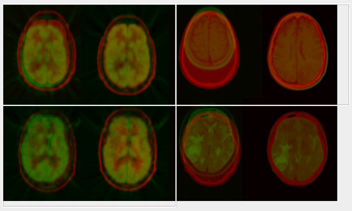
Computer Aided Medical Procedures
Prof. Dr. Nassir Navab



Dissertation

Machine Learning Methods for Computer Assisted Diagnosis and Medical Image Registration

Benjamín Gutiérrez Becker

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin

Machine Learning Methods for Computer Assisted Diagnosis and Medical Image Registration

Benjamín Gutiérrez Becker

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):	Prof. Dr. Julien Gagneur
Prüfer der Dissertation:	1. Prof. Dr. Nassir Navab 2. Prof. Dr. Martin Reuter 3. Prof. Dr. Diana Mateus

Die Dissertation wurde am 25.02.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 17.09.2019 angenommen.

Benjamín Gutiérrez Becker
Machine Learning Methods for
Computer Assisted Diagnosis and Medical Image Registration
Dissertation, Version 0.0

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 and Garching bei München

Abstract

In the last years, machine learning based approaches have seen tremendous success at solving challenging tasks in a plethora of disciplines due to their potential to generate accurate prediction models, leveraging on empirical information extracted from large amounts of data. Medical image analysis has not been the exception, and machine learning based approaches now dominate a variety of tasks including computer assisted diagnosis, image segmentation, image registration and computer assisted interventions. The success of machine learning paired with an increase on the availability of digital image records and computational power calls for an exploration of how novel machine learning approaches can be developed to address key questions in the medical image analysis field.

In this thesis, we present a number of novel machine learning based methods for diverse medical image analysis tasks. Our first contribution is a novel framework based on an age estimation model, used to detect brain abnormalities caused by neuropathologies. We propose a model which measure deviations from the mean healthy aging trajectory using uncertainty based metrics and we showcase its ability to measure brain abnormality caused by autism, mild cognitive impairment and Alzheimer's disease. In our second contribution, we tackle the problem of training machine learning models using limited amounts of labeled data. We formulate the efficient selection of a training dataset from big repositories of medical data as a multi-armed bandit problem. Our method is able to select relevant samples, based solely on meta information associated to the images leading to accurate models trained with only a fraction of the available data. Our third and final contribution is a machine learning based framework for the registration of multimodal images. Our approach models the problem of multimodal image registration not as that of learning a similarity metric between images from different modalities, but it rather aims at learning directly the transformation parameters bringing the images into spatial alignment. This approach leads to a fast and accurate multimodal registration metric, which can be easily optimized using simple gradient descent optimization.

This dissertation aims to generate discussion about the potential of machine learning techniques to solve a variety of medical image analysis tasks, and also to highlight and address some of the critical challenges which have so far limited the success of machine learning in medical image analysis when compared to other disciplines.

Zusammenfassung

In den letzten Jahren hat maschinelles Lernen einen enormen Erfolg in vielen Disziplinen erzielt. Die Verfügbarkeit großen Datenmengen und die Erhöhung der Rechenleistung haben zum Erfolg von Modellen für maschinelles Lernen beigetragen.

Die medizinische Bildanalyse ist nicht die Ausnahme, und maschinelles Lernen dominiert heute Aufgaben wie computergestützte Diagnose, Bildsegmentierung, Bildregistrierung und computergestützte Interventionen. Der Durchbruch des Maschinellen Lernens, die zunehmende Verfügbarkeit von digitalen Bildaufzeichnungen und die erhöhte Rechenleistung erfordern neue Ansätze im Bereich der medizinischen Bildanalyse.

In dieser Thesis stellen wir neuartige maschinelle Lernmethoden für die medizinische Bildanalyse vor.

Unser erster Beitrag ist eine Methode zur Beurteilung von Hirnanomalien, die durch Autismus oder Alzheimer verursacht werden. Diese Anomalien werden mit der Messunsicherheit eines Gauß-Prozesses gemessen. Wir zeigen die Vorteile unseres Ansatzes verglichen mit Standardmethoden zur Altersbestimmung. In unserem zweiten Beitrag befassen wir uns mit folgender Problematik: die Erstellung von Modellen für maschinelles Lernen mit begrenzten Mengen an vorklassifizierten Daten. Wir formulieren die effiziente Datenauswahl eines Trainingsdatensatzes aus großen Repositorien medizinischer Daten als mehrarmiges Banditenproblem.

Unsere Methode ermöglicht, relevante Beispiele auszuwählen. Diese Auswahl basiert ausschließlich auf Metainformationen, die den Bildern zugeordnet sind. Diese Methode zur Datenauswahl schafft präzise Modelle, die mit einem Bruchteil der verfügbaren Daten trainiert werden.

Unser dritter und letzter Beitrag ist ein maschinell lernbasierter Rahmen für die Registrierung multimodaler Bilder. Unsere Methode lernt unmittelbar die Transformationsparameter unter Verwendung eines Datensatzes von ausgerichteten multimodalen Bildern. Dieser Ansatz führt zu einer schnellen und präzisen multimodalen Registrierungsmetrik, die durch die Optimierung des Gradientenabfalls optimiert werden kann.

A mis padres y hermana por su interminable apoyo.
A Elise, por tu invalorable amor e inspiración.

Acknowledgments

This dissertation has been the result of a very long and winding road, which was only possible to go through with the support of many people. Thanks to Prof. Dr. Nassir Navab for giving me the opportunity to pursue my PhD at CAMP, for his patience and trust, and for being an inspiration and a role model.

I would like to thank Julien Gagneur, Martin Reuter and Diana Mateus for taking the time of being part of my PhD committee, and for the great feedback and discussion which took part during my PhD defense. In the case of Diana Mateus, I also would like to express my most sincere thanks for being a great source of support during my PhD. If the work presented in this thesis is any good at all it is only because of your feedback and for all the lessons I learned working with you during my years at CAMP. Special thanks are also due to Christian Wachinger, who gave me the opportunity to continue with my research at his lab. It was an incredible experience to be one of the founding members of AI-med.A great deal of the work presented in this dissertation has also been the result of many discussions with Loic Peter, either in front of a whiteboard, a darts board or during a tea break. Thanks for becoming not only a consistent co-author but a close friend.

My time at CAMP was marked by all the extremely talented people and great human beings I had the privilege to share my time with. In particular I would like to thank Christian Schulte zu Berge for being a co-founder of the Christkindlmarktherausforderung, Tobias Benz for setting up the CAMP kicker league, Ralf Stauder for hosting many board game events at his place, Pierre Chatelain for introducing me to Lillet, Julia Rackerseder and Nicola Rieke for keeping a great spirit at our shared office, Iro Laina and Christian Rupprecht for the food comma events, Ahmad Ahmadi for helping me dealing with the infinite paperwork required to join CAMP, Christoph Hennersperger for driving me on a bike in the streets of Beijing, Jose Gardiazabal for repairing my laptop and showing me that you can always complain about German weather, Richard Brosig for giving me my first chances to play football in Germany; to Sebastian, Abhijit and Ignacio for helping me explore all the food posibilities around Goetheplatz and many other people! Also many thanks to Martina Hilla for always finding a solution for even the most strange and difficult administrative problems I found during my PhD.

Finally thanks to CONACYT and the DAAD, which funded my PhD studies.

Contents

Introduction

1

Contents

1.1 A Brief History of Medical Image Analysis

1.1.1 Medical Imaging

Although ever present in current clinical settings, the use of medical images as a tool assisting diagnosis and medical interventions is relatively novel when seen in the larger context of the long history of medical care. The now taken for granted ability to *look* inside the human body and take images of it in a non-invasive way was only possible for the first time by the end of the XIX century after the breakthrough discovery of x-rays by Wilhelm C. Röntgen [78]. The potential of using x-rays as a new tool to visualize the human body was almost immediately recognized by the medical community which started to make significant efforts to understand how this new technology could help the diagnosis and treatment of multiple medical conditions. As a matter of fact it took less than a year for the scientific community to produce the first *medical imaging* studies, appearing as part of the journal "Archives of (Clinical) Skiagraphy". This journal mainly contained illustrations of different orthopedic cases [70], and served to summarize the first approaches that opportunities this new discovery presented to assist diagnosis and interventions.

From these early applications , it became obvious that the benefits that could be obtained from medical images were highly dependent both in the quality of the acquired images and on the ability of the radiologist of the clinician to extract relevant information from them [39]. In order to unleash the huge potential that medical images offered, the scientific community faced the challenge to exploit the informaton contained in x-ray images. These images are only a *representation* of the real anatomy of the patient, which means that in order to obtain valuable insights, clinicians need to have a clear understanding on how the information encoded in the image relates to the real anatomy of the patient. Therefore early radiological research in the medical field focused on reporting diagnostic cases [8, 11, 83] and in obtaining x-ray images of either phantoms [95] or ex-vivo organs [4]. A further impulse was given to the medical imaging field in the 1940's with the introduction of ultrasound (US) images as a further tool to look into the human body [72]. US served as a complement to x-ray images due to its ability to image soft tissue, and it gained popularity in many medical disciplines, particularly in obstetrics due to its safety and non-invasiveness.

1.1.2 Digital Medical Image Analysis

After the birth of medical imaging as a discipline and the successful first adoption of x-ray and US images as new tools to assist medical care, a major technological breakthrough in the medical imaging field arrived with the introduction of computers able to process medical images. Different to the introduction of x-rays and US, which had an almost immediate impact in medicine, the use of digital computers to process and analyze medical images took a relatively long time to mature and its utility was exploited at a slower pace. Although as early as 1955, Lusted discussed the potential use of computers in the analysis of medical images [60], most of the first attempts to use computers to perform analysis on digital images date to the beginning of the 1960's. These first attempts mainly aimed at image enhancing tasks such as background subtraction, contrast correction and filtering [7, 86], which mainly helped in producing images more suitable to be analyzed by radiologists. However, even at these early years with limited computing power and rudimentary image digitalization techniques, there existed early attempts of aiding diagnosis based on quantitative measures obtained from medical images. For example, Becker [5] describes a method to automatically measure the cardiothoracic ratio based on digitally scanned chest radiographies. By the end of the decade, the first conferences dedicated exclusively to the field of medical image processing were born; most notably the "International Conference in Information Processing in Sinctigraphy" born in 1969 , which later changed its name to "International Conference in Information Processing in Medical Imaging" and is still one of the leading conferences in the field today [96].

The use of computers to perform medical image processing gained momentum by the start of the 1970's. An important reason behind this was the rapid increase in computational power available to researchers. But the real breakthrough in the use of digital processing techniques was the development of the Computed Tomography (CT) scanner [50]. CT scanners offered the possibility to obtain 3D tomographic images of the body and were quickly developed and commercialized, with up to 400 scanners existing in the United States by the mid 1970's [75]. This led the medical imaging community to realize the high impact digital processing could have in the medical imaging field, and also lead to an increase on the available amount of digital images.

The period between 1970's and the 1990's saw an increase in the digitalization of medical images, and a group of researchers with backgrounds in computer vision started to apply their methods and algorithms to the analysis of digital medical images [96]. This led to the creation of a new sub-discipline of computer science now known as *medical image analysis*. Medical image analysis was born as an area of study which focuses on the development of computational and mathematical methods assisting the interpretation of medical images. These methods include the delineation of organs or regions of interest (segmentation), spatial alignment of two or more images representing the same anatomy (registration), computer assisted diagnosis tools, shape analysis, among several others. Thanks to the efforts of the medical image analysis community, digital medical images are now a quintessential part of routine clinical practices and medical image analysis is not only a gimmick but a fundamental tool in all stages of clinical care.

The successful adoption of medical images, and the increased availability of image databases obtained all over the world has brought new areas of opportunity to improve the medical image analysis, and with it facilitate the labor of extracting meaningful information from images. A radiologist is now able not only to obtain images of a patient using different imaging modalities (CTs, Magnetic Resonance Images (MRI), Ultrasound US,Possitron Emision Tomography (PET), etc.), but also potentially has an almost endless amount of similar images obtained in different clinics or hospitals. Needless to say, it is not only impractical but also close to impossible for even the most qualified physician to leverage on information at such scale; the task of extracting relevant information from clinical images is no longer one that can be solely be performed by direct human interaction.

1.1.3 Machine Learning for Medical Image Analysis

The discovery of x-ray and ultrasound allowed us the ability to look into the human body. Digital image processing has provided ways to improve the quality of these images, and tools to make the extraction of information from them an easier task for radiologists and clinicians. In a similar way, a recent change of the digital medical image processing paradigm has opened new possibilities to the way we interact with medical images, and how we can extract valuable information from them. An increased availability of medical images, paired with an explosion in computational resources has led to the popularity of machine learning based methods for medical image analysis. Machine learning - which we explore in chapter 2 - is the study of algorithms which are able to learn complex non-linear relationships or patterns based on empirical data [69]. Machine learning algorithms have nowadays outperformed rule-based methods in several medical image analysis applications and are nowadays the main area of research in medical image analysis. Machine learning brings the promise of accurate models which can leverage on the large amounts of available medical data and computational power available today, and which could potentiate the utility of medical image analysis tasks in routine clinical practice.

Applying machine learning to the analysis of medical images is not a trivial task due to the complexity of medical images, the large variability of anatomies between populations, the differences in scanner protocols and the relative lack of curated and manually annotated medical image datasets. Inspired by these challenges, and by recent developments in the application of machine learning in medical imaging, we present in this dissertation three contributions which constitute novel machine learning approaches to solve diverse challenges faced in typical medical image analysis tasks.

1.2 Summary of contributions

1.2.1 Gaussian Process Uncertainty in Age Estimation as a Measure of Brain Abnormality

One of the most common tasks of medical image analysis is that of creating Computer Aided Diagnosis (CADx) system. A CADx system aims at assisting the clinician in the task of summarizing image based information in order to perform a diagnosis *i.e.* defining if a patient is healthy or suffers from a disease. A recent approach to produce a CADx model is that of training age prediction models. These models are built on cohorts of healthy subjects to reflect normal aging patterns. The application of these age prediction models to diseased subjects usually results in high prediction errors, under the hypothesis that diseases follow similar patterns as those of accelerated or decelerated aging. In our contribution we propose the use of metrics based on uncertainty in a Gaussian process age regression model as a way to measure abnormalities associated to neuropathologies.

1.2.2 Guiding Multimodal Registration with Learned Optimization Updates

. Multimodal image registration is a critical task in medical image analysis which consists on the spatial alignment of images acquired using different imaging modalities. Multimodal image registration is a challenging task because the relationship between the intensities of both images is unknown a priori, and therefore defining an energy function capable of relating both modalities is difficult. Our first contribution - first presented as a conference paper at the International Conference on Medical Image Computing and Computer-assisted Intervention in 2016 [46], where it was awarded the Young Scientist Award, and then expanded as a journal paper in the special issue of Medical Image Analysis [45] - presents a novel approach where the multimodal registration problem is posed as a supervised regression task, with joint image descriptors defining the relationship between both modalities as input and the parameters of the transformation that guides the moving image towards alignment as an output.

1.2.3 A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data

The increasing availability of medical images and the emergence of machine learning techniques has brought an increasing interest in applying these techniques to medical image analysis tasks. However, one of the main difficulties these approaches encounter is that although big imaging datasets exist, the availability of annotated data is scarce. In an effort to tackle this problem, we present a method to smartly select training samples from large imaging datasets. Our approach is based on exploiting meta information associated to medical images such as phenotypic information of the patient (age, gender, weight, etc), or information about the image acquisition process (scanner, site, etc) in order to define which images are relevant to train a model for a specific task. The full text of these contributions can be found in the Appendix of this dissertation.

1.3 Thesis organization

This thesis is structured as follows. This chapter has presented a quick introduction to the medical image analysis field, and presented an overview of the main contributions of this thesis. Chapter 2 constitutes an introduction to machine learning in medical image analysis; machine learning is a broad field, and covering all possible machine learning paradigms and methods escapes the scope of this thesis. We focus instead on presenting an introduction to machine learning as a discipline, and we give an overview of the machine learning algorithms used in the contributions presented on this thesis. In Chapter 3 we present a introduction to medical image registration; we start by defining the image registration problem and we describe the main components of intensity based image registration algorithms. The last subsection of this chapter is an overview of the evolution of the medical image registration field, starting from the very first approaches to manually register images and finishing with machine learning based methods.In chapter 4, we present the conclusions of the dissertation and an outlook of the current status of machine learning in medical image analysis. The full texts of the contributions can be found in the appendix.

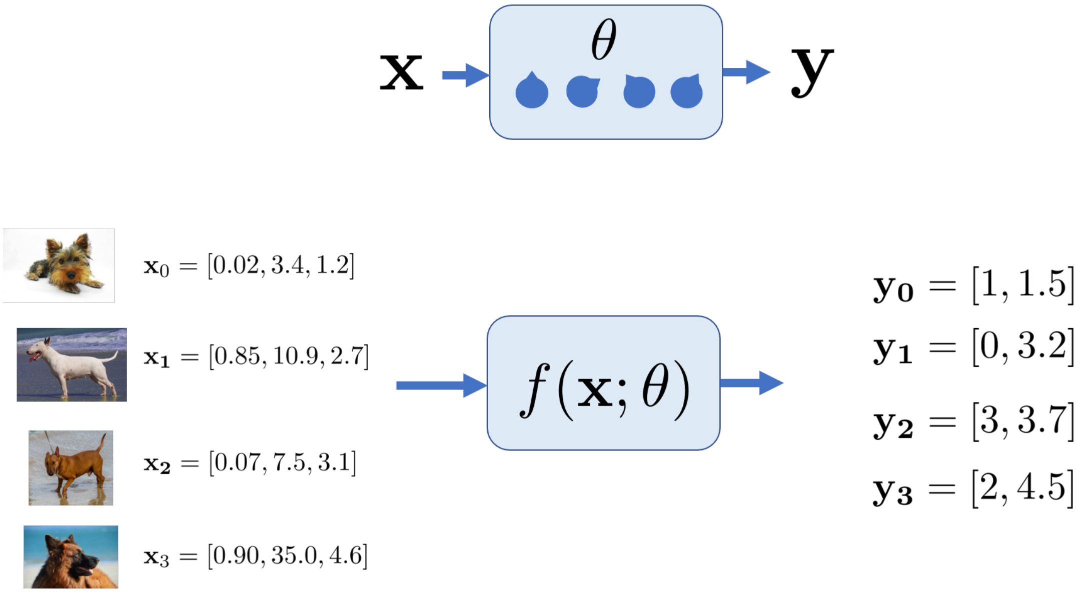
2 Machine Learning in Medical Image Analysis

Contents

2.1 Introduction to Machine Learning

The term *machine learning* first coined by Arthur Samuel in 1959 [81] can be defined as the use of statistics to give computers the ability to *learn* to perform a given task. Machine learning is different to other types of algorithms, since instead of directly coding a set of rules to solve a particular task, these rules are derived in a statistical data-driven manner. Lets walk through a simple example to illustrate the difference between these two approaches: imagine that we are given the task to design an algorithm that takes as an input a set of characteristics of a dog - for example its weight, tail length, fur color, etc. - and gives as an output a *guess* of the breed of the dog. If we approach this problem using a rule based approach we would need to manually define a set of rules which characterize each breed: for example, we could design a rule stating that a very small brown and black dog is a Yorkshire terrier, and another rule defining a big white dog as an Akita. This approach could however be problematic, since defining the rules that distinguish between different dog breeds may not straight forward or we may lack the required expertise on dogs to define clear rules. An alternative approach would be to design and algorithm which *learns* these rules directly from the data: this would involve creating a dataset of measurements of dogs, and extracting meaningful statistics from them. With this statistical approach our algorithm may learn *by itself* that a Yorkshire's mean weight is 1.5 kg and is 80% of the time brown, while the mean weight of an Akita is 40 kg and is always white. This statistical information can afterwards be used to define meaningful rules to make a decision.

Machine learning, as defined by Tom Mitchell [69] is the ability of a computer program to learn from experience E to perform a particular task T according to a performance measure P. Following our example, a computer can learn the task T of identifying the breed of a dog through the experience E of *observing* a dataset of measurements associated to different dog breed. A measure P would indicate how accurate is the prediction of our algorithm when predicting the breed of a dog given its characteristics. Depending on the amount of interaction between the algorithm learning to perform a task and the user, machine learning techniques can be classified between *supervised* and *unsupervised* methods.



$\mathbf{x}$
θ
$\mathbf{y}$
$\mathbf{x}_0 = [0.02, 3.4, 1.2]$
$\mathbf{x_1} = [0.85, 10.9, 2.7]$
$\mathbf{x_2} = [0.07, 7.5, 3.1]$
$\mathbf{x}_3 = [0.90, 35.0, 4.6]$
$f(\mathbf{x}; \theta)$
$\mathbf{y_0} = [1, 1.5]$
$\mathbf{y_1} = [0, 3.2]$
$\mathbf{y_2} = [3, 3.7]$
$\mathbf{y_3} = [2, 4.5]$

Figure 2.1 Supervised learning can be seen as the problem of finding a function $f(\mathbf{x})$ mapping a feature vector $\mathbf{x}$ to a label $\mathbf{y}$. The function $f(\mathbf{x})$ is usually defined using a vector of parameters $\boldsymbol{\theta}$ which act as a set of *knobs* adjusting the behavior of the function f. For example, a predictor function $f(\mathbf{x})$can map a set of characteristics of a dog to a vector of labels indicating its breed and age.

2.1.1 Supervised Learning

Supervised learning algorithms are those where an algorithm is trained to perform a task based on experience extracted from previous observations given by the algorithm in the form of a training set. This training set has the form $\{\mathbf{X}^i, \mathbf{Y}^i\}_{i=1}^n$, where $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a matrix containing N different *observations*, and $\mathbf{Y} \in \mathbb{R}^{p \times n}$ is a matrix of *labels* associated to each one of these observations. The numbers m and n correspond to the size of each feature vector $\mathbf{x}$ and n is the number of observations.

Supervised learning algorithms can be formalized as the problem of finding the mapping:

$$\hat{\mathbf{y}} = f(\mathbf{x}) \tag{2.1}$$

where $\hat{\mathbf{y}}$ corresponds to a prediction and $\mathbf{x}$ is a feature vector. The operator f can be seen then as a machine that is able to process input $\mathbf{x}$ in order to give a prediction. The response of this machine can be changed by adjusting a set of "knobs", which adjust the behaviour of the machine. These knobs are the model parameters $\boldsymbol{\theta} \in \mathbb{R}$. The problem of supervised machine learning then can be reduced to that of finding the optimal values for the parameters $\boldsymbol{\theta}$ such that we obtain accurate values of $\mathbf{y}$, according to a performance metric.

Lets illustrate this by continuing with our simple example 2.1. Given the task of predicting the breed of a dog given a set of characteristics describing it such as its height, weight, snout size, legs length, color and age. An observation $\mathbf{x}$ would correspond to a vector containing these values for a particular dog, and the label vector $\mathbf{y}$ corresponds to the breed and age of the dog. We can build a training dataset by acquiring many of these observations *i.e.* dog characteristics paired with their age and breed. During training, a supervised learning algorithm would adjust its parameters $\boldsymbol{\theta}$ by optimizing an objective function, which in this case would be a measure of the proportion of times the algorithm predicts the breed of the dogs in the training set accurately and the mean squared error of the age prediction. Depending on the nature of the prediction, supervised learning algorithms are divided in *classification*, when the predicted value $\mathbf{y}$ corresponds to a categorical value (*i.e.* the breed), and *regression* when the prediction corresponds to a real number $\mathbf{y} \in \mathbb{R}$ (*i.e.* the age).

