.859. To address these limitations, we opt to study another family of methods that have been received much attention from the research community: *Deep Learning*.

4.2 Deep learning for image segmentation

In recent years, the domain of image processing has witnessed an explosion of deep learning approaches, or also called *Artificial Neural Networks* (ANN), the result of the confluence of the *big data* and *computer vision* communities with the democratized utilization of GPU-type computing resources.

Traditional machine learning classifiers regroups the problem of features engineering and decision model optimization, which are two separate complicated steps with, at the same time, a complicated relationship. The ANNs have *simplified* this problematic by fusing these two steps into one model, which learned automatically and iteratively image features using back-propagation of the prediction error. Concerning the segmentation of medical images by ANNs, the different approaches and difficulties are widely discussed in the works of Rizwan I Haque and Neubert (2020) and Tajbakhsh et al. (2020).

4.2.1 Convolutional Neural Networks

Since the breakthrough of Lecun et al. (1998), Convolutional Neural Networks (CNNs) strived through as the most popular type of ANNs in image processing with numerous applications, among which is the record-breaking result on the ImageNet dataset (Krizhevsky et al., 2012). The classical CNN architecture is presented in Figure 4.3. A CNN usually consists of three main type of layers, the convolution followed by an activation function, then a subsampling (or pooling) of feature maps.

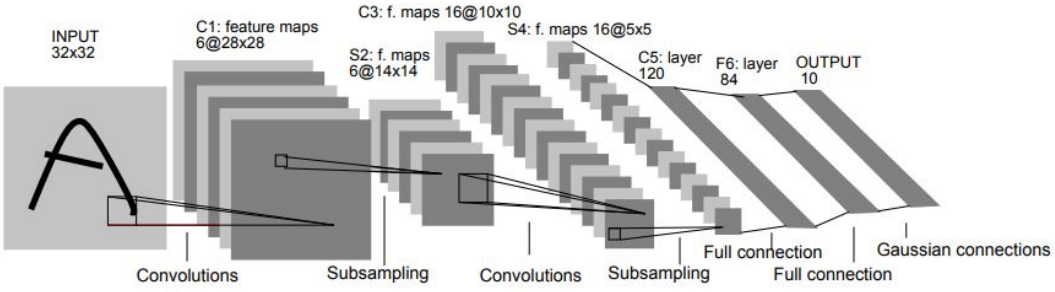

FIGURE 4.3: *CNN architecture by Lecun et al.*

4.2.1.1 Convolution

The *convolution* layer uses kernels of type $\mathcal{K}$, also called filters, that perform convolution operations as it is scanning the input $\mathcal{I}$ with respect to its dimensions. In two-dimensional case (2D), it can be formulated as:

$$(\mathcal{I} * \mathcal{K})(i,j) = \sum_a \sum_b \mathcal{I}(a,b)\mathcal{K}(i-a, j-b) \tag{4.18}$$

A simple illustration of a 2D convolution operation is presented in Figure 4.4. The output is called a feature map. Along each axis, the size of the feature map

is slightly smaller than the input size. Since the kernel has its width and height greater than one and the operation can only be computed where the kernel fits wholly within the input, the size $w_{\mathcal{F}} \times h_{\mathcal{F}}$ of the feature map $\mathcal{F} = \mathcal{I} * \mathcal{K}$ are given by:

$$w_{\mathcal{F}} \times h_{\mathcal{F}} = (w_{\mathcal{I}} - w_{\mathcal{K}} + 1) \times (h_{\mathcal{I}} - h_{\mathcal{K}} + 1) \tag{4.19}$$

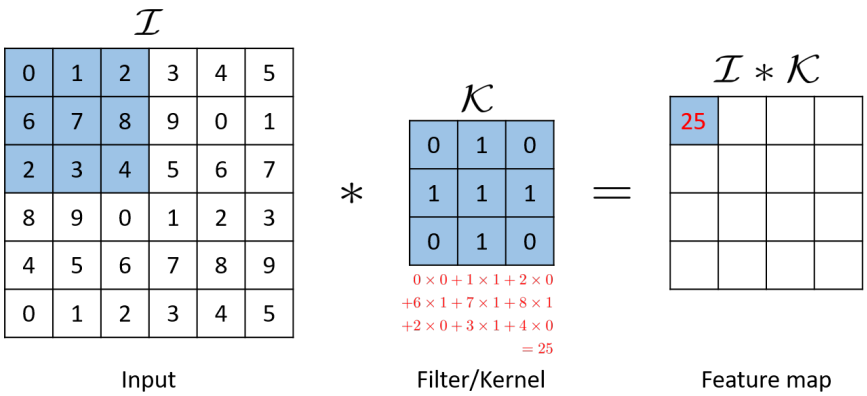

FIGURE 4.4: *Two-dimensional convolution operation with a kernel of size* 3×3*. The output is called a feature map. The elements used to compute the first output element (in red) are shaded in blue.*

To make the size unchanged, a conservative padding can be applied to the input image by adding zeros around its boundary so that there is enough space to fit the kernel when the center of the kernel overlaps the first element of the original input.

It is possible to stack multiple convolution layers by repeating the convolution operation on the resulted feature maps to obtain *deeper* feature maps. Since each filter scans the input in parallel and independently, the number of filters used corresponds to the number of feature maps, i.e., the number of output channels. With the same number of filters, for multi-channel input, the number of output channels remains the same as filters usually mix information from all the input channels. The number of learnable parameters of a convolution layer would be:

$$|\mathbf{w}| = (w_{\mathcal{K}} \times h_{\mathcal{K}} \times |\mathcal{I}| + 1) \times |\mathcal{K}| \tag{4.20}$$

where $|\mathcal{I}|$ is the number of kernels of the input image and $|\mathcal{K}|$ is the number of kernels used in this convolution layer. The number 1 added corresponds to the bias term for each filter.

4.2.1.2 Activation function

Activation function is used to determine the mapping between the input and the output of a layer or of the entire neural network. Traditionally, the value given is either 1 or 0 corresponding to the neuron being activated or not, hence the name activation function. To adapt to the complexity of ANNs and the variability of training data, nonlinear functions are used for most problems. The three most popular nonliear activation function are *sigmoid*, *hyperbolic tangent* and *Rectified Linear Unit* (ReLU) (Nair and Hinton, 2010). These functions are illustrated in Figure 4.5 and formulated as in Equation 4.21, 4.22, and 4.23 respectively.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.21}$$

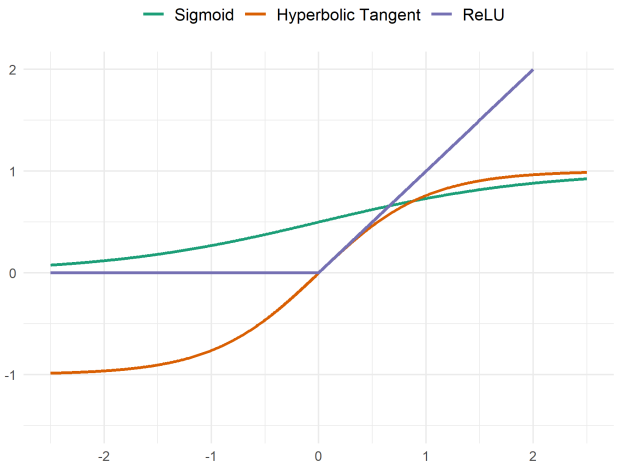

FIGURE 4.5: *Function curves of sigmoid, hyperbolic tangent and ReLU.*

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{4.22}$$

$$\text{ReLU}(x) = \max(0; x) \tag{4.23}$$

Since hyperbolic tangent can be reformulated from the sigmoid function, their performance is similar. Both functions are mainly used for binary classification problems. The disadvantage of these functions is that they tend to saturate, which causes a phenomenon called *vanishing gradient* where the gradient becomes vanishingly small, preventing the weights from changing and making the model stuck at the training time. *ReLU* function was built to resolve this problem (Nair and Hinton, 2010) and, at the same time, accelerate the convergence. Despite the possibility of activating *exploding gradient* problem, where, because of the inappropriate mapping of negative values, the model cannot fit or train from the data properly, ReLU in combination with Adam optimizer (Ruder, 2016) (Sec. 4.2.1.5) is used in almost all the recent CNNs.

For multiclass classification, an activation function called *softmax* is often appended to the last layer of the network. It is formulated as follow:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j^L e^{x_j}} \tag{4.24}$$

where L is the number of classes/labels. This returns, for each pixel of the input image, a vector of values between 0 and 1 whose sum is equal to 1. Each value can be interpreted as the probability that a pixel belongs to a label; we then distribute the label with the highest probability to that pixel.

4.2.1.3 Pooling

The pooling layer, also called downsampling or subsampling, is usually the last layer of a convolution block. It transforms the output of the activation function by scanning the feature maps patch by patch and distributing a single statistical value for each patch, often maximal, minimal, or average value of all the values in the patch. This operation progressively reduces the spatial resolution of feature maps, thus

reducing computation time and memory consumption, keeping compact representative information, and providing invariance to translation. The most common pooling operation is *MaxPooling* which extracts the largest value of a given image patch to represent it.

4.2.1.4 Loss function

A network needs to be optimized accordingly to a *loss function* (also called cost function) so that its parameters are updated iteratively to search for the minimal value of this function. The most frequent loss functions for multiclass segmentation are *categorical cross-entropy* (CCE) and *derivable multiclass DICE loss* (DDL) (Milletari et al., 2016). While the CCE evaluates directly each pixel in multiple-class manner, the DDL separates the problem into multiple binary classifications. The two loss functions are formulated in Equation 4.25 and 4.26, respectively.

$$\mathcal{L}_{CCE}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{i}^{L}\sum_{j}^{N} y_j^{(i)} \log \hat{p}_j^{(i)} \tag{4.25}$$

$$\mathcal{L}_{DDL}(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \sum_{i}^{L}\sum_{j}^{N} \frac{2\hat{p}_j^{(i)} \text{onehot}^{(i)}(y_j)}{(\hat{p}_j^{(i)})^2 + \text{onehot}^{(i)}(y_j)} \tag{4.26}$$

where L is number of labels, N is the size of the output, $\mathbf{y}$ is the ground truth, $\hat{\mathbf{y}}$ is the prediction by model, $\hat{p}_j^{(i)}$ is the predicted probability that the j^{th} element of $\hat{\mathbf{y}}$ correspond to the label i. The one-hot encoding of y_j based on label i consists of transforming the value of this element into 1 if $y_j = i$ and into 0 otherwise.

4.2.1.5 Optimization of parameters

To improve the opitimisation of multi layers networks, the parameters $\mathbf{w}$ need to be updated in respect to the error on the loss function $\mathcal{L}$. The iterative evolution of $\mathbf{w}$ is done by gradient descent:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \tag{4.27}$$

with η the learning rate, which determines how much the weights change after each iteration t. This approach converges to a solution when the gradient becomes close to zero.

Back-propagation Introduced in Rumelhart et al. (1995), the back-propagation algorithm allows an efficient implementation of the calculation of the gradient, based on the chain rule. The chain rule allows computing the gradient of each layer to update the parameters of each layer independently. The step of updating the parameters is determined by an optimization algorithm (i.e., optimizer).

Stochastic gradient descent The stochastic gradient descent (SGD) (Ruder, 2016) is a gradient descent optimization technique, adapted to the neuron networks for supervised learning problems with a large database. The SGD hypothesizes that the gradient can be approximated using only one data point to reduce computation time enormously. It is also common to use a small number of data points (mini-batches) instead of one to denoise the gradient. The algorithm proposes to sample without replacement, at each iteration, a set of mini-batches. The larger the size

of the mini-batches, the more the variance of the parameter updates are reduced under the effect of averaging gradients. Including a momentum β to the process can accelerate the convergence even more with the velocity adapted to the slope of the landscape depicted by the loss of the training set. Another way of formulating the standard gradient descent in Equation 4.27 is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{v}_{t+1} \quad \text{with} \quad \mathbf{v}_{t+1} = \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \tag{4.28}$$

when adding a momentum, $\mathbf{v}_{t+1}$ becomes:

$$\mathbf{v}_{t+1} = \beta \mathbf{v}_t + \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \tag{4.29}$$

Adaptive Moment Estimation Adaptive Moment Estimation (Adam) (Kingma and Ba, 2015) is the most popular optimization method recently proposed. Adam intends to improve the gradient momentum while adapting the learning rate to the magnitude of gradients. Considering the gradient of the cost function of a neural network as random variable, its first moment is the mean:

$$m_t = \gamma_1 m_{t-1} + (1 - \gamma_1) \frac{\partial \mathcal{L}}{\partial w} \tag{4.30}$$

and its second moment is the uncentered variance:

$$v_t = \gamma_2 v_{t-1} + (1 - \gamma_2) \frac{\partial \mathcal{L}^2}{\partial w} \tag{4.31}$$

where γ_1 and γ_2 are two newly introduced hyper-parameters of the algorithm. The values of these moments are then corrected to avoid bias to zero:

$$\hat{m}_t = \frac{m_t}{1 - \gamma_1^t} \qquad \hat{v}_t = \frac{v_t}{1 - \gamma_2^t} \tag{4.32}$$

The model weight update can be performed as follow:

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{4.33}$$

4.2.1.6 Regularization

In machine learning, one of the most common problems is *overfitting*, where the model performs exceptionally well on training data but cannot predict expected results on test data. Assuming sufficient training data, the more complex the model, the more prone it is to overfitting, hence addressing this problem in deep learning with regularization methods.

Early stopping When training a supervised model, one usually looks at the evolution of the values of the loss function on the training data and a set of validation data. One can then observe the improvement in performance over iterations and stop when a plateau is reached. Meanwhile, if the model starts to overfit, the loss on the validation data will increase. The most straightforward method to avoid this problem is to stop the training process after a certain number of iterations without improvement on validation data and conserve the weights of the iterations with the best performance on validation.

Weight penalization Adding a constraint of the parameters is a classical approach to limit the overfitting phenomenon of a supervised model. L1 and L2 norms on w are the most common: They update the general cost function by adding a regularization term on the network parameters. The term differs in L1 and L2, and it has a specific effect on the evolution of the parameters, either in terms of parsimony or amplitude.

Dropout The dropout layer (Srivastava et al., 2014) is a simple method of regularization, which limits the overfitting by encouraging the activity of a random partition of the weights, rather than centralizing the influence of the prediction on the same set of parameters. During the training phase, the dropout layer cancels a specific amount of neurons of its preceded layer to promote the creation or the development of other features.

Data augmentation Another simple but effective way to prevent overfitting is to increase the size and the variability of the training data. While adding annotated data can be too costly, data augmentation generates new training data, either based on the existing training data or a specific model. Depending on the problem to be solved, the simulation may be more or less complicated.

Batch Normalization While mini-batch is indispensable when training large networks on a GPU with limited memory, the distribution of inputs to network layers may change after weight updates (also called *internal covariate shift*), which makes the training process unstable and slows down the convergence of the loss function. Batch normalization is applied to the activation of a mini-batch to standardize the inputs to the next layer (Ioffe and Szegedy, 2015), making the training of that layer less dependent on the previous one. Considering a mini-batch $\mathcal{B}$ of N values: $\mathcal{B} = \{x_i, i = 1..N\}$, we first normalize all the values in $\mathcal{B}$:

$$x_{i,norm} = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \tag{4.34}$$

with $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ mean and variance of $\mathcal{B}$. The values are then scaled and shifted with 2 learnable parameters γ and β:

$$\hat{x_i} = \gamma x_{i,norm} + \beta \tag{4.35}$$

Batch normalization stabilizes the training process, accelerates the loss convergence, and improves the robustness to high learning rate and random initialization (Santurkar et al., 2018).

4.2.2 UNet architecture

Since its introduction in 2015 by Ronneberger et al., UNet has become one of the most popular CNN architecture for medical image segmentation. The network is based on the encoder-decoder architecture, composed of two distinct parts, an encoder, and a decoder. The former has the role of encoding the visual and semantic features by compressing the representation while the latter progressively reconstructs the feature maps up to the input resolution. UNet distinguishes itself by the use of skip connection at each resolution (Fig. 4.6), which transmits the feature maps

from the encoder block to the corresponding decoder block, in order to improve the localization of high-level features.

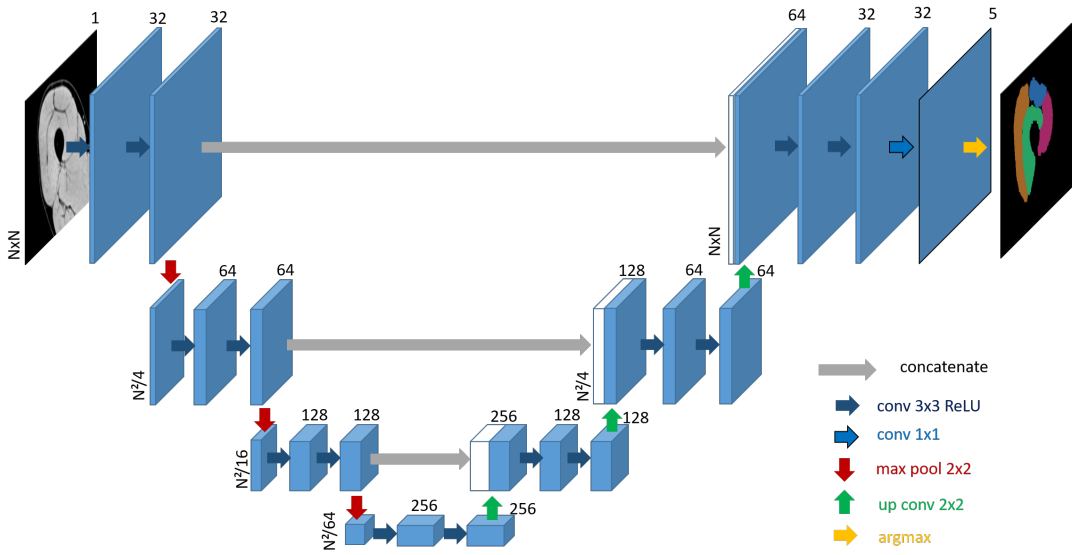

FIGURE 4.6: *Example of a 4-levels UNet architecture with 32 feature maps at the first level.*

Many variants have been proposed to improve the performance of UNet. They often consist of a modification of the architecture and the proposal of a cost function adapted to the learning of this new architecture. Four types of improvements can be distinguished:

The original UNet is two-dimensional (2D); while it can be applied to 3D images by separating the volumes into stacks of 2D slices, the network has difficulties integrating spatial contexts along the last axis. The first type of improvement consists of reintegrating these contexts into the network. While 3D networks, such as 3D UNet (Çiçek et al., 2016) and V-Net (Milletari et al., 2016), suffer from high computational costs, more practical approaches were proposed, among which are Li et al. (2019)'s Z-Net (new patch division strategy and separation of 3D convolution into 2D then 1D convolutions), Li et al. (2019)'s H-DenseUNet (the intra-slice 2D representations and inter-slice 3D features are jointly optimized through a hybrid feature fusion layer), Alkadi et al. (2019)'s 2.5D UNet (upper and lower slices of each input 2D slice are added as supplemental channels), Haque et al. (2019)'s multi-directional UNet (3 differents 2D UNets are trained with slices extracted from 3 different directions then the final segmentation is voted with the *winner takes it all* principle), and Perslev et al. (2019)'s novel weighted multi-directional UNet (similar to Haque et al.'s. However, the weights of UNets are learned automatically).

The second one consists of improving the encoder in particular by taking inspiration from classification network architectures. For example, ResUNet (Zhang et al., 2018) is based on residual blocks (introduced in the classification network ResNet (He et al., 2016)) for its encoder part and uses the mean square error as a cost function. The residual blocks facilitate training by allowing the layers to model the residuals and not the complete model. The gradient propagation is also less attenuated by these blocks, making it possible to create extremely deep networks (more than 100 layers) and thus increase the network's capacity to acquire high-level concepts. An extension of ResUNet, ResUNet++ (Jha et al., 2019) adds, among other things,

the use of attention units to ResUNet. These units determine which part of the data should have more attention and reduce the computing time.

The third type of improvement is to increase the frequency of skip connections. Notably, UNet++ (Zhou et al., 2020) and UNet3+ (Huang et al., 2020) propose to redesign skip connections for aggregating features at varying scales. UNet++ introduces the focal loss and redesigns skip connections with dense connectivity to allow better optimization and attain lower validation loss, while UNet3+ tries to explore more information from full scales and via a hybrid loss function that captures both fine and large scale structure with clear boundaries. Both UNet++ and UNet3+ take advantage of deep supervision (Lee et al., 2015) to learn hierarchical representation from aggregated full-scale features.

The last type is using a cascade of networks such as the Stacked Hourglass Model (SHG) (Newell et al., 2016; Vigneault et al., 2018), which integrates a succession of several encoder-decoder networks, which could be UNet, in a single large network. The first sub-networks are used as residual blocks, i.e., the input of a subnetwork is the concatenation of the previous segmentation and the previous input. Each subnetwork output is associated with an intermediate objective segmentation according to a deep supervision strategy, which, combined with the residual connections, forces subnetworks to learn to refine the previous segmentation.

CHAPTER 5

Muscle segmentation from MR Images

In this chapter, we remind the readers about the main difficulties of our task, which is the segmentation of muscles from MR images, and review some of the existing popular approaches in medical image segmentation.

5.1 Segmentation challenges

Segmentation of the muscles from MR images has always been a challenge due to the lack of defined muscle boundaries, intensity inhomogeneity, bias field, or acquisition artifacts (Prescott et al., 2011) (Fig. 5.1). In the case of the MUM runners who participated in the project MUST, professional athletes with highly developed quadriceps muscles and a minimal amount of body fat, most of the time, the determination of muscle boundaries is not apparent even for medical experts. These problems make the use of standard automatic segmentation methods such as thresholding or clustering impractical.

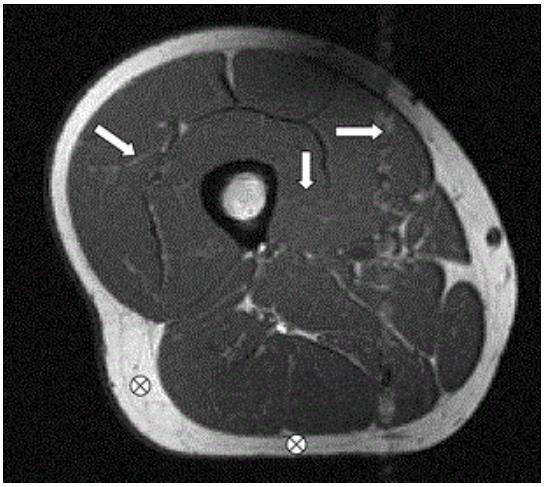

FIGURE 5.1: *Image showing features which cause segmentation difficulties: lack of delineating landmarks (vertical arrow), high contrast intramuscular adipose tissue (diagonal arrow), flow artifacts (horizontal arrow), and bias fields (marks showing low and high intensity subcutaneous fat regions, which should be same intensity). Image extracted from Prescott et al. (2011)*

Even in an apparently homogeneous population as the athletes of Tor des Géants, there is a large morphological variation since the quadriceps is the largest muscle

group in the human body. For instance, the ratio muscle/fat in women is very different from men. It has been shown that pelvis and femur structure is also very different between men and women (Delavier, 2003). Women tend to have a wider pelvis, therefore larger Q-angle (the angle that the femur makes between the hip and the knee) (Fig. 5.2). These differences might need special attention in the next part of our study.

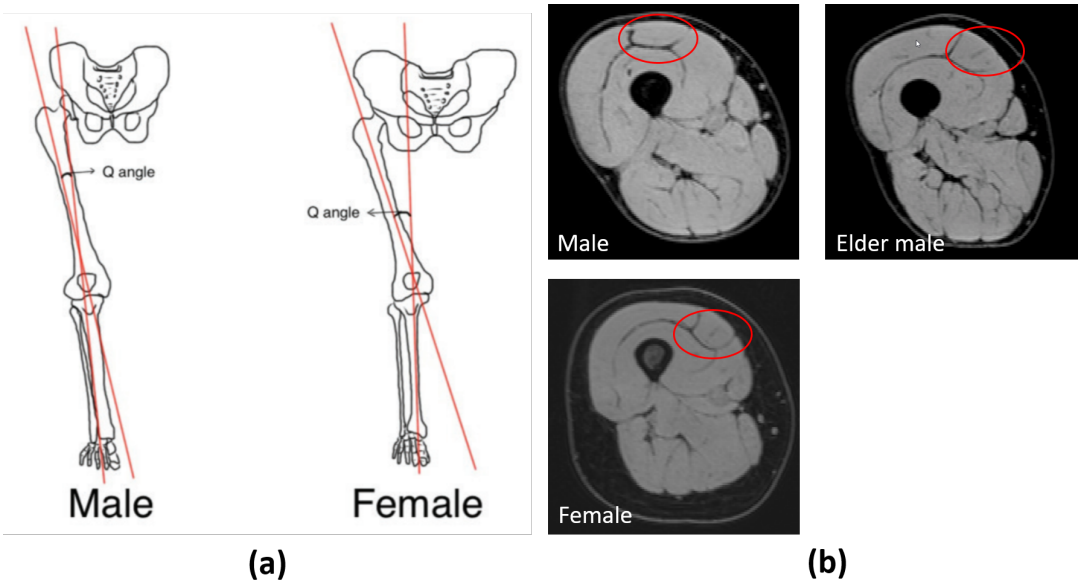

FIGURE 5.2: *(a) Difference between male and female pelvis - femur structure. Women have wider pelvis, therefore larger Q-angle (Canbolat et al., 2018). (b) Images of a young male (ALB-2725), an older male (ANG-2014), and a female (CAL-4223) runner in MUST dataset, red circles point to the rectus femoris (RF) muscle head.*

We can also see in Figure 5.2 that the older male (ANG-2014) and the female (CAL-4223) have similar muscle distribution: the RF is more on the right compared with the other runners. ANG-2014 is the oldest runner in the competition; at the time of the race, other runners were around 30-40 years old, while ANG-2014 was 75 years old. He has a smaller muscle volume and is anatomically quite different from the other male runners.

5.2 Human quadriceps segmentation

Some recent studies have addressed the automatic segmentation of quadriceps muscles (Gilles et al., 2016; Prescott et al., 2011; Ahmad et al., 2014; Andrews and Hamarneh, 2015; Le Troter et al., 2016) but none has archived an accurate segmentation at the boundaries of the muscles, which is very important in the quantification of volume change in our longitudinal project.

As mentioned above, we have in hand the results of an automatic segmentation method for the entire dataset (Fig.5.3). The method is based on deformable model and shape-matching with the necessity of a preliminary bone segmentation, initial seed points, and sometimes manually set boundary constraints (Gilles and Magnenat-Thalmann, 2010).

Briefly, an initial model was defined by manually segmenting all quadriceps heads of interest (vastus medialis (VM), vastus lateralis (VL), vastus intermedius

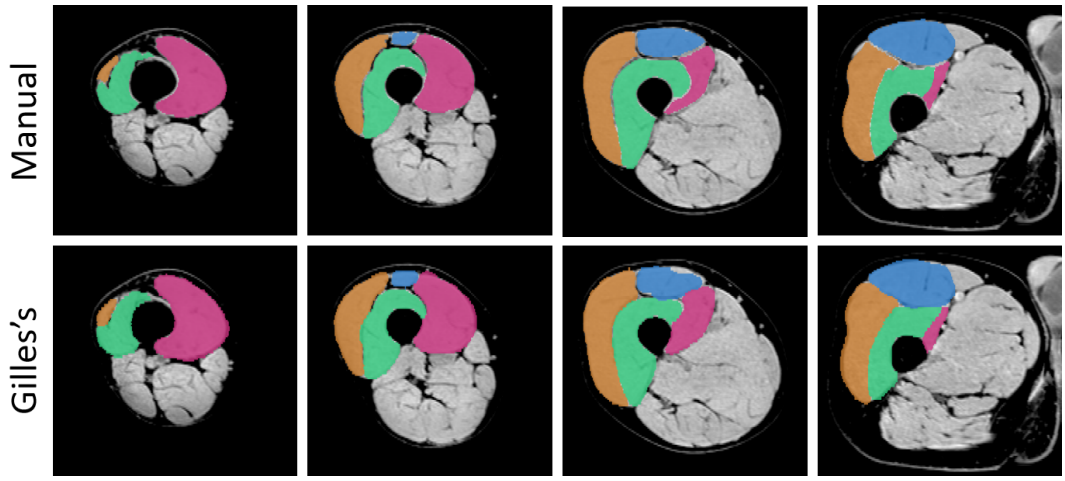

FIGURE 5.3: *Visual comparaison between manual segmentation and a segmentation by Gilles et al. of ALB-2725.*

(VI) and rectus femoris (RF)) and bones (femur, patella and pelvis) from one subject. After conversion to a 3D triangle mesh, this model was considered as a reference template for the registration process. This template was iteratively deformed to match contours in target images from other subjects. The deformation process was driven by external forces to maximize the correlation between reference and target images around the surface, and internal forces to maintain smooth surfaces. The contribution of external forces was iteratively increased to perform a robust coarse-to-fine alignment. For computing image correlation during registration, all four contrast water, fat, IN-phase and OUT-phase images were used. 3D volumes were then computed using the final meshes obtained for each quadriceps head. This process was repeated for all individual subjects enrolled in the study, at all time-points of the race. For all quadriceps head of interest, the best automatic segmentation accuracy was obtained when using the calculated Water image.

The segmentations are not totally accurate, especially at the circumference of the muscles (Fig. 5.3). With some postprocessing, they can still be used as a basis for a statistical study of the image dataset (Nguyen et al., 2021b). However, we are still not satisfied with the overall accuracy of these segmentations as, for now, we cannot quantify, with precision, the muscles volume. The quantitative evaluation of these segmentations is presented in Table 5.1, the results may differ from Gilles et al. (2016) since we have acquired two more manual segmentation since its publication.

Based on this study of Gilles et al. and the observation of our radiologists, the most suitable MR images used for the segmentation task are the Dixon Water only images (see Sec. 2.3.1), here denoted T1W.

Subject	DSC	DSCw	HD (mm)	MAD (mm)	VS
ALB-2725	.913	.914	15.72	1.48	.115
ALF-4529	.897	.898	48.06	1.74	.075
ANG-2014	.778	.792	32.84	4.74	.187
ANS-3229	.901	.903	25.23	1.75	.047
ARS-4026	.915	.917	25.00	1.47	.048
CAL-4223	.756	.772	26.26	4.28	.233
OUK-2927	.890	.895	39.93	2.23	.092
mean±sd	$.864 \pm .067$	$.870 \pm .061$	30.43 ± 10.78	2.53 ± 1.39	$.114 \pm .071$

TABLE 5.1: *Quantitative evaluation of Gilles et al.'s segmentations on 7 subjects with full segmentation of right leg.*

Conclusion

Medical image segmentation has been one of the fastest developing sections of medical image analysis, with methods varying from the most straightforward, such as threshold, to the most complex, such as neural networks. Based on our observation on the state-of-the-art of muscle segmentation, we have decided to lean our next steps on the Wang and Yushkevich (2013)'s multi-atlas segmentation with joint label fusion and corrective learning and the most popular network for image segmentation - UNet (Ronneberger et al., 2015).

In Part III, we will present our evaluation of these segmentation methods on the MUST dataset and our propositions of improvement based on their strengths and weaknesses. Furthermore, in Part IV, the proposed methods are applied to different muscle study projects with further statistical analysis for local functional variation.

PART III

Contributions to MRI muscle segmentation

Contents

Résumé

Dans cette partie, nous présentons nos contributions à la segmentation des quadriceps images par IRM, appliquée explicitement sur la base des données MUST des jambes des athlètes d'ultra-marathon : l'analyse de la méthode de segmentation par recalage multi-atlas, l'application du réseau UNet avec augmentation de données et l'intégration des mesures morphologiques pour optimiser les méthodes de segmentation automatique.

Le chapitre 6 décrit la procédure de prétraitement appliquée aux images avant la segmentation automatique. Il s'agit de correction d'inhomogénéité des signaux IRM, et d'adaptation des jambes gauches sur les jambes droites. On compare aussi les performances de segmentation manuelle des experts afin de définir les objectifs pour nos algorithmes et identifier les difficultés.

Au chapitre 7, on fournit une analyse complète de la méthode de Wang and Yushkevich, une segmentation multi-atlas avec une méthode de vote basée sur la similitude des patchs (fusion d'étiquettes communes - Joint Label Fusion), et une étape d'apprentissage correction d'erreur par AdaBoost. On étudie les impacts des paramètres de la méthode et du nombre d'atlas sur la qualité de la segmentation et le temps de calcul afin de déterminer ses limites.

Nous proposons, dans le chapitre 8, de remplacer l'étape de segmentation multi-atlas de Wang and Yushkevich par un réseau UNet 2D faiblement supervisée, qui est entrainé avec des données annotées manuellement et d'autres générées artificiellement à l'aide de déformation aléatoire. Les résultats sont légèrement meilleurs que ceux obtenues précédemment mais nécessitent un temps de calcul bien inférieur.

Enfin, dans le chapitre 9, nous présentons nos descripteurs morphologiques dédiés à la segmentation des quadriceps. Les descripteurs morphologiques permettent d'améliorer les résultats de la segmentation automatique obtenus par la segmentation multi-atlas avec une approche d'apprentissage correctif utilisant une sélection d'atlas basée sur la similarité morphologique de l'image à traiter. En outre, une stratégie d'augmentation de données basée sur la morphologie est proposée et permet d'augmenter la capacité de généralisation de notre réseau.

Introduction

In this part, we present our contributions to quadriceps muscle segmentation based on MRI data, applied explicitly on the MUST dataset of ultra-marathon athletes' upper legs: the analysis of the interested multi-atlas segmentation method, the application of UNet with data augmentation, and finally, the integration of morphological features to optimize the automatic segmentation methods.

Chapter 6 describes the preprocessing procedure applied to the data before the automatic segmentation. Though it can be used as a reference, this procedure is specific for the MUST data set and might vary when applying to a new dataset.

In chapter 7, we provide a complete analysis of Wang and Yushkevich's method, a multi-atlas segmentation with a patch-similarity-based voting method (joint label fusion), and a step of corrective learning by AdaBoost. We study the impacts of its parameters and the number of atlases on the segmentation quality and computation time in order to identify its limitations.

We propose, in Chapter 8, to replace the multi-atlas segmentation step in the framework of Wang and Yushkevich with a 2D weakly-supervised UNet, which is trained with manually annotated and artificially generated data.

Finally, in Chapter 9, we introduce our morphological features dedicated to quadriceps segmentation. The morphological features help improve the automatic segmentation results obtained by multi-atlas segmentation with a corrective learning approach using a selection of atlases based on morphological similarity to the image to process. Furthermore, a morphology-based data augmentation strategy is proposed with the objective of increasing the generalization capability of our network.

CHAPTER 6

Preprocessing & Manual segmentation

Here we present the preprocessing procedure applied to MR images of the MUST dataset (Sec. 6.1) and how manual segmentations were obtained (Sec. 6.2). The preprocessing procedure is specific to the MUST dataset and needs to be revised when applying to a new dataset.

6.1 Preprocessing

MR sequences were acquired with DICOM format then were converted to NIFTI for faster and simplified manipulation. All the subjects were anonymized and were assigned a subject code.

Each T1W image in the MUST dataset was halved in the middle in coronal view to get 2 separated image volumes of right and left leg. The left leg images were flipped along the coronal axis to resemble the right leg ones. The images were then processed with N4 algorithm to correct bias field (Tustison et al., 2010) and then rescaled to the same intensity range as the alphabetically first image (right leg image of the subject ALB-2725 at time point Pre) (Fig. 6.1).

6.2 Manual segmentation

The first five manual segmentations were done by medical experts using Horos, a specialized DICOM viewer for MacOs exclusively. To facilitate the manipulation and visualization of the segmentation, we have established a manual segmentation protocol using 3DSlicer (Kikinis et al., 2014). With the help of a designer tablet and the built-in manual segmentation tool of 3DSlicer, the task can be done more efficiently and precisely.

Figure 6.2 shows how the right and left legs mentioned in Table 2.1 were segmented. The right leg images of 7 subjects were segmented manually with a step of 10 slices since, with the high resolution of our images, the morphology does not change much within 10 consecutive slices. Each segmentation was then interpolated between slices followed by a precise manual correction by medical experts to obtain a complete 3D segmentation of the image volume of size $280 \times 640 \times 160$. As the correction phase is exhausting, for the 5 left legs manually segmented, we use only 50 segmented slices per subject to validate our automatic segmentation methods.