2.1.2 Unsupervised Learning

Different to supervised learning, in unsupervised learning our training set lacks any sort of labels $\mathbf{y}$. This means that unsupervised learning algorithms need to acquire information about the structure of the data without any intervention or guidance from the user. The most common tasks in unsupervised learning are finding the probability distribution of the feature vectors $P(\mathbf{X})$, or clustering which corresponds to finding homogeneous groups of data that have similar features $\mathbf{x}$. In this thesis we limit the discussion to *only* supervised learning algorithms.

2.2 Machine Learning algorithms

2.2.1 Linear regression

Linear regression is arguably the simplest machine learning algorithm to perform regression tasks. In linear regression the mapping function f corresponds to a simple linear combination of the elements of the feature vector $\mathbf{x}$:

$$\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} + \boldsymbol{\theta}_0. \tag{2.2}$$

in linear regression the model parameters θ_i correspond to a weigh for each feature x_i and an intercept term θ_0. In order to find the vector of parameters $\boldsymbol{\theta}^T$ which best fits the training data $\mathbf{X}, \mathbf{Y}$, we can minimize the mean squared error loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2. \tag{2.3}$$

the parameters $\boldsymbol{\theta}$ minimizing $\mathcal{L}$ can be easily obtained by using the closed form:

$$\boldsymbol{\theta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.4}$$

An important property of linear regression models that makes them popular in many application is that the learned parameters θ have a meaning directly related to each one of the features θ and are therefore easily interpretable. As an example, we show a regression model based on the popular boston dataset [1]. This dataset contains information concerning housing in the area of Boston. In figure 2.2 we show a linear regression model fitted to predict a variable y_{val} corresponding to the median value of homes in \$ 1000's, and which takes as features x_{rooms} which indicates the number of rooms per dwelling and x_{crime} which measures the per capita crime rate of the area. By fitting a linear regression model using equation 2.4 we obtain parameters $\theta_{rooms} = 3.80$ and $\theta_{crime} = -0.10$. This means that adding an extra room to the house increases a value by 3800 and that an 0.10 increase per capita crime decreases the house value by 100. Due to their fast training and their easy interpretability, we use linear regression to build age prediction models as part of our contribution "A Multi-armed Bandit to Smartly Select a Training Set from Big Medical Data".

Although linear regression offers a simple and easy to interpret machine learning model, relationships between feature vectors $\mathbf{x}$ and labels $\mathbf{y}$ are often non-linear. This means that linear regression can be insufficient to build accurate machine learning models, and therefore several strategies have been devised to build non-linear models.

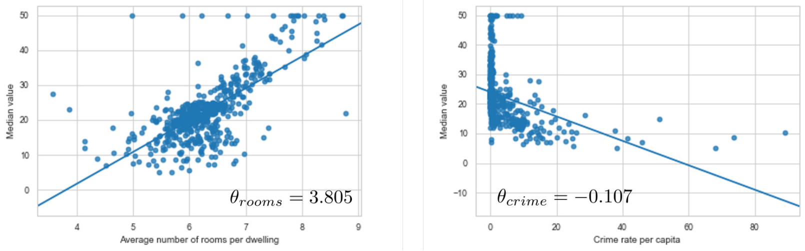


Figure 2.2 A linear regression model fitted on the boston dataset. The label y corresponds to the mean price of properties, while the feature vector $\mathbf{x}$ contains the average number of rooms per dwelling. The parameters θ correspond to the weight each one of these elements has on predicting the outcome variable.

[1] https://www.cs.toronto.edu/delve/data/boston/bostonDetail.html

2.2.2 Decision trees

Decision trees are a machine learning algorithm based on partitioning the feature space in a hierarchical manner. The structure of a decision tree is a a graph consisting of nodes and edges. In the case of nodes they can correspond to either internal nodes, or terminal nodes. The graph structure of the tree has two important characteristics: 1) in a decision tree there are no loops and 2) each node in the tree has exactly one incoming edge. In binary decision trees, which are the most commonly used and the ones presented in this thesis each node has two outcoming edges.

A decision tree can make a prediction by *passing* a feature vector through the tree structure. Starting from the root node, the decision tree algorithm makes a test at each internal node. Each one of these tests are defined by a split function g of the form:

$$g(\mathbf{x}, h, T) = x_h \geq T \tag{2.5}$$

where $\mathbf{x}$ is a feature vector, $h \in \mathcal{N}$ is a position of the feature vector and $T \in \mathbb{R}$ is a threshold. If the result of the split function is 1, the feature vector will be passed to the right out-coming edge; if the condition is false, the feature vector is passed to the left out-coming edge. This operation is repeated until the feature vector reaches a terminal node. When performing classification or regression tasks, each terminal node has assigned a particular label y. A prediction is made by assigning the label of the terminal node to the feature vector reaching it.

We show an example of a decision tree in figure 2.3. Following our initial example, we want to predict the breed of a dog using a decision tree. A decision tree would *ask* a series of questions based on the attributes of each dog, corresponding to the split functions g. Starting at root node 0 we would first evaluate the split function $g_0 = x_1 > 25$ where x_1 is an element of the feature vector $\mathbf{x}$ corresponding to the weight of the dog. Since the dog weights 35 kg, the feature vector $\mathbf{x}$ is passed to the right edge and arrives to internal node 4. Node 4 has assigned the split function $g_4 = x_0 > 0.5$ where x_0 is a feature measuring the proportion of the body of the dog that is white. Again the condition is true, and therefore the feature vector would be passed to the right edge. After this, we would arrive to a terminal node. This means that no further split function will be made, and the dog will be assigned label $y = 4$, which corresponds to the breed akita.

The most important factor in the design of a decision tree is the definition of the split functions g. In this simple scenario, we could achieve a good result by simply engineering split functions based on our previous knowledge about dog breeds. However, a machine learning approach would involve finding this split functions based on observations of the training data.

The procedure to define these split functions - or *train* a decision tree - is to pass grouped training vectors $\mathbf{x}$ together with their ground truth labels y. Borrowing the notation by Criminisi *et al.* [24], the subset of data points arriving a node is denoted as S_a. Each decision node splits subset S_a assigning some of its elements to the right S_r or the left S_l subset. At

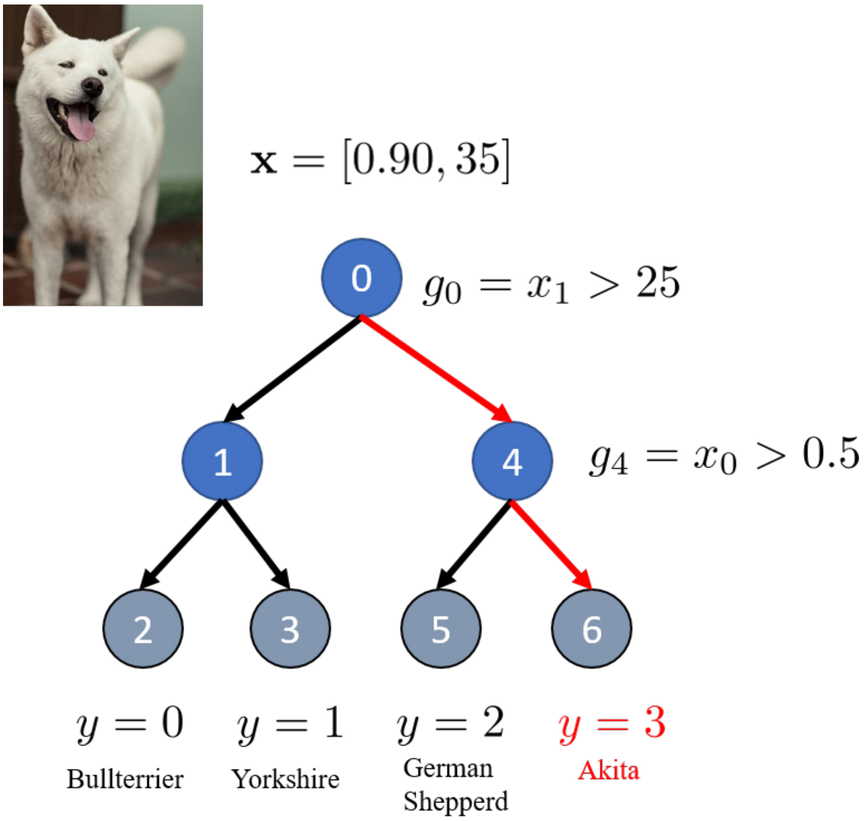


Figure 2.3 A decision tree is an algorithm which performs predictions based on a sequence of hierarchical split functions. In this example a decision tree determines the breed y of a dog based on a feature vector $\mathbf{x}$ describing its amount of white fur and its weight.

each node, a decision function g as that maximizing the energy $I = I(S_r, S_l, S_a, g)$. This means that the optimal parameters $\theta = [h, T]$ at each decision node are obtained by doing:

$$\boldsymbol{\theta}^* = [h, T] = \arg\max_{\boldsymbol{\theta}}(I) \tag{2.6}$$

During training, the decision tree is grown and decision functions are optimized iteratively splitting the training data until a stopping criterion is reached. Typical stopping criteria are reaching a predetermined tree depth or a minimum number of training examples reaching a node. Obtaining the optimal features $\theta = [h, T]$ at each node can be costly, particular in situations where training is done on large dimensional vectors $\mathbf{x}$. A typical strategy to mitigate the cost of this computation is to perform randomized node optimization [24]. Randomized node optimization consists on making available only a subset $\theta_j \in \theta$ of the parameters at at each node j. This strategy not only reduces the training time of decision trees, but also introduces a random element during training which is useful in the training of random forests (more on this in the following subsection where we discuss ensembles of decision trees).

A decision tree can be used to perform different tasks, including regression, classification or even unsupervised tasks like density estimation. The decision tree structure can remain unchanged for each one of this tasks, and the only thing that needs to be altered is the choice of the energy function I. In the case of classification tasks a common choice is *information gain* defined as:

$$I_{IG} = H(S_a) - w_r H(S_r) - w_l H(S_l) \tag{2.7}$$

where $H(S) = \sum_{k=1}^{K} P(y_k|S) log(P(y_k|S))$ is a measure of the entropy of the set S and $w_l = |S_l|/|S_a|$ and $w_r = |S_r|/|S_a|$. For regression tasks, minimization of the trace of the sample covariance is a typical choice:

$$I_{SC} = tr\mathbf{\Sigma}_{S_l} + tr\mathbf{\Sigma}_{S_r} \tag{2.8}$$

where $\mathbf{\Sigma}_{S_l}$ and $\mathbf{\Sigma}_{S_r}$ are the covariance matrices of the labels $\mathbf{Y}$ assigned to the left and right edge respectively.

An important characteristic of decision trees is that different to other common machine learning algorithms such as linear models or support vector machines, they do not require the computation of all possible features $x_h \in \mathbf{x}$ in advance. This makes them particularly useful in situations where feature spaces are of very high dimensionality, or when the computation of features is costly. Thanks to this property, we make use of decision tress in our multimodal registration approach, where infinitely dimensional feature vectors describing images to be registered are processed using decision trees.

Ensembles of decision trees

Decision trees are seldom used on their own to perform predictions in machine learning scenarios. Instead they are usually combined through the use of an ensemble method. Ensemble methods are not machine learning algorithms by themselves, but rather a procedure to combine the predictions of several methods in order to obtain a more accurate predictor or one less prone to overfitting. Two ensemble methods to aggregate predictions made by individual trees are random forests and gradient boosting.

Random forests are based on the observation that several decorrelated trees provides better generalization than a prediction made by a single decision tree [36]. Decorrelated trees are obtained by either randomly feeding subsets of training data to each tree and/or by performing randomized node optimization of the objective function at each node as we discussed previously. During testing, a single prediction $\hat{y}$ obtained by averaging the individual predictions $\hat{y}_t$ of each decision tree :

$$\hat{y} = \frac{1}{T} \sum_{t \in \mathcal{T}} \hat{y}_t \tag{2.9}$$

A second approach to aggregate predictions from individual decision trees is gradient boosting. Gradient Boosted Trees (GBT) have shown to have lower prediction errors to random forests in a variety of scenarios [12]. Different to random forests where each decision tree is trained independently, gradient boosted trees are trained sequentially. The main concept behind GBTs is that each sequentially trained weak predictor tries to minimize the error of the previously trained tree.

The prediction $\hat{y}$ in a GBT predictor is obtained via the weighted sum of functions:

$$\hat{y} = \sum_{t \in \mathcal{T}} \beta \hat{y}_t), \tag{2.10}$$

where β is a scalar weighting each regression tree. Different to regression forests, where decision trees are trained independently, in gradient boosted trees the prediction is built sequentially as:

$$\hat{y} = \hat{y}_{t-1} + \beta(\hat{y}_t). \tag{2.11}$$

where each tree minimizes a loss between the current prediction and the ground truth.

2.2.3 Gaussian processes

Gaussian processes (GP) are defined as *a collection of random variables, any finite number of which have a joint Gaussian distribution*[76]. Different to the machine learning algorithms we have previously described, GPs are *non parametric* models. What this means is that instead of having a fix number **theta** of parameters to optimize, GPs make predictions by directly map the training set to the observations for which we intend to make predictions.

Gaussian processes are therefore defined not by a set of fixed of parameters but by the following two elements: a mean function $m(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[f(\mathbf{x}) - m(\mathbf{x})f(\mathbf{x}') - m(\mathbf{x}')]$, where $\mathbf{x}$ and $\mathbf{x}'$ are two feature vectors. For simplicity, we will assume the mean function to be zero. Therefore a GP is solely defined by its covariance function k. The covariance function can be seen as a way to measure the similarity between two different data points. A common choice for the covariance function is the squared exponential function of the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \delta_f^2 \sum_{k=1}^{K} \exp\left[\frac{-(x_i^k - x_j^k)^2}{2l_k^2}\right] + \sigma_n^2 \delta(\mathbf{x}_i, \mathbf{x}_j), \tag{2.12}$$

When training a GP model, we aim at defining this covariance function. Vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ simply correspond to the feature vectors in the training set. The rest of the parameters allow GP models to include prior information about the expected behaviour of the predictive function. The maximum allowable covariance δ_f^2 controls the range of expected variance in the labels y. The variance δ_n^2 is the expected noise in the observations $\mathbf{x}$; the parameter l is known as *length scale*. Intuitively, we want the prediction of a test point $\mathbf{x}'$ to be influenced only by those examples which have similar feature vectors, and we want very dissimilar observations to have a negligible influence on the prediction. Therefore the length scale parameter controls how much influence a sample with feature vector $\mathbf{x}$ has in the prediction of $\mathbf{x}'$. Additionally, the smaller an element l_k is, the more dependent the label y is to the feature element x_k.

Given the covariance matrix, we can model the joint distribution between the training and testing data as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}' \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) & K(\mathbf{X},\mathbf{X}') \\ K(\mathbf{X}',\mathbf{X}) & K(\mathbf{X}',\mathbf{X}') \end{bmatrix}\right). \tag{2.13}$$

The elements of the joint distribution in Eq.(2.13) can be summarized as follows:

- an intra-covariance matrix of the training set $K(\mathbf{X},\mathbf{X}) \in \mathbb{R}^{mxm}$,
- an intra-covariance matrix of the testing set $K(\mathbf{X}',\mathbf{X}') \in \mathbb{R}^{nxn}$,
- an inter-covariance matrix between the training and testing set $K(\mathbf{X},\mathbf{X}') \in \mathbb{R}^{mxn}$,
- a training labels vector $\mathbf{y} \in \mathbb{R}^{m}$, and
- a testing labels vector $\mathbf{y}' \in \mathbb{R}^{n}$,

where m corresponds to the number of training samples and n to the number of testing samples. Using this conditional distribution, a GP model can perform predictions on unseen data by:

$$\hat{\mathbf{y}'} = \mathbb{E}[\hat{\mathbf{y}'}|\mathbf{X},\hat{\mathbf{y}},\mathbf{X}'] = K(\mathbf{X}',\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}\mathbf{y} \tag{2.14}$$

and

$$\mathbf{cov}(\hat{\mathbf{y}}) = K(\mathbf{X}',\mathbf{X}') - K(\mathbf{X}',\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}K(\mathbf{X},\mathbf{X}'),, \tag{2.15}$$

which correspond to the estimation of the labels and an estimated covariance. The estimated covariance can also be seen as a measure of uncertainty of the predicted values. This property of Gaussian process models to provide uncertainty measurements is the core of "Gaussian process uncertainty in age estimation as a measure of brain abnormality" where uncertainty of Gaussian process regression for age estimation is used to detect brain abnormalities caused by neuropathologies.

2.3 Machine Learning in Medical Imaging

Not surprisingly, applying machine learning methods to medical imaging analysis problems is a much more challenging task when compared to the simple examples we have analyzed in our introduction to machine learning algorithms. First, medical images are more scarcely available when compared to other types of data like natural images in computer vision or pieces of text for natural language processing. One reason behind this is the sensitive nature of medical data, and therefore it is in general not possible to make data publicly available without the informed consent of the patient, anonymization and previous ethical approval. Additionally, acquiring medical images is a costly process which involves expensive equipment and the involvement of highly qualified personal. For some particular imaging modalities such as CT, a risk to the patient health is also involved. Second, manual annotations of medical images required for supervised learning methods are hard to get due to the tedious and complicated nature of the task, and crowd sourcing annotations common in computer vision applications are not feasible due to the training required to properly annotate medical images. Third, the correct interpretation of medical images involves not only the observation of the images themselves, but should also integrate knowledge about the patient such as their demographic information, phenotype, symptoms, diagnosis or even other complex sources of information such as genetic data. Finally, compared to computer vision applications, in many cases medical images are not 2-dimensional but rather present a tomographic 3-D representation of the anatomy. This increased dimensionality make most common image processing algorithms to require a considerably larger amount of computational resources.

Despite these challenges, machine learning has found its place in most medical imaging analysis applications and in recent years, machine learning methods have dominated the research in the field. There are two key factors which have proved critical in this increase of popularity: the first one is the rise of deep learning as a tool which has proven to outperform most traditional image analysis approaches in a variety of tasks, and the second one is an increased effort by the scientific community to make available large collections of medical data. This is particularly true in the case of neuroimaging, where many publicly available datasets are now available such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [53], the Open Access Series of Imaging Studies (OASIS) [64], the Autism Brain Imaging Data Exchange (ABIDE) [25], among others.

In recent years, machine learning has had a notable influence in several medical image analysis applications. In the following section we will present an introduction of the applications covered by the contributions of this thesis: computer assisted diagnosis and image registration.

2.3.1 Computer assisted diagnosis

Diagnosis can be understood as the identification of an illness or other medical problem by examination of the symptoms. Before the introduction of medical images, it was only possible to observe these *symptoms* by simple observation, physically examination of the patient or through simple interviews to the patient (does it hurt? are you feeling dizzy?). The main objective of these observations is to observe a pattern in the body function of the patient that deviates from the *normal* function of the healthy population. Medical images opened the possibility to observe these abnormalities by opening a window to observe directly into the patients bodies, allowing them to compare the appearance of the healthy population against that of patients diagnosed with a particular disease.

Performing these comparisons between images obtained from healthy or diseased populations is extremely challenging and highly dependent on the experience level of the person doing the observation and the amount of time allowed to perform a diagnosis. Computer Aided Diagnosis (CADx) systems, aim at assisting clinicians in this task by summarizing large amounts of imaging data paired with corresponding meta information of the patient such as their phenotype (sex, gender, age), their environmental conditions or even genetic information. As mentioned in the review of CADx tools in medical imaging by Doi [26] CADx diagnosis was already attempted with little success as early as the 1960's [59, 67], but CADx systems only start to gain steam after the usefulness of CADx systems was demonstrated for diagnosis in chest radiographies in the early 2000's [1].

A typical CADx system is based on the elements shown in figure 2.4. A medical image is acquired and regions of interest are segmented. This segmentation can either correspond to a delineation of a particular organ, to a structure of interest or to an abnormality in the scan (tumor, micro-calcifications, etc.). Given these segmentations and the acquired image, a set of features can be extracted such as morphological features of the structure of interest [91] or texture descriptors [65]. The extracted features can be summarized in a feature matrix $\mathbf{X}$ which can then be analyzed using machine learning methods as the ones described in this chapter.

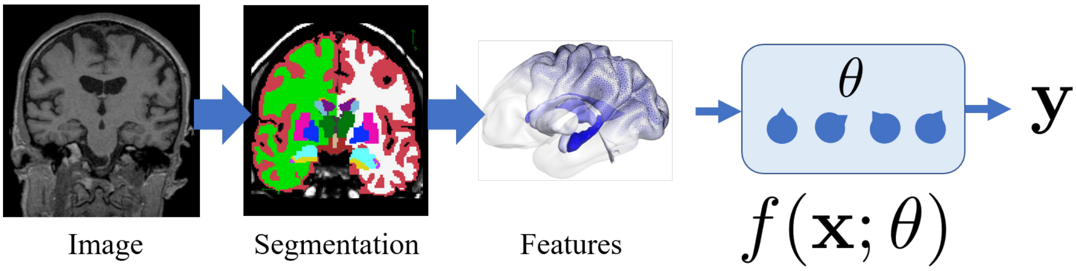


Figure 2.4 A typical Computer Aided Diagnosis (CADx) pipeline consists of 1) the acquisition of an image 2) the segmentation of structures of interest 3) extraction of features given the segmentation and 4) an analysis of the generated features.

2.3.2 Computer assisted diagnosis models based on age estimation

Most CADx methods based on machine learning models discussed so far are based on either a classification model which aims at predicting a discrete label corresponding to different diagnosis or in a regression model which gives an estimate of a particular variable that can be associated to a diagnosis. For example, several classification algorithms have been applied to the task of discriminating between structural MRI images of the brain corresponding to healthy controls (HC), individuals suffering from Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD) [27, 32, 66] while other approaches aim at estimating continuous clinical variables associated to the AD pathology [29, 94, 102].

One limitation of these discriminative models is that they require the acquisition of images corresponding to subjects at different stages of the disease as well as healthy controls. As we have mentioned earlier in this chapter, the acquisition and labelling of these images poses important challenges, and for this reason images corresponding to individuals affected by a particular disease are typically scarce, increasing the risk of overfitting.

An alternative to this direct discriminative models has been proposed in the form of age estimation models, first introduced as a method to measure abnormalities in brain development caused by the onset of Alzheimer's disease [34]. The main assumption behind age estimation models is that disease patterns associated to several diseases are actually similar to an accelerated aging pattern. This would mean, for example, that the brain of a 60 year old individual suffering of Alzheimer's disease would be similar to that of a healthy 80 year old person. A big advantage of age estimation based models is that they require *only* images obtained from healthy individuals to be trained. This means that on the one hand images are easier to get by and that the same trained model could potentially be used to measure abnormal development caused by different diseases, even when they were not included in the training set. CADx methods based on the age estimation model paradigm have shown to have the ability to measure brain abnormalities caused not only by Alzheimer's disease but also bipolar disorder [71],diabetes mellitus [35], schizophrenia, depression [56] among others.

An overview of an age estimation based CADx model used to predict the onset of Alzheimer's disease is shown in figure 2.5. First, a prediction model corresponding to the function $y = f(\mathbf{x})$ is trained. The label y corresponds to the chronological age of the subject and $\mathbf{x}$ is a feature vector extracted from an image. All feature vectors $\mathbf{x}$ are obtained from subjects which are deemed healthy *i.e.* have not been diagnosed with a particular condition. Once the age prediction model is trained it can be used to predict an estimation $\hat{y}$ of the age of previously unseen subjects. Since the chronological age of the subjects is usually known, we can easily obtain the prediction error $\epsilon = y - \hat{y}$, corresponding to the difference between the real age of the subject and the one estimated by the model. Following the assumption that disease processes are similar to that of accelerated aging, an age estimation based model uses the prediction error ϵ a proxy measure of brain abnormality. Higher values of ϵ indicate an accelerated aging process different to the typical pattern of healthy aging.

Although this approach of modeling pathologies as accelerated or decelerated aging has gained increased popularity due to its simplicity, it unfortunately does not capture accurately the underlying mechanisms relating ageing and pathologies [98]. The reason behind this is that although changes of morphology caused by disease and ageing are partially overlapping, they are distinct and affect different brain areas at different ratios [32]. In "Gaussian process uncertainty in age estimation as a measure of brain abnormality" we present an alternative approach to detect abnormalities based on age regression models. Instead of using prediction error as a measure of abnormality, we propose to use of prediction uncertainty. This alternative approach avoids making the strong assumption that accelerated aging and disease are equivalent processes, and leads to more accurate separation between scans obtained from healthy individuals and diseased subjects.

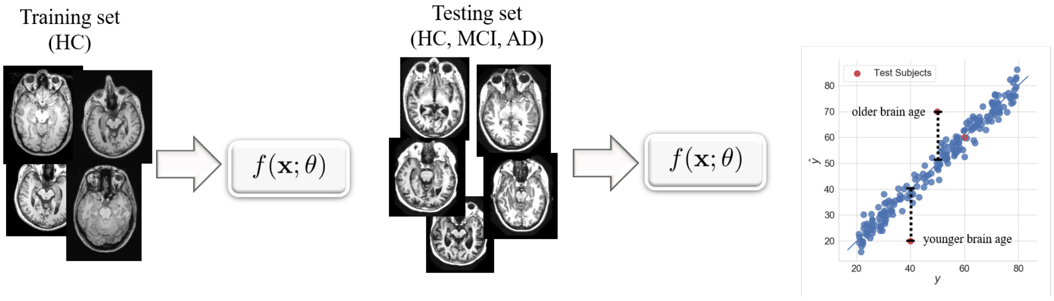


Figure 2.5 An overview of a CADx pipeline based on the age estimation paradigm. A prediction model $f(\mathbf{x};\,`)$ is trained using *only* images of healthy subjects. During testing, disease prediction is performed by using age prediction error as a measure of accelerated aging.

2.3.3 Learning from limited label data

The digitalization of medical images and the ever increasing willingness of research institutions and health centers to share data has dramatically increased the availability of medical data in the last years, the implementation of machine learning algorithms in CADx systems is still constrained by the small availability of annotated data. Labeled medical images are particularly scarce because they require manual annotations performed by highly trained radiologist or clinicians.

The problem of learning from limited amounts of labeled data has been explored using different paradigms. As we discussed on section 2.1.2, unsupervised learning algorithms build prediction models without the use of annotated data. Unsupervised models however lack a mechanism to introduce limited amounts of training data when available. Weakly supervised learning covers methods that train models feeding them with incomplete, inexact or inaccurate labels [103]. Active learning eases the annotation of data by implementing algorithms which recommend to the manual annotator which instances of the training data are more relevant for the problem at hand [20]. We will focus our discussion on learning from limited amount of data to the active learning paradigm, since it is the one closely related to the contributions proposed in this dissertation.

Several active learning approaches have been proposed to assist medical image analysis tasks. One of the most common applications of active learning in the field has been that of interactive segmentation frameworks such as the one by Wang *et al.* [93] to perform segmentation of the placenta in fetal MRI, or [89] where active learning is used to assist interactive 3D segmentation of CT images. Active learning approaches have also been applied as a way to better train data hungry deep learning approaches. Cicek *et al.* present an approach to train a model to perform 3D segmentations based on 2D sparse segmentations [18], Yan *et al.* present an active learning approach to discover local regions in an image which are highly discriminative [99] and [92].

The problem of selecting which samples have to be annotated in an active learning framework is a challenging one. One of the main reasons for this is that in order to select an appropriate set of samples to be annotated we have to balance two contradicting properties of the training set. On one side we want the training samples to be annotated to be *general*, meaning that we want to have a diverse set of samples that can cover as much as variation as possible so that the trained model generalizes to unseen data. On the other hand we have a limited budget of samples that can be annotated, and therefore we would like to annotate those which improve the most the accuracy of our model. These contradictive conditions are known as an exploration/exploitation dilemma, where we aim to have an optimal balance between exploring the space of possible solutions and exploiting those which are more rewarding. Solving an exploration/exploitation problem constitutes an active area of research [6].

Multi-armed bandit. One of the most known instances of the exploration/exploitation dilemma is the multi-armed bandit problem. The multi-armed bandit problem owns its name to an hypothetical situation, where a gambler arrives to a casino with a limited amount of playing tokens and wants to spend them playing in some slot machines (also known as one-armed bandits). Thanks to an insider tip from a friend working at the casino, he knows that each slot machine gives rewards at a different rate: this means that while some machines rarely give a prize, there are others which give prices more often. Naturally, he now wants to play with the machine that gives rewards more often, but unfortunately for him, he has no idea of which machine is the one that gives the higher rewards. This is a typical exploration/exploitation dilemma where he has a limited amount of resources (his playing tokens), that he has to use to explore which machines give more rewards while at the same time playing as often as possible on the machines that give the maximum reward.

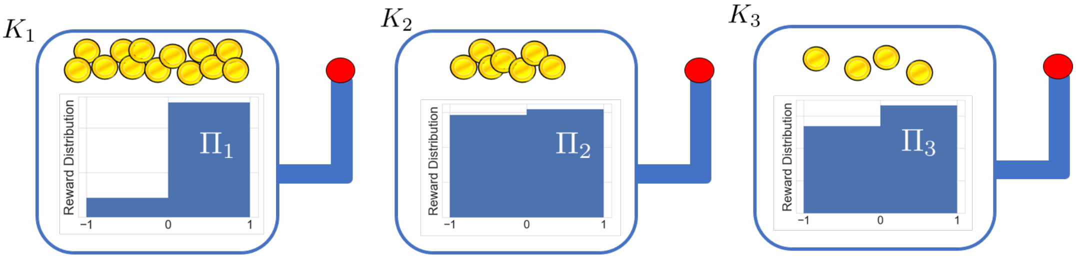


Figure 2.6 Illustration of the Multi-armed bandit problem. A gambler plays different slot machines K_i each one with a fixed reward distribution Π_i. The goal of the gambler is to play the machines in order to maximize his earnings, by exploring which machines are the ones that give larger rewards and exploiting this information to play those machines with higher reward distributions.

The multi-armed bandit problem (see Fig. 2.6) can be modeled as follows: we have a set of arms $\mathcal{K} = \{1, ..., k\}$, each one with a fixed reward distribution $\Pi_1, ..., \Pi_k$ and mean expected rewards $\mu_1, ..., \mu_k$. Since the gambler does not know which ones are the machines that give more money, these reward distributions are unknown to him *a priori*. Every time t, the gambler plays an arm an obtains a reward r_t^k according to distribution Π_k. In a simple scenario where rewards are binary - the slot machines either give a fix amount of money, or they do not give any return at all - the reward distribution can be modeled using a Bernoulli distribution where the arm gives a reward 1 with probability p or -1 otherwise.

An algorithm that aims at solving the multi-armed bandit problem has to make a decision at each time step t on which arm to play. Some naive strategies would be to play any arm K with equal probability at each time t, or to pick an arm K and play it all the time. However, smarter strategies can be designed if we take into account that every time an arm K_i is played, the gambler obtains some information about the hidden distributions Π_i. Intuitively, the gambler would like to keep playing with those machines which tend to give more money, and spend fewer tokens on those machines that do not give rewards. Therefore the decision on which arm K to play can also be designed based on the previous experience obtained from the $t-1$ plays.

Every time the glamber plays arm K_i, he obtains a reward $r(t)$ Π_k. and the objective of the glamber is to minimize a regret function

$$\rho(t,k) = T\mu^* - \Sigma_{t=1}^{T} r_t^k \tag{2.16}$$

which is defined as the expected difference between the reward sum associated with an optimal strategy (always playing the bandit with the higher expected reward μ^*) and the sum of the collected rewards. The goal of the gambler should consequently be to minimize this regret function.

Several strategies have been proposed to minimize the regret function including the ϵ-greedy algorithm [13], the Upper Confidence Bounds (UCB) family of algorithms [2] and Thompson Sampling [88]. The latter has shown to be a very effective heuristic to approach the multi-armed bandit problem due to its easiness of implementation and lack of tuning parameters [14].

Thompson sampling. The reasoning behind the Thompson sampling algorithm for binary rewards is to keep track of the acquired knowledge about the reward distributions Π by maintaining a prior on the Bernoulli means μ. A natural choice of priors for Bernoulli rewards corresponds to the Beta distribution, which corresponds to the conjugate distribution of the Binomial distribution. The prior distribution P tracking the rewards for each arm is then defined as:

$$P_i(\pi_i|\alpha_i,\beta_i) = \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}(1-\pi_i)^{(\beta+i-1)}\pi_i^{(\alpha_i-1)}. \tag{2.17}$$

where Γ is the gamma function.

In the Thompson sampling algorithm, these prior distributions are the mechanism that the gambler has to keep track of the experience he has acquired by playing the slot machines. Every time an arm is played, the prior distribution is updated by changing the parameters α_i and β_i by making α_i the number of times a reward was obtained and β_i the number of failures. In figure 2.7 we can observe how the updates of the parameters α_i and β_i model the knowledge acquired by playing the arms. The top row corresponds to the updated prior distributions resulting from playing an arm which returns a reward with $p = 0.75$, while the bottom one shows an arm which returns a reward with $p = 0.1$. At $t = 0$ the gambler knows nothing about the hidden reward distributions Π and therefore the estimated priors are equivalent to an uniform distribution. By playing the arms an updating the prior distributions, the gambler is able to approximate the real mean of the hidden reward distribution with a high degree of uncertainty.

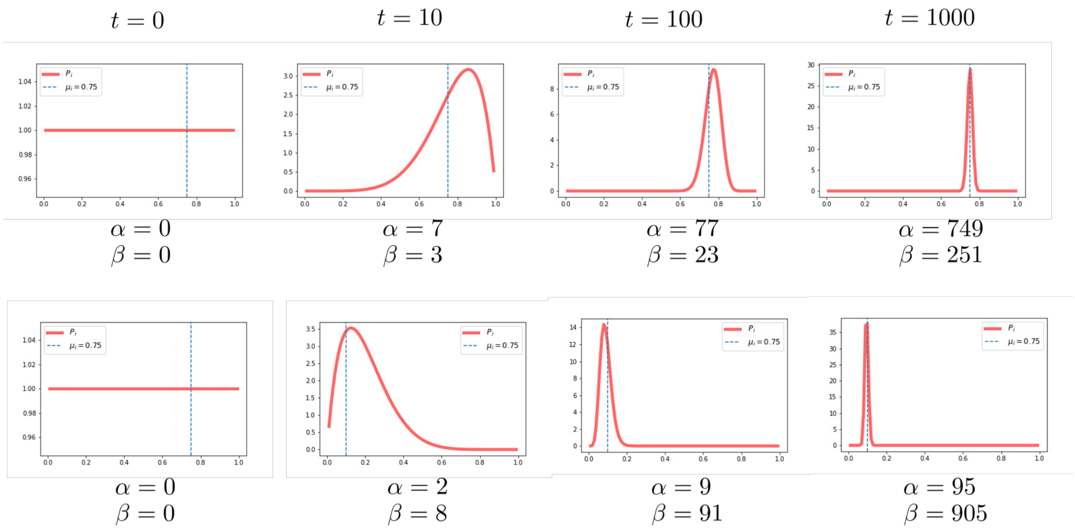


Figure 2.7 Updates of the prior distribution P_i using the Thompson sampling algorithm. Each row corresponds to a different arm, each one with a hidden reward probability μ. By playing each arm, and updating parameters α and β according to the observed rewards, we can approximate the real hidden probability distribution.

Given that the priors P_i are an approximation of the hidden distributions Π_i they can be used to guide the decision process of the gambler. At any time t the gambler can draw a sample π_i at random for each one of the prior distributions Π_i, and play which yields the maximum value. With this stochastic strategy the gambler will always tend to play those bandits that have shown a higher probability to yield a reward in the past, while at the same time keeping his options open to keep exploring the rest of the arms. The algorithm for thompson sampling for binary rewards are summarized in algorithm 1.

Algorithm 1 Thompson Sampling for Binary Rewards

```
$\alpha_i = 1, \beta_i = 1, \forall i \in \{1, \dots, N\}$
for $t = 1, 2, \dots$ do
    for $i = 1, \dots, N$ do
        Draw sample $\hat{\pi}_i$ from $P_i(\alpha_i, \beta_i)$.
        Play arm $i(t) := argmax_i \pi_i(t)$ and observe reward $r_t$.
        if $r_t == 1$ then $\alpha_j = \alpha_j + 1$
        else $\beta_j = \beta_j + 1$
```

We can draw parallels between the exploration/exploitation dilemma of the gambler which aims at maximizing his profit and a computer scientist looking to train a powerful machine learning model for medical image analysis using limited amounts of training data. While the gambler wants to find a way to effectively find those slot machines which are likely to give the higher rewards, the scientist requires to find which images are going to be helpful in the training of a model. Both have a limited budget, since the gambler has a finite amount of tokens and the scientist does not have the time to manually annotate all possible images.

Our contribution "A Multi-armed Bandit to Smartly Select a Training Set from Big Medical Data" follows this reasoning. We pose the problem of an efficient selection of a training set from big medical image data as a multi-armed bandit problem, and we propose a sample selection strategy based on the Thompson sampling algorithm. Our algorithm is based on an exploration/exploitation paradigm, since we aim at simultaneously exploiting data sources high chances of yielding useful images and exploring varied data sources to avoid overfitting during training.

3 Machine Learning for Medical Image Registration

Contents

3.1 Introduction to Medical Image Registration

The ultimate goal of medical image analysis as a discipline is to try to understand the anatomy of a patient through the information obtained from images of their body. Although medical imaging is an invaluable tool in modern routine clinical practice, any given image is limited in the amount of information that can provide to the clinician; the main reason behind this limitation is that an image is a partial *view* of the anatomy of the patient. These limitations are mostly caused by the physics behind the imaging process; in general medical images can only provide images of a spatially constrained region of interest, at a finite resolution and have different sensitivities depending on the type of tissue being imaged.

Due to the limited view provided by any possible image source, integrating information obtained from different images is of critical importance to aid diagnosis and to assist interventions with image based information. These different sources can either be images corresponding to the same anatomical region from different subjects, images of the same subject at different time points in a longitudinal study or images of the same anatomy obtained using different imaging modalities. In all these cases where information coming from different images has to be integrated in order to assist diagnosis or interventions, it is of critical importance to bring the images involved in the analysis into spatial alignment and/or to find anatomical correspondences between images.

This process of bringing two or more images into spatial alignment is commonly known as *image registration*. Image registration is the process of aligning two images so that corresponding features in both images correspond to the same anatomical position. Image registration is a critical step in many medical image processing task such as combining information obtained from different imaging modalities (CT, MRI, US,etc), to monitor changes in anatomy across

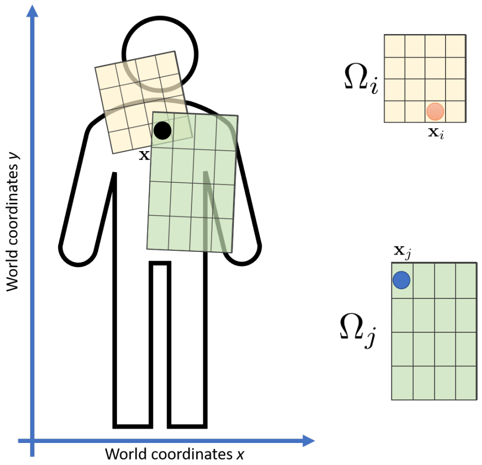


Figure 3.1 The anatomy of a patient exists within a coordinate framework defined by world coordinates. An image $\mathbf{I}(\mathbf{x})$ corresponds to the intensities obtained at certain world coordinates defined in a domain Ω. In the figure a world coordinate $\mathbf{x}$ is mapped to location $\mathbf{x}_i$ for image $\mathbf{I_i}$ and location $\mathbf{x}_j$ for image $\mathbf{I_j}$. Although $\mathbf{x}_i$ and $\mathbf{x}_j$ correspond to a similar spatial location, intensities $\mathbf{I_i}(\mathbf{x_i})$ and $\mathbf{I_j}(\mathbf{x_j})$ are completely different.

time for a particular individual, or to relate the anatomy of one patient to that of another subject or to an standarized atlas [47].

3.1.1 Components of an Image Registration Algorithm

As a starting point to explain a typical image registration algorithm we will present some basic definitions and notations. The first concept we present is that of *world coordinates*. As their name implies, world coordinates represent spatial locations in the *real world*, which means they represent specific positions at the patient anatomy. A world coordinate $\mathbf{x}$ is therefore a spatial point represented by their cartesian coordinates (x, y, z) with respect to an arbitrary origin (Fig. 3.1).

The second important concept to introduce is that of an *image*. If a world coordinate corresponds to a specific location of the anatomy of a patient, when a medical device is used to take an image, what we are actually doing is to obtain a mapping of some of these world coordinates $\mathbf{x}$ to intensity values. These points correspond to those within the field of view (or domain Ω) of the image. In this context, an image can be formalized as

$$\mathbf{I} : \mathbf{x} \in \Omega \subset \mathbb{R}^3 \mapsto \mathbf{I}(\mathbf{x}) \subset \mathbb{R}. \tag{3.1}$$

In image registration we aim at aligning two images. These images are usually known as the fixed image $\mathbf{I_f} \in \Omega_f$ and the moving image $\mathbf{I_m} \in \Omega_m$. Although each point $\mathbf{x}$ is unique since it corresponds to a specific spatial location of the anatomy of the patient, the mapping $\mathbf{I}(x)$ will in general have different values for $\mathbf{I_f}$ and $\mathbf{I_m}$. For example, while the skull of a patient can correspond to a bright value in a CT scan it would correspond to a dark value in an US scan due to the high reflection of US in bone . Also, an image $\mathbf{I}(\mathbf{x})$ has a limited domain $\boldsymbol{\Omega}$ both in terms of its field of view an the resolution of the imaging device. This means that an image is defined only for a finite number of world coordinates.

The next element to define is that of an *image transformation*. An image transformation, which we define using the operator $\mathcal{T}$, is a displacement of the world coordinates corresponding to an image.

$$\mathcal{T} : \mathbf{x} \mapsto \mathbf{x}' \Leftrightarrow \mathcal{T}(\mathbf{x}) = \mathbf{x}'. \tag{3.2}$$

The transformation operator $\mathcal{T}$ is therefore the mechanism that an image registration algorithm has to move $\mathbf{I_m}$ in order to bring it into alignment with $\mathbf{I_f}$. The nature of the transformation operator $\mathcal{T}$ can vary depending on the application and can range from rigid registration where only rotations and translations are allowed to a free-form deformation, where every point $\mathbf{x}$ is allowed to move to any position without any restriction. These transformations can also be represented by a parameter vector $\mathbf{p} \in \mathbb{R}^{N_p}$, which fully describes the transformation operator. For example, in the simple case of a 2D rigid transformation $\mathbf{p}$ corresponds to a vector containing the spatial translations and rotation $[t_x, t_y, \theta]$ which have to be applied to each point $\mathbf{x}$. On the case of the complex case of a free-form deformation $\mathbf{p}$ contains a series of position updates $[\delta(\mathbf{x}_0), \ldots, \delta(\mathbf{x}_n)]$ for each point in an image. In general for larger cardinalities N_p of the parameter vector, the transformation operator has more freedom to transform the moving image.

Another element in image registration which is often implied and not given much attention is the *resampling* operation. Although world coordinates correspond to real positions in the anatomy of the patient and therefore are continuous, the domains of the images Ω_f and Ω_m are discrete, and therefore are only defined for a limited subset of world coordinates. This is problematic since this means that points sampled by the moving and fixed image would rarely correspond to the exact same position in world coordinates. In order to solve this, a resampling operation is needed.

$$\mathbf{I} : \mathbf{I}(\mathbf{x}) \in \Omega \subset \mathbb{R}^3 \mapsto \mathbf{I_{mf}}(\mathbf{x_{mf}}) \in \Omega_{mf} \subset \mathbb{R}^3. \tag{3.3}$$

3.2 The standard image based registration algorithm

Given the elements of image based registration described in section 3.1.1, we can now present the standard image registration algorithm. An image registration algorithm aims at aligning $\mathbf{I_f}$ and $\mathbf{I_m}$ so that corresponding points in both images correspond to the same world coordinates x. A typical image-based registration algorithm (see Fig. 3.2) iteratively performs the following operations:

- *Transform* the moving image $\mathbf{I_m}$ using the parameters $\mathbf{p}$.

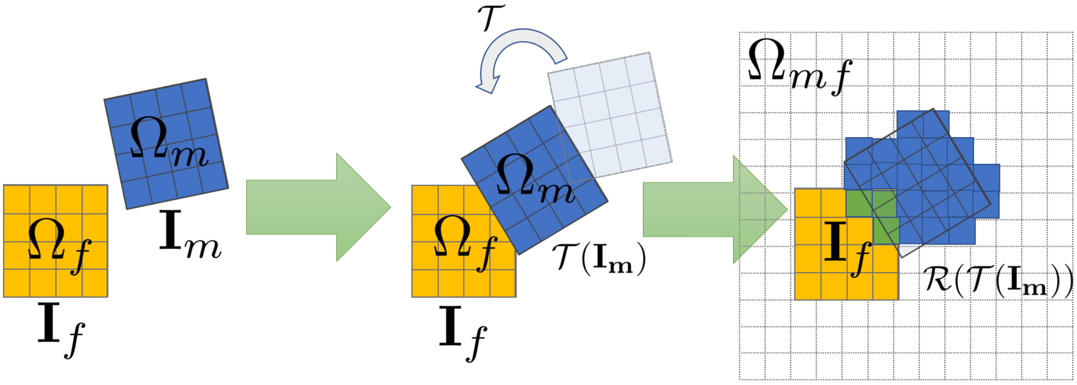


Figure 3.2 Image registration is based on two main operations: A spatial transformation $\mathcal{T}$ which corresponds to a movement of the spatial coordinates assigned to the moving image $\mathbf{I_m}$, and resampling $\mathcal{R}$ which consists on obtaining the values of an image in another domain with a different set of spatial coordinates.

- *Resample* $\mathbf{I_m}$ and $\mathbf{I_f}$ into a common domain Ω.
- *Measure* the similarity between $\mathbf{I_m}$ and $\mathbf{I_f}$ in this domain Ω.
- *Update* the transformation parameters $\mathbf{p}$ in order to maximize the similarity measurement between the two images.

We have already presented the transformation and resampling operations in the previous section. The main problems to solve in an image registration problem are: how to perform the similarity measurement between $\mathbf{I_m}$ and $\mathbf{I_f}$ and how to update the transformation parameters $\mathbf{p}$ accordingly. These problems are solved through the optimization of an energy function (also known as similarity metric) comparing the intensity values of the fixed image and the resampled moving image given a transformation $\mathcal{T}$. The optimal alignment between these two images can be obtained by finding the parameter vector $\mathbf{p}$ maximizing this energy function:

$$\mathbf{p}^* = arg\max_{\mathbf{p}} E(\mathbf{I_f}, \mathcal{R}(\mathcal{T}(\mathbf{I_m}))), \tag{3.4}$$

3.3 Classification of Image Registration Algorithms

Most image-based registration algorithms follow roughly the algorithm described in section 3.2. However, the field of medical image registration has been divided in several categories to study particular image registration problems (See figure 3.3 for an overview). According to the classification shown in figure 3.3, which is in turn an adaptation of the classification first proposed in [63], image based registration methods can be classified according to three different criteria:

- The characteristics of the images $\mathbf{I_f}$ and $\mathbf{I_m}$.
- The type of transformation $\mathcal{T}$.