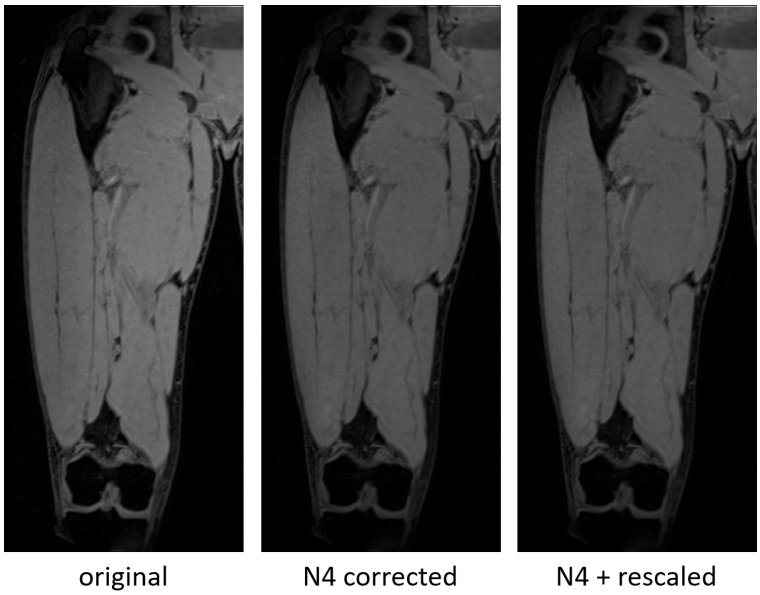

FIGURE 6.1: *Image of a subject in our dataset before and after each step of preprocessing*

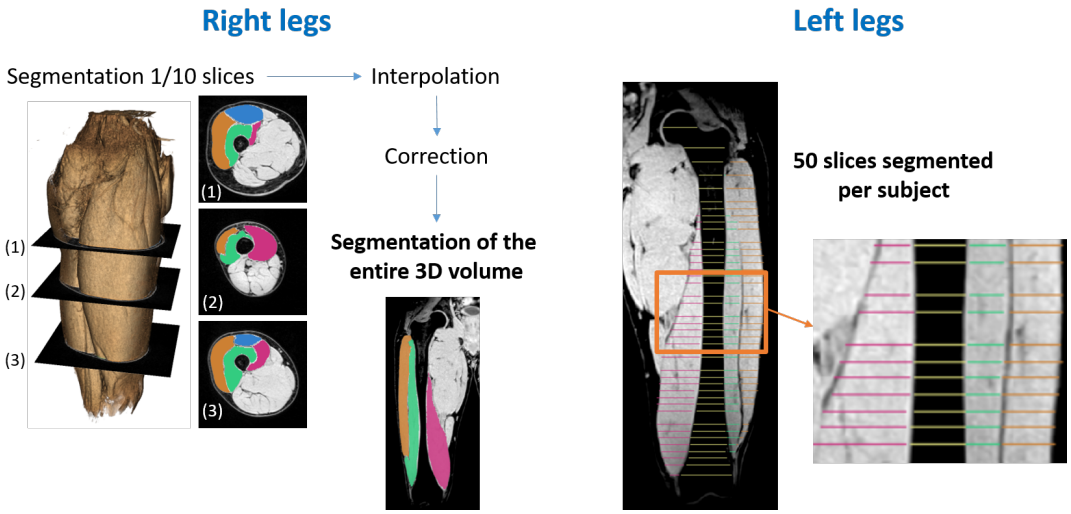

FIGURE 6.2: *Manual segmentation on right and left legs of MUST subjects*

With the time limitation, the longer task, which is the 3D segmentation of the right leg volumes, was assigned to 4 different medical experts, while for each left leg segmented, we have two distinct segmentation by two different experts. The segmentation evaluation metrics between the segmentations done by different medical experts are presented in Table 6.1. Since the segmentation of the left legs is not completed in 3D, only DSC and VS are comparable with 3D segmentation. Furthermore, VS are reported in absolute values as we do not have a fixed reference segmentation in this case.

The difference in the manual segmentations by two medical experts illustrates the complexity of the task (average DSC at 0.91) and the variability of the difficulty level among the muscle heads (DSC varies from 0.87 to 0.94). It also allows us to fix our objective regarding the segmentation quality quantified by the segmentation validation metrics.

	ALB-2725	*BRG-1924*	*CAL-4223*	*MAV-526*	*YAG-47*	**Mean**
DSC	.924	.877	.897	.921	.934	.910
VL	.920	.874	.906	.918	.944	.912
RF	.939	.916	.941	.940	.962	.940
VM	.945	.892	.882	.942	.934	.919
VI	.891	.825	.857	.882	.894	.870
VS	.042	.057	.054	.074	.025	.051
VL	.025	.100	.065	.049	.003	.049
RF	.025	.061	.012	.073	.021	.038
VM	.004	.023	.091	.022	.073	.043
VI	.115	.043	.050	.155	.005	.074

TABLE 6.1: *Segmentation evaluation metrics (DSC & Volume Similarity) between manual segmentations done by 2 different medical experts for the left legs of 5 runners. Muscle head abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius.*

The 7 right leg image volumes were used as references for automatic segmentation methods, with the Leave-One-Out strategy (see Sec. 3.2). They were used as atlases for the multi-atlas segmentation method and contributed to training and validation sets for deep learning methods.

6.3 Conclusion

We present in this chapter the preprocessing of the MUST dataset, which involves the correction of inhomogeneity of the MRI signals and the adaptation of the left legs to the right legs. We also compare the medical experts' manual segmentation performances to define the objectives for our algorithms and identify the difficulties.

The processed images and the manual segmentations will be the materials for our experiments, and the evaluation of the automatic segmentation methods is detailed in the following chapters of this part.

CHAPTER 7

Multi-atlas segmentation with joint label fusion and corrective learning

As detailed in Section 4.1.3, Wang and Yushkevich's multi-atlas segmentation method introduced an intelligent way to combine segmentation results from multiple atlases and to correct systematic segmentation errors, taking into account the neighboring image patch around each pixel. Based on image registration, this method has immense adaptation potential to datasets with large morphological variation while benefiting to the maximum of intensity and spatial information in the image volumes through joint label fusion (JLF) and corrective learning (CL).

This chapter provides a complete analysis of the method: the impacts of its parameters and the number of atlases on the segmentation quality and computation time.

The experiments were implemented using `elastix` (Klein et al., 2010) and `C++/ITK` (Wang and Yushkevich, 2013; Yoo et al., 2002; Tustison et al., 2017) and were run on 16 CPUs for JLF step and 1 CPU for CL step. Since the computation time depends heavily on the quality of the processors to which the task is assigned, for comparison purposes, the time reported here is the average of tasks run on the cluster of CREATIS (see Annex E).

7.1 Parameter optimization

The first study consists of studying the impact of the method's parameters on the segmentation results. Parameters that might have a significant influence are the size of the patches N, the size of the research neighborhood N_r, the regularization parameters α and β, then the radius r_d of label dilation and patch size N_f for features extraction. Multiple values of each parameter were tested by fixing the others at a *reasonable* value. The default value of N was fixed at $5 \times 5 \times 5$ pixels, the size of the smallest structure in the images. N_r was fixed at $8 \times 8 \times 8$ pixels, around 50% larger than N. We also used the default values of α (0.1) and β (2.0) suggested in Wang and Yushkevich (2013). Based on the size of N and N_r, we fixed r_d at 5 pixels and N_f at $8 \times 8 \times 8$ pixels.

All results are reported in average value after Leave-One-Out experiments (see Sec. 3.2) with 7 right leg images with manual segmentation.

7.1.1 Joint label fusion parameters

Regularization terms After the results of our experiments, the regularization terms do not have much effect either on the segmentation quality or on the computation time. We conserve the recommended value of 0.1 for α and 2.0 for β.

Image patch size N & Research neighborhood size N_r These two parameters have a significant impact on the computation time of the JLF step. The computation time increases rapidly with a small increase of each parameter, especially for N. The computation time increases from 50 hours with $N = 5 \times 5 \times 5$ to 200 hours with $N = 8 \times 8 \times 8$ and exceeds 500 hours of calculation with $N = 12 \times 12 \times 12$. Here, we do not report the average segmentation results of our experiments with different radius of N since some of the tests exceed the wall time limit of our computational resources. In the meantime, the finished tests do not show any improvement, if not declination, in segmentation quality.

In the case of the research neighborhood radius N_r, contrary to our expectation that a larger neighborhood would provide more candidates for the most similar patch and result in better classification, the best segmentation quality is obtained with the default value of 8 pixels (Fig. 7.1). Considering the similar image texture in different muscle heads, one possible explanation is that enlarging the research neighborhood size increases the chance of mistaking an image patch from a muscle head with one from another muscle head. A larger neighborhood might necessitate a larger image patch size. Since increasing both neighborhood size and patch size will cause an exponential increase in computation time, we opt for a computationally lighter solution - working with lower resolution (Sec. 7.4).

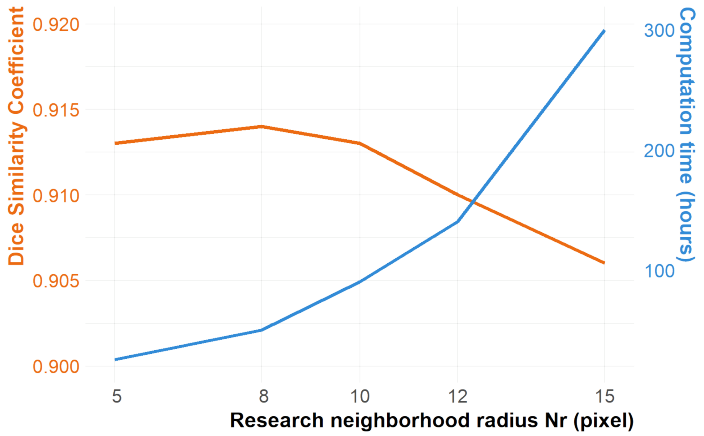

FIGURE 7.1: *Influence of the size of the research neighborhood N_r on the segmentation results and computation time. DSCs reported here are the mean value of 7 Leave-One-Out tests.*

7.1.2 Corrective learning parameters

The results of our experiments with corrective learning parameters are presented in Figure 7.2. With the dilatation radius, the best result is obtained with $r_d = 15$ pixels. The size of image patch for feature learning N_f has more influence on the

computation time than the dilatation radius r_d while having almost no influence on the segmentation quality. After testing $r_d = 15$ with $N_f = 5$ and $N_f = 8$, we decide to keep the value of N_f at 8 pixels.

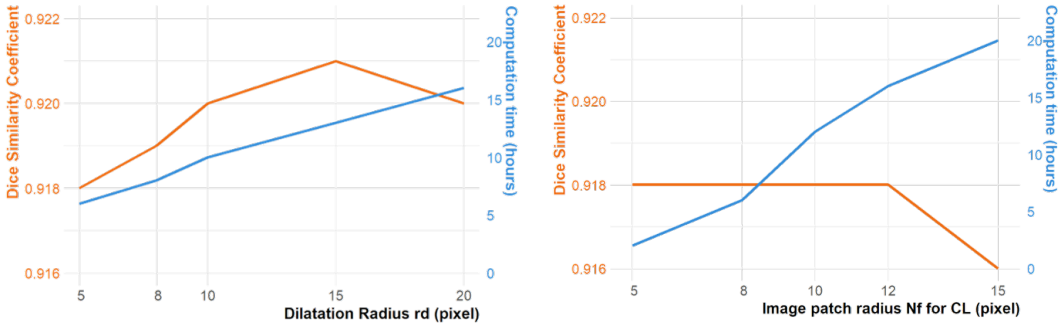

FIGURE 7.2: *Influence of the dilatation radius r_d and image patch radius N_f for feature learning on the segmentation results and computation time. DSCs reported here are the mean value of 7 Leave-One-Out tests.*

7.2 Number of atlases

The number of atlases is a critical factor in multi-atlas segmentation methods. Theoretically, the more atlases we have, the better the segmentation results. The impact of the number of atlases on the segmentation quality and the computation time is presented in Figure 7.3. The atlases for each test were selected randomly, and the reported time is for the JLF step only since the number of atlases does not affect the inference time of the CL step.

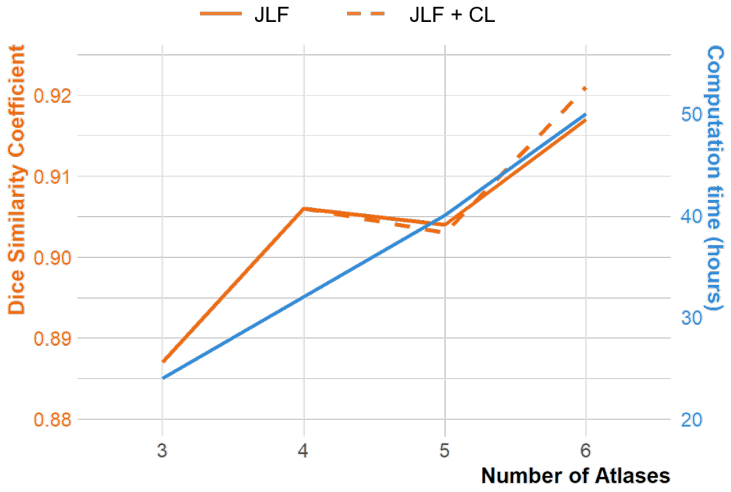

FIGURE 7.3: *Influence of the number of atlases on segmentation quality and computation time. The computation time is reported for the joint label fusion (JLF) step only. DSCs reported here are the mean value of 7 Leave-One-Out tests.*

As expected, between 6 and 3 atlases, the computation time is reduced by half with a decreased segmentation quality. We can also observe that from 4 to 5 atlases, the result is slightly better with 4 atlases, which means that adding information randomly is not necessarily in favor of our task. We hypothesize that, with a specific

atlas selection strategy, we might conserve the segmentation quality while reducing the number of atlases (thus reducing the computation time). Meanwhile, the CL step does not improve the results except for the case with 6 atlases, which might be due to the lack of information when working with few atlases.

7.3 Segmentation results with 6 atlases

Table 7.1 shows the results of Wang and Yushkevich's method with 6 atlases compared to Gilles et al.'s and with deformable registration with one atlas (Nguyen et al., 2018). The JLF + CL yields the best results for all metrics except HD, which is the smallest for Gilles et al.'s. The DSC and VS of JLF + CL is in the same value range of the inter-expert evaluation presented in Table 6.1. The segmentation by JLF is a lot more precise at muscle boundary, compared to Gilles et al.. Most of the time, the CL seems to identify the errors successfully but sometimes fails to correct them entirely (circled in Fig. 7.4), creating some noisy and aberrant voxels, hence a larger HD. Meanwhile, with a much smaller MAD, the larger HD of JLF + CL is not concerning since it probably resulted from these aberrant errors that can be easily removed with simple postprocessing.

Method	DSC	DSCw	HD (mm)	MAD (mm)	VS
Deformable registration	.821	.829	39.13	3.69	.139
Gilles et al.'s	.864	.870	30.43	2.53	.114
Wang and Yushkevich's JLF	.914	.917	34.77	1.65	.080
Wang and Yushkevich's JLF + CL	.921	.923	33.44	1.46	.056

TABLE 7.1: *Quantitative evaluation of Wang and Yushkevich's method on 7 subjects of the MUST dataset with full right leg segmentation (values are averaged over 7 Leave-One-Out tests), compared with Gilles et al.'s segmentation and deformable registration with one atlas.*

The method presented in Wang and Yushkevich (2013) were initially developed for MICCAI 2012 challenge's brain MR images with 207 regions in a volume of $128 \times 128 \times 60$ voxels, while our images have 4 regions in a volume of $280 \times 160 \times 640$ voxels. The regions to segment in brain images are much smaller than the muscle regions in our images, with less morphological variation among the subjects; the multi-atlas segmentation errors are smaller and only in a radius of several pixels. Although still bringing improvement to the segmentation in our case, the corrective learning method proposed is likely adapted for small errors, explaining the seemingly *incomplete* corrections observed in our results.

Table 7.2 shows the details on the DSC of segmentations by Wang and Yushkevich's method, by subject, and by muscle head, before and after CL.

We can observe a non-negligible difference among the subjects and the muscle heads. One muscle, the *rectus femoris*, is less well-segmented than the other muscles, with a significant disparity of results (0.737 to 0.950). Since the global DSC is the average DSC of all the muscle heads, the smaller the head, the more significant the impact an error in it made on the global DSC. Since all the muscle heads share boundaries, one poorly segmented muscle head can cause a decrease in DSC for all the others - a phenomenon we can observe in ANG-2014 and CAL-4223. Observing the variation in the shape and position of RF, VM, and VL in the database (Fig. 7.5), we can see a clear difference between these two subjects and the rest.

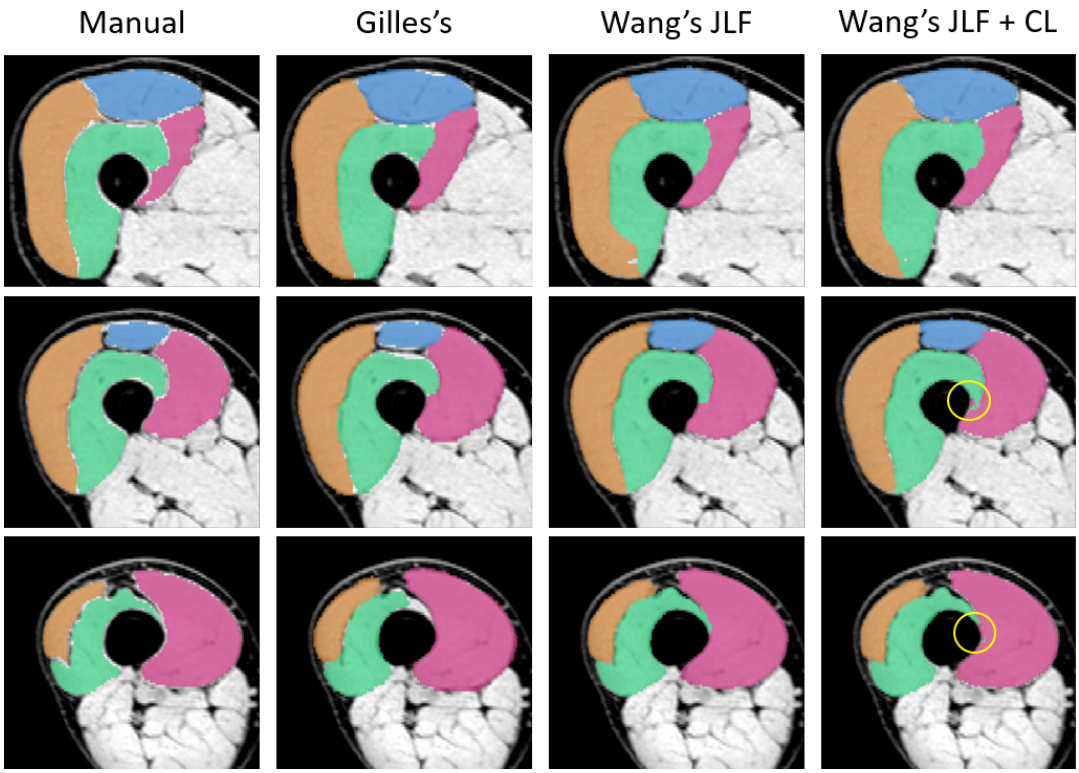

FIGURE 7.4: *Visual comparison between the segmentation of Gilles et al., the segmentation with the method of Wang and Yushkevich and the manual segmentation. Yellow circles indicate the zone where CL successfully identified the errors but failed to correct them entirely.*

	ALB-2725	ALF-4529	ANG-2014	ANS-3229	ARS-4026	CAL-4223	OUK-2927
VL							
JLF	.931	.914	.876	**.951**	**.944**	.883	.938
JLF+CL	.931	**.932**	**.896**	.942	.943	**.894**	**.939**
RF							
JLF	.928	.926	.786	**.946**	.937	.737	.950
JLF+CL	**.936**	**.941**	**.824**	.939	**.948**	**.791**	**.951**
VM							
JLF	.924	.953	.923	**.954**	.953	.880	**.950**
JLF+CL	**.937**	**.959**	**.930**	.952	**.956**	**.886**	.945
VI							
JLF	**.903**	.917	**.906**	**.932**	.934	.890	.918
JLF+CL	.899	**.935**	.901	.927	**.937**	.890	**.920**
Global							
JLF	.921	.927	.873	**.945**	.942	.848	.939
JLF+CL	**.926**	**.942**	**.888**	.940	**.946**	**.865**	.939

TABLE 7.2: *Details on the Dice Score Coefficients of the automatic segmentations by Wang and Yushkevich's method. Bold values signifies better DSC between before and after corrective learning. Muscle head abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius.*

Since the variation is too large, the deformable registration cannot find a satisfying solution for our problems. For example, in Fig. 7.6, all atlases except for ANG-2014, whose morphology is quite similar to the reference, are poorly registered on CAL-4223, which leads to a mediocre multi-atlas segmentation result.

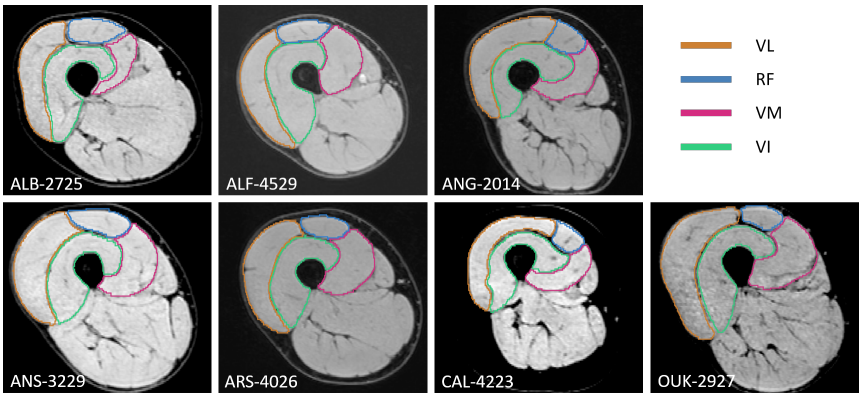

FIGURE 7.5: *Center axial slice of 7 subjects having their right leg manually segmented.*

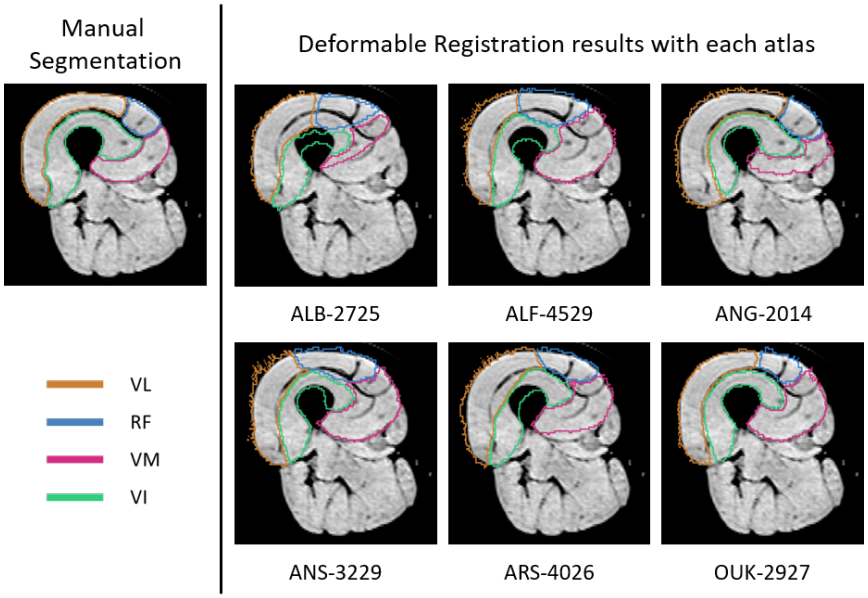

FIGURE 7.6: *Segmentation obtained by deformable registration of 6 other atlases on CAL-4223. Due to the large morphological difference, only one segmentation (issued from ANG-2014) seems acceptable.*

Furthermore, the *corrective learning* is also unstable as it does not always improve the results (see Tab. 7.2, the DSC of ANS-3229 was reduced by CL). An example of a segmentation error that was extended by the corrective learning is shown in Fig. 7.7. The interested muscle head is located on the left of the yellow line in each image. As we can see, the boundary of muscle here is not clear. Additionally, there is a structure with lower intensity near the target boundary (red-circled zone); it could be a blood vessel or other biological structure. Since the muscle boundary usually has lower intensity than the muscle, it is understandable that the JLF chose to place the muscle contour there. The corrective learning then made it worse as it extended the contour to cover that zone completely, which visually seems reasonable.

7.4 Optimizing with lower resolution

Two main disadvantages of Wang and Yushkevich's method are the high computation time and the lack of large-scale information in JLF. Resolving these problems is

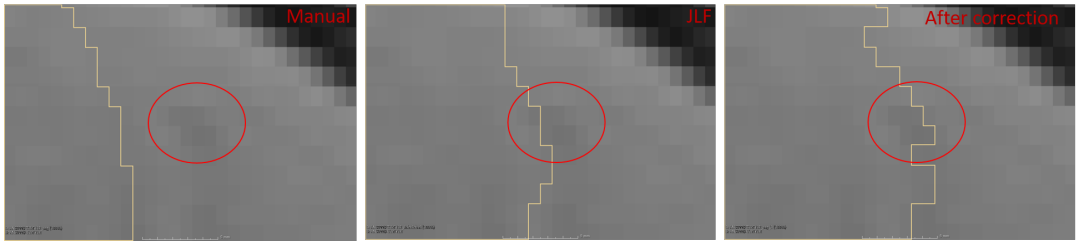

FIGURE 7.7: *From left to right: zoom on image of ANS-3229 with the manual delineation, with the result of JLF and with the result after correction with CL. The circle points out the abnormal zone. The interested muscle head is located on the left side of the yellow line in each image.*

not easy since these two factors are closely related: to include large-scale information, we have to increase the JLF parameters (image patch size and research neighborhood size), which will lead to an undesirable increase in computation time. Our idea here is to reduce the resolution of our image in order to, at the same time, reduce the computational time and improve the segmentation by including large-scale information.

Our images were shrunk by factors of 2, 4, and 8 with a first Gaussian smoothing. The JLF process was then computed at these lower resolutions. To visualize and quantify the results, we resampled the segmentations to the original resolution. Figure 7.8 shows the results of JLF and the average DSC for all resolutions. Despite having lower DSC, shrink factor 2 seems to have a more consistent anatomical structure (no hole in the middle of a muscle head, no unrealistic curvature, ...). The shrink factor of 4 and 8 seems too large and loses too much information at the muscle boundary. While not suitable for our application, with the computation time reduced from 50 hours to 1 hour and 10 minutes, respectively, the segmentation at these resolutions can be helpful to detect rapidly muscular zones either for quick analysis or to get a bounding box for a more precise segmentation method that needs preliminary segmentation.

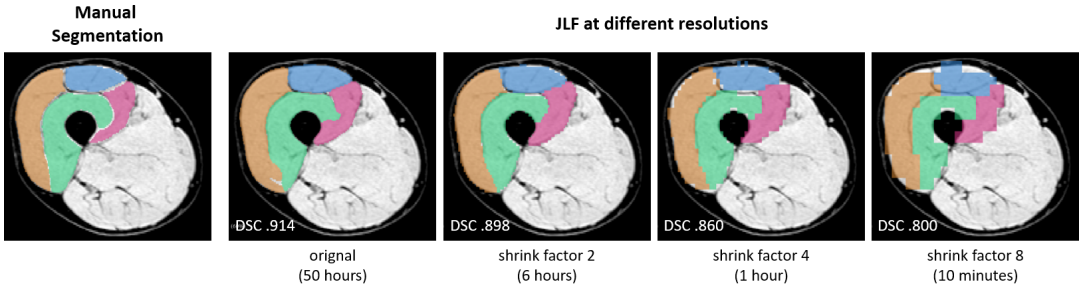

FIGURE 7.8: *Joint Label Fusion results at different resolutions, in the parentheses is the computation time. Reported DSC is the average value of 7 subjects.*

We then applied the CL step to the segmentation with shrink factor 2 since they are the most accurate among the lower resolutions. Our experiments included:

- correcting the segmentation at the lower resolution than at the original resolution.
- resampling the JLF segmentation result to the original resolution then correcting them at the original resolution as proposed in Wang et al. (2017)

Regardless of the promising results reported in Wang et al. (2017), the CL at original resolution gives an average DSC value at 0.912, far lower than our previous result (.921, Tab. 7.2). The only metric that trumps the segmentation done entirely at original resolution is HD: 30.68 mm on average compared to 33.44 mm, which confirms our above statement that the segmentation with shrink factor 2 has a more coherent anatomical structure, meaning less aberrant errors. The second experiment with successive CL at two different levels does not yield promising results, reducing the DSC. While the segmentation at lower resolution provides some expected improvement (lower computation time, anatomically more coherent results), it seems that we still need to figure out a more efficient way to incorporate the information of this resolution into the final segmentation.

7.5 Conclusion

In this chapter, we have studied the approaches of segmentation by atlas registration and, in particular, the approach of Wang and Yushkevich. The approach allows us to obtain good segmentation of one or several organs from a small number of atlases because they take advantage of all the anatomical knowledge correlated to the image information contained in the atlases. It is to be favored to quickly provide high-quality segmentation when few manually segmented data are available. When the number of atlases becomes large, a very significant increase in computing time is quickly observed and also a limit to the improvement of results.

A quick modification of the method (Wang et al., 2017) with joint label fusion at a lower resolution can reduce the computation time significantly and aberrant anatomical errors but does not improve the validation metrics.

Despite the use of joint label fusion and corrective learning such as those proposed by Wang and Yushkevich, the quality of segmentation depends significantly on the ability to register the atlases on the image to be segmented. Moreover, this approach does not perform well when the anatomy is not consistent or simply if the anatomical variability is high between the atlases and the image to be segmented. Indeed, they use the atlases in their entirety and can hardly take into account a local and specific modification (lesions, anatomical anomalies, hypo or hypertrophy). An alternative to MAS that can reduce both the computation time and the sensibility to registration is using deep learning networks such as UNet. However, for the training stage, these approaches require a large amount of segmented data. The next chapter presents this and a strategy to increase the amount of data without additional manual segmentation.

Another way to reduce the computation time of MAS is to reduce the number of atlases. Meanwhile, preserving the segmentation quality imposes an optimized selection of atlases. From our previous observation, such selection should be based on the morphological similarity with the data to be segmented. The development of this perspective will be presented in Chapter 9.

CHAPTER 8

UNet-based approach

The results of Wang and Yushkevich's method were encouraging, but this method is costly in terms of computational time, especially the joint label fusion (JLF) step. Moreover, the method is not robust in the case of large morphological variations among subjects. Here, we proposed to conserve the principal structure of Wang and Yushkevich's framework while replacing the multi-atlas segmentation with JLF step with UNet (Ronneberger et al., 2015) (Sec. 8.1). The UNet is trained and validated with both manually annotated and automatically generated data (Sec. 8.2.2) in order to adapt to various morphologies.

We also expect that the CL step will replace all ad-hoc post-processing often added after a UNet segmentation in removing aberrant segmentation errors.

8.1 Architecture to replace joint label fusion

We are conserving the idea behind Wang and Yushkevich's algorithm that employed a *host* automatic multi-atlas segmentation method to segment a test image and each one of the reference images using all the other reference images as atlases (8.1). The automatic segmentations of reference images were fed to the corrective learning (CL) algorithm that learns and then corrects the typical errors made by the automatic multi-atlas segmentation method (Nguyen et al., 2019b).

Here, the segmentation method by multi-atlas deformable registration and JLF is replaced with UNet (Nguyen et al., 2019b). The original 2D UNet of (Ronneberger et al., 2015) with 64 filters at the first level is used with a Batch Normalization layer after each convolutional layer and ReLU activation function. Before ending with a softmax layer, we add a dropout layer (Li et al., 2018). The learning rate, the batch size, and the dropout coefficient are optimized for each experiment. The axial slices are considered separately and are re-stacked at the end to get the full 3D volume.

We trained and validated a *complete* UNet with the entire training and validation sets, which will be used to segment the test set. For the CL step, we trained multiple UNets by removing each atlas and its derivations and using the rest of the training and the validation sets. The segmentation was evaluated using qualification metrics presented in section 3.1.

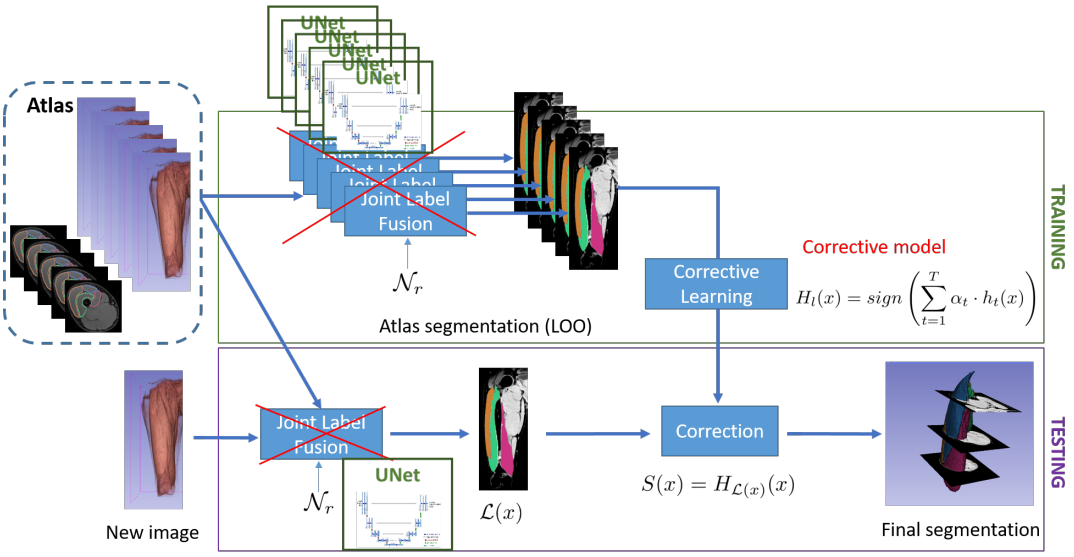

FIGURE 8.1: *Our segmentation framework based on Wang and Yushkevich's with UNet as host segmentation method and corrective learning (UNet + CL).*

8.2 Experiments & Results

8.2.1 First result of UNet without data augmentation

With LOO strategy (Sec. 3.2), each subject among the 7 with full manual right leg segmentation served as the test subject. Among the rest, four were selected randomly and served as training data, while the other two served as validation data. For each test, excluding the slices without interested anatomical structure, there are around 2400 slices for training and 1200 slices for validation.

Quantification of segmentation quality gives very high HDs, with the average value at 109.06 mm. This is comprehensible as we are working with 2D UNet: since the network does not consider the information of the third axis, large spatial errors appear frequently. The average DSC is at 0.854, while the individual scores vary from 0.700 to 0.927. The worst segmentation is of CAL-4223 (Fig. 8.2), the same as with JLF, with 0.2 of DSC difference with the second-worst. It is clear that the network does not have enough training data to adapt to various morphologies.

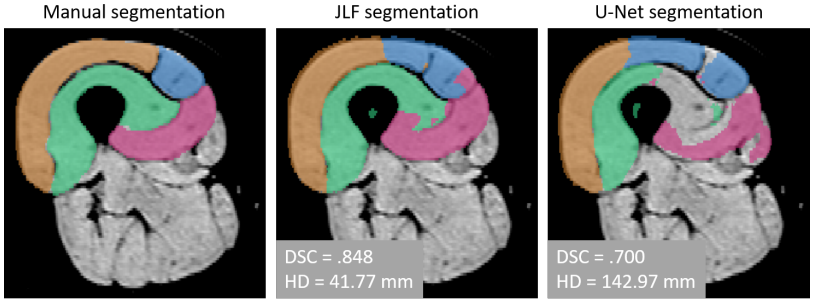

FIGURE 8.2: *Visualization and validation metrics of the CAL-4223's segmentation with UNet (without data augmentation), comparing with the manual segmentation by medical expert and the segmentation with Wang et al.'s JLF.*

8.2.2 Weakly-supervised UNet

In order to enrich our annotated data with different types of morphology, we effectuated a data augmentation step with the help of deformable registration and random B-Spline warping, hence the name *weakly-supervised UNet*: our UNet is trained with experts' segmentations and also weak automatically generated ones.

8.2.2.1 Data augmentation

To obtain morphologically diverse training data, the registration was performed using the athletes without manual right leg segmentation in our dataset (41 subjects). Each of our 7 annotated right leg image volumes was:

- registered to 5 other images (selected randomly without replacement). We used an affine transform followed by a B-spline deformable one to perform the registration. The results were smoothed using a mathematical morphological opening operation to avoid unrealistic deformation.
- randomly B-splines-warped 5 times to produce 5 more image volumes with a minor modification to the original image. The warped image must reduce at least 30% of original mutual information metrics (see Annex B).

The data augmentation was done using `elastix` and `C++/ITK`. Figure 8.3 shows the data derived from one of the atlases (ALB-2725).

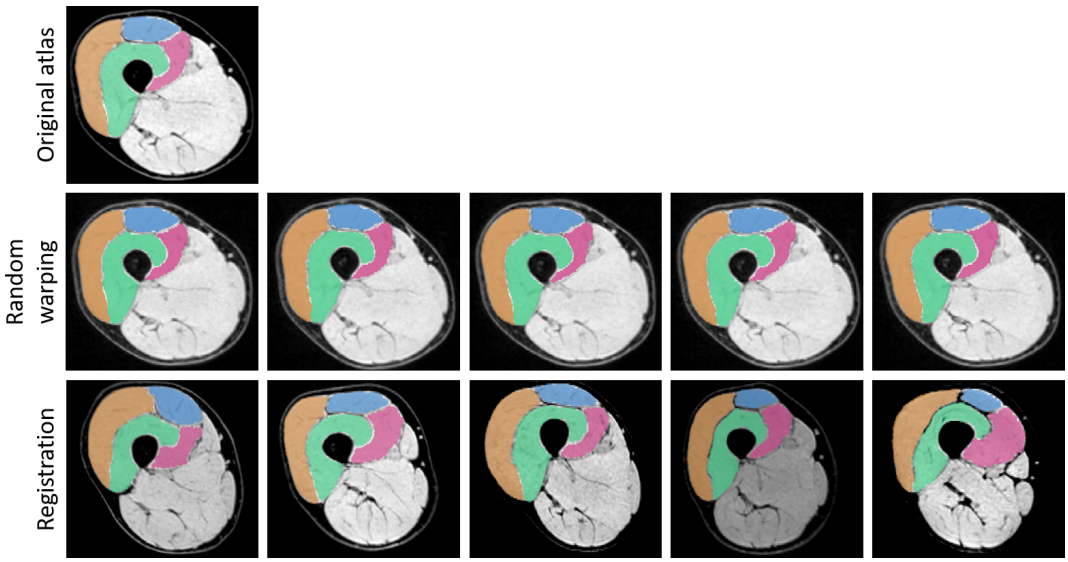

FIGURE 8.3: *ALB-2725 atlas and its derivations: 5 random-B-spline-warped images and 5 registrations to 5 different non-annotated images.*

8.2.2.2 Experiments

We used the augmented datasets to train our UNets as described in Section 8.1. Since we were working with 2D axial slices of very high-resolution images, the anatomical difference among several consecutive slices was minimal. Therefore, only 20% of each volume were used to train and validate our networks to avoid redundant information and reduce the training time. LOO strategy was also employed here, for each test image, all the automatically generated images originated from the test

image were excluded from training and validation sets. We experimented with training and validation data involving data generated by either one or both deformable registration and random B-spline warping.

- *Only random B-spline warping*: All original images were used as training data. For each training image, three among five random-warped images were used for training while the other two were used for validation. There were approximately 1500 image slices in the training set and 750 in the validation set.
- *Only deformable registration*: All original images were used as training data. Three among five registered images originated from each training image were used for training while the other two were used for validation. There were approximately 1500 image slices in the training set and 750 in the validation set.
- *Both random B-spline warping and deformable registration*: All original images were used as training data. Three random-warped images were used for training. Three among five registered images originating from each training image were used for training, while the other two were used for validation. There were approximately 2500 image slices in the training set and 750 in the validation set.

The UNet was implemented in Python language with Keras/Tensorflow (Chollet, 2015) and was run on an NVIDIA Tesla P100 PCIE 16GB.

8.2.2.3 Results

Compared to JLF, which took 48h to segment an image volume, the weakly-supervised UNet considerably reduced the execution time. Our UNet took 1 hour 30 minutes to 2 hours for training and 45 seconds for inference on a whole image volume. The data augmentation step took around 5h with 36 CPUs. In our experiments, the data augmentation was performed only once and the training was performed once for each training set of the LOO scheme. The quantitative results of our experiments are reported in Table 8.1, in comparison with the results of Wang and Yushkevich's method. We would like to remind that the inter-expert score is .910 in DSC and .051 in VS.

Method	DA	DSC	HD (mm)	MAD (mm)	VS
JLF	None	.914 [.848, .945]	34.77 [18.67, 48.59]	1.65 [0.82, 3.27]	.080 [.028, .141]
JLF + CL	None	.921 [.866, .946]	33.44 [20.73, 40.40]	1.46 [0.88, 2.74]	.056 [.024, .104]
UNet	W	.892 [.774, .946]	98.73 [79.66, 141.49]	2.22 [0.87, 3.78]	.097 [.047, .172]
UNet	R	.915 [.842, .947]	79.49 [32.97, 141.62]	1.77 [0.82, 3.72]	.095 [.043, .190]
UNet	W + R	.921 [.874, .945]	85.32 [52.97, 132.57]	1.46 [0.95, 1.99]	.064 [.034, .095]
UNet + CL	W + R	.917 [.842, .947]	48.83 [13.43, 89.84]	1.48 [0.84, 2.70]	.061 [.022, .140]

TABLE 8.1: *Quantitative evaluation of different automatic segmentation methods on MUST dataset. UNet was tested with different data augmentation (DA) strategies: W - random warping, R - registration, W + R - both random warping and registration. Results are reported as mean[min, max] over 7 subjects.*

Among the three experiments with different data augmentation strategies, the one with both random warping and deformable registration (W+R) yields the best average results for almost all metrics. On the other hand, the average HD of the experiment with registration-only data augmentation is the smallest, but the value

range is the largest. Therefore, the W+R data augmentation is conserved for further experiments. From this point onwards, except when specified otherwise, the 2D UNet is always trained with W+R data augmentation.

In terms of DSC, the segmentation with 2D UNet gave similar results as Wang and Yushkevich's JLF + CL. The HD is exceptionally high since the method has created some distanced structures, probably due to missing information of the third axis. These structures are often removed with CL since they are too far from the barycenter of the other voxels with the same label. However, the CL did not have an overall positive impact on the segmentation of UNet as it reduced all metrics except for HD and VS: While CL is efficient at correcting these aberrant distanced or small boundary errors, it is not as efficient in the case of large morphological ones.

Figure 8.4 shows the visual results of the methods in the experiment for three individuals: one with highly accurate automatic segmentation (ARS-4026), one with average validation metrics (ALB-2725), and one with the lowest validation metrics (CAL-4223). For ARS-4026, both JLF and UNet produced satisfying segmentations, and CL worked well with both of these host segmentation methods and corrected minor errors at muscle boundary. For ALB-2725, UNet created a distanced structure outside of the quadriceps regions, which was removed by CL. However, the CL in this case also removed some well-segmented voxels and lowered the validation metrics. UNet seemed to work better for CAL-4223 than JLF, but the results are still not up to par. The JLF could not adapt to the size and position of RF (blue) and VI (rose) (cf. Fig. 5.2), and while UNet made smaller errors for RF, it failed to recognize the VI. The CL also failed to correct the large errors entirely and even created more errors in the case with UNet.

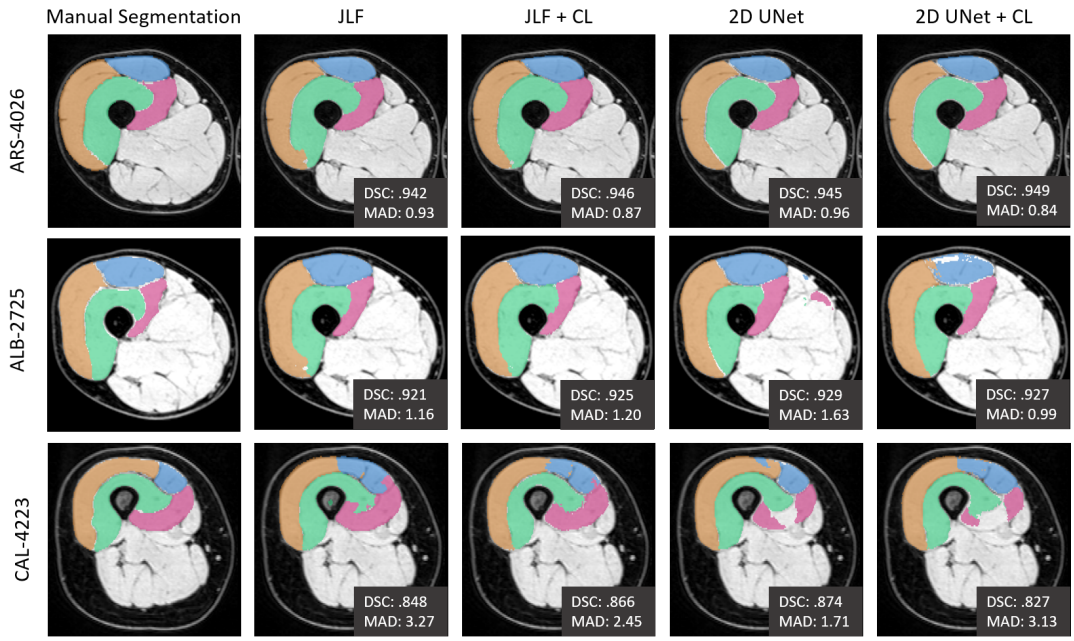

FIGURE 8.4: *Results of 4 different automatic segmentation methods, compared with the manual segmentations, of 3 subjects with visually different morphology.*

8.2.3 UNet variants

As we have observed, although the *corrective learning* does have some positive impacts on the segmentation quality, the host segmentation method must not make

large errors in the first place. We have looked into different variants of UNet, including ResUNet (Zhang et al., 2018) - UNet combined with a residual type architecture and 2.5D UNet (Haque et al., 2019) - differents UNets trained with different image slice directions followed by *winner takes it all* voting. The quantitative results are reported in Table 8.2.

Method	DSC	HD (mm)	MAD (mm)	VS
2D UNet	.921 [.874, .945]	85.32 [52.97, 132.57]	1.46 [0.95, 1.99]	.064 [.034, .095]
2D ResUNet	.920 [.854, .948]	94.00 [45.32, 151.29]	1.50 [0.83, 2.40]	.070 [.036, .142]
2.5D UNet	.915 [.873, .945]	59.24 [27.88, 103.87]	2.13 [0.82, 4.24]	.077 [.047, .129]

TABLE 8.2: *Quantitative evaluation of different architecture based on UNet tested on our quadriceps dataset. Results are reported as mean[min, max] over 7 subjects.*

While the 2.5D UNet did help eliminate aberrant distanced errors, which is reflected by much smaller HD, both ResUNet and 2.5D UNet did not improve the segmentation accuracy compared to the classic UNet. Considering the best values for all validation metrics are similar among different methods, we need to focus on improving our network's adaptation capability to a morphology that is less represented in the dataset. These networks will be re-investigated with another database (Sec. 10.1).

8.3 Conclusion & Perspectives

In this chapter, we proposed to replace the Joint Label Fusion (JLF) segmentation method in Wang and Yushkevich's framework with 2D UNet. The UNet was trained and validated with manually segmented data and automatically generated ones to enhance the training and validation sets' morphological diversity. The computation time is reduced considerably compared with JLF, passing from 48h of inference time to 45s with similar segmentation qualities, reflected by validation scores in the same range as the inter-expert scores. Testing multiple data augmentation methods, the best results were obtained with a network trained and validated by a dataset augmented with both random B-spline warping and deformable registration to non-annotated data.

While the segmentation errors made by JLF and UNet are not the same as JLF works directly with 3D volume, and UNet works with separated 2D slices, the largest errors are often made on some certain subjects whose morphological type is the minority in the dataset. In the next chapter, we will focus on improving our methods' adaptation capability by paying more attention to the morphology of the subjects used as training data or atlases.

CHAPTER 9

Morphological features

Chapter 7 and Chapter 8 presented the results of Wang and Yushkevich's multi-atlas segmentation method and weakly-supervised 2D UNet. These automatic approaches produced segmentations with average Dice Score Coefficients (DSC) in the same value range as inter-experts score (see Table 6.1). However, both methods failed to produce satisfying results for some subjects with particular morphology.

In this chapter, we propose i) a morphology measurement in order to validate our observation; ii) 3 strategies to improve automatic segmentation results. Thus, we first introduce our morphological measurement dedicated to quadriceps segmentation. Then, we improve automatic segmentation results obtained by multi-atlas segmentation with a corrective learning approach using a selection of atlases based on morphological similarity to the image to process. Our results show that using few atlases (3 in lieu of 6) based on our morphological measurement improves segmentation quality and decreases computation time for multi-atlas segmentation with CL. Based on the proposed measurements, we also defined a data augmentation strategy for the weakly-supervised UNet, expecting better generalization capability, with encouraging results.

9.1 Morphological measurement

We propose here a measurement dedicated to morphology (i.e. only take into account the region of interest, not the MR signals). Morphological features are computed on the segmentation of a specific image slice. In our case, this slice is the central axial slice of 3D GRE sequences located at 15 cm from the upper part of the patella (Sec. 2.3.1). For each muscle head of the quadriceps, we measure 3 features:

- the muscle surface S, in mm^2.
- the polar coordinates (θ, r) of the center of the muscle head with the center of the femur (FM) as the pole and the vector between the center of the femur and the center of the vectus intermedius muscle head (VI) as the polar axis (Fig. 9.1).

Each quadriceps group has 12 features in total: 3 features x 4 muscle heads. The angle θ of VI is set to 0 for every leg and serves as a reference angle to let our features be rotation invariant (the proposed features are also shift-invariant). Thus, the feature vector for a leg X can be noted:

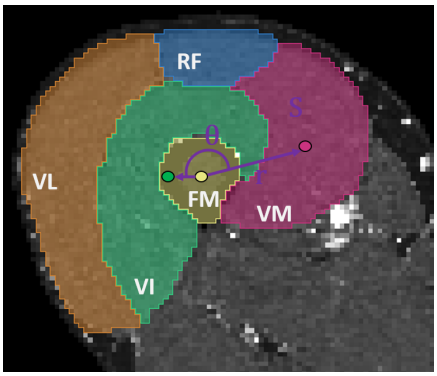

FIGURE 9.1: *Morphological features of the vectus medialis on the right leg of a runner. Abbreviations: FM - femur, VM – vectus medialis, VL – vectus lateralis, VI – vectus intermedius, RF – rectus femoris.*

$$f_X = [S_X^{VI}, 0, r_X^{VI}, S_X^{VM}, \theta_X^{VM}, r_X^{VM}, S_X^{VL}, \theta_X^{VL}, r_X^{VL}, S_X^{RF}, \theta_X^{RF}, r_X^{RF}] \tag{9.1}$$

The measurement is based on the manual segmentation of the central axial slice of 3D GRE sequences. In the case of our dataset, after excluding runners with low-quality 3DGre acquisition, we have **48 subjects** in total.

Next, all features were centered and scaled to calculate unbiased distances between subjects and be more robust to acquisition properties. The morphological difference d_{AB} between 2 legs A and B was computed with:

$$d_{AB} = \|f_A^* - f_B^*\|_2$$

where f_X^* the vector of standardized morphological features of the leg X, $f_X^* \in \mathbb{R}^{11}$.

Figure 9.2 shows the right leg's features of 48 subjects projected on the plan of the first two PCA axis, which represent 70.05% of the dataset's variance) with examples of T1w images and segmentations of 4 different subjects, among which two are close in distance (in yellow and pink). The subjects (in blue and green) distanced from the main cluster show indeed visually different morphology to the others.

By including the center of one muscle head (VI) in the polar axis, our features are image rotation invariant. We investigated the case of a biomechanical rotation of a leg during an MRI acquisition. Images of the right leg of a control subject were taken in 2 different positions: normal (relaxed) position and turned inward (9.2, on the right). The morphological distance between the two positions of this subject is the smallest compared to the distances from them to the other subjects, being approximately .68 the distance to the closest and at .10 the distance to the furthest subject. The distance is also the smallest in the distance matrix of all subjects' right legs, confirming our hypothesis that our features are invariant to a morphological rotation that can occur during acquisition.

With this new measurements, we propose 3 strategies to optimize the atlas choice for the previously proposed segmentation methods.

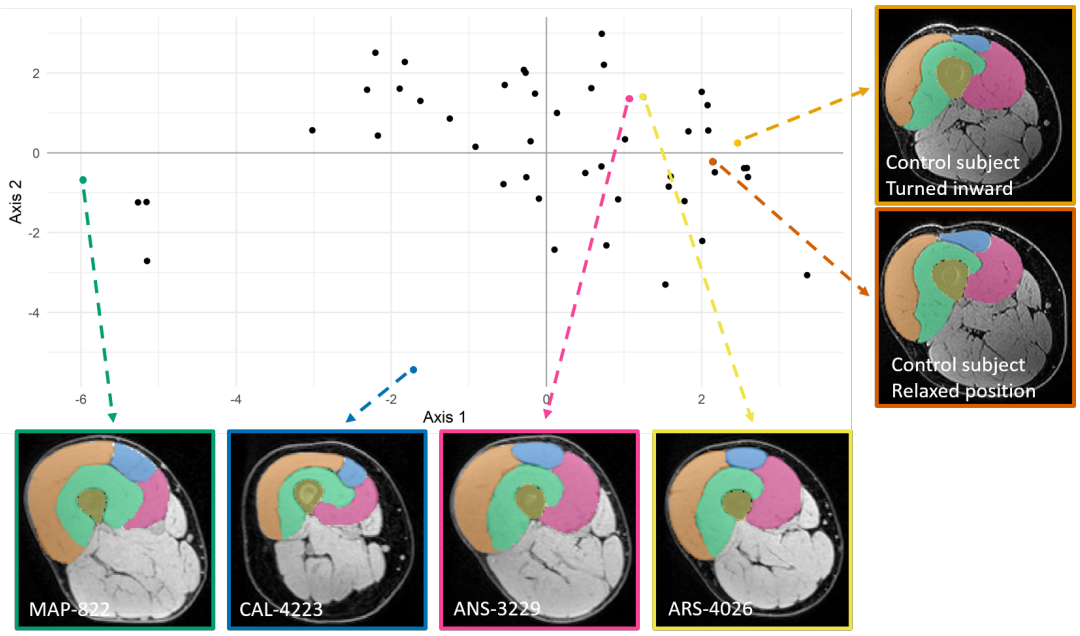

FIGURE 9.2: *Representation of subjects on the plan of the first two Principal Components Analysis (PCA) axis of their right legs' morphological features with examples of T1w images and their segmentations at the axial slice where morphological features were extracted. In green, blue, pink, and yellow (at the bottom) are 4 different dataset subjects. In light and dark orange (on the right) are images of a control subject (not originally in our database) in two different positions.*

9.2 Atlas selection for multi-atlas Segmentation

Two major drawbacks of JLF+CL are the computation time required when using 6 atlases and reduced performance when applied to subjects with morphology differents from the most of atlases. Here, we propose to optimize the method with a strategic choice of 3 atlases to reduce the computation time and improve the segmentation quality. When introducing new image volume for segmentation, we only need to define roughly the quadriceps muscle heads boundary at one axial slice to employ this strategy.

9.2.1 Experiments

With 7 manual segmentations, we adopted the Leave-One-Out (LOO) scheme to evaluate the segmentation method: each subject among the 7 subjects with manual segmentations was used as test using the 6 others as atlases. Based on the morphological features, we could sort the atlases from the closest to the furthest to the test image and choose to use either all 6 atlases for the segmentation or only the 3-5 atlases the closest to the test subject.

The experiments were implemented using `elastix` (Klein et al., 2010) and `C++/ITK` (Wang and Yushkevich, 2013; Yoo et al., 2002; Tustison et al., 2017).

9.2.2 Results & Discussion

As reported in Section 7.2, the computation time of Joint Label Fusion (JLF) step increases rapidly with the increase in number of atlases, from around 24h with 3 atlases

to around 50h with 6 atlases. Figure 9.3 shows the results in terms of DSC of Wang and Yushkevich's method with random and morphology-based atlas selection.

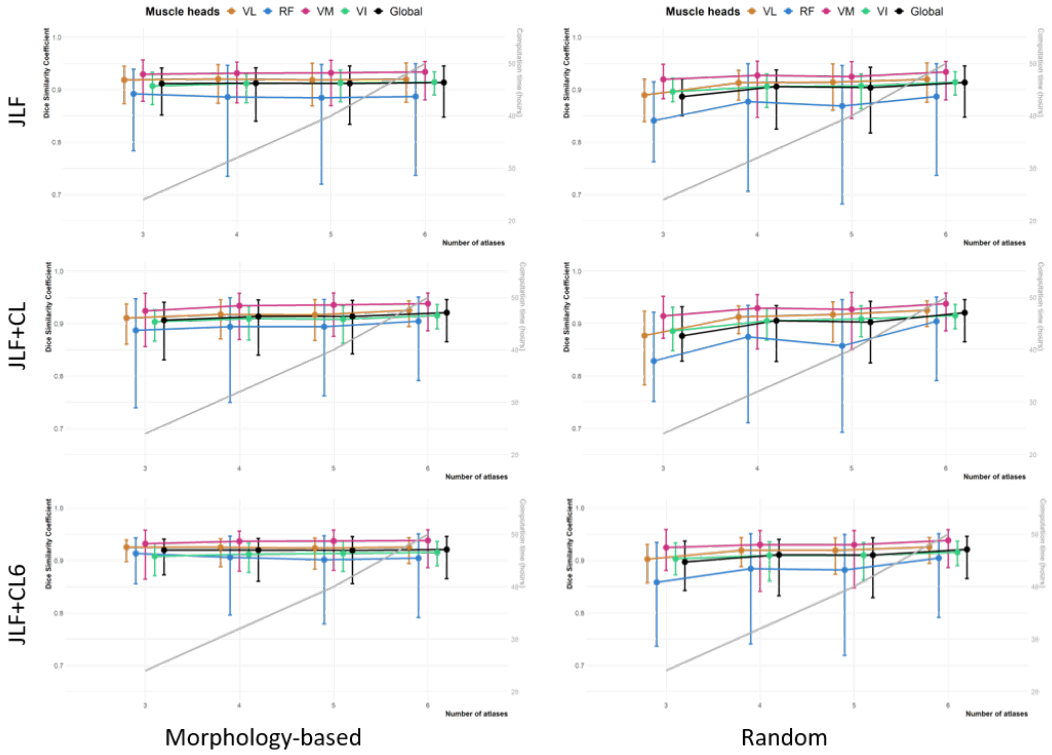

FIGURE 9.3: *Results in terms of Dice Similar Coefficient of Wang and Yushkevich's method with random and morphology-based atlas selection. The results are reported for joint label fusion (JLF), JLF and corrective learning (CL), and JLF and 6-atlases-based CL (CL6). Each color represents a muscle head, with the color black representing the global score over the entire quadriceps. Each muscle head is presented with the average value and a vertical bar limited by the minimal and the maximal values over 7 subjects.*

Compared with the random one, the morphology-based atlas selection has reduced the impact of a smaller number of atlases on the average quality, from .877 in average DSC with 3 random atlases to .912 with 3 morphologically closest atlases (see Table 9.1). Corrective learning (CL) with 3 closest atlases did not improve the segmentation, which confirmed the observation in previous studies Nguyen et al. (2018, 2019b) that the CL in its current state is not suitable to correct large errors caused by morphological variation in the quadriceps. Meanwhile, since increasing the number of atlases in CL will only increase the training time (learning the corrective model) but not the inference time, we applied the corrective model learned on 6 atlases (CL6) on the results of JLF with 3 closest atlases and obtained an average DSC similar to JLF + CL entirely with 6 atlases. Moreover, when introducing a new image volume to segment, correcting the automatic segmentation with a model pre-trained on all of our available atlases is more convenient than re-train a corrective model based on the 3 closest atlases.

Overall, the JLF + CL with 6 atlases has the best performance in terms of DSC and VS, but not by far. In the meantime, the JLF with 3 closest atlases + CL based on 6 atlases outperformed the JLF + CL with 6 atlases regarding the MAD metrics and regarding the robustness (based on the smaller value range in all metrics, cf. Table 9.1 and Figure 9.3). Smaller MAD means smaller errors in 3D physical space,

	Random			Morphology-based		
	JLF	*JLF + CL*	*JLF + CL6*	*JLF*	*JLF + CL*	*JLF + CL6*
DSC						
3 atlases	.887 [.850, .921]	.877 [.828, .932]	.897 [.842, .937]	.912 [.852, .942]	.906 [.831, .941]	.920 **[.873, .941]**
4 atlases	.906 [.825, .938]	.906 [.828, .934]	.911 [.833, .940]	.912 [.840, .942]	.914 [.840, .945]	.920 [.860, .942]
5 atlases	.904 [.817, .943]	.903 [.825, .943]	.910 [.829, .943]	.912 [.834, .945]	.914 [.842, .944]	.919 [.856, .946]
6 atlases	.914 [.848, .945]	**.921** [.866, .946]				
MAD (mm)						
3 atlases	2.15 [1.32, 3.46]	2.50 [1.09, 3.75]	1.94 [1.01, 3.29]	1.62 [0.97, 2.90]	1.72 [1.03, 3.13]	**1.43 [0.99, 2.66]**
4 atlases	1.85 [1.14, 3.87]	1.91 [1.09, 3.82]	1.79 [1.06, 3.09]	1.67 [0.95, 3.21]	1.61 [0.88, 3.00]	1.48 [0.99, 2.69]
5 atlases	1.95 [0.90, 4.36]	1.93 [0.91, 4.05]	1.75 [0.91, 3.55]	1.72 [0.88, 3.57]	1.58 [0.90, 2.97]	1.53 [0.87, 3.01]
6 atlases	1.65 [0.84, 3.27]	1.45 [0.88, 2.74]				
VS						
3 atlases	.104 [.069, .143]	.117 [.046, .238]	.085 [.044, .169]	.076 [.030, .138]	.079 [.035, .121]	.064 **[.037, .088]**
4 atlases	.080 [.045, .152]	.079 [.051, .121]	.062 [.019, .090]	.079 [.032, .151]	.076 [.036, .137]	.060 [.033, .091]
5 atlases	.087 [.025, .175]	.082 [.031, .148]	.066 [.028, .136]	.083 [.025, .164]	.077 [.029, .155]	.062 [.030, .118]
6 atlases	.080 [.028, .141]	**.056** [.024, .104]				

TABLE 9.1: *Validation metrics of segmentations with Wang and Yushkevich's method, with and without morphology-based atlas selection. The results are reported for joint label fusion (JLF), JLF and corrective learning (CL), and JLF and 6-atlases-based CL (CL6). Each metric is reported with the average, the minimal, and the maximal values over 7 subjects. Bold values mark the best average value or the smallest value range.*

and thus, even with similar DICE, these results suggested that the method has made fewer errors related to morphological variation. Figure 9.4 shows the segmentation results of CAL-4223, the subject always with a low-quality segmentation due to their morphological difference with the other subjects (Sec. 7.3, Fig. 7.6): The Vectus Intermedius (VI) muscle head is always the most difficult to segment, the errors made on this muscle head is not corrected entirely but still noticeably smaller when working with 3 closest atlases.

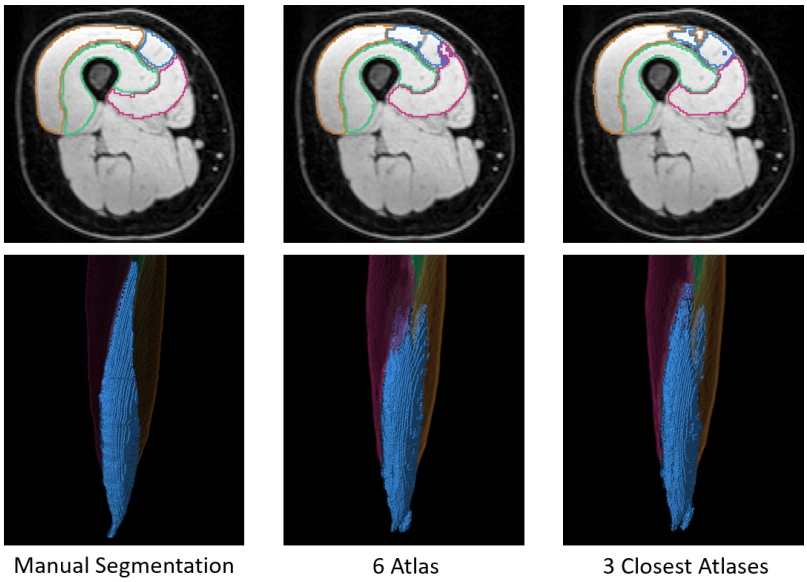

FIGURE 9.4: *Segmentation results of CAL-4223 based on 3 closest atlases and on 6 atlases. CAL-4223 is always the subject with the worst segmentation quality since their Vectus Intermedius (VI) muscle head (blue color) is much smaller and is at a different position compared with the other 6 subjects (cf., Fig. 7.5 & 7.6)*

This strategy reduces by half the execution time while conserving or even improving the segmentation quality. Next, we will present our strategy for weakly-supervised UNet.

9.3 Selective data augmentation for weakly-supervised UNet

As discussed in Section 8.2.1, training with 6 atlases is not enough for UNet since even when in 2D with hundreds of axial slices in the training set, we have only 6 different morphologies, among which some are quite similar to each other. When the redundant information is fed to UNet, it will be inefficient in treating new data: data augmentation is indispensable in this case.

Initially, the images used as targets for data augmentation by registration were selected randomly for each reference image (Sec.8.2.2). By introducing the morphological features, we suggested a specific data augmentation strategy selection to exploit as much as possible the morphological diversity in our dataset. Our objective is to train only one model that can accurately segment different morphologies.

9.3.0.1 Experiments

We computed the morphological distances from each annotated image to all the other 41 images in the data set (48 minuses 7 atlases already in the training set) and sorted these morphological distances in ascending order (i.e., from the closest to the furthest). Each reference image would have 41 candidates for registration: we excluded the first 11 since they would be too similar to the reference image and then divided the other 30 into five groups of 6 images in order of distance. We selected randomly and consecutively, for each reference image, without replacement, one image from the first, the third, and the last group for training and one from the second and the fourth groups for validation.

LOO scheme along with presented segmentation validation metrics (Sec. 3.1) was also applied here to evaluate the segmentation method. For each test, including the original atlases and their randomly-warped images, there were 42 images volumes in the training set and 12 in the validation set, which leads to a training set of approximately 2700 slices and a validation set of 770 slices.

We also rerun the experiment with random data augmentation of Section 8.2.2 four more times to compare with the morphology-based method.

The data augmentation was done using `elastix` and `C++/ITK`. The UNet was implemented in Python language with Keras/Tensorflow (Chollet, 2015) and was run on an NVIDIA Tesla P100 PCIE 16GB.

9.3.0.2 Results & Discussion

The quantitative results of the random and selective data augmentation strategies for UNet are reported in Table 9.2. Results are reported as *mean[min, max]* over 7 subjects. Overall, the two strategies give similar results, which is comprehensible since, for each experiment, we used 35 in 41 non-annotated subjects as targets for registration, so there is little difference among experiments. In the meantime, observing the value ranges of all metrics shows inclination towards the morphology-based strategy since its metric ranges are smaller, and its worst values are often better than of random selection.

Strategy	DSC	HD (mm)	MAD (mm)	VS
Random	.918 [.822, .954]	76.19 [37.26, 139.96]	1.51 [0.84, 3.60]	.076 [.036, .168]
Morphology-based	.920 [.850, .951]	77.34 [47.62, 105.82]	1.37 [0.76, 2.68]	.084 [.042, .151]

TABLE 9.2: *Quantitative evaluation of different data augmentation strategies for UNet. Results are reported as mean[min, max] over 7 subjects. The random data augmentation was run five times, and for each time, the UNet was retrained entirely for each test image.*

The lowest DSC at .850 is not satisfying as our objective is to train a network that can produce accurate segmentation for all morphology. There is still not enough information in the training set to improve the segmentation of *the most challenging case* (CAL-4223): We might need to aim for a more specific training set for each target image.

9.4 Target-driven UNet

Based on the conclusion above, where the training set did not have enough relevant information to work with certain types of morphology, we propose two strategies to train UNet based on the target image volume's morphology.

9.4.1 Target-trained UNet

Our idea here is to overtrain UNet with data the most similar to the target image:

- *Training set* (10 volumes): The two closest atlases registered on the two closest non-annotated volumes, respectively. The target volume was randomly warped four times to produce 4 target volumes on which registered the two closest atlases, thereby 2 per atlases.
- *Validation set*: The two closest atlases.

The term *closest* here refers to the morphological distance to the target image. The relative distances from each atlas to the other subjects are illustrated in Figure 9.5.

9.4.2 Fine-tuned UNet

Another option is to fine-tune our pre-trained UNet from Sec. 9.3 for each target image. Before being applied on the target image, the trained model is tuned with:

- *Training set*: The two closest atlases registered on the target image.
- *Validation set*: The third closest atlas registered on the target image.

Most of the fine-tuning processes are optimized before the 5th epochs and take only 5 to 10 minutes.

9.4.2.1 Results

The quantitative results of our experiments are reported in Table 9.3.

Based on the validation scores, target-trained UNet has the weakest performance among the four methods. However, it helps validate our hypothesis that the information fed to UNet must be chosen carefully: While being trained with only 8 image volumes instead of 42 as for generic UNet, the target-trained UNet did not reduce much the average DSC in comparison with the latter (.916 compared to

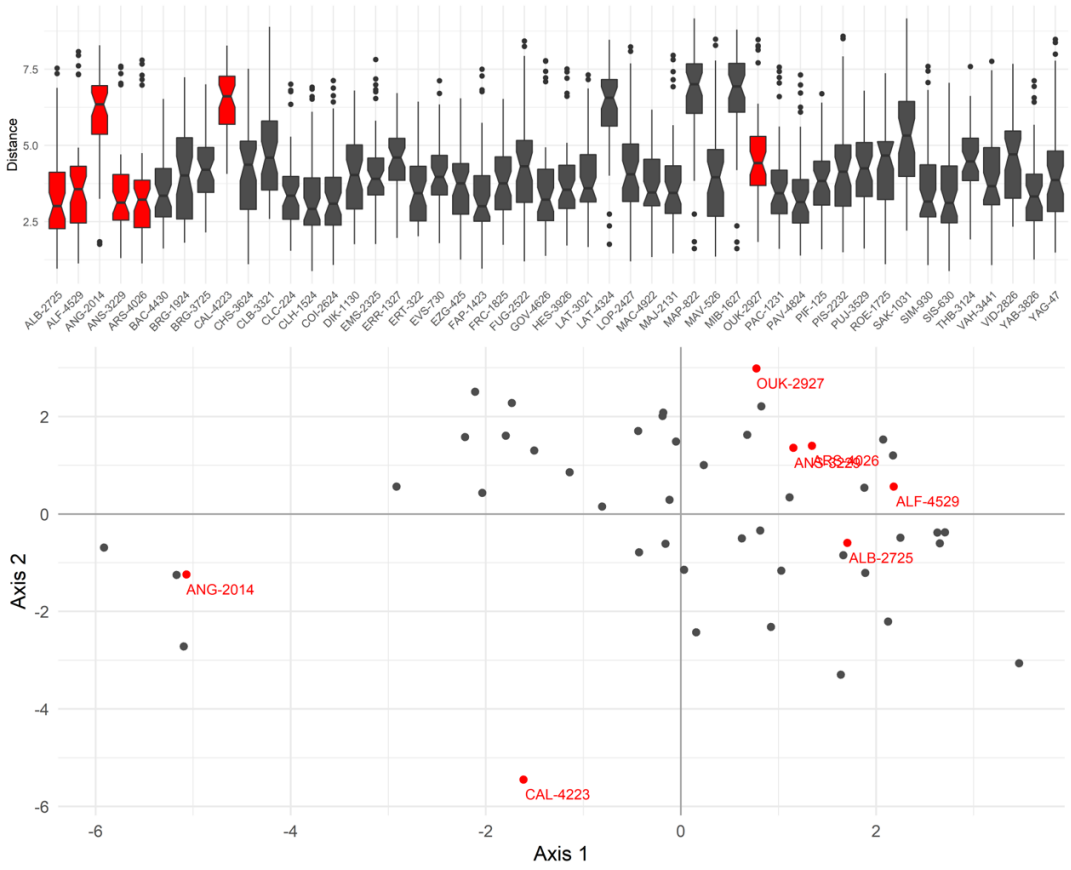

FIGURE 9.5: ***Top figure**: The boxplot of relative morphological distance from each subject in the dataset to the others. **Bottom figure**: The subjects' position in the morphological space projected on the plan of the first two Principal Components Analysis axis. The seven atlases (image volumes with manual segmentation) are colored in red.*

	ALB-2725	ALF-4529	ANG-2014	ANS-3229	ARS-4026	CAL-4223	OUK-2927	**Mean**
DSC								
JLF+CL	.917	.941	.900	.934	.941	.872	.933	.920
Generic UNet	.927	.893	.924	.951	.949	.849	.943	.920
Target-trained UNet	.913	.920	.902	.937	.931	.880	.934	.916
Fine-tuned UNet	.929	.926	.925	.948	.947	.915	.945	.934
HD (mm)								
JLF+CL	44.80	18.03	36.47	23.89	34.74	28.37	43.75	32.86
Generic UNet	100.00	63.33	96.01	70.55	47.62	105.82	58.02	77.34
Target-trained UNet	103.96	137.11	96.61	120.50	94.55	132.27	77.77	108.40
Fine-tuned UNet	42.48	102.89	100.64	56.60	46.81	57.77	58.73	66.56
MAD (mm)								
JLF+CL	1.33	0.99	2.66	1.13	1.01	1.72	1.19	1.43
Generic UNet	1.23	1.91	1.01	0.76	0.85	2.68	1.11	1.37
Target-trained UNet	2.79	2.29	1.88	1.00	1.18	3.05	2.18	2.05
Fine-tuned UNet	1.14	1.20	1.08	0.81	0.96	0.89	1.00	1.01
VS								
JLF+CL	.088	.037	.078	.049	.045	.088	.065	.064
Generic UNet	.106	.129	.073	.042	.045	.151	.042	.084
Target-trained UNet	.115	.071	.121	.041	.055	.070	.020	.070
Fine-tuned UNet	.095	.103	.088	.044	.035	.064	.026	.065

TABLE 9.3: *Quantitative evaluation of Joint Label Fusion with 3 closest atlases + Corrective Learning, generic 2D UNet with morphology-based data augmentation, target-trained UNet and fine-tuned UNet. Values with gray background marks the best validation score for each subject.*

.920). Although for most subjects, adding information is profitable and helps boost the precision of the model, for a subject considered *outlier* as CAL-4223 (the projection of the subject in the first 2 PCA plan is clearly at a relatively large distance from the others, cf. Fig. 9.2), it only over-fits the model to the cluster that is far from the test subject.

Figure 9.6 presents the visual results with generic UNet and fine-tuned UNet for CAL-4223, whose DSC increased from .849 to .915. The result of generic UNet is very noisy, with each muscle head separated into multiple parts, especially the *rectus femoris* (colored blue in the figure). The result is much more smooth and coherent after fine-tuning, morphology-wise, with less distanced errors.

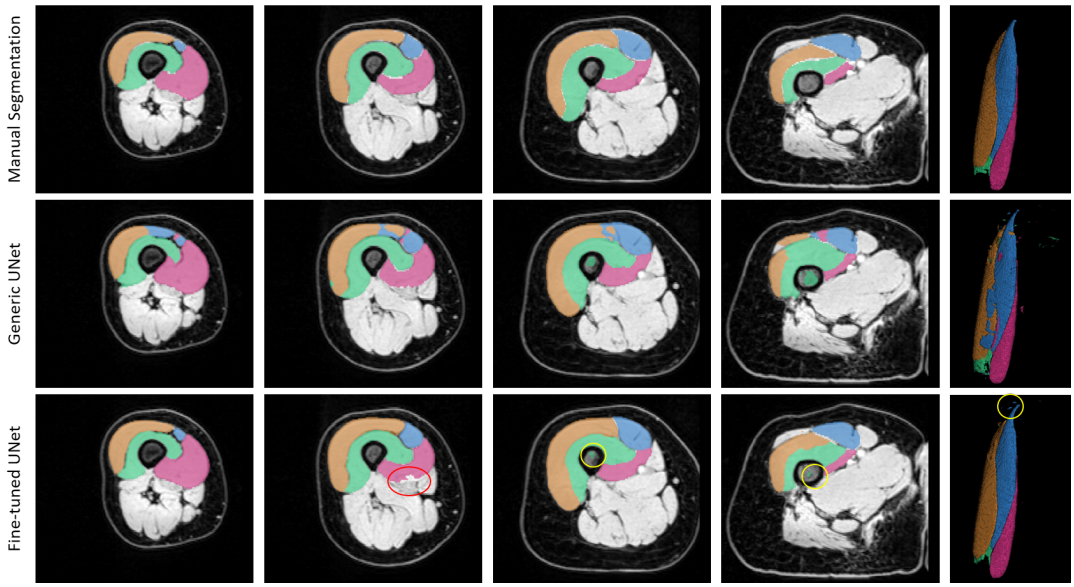

FIGURE 9.6: *Visual results of Generic UNet and Fine-tuned UNet, compared with the manual segmentation of CAL-4223. The last column is the 3D views of the three segmentations. Red circle marks the most visible error made by fine-tuned UNet. Yellow cirles mark the distanced errors that might be the source of a large HD.*

In terms of Hausdorff Distance (HD), the JLF+CL demonstrated the advantage of working fully in 3D, with a limited amount of distanced errors. However, the lower MAD of fine-tuned UNet shows that these errors that amplified HD are aberrant and can be corrected with simple postprocessing (Fig. 9.6, yellow circles).