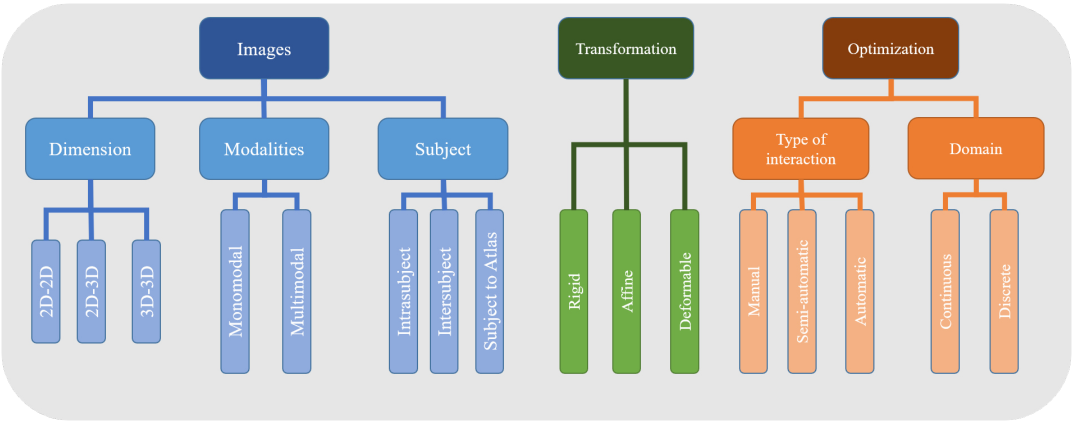


Figure 3.3 Classification of medical image registration algorithms. Methods can be mainly classified according to the type of images involved in the registration (blue), the type of transformation applied to the moving image (green), and the strategy used for optimization of the energy function (orange).

- The amount of interaction or feedback required by the human operator during the registration process.

Since most image registration tasks can be summarized to the formulation of equation 3.4, most research in medical image registration has also focused on a few common trends. Most fundamental research on the field of medical image registration has focused on the following trends:

- The design or choice of a *similarity metric*. A similarity metric can be defined as a measure of how compatible are images $\mathbf{I_f}$ and $\mathbf{I_m}$.
- The *parametrization* used to define the transformation $\mathcal{T}$.
- The optimization strategy to find the maximum of the energy function E.

3.4 Evolution of the Medical Image Registration Field

In 1896, less than a year after the discovery of x-rays, the first case of the use of an x-ray image in a clinical application was documented [33]. This case was that of a woman with a needle in her hand which had to be extracted. A x-ray image of the hand was obtained and the physician proceeded to *align* this image to the hand of the patient in order to guide the extraction. This alignment between the hand and the x-ray image was evidently performed manually, only with the visual information of the hand of the patient and the image. Although this image-patient alignment was good enough for minor procedures, it soon became evident that precise methods for image registration would enhance the use of medical images. Due to this need, several techniques were proposed for the alignment of images to patients. The first widely used method for image to patient registration was the stereotactic frame [84], which allowed the alignment first of x-ray images and later of CT scans to the head of the patient during neurosurgery. A stereotactic frame is rigidly attached to the skull and can be used to define a coordinate system for both the images and the patient. Although stereotactic frames provide reasonably accurate registration between patient and images, attaching a frame to the patient is invasive and restricts the movement of the neurosurgeons during a surgery and the types of images that can be acquired. Therefore, a need existed to develop methods which were able to perform registration based only on the information contained in the images themselves, without the need of any external device.

With the invention of digital computers, the medical imaging field saw an increased use of computational methods to assist in a variety of image processing tasks. As mentioned in chapter I, computers were first used to perform easy preprocessing tasks such as background subtraction and contrast correction, but by the 1980's they started to be used to perform basic first retrospective computer based image registration methods started to be proposed. Retrospective image registration methods aim at aligning images after they have been acquired, without any aid of markers or external devices. Therefore they rely *only* in the information contained in the images themselves [97]. Due to the computational cost of performing operations across images and the relatively lack of computational power at the time, these first methods required larges amount of user interaction in order to guarantee reliable registration results [15]. Also, early methods mainly relied on reducing the search space by not comparing the intensity values of all the pixels in the images to be aligned, but rather on performing a geometric alignment of either a few landmarks on each image corresponding to salient locations [31, 49] or curves describing the surface of regions of interest [3, 41].

With an increase of computational power, image-based registration techniques gained increasing popularity. Research in image-based registration methods focused mainly on two different areas: i) the development of cost functions able to compute similarities between images based solely on image information and ii) on improvements on the optimization algorithm of these cost functions. Research in cost functions saw the development of image-based similarity measures based on the sum of squared differences (SSD) and its variations [37, 48]. The main assumption of these measures is that the same object in both images should have the same intensities. These first approaches soon proved to be limited, since this requires not only that both images correspond to the same imaging modality, but is also dependent on acquisition settings which guarantee that intensity levels for both images are comparable. To try to overcome this limitation, approaches based on computing cross-correlations of image intensities between both images were proposed [19, 22, 58]. Cross correlation based metrics do not assume that the same intensities should be observed at both images, but rather rely on observing linear relationships between the intensities of the same object at each image. Although these approaches proved to be effective and fast metrics to perform mono-modal registration, the mentioned assumptions meant that these metrics were not suitable in the multi-modal registration setting. A breakthrough in the registration of multimodal images arrived with the introduction of mutual information as a similarity metric by Viola and Wells [90] and successive modifications to their approach [61, 85]. Mutual information proved to be so reliable that it became the de facto standard metric for multi-modal registration. Only a few years after its introduction, a review dedicated exclusively to the use of mutual information for medical image registration [74] included 165 papers.

In parallel to the development of novel similarity metrics, the field pushed towards optimization methods which could leverage on these similarity metrics to perform more accurate image registrations, with a reduced need of human interaction. Of particular importance was the development of methods which were able to perform non-rigid registration between images. Arguably the first popular approach to perform automatic non-rigid registration between medical images was the Automated Nonlinear Image Matching and Anatomical Labelling (ANIMAL) algorithm [23]. This approach became popular not only because of its high accuracy, but also because of its availabilaity as an out of the box registration tool to perform brain segmentation based on registering images to an atlas, and then transferring the labels. Given the success of ANIMAL, the following years saw the development of new algorithms to perform efficient image registration. In 1998, Thirion presented the demons algorithm, which provided fast non-rigid registration based on optical flow using an image based similarity metric (SSD) [87]. One year after, Rueckert *et al.* presented a new approach to perform image based non rigid registration using a parametrization based on free form deformation [80]. In contrast to the demons method, free-form deformations do not require that the intensities of tissues between images remain constant. An alternative formulation to the free form deformation is that of the large deformation registration framework [17]. The idea is to model the problem of image registration as a simulation of a viscous fluid. In contrast to the free form deformation approach, this formulation allows large deformations to be achieved. The large deformation framework has also been widely used in the field of computational anatomy [40] which studies anatomical shape variations within a population.

3.5 Machine Learning for Medical Image Registration

In the last few years, the field of medical image registration has been greatly influenced by the introduction of machine learning methods. These methods aim at introducing prior information about the images to be aligned in order to aid the registration algorithm. In a general sense, two strategies have been used to introduce machine learning methods in the medical image registration field: 1) using machine learning methods to estimate similarity measures between two images or 2) to directly predict deformation fields or transformation parameters.

The first family of approaches, commonly known as similarity learning obtain *a priori* information in the form of a training set of aligned examples. For example, the approaches by Jiang *et al.* [54] Lee *et al.* [57] and Michel *et al.* [68] posed the similarity learning problem in a discriminative manner by training a classification model able to discriminate between aligned and misaligned examples. With the advent of deep learning techniques, several approaches have been proposed for the use of deep learning architectures to learn similarity metrics. Among them Cheng *et al.* [16] proposed the use of stacked auto-encoders to learn similarities between CT and MRI images. In 2016 we proposed the use of convolutional neural networks to directly learn similarities between multimodal images [82].

The second group of approaches aim at estimating directly the transformation parameters given the joint appearance of the fixed and moving images. Due to the increased difficulty of this task, early learning based approaches were limited to a monomodal case. Kim *et al.* [55] learn updates of the transformation parameters to perform 2D-3D registration. Hu *et al.* propose to learn a deformation field to register monomodal images of the fetal brain [51].

Learning based registration methods for multimodal registration focused mainly on learning similarity metrics relating the intensities of one image modality to another. Although these approaches shown that it was possible to learn complex relationships between two different modalities they had two major limitations: 1) they relied on local relationships at a patch level between the two modalities and 2) they lack a mechanism to obtain the gradient of the metric directly, requiring the need of local gradient approximations or gradient free methods. As part of this thesis we propose to address multimodal registration as a supervised regression task where joint image descriptors are used as the input, and the parameters of the transformation aligning both images are obtained as the output. We model the joint appearance between the two images using context aware descriptors that capture both local an global clues simultaneously in both modalities. This method was presented first as a conference submission at MICCAI 2016 [46] and was further expanded as a full journal version in the special issue of medical image analysis [45]. The text of the latter is included in its full version as part of this thesis.

4 Conclusion and Outlook

In this dissertation we have aimed at discussing the impact of machine learning based approaches for medical image analysis tasks, in particular to the fields of computer assisted diagnosis and multimodal image registration. We have also presented three novel contributions, which leverage on machine learning approaches to tackle these tasks. We would now like to summarize these contributions, to present a discussion on the topics addressed in this thesis, and to present a short outlook on current trends related to our contributions.

As a first contribution we presented a CADx model based on an age estimation model. Age estimation models based on imaging data have gained increased popularity as a tool to measure brain abnormalities caused by heteregeneous factors, ranging from obvious factors such as neuropathology to systemic diseases (*e.g.* diabetes, obesity) or even environmental conditions [21]. However, we have demonstrated in our contribution that these approaches are limited by their use of age estimation error as a measure of deviation from healthy aging. Our model uses uncertainty based measures to assess abnormality caused by neuropathologies. Different to purely discriminative approaches, our method can be trained using *only* data acquired from healthy subjects. This allows us to use the same model to assess brain abnormalities caused by different unrelated diseases, as demonstrated in our experiments where we use our model to find morphological changes caused by autism or by Alzheimer's disease. Measuring and using model uncertainty has become an active area of research, particularly for deep network models for classification [38] and segmentation [79].

The second contribution consisted is a method for the efficient and intelligent acquisition of samples to construct a training dataset obtained from large scale medical records. Our method is based on the multi-armed bandit problem, and was solved using Thompson sampling. The main characteristic that separates our approach when compared to previous active learning approaches([103], [93], [89]) is that our approach relies *only* on meta data assigned to each one of the images and does not require the access and processing of imaging data. This approach can be particularly advantageous on situations where data acquisition and processing is cumbersome or expensive, but also in cases where imaging data is censored due to ethical constraints. Given the explosion on the availability of medical data we consider that active learning will be required to smartly train models in a smarter directed manner. The medical image analysis community has already shown interest in pursuing this directions as demonstrated in recent approaches. We highlight in particular the work by Maicas *et al.* [62] who presents an approach to train tumor detection models using small training sets is demonstrated, Wang *et al.* [92] shows the use of active learning models to train deep learning models for medical image segmentation and Hu *et al.* [52] who demonstrate the potential to deploy learning based multimodal image registration using weakly labelled data.

In our third contribution we presented a novel formulation using ensembles of decision trees to perform multimodal registration. We shown that our approach based on learning optimization updates is able to perform accurate segmentations between images coming from very different modalities such as US, MRI or histology. Additionally, our method based on estimating directly the transformation parameters is able to perform fast registration using different parametrizations. Although not included as part of this thesis, we further explored the idea of learning directly the gradient of the transformation using a deep learning architecture [82] and more recent works have further explored this direction [9, 10, 30, 100, 101]. We hope that further advancements in machine learning algorithms, particularly in regression methods based on deep learning architectures lead to further developments in this direction.

Medical imaging analysis is a discipline that has been highly impacted first by the introduction of digital imaging and in recent years by the introduction of machine learning methods. The current trend of the medical image analysis community has been the adoption of machine learning algorithms to approach segmentation, registration, shape analysis and diagnosis, with a large fraction of them being computer vision approaches adapted to medical image analysis tasks. Despite the ample success of these approaches in scientific venues, major hurdles still exist to translate these methods into applications which can be used in routine clinical care. First, it is of critical importance to understand that although medical image analysis tasks share common challenges with computer vision, they present particular characteristics that require holistic approaches which leverage on the ample knowledge of radiologists, clinicians, biologists and other related disciplines. Second, the community has to put concrete and strong efforts to be able to communicate the strengths and limitations of machine learning models and to make sure that experts from other areas do not not see machine learning boxes as a *magical black box* able to perform predictions out of thin air.

5 List of Authored and Co-authored Publications

2018

[44] **Benjamín Gutiérrez-Becker**, Tassilo Klein and Christian Wachinger. “Gaussian process uncertainty in age estimation as a measure of brain abnormality”. *Neuroimage* (175) 2018, pp. 246-258.

[73] Jorge Perez-Gonzalez, Fernando Arambula-Cosio, Mario Guzman, Lisbeth Camargo, **Benjamín Gutiérrez-Becker**, Diana Mateus, Nassir Navab, Veronica Medina-Banuelos. “Spatial compounding of 3-D fetal brain ultrasound using probabilistic maps”. *Ultrasound in Medicine and Biology 44.1, 278-291.*.

2017

[45] **Benjamín Gutiérrez Becker**, Diana Mateus, Loic Peter and Nassir Navab. “Guiding multimodal registration with learned optimization updates”. *Medical Image Analysis, 41, 2-17*.

[42] **Benjamín Gutiérrez Becker**, Loic Peter, Tassilo Klein and Christian Wachinger. “A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data”. *International Conference on Medical Image Computing and Computer-Assisted Interventions, 2017, Quebec, CA*.

[77] Marco Riva, Christoph Hennersperger, Fausto Milletari, Amin Katouzian, Federico Pessina **Benjamín Gutiérrez Becker**, Antonella Castellano, Nassir Navab and Lorenzo Bello. “3D intra-operative ultrasound and MR image guidance: pursuing an ultrasound-based management of brainshift to enhance neuronavigation”. *International journal of computer assisted radiology and surgery, 12.10, 1711-1725*.

2016

[82] Martin Simonovsky, **Benjamín Gutiérrez Becker**, Diana Mateus, Nassir Navab and Nikos Komodakis. “A deep metric for multimodal registration”. *International Conference on Medical Image Computing and Computer-Assisted Interventions, 2016, Athens, GR*.

[28] Florian Dubost, Loic Peter, Christian Rupprecht, **Benjamín Gutiérrez Becker** and Nassir Navab. “Hands-Free Segmentation of Medical Volumes via Binary Inputs”. *Deep Learning and Data Labeling for Medical Applications. 259-268*.

[46] **Benjamín Gutiérrez Becker** , Diana Mateus, Loic Peter and Nassir Navab. "Learning optimization updates for multimodal registration" *International Conference on Medical Image Computing and Computer-Assisted Interventions, 2016, Athens, GR* .

2014

[43] **Benjamín Gutiérrez Becker**, Diana Mateus, Ehab Shiban, Bernhard Meyer, Jens Lehmberg and Nassir Navab. “A sparse approach to build shape models with routine clinical data”. *International Conference on Medical Image Computing and Computer Assisted Interventions 2014. Beijing, CH.*

Bibliography

[1] H. Abe, H. MacMahon, R. Engelmann, et al. “Computer-aided diagnosis in chest radiography: results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies”. In: *Radiographics* 23.1 (2003), pp. 255–265 (cit. on p. 17).

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47.2-3 (2002), pp. 235–256 (cit. on p. 21).

[3] J. M. Balter, C. A. Pelizzari, and G. T. Chen. “Correlation of projection radiographs in radiation therapy using open curve segments and points”. In: *Medical physics* 19.2 (1992), pp. 329–334 (cit. on p. 30).

[4] C. Bardeen. “Determination of the size of the heart by means of the x-rays”. In: *American Journal of Anatomy* 23.2 (1918), pp. 423–487 (cit. on p. 2).

[5] H. Becker, W. Nettleton, P. Meyers, J. Sweeney, and C. Nice. “Digital computer determination of a medical diagnostic index directly from chest X-ray images”. In: *IEEE Transactions on Biomedical Engineering* 3 (1964), pp. 67–72 (cit. on p. 2).

[6] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz. “The exploration-exploitation dilemma: a multidisciplinary framework”. In: *PloS one* 9.4 (2014), e95693 (cit. on p. 20).

[7] D. W. Brown. “Digital computer analysis and display of the radioisotope scan”. In: *J Nucl Med* 5.Oct (1964), pp. 802–806 (cit. on p. 2).

[8] S. Brown. “Differential diagnosis of abdominal tumors by the roentgenological method”. In: *Radiology* 10.1 (1928), pp. 48–56 (cit. on p. 2).

[9] X. Cao, J. Yang, J. Zhang, et al. “Deformable image registration based on similarity-steered CNN regression”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 300–308 (cit. on p. 34).

[10] X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen. “Deformable image registration using a cue-aware deep regression network”. In: *IEEE Transactions on Biomedical Engineering* 65.9 (2018), pp. 1900–1911 (cit. on p. 34).

[11] R. D. Carman and S. Fineman. “The roentgenologic diagnosis of diaphragmatic hernia, with a report of seventeen cases”. In: *Radiology* 3.1 (1924), pp. 26–45 (cit. on p. 2).

[12] R. Caruana and A. Niculescu-Mizil. “An empirical comparison of supervised learning algorithms”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 161–168 (cit. on p. 13).

[13] N. Cesa-Bianchi and P. Fischer. “Finite-Time Regret Bounds for the Multiarmed Bandit Problem.” In: *ICML*. Citeseer. 1998, pp. 100–108 (cit. on p. 21).

[14] O. Chapelle and L. Li. “An empirical evaluation of thompson sampling”. In: *Advances in neural information processing systems*. 2011, pp. 2249–2257 (cit. on p. 21).

[15] G. T. Chen, M. Kessler, and S. Pitluck. “Structure transfer between sets of three dimensional medical imaging data”. In: *Computer graphics* 24 (1985), pp. 172–175 (cit. on p. 30).

[16] X. Cheng, L. Zhang, and Y. Zheng. “Deep similarity learning for multimodal medical images”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* (2016), pp. 1–5 (cit. on p. 31).

[17] G. E. Christensen, R. D. Rabbitt, M. I. Miller, et al. “Deformable templates using large deformation kinematics”. In: *IEEE transactions on image processing* 5.10 (1996), pp. 1435–1447 (cit. on p. 31).

[18] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 424–432 (cit. on p. 19).

[19] A. V. Cideciyan. “Registration of ocular fundus images: an algorithm using cross-correlation of triple invariant image descriptors”. In: *IEEE Engineering in Medicine and Biology Magazine* 14.1 (1995), pp. 52–58 (cit. on p. 30).

[20] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. “Active learning with statistical models”. In: *Journal of artificial intelligence research* 4 (1996), pp. 129–145 (cit. on p. 19).

[21] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary. “Brain age and other bodily ‘ages’: implications for neuropsychiatry”. In: *Molecular psychiatry* (2018), p. 1 (cit. on p. 33).

[22] D. L. Collins and A. C. Evans. “Animal: validation and applications of nonlinear registration-based segmentation”. In: *International journal of pattern recognition and artificial intelligence* 11.08 (1997), pp. 1271–1294 (cit. on p. 30).

[23] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. “Automatic 3-D model-based neuroanatomical segmentation”. In: *Human brain mapping* 3.3 (1995), pp. 190–208 (cit. on p. 31).

[24] A Criminisi, J Shotton, and S Bucciarelli. “Decision forests with long-range spatial context for organ localization in CT volumes”. In: Citeseer. 2009, pp. 69–80 (cit. on pp. 11, 12).

[25] A. Di Martino, C.-G. Yan, Q. Li, et al. “The autism brain imaging data exchange: towards large-scale evaluation of the intrinsic brain architecture in autism”. In: *Molecular psychiatry* 19.6 (2014), p. 659 (cit. on p. 16).

[26] K. Doi. “Computer-aided diagnosis in medical imaging: historical review, current status and future potential”. In: *Computerized medical imaging and graphics* 31.4-5 (2007), pp. 198–211 (cit. on p. 17).

[27] A.-T. Du, N. Schuff, J. H. Kramer, et al. “Different regional patterns of cortical thinning in Alzheimer’s disease and frontotemporal dementia”. In: *Brain* 130.4 (2007), pp. 1159–1166 (cit. on p. 17).

[28] F. Dubost, L. Peter, C. Rupprecht, B. Gutiérrez-Becker, and N. Navab. “Hands-Free Segmentation of Medical Volumes via Binary Inputs”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 259–268 (cit. on p. 36).

[29] S. Duchesne, A. Caroli, C. Geroldi, D. L. Collins, and G. B. Frisoni. “Relating one-year cognitive change in mild cognitive impairment to baseline MRI features”. In: *Neuroimage* 47.4 (2009), pp. 1363–1370 (cit. on p. 17).

[30] K. A. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, and J. P. Pluim. “Deformable image registration using convolutional neural networks”. In: *Medical Imaging 2018: Image Processing*. Vol. 10574. International Society for Optics and Photonics. 2018, 105740S (cit. on p. 34).

[31] D. L. Fill, D. J. Hawkes, Z. Hussain, S. E. Green, C. F. Ruff, and G. P. Robinson. “Accurate combination of CT and MR data of the head: validation and applications in surgical and therapy planning”. In: *Computerized medical imaging and graphics* 17.4-5 (1993), pp. 357–363 (cit. on p. 30).

[32] A. M. Fjell, K. B. Walhovd, C. Fennema-Notestine, et al. “CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer’s disease”. In: *Journal of Neuroscience* 30.6 (2010), pp. 2088–2101 (cit. on pp. 17, 18).

[33] M. A. Flower. *Webb’s physics of medical imaging*. CRC Press, 2012 (cit. on p. 29).

[34] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A. D. N. Initiative, et al. “Estimating the age of healthy subjects from T 1-weighted MRI scans using kernel methods: Exploring the influence of various parameters”. In: *Neuroimage* 50.3 (2010), pp. 883–892 (cit. on p. 18).

[35] K. Franke, C. Gaser, B. Manor, and V. Novak. “Advanced BrainAGE in older adults with type 2 diabetes mellitus”. In: *Frontiers in aging neuroscience* 5 (2013), p. 90 (cit. on p. 18).

[36] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 13).

[37] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. Frackowiak. “Spatial registration and normalization of images”. In: *Human brain mapping* 3.3 (1995), pp. 165–189 (cit. on p. 30).

[38] Y. Gal and Z. Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. 2016, pp. 1050–1059 (cit. on p. 33).

[39] M. L. Giger, H.-P. Chan, and J. Boone. “Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM”. In: *Medical physics* 35.12 (2008), pp. 5799–5820 (cit. on p. 1).

[40] U. Grenander and M. I. Miller. “Computational anatomy: An emerging discipline”. In: *Quarterly of applied mathematics* 56.4 (1998), pp. 617–694 (cit. on p. 31).