9.5 Conclusion

In this chapter, we have presented our contributions on morphological features: the measurement and the strategies for both JLF and UNet. The features are computed from the center slice of a large 3D MR image volume of quadriceps muscles representing the subject's morphology. These features can be integrated into multiple automatic segmentation methods to either efficiently select atlases or artificially generate training data while conscious of morphological variance.

The morphological atlas selection helps reduce the number of atlases necessary to achieve the same results as using all available atlases, which is crucial in multi-atlas segmentation since the computation time is linearly proportional with the number of atlases. Meanwhile, a generic 2D UNet trained with data augmented based on morphological variation gives segmentations with similar quality as multi-atlas

segmentation upon reducing inference time from 24 hours to 45 seconds. In some cases where the generic UNet fails to produce adequate results, an extra step of fine-tuning that takes 5 to 10 extra minutes, based on morphological similarity, is advised and was proven to give remarkably improvement.

Conclusion

In this part, we have presented our contributions to the automatic segmentation of quadriceps using the MUST dataset. The MR images are first preprocessed to correct the signal inhomogeneity and adapt the left legs to the right legs. The processed images and the manual segmentations by medical experts served as the materials for our experiments.

First of all, we provided a complete analysis of Wang and Yushkevich's muti-atlas segmentation method with joint label fusion and corrective learning (JLF+CL). With optimized parameters, the method gave high-quality segmentation for most test subjects but necessitated a very high computation time. Meanwhile, this method showed some limitations when applied to subjects with a morphology different from those of the atlases.

Secondly, we proposed to replace the time-consuming JLF step in the framework of Wang and Yushkevich with weakly-supervised 2D UNet. The UNet is trained and validated with both manually annotated and automatically generated data. The data augmentation using random B-spline warping and deformable registration enhanced the morphological diversity in the training and validation sets and improved the segmentation quality. As a result, the weakly supervised UNet provided similar validation metric values as JLF+CL while reducing the inference time from 48h for JLF to 45s.

Finally, we introduced a morphological measurement and its applications to optimize presented segmentation methods: an atlas selection strategy for JLF to reduce computation time while conserving the segmentation quality, a data augmentation strategy to maximize the morphological variation in training and validation sets for UNet, a fine-tuning process added to UNet to improve the segmentation quality further.

In the next part, we will evaluate our proposed methods on different applications based on the MUST dataset and two other datasets from two other muscle studies.

PART IV

Applications and further analysis

Contents

Résumé

Alors que la partie III a détaillé nos contributions à la segmentation des muscles, dans cette partie, nous les appliquons à des problèmes proches nous permettant de mieux appréhender les capacités et limites de nos approches.

Ainsi, dans le chapitre 10 nous étudions les questions suivantes :

- Est-ce que nos méthodes permettent d'améliorer la segmentation quand un grand nombre d'atlas est disponible ?
- Comment se généralisent nos approches à la segmentation de l'autre jambe ?
- Nos segmentation sont elles assez reproductibles pour permettre une étude longitudinale ?
- Nos approches sont elles robustes à un changement dans la configuration d'acquisition IRM ?

Pour répondre à ces questions, nous utilisons, en plus des images de MUST, deux autres jeux de données issus de projets connexes :

- HAMMER : Une étude des lésions du muscle ischio-jambier du Centre Hospitalier Universitaire (CHU) de Saint-Étienne et le point central du projet de doctorat du docteur Sylvain Grange, radiologue très impliqué dans le projet MUST.
- SHOULDER : Segmentation des muscles de l'épaule. Un projet réalisé en tant que stage de Malick Kandji, étudiant en Master de l'INSA Lyon, en collaboration avec les Hospices Civils de Lyon et Hôpitaux universitaires de Genève.

Ce chapitre montre la pertinence de nos développements pour la segmentation de différents groups musculaires en IRM. Il montre aussi que les erreurs de segmentation obtenues en longitudinal, sur les images de MUST, sont plus faibles que la segmentation par recalage d'un instant sur un autre instant, et que nos développements restent pertinents, même quand on dispose d'un grand nombre d'atlas en termes de diminution de temps de calcul.

Enfin, le chapitre 11 résume l'étude longitudinale, sur la base MUST, de l'évolution des caractéristiques musculaires observées à l'aide de l'IRM. D'abord en vérifiant la pertinence de la mise en correspondances de caractéristiques radiomiques extraites localement sur des séquences IRM différentes et donc potentiellement non alignées. Puis en analysant muscle par muscle les corrélations observées au cours du temps ainsi que leurs significativités statistiques. Ce dernier chapitre s'appuie sur 2 contributions personnelles dans des journaux internationaux.

Introduction

In this part, we will apply our contributions proposed in Part III to related problems that allow us to better understand the capacities and limitations of our proposition. In Chapter 10, we study the following cases:

- How can our approaches improve segmentation when more atlases are available?
- How do our methods perform when dealing with both legs?
- Is our segmentation precise enough to allow longitudinal study?
- Can we segment, with few atlases, other muscles than quadriceps?
- Is our approaches robust to small MRI acquisition setting changes?

To answer these questions, we will evaluate the performance of our methods, not only on the MUST dataset but also on the datasets of two other projects that involved muscle study:

- HAMMER (Sec. 10.5): A study of hamstring muscle injury of Centre Hospitalier Universitaire (CHU) de Saint-Étienne and the focus of the Ph.D. project of Sylvain Grange, M.D., a radiologist highly involved in the MUST project.
- Shoulder muscle segmentation (Sec. 10.1): a project carried out as an internship of Malick Kandji, a Master student of INSA Lyon, where I acted as co-supervisor, in collaboration with Hospices Civils de Lyon and Hôpitaux universitaires de Genève.

The final chapter, Chapter 11, is a summary of 2 papers that analyze locally the evolution of the quadriceps muscles during an ultra-marathon using the MUST dataset.

CHAPTER 10

Muscle segmentation based on MRI data

This chapter presents different applications of the studied automatic segmentation methods on different MRI datasets of human muscle groups.

First, we employed our approaches in an automatic segmentation study to assess fat level in muscles of patients with lesions in the rotator cuff tendon (Sec. 10.1). This application allows us to investigate the importance of the number of atlases on the proposed methods and the generalization to other muscle groups.

Second, using few atlases, we evaluate the generalization capability of our approaches to the quadriceps segmentation on both legs (Sec. 10.2), then in longitudinal context (Sec. 10.3), and to the segmentation of another upper leg muscle group (hamstring, Sec. 10.4), all with the MUST dataset.

Finally, we tested our approaches on another dataset where the MRI acquisition system was different from the MUST dataset (Sec. 10.5).

10.1 Dataset with more atlases: Rotator cuff segmentation

10.1.1 Context

Aging and repeated stresses can cause partial or complete rupture of the tendon of one or more shoulder muscles, resulting in pain and sometimes severe functional limitations. The treatment of these lesions can be medical or surgical, depending on the age, the pains, the quality of the injured muscle, and the functional deficit. If it is a surgery, it is proven that its success will largely depend on the muscle trophicity and its fatty ratio.

A project was proposed with the objective of automatically quantifying muscle/fat ratio and assessing muscle quality before taking charge of patients, based on muscle segmentation. This project was carried out as an internship of Malick Kandji, a Master student of INSA Lyon, in collaboration with Hospices Civils de Lyon and Hôpitaux universitaires de Genève.

10.1.2 Data

A series of patients with lesions in the rotator cuff tendon was evaluated, before treatment, using MRI Dixon sequences which allow intramuscular fat quantification. The project's dataset consists of 51 patients noted from P1 to P51. For each patient,

we have eight types of 3D MRI sequences with their phases. P40 (patient N°40) was removed since many of their sequences were missing. For now, we are interested only in the Dixon e8 sequences, identified by our radiologist as the *easiest* to segment muscles manually and covering the adequate volume to observe muscles of interest in all patients. The images are of size $320 \times 320 \times 60$ voxels with the voxel size of $0.6875 \times 0.6875 \times 2\,\text{mm}^3$.

10.1.2.1 Manual segmentation

Manual segmentation were done based on these sequences for **27 patients** among the 50 (e.g., Figure 10.1). There are 5 muscle heads in the rotator cuff muscle group: *Subscapularis* (SBC), *Supraspinatus* (SPR), *Infraspinatus* (IFR), *Teres minor* (TMN), and *Deltoid* (DTD).

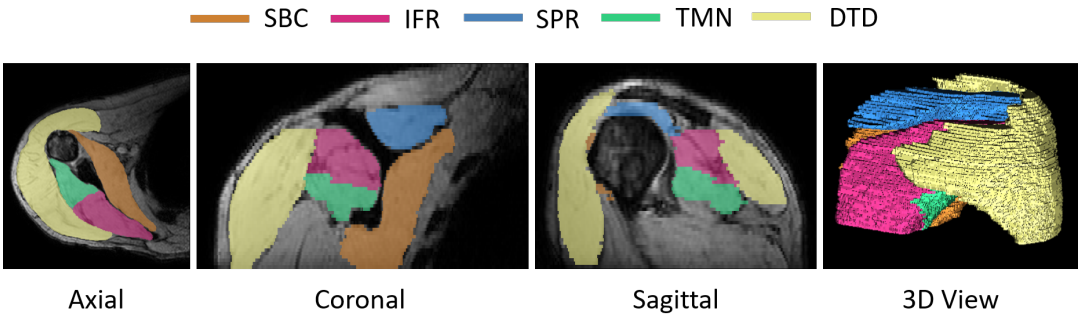

FIGURE 10.1: *Manual segmentation of the rotator cuff muscle group of a patient in the rotator cuff dataset. Abbreviation: SBC - Subscapularis, SPR - Supraspinatus, IFR - Infraspinatus, TMN - Teres minor, DTD - Deltoid.*

10.1.2.2 Preprocessing

Compared to the MUST dataset, extra preprocessing steps were necessary for the rotator cuff dataset as the images are much more heterogeneous (images acquired from different angles, strong bias field, ...). Since each patient has only one side of their shoulders examined depending on the position of the lesions, we have images of both right and left shoulders in the dataset. Therefore, one preprocessing step is to transform all images to right shoulder images to increase the size of the dataset. The preprocessing pipeline consists of 5 steps and is presented in Figure 10.2.

1. *Bias field correction* with N4 algorithm (Tustison et al., 2010)
2. *Normalization*: All images were normalized to have the same data dynamics as P12 (randomly chosen as reference). The normalization was done only on a region of interest (ROI) defined by a simple process of threshold and mathematical operations (see Fig. 10.2).
3. *Resampling & Padding*: To prepare for the registration step, the images were resampled to isotropic resolution ($0.6875 \times 0.6875 \times 0.6667\,\text{mm}^3$) and padded with large 0 border. This prevents the lost of information due to registration.
4. *Rigid registration*: Since images in the dataset were not acquired in the same direction; registration is necessary to employ automatic segmentation approaches, more particularly, deep learning methods which caught context information. Right shoulder images were rigidly registered on a template patient (chosen

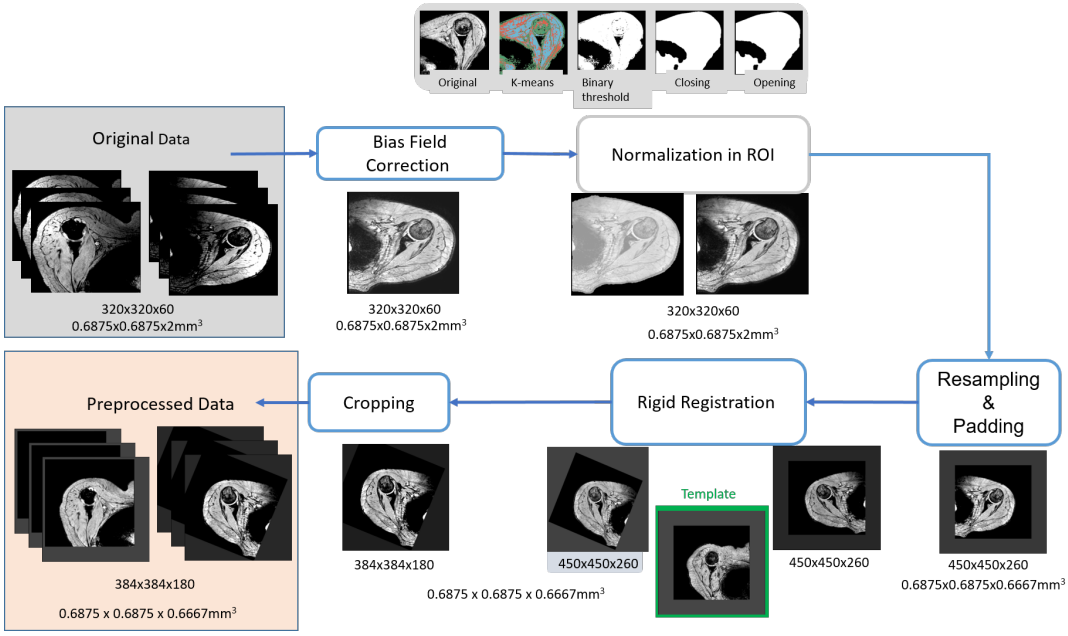

FIGURE 10.2: *Preprocessing pipeline for the rotator cuff dataset, with the size and the resolution of images at each step.*

randomly) - P15. Left shoulder images were first registered on P4 (also a left shoulder image), then mirrored to resemble a right shoulder, and finally, registered on P15.

5. *Cropping*: After registration, images were cropped to eliminate the part where locates no helpful information, thus accelerating the training process and reducing memory usage.

10.1.3 Automatic segmentation

Since there are 27 manual segmentations, the LOO strategy and data augmentation (see Sec. 3.2) is not necessary. We divided the dataset into two subsets: testing subset with 10 patients and training subset with 17 patients.

One of the objectives here is to implement and analyze different deep learning networks' performance on the dataset. The tested networks are 2D UNet (Ronneberger et al., 2015), 2.5D UNet (Haque et al., 2019), Mask R-CNN (He et al., 2020), ResUNet (Zhang et al., 2018) and UNet 3+ (Huang et al., 2020). The detailed results is not relevant to this dissertation. Briefly, despite the high expectation for UNet 3+, among the 2D networks applied on axial slices, the ResUNet with ResNet101 backbone and deep supervision (Lee et al., 2015) gave the best performance in term of DSC ($.892 \pm .040$) and MAD (3.22 ± 2.80 mm). Overall, the best result obtained was with 2.5 UNet (2D UNets on 3 axes followed by majority voting), with DSC at $.893 \pm .040$and MAD at 2.12 ± 0.70 mm.

This dataset with a larger number of manual segmentations allowed us to further analyze the impact of the number of atlases on Wang et al.'s multi-atlas segmentation method, with and without the help of morphological measures. The morphological features were measured at the same axial slice after preprocessing, where 4 over 5 muscle heads are visible for all patients (the 5 muscle heads do not appear

together at any axial slice). Among the 10 test subjects, there are 3 subjects that always have non-adequate results with JLF due to deformable registration failure. We have not yet been able to identify the reason behind these failures, so we removed these 3 patients from the test set for now. We observed the results of JLF with 3, 4, 5, 7, 9, and 11 atlases, selected randomly or based on our morphological distance. The results are summarized in Figure 10.3.

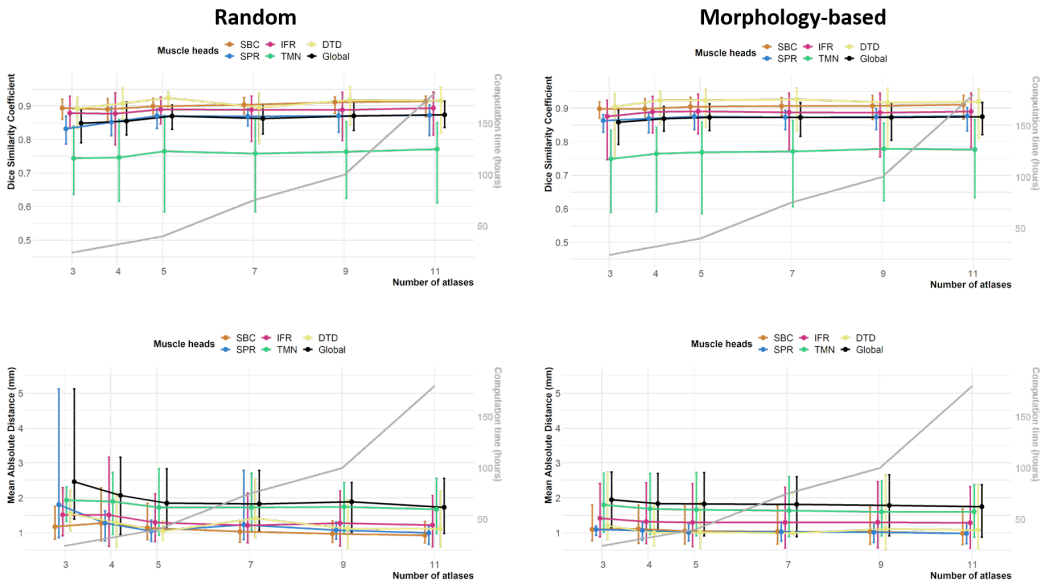

FIGURE 10.3: *Results of Wang et al.'s JLF segmentation method on rotator cuff dataset, with and without morphology-based atlas selection. Each color represents a muscle head, with the color black representing the global score over the entire muscle group. Each muscle head is presented with the average value and a vertical bar limited by the minimal and the maximal values over 7 subjects.*

The evolution of DSC and MAD in function of the number of atlases shows that the morphology-based atlas selection helps the results converge faster and reduces the quality gap among the different numbers of atlases. With a random selection, adding atlases does not always mean improving the result, while it is the case for morphology-based selection. With 3 atlases, we can observe a clear improvement in MAD when using morphology measure, meaning smaller spatial errors. The morphological features have been once again proven to be pertinent in multi-atlas segmentation of muscle groups with large morphological variance in the dataset.

10.2 Generalization to quadriceps segmentation on both legs - MUST dataset

Our experiments in Part III exploited only right leg images of the MUST dataset since the 7 volumes with manual segmentation of quadriceps are of right legs. This section will evaluate automatic segmentation of quadriceps on both right and left legs using the manually segmented center slice of each image. From this section onwards, JLF+CL is short for Wang and Yushkevich's multi-atlas segmentation with joint label fusion and corrective learning. Except for when specified differently, the multi-atlas segmentation is based on three morphologically closest atlases, and corrective learning is based on the model learned on all atlases (see Sec. 9.2). Additionally, UNet here refers to the fine-tuned UNet presented in Section 9.4.2.

To extract the morphological features of 48 subjects in the MUST dataset for the study presented in Chapter 9, a medical expert in our team has manually segmented the quadriceps, the hamstrings, and the femurs on both legs at the center axial slice of a 3D spoiled gradient echo (3D GRE) sequence, acquired at the time point Pre (before the race, see Sec. 2.3). This center position is planned at a 15 cm distance from the upper part of the patella (see Sec. 2.3). In total, we have $48 \times 2 = 96$ axial slices with manual segmentation at 3D GRE sequences' resolution. These manual segmentations can now be used for quality control of the automatic segmentation methods.

All of the atlases for JLF+CL and the training data for UNet are images of right legs. For the multi-atlas approach, the atlases used to segment a left leg are chosen independently from the right leg of the same subject since we cannot assume that both legs have the same morphology. The projection of left and right legs of all subjects on the plan of the first two PCA axis (Fig. 10.4) confirms that the two legs of the same subjects are not always *morphologically close* during the MRI acquisition. The validation metrics of the two automatic segmentation methods are reported in Table 10.1.

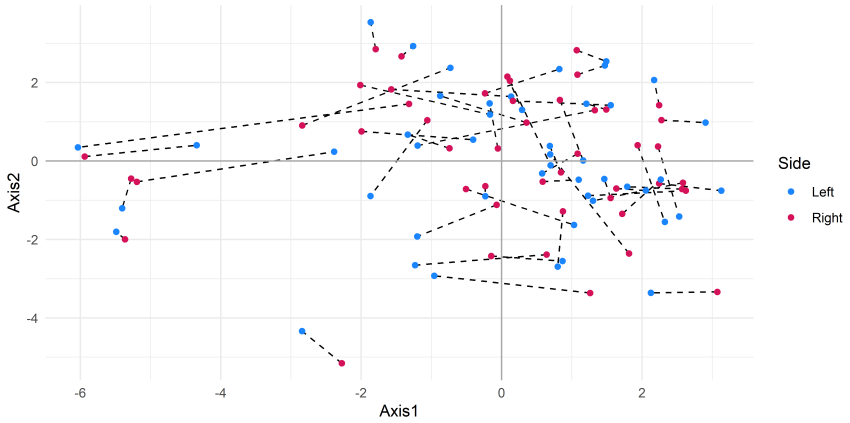

FIGURE 10.4: *Representation of both legs of the MUST dataset's subjects on the plan of the first two Principal Components Analysis (PCA) axis of morphological data. The two legs of a subject are linked with a dash line.*

Method	*DSC*	*DSCw*	*HD (mm)*	*MAD (mm)*	*VS*
JLF + CL					
Left legs	$.937 \pm .027$	$.939 \pm .023$	10.43 ± 6.26	1.30 ± 0.79	$.066 \pm .046$
Right legs	$.934 \pm .021$	$.937 \pm .019$	11.62 ± 5.49	1.38 ± 0.69	$.074 \pm .036$
Both	$.936 \pm .024$	$.938 \pm .021$	10.98 ± 5.92	1.34 ± 0.74	$.070 \pm .041$
UNet					
Left legs	$.937 \pm .024$	$.939 \pm .024$	10.39 ± 6.01	1.32 ± 0.69	$.069 \pm .044$
Right legs	$.940 \pm .024$	$.941 \pm .022$	12.33 ± 8.45	1.43 ± 1.27	$.059 \pm .039$
Both	$.938 \pm .024$	$.940 \pm .023$	11.28 ± 7.26	1.37 ± 0.99	$.065 \pm .042$

TABLE 10.1: *Quantitative evaluation of quadriceps segmentation at center axial slice for all 48 subjects of MUST dataset with the training data consist only of right leg images. Results are reported with mean $\pm$ sd.*

The results are slightly superior to the results obtained in Chapter 9 (average DSC at .934 for UNet and .920 for JLF+CL over 7 image volumes) since the center

axial slice passes by the middle part of all muscle heads, where the boundary is better defined than at the extremities. The average validation scores are very close to the inter-expert scores reported in Section 6.2 (Average DSC at .910). With simple postprocessing, such as a morphological opening with a structuring element of minimal size to remove aberrant errors, these automatic segmentations would be accurate enough for image features extraction and statistical analysis (see Chapter 11).

10.3 Robustness study on segmentation of longitudinal images of quadriceps - MUST dataset

In order to evaluate the robustness of the automatic segmentation methods on longitudinal data, we compared the segmentation validation metrics of the race finishers' segmentation at three different time points (see Sec. 2.3): Pre (before the race), Post (at the arrival) and Post+3 (48-72h after the race). There are 4 finishers among the 7 subjects with manual right quadriceps segmentation at the 3 time points. The segmentation of a subject's image at Post and Post+3 with UNet and JLF+CL is entirely independent of their manual segmentation at Pre.

Since we have 17 manually segmented 2D slices for each volume at Post and Post+3 (see Tab. 2.1), only DSC and VS are comparable as the other metrics are computed in 3D. Moreover, for Post and Post+3, we also compared the segmentation obtained by applying the automatic segmentation methods on the image volume at the time point and the one obtained by registering the segmentation at Pre to the time point in question. The results are reported in Table 10.2.

	DSC				VS			
Time point/Method	*ALB*	*ALF*	*ARS*	*OUK*	*ALB*	*ALF*	*ARS*	*OUK*
Pre								
JLF+CL	.912	.950	.951	.938	.101	.036	.041	.072
UNet	.918	.936	.950	.948	.112	.089	.043	.033
Post								
Manual Pre registered	**.914**	**.929**	**.934**	**.921**	**.111**	**.070**	**.066**	**.039**
JLF+CL Pre registered	.926	.921	.933	.914	.022	.060	.077	.067
UNet Pre registered	.939	.926	.940	.926	.026	.044	.028	.040
JLF+CL	.924	.922	.939	.909	.077	.052	.070	.090
UNet	.939	.903	.939	.927	.020	.103	.027	.062
Post+3								
Manual Pre registered	**.919**	**.924**	**.934**	**.937**	**.085**	**.060**	**.058**	**.038**
JLF+CL Pre registered	.923	.917	.933	.933	.055	.070	.078	.062
UNet Pre registered	.935	.925	.945	.942	.034	.041	.023	.044
JLF+CL	.921	.918	.941	.937	.044	.072	.076	.050
UNet	.935	.903	.946	.944	.037	.080	.016	.053

TABLE 10.2: *Validation metrics on longitudinal data of 4 subjects (ALB-2715, ALF-4529, ARS-4026, and OUK-2927). Only the first three letters of the subjects' code names are inscribed here to reduce the table's size for the sake of readability. The metrics were computed based on the same 17 slices for Pre, Post, and Post+3. Gray cell indicates the best automatic segmentation result for a subject at a time point. Bold values indicate the metrics of the manual segmentation at Pre registered to the time point in question.*

Overall, all the DSC scores are in the same value range (over .900). In all the cases, either the segmentation at Pre of UNet registered to Post and Post+3 or the segmentation of UNet directly at these two time points gives the best validation scores. These scores are mostly close to or better than the manual segmentation at Pre registered to the succeeding time points (bold values in Table 10.2).

These results confirm the applicability of the proposed automatic segmentation methods to longitudinal data.

10.4 Application to hamstrings segmentation - MUST dataset

We have been focusing on quadriceps segmentation of the MUST dataset since the quadriceps is known as the most affected muscle group due to eccentric effort during downhill running (see Sec. 2.1.1). Meanwhile, another skeletal muscle group in the upper thigh which is the hamstrings is also in the interest of the researchers at CHU Saint-Étienne and CREATIS laboratory. To fully exploit the MUST dataset and to prepare for another research project (see Sec. 10.5), the hamstrings in 7 left leg images were manually segmented. To accelerate the process, we first estimated the segmentations with an automatic segmentation method, such as deformable registration or multi-atlas segmentation, then they were precisely corrected by the medical expert.

Hamstrings is an upper thigh muscle group consisted of 4 muscle heads: short head of *Biceps Femoris* (SHBF), long head of *Biceps Femoris* (LHBF), *Semitendinosus* (ST), and *Semimembranosus* (SM). Figure 10.5 shows an example of a manual segmentation of hamstrings superposed on a T1W image.

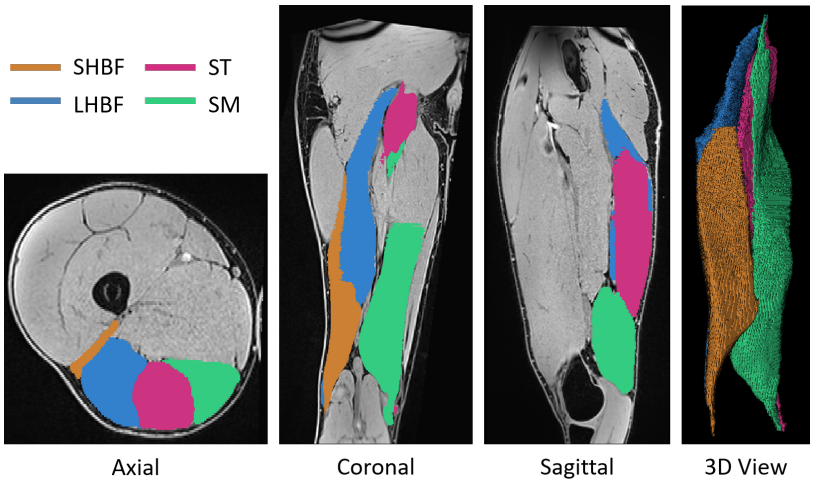

FIGURE 10.5: *Manual segmentation of hamstrings muscle group, superposed on T1W image, of a patient in the HAMMER dataset (Sec. 10.5). Abbreviation: SHBF - short head of Biceps Femoris, LHBF - long head of Biceps Femoris, ST - Semitendinosus, SM - Semimembranosus.*

With 7 manual segmentations from 7 left leg images, we employed the Leave-One-Out strategy to evaluate the performance of the automatic segmentation methods on hamstrings. All the images used in this experiment, including atlases, targets for data augmentation, and test images, are all images of left legs. The results are reported in Table 10.3.

We achieved results of similar quality as in the case of quadriceps segmentation with an average DSC of .949 and .926 for morphology-based fine-tuned UNet and

	ALB-2725	ALF-4529	ANG-2014	ANS-3229	BRG-1924	CAL-4223	YAG-47	**Mean**
JLF+CL								
DSC	.928	.933	.935	.940	.921	.919	.907	.926
HD (mm)	24.72	21.16	37.59	32.84	25.99	54.58	63.51	37.20
MAD (mm)	0.90	0.75	1.27	0.86	1.04	0.97	1.92	1.10
VS	.018	.037	.049	.035	.046	.048	.077	.045
UNet								
DSC	.940	.955	.953	.955	.938	.950	.949	.949
HD (mm)	60.29	21.51	24.42	221.96	81.20	63.78	220.88	99.15
MAD (mm)	1.15	0.47	0.50	0.52	0.97	0.57	0.59	0.68
VS	.015	.019	.033	.027	.043	.018	.033	.027

TABLE 10.3: *Quantitative evaluation of Joint Label Fusion with 3 closest atlases + Corrective Learning and fine-tuned UNet for hamstrings segmentation.*

JLF+CL, respectively. While HD value is high for UNet, the observed small MAD reassures the segmentation quality and suggests that the high HD is due to distanced aberrant errors. Indeed, the SHBF is not present in the top one-third of the image volume (see Fig. 10.5, coronal view), and as the UNet works in axial 2D, it sometimes creates a small island of SHBF label in the slices where the SHBF must not be present.

A similar evaluation as in Section 10.2 was done with the hamstrings at the center axial slice. The results are reported in Table 10.4. The average metrics at the center slices are, on average, inferior to the ones evaluated for the 3D volumes reported in Table 10.3. It is due to the fact that the center axial slice is very close to the top of the SHBF where the muscle head's axial surface is small, which makes every mis-segmented pixels have a larger weight in the global evaluation metrics. Moreover, from observation, the extremities of a muscle head are often where the muscle boundary is unclear, complicating even the manual segmentation. Figure 10.6 shows the MRI images of a subject with the segmentation of hammer muscle heads, both manual and automatic. The SHBF (colored orange) is much smaller than the others, and its extremity zone (circled in red) with slightly lower intensity is where the automatic segmentation methods made mistakes. The errors are small, but due to the small surface of SHBF at this specific axial slice, the DICE score of this muscle head is much lower than the global DSC (average of all muscle heads' DICE scores).

Method	*DSC*	*DSCw*	*HD (mm)*	*MAD (mm)*	*VS*
JLF + CL					
Left legs	$.864 \pm .047$	$.891 \pm .031$	12.98 ± 5.38	2.16 ± 0.99	$.148 \pm .067$
Right legs	$.857 \pm .046$	$.882 \pm .040$	12.62 ± 5.15	2.29 ± 0.96	$.155 \pm .079$
Both	$.860 \pm .046$	$.886 \pm .037$	12.78 ± 5.23	2.23 ± 0.97	$.152 \pm .074$
UNet					
Left legs	$.880 \pm .060$	$.902 \pm .046$	10.19 ± 4.98	1.83 ± 0.92	$.116 \pm .068$
Right legs	$.873 \pm .053$	$.899 \pm .034$	11.00 ± 6.07	2.02 ± 0.84	$.131 \pm .067$
Both	$.876 \pm .056$	$.900 \pm .040$	10.63 ± 5.58	1.94 ± 0.88	$.124 \pm .068$

TABLE 10.4: *Quantitative evaluation of hamstring segmentation at center axial slice for all 48 subjects of MUST dataset. Results are reported with mean $\pm$ sd.*

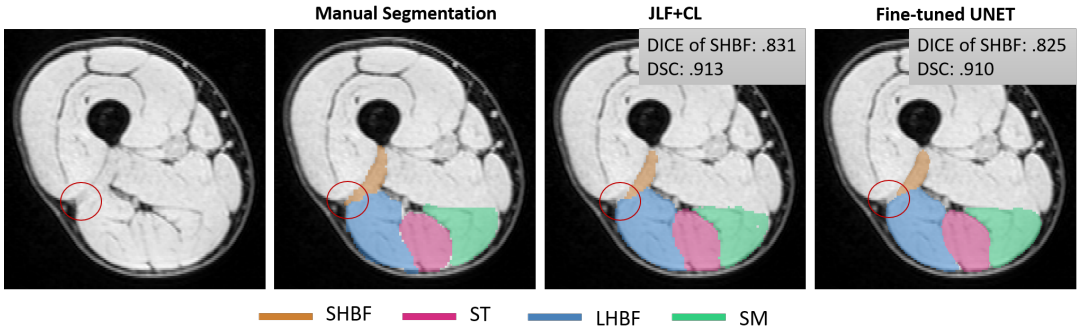

FIGURE 10.6: *Center slice of a subject with the hammers segmented with different methods. Red circles mark the zone where the automatic segmentation methods often make mistakes. Abbreviation: SHBF - short head of Biceps Femoris, LHBF - long head of Biceps Femoris, ST - Semitendinosus, SM - Semimembranosus.*

10.5 Generalization to new data with different acquisition parameters: HAMMER dataset

This section's experiment uses the atlases, the corrective models, and UNet models trained with the MUST dataset to segment a new dataset called HAMMER, whose subjects are also athletes but of different sports. The major difference here with the previous sections is the change in the acquisition system from 1.5T to 3T.

10.5.1 HAMMER case study

HAMMER (*HAMstring MEchanics and mRi*) is a prospective, multi-centre, non-interventional cohort study (Nguyen et al., 2019a). The main objective is to analyze the association between hamstring injury location and injury mechanism, hence the necessity of muscle segmentation.

The Comité de Protection des Personnes d'Ile de France V (CPP IDF 5: 17059) approved the study protocol (No. ID-RBC: 2017-A03433-50). It consisted of performing an MRI scan of the thighs in a patient with a clinical suspicion of hamstring muscle injury and requesting the patient to fill a questionnaire describing their injury's circumstances and mechanisms.

Patients were recruited from two radiology centers: the Clinique Mutualiste and the Centre Hospitalier Universitaire (CHU) de Saint-Étienne. Patients were referred for an MRI scan of the thigh by their attending physician or by a sports physician following suspicion of a hamstring muscle injury during sports practice. The inclusion criteria were as follows: leisure/amateur, high-level or professional athletes, aged 18 to 50 years old, referred by their attending physician, their sports physician or coming of their own free will to a medical imaging facility to perform an MRI scan of the thigh following the suspicion of an acute hamstring injury that occurred during the practice of sports and that was less than 21 days old. The exclusion criteria were as follows: female athletes (to limit bias related to potential differences in injury between male and female athletes (Edouard et al., 2016)), recurrences, patients who had a history of anterior cruciate ligament surgery with hamstring tendon removal, a hamstring injury occurring outside of sports practice, an inability to understand the French language, and patients with no MRI abnormality and grade 0 of the modified Peetrons MRI classification (Peetrons, 2001).

The number of patients to date is 54, and the data collection is still in progress.

10.5.2 Data

MRI acquisition

The MRI examination takes place within a maximum of 21 days after the onset of the lesion. For patients' data to be included in the study, MRI imaging must have been performed on MRI machines less than 5 years old and with a magnetic field of at least 1.5 T.

For the MRI examination, axial and coronal slice sequences with T1, T2-weighting, and fat saturation are used. The field of view covers the proximal musculotendinous insertion at the ischial tuberosity and the distal insertion at the fibula's head and the proximal end of the tibia. Considering the coronal T1 Dixon Water-only images, which must be at a similar resolution as the T1-Water (T1W) images of the MUST dataset, the image volume dimension is $384 \times 456 \times 224$ voxels with the voxel size of $1.1 \times 1.1 \times 1.1\,\text{mm}^3$. Figure 10.7 shows, as an example, a T1W image of the HAMMER dataset, acquired with the 3T MRI scanner of CHU de Saint-Étienne.

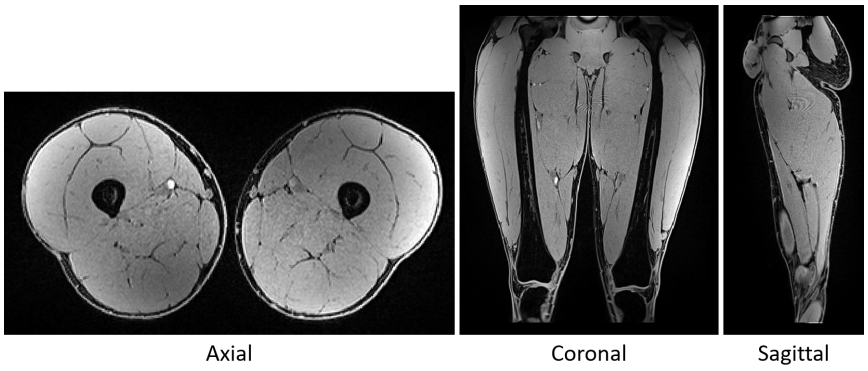

FIGURE 10.7: *Three views of a coronal 3D T1 Dixon Water-only sequence of the HAMMER dataset, acquired with a 3T MRI machine.*

Preprocessing

As the HAMMER dataset is very similar to the MUST dataset regarding the quality of images and the acquisition position (upper thigh), we aim to segment the former's images based on the manual segmentations from the latter. The main difference is that MUST images were acquired with a 1.5 T MRI machine instead of 3T as for the HAMMER images.

A similar preprocessing procedure as the one described in Section 6.1, was applied, including left-right separation, flipping of left leg images, N4 bias field correction, and gray-scale standardization based on the right leg image of ALB-2725 of the MUST dataset. An extra step of resampling was added to obtain images with the same resolution as in the MUST dataset.

Manual segmentations

An automatically-obtained segmentation of an image in the HAMMER dataset was manually corrected by a medical expert to evaluate our methods' performance. The segmentation includes both quadriceps and hamstrings muscle groups.

10.5.3 Experiments & Results

We employed both Joint Label Fusion with three closest atlases and fine-tuned UNet, using atlases from MUST dataset, to segment the quadriceps and the hamstrings of a subject from the HAMMER dataset, whose segmentation was manually corrected for qualitative evaluation. The results are presented in Figure 10.8.

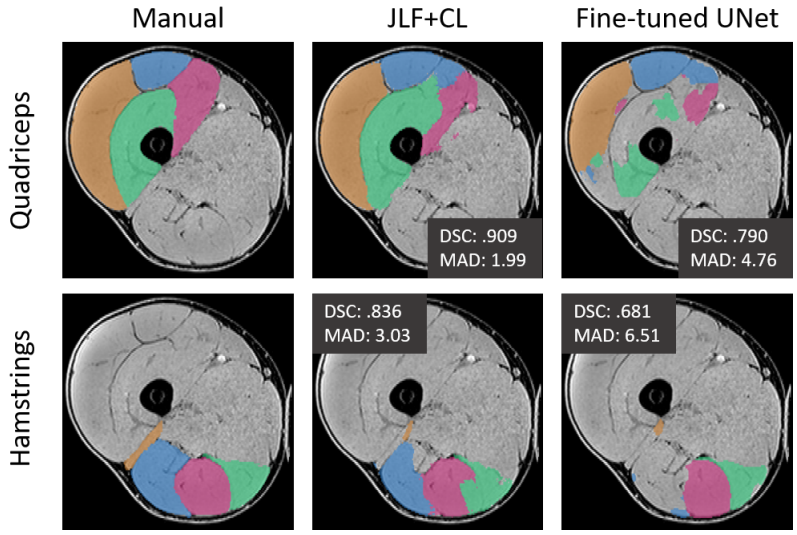

FIGURE 10.8: *Quadriceps and hamstrings automatic segmentation, using atlases from MUST dataset for training, of a subject in HAMMER dataset, compared with medical expert's manual segmentation.*

The automatic segmentation methods did not perform well on the target image: the UNet failed to create coherent segmentation while the JLF+CL made some big mistakes at the muscle boundary. To understand the results, we investigated the projection of the target subject on the PCA plans (the plan of the 2 first axis and the plan of the first axis and the third axis) of the MUST dataset's morphological data (Fig. 10.9): The HAMMER subject is close to the atlases in terms of quadriceps morphology and is separated from the cluster of MUST subjects along the third axis in terms of hamstring morphology. Subsequently, for this specific subject, it is coherent that the JLF+CL had a more favorable outcome with the quadriceps than with the hamstring. However, in the case of quadriceps segmentation, given the close distance of the target subject to the atlases, the results must have been of higher quality, without such large morphological error. Furthermore, this analysis of the subject's morphological position among the MUST subjects does not justify the non-adequate results of UNet.

To make sure that the preprocessing does not have a negative effect on the results, we rerun the methods on the image after each preprocessing step. The results displayed in Figure 10.10 indicate that the failure of the segmentation methods does not originate from the preprocessing.

It is important to keep in mind that the images of the target HAMMER subject are acquired with a 3T MRI scanner, while the images of the MUST dataset came from a 1.5T machine. One possible explanation here is that the underlying difference in texture between 3T images and 1.5T images invalidated the UNet models trained with the MUST dataset. This difference may also reduce the performance of the registration process, which leads to less accurate results for JLF+CL.

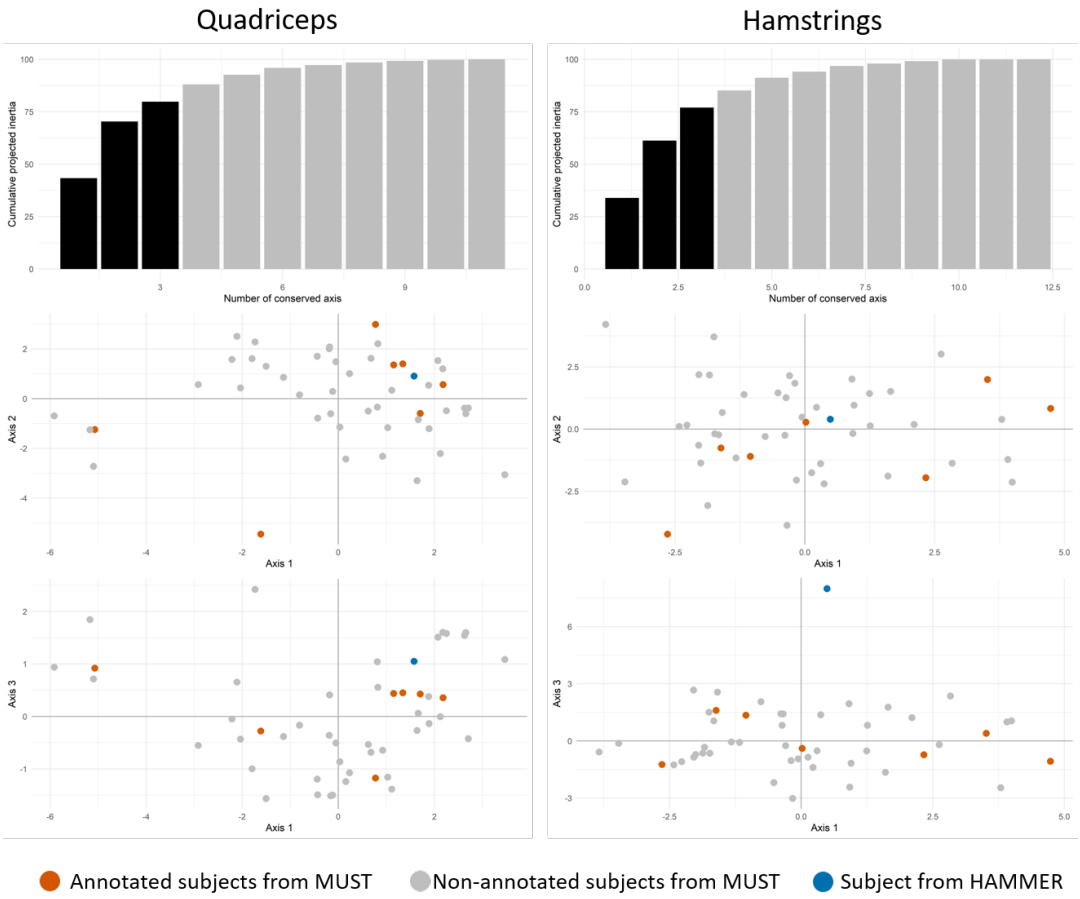

FIGURE 10.9: *Projection of the target subject (blue point) on the PCA plans of the MUST dataset's morphological data. First row: cumulative projected inertia plots. Second row: Projection on the plan of the first two PCA axis. Third row: Projection on the plain of the first and the third PCA axis.*

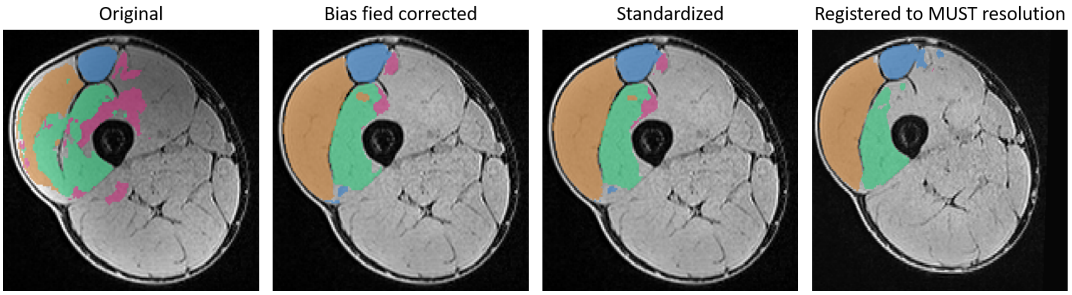

FIGURE 10.10: *Quadriceps segmentation produced by fine-tuned UNet for the original T1W image of the HAMMER subject and for the image resulted from each preprocessing step*

From the experiments, we understand that our approaches are sensitive to changes in MRI acquisition and that our morphology measures cannot evaluate this difference as they are based on the segmentation regions of MR images and not on MR images themselves. We can also assume that the morphology measure of quadriceps is not generalizable to hamstring and should be computed in a specific manner to adapt to hamstrings morphology.

10.6 Conclusion

This chapter presents the results of our automatic segmentation methods based on morphological features on multiple MRI muscle datasets. In general, the results proved the robustness of the methods in a longitudinal setting and on different muscle groups, as observed on shoulder and longitudinal segmentation of quadriceps.

The segmentation of a 3T MR image using atlases and models trained with the 1.5T images from the MUST dataset showed that the textural difference between 3T and 1.5T prevents the transfer of our pre-trained model to the new dataset. A *translation* between these two different types of signal needed to be studied to avoid the manual segmentation process each time there is a new dataset from an MRI scanner of different magnetic field strength.

CHAPTER 11

Longitudinal study on the MUST dataset

This chapter resumes two of our contributions: Nguyen et al. (2021a), and Nguyen et al. (2021b), which involve the analytical studies based on the muscle segmentations. The study flowchart was previously presented in Chapter 2, Figure 2.4.

First, a preprocessing framework is built based on the quality of the segmentation (Sec. 11.1). The radiomic features are then extracted from each muscle head based on the segmentation. Finally, a longitudinal statistical analysis based on these features was done to explore the link between image information and physiological change in the subjects (Sec. 11.3).

11.1 Data preparation & Feature extraction

Radiomics is a process that involves extracting and analyzing a large number of features from medical images. The features can be first-order statistics (distribution of voxel intensities), shape-based, or textural (Hatt et al., 2018). While all features can be analyzed with a standardized framework, an exhaustive interpretation of a textural feature requires a specific analytical problem for optimizing feature extraction parameters. In MUST dataset, radiomic features were extracted from quantitative MRI (qMRI) maps (T2, T2*, PDFF and χ) for each runner at each time point.

This section provides the data preparation steps that we judge necessary to extract image features accurately.

11.1.1 Distorsion among image sequences

A localized longitudinal analysis requires accurate segmentations. While radiomic features need to be extracted from many different image sequences, the segmentation is generally best performed on a given contrast. The correction of MRI distortion is a well-known challenge in MRI analysis (Walker et al., 2014; Rizzo et al., 2018), which makes feature extraction by superposing automatic segmentation obtained on one sequence on the other sequences questionable. Moreover, given the errors that could occur during the automatic segmentation, preprocessing the automatic segmentations is necessary to extract accurate image features.

In Nguyen et al. (2021a), we investigated the impact of the MRI distortion that appears between 3D T1 Dixon Water-only images and both spoiled gradient echo

and multi-echos T2 weighted spin-echo images when sequentially acquired (Fig. 11.1).

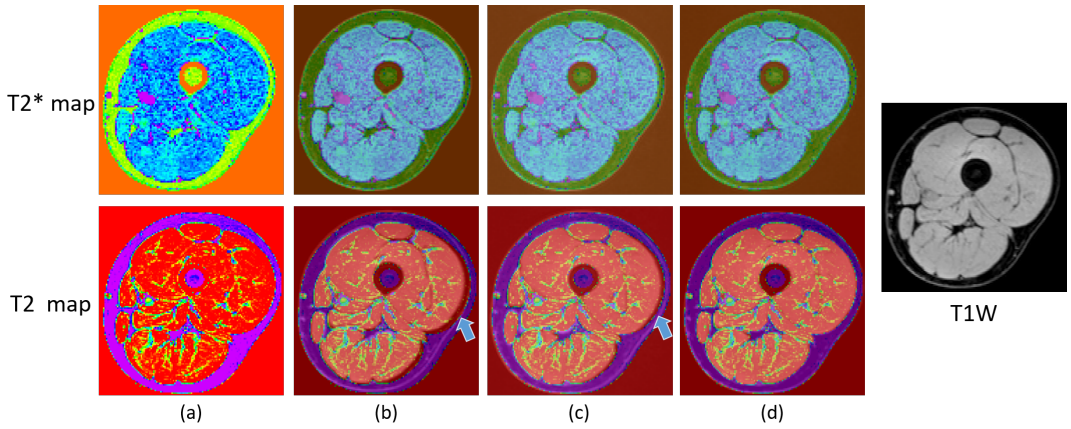

FIGURE 11.1: *T2 and T2* maps (a) and T1W image (on the right) of the left leg of a runner. By superposing T1W image on the quantitative maps, we can observe the original distortions between T1w and T2 maps (b), distortions after rigid registration (c), and after deformable registration (d). The arrows guides to the visually noticeable distortions. The distortions between T1w and T2* map are not as visible. T2 and T2* maps are displayed in colors for better visualization. Figure extracted from Nguyen et al. (2021a).*

The results suggest that classically used rigid registration is not optimal and that deformable registration should be preferred and should limit significant error in radiomic feature extraction. However, from the experiments on the MUST dataset, no significant change in radiomic statistic is observed whatever segmentation correction approach was applied; which indicates that radiomic features are not sensitive to segmentation refinement when considering large 3D regions.

11.1.2 Postprocessing of automatic segmentations

Proper segmentation of each muscle head is a crucial step of the analysis since it directly constrains the quality of the extraction of the quantitative data, on which further statistical analysis would be done. Thus, the segmentation must be accurate for all the data sets and all the muscles. Depending on the accuracy of the automatic segmentation method, more or less complicating of postprocessing procedure must be applied to obtain statistically accurate data for later analysis. Figure 11.2 shows an example of a segmentation refinement framework for Gilles et al.'s segmentations.

The framework consists of 4 steps: i) thresholding, ii) mathematical morphology application, iii) spatial resolution adaptation of the labels, and iv) histogram-based noise removal. These operations were applied to Gilles et al.'s automatic segmentation of each runner quadriceps at each time point to produce the final refined and more conservative segmentation to be used to retrieve statistical data from quantitative maps:

- *Thresholding*: Co-registration of the PDFF maps, derived from the me-3DGRE sequence, allows one to re-assigned all corresponding labels, erroneously classified in the fat tissue by the automatic segmentation, to a null value (same label as background mask) by simple thresholding if the corresponding PDFF was higher than 0.60.

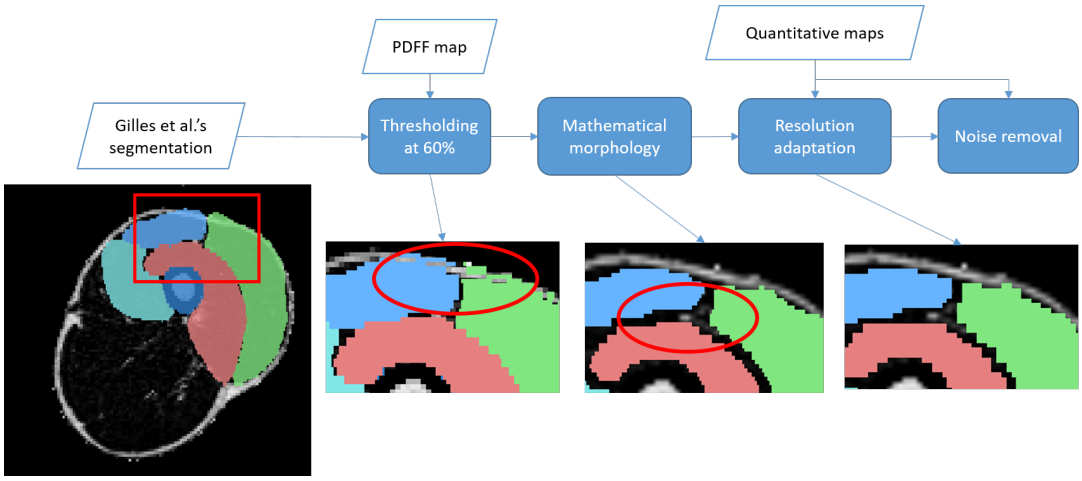

FIGURE 11.2: *Segmentation refinement framework for Gilles et al.'s segmentations. A segmentation is first cleaned from residual contamination by fatty tissues using thresholding. Mis-segmented pixels are further removed with a mathematical morphology step. The corrected labels are adapted to fit to the resolution of the studied quantitative map prior to the final step consisting of histogram-based noise removal. Figure extracted from Nguyen et al. (2021b).*

- *Mathematical morphology application*: The final segmentation map contained 8 label objects corresponding to the 8 studied muscles in the left and right quadriceps. We used a morphological erosion followed by a morphological opening with the smallest spherical structuring element possible (radius of 1 pixel) to smooth the label objects' contours, remove aberrant pixels, and separate the objects from each other.
- *Resolution adaptation*: Originally, the segmented label dimensions were these of the isotropic water image, calculated from the first anatomical isotropic 3D GRE. All quantitative maps from which we aim at extracting the information have 2 different and lower resolutions. Thus, it is necessary to create 2 other sets of labels with the same resolutions as the corresponding quantitative maps to correctly decipher the information from each pixel classified in each muscle head. The adaption is done by using the physical coordinates of each voxel in the quantitative maps and finding the label of the same coordinates in the label images.
- *Noise removal*: We noticed the random appearance of noisy pixels in our images and quantitative maps. Based on the histograms computed on the labeled pixels, we remove the labels of the 1% of pixels the less representative (which can be considered aberrant) from the two extremities of the histograms. The histograms of different labels are processed separately. The example of histogram before and after noise removal is in Figure 11.3.

This framework is specific to post-process Gilles et al.'s segmentations. For other automatic segmentation methods, one or more steps can be removed or added depending on the desired segmentation accuracy.

The processed automatic segmentations were then used for local radiomic feature extraction. The feature extraction can be done using either Vallières et al.'s MATLAB toolkit or Van Griethuysen et al.'s python package. The analyse presented in this thesis is based on the feature extracted using Vallières et al.'s MATLAB toolkit.

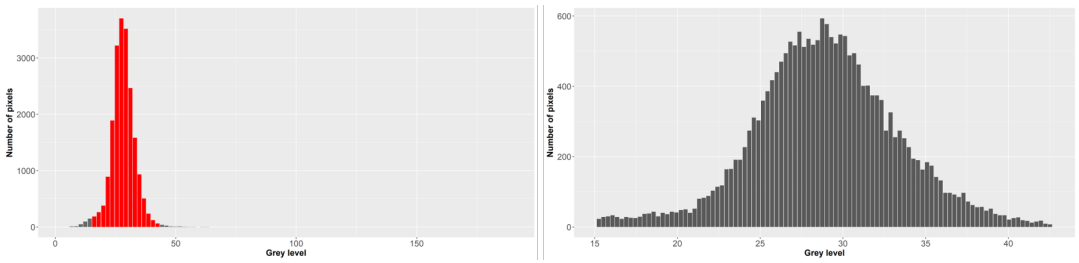

FIGURE 11.3: *Histogram of pixels labeled as right Vastus Lateralis in a T2*-maps before (right) and after (left) noise removal. The red part in the before histogram is conserved and resampled to compute the after histogram.*

11.2 Difference among muscle heads

The first step was to explore differences among all quadriceps muscle heads. Figure 11.4 shows a matrix of the t-test results when comparing quantitative MRI (qMRI) metrics (χ, PDFF, T2, T2*) between all muscle heads at the three MR acquisition time points (Pre, Post, Post+3).

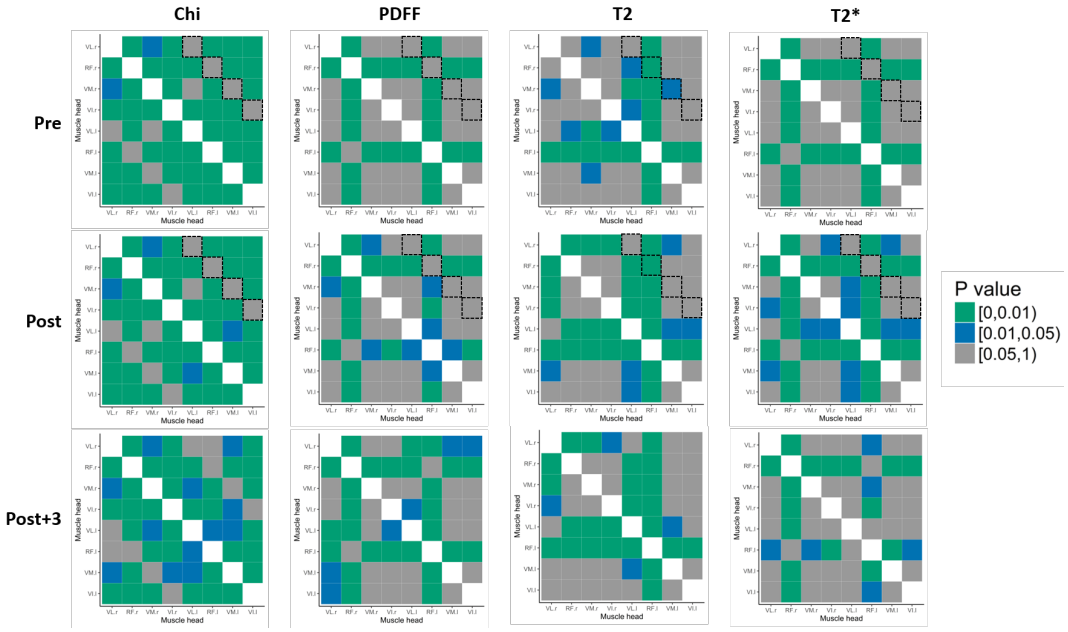

FIGURE 11.4: *t-test matrix with color-coded P-values for multiple comparisons of qMRI metrics (χ, PDFF, T2, T2*) between muscle heads at all three acquisition time points (Pre, Post, Post+3). Abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius, r – right, l – left. For easy viewing, dotted diagonals highlight inter-leg (right/left) comparisons of the same muscle heads at each time point, while all other boxes are intra-leg and/or inter-head comparisons. A P-value less than .05 indicates a significant difference between two muscle heads.*

When focusing only on right/left differences of the corresponding muscle heads, χ, T2* and PDFF showed similar tendency. At the same time, there was no significant difference between right and left muscle heads for χ and T2* at any time point, and PDFF exhibited a significant difference between vastus lateralis (VL) heads only at a single time point (Post+3). On the other hand, the T2 metric showed a different pattern with a significant difference between the rectus femoris (RF) heads of the two legs at the three time points (Pre, Post, Post+3) and a significant difference between

the left and right vastus medialis (VM) at time point Pre. Different muscle heads also had different qMRI values, especially the RF, which showed significant differences compared to all the other muscle heads most of the time. These results highlight the need to consider individual muscle behavior separately and that pooling quadriceps muscle heads may result in a loss of information.

11.3 Longitudinal analysis

The longitudinal statistical analysis is inspired by the procedure of Froeling et al. (2015).

We used a statistical analysis of repeated measures with adaptation to the data normality, as the normal distribution could not be assumed. For each qMRI calculated index of each of the 9 muscle volumes (4 quadriceps muscle heads per leg and the total quadriceps volume), we tested the normality of the data at 3 time points with the *Shapiro-Wilk* test. For the global effect test, *one-way ANOVA* designed for repeated measures was conducted for all data normally distributed at all the time points; otherwise, the *Friedman* test was employed. While performing ANOVA, the sphericity of the data was verified by using *Mauchly*'s test. If the sphericity assumption was violated, the *Greenhouse-Geisser* correction method was used on the *P-value* of ANOVA. After the global effect test, a *post hoc* test was performed to compare each time point pair. The type of post hoc test depended on the normality of differences between two time points: *dependent t-test* for normally distributed differences and *Wilcoxon signed rank test* otherwise. The obtained P-values were adjusted with the *Bonferroni* adjustment method for multiple comparisons. A similar strategy was applied to the biological marker data with 58 variables at 4 time points (Appendix A.3.3).

When analyzing temporal changes on the repeated measures of qMRI metrics, we observed a significant time effect (i.e., race effect) on T2* and T2 mean values (Fig. 11.5 & 11.6).

Both T2* and T2 were significantly longer at arrival for most of the muscle heads. They significantly decreased after recovery for the VM and the vastus intermedius (VI) while not returning to baseline values at that time of measurement. PDFF and χ showed only a small time effect (details in Nguyen et al. (2021b)). Pooling all muscles, PDFF had a tendency to decrease after the race (2.88 ± 0.53) compared to baseline (3.16 ± 0.45) while increased to higher values than baseline (3.17 ± 0.46) after 48 h of recovery. When considering each muscle, the time effect reached significance for VI, right VM, and left VL muscle heads. Despite a similar trend as the PDFF, time-effect changes in χ did not reach significance for most muscle heads, except for left VL and left VM. When focusing on T2* and T2 findings, the vastus group exhibited more substantial variations than the RF. The VM and VI had more extensive changes than the VL or the average of all muscles. T2* values appeared less sensitive to muscle changes than T2, as illustrated in Figure 11.4, but most of the statistical tests were significant except for some muscle heads on both legs between time points Post and Post+3.

We propose below, in Figure 11.7, the histograms of T2* metric for the subject used as illustration in Figure 11.5, where the reader should have the impression that the lesion is very focal. The histograms are generated on the right Vastus Intermedius (VI) T2* images where we observed a large inflammation area (also in Fig.

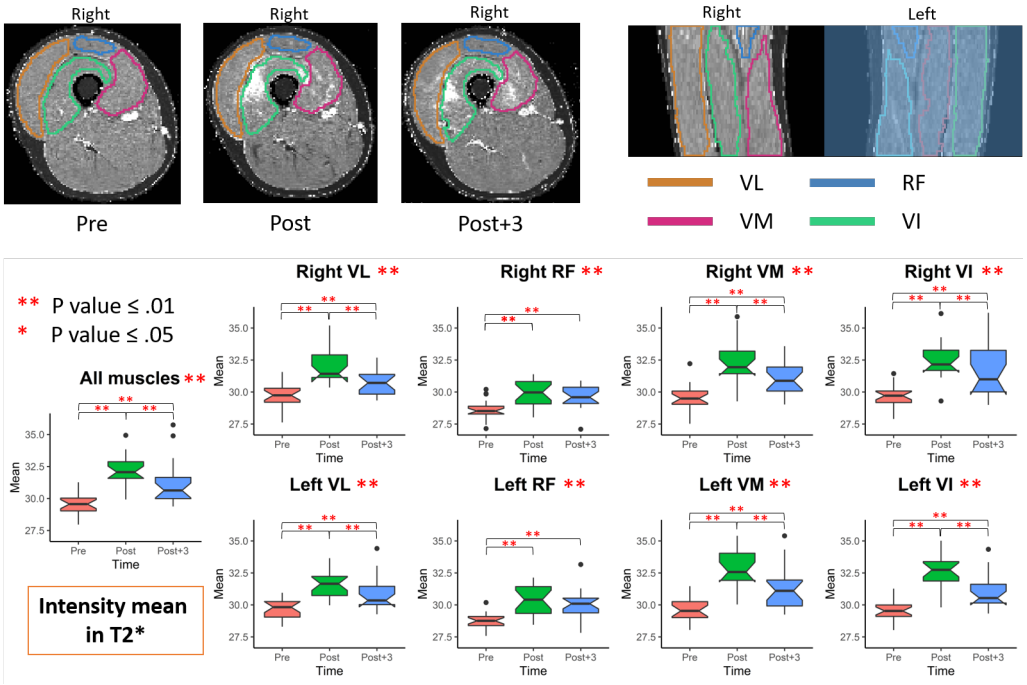

FIGURE 11.5: *Variation of T2* mean in the individual muscle heads of all finishers with an example of T2* maps at the three MR acquisition time points relative to the race of the same subject. A P-value less than .05 indicates a significant change between two time points. Abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius.*

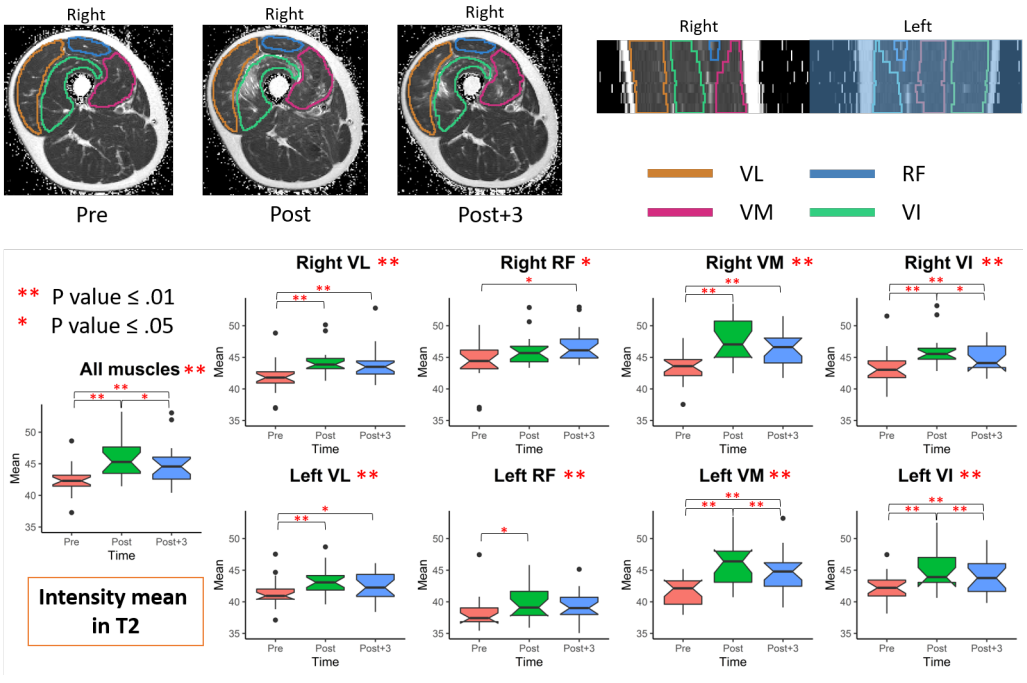

FIGURE 11.6: *Variation of T2 mean in the individual muscle heads of all finishers with an example of T2* maps at the three MR acquisition time points relative to the race of the same subject. A P-value less than .05 indicates a significant change between two time points. Abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius.*

11.5). We can see that while the mode did not change much (always around 30), the shape and the histogram range changed drastically: from a nearly symmetric and normal distribution in Pre, the histogram changed to a log-normal like distribution (right-tailed distribution) in Post and Post+3.

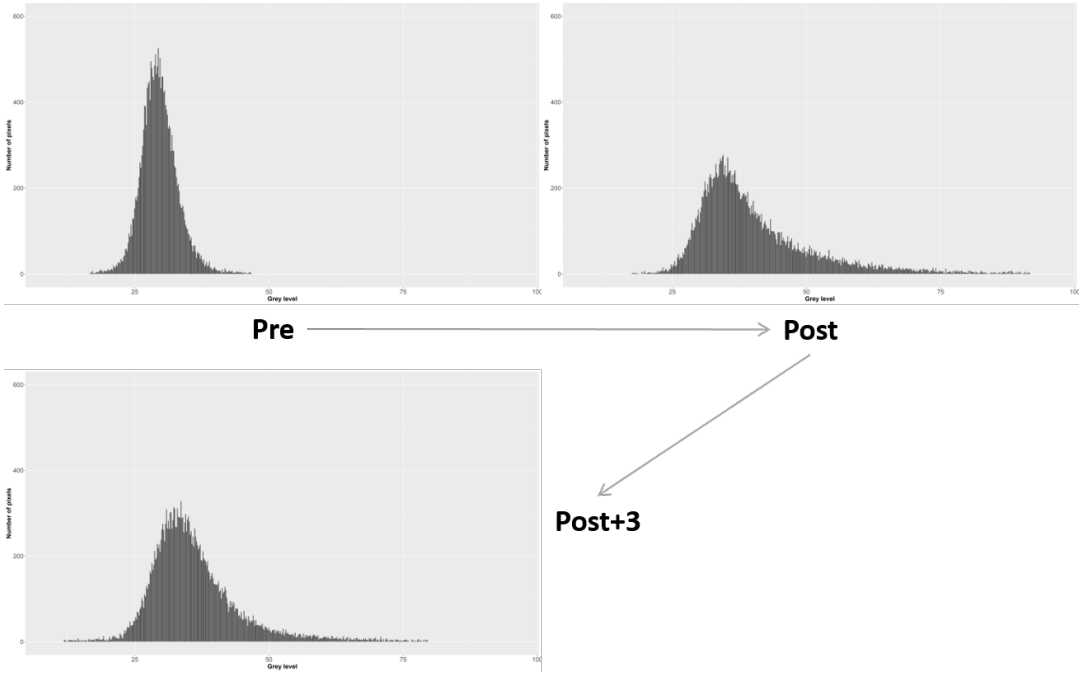

FIGURE 11.7: *Histogram of T2* metrics of a subject's vastus intermedius at 3 different time points of the race.*

We then separate the right VI into 2 regions depending on the T2* intensity at Post: a hyper-intensity region and the apparently *unchanging* region (Fig. 11.7). Table 11.1 gives the mean intensities of these two regions at the 3 time-points.

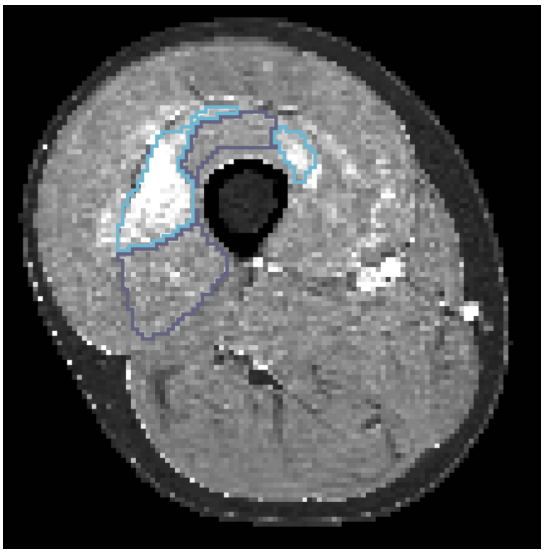

FIGURE 11.8: *T2* map of a right leg at the time point Post with the vastus intermedius separated into 2 regions depending on the intensity: a hyper-intensity region and the apparently unchanging region compared to the images at Pre.*

Sub-regions in VI muscle (based on T2* behavior at Post)	**Pre**	**Post**	**Post+3**
hyper-intensity region	29.86	50.80	41.68
unchanging region	29.69	35.49	30.22

TABLE 11.1: *Intensity mean of the 2 regions of vastus intermedius (illustrated in Fig. 11.8) at the three time points.*

We can see that, at Pre (baseline), the mean intensity is almost the same for both regions. At Post, both regions show an increase in mean intensity. While the hyper-intensity region demonstrates the largest change that indicates severe inflammation, the remaining of the muscle is also showing a smaller but not negligible change. As expected, both regions have a decrease in mean intensity at Post+3 compared to Post, which indicates the occurrence of the recovery process.

In addition to this individual case analysis, we have also studied the histogram evolution of each subject and analyzed globally four histogram features (mean, median, kurtosis, and skewness) in order to evaluate the homogeneity vs. the heterogeneity. We provide in Figure 11.9 the median, kurtosis and skewness boxplots computed from T2* maps. We can observe that all the histogram features increase for most of the muscle heads (except RF), meaning that the trend observed in the individual above (going from a symmetric distribution to a more right-tailed distribution) is observed in all finishers. If focal areas first draw attention in the T2* maps, there is also an overall inflammation of the tissue that statistical analysis undoubtedly helps to establish.

Intensity mean, median, histogram kurtosis, and skewness are first-order radiomic features. Similar analysis can be done on the higher-order features extracted from the qMRI maps. An example of the results obtained with T2* maps is presented in Figure 11.10. All the textural features (GLCM, GLRLM, GLSZM and NGTDM groups in the figure) are extracted with default parameters.

Finally, the potential associations between image markers and biological markers were evaluated by calculating the repeated measure correlation coefficient (Bakdash and Marusich, 2017) between each image marker and each biological marker. The correlation heatmap between the radiomic features extracted from T2* maps and the biological markers are shown in Figure 11.11.

Further discussion on the first-order image features and the biological markers for the MUST dataset, in particular their physiological aspects, can be found in Nguyen et al. (2021b). This longitudinal analysis procedure can also be applied to other anatomical structures, such as the femur bone marrow (Nguyen et al., 2019c).

11.4 Conclusion

In this chapter, we have presented the analytical application of automatic segmentation by resuming two of our contributions Nguyen et al. (2021a) and Nguyen et al. (2021b). The segmentation allows the study of local changes in subjects based on longitudinal quantitative MR data. The analysis framework is the same for manual, semi-automatic, or automatic segmentation. As mentioned above, considering the time-consuming and mentally exhausting nature of the manual segmentation task, a robust automatic segmentation method is indispensable. The more precise the

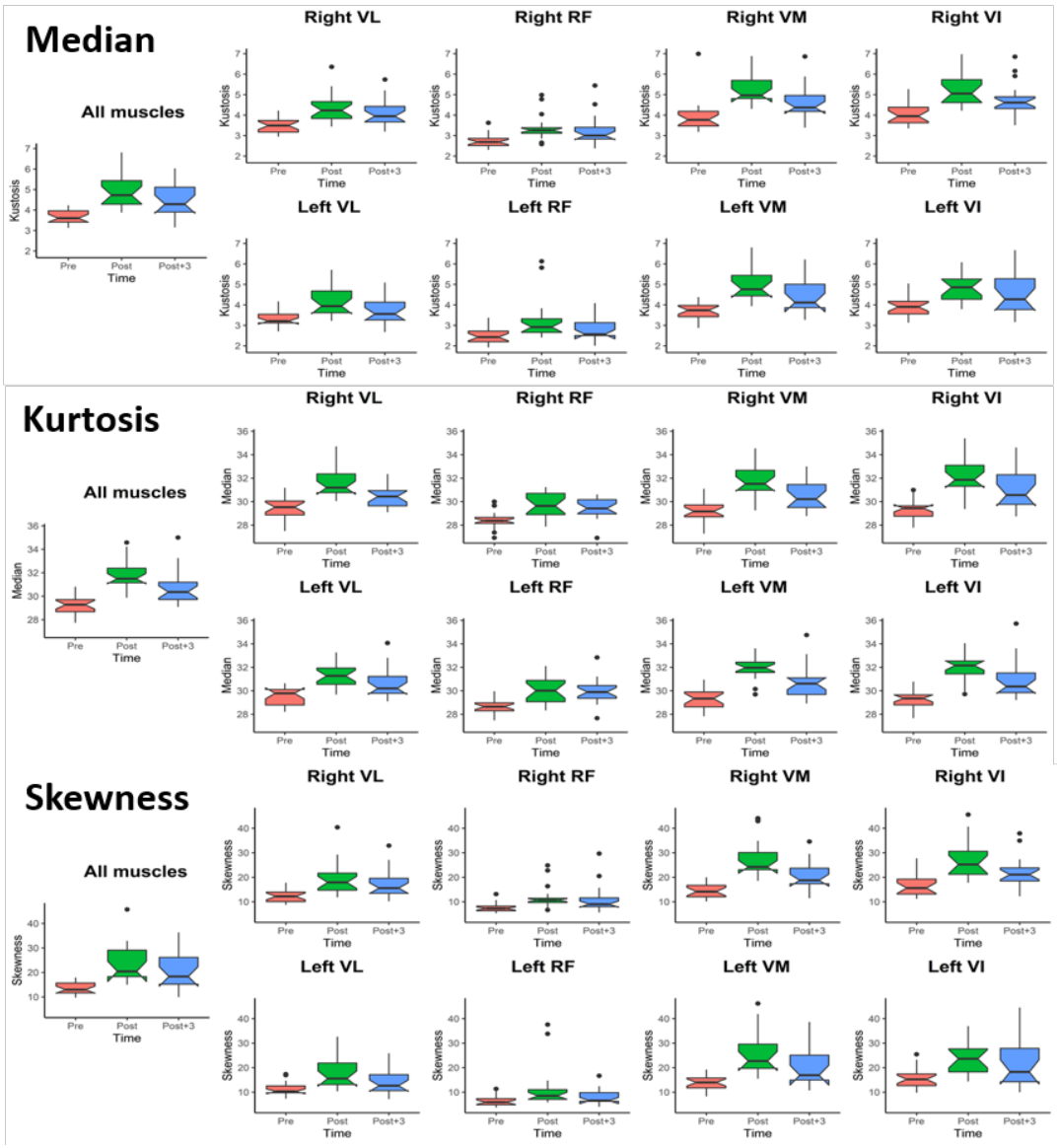

FIGURE 11.9: *Variation of T2* median and its histogram kurtosis and skewness in the individual muscle heads of all finishers. A P-value less than .05 indicates a significant change between two time points. Abbreviations: VL – Vastus Lateralis, RF – Rectus Femoris, VM – Vastus Medialis, VI – Vastus Intermedius.*

segmentation, the fewer postprocessing operations need to be applied for accurate analytical results.