[41] A. Guéziec and N. Ayache. “Smoothing and matching of 3-D space curves”. In: *International Journal of Computer Vision* 12.1 (1994), pp. 79–104 (cit. on p. 30).

[42] B. Gutiérrez-Becker, L. Peter, T. Klein, and C. Wachinger. “A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 38–45 (cit. on p. 35).

[43] B. Gutiérrez-Becker, D. Mateus, E. Shiban, B. Meyer, J. Lehmberg, and N. Navab. “A sparse approach to build shape models with routine clinical data”. In: *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*. IEEE. 2014, pp. 258–261 (cit. on p. 36).

[44] B. Gutierrez-Becker, T. Klein, and C. Wachinger. “Gaussian process uncertainty in age estimation as a measure of brain abnormality”. In: *NeuroImage* 175 (2018), pp. 246–258 (cit. on p. 35).

[45] B. Gutiérrez-Becker, D. Mateus, L. Peter, and N. Navab. “Guiding multimodal registration with learned optimization updates”. In: *Medical image analysis* 41 (2017), pp. 2–17 (cit. on pp. 4, 32, 35).

[46] B. Gutiérrez-Becker, D. Mateus, L. Peter, and N. Navab. “Learning optimization updates for multimodal registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 19–27 (cit. on pp. 4, 32, 36).

[47] J. V. Hajnal and D. L. Hill. *Medical image registration*. CRC press, 2001 (cit. on p. 26).

[48] J. V. Hajnal, N. Saeed, A. Oatridge, E. J. Williams, I. R. Young, and G. M. Bydder. “Detection of subtle brain changes using subvoxel registration and subtraction of serial MR images.” In: *Journal of computer assisted tomography* 19.5 (1995), pp. 677–691 (cit. on p. 30).

[49] M. Herbin, A. Venot, J. Devaux, et al. “Automated registration of dissimilar images: application to medical imagery”. In: *Computer vision, graphics, and image processing* 47.1 (1989), pp. 77–88 (cit. on p. 30).

[50] G. N. Hounsfield. *Method of and apparatus for examining a body by radiation such as X or gamma radiation*. Tech. rep. 1975 (cit. on p. 2).

[51] S. Hu, L. Wei, Y. Gao, Y. Guo, G. Wu, and D. Shen. "Learning-based Deformable Image Registration for Infant MR Images in the First Year of Life". In: *Medical Physics* 44 (1 2016), pp. 158,170 (cit. on p. 32).

[52] Y. Hu, M. Modat, E. Gibson, et al. "Weakly-supervised convolutional neural networks for multimodal image registration". In: *Medical image analysis* 49 (2018), pp. 1–13 (cit. on p. 33).

[53] C. R. Jack, M. A. Bernstein, N. C. Fox, et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of magnetic resonance imaging* 27.4 (2008), pp. 685–691 (cit. on p. 16).

[54] J. Jiang, S. Zheng, A. Toga, and Z. T. "Learning based coarse-to-fine image registration". In: 2008, pp. 1–7 (cit. on p. 31).

[55] M. Kim, G. Wu, P. T. Yap, and D. Shen. "A General Fast Registration Framework by Learning Deformation Appearance Correlation". In: *IEEE Transactions on Image Processing* 21.4 (2012), pp. 1823–1833 (cit. on p. 32).

[56] N. Koutsouleris, C. Davatzikos, S. Borgwardt, et al. "Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders". In: *Schizophrenia bulletin* 40.5 (2013), pp. 1140–1153 (cit. on p. 18).

[57] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N. Cahill, and B. Scholkopf. "Learning similarity measure for multi-modal 3D image registration". In: 2009 (cit. on p. 31).

[58] L Lemieux, R Jagoe, D. Fish, N. Kitchen, and D. Thomas. "A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs". In: *Medical physics* 21.11 (1994), pp. 1749–1760 (cit. on p. 30).

[59] G. S. Lodwick, C. L. Haun, W. E. Smith, R. F. Keller, and E. D. Robertson. "Computer diagnosis of primary bone tumors: A preliminary report". In: *Radiology* 80.2 (1963), pp. 273–275 (cit. on p. 17).

[60] L. B. Lusted. "Medical electronics". In: *New England Journal of Medicine* 252.14 (1955), pp. 580–585 (cit. on p. 2).

[61] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. "Multimodality image registration by maximization of mutual information". In: *IEEE Transactions on Medical Imaging* 16.2 (1997), pp. 187–198 (cit. on p. 30).

[62] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro. "Training Medical Image Analysis Systems like Radiologists". In: *arXiv preprint arXiv:1805.10884* (2018) (cit. on p. 33).

[63] J. A. Maintz and M. A. Viergever. "A survey of medical image registration". In: *Medical image analysis* 2.1 (1998), pp. 1–36 (cit. on p. 28).

[64] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults". In: *Journal of cognitive neuroscience* 19.9 (2007), pp. 1498–1507 (cit. on p. 16).

[65] A. Materka. "Texture analysis methodologies for magnetic resonance imaging". In: *Dialogues in clinical neuroscience* 6.2 (2004), p. 243 (cit. on p. 17).

[66] L. K. McEvoy, C. Fennema-Notestine, J. C. Roddey, et al. "Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment". In: *Radiology* 251.1 (2009), pp. 195–205 (cit. on p. 17).

[67] P. H. Meyers, C. M. Nice Jr, H. C. Becker, W. J. Nettleton Jr, J. W. Sweeney, and G. R. Meckstroth. "Automated computer analysis of radiographic images". In: *Radiology* 83.6 (1964), pp. 1029–1034 (cit. on p. 17).

[68] F. Michel, M. Bronstein, A. Bronstein, and N. Paragios. “Boosted metric learning for 3D multimodal deformable registration”. In: IEEE. 2011, pp. 1209–1214 (cit. on p. 31).

[69] T. M. Mitchell et al. *Machine learning. WCB*. 1997 (cit. on pp. 3, 8).

[70] R. Murray. “Orthopaedic radiology: an expanding discipline.” In: *Journal of the Royal Society of Medicine* 73.5 (1980), p. 320 (cit. on p. 1).

[71] I. Nenadic, M. Dietzek, K. Langbein, H. Sauer, and C. Gaser. “BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder”. In: *Psychiatry Res. Neuroimaging* 266.March (2017), pp. 86–89 (cit. on p. 18).

[72] P. G. Newman and G. S. Rozycki. “The history of ultrasound”. In: *Surgical clinics of north America* 78.2 (1998), pp. 179–195 (cit. on p. 2).

[73] J. Perez-Gonzalez, F. Arámbula-Cosío, M. Guzmán, et al. “Spatial Compounding of 3-D Fetal Brain Ultrasound Using Probabilistic Maps”. In: *Ultrasound in medicine & biology* 44.1 (2018), pp. 278–291 (cit. on p. 35).

[74] J. P. Pluim, J. A. Maintz, and M. A. Viergever. “Mutual-information-based registration of medical images: a survey”. In: *IEEE transactions on medical imaging* 22.8 (2003), pp. 986–1004 (cit. on p. 30).

[75] K. Preston. “Computer processing of biomedical images”. In: *Computer* 9.5 (1976), pp. 54–68 (cit. on p. 2).

[76] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005 (cit. on p. 14).

[77] M. Riva, C. Hennersperger, F. Milletari, et al. “3D intra-operative ultrasound and MR image guidance: pursuing an ultrasound-based management of brainshift to enhance neuronavigation”. In: *International journal of computer assisted radiology and surgery* 12.10 (2017), pp. 1711–1725 (cit. on p. 35).

[78] W. C. Röntgen. “On a new kind of rays”. In: *Science* 3.59 (1896), pp. 227–231 (cit. on p. 1).

[79] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger. “Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-wise Quality Control”. In: *arXiv preprint arXiv:1811.09800* (2018) (cit. on p. 33).

[80] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. “Nonrigid registration using free-form deformations: application to breast MR images”. In: *IEEE transactions on medical imaging* 18.8 (1999), pp. 712–721 (cit. on p. 31).

[81] A. L. Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229 (cit. on p. 7).

[82] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis. “A deep metric for multimodal registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 10–18 (cit. on pp. 31, 34, 36).

[83] M. C. Sosman. “Radiology as an aid in the diagnosis of skull and intracranial lesions”. In: *Radiology* 9.5 (1927), pp. 396–407 (cit. on p. 2).

[84] E. A. Spiegel, H. T. Wycis, M Marks, and A. Lee. “Stereotaxic apparatus for operations on the human brain”. In: *Science* 106.2754 (1947), pp. 349–350 (cit. on p. 29).

[85] C. Studholme, D. L. Hill, and D. J. Hawkes. “An overlap invariant entropy measure of 3D medical image alignment”. In: *Pattern recognition* 32.1 (1999), pp. 71–86 (cit. on p. 30).

[86] W. N. Tauxe, D. W. Chaapel, and A. C. Sprau. “Contrast enhancement of scanning procedures by high speed digital computer”. In: *J Nucl Med* 7.9 (1966), pp. 647–656 (cit. on p. 2).

[87] J.-P. Thirion. “Image matching as a diffusion process: an analogy with Maxwell’s demons”. In: *Medical image analysis* 2.3 (1998), pp. 243–260 (cit. on p. 31).

[88] W. R. Thompson. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. In: *Biometrika* 25.3/4 (1933), pp. 285–294 (cit. on p. 21).

[89] A. Top, G. Hamarneh, and R. Abugharbieh. “Active learning for interactive 3D image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 603–610 (cit. on pp. 19, 33).

[90] P. Viola and W. M. Wells III. “Alignment by maximization of mutual information”. In: *International journal of computer vision* 24.2 (1997), pp. 137–154 (cit. on p. 30).

[91] C. Wachinger, P. Golland, W. Kremen, B. Fischl, and M. Reuter. “BrainPrint: a discriminative characterization of brain morphology”. In: *NeuroImage* 109 (2015), pp. 232–248 (cit. on p. 17).

[92] G. Wang, W. Li, M. A. Zuluaga, et al. “Interactive medical image segmentation using deep learning with image-specific fine-tuning”. In: *IEEE Transactions on Medical Imaging* (2018) (cit. on pp. 19, 33).

[93] G. Wang, M. A. Zuluaga, R. Pratt, et al. “Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views”. In: *Medical image analysis* 34 (2016), pp. 137–147 (cit. on pp. 19, 33).

[94] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos. “High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables”. In: *Neuroimage* 50.4 (2010), pp. 1519–1535 (cit. on p. 17).

[95] J. Weatherwax and R. Charles. “Determination of radiation values in lung tissue with variable qualities of radiation”. In: *Radiology* 14.4 (1930), pp. 401–410 (cit. on p. 2).

[96] W. M. Wells III. *Medical Image Analysis–past, present, and future*. 2016 (cit. on p. 2).

[97] J. West, J. M. Fitzpatrick, M. Y. Wang, et al. “Retrospective intermodality registration techniques for images of the head: surface-based versus volume-based”. In: *IEEE Transactions on Medical Imaging* 18.2 (1999), pp. 144–150 (cit. on p. 30).

[98] T. Wyss-Coray. “Ageing, neurodegeneration and brain rejuvenation”. In: *Nature* 539.7628 (2016), p. 180 (cit. on p. 18).

[99] Z. Yan, Y. Zhan, Z. Peng, et al. “Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1332–1343 (cit. on p. 19).

[100] X. Yang, R. Kwitt, M. Styner, and M. Niethammer. “Fast predictive multimodal image registration”. In: *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE. 2017, pp. 858–862 (cit. on p. 34).

[101] X. Yang, R. Kwitt, M. Styner, and M. Niethammer. “Quicksilver: Fast predictive image registration–a deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396 (cit. on p. 34).

[102] D. Zhang, D. Shen, A. D. N. Initiative, et al. “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease”. In: *NeuroImage* 59.2 (2012), pp. 895–907 (cit. on p. 17).

[103] Z.-H. Zhou. “A brief introduction to weakly supervised learning”. In: *National Science Review* 5.1 (2017), pp. 44–53 (cit. on pp. 19, 33).

List of Figures

Part I

Appendix: Full text of Contributions

Gaussian Process Uncertainty in Age Estimation as a Measure of Brain Abnormality

Benjamin Gutierrez,Loic Peter, Tassilo Klein, Christian Wachinger

Copyright statement Originally published as: "B.Gutierrez-Becker, T. Klein, C. Wachinger. Gaussian process uncertainty in age estimation as a measure of brain abnormality In *Neuroimage* 175 (2018), pp. 246-258". Reprinted under the consent of Elsevier.

Contributions The author of this thesis was responsible for the main idea of using uncertainty based measures for the measuring of abnormality caused by neuropathology, as well as the implementation of the algorithms, the experimental validation and the writing of the manuscript. Co-authors contributed with the discussion of the main ideas of the paper, collaborated with the wri

Guiding Multimodal Registration with Learned Optimization Updates

Benjamin Gutierrez, Diana Mateus, Loic Peter, Nassir Navab

Copyright statement Originally published as: "B.Gutierrez-Becker, D. Mateus, L.Peter, N. Navab. Guiding Multimodal registration with learned optimization updates In *Medical Image Analysis* 41 (2017), pp. 2-17". Reprinted under the consent of Elsevier.

Contributions The author of this thesis was responsible for the main idea of developing a machine learning based method applied to the mutimodal registration task as well as the implementation of the algorithms, the experimental validation and the writing of the manuscript. Co-authors contributed with the discussion of the main ideas of the paper, collaborated with the writing and revision of the manuscript.

A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data

Benjamin Gutierrez,Loic Peter, Tassilo Klein, Christian Wachinger

Copyright statement Originally published as: **Benjamín Gutiérrez Becker**, Loic Peter, Tassilo Klein and Christian Wachinger. “A Multi-Armed Bandit to Smartly Select a Training Set from Big Medical Data”. *International Conference on Medical Image Computing and Computer-Assisted Interventions, 2017, Quebec, CA*. Reprinted under the consent of Elsevier.

NeuroImage 175 (2018) 246–258

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Gaussian process uncertainty in age estimation as a measure of brain abnormality☆

Benjamin Gutierrez Becker [a,c,*], Tassilo Klein [b], Christian Wachinger [a], for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[1]

[a] Artificial Intelligence in Medical Imaging (AI-Med), Department of Child and Adolescent Psychiatry, LMU München, Germany
[b] SAP SE Berlin, Germany
[c] CAMP, Technische Universität München, Germany

ARTICLE INFO

Keywords:
Age estimation
Uncertainty
Alzheimer's disease
Autism

ABSTRACT

Multivariate regression models for age estimation are a powerful tool for assessing abnormal brain morphology associated to neuropathology. Age prediction models are built on cohorts of healthy subjects and are built to reflect normal aging patterns. The application of these multivariate models to diseased subjects usually results in high prediction errors, under the hypothesis that neuropathology presents a similar degenerative pattern as that of accelerated aging. In this work, we propose an alternative to the idea that pathology follows a similar trajectory than normal aging. Instead, we propose the use of metrics which measure deviations from the mean aging trajectory. We propose to measure these deviations using two different metrics: uncertainty in a Gaussian process regression model and a newly proposed age weighted uncertainty measure. Consequently, our approach assumes that pathologic brain patterns are different to those of normal aging. We present results for subjects with autism, mild cognitive impairment and Alzheimer's disease to highlight the versatility of the approach to different diseases and age ranges. We evaluate volume, thickness, and VBM features for quantifying brain morphology. Our evaluations are performed on a large number of images obtained from a variety of publicly available neuroimaging databases. Across all features, our uncertainty based measurements yield a better separation between diseased subjects and healthy individuals than the prediction error. Finally, we illustrate differences in the disease pattern to normal aging, supporting the application of uncertainty as a measure of neuropathology.

Introduction

The brain is a complex organ whose morphology varies substantially across the population. The causes of morphological variation have not yet been fully understood, but several studies have reported on potential causal factors including age (Guttmann et al., 1998; Franke et al., 2010; Ziegler et al., 2012; Wachinger et al., 2015), sex (Ingalhalikar et al., 2014), pathologies like dementia (Gaser et al., 2013; Wachinger et al., 2016), and even environmental factors such as education and physical activity (Steffener et al., 2016). Among all these variables, age was shown to be the main factor determining brain morphology (Potvin et al., 2017). Due to the wide impact of aging on brain morphology, multivariate regression methods using features based on brain morphology can in turn be used to estimate a subject's age. A recent volume of work has focused on modeling the normal aging of healthy individuals to predict a subject's age. Obtaining a prediction of the age with imaging features was

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

* Corresponding author. Waltherstrasse 23, KJP Forschungsabteilung, 80337, Munich, Germany

E-mail addresses: benjamin.gutierrez_becker@med.uni-muenchen.de (B. Gutierrez Becker), tassilo.klein@sap.com (T. Klein), christian.wachinger@med.uni-muenchen.de (C. Wachinger).

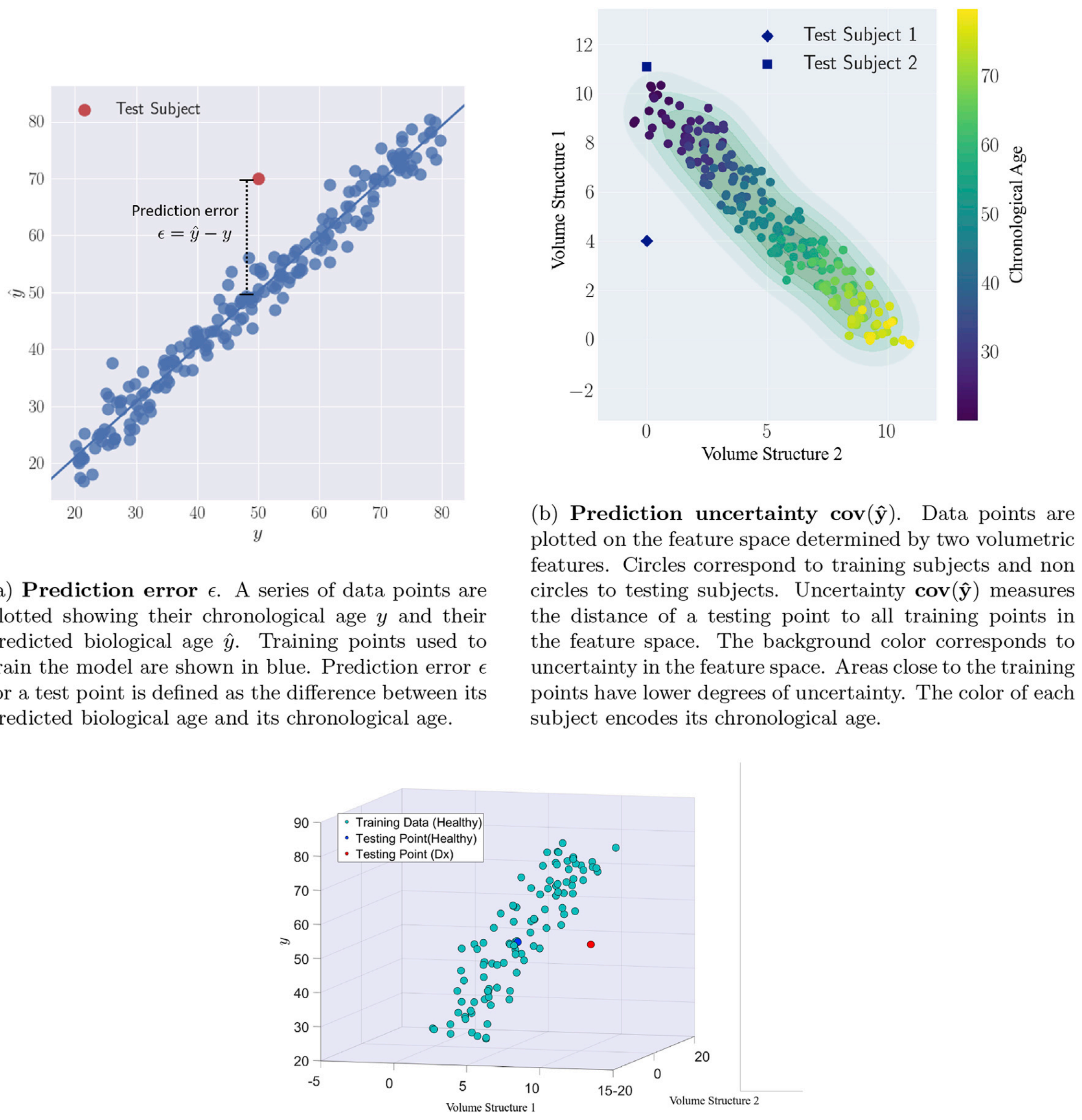
[1] Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

https://doi.org/10.1016/j.neuroimage.2018.03.075
Received 17 November 2017; Received in revised form 28 March 2018; Accepted 30 March 2018
Available online 5 April 2018

shown to be useful to derive imaging biomarkers, which can potentially be used to predict brain anomaly caused by disease (Cole and Franke, 2017).

The task of predicting age from brain images has been formulated as multivariate regression, where a predictive model is trained to relate structural information obtained from brain MR images to the chronological age of healthy subjects (Gaser et al., 2013; Wang et al., 2014; Kondo et al., 2015; Valizadeh et al., 2017; Liem et al., 2017). The prediction from these models is interpreted as an estimate of a subject's biological age, in contrast to a subject's chronological age. Of particular interest is the *prediction error*, which is defined as the difference between the biological and chronological age (Fig. 1). When predicting the age of healthy subjects, the prediction error is assumed to be small, while the prediction on subjects with neuropathology is assumed to result in large positive prediction errors. The error could therefore serve as a personalized marker of pathological processes (Franke et al., 2010; Gaser et al., 2013). The main assumption behind using the prediction error as a measure of pathology is that changes caused by neuropathology are equivalent to an accelerated aging process. Following this hypothesis, Gaser et al. (2013) and Habes et al. (2016) showed that changes related to Alzheimer's disease (AD) resemble *accelerated aging*, since differences between biological and chronological age are larger for individuals with AD than for healthy controls. Similar results on age differences have also been reported for individuals diagnosed with schizophrenia (Nenadic et al., 2017) and depression (Koutsouleris et al., 2013). In these studies, the age prediction model is trained using only images from healthy individuals. This means that contrary to their discriminative counterparts, a single age prediction model can be used to assess differences between healthy controls and individuals diagnosed with different conditions.

Although the findings of these studies show the big potential of using models of healthy aging to assess brain abnormality, a potentially limiting factor when quantifying neuropathology through the difference between chronological and predicted age, is the assumption that morphological changes caused by disease follow an accelerated aging



(a) **Prediction error** ϵ. A series of data points are plotted showing their chronological age y and their predicted biological age $\hat{y}$. Training points used to train the model are shown in blue. Prediction error ϵ for a test point is defined as the difference between its predicted biological age and its chronological age.

(b) **Prediction uncertainty** $\mathbf{cov}(\hat{\mathbf{y}})$. Data points are plotted on the feature space determined by two volumetric features. Circles correspond to training subjects and non circles to testing subjects. Uncertainty $\mathbf{cov}(\hat{\mathbf{y}})$ measures the distance of a testing point to all training points in the feature space. The background color corresponds to uncertainty in the feature space. Areas close to the training points have lower degrees of uncertainty. The color of each subject encodes its chronological age.

(c) **Weighted Uncertainty** $\mathbf{cov}_w(\hat{\mathbf{y}})$. Data points are shown on an extended feature space given by two volumetric features and their corresponding chronological age y. Weighted uncertainty $\mathbf{cov}_w(\hat{\mathbf{y}})$ measures the distance of a testing point to all training points in this extended feature space.