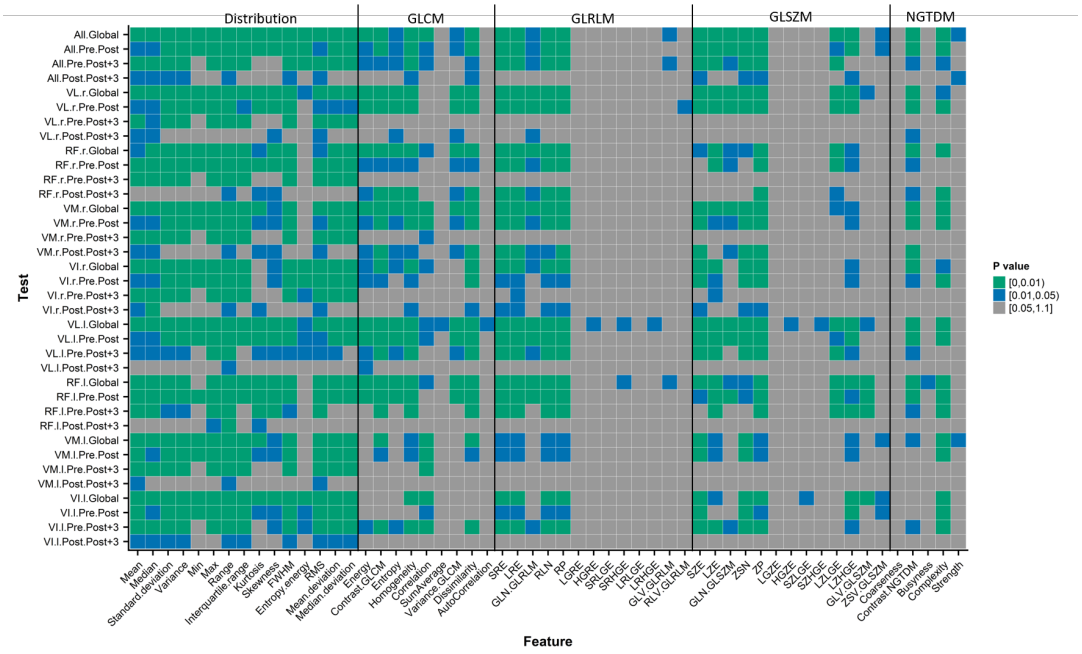

FIGURE 11.10: *Results of statistical tests in our longitudinal analysis of T2* on the entire set of radiomic features. A P-value inferior to .05 indicates a siginificant change between two time points. Abbreviations: VL - Vastus Lateralis, RF - Rectus Femoris, VM - Vastus Medialis, VI - Vastus Intermedius, r – right, l – left.*

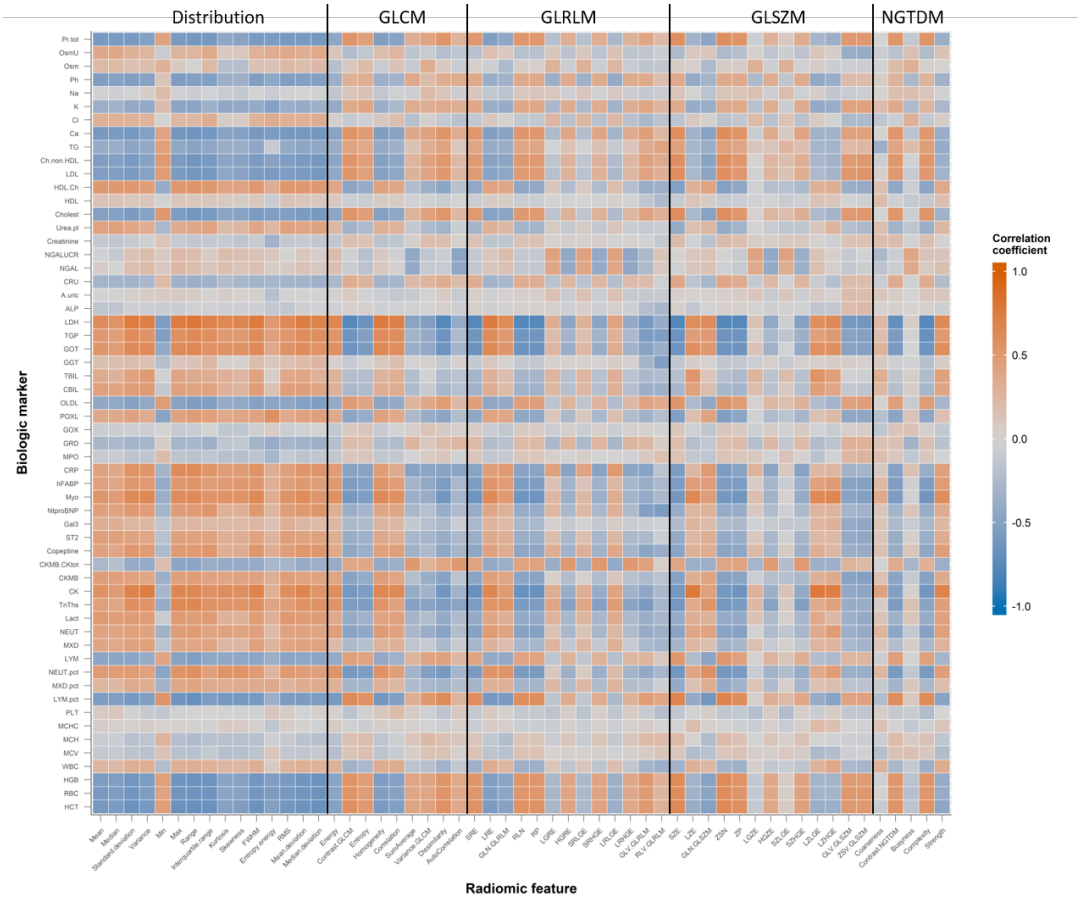

FIGURE 11.11: *Correlation between biologic markers and radiomic features extracted from the entire quadriceps volume in T2* maps.*

Conclusion

In this part, we have presented the applications of our segmentation methods on different muscle datasets and the longitudinal analysis of quadriceps muscles during an ultra-marathon based on automatic segmentation.

We applied our automatic segmentation methods optimized with morphological features on multiple MRI muscle datasets, which involve the studies of quadriceps, hamstrings, and rotator cuff. Overall, the methods remained robust for all the muscle groups and in longitudinal settings. However, our experiments of segmentation of a 3T MR image using atlases and pre-trained models from a 1.5T dataset showed that the textural difference due to different acquisition settings prevents pre-trained models' transfer to new datasets. A two-way translation between these two different types of signals needed to be studied to avoid repeating the manual segmentation process each time there is a new dataset.

Finally, we presented our longitudinal analysis of ultra-marathoners' quadriceps during and after an ultra-marathon based on the MUST dataset, which is the principal clinical application of our contributions. The detailed analysis can be found in Nguyen et al. (2021b), Nguyen et al. (2021a) and Nguyen et al. (2019c).

General conclusion

CHAPTER 12

Conclusion

This last chapter summarizes our key contributions (Sec. 12.1) and the conclusions (Sec. 12.2) drawn from our experiment results and our methodological and clinical applications. Section 12.3 offers our perspectives on the method improvement and a further investigation of the clinical study.

12.1 Key contributions

Our key contributions presented in this thesis are:

- We provided a complete analysis of Wang and Yushkevich's muti-atlas segmentation method with joint label fusion and corrective learning (JLF+CL) with a proposition of replacing the JLF step with 2D UNet coupling with data augmentation to tackle the method's limitations (Nguyen et al., 2018, 2019b).
- We introduced a morphological measurement and its applications to optimize presented segmentation methods: an atlas selection strategy for JLF to reduce computation time while conserving the segmentation quality, a data augmentation strategy to maximize the morphological variation in training and validation sets for UNet, and finally, a fine-tuning process added to UNet to improve the segmentation quality further.
- We proposed a robust longitudinal analysis procedure to study the effect of an ultra-marathon on the quadriceps, which involved pretreatments for the MRI data and their segmentations and longitudinal statistical analysis (Nguyen et al., 2019c, 2021a,b).

12.2 Conclusions

Based on the MUST dataset of upper leg MRI of ultra-marathon runners, the objective of this thesis is to propose a robust method that necessitates few annotated data for MRI quadriceps automatic segmentation, which will be the base for longitudinal muscle inflammation analysis.

Firstly, we have studied the multi-atlas segmentation approach (MAS) of Wang and Yushkevich with joint label fusion by patch-based weighting (JLF) and corrective learning by AdaBoost (CL) (Nguyen et al., 2018). The approach can provide high-quality segmentation even with a minimal number of manual segmentation as

atlases, making it the go-to method to quickly obtain an automatic segmentation to identify problematic difficulties. Since the approach is based on comparing the image patches in the target image with those in the atlases, naturally, the computation time is proportional to the image size and the number of atlases. However, the performance of this method depends mainly on the capability to register the atlases to the target image. Moreover, when the morphological difference between two subjects is too large, it is impossible to completely register their muscles without creating unrealistic deformation to the other anatomical parts and disturbing the image texture. A modification of the method at a lower resolution (Wang et al., 2017) can significantly reduce the computation time and remove aberrant anatomical errors but does not improve the validation metrics since it lacks precision around the muscle boundary.

An alternative to MAS that can reduce both the computation time and the sensibility to registration is deep learning. We proposed to replace the JLF step of Wang and Yushkevich's framework with 2D UNet (Nguyen et al., 2019b). To enhance the morphological diversity in the training and validation sets of the network, a data augmentation step was effectuated using random B-spline warping and deformation registration to non-annotated data. The network is called *weakly supervised U-Net* since it was trained and validated using both manually annotated data and automatically generated data. The main advantage of this approach is that it reduces the inference time from 24-48 hours to 45 seconds while conserving and, in some cases, improving the segmentation precision.

In both cases of JLF and UNet, the CL step was executed in the same manner. Most of the time, CL performs well in removing aberrant distanced errors and refining the precision at the muscle boundary. However, the correction is done voxel by voxel, so the final result is often noisy around the border. Furthermore, this method assumes that the segmentation to refine is very close to the target accuracy; therefore, it does not work with large errors.

For a large anatomical group such as the quadriceps, the morphological difference among subjects are the main challenge in the automatic segmentation process. Our last contribution for the automatic segmentation methods in this thesis is introducing morphological measures that help optimize the atlas selection for MAS and the data augmentation for UNet. With a strategic atlas selection, the number of atlases needed to achieve the same performance was reduced by half, along with the computation time. The 2D UNet trained with morphological-measure-based augmented data gives similar segmentation validation metrics as MAS. In the cases where the UNet fails to produce adequate results, an extra step of fine-tuning, also based on morphological similarity, is recommended to get remarkable improvement in segmentation quality. The optimized MAS and UNet remain robust when being applied on different muscle datasets and in longitudinal settings. However, when using atlases and trained networks of a 1.5T MRI dataset to segment a 3T image, the textural difference between images from different acquisition settings prevent a smooth transfer.

Finally, we presented the principal clinical applications of our above contributions, which is a longitudinal analysis of Tors de Géants 2014 ultra-marathoners' quadriceps. Based on the automatic segmentation of quadriceps, we extracted and analyzed the radiomic features of each muscle head from quantitative MR images (T2, T2*, PDFF and χ). The results were published in Nguyen et al. (2021a) and Nguyen et al. (2021b).

12.3 Perspectives

Based on the identified limitations of the methods proposed in this thesis, this section offers our perspectives on the possible improvements upon the segmentation methods for better generalization and the possibility of further investigation of the clinical study.

12.3.1 Segmentation methods

Several aspects of the segmentation algorithms can be further investigated and experimented with in the follow-up studies:

- *Deep learning networks*: Deep learning is currently at the center of the image processing research area, with a significant number of new networks or improvements of existing networks introduced each year. Therefore, a regular update of the literature and testing of new networks can be done to improve the segmentation quality. This was also an objective of the internship of Malick Kandji, which revolved around a new MRI dataset of shoulder muscles presented in Section 10.1.
- *Corrective learning*: Corrective learning of Wang and Yushkevich is based on the image features computed at each pixel, including the physical position, the gray level of the pixel and its neighbors, the texture of the image patch centered at the pixel. Technically, we can incorporate any type of image features; one of the most common is SIFT (Scale Invariant Feature Transform, Lowe (2004)). A more novel approach is getting the image encoded into patch-based image features by CNN-based networks as in Simo-Serra et al. (2015) and Ono et al. (2018). In both papers, SIFT-like 128 dimensional features of a *keypoint* are generated from the image patch centered at the keypoint using 3 convolutional layers. Another possibility is to replace the current corrective learning algorithm entirely with another segmentation refinement method such as ErrorNet (Tajbakhsh et al., 2020), which used UNet to learn the systematic segmentation errors injected by a variational autoencoder and generated, for the to-be-corrected automatic segmentation, the pixel-wised cards of the error probability. Finally, an end-to-end solution is to create an architecture based on Generative Adversarial Networks (Goodfellow et al., 2014), for which UNet-GAN (Dong et al., 2019) could be a reference.
- *Morphological features*: More descriptors of the size and shape of each segmentation label can be included for a more complex morphological representation of the images while keeping it gray-level-independent. Numerous morphological features were presented in Zwanenburg et al. (2020), which are derived from the approximated shape defined by the circumference mesh generated using an adapted version of marching cubes algorithm of Lorensen and Cline. However, with our limited number of subjects, introducing many features could make the distribution of the subjects in the feature spaces sparse and noisy, which might necessitate a weighted distance for morphological similarity measure instead of euclidean distance as used in our experiments. Furthermore, the morphological features can be used for failure prediction of our segmentation methods as the morphological-distanced subjects were observed with lower segmentation accuracy in our experiments and for the active building of our dataset (see Sec. 12.3.2 below).

- *Transfer to different acquisition settings*: To transfer our pre-trained models and atlases from a 1.5T dataset to a 3T dataset, a translation between these two types of signals needed to be studied to avoid the repetition of the manual segmentation process. We can start with some existing image synthesis methods such as in Jog et al. (2013) and Qu et al. (2020), which synthesized an MRI sequence of a specific modality from a sequence of a different modality: Jog et al. synthesized T2-weighted images from T1-weighted ones and 3T T1-weighted images from 1.5T MPRAGEs while Qu et al. synthesized 7T T1-weighted images from their 3T counterparts. The application of these methods might not be direct as in both cases, data of lower field were translated into higher field, which is the reverse of our application - using a model trained with lower-field images to segment a higher-field image. Moreover, another possibility is to transform all the data to a common space, which open a new direction in domain adaptation.

12.3.2 Clinical application

There are two aspects in the clinical application that can profit from further research: the extension of the annotated data set and the longitudinal analysis.

- *Active building of the dataset*: As the importance of morphological variation in the annotated dataset has been proven, adding an atlas with a significant morphological difference from the existing dataset will be extremely valuable. When a new image is presented and in the case where our proposed segmentation methods cannot provide a satisfying segmentation, it is recommended to study the morphological measure of the new image and its position compared to the existing atlases. If the lack of similar morphology in the training data is the reason for segmentation failure, manual segmentation of this new image should be done and added to the dataset for future application.
- *Radiomic feature analysis*: Section 11.3 proved that the muscle inflammation is not only focal but appears in the entire muscle head. However, it is undeniable that some regions in a muscle head are more affected than others. At the same time, the radiomic features are averaged over all pixels when computed on a large volume, which could be the cause of the non-significant statistical test results on the longitudinal effect of the race, especially for the textural features. A further study on the more affected regions might be a direction worth exploring.

12.4 Personal bibliography

[Nguyen et al. 2018] NGUYEN, Hoai-Thu; CROISILLE, Pierre; VIALLON, Magalie; DE BOURGUIGNON, Charles; GRANGE, Rémi; GRANGE, Sylvain; GRENIER, Thomas: Robust multi-atlas MRI segmentation with corrective learning for quantification of local quadriceps muscles inflammation changes during a longitudinal study in athletes. In: *ISMRM: International Society for Magnetic Resonance in Medicine*. Paris, 2018

[Nguyen et al. 2019b] NGUYEN, Hoai-Thu; CROISILLE, Pierre; VIALLON, Magalie; LECLERC, Sarah; GRANGE, Sylvain; GRANGE, Rémi; BERNARD, Olivier; GRENIER, Thomas: Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised U-Net. In: *MIDL: International Conference on Medical Imaging with Deep Learning*. London, 2019

[Nguyen et al. 2019c] NGUYEN, Hoai-Thu; GRENIER, Thomas; LEPORQ, Benjamin; BEY, Loïc; VIALLON, Magalie; CROISILLE, Pierre: Evaluation of local changes in femoral bone marrow during a mountain ultra-marathon with quantitative MRI. In: *ISMRM: International Society for Magnetic Resonance in Medicine*. Montréal, 2019

[Nguyen et al. 2021a] NGUYEN, Hoai-thu; GRANGE, Sylvain; LEPORQ, Benjamin; VIALLON, Magalie; CROISILLE, Pierre; GRENIER, Thomas: Impact of Distortion on Local Radiomic Analysis of Quadriceps Based on Quantitative Magnetic Resonance Imaging Data. In: *International Journal of Pharma Medicine and Biological Sciences* 10 (2021), Vol. 2

[Nguyen et al. 2021b] NGUYEN, Hoai-Thu; GRENIER, Thomas; LEPORQ, Benjamin; LE GOFF, Caroline; GILLES, Benjamin; GRANGE, Sylvain; GRANGE, Rémi; MILLET, Grégoire P.; BEUF, Olivier; CROISILLE, Pierre; VIALLON, Magalie: Quantitative Magnetic Resonance Imaging Assessment of the Quadriceps Changes during an Extreme Mountain Ultramarathon. In: *Medicine & Science in Sports & Exercise* 53 (2021), Vol. 4, pp. 869–881. – DOI 10.1249/mss.0000000000002535. – ISSN 0195–9131

CHAPTER 13

Conclusion en Français (Conclusion in French)

Ce dernier chapitre résume nos principales contributions (Sec. 13.1) et les conclusions (Sec. 13.2) déduites de nos résultats expérimentaux et de nos applications méthodologiques et cliniques. La section 13.3 présente nos perspectives sur l'amélioration des méthodes, et des approfondissements de l'étude clinique.

13.1 Contributions clés

Nos principales contributions présentées dans cette thèse sont :

- Nous avons fourni une analyse complète de la méthode de segmentation mutiatlas de Wang and Yushkevich avec la fusion conjointe d'étiquettes et l'apprentissage correctif (JLF+CL) avec une proposition de remplacement de l'étape JLF par un couplage UNet 2D incluant une augmentation des données pour remédier aux besoins d'entraînement de la méthode (Nguyen et al., 2018, 2019b).
- Nous avons proposé une mesure morphologique et ses applications pour optimiser les méthodes de segmentation présentées : une stratégie de sélection d'atlas pour JLF afin de réduire le temps de calcul tout en conservant la qualité de la segmentation, une stratégie d'augmentation des données pour maximiser la variabilité morphologique dans les bases de données d'entraînement et de validation pour UNet, et finalement, un processus de réglage fin ajouté à UNet pour améliorer encore la qualité de la segmentation.
- Nous avons proposé une procédure d'analyse longitudinale robuste pour étudier l'effet d'un ultra-marathon sur le quadriceps, qui implique des prétraitements pour les données IRM et leurs segmentations ainsi qu'une analyse statistique longitudinale (Nguyen et al., 2019c, 2021a,b).

13.2 Conclusions

En s'appuyant sur le jeu de données MUST d'IRM de la partie supérieure de la jambe de coureurs d'ultra-marathon, l'objectif de cette thèse est de proposer une méthode robuste de segmentation automatique du quadriceps par IRM qui nécessite peu de données annotées, qui sera la base de l'analyse longitudinale de l'inflammation musculaire.

Nous avons étudié l'approche de segmentation multi-atlas (MAS) de Wang and Yushkevich avec une fusion conjointe des étiquettes par pondération basée sur les patchs d'image (JLF) et un apprentissage correctif par AdaBoost (CL). L'approche peut fournir une segmentation de haute qualité, même avec un nombre minimal de segmentations manuelles, ce qui en fait la méthode de référence pour obtenir rapidement une segmentation automatique afin d'identifier les difficultés de segmentation. Puisque l'approche est basée sur la comparaison des patchs de l'image cible avec ceux des atlas, le temps de calcul est naturellement proportionnel à la taille de l'image et au nombre d'atlas. La performance de cette méthode dépend principalement de la capacité à recaler les atlas sur l'image cible. Lorsque la différence morphologique entre deux sujets est trop importante, il est impossible de recaler complètement leurs muscles sans créer une déformation irréaliste de l'autre partie des structures anatomiques et sans perturber la texture de l'image. Une modification de la méthode JLF+CL utilisant une résolution inférieure (Wang et al., 2017) peut réduire considérablement le temps de calcul et supprimer les erreurs anatomiques aberrantes mais n'améliore pas les métriques de validation car elle manque de précision autour du contour du muscle.

L'apprentissage profond est une alternative au MAS qui peut réduire à la fois le temps de calcul et la sensibilité au recalage. Nous avons proposé de remplacer l'étape JLF de la procédure de Wang and Yushkevich par un UNet 2D. Pour améliorer la diversité morphologique dans les jeux de données d'apprentissage et de validation du réseau, une étape d'augmentation de données a été effectuée en utilisant une déformation B-spline aléatoire et un recalage déformable sur des données non annotées. Cette approche est appelée *weakly supervised UNet* car le réseau a été entrainé et validé en utilisant à la fois des données annotées manuellement et des annotations générées automatiquement. Le principal avantage de cette approche est qu'elle réduit le temps d'inférence de 24-48 heures à 45 secondes tout en conservant et, dans certains cas, en améliorant la précision de la segmentation.

Dans les deux cas de JLF et UNet, l'étape CL a été exécutée de la même manière. La plupart du temps, CL donne de bons résultats en supprimant les erreurs de distance aberrantes et en affinant la précision au contour du muscle. Cependant, la correction est effectuée voxel par voxel, de sorte que le résultat final est souvent bruité autour de la frontière. De plus, cette méthode assume que la segmentation à affiner est très proche de la solution, elle ne fonctionne donc pas avec des erreurs importantes.

Pour un groupe anatomique tel que le quadriceps, les différences morphologiques entre les sujets constituent le principal défi du processus de segmentation automatique. Notre dernière contribution aux méthodes de segmentation automatique dans cette thèse est l'introduction de descripteurs morphologiques qui aident à optimiser la sélection d'atlas pour MAS et l'augmentation de données pour UNet. Avec une stratégie de sélection des atlas, le nombre d'atlas nécessaires pour atteindre la même performance a été réduit de moitié, ainsi que le temps de calcul. UNet 2D entraîné avec des données augmentées basées sur la mesure morphologique donne des métriques de validation de segmentation similaires à celles du MAS. Dans les cas où UNet ne produit pas de résultats adéquats, une étape supplémentaire de *fine tuning*, également basée sur la similarité morphologique, est recommandée pour obtenir une amélioration remarquable de la qualité de la segmentation. Le MAS et

UNet optimisés restent robustes lorsqu'ils sont appliqués à différents jeux de données musculaires et dans des configurations longitudinales. Cependant, lors de l'utilisation d'atlas et de réseaux entraînés sur un jeu de données IRM 1,5T pour segmenter des images 3T, la différence de texture entre les images provenant de différents paramètres d'acquisition, empêche un transfert efficace.

Finalement, nous avons présenté les principales applications cliniques de nos contributions ci-dessus, à savoir une analyse longitudinale des quadriceps des ultra-marathoniens du Tors de Géants 2014. Sur la base de la segmentation automatique du quadriceps, nous avons extrait et analysé les relations des caractéristiques radiomiques de chaque chef musculaire à partir d'images RM quantitatives (T2, T2*, PDFF et χ). Les résultats ont été publiés dans Nguyen et al. (2021a) et Nguyen et al. (2021b).

13.3 Perspectives

Sur la base des limites identifiées des méthodes proposées dans cette thèse, cette section présente nos perspectives d'améliorations possibles des méthodes de segmentation pour obtenir une meilleure généralisation et la possibilité d'approfondir les études cliniques.

13.3.1 Méthodes de segmentation

Plusieurs aspects des algorithmes de segmentation peuvent être étudiés et expérimentés dans les études de suivi :

- *Réseaux d'apprentissage profond* : L'apprentissage profond est actuellement au centre de la recherche en traitement d'images, avec un nombre important de nouveaux réseaux ou d'améliorations de réseaux existants introduits chaque année. Par conséquent, une mise à jour régulière sur la littérature et le test de nouveaux réseaux sont nécessaires afin de voir si une amélioration peut être observée. C'était également l'un des objectifs du stage de Malick Kandji, qui portait sur un nouveau jeu de données IRM des muscles de l'épaule présenté dans la section 10.1.
- *Apprentissage correctif* : L'apprentissage correctif de Wang and Yushkevich est basé sur les descripteurs de l'image calculées à chaque pixel, y compris la position physique, le niveau de gris du pixel et de ses voisins, la texture du patch d'image centré sur le pixel. Techniquement, nous pouvons incorporer n'importe quel type de descripteurs d'images, l'une des plus courantes étant SIFT (Scale Invariant Feature Transform, Lowe (2004)). Une approche plus récente consiste à coder l'image en descripteurs d'images basés sur les patchs à l'aide de réseaux CNN, comme dans Simo-Serra et al. (2015) et Ono et al. (2018). Dans ces deux articles, les caractéristiques de 128 dimensions sont similaires aux SIFT d'un *point clé* et sont générées à partir du patch d'image centré sur le point clé à l'aide de 3 couches convolutionelles. Une autre possibilité consiste à remplacer entièrement l'algorithme d'apprentissage correctif actuel par une autre méthode de raffinement de la segmentation, comme ErrorNet (Tajbakhsh et al., 2020), qui a utilisé UNet pour apprendre les erreurs systématiques de segmentation injectées par un auto-codeur variationnel (VAE) et a généré, pour la segmentation automatique à corriger, les cartes pixellisées de la probabilité

d'erreur. Finalement, une solution de type *end-to-end* consiste à créer une architecture basée sur les Generative Adversarial Networks (Goodfellow et al., 2014), pour laquelle UNet-GAN (Dong et al., 2019) pourrait être une référence.

- *Descripteurs morphologiques* : Il est possible d'inclure davantage de descripteurs de la taille et de la forme de chaque étiquette de segmentation pour obtenir une représentation morphologique plus complexe des images tout en la maintenant indépendante du niveau de gris. De nombreux descripteurs morphologiques ont été présentés dans Zwanenburg et al. (2020), qui sont dérivés de la forme approximative définie par le maillage de la circonférence généré à l'aide d'une version adaptée de l'algorithme *marching cube* de Lorensen and Cline. Cependant, avec notre nombre limité de sujets, l'introduction de nombreux descripteurs pourrait rendre la distribution des sujets dans les espaces de descripteurs clairsemée et bruitée, ce qui pourrait nécessiter l'introduction d'une distance pondérée pour la mesure de similarité morphologique au lieu de la distance euclidienne utilisée dans nos expériences. En outre, les descripteurs morphologiques peuvent être utilisés pour la prédiction de l'échec de nos méthodes de segmentation, car pour les sujets morphologiquement éloignés des autres, nous avons observés une précision de segmentation inférieure dans nos expériences. Les descripteurs pourraient aussi permettre la construction active de notre jeu de données (voir Sec. 13.3.2 ci-dessous).

- *Transfert de jeux de données avec différents paramètres d'acquisition* : Pour transférer nos modèles pré-entraînées sur un jeu de données 1,5T à un jeu de données 3T, une traduction entre ces deux types de signaux devrait être étudiée pour éviter la répétition du processus de segmentation manuelle. Nous pouvons commencer par certaines méthodes de synthèse d'images existantes, comme dans Jog et al. (2013) et Qu et al. (2020), qui ont synthétisé une séquence IRM spécifique à partir d'une séquence différente. Jog et al. a synthétisé des images pondérées en T2 à partir de celles pondérées en T1 et des images pondérées en T1 à 3T à partir de MPRAGEs à 1,5T, tandis que Qu et al. a synthétisé des images pondérées en T1 à 7T à partir de leurs homologues à 3T. L'application de ces méthodes n'est peut-être pas directe car, dans les deux cas, les données de champ inférieur ont été traduites en champ supérieur, ce qui est l'inverse de notre application : utiliser un modèle entraîné avec des images de champ inférieur pour segmenter une image de champ supérieur. En outre, une autre possibilité consiste à transformer toutes les données dans un espace commun comme ce que peut être fait en *domain adaptation*.

13.3.2 Application clinique

Deux aspects de l'application clinique peuvent bénéficier de recherches supplémentaires : l'extension du jeu de données annotées et l'analyse longitudinale.

- *Construction active du jeu de données* : L'importance de la variation morphologique dans le jeu de données annotées ayant été prouvée, l'ajout d'un atlas présentant une différence morphologique significative par rapport à l'ensemble de données existant sera extrêmement précieux. Ainsi, lorsqu'une nouvelle image est présentée et dans le cas où nos méthodes de segmentation ne peuvent fournir une segmentation satisfaisante, il est recommandé d'étudier la mesure morphologique de la nouvelle image et sa position par rapport aux atlas existants. Si l'absence de morphologie similaire dans les données d'entraînement est la raison de l'échec de la segmentation, une segmentation manuelle

de cette nouvelle image devrait être effectuée et ajoutée au jeu de données pour les prochaines applications.

- *Analyse des descripteurs radiomiques* : La section 11.3 a prouvé que l'inflammation musculaire n'est pas seulement focale mais apparaît dans tout le chef musculaire. Cependant, il est indéniable que certaines régions du chef musculaire sont plus touchées que d'autres. En même temps, les descripteurs radiomiques sont moyennés sur tous les pixels. Lorsqu'ils sont calculés sur un grand volume, ce moyennage pourrait être la cause des résultats non significatifs des tests statistiques sur l'effet longitudinal de la course, en particulier pour les descripteurs texturaux. Une étude plus approfondie sur les régions les plus touchées pourrait être une direction à explorer.

13.4 Bibliographie personnelle

[Nguyen et al. 2018] NGUYEN, Hoai-Thu ; CROISILLE, Pierre ; VIALLON, Magalie ; DE BOURGUIGNON, Charles ; GRANGE, Rémi ; GRANGE, Sylvain ; GRENIER, Thomas : Robust multi-atlas MRI segmentation with corrective learning for quantification of local quadriceps muscles inflammation changes during a longitudinal study in athletes. In : *ISMRM : International Society for Magnetic Resonance in Medicine*. Paris, 2018

[Nguyen et al. 2019b] NGUYEN, Hoai-Thu ; CROISILLE, Pierre ; VIALLON, Magalie ; LECLERC, Sarah ; GRANGE, Sylvain ; GRANGE, Rémi ; BERNARD, Olivier ; GRENIER, Thomas : Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised U-Net. In : *MIDL : International Conference on Medical Imaging with Deep Learning*. London, 2019

[Nguyen et al. 2019c] NGUYEN, Hoai-Thu ; GRENIER, Thomas ; LEPORQ, Benjamin ; BEY, Loïc ; VIALLON, Magalie ; CROISILLE, Pierre : Evaluation of local changes in femoral bone marrow during a mountain ultra-marathon with quantitative MRI. In : *ISMRM : International Society for Magnetic Resonance in Medicine*. Montréal, 2019

[Nguyen et al. 2021a] NGUYEN, Hoai-thu ; GRANGE, Sylvain ; LEPORQ, Benjamin ; VIALLON, Magalie ; CROISILLE, Pierre ; GRENIER, Thomas : Impact of Distortion on Local Radiomic Analysis of Quadriceps Based on Quantitative Magnetic Resonance Imaging Data. In : *International Journal of Pharma Medicine and Biological Sciences* 10 (2021), Vol. 2

[Nguyen et al. 2021b] NGUYEN, Hoai-Thu ; GRENIER, Thomas ; LEPORQ, Benjamin ; LE GOFF, Caroline ; GILLES, Benjamin ; GRANGE, Sylvain ; GRANGE, Rémi ; MILLET, Grégoire P. ; BEUF, Olivier ; CROISILLE, Pierre ; VIALLON, Magalie : Quantitative Magnetic Resonance Imaging Assessment of the Quadriceps Changes during an Extreme Mountain Ultramarathon. In : *Medicine & Science in Sports & Exercise* 53 (2021), Vol. 4, pp. 869–881. – DOI 10.1249/mss.0000000000002535. – ISSN 0195–9131

Bibliography

[Ababneh et al. 2008] ABABNEH, Zaid Q.; ABABNEH, Riad; MAIER, Stephan E.; WINALSKI, Carl S.; OSHIO, Koichi; ABABNEH, Anas M. ; MULKERN, Robert V.: On the correlation between T2 and tissue diffusion coefficients in exercised muscle: Quantitative measurements at 3T within the tibialis anterior. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 21 (2008), pp. 273–278. – DOI 10.1007/s10334–008–0120–8. – ISSN 09685243

[Ahmad et al. 2014] AHMAD, Ezak; YAP, Moi H.; DEGENS, Hans ; MCPHEE, Jamie S.: *Atlas-registration based image segmentation of MRI human thigh muscles in 3D space*. Version: 2014

[Alkadi et al. 2019] ALKADI, Ruba; EL-BAZ, Ayman; TAHER, Fatma ; WERGHI, Naoufel: A 2.5D Deep Learning-Based Approach for Prostate Cancer Detection on T2-Weighted Magnetic Resonance Imaging. In: LEAL-TAIXÉ, Laura (Hrsg.); ROTH, Stefan (Hrsg.): *Computer Vision – ECCV 2018 Workshops*. Cham : Springer International Publishing, 2019. – ISBN 978–3–030–11018–5, pp. 734–739

[Andrews and Hamarneh 2015] ANDREWS, S.; HAMARNEH, G.: The Generalized Log-Ratio Transformation: Learning Shape and Adjacency Priors for Simultaneous Thigh Muscle Segmentation. In: *IEEE Transactions on Medical Imaging* 34 (2015), Vol. 9, pp. 1773–1787

[Artaechevarria et al. 2008] ARTAECHEVARRIA, Xabier; MUÑOZ-BARRUTIA, Arrate ; SOLÓRZANO, Carlos Ortiz-de: Efficient classifier generation and weighted voting for atlas-based segmentation : Two small steps faster and closer to the Combination Oracle. In: *Proc. SPIE Int. Soc. Opt. Photonics* (2008). – DOI 10.1117/12.769401. – ISBN 978–0–8194–7098–0

[Artaechevarria et al. 2009] ARTAECHEVARRIA, Xabier; MUÑOZ-BARRUTIA, Arrate ; SOLÓRZANO, Carlos Ortiz-de: Combination strategies in multi-atlas image segmentation: Application to brain MR data. In: *IEEE Trans. Med. Imaging* (2009). – DOI 10.1109/TMI.2009.2014372. – ISBN 1558–254X (Electronic)\n0278–0062 (Linking)

[Azzabou et al. 2015] AZZABOU, Noura; DE SOUSA, Paulo L.; CALDAS, Ericky ; CARLIER, Pierre G.: Validation of a generic approach to muscle water T2 determination at 3T in fat-infiltrated skeletal muscle. In: *Journal of Magnetic Resonance Imaging* (2015). – DOI 10.1002/jmri.24613. – ISSN 15222586

[BachCuadra et al. 2015] BACHCUADRA, M.; DUAY, V. ; THIRAN, J. P.: Atlas-based segmentation. Version: 2015. In: *Handbook of Biomedical Imaging: Methodologies and Clinical Research*. 2015. – DOI 10.1007/978–0–387–09749–7_12. – ISBN 9780387097497, pp. 221–224

[Bakdash and Marusich 2017] BAKDASH, Jonathan Z.; MARUSICH, Laura R.: Repeated measures correlation. In: *Frontiers in Psychology* 42 (2017), Vol. 2, pp. 261–7. – DOI 10.3389/fpsyg.2017.00456. – ISBN 1664–1078

[Balafar et al. 2008] BALAFAR, M a.; RAMLI, Abd. R.; SARIPAN, M.Iqbal; MAHMUD, Rozi ; MASHOHOR, Syamsiah: Medical image segmentation using Fuzzy C-Mean (FCM) and dominant grey levels of image. In: *Vis. Inf. Eng. 2008. VIE 2008. 5th Int. Conf.* (2008). – DOI 10.1049/cp:20080329. ISBN 0537–9989 VO –

[Cai et al. 2020] CAI, L.; GAO, Jing-Yang ; ZHAO, Di: A review of the application of deep learning in medical image classification and segmentation. In: *Annals of translational medicine* 8 11 (2020), pp. 713

[Canbolat et al. 2018] CANBOLAT, Mustafa; SENOL, Deniz; CEVIRGEN, Furkan ; OZBAG, Davut: An anatomic overview to "manspreading" campaign. In: *Annals of Medical Research* 25 (2018), Vol. 3, pp. 499–502. – DOI 10.5455/annalsmedres.2018.06.106. – ISSN 2636–7688

[Chollet 2015] CHOLLET, François: Keras: The Python Deep Learning library. In: *Keras.Io* (2015)

[Çiçek et al. 2016] ÇIÇEK, Özgün; ABDULKADIR, Ahmed; LIENKAMP, Soeren S.; BROX, Thomas ; RONNEBERGER, Olaf: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: OURSELIN, Sebastien (Hrsg.); JOSKOWICZ, Leo (Hrsg.); SABUNCU, Mert R. (Hrsg.); UNAL, Gozde (Hrsg.) ; WELLS, William (Hrsg.): *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Cham : Springer International Publishing, 2016. – ISBN 978–3–319–46723–8, pp. 424–432

[Clarke et al. 1995] CLARKE, L P.; VELTHUIZEN, R P.; CAMACHO, M A.; HEINE, J J.; VAIDYANATHAN, M; HALL, L O.; THATCHER, R W. ; SILBIGER, M L.: MRI segmentation: Methods and applications. In: *Magnetic Resonance Imaging* 13 (1995), Vol. 3, 343–368. – DOI 10.1016/0730–725X(94)00124–L. – ISSN 0730–725X

[Conn 2009] CONN, P M.: *Essential Bioimaging Methods*. Elsevier Science, 2009 (ISSN). https://books.google.fr/books?id=luXzcplngjcC. – ISBN 9780080963426

[Degache et al. 2014] DEGACHE, Francis; VAN ZAEN, Jérôme; OEHEN, Lukas; GUEX, Kenny; TRABUCCHI, Pietro ; MILLET, G??goire: Alterations in postural control during the world's most challenging mountain ultra-marathon. In: *PLoS ONE* 9 (2014), Vol. 1

[Delavier 2003] DELAVIER, F.: *Women's Strength Training Anatomy*. Human Kinetics, 2003 (Anatomy series). https://books.google.fr/books?id=mKX9tAxwpG4C. – ISBN 9780736048132

[Deng et al. 2010] DENG, Wankai; XIAO, Wei; DENG, He ; LIU, Jianguo: MRI brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve. In: *Proc. - 2010 3rd Int. Conf. Biomed. Eng. Informatics, BMEI 2010*, 2010. – ISBN 9781424464968

[Depa et al. 2010] DEPA, Michal; SABUNCU, Mert R.; HOLMVANG, Godtfred; NEZAFAT, Reza; SCHMIDT, Ehud J. ; GOLLAND, Polina: Robust atlas-based segmentation of highly variable anatomy: Left atrium segmentation. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2010. – ISBN 364215834X

[Dice 1945] DICE, Lee R.: Measures of the Amount of Ecologic Association Between Species. In: *Ecology* 26 (1945), July, Vol. 3, 297–302. http://www.jstor.org/pss/1932409

[Dong et al. 2019] DONG, Xue; LEI, Yang; WANG, Tonghe; THOMAS, Matthew; TANG, L.; CURRAN, W.; LIU, T. ; YANG, Xiaofeng: Automatic multiorgan segmentation in thorax CT images using U-net-GAN. In: *Medical physics* 46 5 (2019), pp. 2157–2168

[Edouard et al. 2016] EDOUARD, Pascal; BRANCO, Pedro ; ALONSO, Juan-Manuel: Muscle injury is the principal injury type and hamstring muscle injury is the first

injury diagnosis during top-level international athletics championships between 2007 and 2015. In: *British Journal of Sports Medicine* 50 (2016), Vol. 10, 619–630. – DOI 10.1136/bjsports–2015–095559. – ISSN 0306–3674

[Freund and Schapire 1996] FREUND, Yoav; SCHAPIRE, Robert E.: Experiments with a New Boosting Algorithm. In: *Mac hine Learn. Proc. Thirteen. Int. Confer ence, 1996*, 1996

[Froeling et al. 2015] FROELING, Martijn; OUDEMAN, Jos; STRIJKERS, Gustav J.; MAAS, Mario; DROST, Maarten R.; NICOLAY, Klaas ; NEDERVEEN, Aart J.: Muscle Changes Detected with Diffusion-Tensor Imaging after Long-Distance Running. In: *Radiology* 274 (2015), Vol. 2, 548–562. – DOI 10.1148/radiol.14140702

[Gavin 2013] GAVIN, Henri P.: The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems. In: *Department of Civil and Environmental Engineering, Duke University* (2013). – DOI 10.1080/10426914.2014.941480. – ISBN 9780898713527

[Gilles et al. 2016] GILLES, Benjamin; DE BOURGUIGNON, Charles; CROISILLE, Pierre; MILLET, Grégoire; BEUF, Olivier ; VIALLON, Magalie: Automatic segmentation for volume quantification of quadriceps muscle head: a longitudinal study in athletes enrolled in extreme mountain ultra-marathon. In: *ISMRM: International Society for Magnetic Resonance in Medicine*. Singapour, Singapore, Mai 2016

[Gilles and Magnenat-Thalmann 2010] GILLES, Benjamin; MAGNENAT-THALMANN, Nadia: Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations. In: *Medical Image Analysis* (2010). – DOI 10.1016/j.media.2010.01.006. – ISBN 1361–8423 (Electronic)\n1361–8415 (Linking)

[Goodfellow et al. 2014] GOODFELLOW, Ian J.; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron ; BENGIO, Yoshua: Generative Adversarial Networks. (2014), Juni. https://arxiv.org/abs/1406.2661

[Gray and Lewis 1918] GRAY, H.; LEWIS, W.H.: *Anatomy of the Human Body*. Lea & Febiger, 1918 https://books.google.fr/books?id=Ur9qAAAAMAAJ

[Grenier et al. 2006] GRENIER, T.; REVOL-MULLER, C.; COSTES, N.; JANIER, M. ; GIMENEZ, G.: 3D Robust Adaptive Region Growing for segmenting [18F] fluoride ion PET images. In: *2006 IEEE Nuclear Science Symposium Conference Record* Bd. 5, 2006, pp. 2644–2648

[Haacke et al. 1999] HAACKE, E M.; BROWN, Robert W.; THOMPSON, Michael R. ; VENKATESAN, Ramesh: *Haacke - Magnetic Resonance Imaging - Physical Principles and Sequence Design.pdf*

[Haider et al. 2011] HAIDER, Waqas; SHARIF, Muhammad ; RAZA, Mudassar: Achieving accuracy in early stage tumor identification systems based on image segmentation and 3D structure analysis. In: *Comput. Eng. Intell. Syst.* (2011)

[Handels et al. 2014] HANDELS, Heinz; SCHMIDT-RICHBERG, Alexander ; EHRHARDT, Jan: A Flexible Variational Registration Framework. (2014), pp. 1–12

[Haque et al. 2019] HAQUE, H.; HASHIMOTO, M.; UETAKE, Nozomu ; JINZAKI, M.: Semantic Segmentation of Thigh Muscle using 2.5D Deep Learning Network Trained with Limited Datasets. In: *ArXiv* abs/1911.09249 (2019)

[Harrison et al. 2001] HARRISON, B C.; ROBINSON, D; DAVISON, B J.; FOLEY, B; SEDA, E ; BYRNES, W C.: Treatment of exercise-induced muscle injury via hyperbaric oxygen therapy. In: *Medicine and science in sports and exercise* 33 (2001), Vol. 1, pp. 36–42. – DOI 10.1097/00005768–200101000–00007. – ISSN 0195–9131

[Hastie et al. 2009] HASTIE, Trevor; TIBSHIRANI, Robert ; FRIEDMAN, Jerome: *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009

[Hatt et al. 2018] HATT, Mathieu; VALLIERES, Martin; VISVIKIS, Dimitris ; ZWANENBURG, Alex: IBSI: an international community radiomics standardization initiative. In: *Journal of Nuclear Medicine* (2018). – ISSN 0161–5505

[He et al. 2020] HE, K.; GKIOXARI, G.; DOLLÁR, P. ; GIRSHICK, R.: Mask R-CNN. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020), Vol. 2, pp. 386–397. – DOI 10.1109/TPAMI.2018.2844175

[He et al. 2016] HE, K.; ZHANG, X.; REN, S. ; SUN, J.: Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778

[Hegadi et al. 2010] HEGADI, Ravindra; KOP, Arpana ; HANGARGE, Mallikarjun: A Survey on Deformable Model and its Applications to Medical Imaging. In: *Int. J. Comput. Appl.* (2010)

[Hoffman et al. 2012] HOFFMAN, Martin D.; INGWERSON, Julie L.; ROGERS, Ian R.; HEW-BUTLER, Tamara ; STUEMPFLE, Kristin J.: Increasing creatine kinase concentrations at the 161-km western states endurance run. In: *Wilderness and Environmental Medicine* 23 (2012), Vol. 1, pp. 56–60

[Huang et al. 2020] HUANG, Huimin; LIN, Lanfen; TONG, Ruofeng; HU, Hongjie; ZHANG, Qiaowei; IWAMOTO, Yutaro; HAN, Xianhua; CHEN, Yen-Wei ; WU, Jian: UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. (2020), Vol. ii, pp. 1055–1059. – DOI 10.1109/icassp40776.2020.9053405

[Ioffe and Szegedy 2015] IOFFE, Sergey; SZEGEDY, Christian: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *32nd International Conference on Machine Learning, ICML 2015*, 2015. – ISBN 9781510810587

[Išgum et al. 2009] IŠGUM, Ivana; STARING, Marius; RUTTEN, Annemarieke; PROKOP, Mathias; VIERGEVER, Max A. ; VAN GINNEKEN, Bram: Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans. In: *IEEE Trans. Med. Imaging* (2009). – DOI 10.1109/TMI.2008.2011480. – ISBN 1558–254X (Electronic)\n0278–0062 (Linking)

[Jha et al. 2019] JHA, Debesh; SMEDSRUD, Pia H.; RIEGLER, Michael A.; JOHANSEN, Dag; DE LANGE, Thomas; HALVORSEN, Pal ; JOHANSEN, Havard D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: *Proceedings - 2019 IEEE International Symposium on Multimedia, ISM 2019* (2019), pp. 225–230. – DOI 10.1109/ISM46123.2019.00049. ISBN 9781728156064

[Jog et al. 2013] JOG, Amod; ROY, Snehashis; CARASS, Aaron ; PRINCE, Jerry L.: Magnetic resonance image synthesis through patch regression. In: *Proceedings - International Symposium on Biomedical Imaging*, 2013. – ISBN 9781467364546

[Kan et al. 2009] KAN, Hermien E.; SCHEENEN, Tom W.; WOHLGEMUTH, Marielle; KLOMP, Dennis W.; LOOSBROEK-WAGENMANS, Ivonne van; PADBERG, George W. ; HEERSCHAP, Arend: Quantitative MR imaging of individual muscle involvement in facioscapulohumeral muscular dystrophy. In: *Neuromuscular Disorders* (2009). – DOI 10.1016/j.nmd.2009.02.009. – ISSN 09608966

[Kasiri et al. 2014] KASIRI, Keyvan; FIEGUTH, Paul ; CLAUSI, David A.: Cross modality label fusion in multi-atlas segmentation. In: *2014 IEEE Int. Conf. Image Process. ICIP 2014*, 2014. – ISBN 9781479957514

[Khare and Tiwary 2005] KHARE, Ashish; TIWARY, Uma S.: Soft-Thresholding for Denoising of Medical Images : A Multiresolution Approach. In: *International Journal of Wavelets Multiresolution and Information Processing* 3 (2005), 12, pp. 477–496. – DOI 10.1142/S021969130500097X

[Kikinis et al. 2014] KIKINIS, Ron; PIEPER, Steve D. ; VOSBURGH, Kirby G.: 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. Version: 2014. In: JOLESZ, Ferenc A. (Hrsg.): *Intraoperative Imaging and Image-Guided Therapy*. New York, NY : Springer New York, 2014. – DOI 10.1007/978–1–4614–7657–3_19. – ISBN 978–1–4614–7657–3, pp. 277–289

[Kingma and Ba 2015] KINGMA, Diederik P.; BA, Jimmy: Adam: A Method for Stochastic Optimization. In: *CoRR* abs/1412.6980 (2015)

[Klein et al. 2010] KLEIN, Stefan; STARING, Marius; MURPHY, Keelin; VIERGEVER, Max A. ; PLUIM, Josien P W.: Elastix: A toolbox for intensity-based medical image registration. In: *IEEE Trans. Med. Imaging* 29 (2010), Vol. 1, pp. 196–205

[Knechtle and Nikolaidis 2018] KNECHTLE, Beat; NIKOLAIDIS, Pantelis T.: *Physiology and pathophysiology in ultra-marathon running*

[Krizhevsky et al. 2012] KRIZHEVSKY, Alex; SUTSKEVER, Ilya ; HINTON, Geoffrey E.: ImageNet Classification with Deep Convolutional Neural Networks. In: PEREIRA, F. (Hrsg.); BURGES, C. J. C. (Hrsg.); BOTTOU, L. (Hrsg.) ; WEINBERGER, K. Q. (Hrsg.): *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105

[Le Troter et al. 2016] LE TROTER, Arnaud; FOURÉ, Alexandre; GUYE, Maxime; CONFORT-GOUNY, Sylviane; MATTEI, Jean-Pierre; GONDIN, Julien; SALORT-CAMPANA, Emmanuelle ; BENDAHAN, David: Volume measurements of individual muscles in human quadriceps femoris using atlas-based segmentation approaches. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 29 (2016), Vol. 2, 245–257. – DOI 10.1007/s10334–016–0535–6. – ISSN 1352–8661

[Lecun et al. 1998] LECUN, Y.; BOTTOU, L.; BENGIO, Y. ; HAFFNER, P.: Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86 (1998), Vol. 11, pp. 2278–2324

[Lee et al. 2015] LEE, Chen-Yu; XIE, Saining; GALLAGHER, Patrick; ZHANG, Zhengyou ; TU, Zhuowen: Deeply-Supervised Nets. San Diego, California, USA : PMLR, 09–12 May 2015 (Proceedings of Machine Learning Research), 562–570

[Lee et al. 1996] LEE, Seungyong; WOLBERG, George; CHWA, Kyung-Yong ; SHIN, Sung Y.: Image Metamorphosis with Scattered Feature Constraints. In: *IEEE Trans. Vis. Comput. Graph.* 2 (1996), Vol. 4, 337–354. http://dblp.uni-trier.de/db/journals/tvcg/tvcg2.html{#}LeeWCS96

[Lee et al. 2008] LEE, T H.; FAUZI, M F A. ; KOMIYA, R: *Segmentation of CT Brain Images Using K-means and EM Clustering*. 2008. http://doi.org/10.1109/cgiv.2008.17. – ISBN 978–0–7695–3359–9

[Leporq et al. 2017] LEPORQ, Benjamin; LE TROTER, Arnaud; LE FUR, Yann; SALORT-CAMPANA, Emmanuelle; GUYE, Maxime; BEUF, Olivier; ATTARIAN, Shahram ; BENDAHAN, David: Combined quantification of fatty infiltration, T 1-relaxation times and T 2*-relaxation times in normal-appearing skeletal muscle of controls and dystrophic patients. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 30 (2017), aug, Vol. 4, 407–415. – DOI 10.1007/s10334–017–0616–1. – ISSN 1352–8661

[Leporq et al. 2013] LEPORQ, Benjamin; RATINEY, Hélène; PILLEUL, Frank ; BEUF, Olivier: Liver fat volume fraction quantification with fat and water T1 and T2*estimation and accounting for NMR multiple components in patients with chronic liver disease at 1.5 and 3.0 T. In: *European Radiology* 23 (2013), Vol. 8, pp. 2175–86. – DOI 10.1007/s00330–013–2826–x. – ISSN 09387994

[Leyendecker et al. 2010] LEYENDECKER, J R.; BROWN, J J. ; MERKLE, E M.: *Practical Guide to Abdominal and Pelvic MRI*. Wolters Kluwer/Lippincott Williams & Wilkins Health, 2010 https://books.google.fr/books?id=sTiJxvWU0mEC. – ISBN 9781605471440

[Li et al. 2006] LI, Jing; ZHU, Shan'an ; BIN, He: Medical image segmentation techniques. In: *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi* (2006). – ISSN 10015515

[Li et al. 2008] LI, Min; HUANG, Tinglei ; ZHU, Gangqiang: Improved fast fuzzy c-means algorithm for medical MR images segmentation. In: *Proc. - 2nd Int. Conf. Genet. Evol. Comput. WGEC 2008*, 2008. – ISBN 9780769533346

[Li et al. 2019] LI, P.; ZHOU, Xiao-Yun; WANG, Z. ; YANG, G.: Z-Net: an Asymmetric 3D DCNN for Medical CT Volume Segmentation. In: *ArXiv* abs/1909.07480 (2019)

[Li et al. 2018] LI, Xiang; CHEN, Shuo; HU, Xiaolin ; YANG, Jian: Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift. (2018). http://arxiv.org/abs/1801.05134

[Liang et al. 2000] LIANG, Z P.; LAUTERBUR, P C.; MEDICINE, IEEE E. ; SOCIETY, Biology: *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*. SPIE Optical Engineering Press, 2000 (IEEE Press series in biomedical engineering). https://books.google.fr/books?id=sRyEQgAACAAJ. – ISBN 9780819435163

[Litjens et al. 2017] LITJENS, G.; KOOI, Thijs; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. V. D.; GINNEKEN, B. ; SÁNCHEZ, C.: A survey on deep learning in medical image analysis. In: *Medical image analysis* 42 (2017), pp. 60–88

[Lorensen and Cline 1987] LORENSEN, William E.; CLINE, Harvey E.: MARCHING CUBES: A HIGH RESOLUTION 3D SURFACE CONSTRUCTION ALGORITHM. In: *Computer Graphics (ACM)* (1987). – DOI 10.1145/37402.37422. – ISSN 00978930

[Lowe 2004] LOWE, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60 (2004), pp. 91–110

[Maeo et al. 2017] MAEO, Sumiaki; ANDO, Yukino; KANEHISA, Hiroaki ; KAWAKAMI, Yasuo: Localization of damage in the human leg muscles induced by downhill running. In: *Scientific Reports* 7 (2017), Vol. 1, pp. 5769. – DOI 10.1038/s41598–017–06129–8. – ISBN 4159801706

[Maes et al. 2015] MAES, F.; LOECKX, D.; VANDERMEULEN, D. ; SUETENS, P.: Image registration using mutual information. Version: 2015. In: *Handbook of Biomedical Imaging: Methodologies and Clinical Research*. 2015. – DOI 10.1007/978–0–387–09749–7_16. – ISBN 9780387097497, pp. 295–308

[MATLAB 2017] MATLAB: *version 9.3.0.713579 (R2017b)*. Natick, Massachusetts : The MathWorks Inc., 2017

[Maufrais et al. 2016] MAUFRAIS, Claire; MILLET, Grégoire P; SCHUSTER, Iris; RUPP, Thomas ; NOTTIN, Stéphane: Progressive and biphasic cardiac responses during extreme mountain ultramarathon. In: *American Journal of Physiology - Heart and Circulatory Physiology* 310 (2016), Vol. 10, H1340–H1348. – DOI 10.1152/ajpheart.00037.2016. – ISSN 0363–6135

[McCully and Faulkner 1985] MCCULLY, K K.; FAULKNER, J A.: Injury to skeletal muscle fibers of mice following lengthening contractions. In: *Journal of Applied Physiology* 59 (1985), Vol. 1, 119–126. http://jap.physiology.org/content/59/1/119. – ISSN 8750–7587

[Metaxas 1996] METAXAS, D.N.: *Physics-Based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging*. Springer US, 1996 (The Springer International Series in Engineering and Computer Science). https://books.google.fr/books?id=liUUSKBsnDoC. – ISBN 9780792398400

[Millet and Millet 2012] MILLET, Grégoire P; MILLET, Guillaume Y.: Ultramarathon is an outstanding model for the study of adaptive responses to extreme load and stress. In: *BMC Medicine* 10 (2012), Vol. 1, 1–3. – DOI 10.1186/1741–7015–10–77. – ISSN 1741–7015

[Milletari et al. 2016] MILLETARI, F.; NAVAB, N. ; AHMADI, S.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571

[Moreau et al. 2016] MOREAU, Baptiste; DICKO, Ali-Hamadi; MAILLIEZ, Pierre; PORTEJOIE, Pierre; LECOMTE, Christophe; BAH, Mamadou; GRENIER, T.; JOLIVET, Erwan; PETIT, Philippe; FRECHEDE, Bertrand; FAURE, Francois; GILLES, Benjamin ; BEILLAS, Philippe: A segmentation pipeline for the creation of statistical shape models in the PIPER project, 2016

[Muda and Salam 2011] MUDA, T. Z. T.; SALAM, R. A.: Blood cell image segmentation using hybrid K-means and median-cut algorithms. In: *2011 IEEE International Conference on Control System, Computing and Engineering*, 2011, pp. 237–243

[Nair and Hinton 2010] NAIR, Vinod; HINTON, Geoffrey E.: Rectified Linear Units Improve Restricted Boltzmann Machines. In: FÜRNKRANZ, Johannes (Hrsg.); JOACHIMS, Thorsten (Hrsg.): *ICML*, Omnipress, 2010, 807-814

[Newell et al. 2016] NEWELL, Alejandro; YANG, Kaiyu ; DENG, Jia: Stacked Hourglass Networks for Human Pose Estimation. In: LEIBE, Bastian (Hrsg.); MATAS,

Jiri (Hrsg.); SEBE, Nicu (Hrsg.) ; WELLING, Max (Hrsg.): *Computer Vision – ECCV 2016*. Cham : Springer International Publishing, 2016. – ISBN 978–3–319–46484–8, pp. 483–499

[Ng et al. 2008] NG, H P.; HUANG, S; ONG, S H.; FOONG, K C.; GOH, P S. ; NOWINSKI, W L.: Medical image segmentation using watershed segmentation with texture-based region merging. In: *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.* (2008). – DOI 10.1109/IEMBS.2008.4650096. – ISBN 9781424418152

[Nguyen et al. 2019a] NGUYEN, A.; GRANGE, S.; COURT, L.; BARRAL, F.G. ; EDOUARD, P.: Localisation en IRM des lésions musculaires des ischio-jambiers survenues lors de la pratique sportive et liens avec le mécanisme lésionnel : résultats préliminaires de l'étude HAMMER (Hamstring mechanics and MRI). In: *Journal de Traumatologie du Sport* 36 (2019), Vol. 2, 86-95. – DOI 10.1016/j.jts.2019.03.006. – ISSN 0762–915X

[Nguyen et al. 2018] NGUYEN, Hoai-Thu; CROISILLE, Pierre; VIALLON, Magalie; DE BOURGUIGNON, Charles; GRANGE, Rémi; GRANGE, Sylvain ; GRENIER, Thomas: Robust multi-atlas MRI segmentation with corrective learning for quantification of local quadriceps muscles inflammation changes during a longitudinal study in athletes. In: *ISMRM: International Society for Magnetic Resonance in Medicine*. Paris, 2018

[Nguyen et al. 2019b] NGUYEN, Hoai-Thu; CROISILLE, Pierre; VIALLON, Magalie; LECLERC, Sarah; GRANGE, Sylvain; GRANGE, Rémi; BERNARD, Olivier ; GRENIER, Thomas: Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised U-Net. In: *International Conference on Medical Imaging with Deep Learning*. London, 2019

[Nguyen et al. 2021a] NGUYEN, Hoai-thu; GRANGE, Sylvain; LEPORQ, Benjamin; VIALLON, Magalie; CROISILLE, Pierre ; GRENIER, Thomas: Impact of Distortion on Local Radiomic Analysis of Quadriceps Based on Quantitative Magnetic Resonance Imaging Data. In: *International Journal of Pharma Medicine and Biological Sciences* 10 (2021), Vol. 2

[Nguyen et al. 2019c] NGUYEN, Hoai-Thu; GRENIER, Thomas; LEPORQ, Benjamin; BEY, Loïc; VIALLON, Magalie ; CROISILLE, Pierre: Evaluation of local changes in femoral bone marrow during a mountain ultra-marathon with quantitative MRI. In: *ISMRM: International Society for Magnetic Resonance in Medicine*. Montréal, 2019

[Nguyen et al. 2021b] NGUYEN, Hoai-Thu; GRENIER, Thomas; LEPORQ, Benjamin; LE GOFF, Caroline; GILLES, Benjamin; GRANGE, Sylvain; GRANGE, Rémi; MILLET, Grégoire P.; BEUF, Olivier; CROISILLE, Pierre ; VIALLON, Magalie: Quantitative Magnetic Resonance Imaging Assessment of the Quadriceps Changes during an Extreme Mountain Ultramarathon. In: *Medicine & Science in Sports & Exercise* 53 (2021), Vol. 4, pp. 869–881. – DOI 10.1249/mss.0000000000002535. – ISSN 0195–9131

[Nie and Shen 2013] NIE, Jingxin; SHEN, Dinggang: Automated segmentation of mouse brain images using multi-atlas multi-ROI deformation and label fusion. In: *Neuroinformatics* (2013). – DOI 10.1007/s12021–012–9163–0. – ISBN 1202101291

[Nosaka and Clarkson 1996] NOSAKA, Kazunori; CLARKSON, P M.: Changes in indicators of inflammation after eccentric exercise of the elbow flexors. In: *Medicine and science in sports and exercise* 28 (1996), Vol. 8, pp. 953–961

[Nurenberg et al. 1992] NURENBERG, P; GIDDINGS, C J.; STRAY-GUNDERSEN, J; FLECKENSTEIN, J L.; GONYEA, W J. ; PESHOCK, R M.: MR imaging-guided muscle biopsy for correlation of increased signal intensity with ultrastructural change and delayed-onset muscle soreness after exercise. In: *Radiology* 184 (1992), Vol. 3, 865–869. – DOI 10.1148/radiology.184.3.1509081

[Ono et al. 2018] ONO, Yuki; TRULLS, Eduard; FUA, Pascal ; YI, Kwang M.: *LF-Net: Learning Local Features from Images*. 2018

[Patten et al. 2003] PATTEN, Carolynn; MEYER, Ronald A. ; FLECKENSTEIN, James L.: T2 mapping of muscle. In: *Seminars in musculoskeletal radiology* 7 (2003), Vol. 4, pp. 297–305. – DOI 10.1055/s–2004–815677. – ISBN 1089–7860

[Peetrons 2001] PEETRONS, P.: Ultrasound of muscles. In: *European Radiology* 12 (2001), pp. 35–43

[Perslev et al. 2019] PERSLEV, Mathias; DAM, Erik B.; PAI, Akshay ; IGEL, Christian: One Network to Segment Them All: A General, Lightweight System for Accurate 3D Medical Image Segmentation. In: SHEN, Dinggang (Hrsg.); LIU, Tianming (Hrsg.); PETERS, Terry M. (Hrsg.); STAIB, Lawrence H. (Hrsg.); ESSERT, Caroline (Hrsg.); ZHOU, Sean (Hrsg.); YAP, Pew-Thian (Hrsg.) ; KHAN, Ali (Hrsg.): *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham : Springer International Publishing, 2019. – ISBN 978–3–030–32245–8, pp. 30–38

[Pham et al. 2000] PHAM, Dzung L.; XU, Chenyang ; PRINCE, Jerry L.: Current Methods in Medical Image Segmentation. In: *Annual Review of Biomedical Engineering* 2 (2000), Vol. 1, 315–337. – DOI 10.1146/annurev.bioeng.2.1.315

[Pizza et al. 2002] PIZZA, Francis X.; KOH, Timothy J.; MCGREGOR, Stephen J. ; BROOKS, Susan V.: Muscle inflammatory cells after passive stretches, isometric contractions, and lengthening contractions. In: *Journal of Applied Physiology* 92 (2002), Vol. 5, 1873–1878. – DOI 10.1152/japplphysiol.01055.2001. – ISSN 8750–7587

[Ploutz-Snyder et al. 1997] PLOUTZ-SNYDER, Lori L.; NYREN, Sven; COOPER, Thomas G.; POTCHEN, E. J. ; MEYER, Ronald A.: Different effects of exercise and edema on T2 relaxation in skeletal muscle. In: *Magnetic Resonance in Medicine* 37 (1997), pp. 676–682. – DOI 10.1002/mrm.1910370509. – ISSN 07403194

[Prescott et al. 2011] PRESCOTT, Jeffrey W.; BEST, Thomas M.; SWANSON, Mark S.; HAQ, Furqan; JACKSON, Rebecca D. ; GURCAN, Metin N.: Anatomically Anchored Template-Based Level Set Segmentation: Application to Quadriceps Muscles in MR Images from the Osteoarthritis Initiative. In: *J. Digit. Imaging* 24 (2011), Vol. 1, 28–43. – DOI 10.1007/s10278–009–9260–2. – ISSN 1618–727X

[Qayyum et al. 2018] QAYYUM, Adnan; ANWAR, S.; MAJID, M.; AWAIS, M. ; ALNOWAMI, M.: Medical Image Analysis using Convolutional Neural Networks: A Review. In: *Journal of Medical Systems* 42 (2018), pp. 1–13

[Qu et al. 2020] QU, Liangqiong; ZHANG, Yongqin; WANG, Shuai; YAP, Pew T. ; SHEN, Dinggang: Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains. In: *Medical Image Analysis* (2020). – DOI 10.1016/j.media.2020.101663. – ISSN 13618423

[Ramus et al. 2010] RAMUS, Liliane; COMMOWICK, Olivier ; MALANDAIN, Grégoire: Construction of patient specific atlases from locally most similar anatomical pieces. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2010. – ISBN 3642157106

[Rizwan I Haque and Neubert 2020] RIZWAN I HAQUE, Intisar; NEUBERT, Jeremiah: Deep learning approaches to biomedical image segmentation. In: *Informatics in Medicine Unlocked* 18 (2020), 100297. – DOI 10.1016/j.imu.2020.100297. – ISSN 2352–9148

[Rizzo et al. 2018] RIZZO, Stefania; BOTTA, Francesca; RAIMONDI, Sara; ORIGGI, Daniela; FANCIULLO, Cristiana; MORGANTI, Alessio G. ; BELLOMI, Massimo: *Radiomics: the facts and the challenges of image analysis*

[Rohlfing et al. 2005] ROHLFING, T.; RUSSAKOFF, Daniel B.; DENZLER, Joachim; MORI, K. ; MAURER, C. R.: Progressive attenuation fields: Fast 2D-3D image registration without precomputation. In: *Medical Physics* 32 (2005), pp. 2870–2880

[Rohlfing et al. 2004] ROHLFING, Torsten; BRANDT, Robert; MENZEL, Randolf ; MAURER, Calvin R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. In: *Neuroimage* (2004). – DOI 10.1016/j.neuroimage.2003.11.010. – ISBN 1053–8119

[Ronneberger et al. 2015] RONNEBERGER, Olaf; FISCHER, Philipp ; BROX, Thomas: U-net: Convolutional networks for biomedical image segmentation. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. – ISBN 9783319245737

[Ruder 2016] RUDER, Sebastian: *An overview of gradient descent optimization algorithms.* http://arxiv.org/abs/1609.04747. Version: 2016. – cite arxiv:1609.04747Comment: Added derivations of AdaMax and Nadam

[Rueckert et al. 1999] RUECKERT, D.; SONODA, L. I.; HAYES, C.; HILL, D. L. G.; LEACH, M. O. ; HAWKES, D. J.: Nonrigid registration using free-form deformations: application to breast MR images. In: *IEEE Transactions on Medical Imaging* 18 (1999), Aug, Vol. 8, pp. 712–721. – DOI 10.1109/42.796284. – ISSN 0278–0062

[Rumelhart et al. 1995] RUMELHART, David E.; DURBIN, Richard; GOLDEN, Richard ; CHAUVIN, Yves: *Backpropagation: The basic theory.*

[Saab et al. 1999] SAAB, George; THOMPSON, R. T. ; MARSH, Greg D.: Multicomponent T2 relaxation of in vivo skeletal muscle. In: *Magnetic Resonance in Medicine* 42 (1999), Vol. 1, pp. 150–7. – DOI 10.1002/(SICI)1522–2594(199907)42:1<150::AID–MRM20>3.0.CO;2–5. – ISBN 0740–3194 (Print)

[Safavian and Landgrebe 1991] SAFAVIAN, S. R.; LANDGREBE, David: A Survey of Decision Tree Classifier Methodology. In: *IEEE Trans. Syst. Man Cybern.* (1991). – DOI 10.1109/21.97458. – ISBN 0018–9472

[Santurkar et al. 2018] SANTURKAR, Shibani; TSIPRAS, Dimitris; ILYAS, Andrew ; MADRY, Aleksander: How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift). In: *Advances in Neural Information Processing Systems 31* (2018). – ISBN 0362–2436

[Saugy et al. 2013] SAUGY, Jonas; PLACE, Nicolas; MILLET, Guillaume Y.; DEGACHE, Francis; SCHENA, Federico ; MILLET, Grégoire P: Alterations of Neuromuscular Function after the World's Most Challenging Mountain Ultra-Marathon. In: *PLoS ONE* 8 (2013), Vol. 6, 1–11. – DOI 10.1371/journal.pone.0065596

[Sayers and Clarkson 2001] SAYERS, P S.; CLARKSON, M P.: Force recovery after eccentric exercise in males and females. In: *European Journal of Applied Physiology* 84 (2001), Vol. 1, 122–126. – DOI 10.1007/s004210000346. – ISSN 1439–6327

[Sayers et al. 1999] SAYERS, S P.; CLARKSON, P M.; ROUZIER, P A. ; KAMEN, G: Adverse events associated with eccentric exercise protocols: six case studies. In: *Medicine and Science in Sports and Exercise* 31 (1999), Vol. 12, 1697–1702. http://www.ncbi.nlm.nih.gov/pubmed/10613417. ISBN 0195–9131 (Print)\r0195–9131 (Linking)

[Sharma and Aggarwal 2010] SHARMA, N.; AGGARWAL, L.: Automated medical image segmentation techniques. In: *Journal of Medical Physics / Association of Medical Physicists of India* 35 (2010), pp. 3 – 14

[Siddique et al. 2006] SIDDIQUE, I.; BAJWA, I. S.; NAVEED, M. S. ; CHOUDHARY, M. A.: Automatic Functional Brain MR Image Segmentation using Region Growing and Seed Pixel. In: *2006 ITI 4th International Conference on Information Communications Technology*, 2006. – ISSN 2329–6364, pp. 1–2

[Simo-Serra et al. 2015] SIMO-SERRA, Edgar; TRULLS, Eduard; FERRAZ, Luis; KOKKINOS, Iasonas; FUA, Pascal ; MORENO-NOGUER, Francesc: Discriminative Learning of Deep Convolutional Feature Point Descriptors. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 118–126

[Srivastava et al. 2014] SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya ; SALAKHUTDINOV, Ruslan: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In: *Journal of Machine Learning Research* 15 (2014), Vol. 56, 1929-1958. http://jmlr.org/papers/v15/srivastava14a.html

[Taha and Hanbury 2015] TAHA, Abdel A.; HANBURY, Allan: Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. In: *BMC Medical Imaging* (2015). – DOI 10.1186/s12880–015–0068–x. – ISSN 14712342

[Tajbakhsh et al. 2020] TAJBAKHSH, Nima; JEYASEELAN, Laura; LI, Qian; CHIANG, Jeffrey N.; WU, Zhihao ; DING, Xiaowei: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. In: *Medical Image Analysis* 63 (2020), 101693. – DOI 10.1016/j.media.2020.101693. – ISSN 1361–8415

[Takahashi et al. 1994] TAKAHASHI, Hideyuki; KUNO, Shin-ya; MIYAMOTO, Toshikazu; YOSHIOKA, Hiroshi; INAKI, Mitsuharu; AKIMA, Hiroshi; KATSUTA, Shigeru; ANNO, Izumi ; ITAI, Yuji: Changes in magnetic resonance images in human skeletal muscle after eccentric exercise. In: *European Journal of Applied Physiology and Occupational Physiology* 69 (1994), Vol. 5, 408–413. – DOI 10.1007/BF00865404. – ISSN 1439–6327

[Tamez-Peña et al. 2012] TAMEZ-PEÑA, Jose; FARBER, J.M.; GONZÁLEZ, Patricia; SCHREYER, Edward; SCHNEIDER, Erika ; TOTTERMAN, Saara: Unsupervised Segmentation and Quantification of Anatomical Knee Features: Data From the Osteoarthritis Initiative. In: *IEEE transactions on bio-medical engineering* 59 (2012), 02, pp. 1177–86. – DOI 10.1109/TBME.2012.2186612

[Tawara et al. 2011] TAWARA, Noriyuki; NITTA, Osamu; KURUMA, Hironobu; NIITSU, Mamoru ; ITOH, Akiyoshi: T2 mapping of muscle activity using ultrafast imaging. In: *Magnetic resonance in medical sciences : MRMS : an official journal of Japan Society of Magnetic Resonance in Medicine* 10 (2011), Vol. 2, pp. 85–91. – ISSN 1880–2206

[Thirion 1998] THIRION, J.-P.: Image matching as a diffusion process: an analogy with Maxwell's demons. In: *Med. Image Anal.* 2 (1998), Vol. 3, 243–260. – DOI 10.1016/S1361–8415(98)80022–4. – ISSN 1361–8415

[Thuny et al. 2012] THUNY, Franck; LAIREZ, Olivier; ROUBILLE, François; MEWTON, Nathan; RIOUFOL, Gilles; SPORTOUCH, Catherine; SANCHEZ, Ingrid; BERGEROT, Cyrille; THIBAULT, Hélène; CUNG, Thien T.; FINET, Gérard; ARGAUD, Laurent; REVEL, Didier; DERUMEAUX, Geneviève; BONNEFOY-CUDRAZ, Eric; ELBAZ, Meier; PIOT, Christophe; OVIZE, Michel ; CROISILLE, Pierre: Post-conditioning reduces infarct size and edema in patients with ST-segment elevation myocardial infarction. In: *Journal of the American College of Cardiology* 59 (2012), Vol. 24, pp. 2175–2181. – DOI 10.1016/j.jacc.2012.03.026. – ISSN 07351097