Fig. 1. Comparison between the three evaluated anomaly metrics: prediction error ε, GPR uncertainty $\mathbf{cov}(\hat{\mathbf{y}})$ and GPR age-weighted uncertainty $\mathbf{cov}_w(\hat{\mathbf{y}})$.

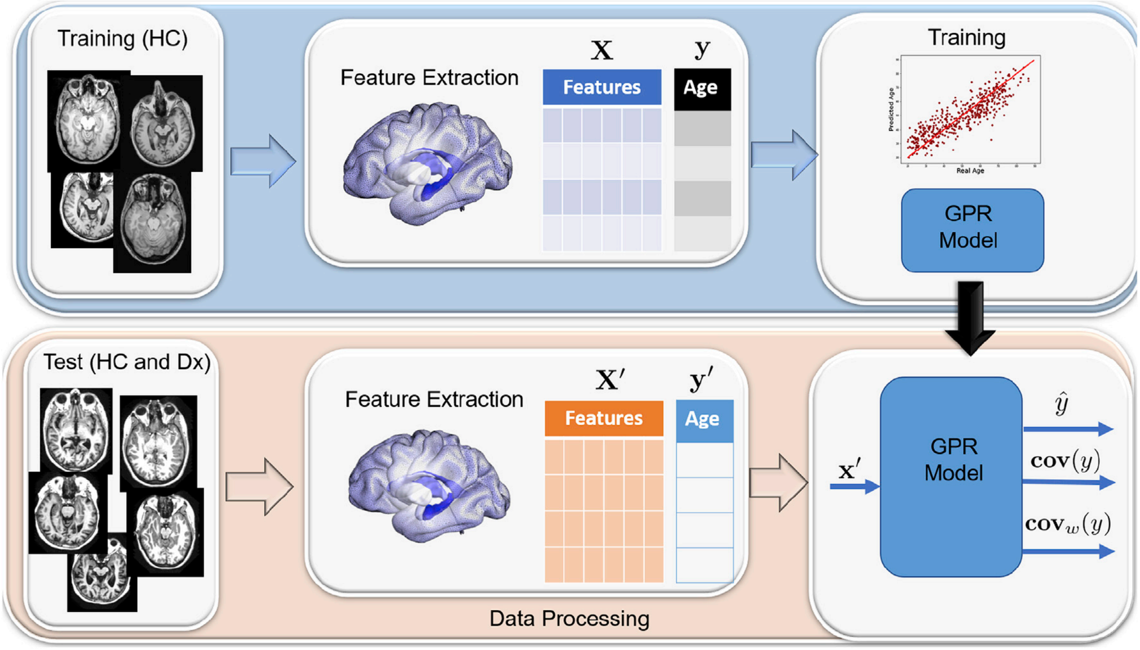


Fig. 2. Overview of our brain anomaly prediction model. The top part corresponds to the training stage where a set of images from healthy individuals is used to build a GPR age prediction model. The bottom part corresponds to the age prediction stage where the GPR model is used to predict the age of a set of test images. A predicted age $\hat{y}$ as well as the uncertainty measures $\mathbf{cov}(\hat{\mathbf{y}})$ and $\mathbf{cov}_w(\hat{\mathbf{y}})$ are obtained. These measurements can be used to find differences between the HC and Dx groups.

process. The assumption that brain anomaly is equivalent to accelerated aging may hold true for specific brain regions that accommodate neural systems with high susceptibility to deleterious factors, which are therefore affected by aging and disease processes. However, the assumption of accelerated aging does likely not extend to the whole brain, given that differences in brain morphology are caused by a variety of neurobiological processes that are complex and non-linear (Fjell et al., 2014; Buckner, 2004; Hedden and Gabrieli, 2004). This is potentially problematic, as current approaches for predicting the brain age are based on multivariate regression models that operate on gray matter maps or morphological features across the entire brain.

In this work, we build on the idea of modeling neuropathology as deviations from the healthy development of the brain. Our main hypothesis is that disease and aging result in brain-wide patterns of change. These patterns are not independent from each other and are essentially similar for several brain regions where disease results in patterns resembling accelerated aging. However, these accelerated aging patterns do not extend to the whole brain, making the assessment of deviations from healthy aging solely through prediction error problematic. We propose instead the use of Gaussian process regression (GPR). GPR can measure how a new subject deviates from previous observations used to construct the model by means of the posterior prediction uncertainty. Different to prediction error, GPR uncertainty is able to measure deviations from the healthy aging model without the implicit assumption of a brain-wide accelerated aging pattern. Particularly for the task of age prediction, we introduce a variation to traditional Gaussian processes regression that takes the known chronological age into account. This modification yields a weighted uncertainty measure. We evaluate our new method on a large collection of images obtained from several public datasets for assessing the variation to normal aging in mild cognitive impairment, Alzheimer's disease, and autism. Our results support the use of Gaussian process uncertainty and the age weighted uncertainty as tools to measure neuropathological patterns that deviate from healthy aging. Similar to previous age prediction models, our evaluations are done with a single age prediction model which is trained only on healthy controls, showing its versatility across different age ranges and diseases.

Materials and methods

Method overview

In this section, we describe our method for assessing neuropathology based on GPR uncertainty. Fig. 2 presents an overview of our method, which consists of two stages. In the training stage (top section of Fig. 2), we build a GPR model that estimates the chronological age of healthy subjects. This model is built using a dataset of MRI scans of healthy controls (section 2.2). Images are processed and segmented to extract a set of features describing brain morphology (section 2.3). Finally, a GPR model mapping the extracted features to a predicted age is trained on these features (section 2.4).

In the testing stage (bottom section of Fig. 2), we use the GPR model trained on healthy subjects to quantify deviations from the normal aging pattern on previously unseen subjects. In this stage, morphological features are extracted from the MR images of the test subjects, and these features are then used to obtain an estimate of the age of the subject using the GPR model. From the GPR model, we obtain the estimated age $\hat{y}$, an uncertainty measure of the estimation $\mathbf{cov}(\hat{\mathbf{y}})$, and a weighted uncertainty measure $\mathbf{cov}_w(\hat{\mathbf{y}})$ (see section 2.4.1 for details on these measurements). We will show in our experiments (section 3) that these measurements based on the uncertainty of the GPR model can be used to assess the similarity between subjects in the testing set and the healthy population in the training set.

Data

Similar to previous work on age prediction, we train an age regression model based on T1-MR images of healthy individuals. The training images for our age regression model are extracted from three different databases: IXI ,[2] ABIDE (Di Martino et al., 2014), and AIBL (Ellis et al., 2009). Details for each training dataset are shown in Table 1. We perform evaluation on 3 different test datasets summarized in Table 2. The

[2] http://brain-development.org/ixi-dataset/.

Table 1
Summary of the datasets used for training the age prediction model.

Training dataset	No. Images	Female/ Male	Age (Min-Max)	Age quantiles
IXI	581	311/270	19–87	33.7–48.6 - 62.2
ABIDE I	573	99/474	6–64	14.6–17.0 - 20.1
AIBL	409	209/200	55–92	67.0–73.0 - 79.0

Table 2
Summary of the datasets used for testing.

Testing dataset	No. Images	Female/ Male	Age (Min-Max)	Age quantiles	Target Dx
ADNI	3591	1422/ 2169	54–90	71.2–75.1 -79.7	MCI - AD
OASIS	196	129/67	60–82	71.0–76.0 - 82.0	MCI - AD
ABIDE II	1032	247/785	5–64	9.5–11.4 - 15.2	Autism

training and testing groups are therefore extracted from different databases and are independent from each other. For the first and second groups, obtained from the OASIS (Marcus et al., 2007) and ADNI (Jack et al., 2008) datasets, we aim at finding differences between Healthy Controls (HC), individuals diagnosed with Mild Cognitive Impairment (MCI), and subjects diagnosed with Alzheimer's disease (AD). For the third dataset, ABIDE II, we look for differences between HC and individuals diagnosed with autism. In total, 1543 images were used for training and 4819 images for testing.

Feature extraction

As mentioned in the overview section of our method, we require to extract features from structural MR images to quantify the brain morphology. Several image based features have previously been used for the age estimation task: Good et al. (2001), Franke et al. (2010) and Gaser et al. (2013) used Voxel Based Morphometry (VBM) features, which identify differences on the composition of brain tissue by registering all structural images to the same space. After segmenting gray and white matter, the voxel values of the gray matter extracted images are used as features. Finally the dimensionality of the feature space is reduced using Principal Component Analysis (PCA) and keeping only the principal modes of variation. Valizadeh et al. (2017) and Wang et al. (2014) use volumetric, thickness and curvature measurements of the brain, derived from the brain segmentation with FreeSurfer (Fischl, 2012). These different sets of features have been presented in separate studies but, to the best of our knowledge, have not yet been directly compared on the same age estimation task. In summary, we use three different types of features in our approach:

- VBM features (50 Principal Components),
- Thickness of 70 cortical structures.
- Volume of 50 brain structures.

Additionally we build a prediction model combining VBM, thickness and volume features together. VBM features were extracted using the CAT12 toolbox [3] together with the SPM12 toolbox [4] for segmentation. The preprocessing of the images, the segmentation of gray matter, and post processing were implemented in line with the pipeline proposed by Franke et al. (2010). Dimensionality reduction was performed using the PCA library included in the scikit-learn toolbox (Pedregosa et al., 2011). The principal component directions were estimated using only the training sample, and the testing data was projected to this estimated lower dimensional space. For all the analyses on thickness and volume features, FreeSurfer version 5.3 was used. The default Deskian/Killiany atlas was used for the parcellation to obtain thickness measurements. We are using all subcortical volume measurements as provided by FreeSurfer and described in the FreeSurfer subcortical segmentation pipeline.

Uncertainty estimation with Gaussian process regression

Several multivariate regression techniques have been previously used for the task of age prediction from brain MR images. A detailed comparison of the performance of neural networks, random forests, k-nearest neighbors, support vector machines, multiple linear regression and ridge regression was presented by (Valizadeh et al., 2017). In the work by Franke et al. (2010) relevance vector regression was preferred.

In our case, we are interested in modeling the age regression problem with a model that does not only provide estimates of the biological age, but also provides uncertainties of these estimates. Gaussian process regression achieves a comparable accuracy in age regression than standard regression techniques, while offering the advantage of providing an estimate of the uncertainty of each prediction. GPR models have been used successfully before as age prediction models (Cole et al., 2015, 2016), but the potential of using uncertainty based measurements as biomarkers has not been explored yet. In this section, we will briefly introduce GPR models, focusing particularly on the calculation of uncertainty, where we refer the reader to (Rasmussen and Williams, 2005) for a more detailed explanation, and introduce our modification for computing an age-weighted uncertainty.

Gaussian process

A Gaussian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen and Williams, 2005) with:

- a mean function $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$,
- and a covariance function $k(\mathbf{x},\mathbf{x}') = \mathbb{E}[f(\mathbf{x}) - m(\mathbf{x})f(\mathbf{x}') - m(\mathbf{x}')]$.

Although not necessary, it is often assumed that the mean function $m(\mathbf{x})$ of the GPR is zero. Therefore the design of a GPR is focused on the selection of an appropriate covariance function $k(\mathbf{x},\mathbf{x}')$ measuring the similarity between data points. This covariance function is equivalent to a similarity measure between two data points, giving small values for points that are close to each other and large values otherwise. In our case we define the covariance function as a squared exponential function of the form:

$$k(\mathbf{x}_i,\mathbf{x}_j) = \sum_{k=1}^{K} \exp\left[\frac{-\left(x_i^k - x_j^k\right)^2}{2l_k^2}\right] + \sigma_n^2\delta(\mathbf{x}_i,\mathbf{x}_j), \tag{1}$$

where σ_n^2 is the noise variance, l^k is the length scale of the k-th feature, and δ is the Kronecker delta function. We can think of the length scale vector $\mathbf{l} \in \mathbb{R}^K$ as a parameter controlling how close should two data points x_i and x_j should be in order to influence each other. In general, the smaller an element l^k is, the more dependent y is to the feature element x_k.

We model the joint distribution of the training and test outputs as:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}'\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}K(\mathbf{X},\mathbf{X}) & K(\mathbf{X},\mathbf{X}')\\ K(\mathbf{X}',\mathbf{X}) & K(\mathbf{X}',\mathbf{X}')\end{bmatrix}\right). \tag{2}$$

The elements of the joint distribution in Eq.(2) can be summarized as follows:

[3] http://www.neuro.uni-jena.de/cat/.
[4] http://www.fil.ion.ucl.ac.uk/spm/.

- an intra-covariance matrix of the training set $K(\mathbf{X},\mathbf{X}) \in \mathbb{R}^{mxm}$,
- an intra-covariance matrix of the testing set $K(\mathbf{X}',\mathbf{X}') \in \mathbb{R}^{nxn}$,
- an inter-covariance matrix between the training and testing set $K(\mathbf{X}, \mathbf{X}') \in \mathbb{R}^{mxn}$,
- a training labels vector $\mathbf{y} \in \mathbb{R}^{m}$, and
- a testing labels vector $\mathbf{y}' \in \mathbb{R}^{n}$,

where m corresponds to the number of training samples and n to the number of testing samples. The matrices $K(X,X)$ have the form:

$$K = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_1,\mathbf{x}_2) & \dots & k(\mathbf{x}_1,\mathbf{x}_n) \\ k(\mathbf{x}_2,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_2) & \dots & k(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m,\mathbf{x}_1) & k(\mathbf{x}_m,\mathbf{x}_2) & \dots & k(\mathbf{x}_m,\mathbf{x}_n) \end{bmatrix}, \tag{3}$$

where each element $k(\mathbf{x}_1,\mathbf{x}_2)$ corresponds to a measure of the similarity between two feature vectors x_i and x_j. We are interested in predicting the values for the test labels $\mathbf{y}'$, which have the following conditional distribution:

$$\mathbf{y}'|\mathbf{y},\mathbf{X},\mathbf{X}' \sim \mathcal{N}\left(K(\mathbf{X}',\mathbf{X}), K(\mathbf{X},\mathbf{X})^{-1}\mathbf{y}, K(\mathbf{X}',\mathbf{X}') - K(\mathbf{X}',\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}K(\mathbf{X},\mathbf{X}')\right). \tag{4}$$

Using this conditional distribution we can derive the predictive equations of a GPR:

$$\hat{\mathbf{y}}' = \mathbb{E}\left[\hat{\mathbf{y}}'\middle|\mathbf{X},\hat{\mathbf{y}},\mathbf{X}'\right] = K(\mathbf{X}',\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}\mathbf{y} \tag{5}$$

$$\mathbf{cov}(\hat{\mathbf{y}}) = K(\mathbf{X}',\mathbf{X}') - K(\mathbf{X}',\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}K(\mathbf{X},\mathbf{X}'), \tag{6}$$

which correspond to the predicted labels and to the estimated covariance, respectively. The estimated covariance can also be thought as a measure of uncertainty for the predicted values. This uncertainty estimate is usually used in the GPR framework to measure the degree of confidence of a predicted value $\hat{y}$' by measuring the similarity of a new observation with respect to the previous observations in the training set.

When training a GPR model, the parameters $\theta = \{\mathbf{l}, \sigma_n\}$ have to be tuned in order to fit the training data. This is done by maximizing the marginal likelihood of the model given by:

$$\log(p(\mathbf{y}|\mathbf{X},\theta)) = -\frac{1}{2}\mathbf{y}^T K(\mathbf{X},\mathbf{X})\mathbf{y} - \frac{1}{2}\log(K(\mathbf{X},\mathbf{X})) - \frac{n}{2}\log(2\pi). \tag{7}$$

By finding the parameters θ that maximize the marginal likelihood, we can obtain a GPR model that best fits the training data.

Age-weighted uncertainty

The uncertainty measurement of the GPR $\mathbf{cov}(\hat{\mathbf{y}})$ is solely defined with respect to the feature vectors x. In a common regression scenario this is a natural approach since the *real* values of the labels are unknown. However, in the case of an age estimation framework, we do possess the real values of the labels, which correspond to the chronological age of the patient.

We can introduce the age information into the GPR framework by creating age-weighted similarity matrices $K_w(\mathbf{X},\mathbf{X}',\mathbf{y},\mathbf{y}')$. Similar to the GPR covariance matrices we can construct three different similarity matrices:

- a weighted intra-similarity matrix for the training samples $K_w(\mathbf{X},\mathbf{X},\mathbf{y},\mathbf{y}) \in \mathbb{R}^{mxm}$,
- a weighted intra-similarity matrix for the testing samples $K_w(\mathbf{X}',\mathbf{X}',\mathbf{y}, \mathbf{y}') \in \mathbb{R}^{nxn}$, and
- a weighted inter-similarity matrix between the training and test samples $K_w(\mathbf{X},\mathbf{X}',\mathbf{y},\mathbf{y}') \in \mathbb{R}^{mxn}$.

These similarity matrices are constructed in the same manner as the covariance matrices presented in section 2.4.1. The only difference consists in a modification of the kernel to take into account differences in age. This is achieved by creating an age weighted similarity kernel of the form:

$$k_w(\mathbf{x}_i,\mathbf{x}_j,\mathbf{y}_i,\mathbf{y}_j) = s(\mathbf{y}_i,\mathbf{y}_j)k(\mathbf{x}_i,\mathbf{x}_j), \tag{8}$$

where $k(\mathbf{x}_i,\mathbf{x}_j)$ corresponds to the kernel defined in Eq.(1) and $s(\mathbf{y}_i,\mathbf{y}_j)$ corresponds to an age similarity term defined as:

$$s(\mathbf{y}_i,\mathbf{y}_j) = \exp\left[\frac{-(\mathbf{y}_i - \mathbf{y}_j)^2}{2l_y^2}\right] + \sigma_y^2\delta(\mathbf{y}_i,\mathbf{y}_j). \tag{9}$$

where l_y corresponds to the age length scale, which is a parameter controlling the effect of the age weighting. By using this updated kernel k_w, we obtain a *weighted uncertainty* term $\mathbf{cov}_w(y)$ which takes into account the age of the subjects to define similarities between subjects. This weighted uncertainty is obtained similar to the regular uncertainty presented in Eq. (6):

$$\mathbf{cov}_w(\mathbf{y}') = K_w(\mathbf{X}',\mathbf{X}') - K_w(\mathbf{X}',\mathbf{X})K_w(\mathbf{X},\mathbf{X})^{-1}K_w(\mathbf{X},\mathbf{X}'). \tag{10}$$

Prediction error, uncertainty and age-weighted uncertainty

In this work, we compare three age regression based metrics in order to measure their usefulness as a biomarker to distinguish between healthy controls and subjects with different neuropathologies. These metrics are the commonly used prediction error $\varepsilon = \hat{y} - y$ (Franke et al., 2010), the GPR uncertainty $\mathbf{cov}(y)$, and the GPR age-weighted uncertainty $\mathbf{cov}_w(y)$. As discussed in the introduction, the prediction error has previously been used to assess differences between healthy and non-healthy populations. The prediction error is the difference between the predicted and chronological age, as shown in Fig. 1. A higher prediction error is assumed to indicate an accelerated aging process (Franke et al., 2010; Gaser et al., 2013).

The computation of the GPR uncertainty $\mathbf{cov}(\hat{\mathbf{y}})$ was presented in section 2.4.1. It can be thought of as a metric on how close a testing point is to all the training points in the feature space, illustrated in Fig. 1. The scatter plot represents a set of subjects in a 2-dimensional space composed of the volume of two different structures. By training a GPR on a set of training points (represented by circles), we obtain a measure of uncertainty $\mathbf{cov}(\hat{\mathbf{y}})$ for every point in the 2D-space. This covariance matrix is represented by the shading of the grid, where darker regions correspond to regions where the predictor has higher confidence on its prediction. When performing prediction on previously unseen points (Test Subject 1 and Test Subject 2), we can obtain both a predicted age $\hat{y}$ and its confidence $\mathbf{cov}(\hat{\mathbf{y}})$. In Fig. 1, we observe that even though both test subjects get similar predicted values, the confidence of the prediction for subject two is higher due to its proximity to the training set.

The third metric, the age-weighted uncertainty $\mathbf{cov}_w(\hat{\mathbf{y}})$ expands upon the notion of uncertainty by taking into account the subject's age. Measuring $\mathbf{cov}_w(\hat{\mathbf{y}})$ is equivalent to adding a further dimension to the distance measured by the normal GPR uncertainty. The reasoning behind this is to give higher similarity to individuals which have similar morphological features to healthy individuals of similar ages. For example, we see in Fig. 1 that a healthy testing point (blue) is close to training points with similar features and age; the testing point would therefore have a high $\mathbf{cov}_w(\hat{\mathbf{y}})$ value. On the other hand, the Testing Point corresponding to a diseased subject (red point) would have a low $\mathbf{cov}_w(\hat{\mathbf{y}})$ value because even though there are individuals in the training set with similar feature values, they correspond to subjects with a different age range. Both proposed metrics are closely related. In fact, $\mathbf{cov}(\hat{\mathbf{y}})$ is equivalent to $\mathbf{cov}_w(\hat{\mathbf{y}})$ for the special case when $l_y = \infty$.

Aging and disease assessment

As discussed in the introduction, several studies have demonstrated that aging is a complex process, which affects different brain structures and regions at different rates of change. It has also been reported that deleterious changes caused by neurodegenerative disease follow a pattern that resembles an accelerated aging process. In order to evaluate how aging and neurodegenerative disease affect different brain regions, we performed an analysis of the volumetric features obtained from our training set. To facilitate this analysis we restricted our analysis to images obtained from the ADNI database. To assess which individual structures are affected either by aging, disease or both factors, a series of simple linear fixed effects model were fitted to our data. In each one of these models, the dependent variable corresponds to the volume of a different brain structure, and the independent variables correspond to age, sex, diagnosis ($0 = $ healthy, $1 = $ MCI/AD) and an interaction term between diagnosis and age.

Results

Assessing the effect of aging and disease on brain development

In this section we present the results obtained after fitting the aging and disease model described in section 2.6. In Table 3, we present the regression coefficients for age, diagnosis and the age/diagnosis interaction term as well as their corresponding p-values for the linear fixed effects model. The table is sorted by descending p-value for the diagnostic coefficient, which means that the structures at the top of the table are those which present more significant volume alterations caused by disease. Age and volume variables were normalized in order to make the coefficients of different structures comparable. In the case of bilateral structures we only show the values for the left hemisphere in order to simplify our analysis. Plots showing the progress of the hippocampus and cerebellum white matter across different ages for both healthy and individuals with MCI/AD are also presented to illustrate our results in Fig. 3. The hippocampus was selected since it was the structure which had the most evident effects of age and disease. On the other hand, the cerebellum white matter was selected as a structure which showed significant effects of aging but is apparently not largely affected by Alzheimer's disease.

There are a couple of relevant observations that can be extracted from Table 3. First, the regression coefficients for age and diagnosis always have the same sign for structures with significant associations, and there exist significant interactions between aging and diagnosis for most of the analyzed structures. This supports the hypothesis that disease and aging are overlapping processes that affect the brain structures in the same direction. However, we can also observe in Table 3 that there exist some structures that although largely affected by aging do not present significant disease effects (*i.e.* cerebellum white matter). This can also be observed on the box plots in the top of Fig. 3, where a clear difference between the HC and Dx groups is evident for all age ranges in the case of the hippocampus, whereas for the cerebellum white matter no significant differences exist between both groups.