[Tidball 2005] TIDBALL, James G.: Inflammatory processes in muscle injury and repair. In: *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 288 (2005), Vol. 2, R345—-R353. – DOI 10.1152/ajpregu.00454.2004. – ISSN 0363–6119

[Tiidus 2008] TIIDUS, Peter M.: *Skeletal Muscle Damage and Repair*. 2008. – 337 S. – ISBN 9780736058674

[Tustison et al. 2010] TUSTISON, N. J.; AVANTS, B. B.; COOK, P. A.; ZHENG, Y.; EGAN, A.; YUSHKEVICH, P. A. ; GEE, J. C.: N4ITK: Improved N3 Bias Correction. In: *IEEE Transactions on Medical Imaging* 29 (2010), Vol. 6, pp. 1310–1320

[Tustison et al. 2017] TUSTISON, Nick; AVANTS, Brian; WANG, Hongzhi; XIE, Long; COUP, Pierrick; YUSHKEVICH, Paul ; MANJÓN, Vicente: A patch-based framework for new ITK functionality: Joint fusion, denoising, and non-local super-resolution. (2017), pp. 1–8

[Vallières et al. 2015] VALLIÈRES, M.; FREEMAN, C. R.; SKAMENE, S. R. ; EL NAQA, I.: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. In: *Physics in Medicine and Biology* 60 (2015), Vol. 14, pp. 5471–5496. – DOI 10.1088/0031–9155/60/14/5471. – ISBN 1361–6560 (Electronic)\r0031–9155 (Linking)

[Van Griethuysen et al. 2017] VAN GRIETHUYSEN, Joost J.; FEDOROV, Andriy; PARMAR, Chintan; HOSNY, Ahmed; AUCOIN, Nicole; NARAYAN, Vivek; BEETS-TAN, Regina G.; FILLION-ROBIN, Jean C.; PIEPER, Steve ; AERTS, Hugo J.: Computational radiomics system to decode the radiographic phenotype. In: *Cancer Research* (2017). – DOI 10.1158/0008–5472.CAN–17–0339. – ISSN 15387445

[Vernillo et al. 2015] VERNILLO, G; RINALDO, N; GIORGI, A; ESPOSITO, F; TRABUCCHI, P; MILLET, G P. ; SCHENA, F: Changes in lung function during an extreme mountain ultramarathon. In: *Scandinavian Journal of Medicine & Science in Sports* 25 (2015), Vol. 4, e374–e380. – DOI 10.1111/sms.12325. – ISSN 1600–0838

[Viallon et al. 2019] VIALLON, Magalie; LEPORQ, Benjamin; DRINDA, Stephan; WILHELMI DE TOLEDO, Françoise; GALUSCA, Bogdan; RATINEY, Helene ;

CROISILLE, Pierre: Chemical-shift-encoded magnetic resonance imaging and spectroscopy to reveal immediate and long-term multi-organs composition changes of a 14-days periodic fasting intervention: A technological and case report. In: *Frontiers in Nutrition* (2019). – DOI 10.3389/fnut.2019.00005. – ISSN 2296861X

[Vigneault et al. 2018] VIGNEAULT, Davis M.; XIE, Weidi; HO, Carolyn Y.; BLUEMKE, David A. ; NOBLE, J. A.: Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. In: *Medical Image Analysis* 48 (2018), 95 - 106. – DOI 10.1016/j.media.2018.05.008. – ISSN 1361–8415

[Walker et al. 2014] WALKER, Amy; LINEY, Gary; METCALFE, Peter ; HOLLOWAY, Lois: MRI distortion: Considerations for MRI based radiotherapy treatment planning. In: *Australasian Physical and Engineering Sciences in Medicine* (2014). – DOI 10.1007/s13246–014–0252–2. – ISSN 18795447

[Wang et al. 2017] WANG, H.; PRASANNA, P. ; SYEDA-MAHMOOD, T.: Fast anatomy segmentation by combining low resolution multi-atlas label fusion with high resolution corrective learning: An experimental study. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 223–226

[Wang et al. 2013] WANG, H.; SUH, J. W.; DAS, S. R.; PLUTA, J. B.; CRAIGE, C. ; YUSHKEVICH, P. A.: Multi-Atlas Segmentation with Joint Label Fusion. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013), March, Vol. 3, pp. 611–623. – DOI 10.1109/TPAMI.2012.143. – ISSN 0162–8828

[Wang et al. 2011] WANG, Hongzhi; DAS, Sandhitsu R.; SUH, Jung W.; ALTINAY, Murat; PLUTA, John; CRAIGE, Caryne; AVANTS, Brian ; YUSHKEVICH, Paul A.: A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. In: *NeuroImage* 55 (2011), Vol. 3, 968–985. – DOI 10.1016/j.neuroimage.2011.01.006. – ISSN 1053–8119

[Wang and Yushkevich 2013] WANG, Hongzhi; YUSHKEVICH, Paul: Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. In: *Frontiers in Neuroinformatics* 7 (2013), 27. – DOI 10.3389/fninf.2013.00027. – ISSN 1662–5196

[Welsch et al. 2014] WELSCH, Goetz H.; HENNIG, Friedrich F.; KRINNER, Sebastian ; TRATTNIG, Siegfried: *T2 and T2* Mapping*

[Yoo et al. 2002] YOO, Terry S.; ACKERMAN, Michael J.; LORENSEN, William E.; SCHROEDER, Will; CHALANA, Vikram; AYLWARD, Stephen; METAXAS, Dimitris ; WHITAKER, Ross: Engineering and algorithm design for an image processing API: A technical report on ITK - The Insight Toolkit. In: *Stud. Health Technol. Inform.* Bd. 85, 2002, pp. 586–592

[Zhang et al. 2018] ZHANG, Z.; LIU, Q. ; WANG, Y.: Road Extraction by Deep Residual U-Net. In: *IEEE Geoscience and Remote Sensing Letters* 15 (2018), Vol. 5, pp. 749–753

[Zhou et al. 2020] ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH, N. ; LIANG, J.: UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. In: *IEEE Transactions on Medical Imaging* 39 (2020), Vol. 6, pp. 1856–1867

[Zhou 2012] ZHOU, Zhi-Hua: *Ensemble Methods: Foundations and Algorithms*. 2012. http://doi.org/10.1201/b12207-2. – ISBN 978–1–4398–3003–1

[Zwanenburg et al. 2020] ZWANENBURG, Alex; VALLIÈRES, Martin; ABDALAH, Mahmoud A.; AERTS, Hugo J. W. L.; ANDREARCZYK, Vincent; APTE, Aditya; ASHRAFINIA, Saeed; BAKAS, Spyridon; BEUKINGA, Roelof J.; BOELLAARD, Ronald ; AL. et: The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. In: *Radiology* 295 (2020), May, Vol. 2, 328–338. – DOI 10.1148/radiol.2020191145. – ISSN 1527–1315

Appendices

APPENDIX A

Supplemental information about the MUST dataset

A.1 Demographic data

Time point	Sex (M/F)	Age (years)	Height (meters)	Weight (kg)	BMI ($kg\,m^{-2}$)	Limb dominance (R/L)
Pre ($n = 50$)	46/4	43 ± 9.1	1.75 ± 0.62	72.2 ± 8.0	23.6 ± 2.0	43/7
Post ($n = 31$)	30/1	43 ± 8.6	1.75 ± 0.64	71.7 ± 8.2	23.4 ± 2.0	27/4
Post+3 ($n = 27$)	27/0	43 ± 8.6	1.75 ± 0.56	70.8 ± 7.3	23.1 ± 2.0	23/4

TABLE A.1: *Demographic data of ultra-marathoners population*

A.2 MRI acquisition parameters

Parameters	**Anatomical isotropic 3D GRE dual-Echo**	**Quantitative 3D GRE multiecho**	**Quantitative T2 spin-echo multiecho**
Sequence	3D gradient echo	3D gradient echo	spin echo
K-space specificity	3D, cartesian	3D, cartesian	2D, cartesian
FOV (ms^2)	437×500	256×160	400×250
Slice thickness (ms^2)	1.3	5	10
Acq. pixel size (ms^2)	1.56×1.56	1.56×1.56	1.25×1.25
Interp. pixel size (ms^2)	0.78×0.78	1.56×1.56	1.25×1.25
Voxel size (ms^3)	0.79	12.17	12.17
Matrix size	320×280	256×160	320×200
Bandwidth (Hz/pixel)	372	1395	1395
Water-fat shift (pixel)	0.59	0.15	0.15
Repetition time (ms)	11.1	22	1580
Echo time (ms)	2.38/4.76	182-20.6	10.9-174.4
Echo spacing (ms)	2.38	2.52	10.9
Average	1	1	1
Flip angle (°)	10	5	90
Number of partitions/slices	16	48	7
3D volume explored (x, y, z)	$437.5 \times 208 \times 500$	$400 \times 280 \times 240$	$400 \times 250 \times 70$
Fat suppression	Dixon	None	None
GRAPPA factor	2	2	-

TABLE A.2: *MRI acquisition parameters. Abbreviations: SE - Spin Eche, GRE - Gradient Echo, Interp - Interpolated, Acq - Accquired*

A.3 Biological data

A.3.1 Data collecting

Blood and urinary samples were collected at each of the four sessions (Pre, Mid, Post, Post+3) within 10 min after arrival at each key point. Blood samples were drawn from an antecubital vein into a dry, heparinized, or EDTA tube according to the analysis to be performed and immediately centrifuged. Since it was not possible to carry out all the investigations on the same day by point-of-care technologies, plasma and serum were frozen at $-80°C$ within 20 min after blood collection for later analysis of muscle injury markers and biochemical variables. The hematology parameters (hemoglobin, red blood cells, white blood cells) were directly analyzed by a pocH-100i^{TM} automated hematology analyzer (Sysmex, Villepinte, France). Cobas 8000 (RocheDiagnostics, Manheim, Germany) was used to perform serial determinations for C-reactive protein (CRP), urinary creatinine, creatinine, calcium, chlorine, potassium, sodium, and cholesterol. The osmolality and urinary osmolality were measured on an Arkray Osmo Station OM-6050 (Menarini, Florence, Italy). All blood biomarkers analyzed in this study are listed in table A.3.

Biological markers (units)	**Abbreviation**
Hematocrit (%)	HTC
Red blood cells ($10^6/\mu L$)	RBC
Hemoglobin (g/dL)	HGB
Mean corpuscular volume (fL)	MCV
Mean corpuscular hemoglobin (pg)	MCH
Mean corpuscular hemoglobin concentration (g/dL)	MCHC
Platelets ($10^3/\mu L$)	PLT
White blood cells ($10^3/\mu L$)	WBC
Lymphocytes ($10^3/\mu L$)	LYM
Lymphocytes fraction (%)	LYM.pct
Mixed cell fraction (%)	MXD.pct
Mixed cell count ($10^3/\mu L$)	MXD
Neutrophil granulocytes ($10^3/\mu L$)	NEUT
Neutrophil granulocytes fraction (%)	NEUT.pct
Lactate (mmol/L)	Lact
Troponin T high sensitive (ng/L)	TnThs
Creatine kinase (UI/L)	CK
Creatine kinase MB isoenzyme (µg/L)	CKMB
CKMB / CK total ratio (µg/100UI)	CKMB.CK
Copeptine (pmol/L)	Copeptine
Suppression Tumorogenicity 2 (ng/mL)	ST2
Galectine-3 (ng/mL)	Gal3
N-Terminal natriuretic peptide (ng/L)	NtptoNBP
Myoglobin (µg/L)	Myo
Heart fatty acid binding globulin	hFABP
C-reactive protein (mg/L)	CRP

Myeloperoxydase (ng/mL)	MPO
Gluthathione reduced (µmol/L)	GRD
Gluthathione oxidase (µmol/L)	GOX
Lipid peroxidase (µmol/L)	POLX
Oxidized LDL (U/L)	OLDL
Conjugated bilirubin (mg/dL)	CBIL
Total bilirubin (mg/dL)	TBIL
Gamma-glutamyl transaminase (U/L)	GGT
Glutamic-oxaloacetic transaminase (U/L)	GOT
Glutamate pyruvate transaminase (U/L)	TGP
Lactate dehydrogenase (U/L)	LDH
Alkaline phosphatase (U/L)	ALP
Uric acid (mg/dL)	A.uric
Urinary creatinine (g/L)	CRU
Urinary neutrophil gelatinase-associated lipocaline (ng/mL)	NGAL
NGAL/ creatinine urinary ratio	NGALUCR
Creatinine (mg/dL)	Creatinine
Plasma urea (mg/dL)	Urea.pl
Cholesterol (mg/dL)	Cholest
High density lipoprotein (mg/dL)	HDL
High density lipoprotein / Cholesterol	HDL.Ch
Low density lipoprotein (mg/dL)	LDL
Non-HDL cholesterol (mg/dL)	Ch.non.HDL
Triglycerides (mg/dL)	TG
Calcium (mmol/dL)	Ca
Chloride (mmol/dL)	Cl
Potassium (mmol/dL)	K
Sodium (mmol/dL)	Na
Phosphate (mmol/dL)	Ph
Blood osmolality (mosm/kg)	Osm
Urinary osmolality (mosm/kg)	OsmU
Total proteins (g/L)	Pr.tot

TABLE A.3: *List of 58 biological markers analyzed in the study with their abbreviation used in R*

A.3.2 Preprocessing for missing data

Kernel density estimation method was used to handle missing individual biological values for correlation analysis of 55 qMRI-extracted features and the 58 blood biomarkers ($n = 72$)/ total 4640). A missing value $x_{(m,n)}$ of a variable (biomarker) X of the subject m at the time point n, with $n \in \{Pre, Mid, Post, Post + 3\}$, was estimated using the presented values of X of the subject m at the others time points and the observed values of X of the other subjects at all the time points.

$$x_{(m,n)} = \frac{\sum_{i=1..N, i\neq m} x_{i,n} \times g\left(\sqrt{\sum_{j=Pre..Post+3, j\neq n}(x_{m,j} - x_{i,j})^2}\right)}{\sum_{i=1..N, i\neq m} g\left(\sqrt{\sum_{j=Pre..Post+3, j\neq n}(x_{m,j} - x_{i,j})^2}\right)} \tag{A.1}$$

where N is the number of subjects and g is a kernel function. To minimize estimation error according to the asymptotic mean integrated squared error, we used the function Epanechnikov as kernel:

$$g(u) = 1 - u^2 \text{ for } |u| < 1,\ 0 \text{ otherwise} \tag{A.2}$$

Values were standardized before estimation and revert to original scale after processing.

A.3.3 Longitudinal analysis

The detailed longitudinal variations in biological markers throughout the race can be found in Nguyen et al. (2021b). As main blood biomarkers of muscle damage, serum creatine kinase (CK) and myoglobin levels peaked at +6598% and +4159% at Mid, while serum lactate dehydrogenase (LDH) levels peaked at +240% at Post and were elevated at Mid at +197% (all P-values of post hoc test against Pre are less than 0.01, see Figure A.1). For most of the biomarkers, the biomarker values at Pre were significantly different from those at the other time points (Mid, Post, Post+3), while the difference between Mid and Post was nonsignificant.

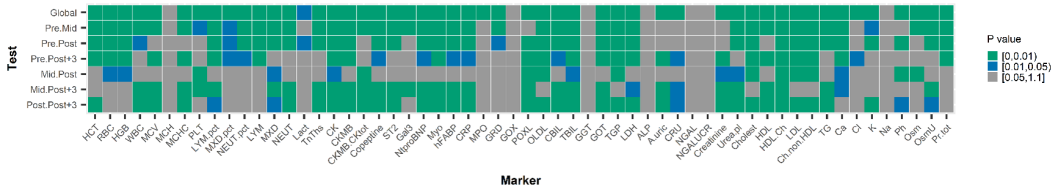

FIGURE A.1: *P-values of statistical tests in our longitudinal analysis on the blood and urinary biomarker data. A P-value less than .05 indicates a significant change between two time points.*

APPENDIX B

Evaluation of registration methods

In addition to the B-spline registration presented in Section 4.1.1, Thirion's demons algorithm (Thirion, 1998) is a very popular non-parametric registration method. The deformation field is optimized using local image forces which were computed independently for each voxel. The displacement estimated is then regularized using Gaussian smoothing. This algorithm is implemented in the Variational Registration Framework in ITK (Handels et al., 2014) (VRF). The VRF is a flexible framework for non-parametric variational image registration. The framework provides many choices of force terms and regularizers and is an open-source framework that can be modified to integrate more options. We can also restrain the transformation to the space of diffeomorphisms by using a stationary velocity field in place of a dense displacement field. The diffeomorphic registration does not allow folding to compute an invertible transformation to preserve the topology in images.

For our tests, we will only use the VRF to run the classic demons algorithm with Normalized Sum of Squared Differences-based force term and Gaussian regularization and take advantage of the diffeomorphic registration option.

B.1 Methods

B.1.1 Preliminary test

For the premilinary test, we programmed a function in C++ with ITK which applied random local B-spline deformations to an image and used the two methods above to register the original image to the deformed image to retrieve the same deformations. To preserve the anatomic structures in our image, the displacements need to satisfy the one-to-one property (Lee et al., 1996): Let $\Delta\phi_{ijk} = \phi_{ijk} - \phi^0_{ijk}$ be the displacement of the ijk^{th} control point from its initial position, $\mid \delta \mid_{\text{inf}} = max(\mid \delta \mid_1, \mid \delta \mid_2, \mid \delta \mid_3)$ where $\delta = (\delta_1, \delta_2, \delta_3)$. The total transformation is one-to-one if $\mid \Delta\phi_{ijk} \mid_{\text{inf}} \leq 0.48$ for all i, j, k.

B.1.2 Registration test

We carried out a registration test with T1W image of a runner at 2 time points Pre and Post. We chose a runner with visually accurate segmentation by Gilles *et al*.Gilles et al. (2016) at the two time points. The images at time point Pre and Post acted

as the moving image and the fixed image, respectively. The test followed the steps below:

1. Rigidly register the moving image to the fixed image with `elastix`(Klein et al., 2010) and apply the computed deformation field to the label image corresponding to the moving image.
2. Register the result image of step (1) to the fixed image with **Bspline deformation** model using `elastix` and apply the computed deformation field to the label image obtained from step (1).
3. Register the result image of step (1) to the fixed image with **demons algorithm** using the VRF and apply the computed deformation field to the label image obtained from step (1).
4. Register the moving image to the fixed image with **demons algorithm** using the VRF and apply the computed deformation field to the label image corresponding to the moving image.

We ended up with 3 different registration results from 3 different procedures: Rigid then Bspline registration, Rigid then Demons registration and Demons registration only.

B.1.3 Methods comparison

To compare the different methods, we computed for each method:

- a checkboard and an image of differences (subtraction of 2 images) of the resulted image and the fixed image
- a checkboard of the resulted label image and the label image corresponding to the fixed image
- different similarity metrics (Mutual Information, Normalized Mutual Information, Sum of Squared Differences and Normalized Cross Correlation). The formulas of the similarity metrics can be found below.

Notations: These notations are used for all the formulas in this section

F: fixed image

M: moving image

T: a transformation (displacement)

Ω_F: domain of fixed image

$\mid \Omega_F \mid$: number of voxels in fixed image

Mutual Information (MI): MI between two images A and B is defined as:

$$MI(A,B) = H(A) + H(B) - H(A,B)$$

where $H(A)$ and $H(B)$ are marginal entropies of images A and B respectively and $H(A,B)$ is their joint entropy. The entropies are defined as:

$$H(A) = -\int p_A(a) \log p_A(a) da$$

$$H(B) = -\int p_B(b) \log p_B(b) db$$

$$H(A,B) = -\int p_{AB}(a,b) \log p_{AB}(a,b) da db$$

where p_A, p_B and p_{AB} are respectively marginal probability density functions for A and B and their joint probability density function. The larger the MI, the more similar the two images.

Normalized Mutual Information (NMI): We computed two type of NMI (noted NMI1 and NMI2):

$$NMI1(A,B) = \frac{2MI(A,B)}{H(A) + H(B)}$$

$$NMI2(A,B) = \frac{H(A) + H(B)}{H(A,B)}$$

Sum of Squared Differences(SSD): The SSD is defined as:

$$SSD(T,F,M) = \sum_{x \in \Omega_F} (F(x) - M(T(x)))^2$$

Normalized Cross-Correlation(NCC): The NCC is defined as:

$$NCC(T,F,M) = \frac{\sum_{x \in \Omega_F} (F(x) - \bar{F})(M(T(x)) - \bar{M})}{\sqrt{\sum_{x \in \Omega_F} (F(x) - \bar{F})^2 \sum_{x \in \Omega_F} (M(T(x)) - \bar{M})^2}}$$

Minimizing this metric will also minimize the Normalized Sum of Squared Differences (NSSD, used as metric in Demons algorithm):

$$NSSD(T,F,M) = \sum_{x \in \Omega_F} \left(\frac{F(x) - \bar{F}}{\sqrt{\sum_{x \in \Omega_F} (F(x) - \bar{F})^2}} - \frac{M(T(x)) - \bar{M}}{\sqrt{\sum_{x \in \Omega_F} (M(T(x)) - \bar{M})^2}} \right)$$

B.2 Results & Discussion

B.2.1 Preliminary test

The figure B.1 show the random deformation field that we applied to our image and the two deformation field obtained by registering the original image to the deformed image with B-spline and demons methods. The Demons method was sensitive to smaller changes when the B-spline method detected larger deformations. Given the fact that the displacements that we applied to our image were quite small, the demons method gave a slightly better results (Tab. B.1).

B.2.2 Registration test

The results of our test are presented in the figures B.2, B.3 and the table B.2. It is obvious that when we used only rigid registration or demons registration, we could not

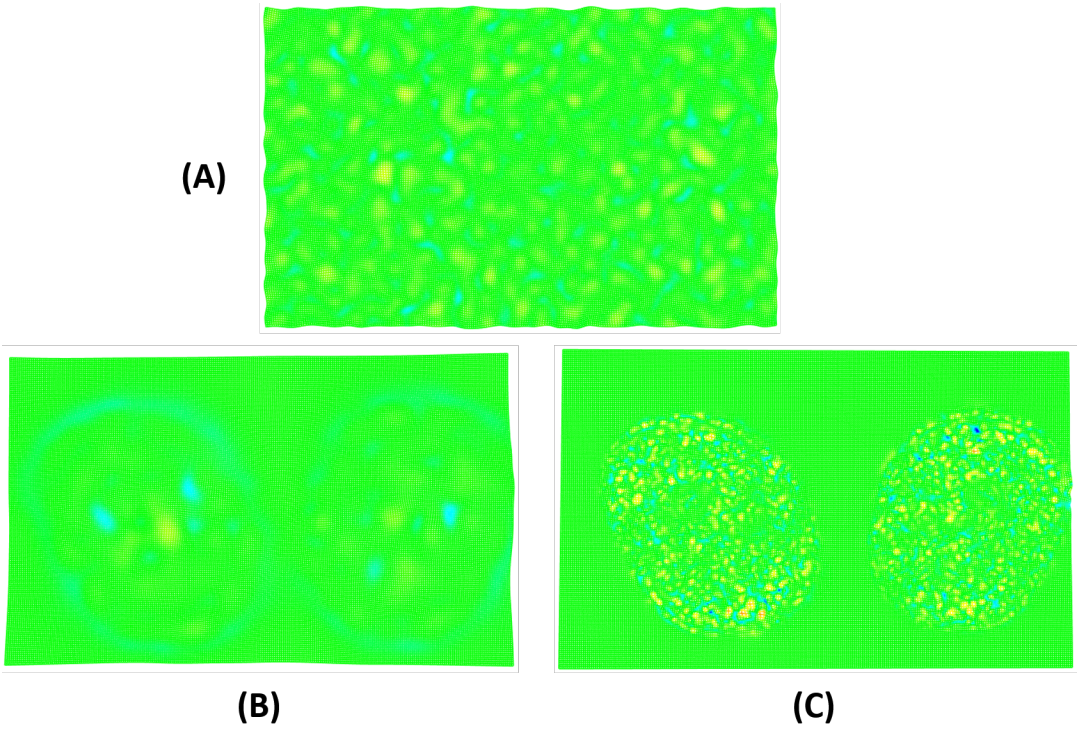

FIGURE B.1: *Deformation fields: (A) applied to the original image and obtained by registering the original image to the deformed image with (B) B-spline registration method and (C) demons registration method*

Method	MI	NMI1	NMI2	SSD	NCC
B-spline	1.50538	0.50049	1.33377	34.3851	-0.96634
Demons	**1.91519**	**0.64697**	**1.47816**	**4.51989**	**-0.99514**

TABLE B.1: *Similarity metrics computed between the fixed image and the final results of B-spline and Demons*

obtain a satisfying results. As we observed the checkboard images, we can see that rigid & demons registration gave the smoothest result while the rigid & Bspline registration still could not correct some intensity differences. The table B.2 also shows that the similarity metrics were in favor of the rigid & demons registration.

Method	MI	NMI1	NMI2	SSD	NCC
Original	0.906125	0.221014	1.12424	1873.89	-0.883636
Rigid	1.3774	0.321789	1.19174	683.436	-0.958544
Demons	1.50538	0.469704	1.30694	125.543	-0.991795
Rigid + Demons	**2.20382**	**0.527451**	**1.35819**	**68.2195**	**-0.995512**
Rigid + B-spline	1.91462	0.463898	1.302	141.023	-0.991956

TABLE B.2: *Similarity metrics computed between the fixed image and the final results of different registration methods*

However, our ultimate objective is to find an accurate segmentation. Thus, we

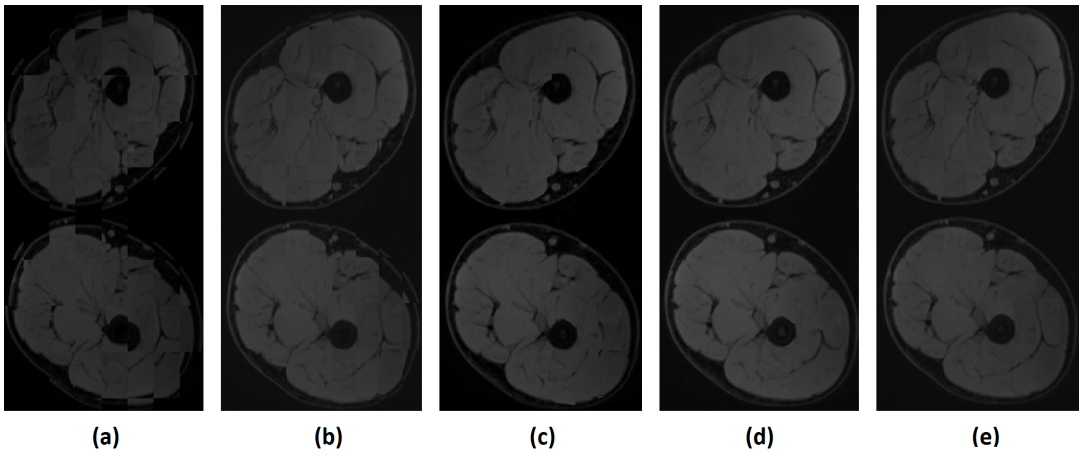

FIGURE B.2: *Checkboard image between the fixed image and (a) the moving image, (b) the result of rigid registration, (c) the result of demons registration, (d) the result of rigid & demons registration and (e) the result of rigid & B-spline registration.*

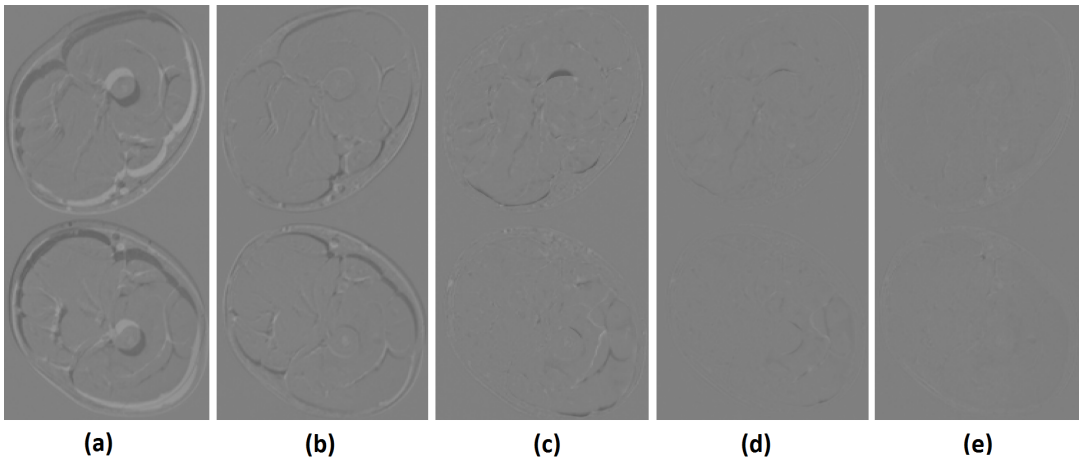

FIGURE B.3: *Subtraction image between the fixed image and (a) the moving image, (b) the result of rigid registration, (c) the result of demons registration, (d) the result of rigid & demons registration and (e) the result of rigid & B-spline registration.*

superposed the transformed label images issued from the registration methods on our fixed image and the result was not at all in favor of the demons method (Fig B.4). We could see that the label of the B-spline method was visually more accurate with smoother borders. The B-spline method conserved the smoothness of the original label image since it seemed to apply larger deformations. The demons method applied smaller deformations thus destroyed the topology of the label images, made it even less accurate than the one issued from the rigid registration. When we looked back at the figure B.3, we noticed that the subtraction of the B-spline method was more homogeneous while the one of the demons method had some small visible differences.

B.3 Perspectives

Both demons and B-spline methods (accompanied by a rigid registration) seem really efficient in the registration of MR images of the quadriceps. While the similarity metrics were in favor of the demons method, the visualization of the label images

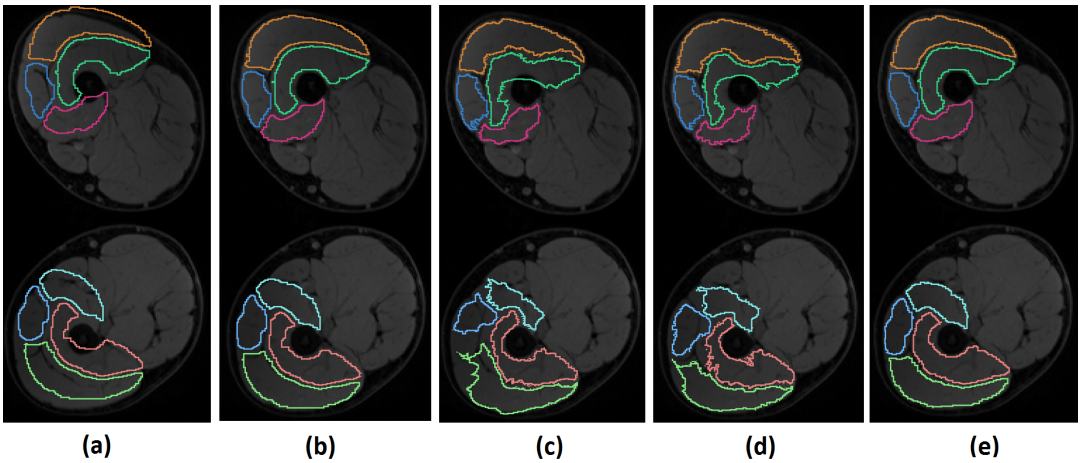

FIGURE B.4: *Label images superposed on the fixed image. The label images displayed are of (a) the moving image, (b) the result of rigid registration, (c) the result of demons registration, (d) the result of rigid & demons registration and (e) the result of rigid & B-spline registration, respectively.*

indicated that the B-spline method might be more suitable for our purpose which was to accurately segment the quadriceps muscles using atlas.

APPENDIX C

AdaBoost

AdaBoost (Adaptive Boosting) Freund and Schapire (1996) is one of the most famous *ensemble learning techniques*. It belongs to a family of methods called *boosting* which works with multiple weak learners and try to boost their performance from weak to strong. A weak learner is just slightly better than random guess (error rate is a little smaller than 50%) while a strong learner has a nearly perfect performance.

At each t iteration, a weak h_t learner is chosen to best rank (smallest ϵ_t error rate) the m samples, where each sample is weighted by $\mathcal{D}_t$. The distribution of the weights $\mathcal{D}_t$ of the m samples is updated by increasing the weights for poorly ranked samples (and vice versa for highly ranked samples). This algorithm stops after a given number of T iterations. The strong learner is formed from all the weak h_t learners obtained during the iterations. They are weighted by a function almost inversely proportional to the error rate ϵ_t.

The AdaBoost algorithm is detailed in Algorithm 1. Weak learners are often a threshold in one of the dimensions, which is simple enough to avoid adapting too much to the training data (*over-fitting*).

Algorithm 1: AdaBoost algorithm Zhou (2012)

Input: Dataset $D = \{(X_1, y_1), (X_2, y_2), ..., (X_m, y_m)\}$
Base learning algorithm $\mathcal{L}$
Number of learning rounds T

$\mathcal{D}_1(x) = 1/m$; /* Initialize the weight distribution */
for $t = 1, ..., T$ **do**
  $h_t = \mathcal{L}(D, \mathcal{D}_t)$; /* Train classifier h_t from D under distribution $\mathcal{D}_t$ */
  $\epsilon_t = P_{x \sim \mathcal{D}_t}(h_t(x) \neq f(x))$; /* Evaluate the error of h_t */
  if $\epsilon_t > 0.5$ **then break**;
  $\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$; /* Determine the weight of h_t */

$$\mathcal{D}_{t+1}(x) = \frac{\mathcal{D}_t(x)}{Z_t} \begin{cases} \exp(-\alpha_t), & \text{if } h_t(x) = f(x) \\ \exp(\alpha_t), & \text{if } h_t(x) \neq f(x) \end{cases}$$
$$= \frac{\mathcal{D}_t(x)}{\exp}(-\alpha_t f(x) h_t(x)) Z_t$$

  /* Update the distribution, where Z_t is a normalization factor which enables $\mathcal{D}_{t+1} to be a distribution$ */
end
Output: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

APPENDIX D

elastix parameters

D.1 Affine registration for MUST dataset

```
(FixedInternalImagePixelType "float")
(FixedImageDimension 3)
(MovingInternalImagePixelType "float")
(MovingImageDimension 3)


// *************** Main Components **************************
(Registration "MultiResolutionRegistration")
(FixedImagePyramid "FixedSmoothingImagePyramid")
(MovingImagePyramid "MovingSmoothingImagePyramid")
(Interpolator "BSplineInterpolator")
(Metric "AdvancedMattesMutualInformation")
(Optimizer "StandardGradientDescent")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "AffineTransform")


// ***************** Transformation **************************
(AutomaticTransformInitialization "true")
(AutomaticScalesEstimation "true")
(HowToCombineTransforms "Compose")
(HowToCombineTransforms "Compose")

// ****************** Similarity measure *********************
(NumberOfHistogramBins 64)
(FixedLimitRangeRatio 0.0)
(MovingLimitRangeRatio 0.0)
(FixedKernelBSplineOrder 1)
(MovingKernelBSplineOrder 3)

// ******************* Multiresolution ***********************
(NumberOfResolutions 4)
```

```
// ******************* Optimizer ****************************
(MaximumNumberOfIterations 1000)
(WriteTransformParametersEachIteration "false")
(WriteTransformParametersEachResolution "false")
(WriteResultImage "true")
(ShowExactMetricValue "false")
(ErodeFixedMask "false")
(ErodeMovingMask "false")
(UseDifferentiableOverlap "false")

// *************** Image sampling **********************
(NumberOfSpatialSamples 2048 2048 5000 5000)
(NewSamplesEveryIteration "true")
(ImageSampler "RandomCoordinate"

// ************* Interpolation and Resampling ***************
(BSplineInterpolationOrder 3)
(FinalBSplineInterpolationOrder 3)
(DefaultPixelValue 0)

(WriteResultImage "true")
(ResultImagePixelType "float")
(ResultImageFormat "nii")
```

D.2 Bspline registration for MUST dataset

```
(FixedInternalImagePixelType "float")
(MovingInternalImagePixelType "float")
(UseDirectionCosines "true")

// *************** Main Components **************************
(Registration "MultiResolutionRegistration")
(Interpolator "BSplineInterpolator")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")

(FixedImagePyramid "FixedRecursiveImagePyramid")
(MovingImagePyramid "MovingRecursiveImagePyramid")

(Optimizer "AdaptiveStochasticGradientDescent")
(Transform "BSplineTransform")
(Metric "AdvancedMattesMutualInformation")

// ***************** Transformation **************************
(FinalGridSpacingInVoxels 25)
(HowToCombineTransforms "Compose")

// ****************** Similarity measure *********************
(NumberOfHistogramBins 32)
(ErodeMask "false")
```

```
// ******************** Multiresolution **********************
(NumberOfResolutions 4)

// ****************** Optimizer ****************************
(MaximumNumberOfIterations 1000)

// *************** Image sampling **********************
(NumberOfSpatialSamples 5000 5000 10000 10000)
(NewSamplesEveryIteration "true")
(ImageSampler "Random")

// ************* Interpolation and Resampling ***************
(BSplineInterpolationOrder 3)
(FinalBSplineInterpolationOrder 3)
(DefaultPixelValue 0)

(WriteResultImage "true")
(ResultImagePixelType "float")
(ResultImageFormat "nii")
```

APPENDIX E

Computational ressources

saki is the DELL machine equipped with 2 GPUs GTX 1080, each with a RAM of 16 GB. Saki also includes 2 CPUs Xeon E5, each of 10 cores with 2 threads to access a RAM of 512 GB.

CREATIS Cluster: computation center reserved for researcher of CREATIS Laboratory that includes 36 heterogeneous computers with 4 GPUs Tesla V100 32GB, number of cores varying from 8 to 32 and RAM from 16 to 128GB.

IN2P3 Computing Center: CNRS (Centre National de Recherche Scientifique) cluster involving 387 machines with 17808 cores and 8 GPUs Tesla V100 32GB