To further illustrate the point that aging and disease are processes that affect different regions of the brain at different rates, we show the progress of pairs of features for both the HC and Dx groups (bottom of Fig. 3). By looking at the central plot, where the volume of left and right hippocampus is shown, we can understand the reasoning behind the accelerated aging hypothesis of previous age estimation works. Indeed, by looking only at these features, we would be tempted to conclude that the brain of a healthy 80 year old is essentially similar to that of a diseased 60 year old. However, this observation contrasts with the left plot, where the left and right cerebellum white matter volumes are shown. By looking at these features alone, we would draw a different conclusion, since it would appear that there are no differences between the brains of healthy and diseased subjects of the same age. By looking at the left hippocampus and left cerebellum white matter simultaneously (right in Fig. 3), we can observe that disease produce changes in the brain that are essentially different to those of accelerated aging, causing the overall appearance of the brain of an average 60 year old diagnosed with AD to be different to a healthy individual of any age. These observations support our hypothesis that morphological changes associated to AD and MCI are complex and that a model of accelerated aging across the whole brain may be too simplistic to model the specific effects of disease and aging at specific brain structures.

Training of the age prediction model

Using the training datasets summarized in Table 1, we train 4 different GPR models, each one with a different set of features as described in section 2.3. Each one of the GPR models is trained to estimate the age of healthy subjects based on either volume, thickness, VBM features or a combination of all features. Our models were implemented using python together with the scikit-learn toolbox. In Table 4, we show the Mean Absolute Error (MAE) and R^2 score for the training set, using a 5-fold cross validation. Our model presents similar MAE and R^2 when compared to previous work on age estimation (Valizadeh et al., 2017; Cole et al., 2016). Similar to previously reported results (Valizadeh et al., 2017), we observed higher R^2 score and lower MAE for the model trained using an ensemble of all available features. The chronological and

Table 3
Coefficients and p-values corresponding to the linear models fitted to predict volume of individual structures. Structures are sorted by descending p-value for diagnostic.

Structure	Age Coefficient	Dx Coefficient	Age × Dx Coefficient	Age p-value	Dx p-value	Age × Dx p-value
Left.Hippocampus	−0.30	−0.53	0.03	9.12×10^{-107}	4.04×10^{-291}	2.87×10^{-02}
Left.Amygdala	−0.21	−0.41	0.07	4.13×10^{-50}	1.53×10^{-169}	1.67×10^{-05}
Left.Inf.Lat.Vent	0.27	0.38	−0.07	1.77×10^{-70}	1.50×10^{-141}	3.00×10^{-05}
Left.Lateral.Ventricle	0.22	0.24	−0.09	8.38×10^{-46}	1.84×10^{-57}	1.32×10^{-07}
CSF	0.16	0.22	−0.11	1.07×10^{-23}	1.85×10^{-46}	5.95×10^{-11}
Left.Accumbens.area	−0.30	−0.22	0.07	2.75×10^{-75}	1.48×10^{-43}	7.89×10^{-05}
3rd.Ventricle	0.27	0.19	−0.13	1.04×10^{-70}	1.53×10^{-36}	3.49×10^{-15}
Left.choroid.plexus	0.14	0.17	−0.04	2.04×10^{-18}	1.59×10^{-27}	2.16×10^{-02}
Left.Putamen	−0.15	−0.14	0.02	9.84×10^{-21}	2.15×10^{-17}	0.30
Left.VentralDC	−0.27	−0.11	−0.03	7.09×10^{-71}	3.38×10^{-14}	3.00×10^{-02}
Left.Thalamus.Proper	−0.32	−0.11	−0.01	6.28×10^{-97}	2.65×10^{-13}	0.48
Left.Cerebellum.Cortex	−0.26	−0.07	−0.03	1.98×10^{-65}	5.01×10^{-06}	9.02×10^{-02}
Brain.Stem	−0.23	−0.06	−0.02	1.15×10^{-48}	1.60×10^{-05}	0.13
Left.Caudate	0.07	0.04	0.03	7.89×10^{-06}	1.93×10^{-02}	0.12
Left.Cerebellum.White.Matter	−0.30	−0.03	0.00	1.60×10^{-74}	0.10	0.86
Left.Pallidum	−0.06	−0.02	−0.04	4.84×10^{-04}	0.23	1.31×10^{-02}
4th.Ventricle	0.09	0.00	−0.08	2.38×10^{-07}	0.83	1.13×10^{-05}
Left.vessel	0.12	0.00	−0.02	1.05×10^{-12}	0.93	0.37

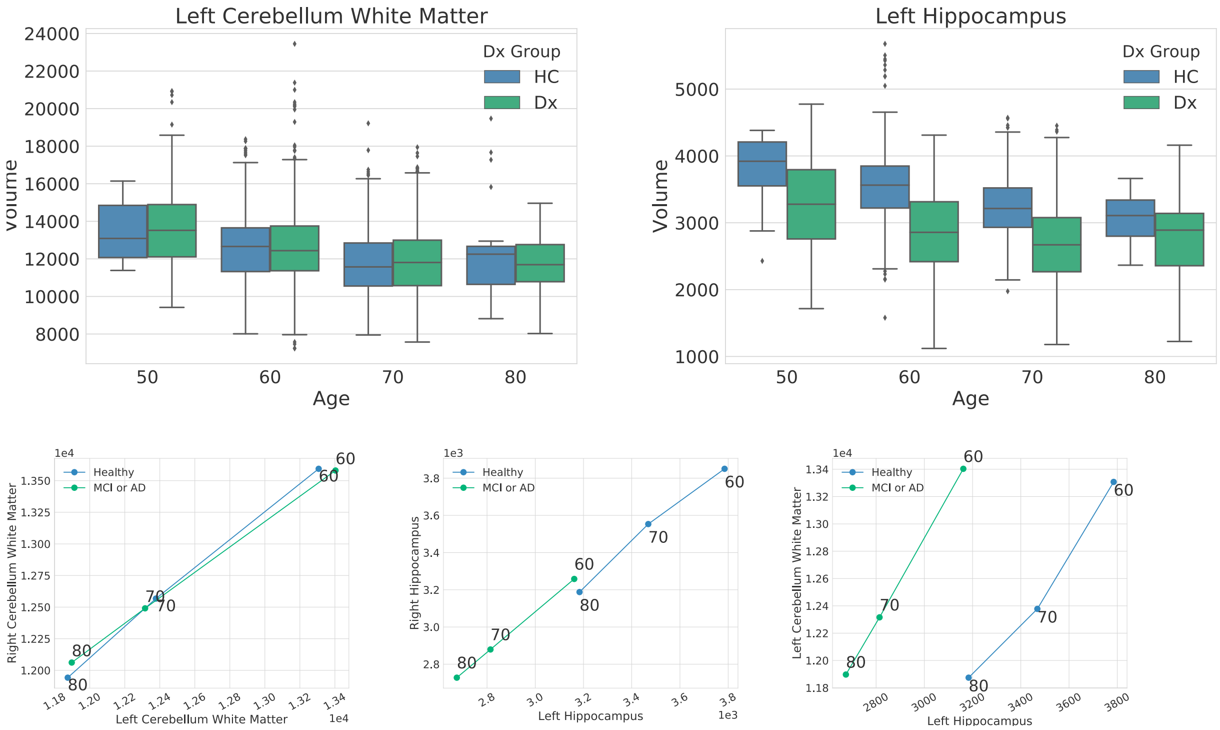


Fig. 3. Top: Box plots showing changes in left cerebellum white matter volume and hippocampus volume for individuals between 50 and 80 years old. Bottom: Plots showing age progression of feature pairs for three different cases; left: structures that are not affected by disease; center: structures affected by disease that show an accelerated aging pattern; right: one structure affected by disease (left hippocampus) and one structure with no significant disease effect (left cerebellum white matter).

Table 4
Mean Absolute Error (MAE) and R^2 score of the age prediction models trained with different sets of features. Measurements are obtained using a 5-fold cross validation on the training set.

Feature Set	MAE	R^2
Volume	5.52	0.87
Thickness	6.50	0.80
VBM	5.65	0.86
All	3.86	0.93

predicted age for each subject in the training set are presented in the scatter plots in Fig. 4.

Evaluation of Gaussian process uncertainty as a measure of brain abnormality

In this section, we present results of our experiments comparing the use of prediction error ε, uncertainty $\mathbf{cov}(\widehat{\mathbf{y}})$ and age-weighted uncertainty $\mathbf{cov}_w(\widehat{\mathbf{y}})$ as a biomarker for differentiating between healthy controls and patients with MCI, AD or autism. We performed three different experiments: the first two experiments are targeted at finding differences between HC, MCI and AD groups in both the ADNI and OASIS databases; the third experiment is performed on the ABIDE II database where differences between autism and healthy groups are evaluated. Note that as summarized in tables 1 and 2, the datasets used for testing are different to those used for training. For all experiments we assessed differences between groups both by performing non-parametric Wilcoxon rank-sum tests (Mann and Whitney, 1947) and by measuring the classification performance in a per subject basis by generating Receiver Operating Characteristic (ROC) curves with their corresponding Area Under the Curve (AUC) values. For all experiments, results are shown for four different sets of features: volume, thickness and VBM, as well as for the combination of all three feature sets. Our proposed GPR uncertainty based metrics are compared to the prediction error ε, obtained in a similar fashion as previous work on age estimation (Franke et al., 2010). An appropriate age length scale parameter l_y for the $\mathbf{cov}_w(\widehat{\mathbf{y}})$ metric was set independently for each experiment by performing evaluations at different scales an keeping the best performing results (See Fig. 5).

Experiment 1: ADNI dataset

For our first experiment we measure the separation between HC, MCI and AD groups for images obtained from the ADNI database. Due to the very large dataset size of this testing scenario, all the p-values reported in Table 5 are statistically significant (p-value $< 6 \times 10^{-5}$). The reported AUC values in Table 6 and the ROC curves in Fig. 6 show consistently a better performance of the uncertainty based metrics $\mathbf{cov}(\widehat{\mathbf{y}})$ and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ with respect to ε. In general $\mathbf{cov}(\widehat{\mathbf{y}})$ and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ presented similar performance, but adding the age term resulted in larger AUC values for the experiments on volume and VBM features. Box plots for each feature set and each diagnostic group are also shown in Fig. 7. The results using uncertainty based measures $\mathbf{cov}(\widehat{\mathbf{y}})$ and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ were strongly correlated (R = 0.95). In contrast, correlations of 0.35 and 0.37 were obtained between ε-$\mathbf{cov}(\widehat{\mathbf{y}})$ and ε-$\mathbf{cov}_w(\widehat{\mathbf{y}})$, respectively.

Experiment 2: OASIS dataset

Our second experiment is similar to experiment 1, but our evaluation is performed on images obtained from the OASIS database. In order to ensure similar age ranges for the HC, MCI and AD groups, all individuals under 60 years were removed from the testing dataset. Tables 7 and 8 summarize the numerical results of the comparisons between HC-MCI and MCI-AD groups. Similar to previous results (Franke et al., 2010), we observed that prediction error ε is a useful biomarker in this particular dataset. According to the results in tables 7 and 8, ε presented larger AUC values and smaller p-values for the models trained using volume and thickness features when discriminating between HC and MCI groups.

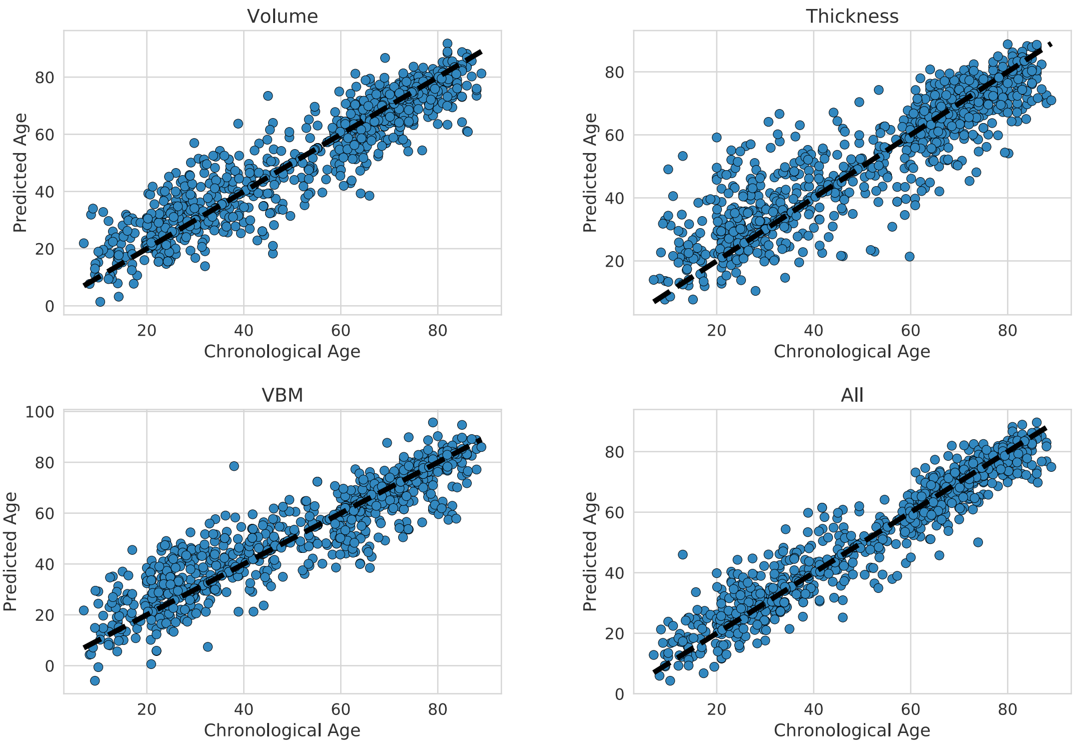


Fig. 4. Scatter plots showing the prediction results of the age prediction models trained using different feature sets.

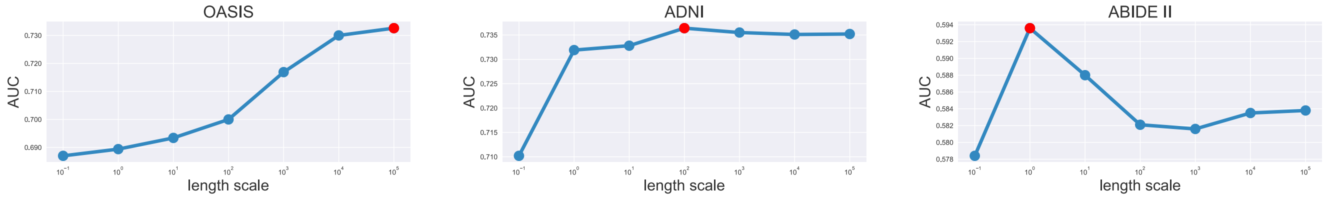


Fig. 5. AUC values obtained by using the $\mathbf{cov}_w(\widehat{\mathbf{y}})$ metric for different age length scales l_y. According to this curves, a different length scale was selected for each experiment (ADNI: $l_y = 1 \times 10^2$, OASIS: $l_y = 1 \times 10^5$, ABIDE: $l_y = 1$). The selected length scales are highlighted in red in each plot.

Table 5
p-values corresponding to the statistical tests performed on the experiments comparing the HC, MCI and AD groups on the ADNI dataset.

	HC-MCI			MCI-AD		
	ϵ	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$	ϵ	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$
Volume	6.82×10^{-19}	2.07×10^{-29}	1.85×10^{-31}1	3.36×10^{-60}	4.54×10^{-69}	2.05×10^{-74}
Thickness	2.18×10^{-18}	3.53×10^{-27}	2.92×10^{-24}	3.98×10^{-38}	6.03×10^{-105}	1.45×10^{-101}
VBM	3.10×10^{-07}	6.12×10^{-30}	1.20×10^{-34}	5.61×10^{-22}	2.92×10^{-77}	1.17×10^{-76}
All	5.63×10^{-05}	5.16×10^{-36}	5.10×10^{-37}	9.46×10^{-23}	2.76×10^{-103}	1.02×10^{-103}

However for all the rest of the evaluations on the OASIS database, $\mathbf{cov}(\widehat{\mathbf{y}})$ and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ presented the best performance amongst the evaluated metrics. Correlation coefficients between the metrics were 0.99 for $\mathbf{cov}(\widehat{\mathbf{y}})$-$\mathbf{cov}_w(\widehat{\mathbf{y}})$, 0.38 for ϵ-$\mathbf{cov}(\widehat{\mathbf{y}})$ and 0.38 for ϵ-$\mathbf{cov}_w(\widehat{\mathbf{y}})$. Notice that in this case the results for $\mathbf{cov}(\widehat{\mathbf{y}})$ and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ are strongly correlated due to the very large value assigned to the age length scale parameter l_y. The ROC curves in Fig. 8 and the box plots in Fig. 9 confirm these observations.

Table 6
Area Under the Curve (AUC) values corresponding to the statistical tests performed on the experiments comparing the HC, MCI and AD groups on the ADNI dataset.

	HC-MCI			MCI-AD		
	ϵ	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$	ϵ	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$
Volume	0.67	0.71	**0.72**	0.73	0.76	**0.77**
Thickness	0.66	**0.71**	**0.71**	0.69	**0.80**	**0.80**
VBM	0.60	0.71	**0.72**	0.64	**0.77**	**0.77**
All	0.59	**0.74**	**0.74**	0.64	**0.81**	**0.81**

Experiment 3: ABIDE II dataset

For our third experiment, we evaluate the age prediction on the ABIDE II dataset that contains subjects with autism. To the best of our knowledge, no previous age prediction based approach has been used for studying autism. However, previous studies have suggested abnormal brain development in patients diagnosed with autism Courchesne et al. (2001). By observing Figs 11 and 10 it is clear that differences between HC and MCI groups are considerably less noticeable than in the previous

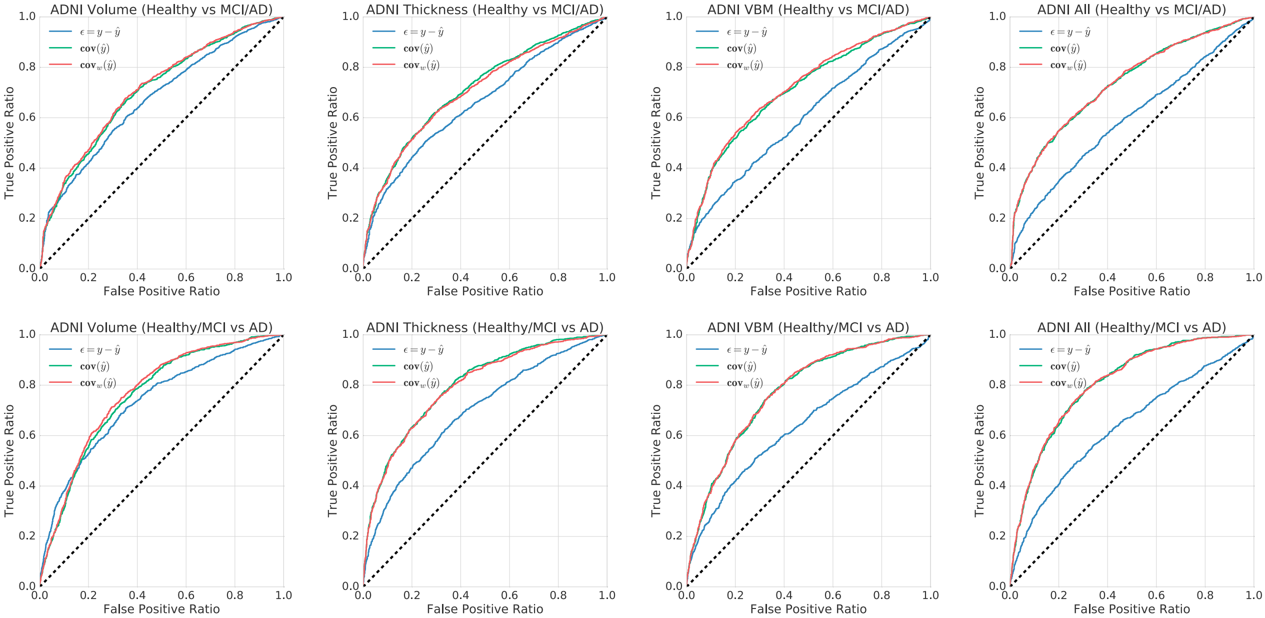


Fig. 6. Receiver Operating Characteristic (ROC) curves for the prediction of the presence of MCI/AD (Top) or the presence of AD (Bottom) evaluated on the ADNI dataset. Columns correspond to the different evaluated features.

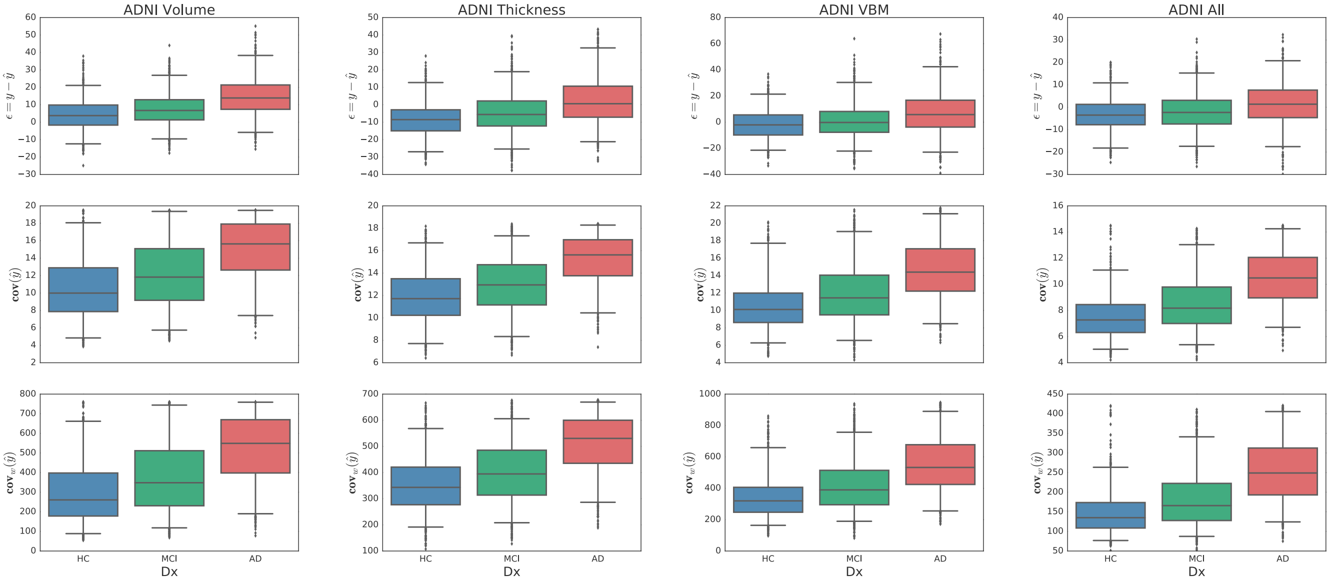


Fig. 7. Box plots showing prediction results for ε (top), $\mathbf{cov}(\hat{\mathbf{y}})$ (middle) and $\mathbf{cov}_w(\hat{\mathbf{y}})$ (bottom) for HC, MCI and AD groups on the ADNI dataset. Columns correspond to the different evaluated features.

two experiments. In fact, by analyzing the AUC results in Table 10 and the p-values in Table 9, we can observe that no significant differences between groups were found using the standard prediction error approach using ε as a predictive variable. In contrast, both uncertainty based measurements showed significant differences between HC and autistic groups. Also different to the previous two experiments, the weighted uncertainty based measurement $\mathbf{cov}_w(\hat{\mathbf{y}})$ showed a better performance than the standard uncertainty $\mathbf{cov}(\hat{\mathbf{y}})$. In this case correlation coefficients between the metrics were 0.64 for $\mathbf{cov}(\hat{\mathbf{y}})$-$\mathbf{cov}_w(\hat{\mathbf{y}})$, 0.30 for ε-$\mathbf{cov}(\hat{\mathbf{y}})$ and 0.49 for ε-$\mathbf{cov}_w(\hat{\mathbf{y}})$. The lower correlation between $\mathbf{cov}(\hat{\mathbf{y}})$-$\mathbf{cov}_w(\hat{\mathbf{y}})$ and the larger correlation between ε-$\mathbf{cov}_w(\hat{\mathbf{y}})$ when compared to the two previous experiments are caused by the smaller value of l_y.

Discussion

In this work, we have proposed to use uncertainty in GPR as a measure of neuropathology. In contrast to previous work based on the prediction error, which assumes similar trajectories between aging and disease processes, the GPR uncertainty handles differences in morphology of diseased brains that do not necessarily lie on a healthy

Table 7
p-values corresponding to the statistical tests performed on the experiments comparing the HC, MCI and AD groups on the OASIS dataset. The highlighted values correspond to p values with significance levels under 0.05 (light background), 0.01 (middle background) and 0.001 (dark background).

	HC-MCI			MCI-AD		
	ε	$\mathbf{cov}(\hat{\mathbf{y}})$	$\mathbf{cov}_w(\hat{\mathbf{y}})$	ε	$\mathbf{cov}(\hat{\mathbf{y}})$	$\mathbf{cov}_w(\hat{\mathbf{y}})$
Volume	<0.0001	<0.0001	<0.0001	0.0611	0.0065	0.0067
Thickness	0.0005	0.1186	0.1123	0.0707	0.0002	0.0002
VBM	0.0075	<0.0001	<0.0001	0.0225	0.0003	0.0003
All	0.0015	<0.0001	<0.0001	0.0707	0.0020	0.0020

Table 8
Area Under the Curve (AUC) values corresponding to the statistical tests performed on the experiments comparing the HC, MCI and AD groups on the OASIS dataset.

	HC-MCI			MCI-AD		
	ε	$\mathbf{cov}(\hat{\mathbf{y}})$	$\mathbf{cov}_w(\hat{\mathbf{y}})$	ε	$\mathbf{cov}(\hat{\mathbf{y}})$	$\mathbf{cov}_w(\hat{\mathbf{y}})$
Volume	**0.77**	0.73	0.73	0.73	**0.75**	**0.75**
Thickness	**0.68**	0.62	0.62	0.68	**0.76**	**0.76**
VBM	0.64	**0.72**	**0.72**	0.68	**0.77**	**0.77**
All	0.67	**0.74**	**0.74**	0.68	**0.77**	**0.77**

aging trajectory. If we consider predicted age as an aging biomarker, GPR uncertainty can be seen as a measure of uncertainty of the aging biomarker. We have evaluated the ability of GPR uncertainty to discriminate subjects with pathology for two very different diseases: Alzheimer's disease where we worked with a cohort of advanced age individuals, and autism where we operated on a younger cohort. To the best of our knowledge, this is the first time that uncertainty in a GPR model is used as a measure of neuropathology and it is also the first application of an age regression model for autism. For both applications, we work with a single model that was trained on healthy subjects from a wide age range. This distinguishes our work from discriminative approaches, which require the inclusion of images of patients diagnosed with a particular disease in the training set.

In this work, we build age prediction models using three different types of brain features: VBM, volume and thickness as well as a combination of all of them. Based on our results, we have observed that none of the features outperformed the others across all our evaluations. However, combining all features resulted on a model with the lowest prediction error and consistently achieved the best results when performing separation between healthy and disease groups. This is in line with previous work (Valizadeh et al., 2017; Liem et al., 2017), where it has been observed that extended feature sets which give models a larger variety of measurements to base the prediction on, result in more accurate age estimation models. Although the main goal of this paper is not to present a state-of-the-art age prediction method, we have observed that our proposed GPR model has a high prediction accuracy, comparable to that of current age estimation approaches (Valizadeh et al., 2017; Cole et al., 2016).

We have also demonstrated the generalization ability of our method by training and testing our model in completely independent datasets. Our training dataset was built based on the IXI, ABIDE and AIBL databases while testing was performed on the OASIS, ADNI and ABIDE II databases. Using different datasets for training and testing complicates the age prediction problem, as undesired dataset biases can impact the result (Wachinger and Reuter, 2016; Gutiérrez et al., 2017). However, such experiments model a scenario that is more realistic, as the translation to the clinic requires the accurate deployment of our method on data that differs from the training set.

Based on GPR uncertainty, we have introduced two metrics to assess the similarity of a test subject to a model of healthy aging: the uncertainty of the predictions of the GPR $\mathbf{cov}(\hat{\mathbf{y}})$ and an age-weighted uncertainty measurement $\mathbf{cov}_w(\hat{\mathbf{y}})$. We have shown in our experiments in section 3 that both measures find statistically significant differences between HC, MCI and AD groups as well as between autism and HC groups. We have compared these results to the commonly used prediction error ε, and we have shown that the proposed metrics yield a better separation between groups. The age-weighted uncertainty measurement can be seen as an extension to the standard uncertainty measure, with the inclusion of a weighting parameter based on the chronological age of the test subject. The effect of this weighting is controlled by the age-length scale parameter l_y. We have analyzed the effect of l_y in the performance of the

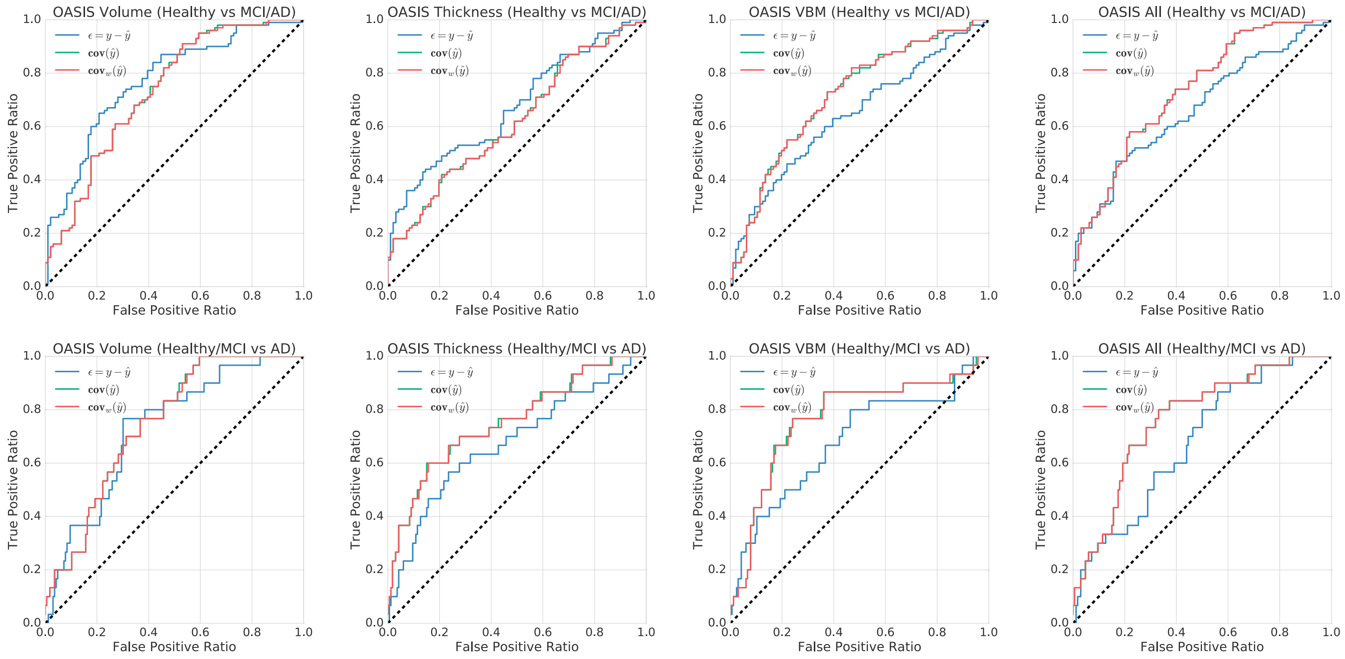


Fig. 8. Receiver Operating Characteristic (ROC) curves for the prediction of the presence of MCI/AD (Top) or the presence of AD (Bottom) evaluated on the OASIS dataset. Columns correspond to the different evaluated features.

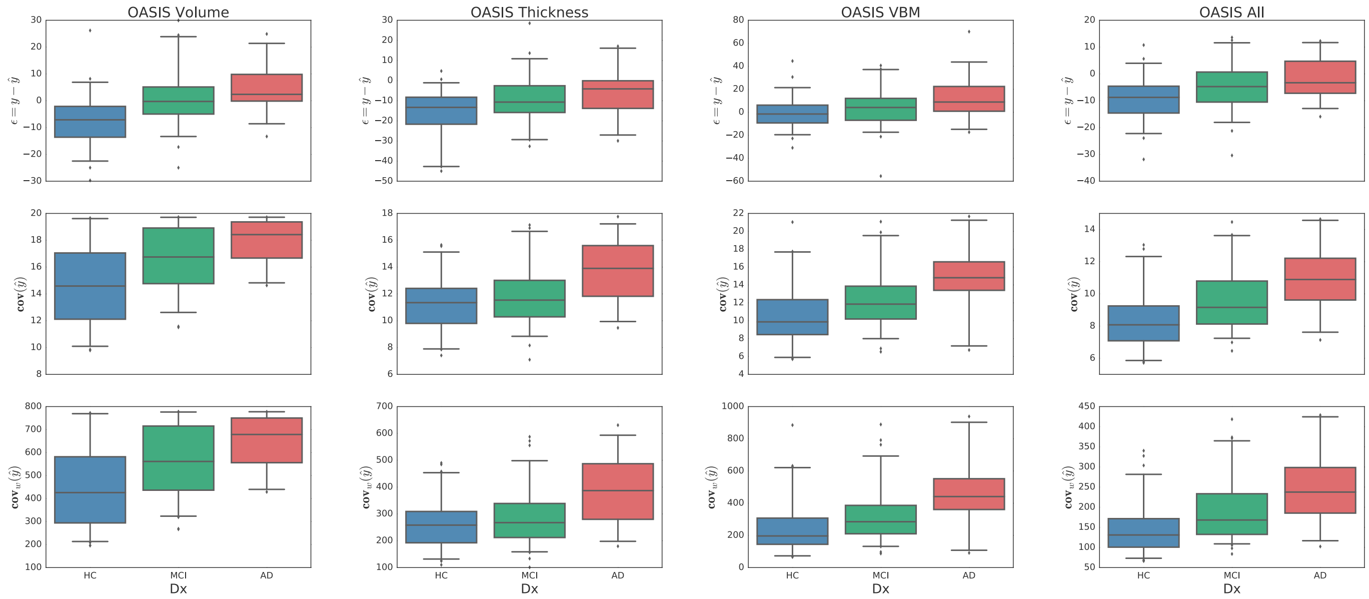


Fig. 9. Box plots showing prediction results for ε (top), $\mathbf{cov}(\hat{\mathbf{y}})$ (middle) and $\mathbf{cov}_w(\hat{\mathbf{y}})$ (bottom) for HC, MCI and AD groups on the OASIS dataset. Columns correspond to the different evaluated features.

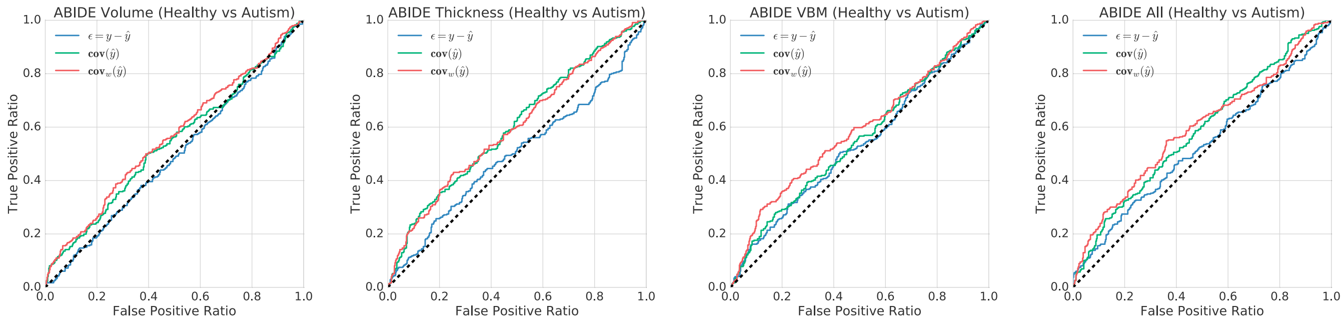


Fig. 10. Receiver Operating Characteristic (ROC) curves for the prediction of the presence of Autism. Columns correspond to the different evaluated features.

age-weighted uncertainty measurement and we have observed that although the use of the age-weighting had a limited effect in the case of the MCI/AD experiments there was a clear improvement in the case of autism. We hypothesize that these differences in performance of $\mathbf{cov}_w(\hat{\mathbf{y}})$ can be attributed to the different age ranges of the testing cohorts, since age prediction models present smaller prediction errors when testing on younger cohorts compared to the prediction error presented on datasets consisting of older individuals (Cole and Franke, 2017).

For the experiment on autism, our proposed uncertainty based metrics showed its ability to discriminate between autistic and healthy groups. We find these results particularly encouraging since the prediction error based approach showed to be insufficient to find differences for this particular disease. Given the analysis performed in section 3.1 and the results of our experiments, we believe that the main reason of the better performance of our uncertainty based measures is that they do not model brain anomaly as an accelerated aging process, but rather as deviations from healthy aging. As discussed before, the complex effects of aging and disease follow trajectories that affect different areas of the brain at different rates. The more relaxed assumptions, which our proposed uncertainty based measures are based on, are therefore better suited to account for the complex impact of aging and disease across the entire brain.

We have not performed direct quantitative comparisons between our uncertainty based measures and discriminative approaches. The main reason behind this is that discriminative approaches require training images not only from healthy individuals but also from patients. This means that separate models have to be trained for each specific disease. In contrast, age-prediction based models are only built on images from healthy individuals. This allows to have a flexible model which can be used for different diseases without any need to retrain or adjust the model. We demonstrated this in our experiments, where we used the same age prediction model to predict brain anomaly both on patients with Alzheimer's disease and patients with autism.

Conclusions

We introduced the prediction uncertainty in age estimation as a measure of neuropathology, based on a multivariate age prediction model based on Gaussian process regression. Our measure does not make a priori specific assumptions about the nature of the changes caused by disease, but rather models these changes as deviations from healthy aging. The method is therefore not limited to a specific pathology or age

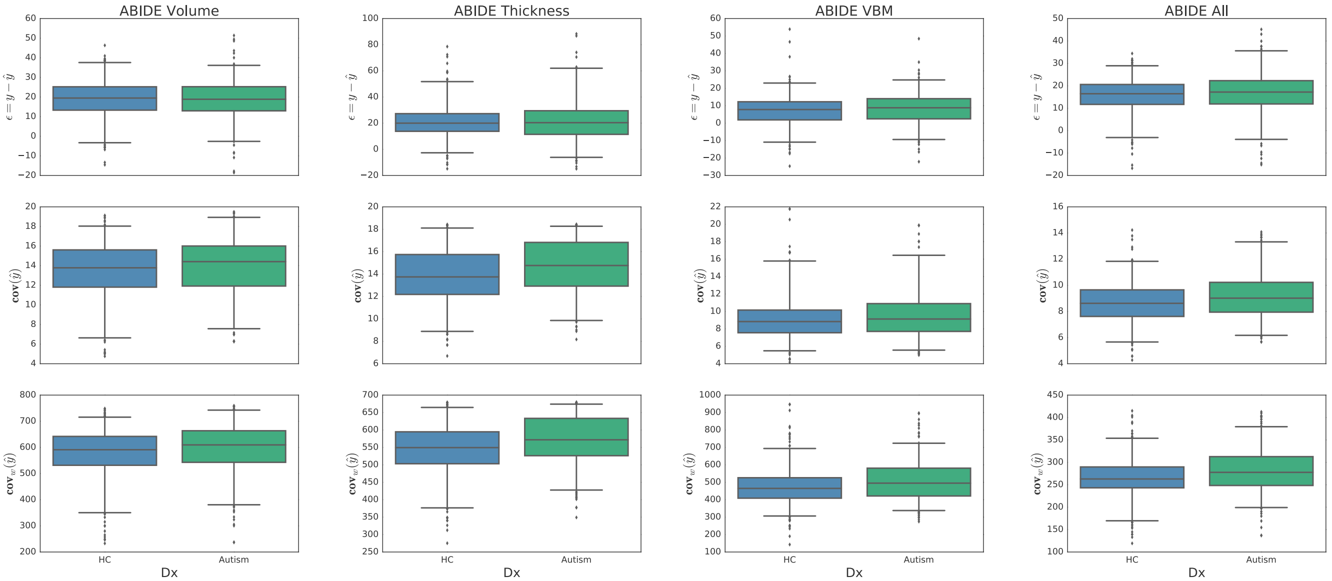


Fig. 11. Box plots showing prediction results for ε (top), $\mathbf{cov}(\widehat{\mathbf{y}})$ (middle) and $\mathbf{cov}_w(\widehat{\mathbf{y}})$ (bottom) for HC and Autism groups on the ABIDE II dataset. Columns correspond to the different evaluated features.

Table 9
p-values corresponding to the statistical tests performed on the experiments comparing the HC, and Autism groups. The highlighted values correspond to p values with significance levels under 0.05 (light background), 0.01 (middle background) and 0.001 (dark background).

	ε	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$
Volume	0.3210	<0.0317	<0.0019
Thickness	0.4664	<0.0001	<0.0001
VBM	0.0556	0.0089	<0.0001
All	0.1060	<0.0001	<0.0001

Table 10
Area Under the Curve (AUC) values corresponding to the statistical tests performed on the experiments comparing the HC, and Autism groups on the ABIDE II dataset.

	ε	$\mathbf{cov}(\widehat{\mathbf{y}})$	$\mathbf{cov}_w(\widehat{\mathbf{y}})$
Volume	0.50	0.53	**0.58**
Thickness	0.50	**0.60**	**0.60**
VBM	0.53	0.55	**0.60**
All	0.53	0.58	**0.60**

range, as demonstrated in our experiments on patients with Alzheimer's disease and patients with autism. Our method is also flexible to work with different sets of features, as we have illustrated in our experiments using volume, thickness, and VBM features. We have introduced an extension of the Gaussian process uncertainty measure for age estimation that also takes the chronological age into account, resulting in a weighted uncertainty measure, and we have demonstrated that the inclusion of this weighted measure can potentially be helpful for some applications. In comparison to the commonly used prediction error, the prediction uncertainty yielded an improved separation of diagnostic groups across all feature types and for different applications. It is also important to point out that in contrast to discriminative approaches, age prediction based models only require images of healthy individuals for training, which may allow for incorporating scans from large population-based studies in the future. The results presented in this paper encourage us to further explore the potential of uncertainty based measures and to apply our method to different diseases or conditions that might have complex effects in the anatomy of the brain. We are further interested in investigating the relationship between the prediction uncertainty and cognitive and clinical characteristics, as well as, future health outcomes.

Acknowledgements

This work was partly funded by SAP SE, the Bavarian State Ministry of Education, Science and the Arts in the framework of the Center Digitisation.Bavaria (ZD.B) and the Faculty of Medicine of the University Munich (Förderprogramm für Forschung & Lehre). Data collection and sharing for this project was funded by the Alzheimer's DiseaseNeuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing

(AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

References

Buckner, R.L., 2004. Memory and executive function in aging and ad. Neuron 44 (1), 195–208.

Cole, J.H., Franke, K., 2017. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends Neurosci.

Cole, J.H., Leech, R., Sharp, D.J., 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann Neurol. 77 (4), 571–581.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G., 2016. Predicting Brain Age with Deep Learning from Raw Imaging Data Results in a Reliable and Heritable Biomarker. http://arxiv.org/abs/1612.02572.

Courchesne, E., Karns, C., Davis, H., Ziccardi, R., Carper, R., Tigue, Z., Chisum, H., Moses, P., Pierce, K., Lord, C., et al., 2001. Unusual brain growth patterns in early life in patients with autistic disorder an mri study. Neurology 57 (2), 245–254.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatr. 19 (6), 659.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al., 2009. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatr. 21 (4), 672–687.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

Fjell, A.M., Westlye, L.T., Grydeland, H., Amlien, I., Reinvang, I., Raz, N., Holland, D., Dale, A.M., Walhovd, K.B., Neuroimaging, D., 2014. Critical Ages in the Life-course of the Adult Brain: Nonlinear Subcortical Aging, 34, pp. 2239–2247 (10).

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50 (3), 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PLoS One 8 (6).

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 14 (1), 21–36. http://linkinghub.elsevier.com/retrieve/pii/S1053811901907864.

Gutiérrez, B., Peter, L., Klein, T., Wachinger, C., 2017. A multi-armed bandit to smartly select a training set from big medical data. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 38–45.

Guttmann, C., Jolesz, F.A., Kikinis, R., Killiany, R.J., Moss, M.B., Sandor, T., Albert, M.S., 1998. White matter changes with normal aging. Neurology 50 (4), 972–978. http://www.neurology.org/cgi/doi/10.1212/WNL.50.4.972.

Habes, M., Janowitz, D., Erus, G., Toledo, J.B., Resnick, S.M., Doshi, J., Van der Auwera, S., Wittfeld, K., Hegenscheid, K., Hosten, N., Biffar, R., Homuth, G., Völzke, H., Grabe, H.J., Hoffmann, W., Davatzikos, C., 2016. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. Transl. Psychiatry 6, e775 (February). http://www.ncbi.nlm.nih.gov/pubmed/27045845.

Hedden, T., Gabrieli, J.D., 2004. Insights into the ageing mind: a view from cognitive neuroscience. Nat. Rev. Neurosci. 5 (2), 87–96.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C., Verma, R., 2014. Sex differences in the structural connectome of the human brain. Proc. Natl. Acad. Sci. 111 (2), 823–828.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, Gunnar, Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's disease neuroimaging initiative (adni): mri methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Kondo, C., Ito, K., Wu, Kai, Sato, K., Taki, Y., Fukuda, H., Aoki, T., 2015. An age estimation method using brain local features for T1-weighted images. 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 666–669. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7318450.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., et al., 2013. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr. Bull. 40 (5), 1140–1153.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., et al., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 50–60.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. Cognit. Neurosci. 19 (9), 1498–1507.

Nenadic, I., Dietzek, M., Langbein, K., Sauer, H., Gaser, C., 2017. BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. Psychiatry Res. Neuroimaging 266, 86–89 (March). http://linkinghub.elsevier.com/retrieve/pii/S0925492717301002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Potvin, O., Dieumegarde, L., Duchesne, S., 2017. Normative morphometric data for cerebral cortical areas over the lifetime of the adult human brain. Neuroimage. http://linkinghub.elsevier.com/retrieve/pii/S1053811917304135.

Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., Stern, Y., 2016. Differences between chronological and brain age are related to education and self-reported physical activity. Neurobiol. Aging 40, 138–144.

Valizadeh, S.A., Hanggi, J., Merillat, S., Jäncke, L., 2017. Age prediction on the basis of brain anatomical measures. Hum. Brain Mapp 38 (2), 997–1008.

Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M., 2015. Brainprint: a discriminative characterization of brain morphology. NeuroImage 109, 232–248.

Wachinger, C., Reuter, M., 2016. Domain adaptation for Alzheimer's disease diagnostics. Neuroimage 139, 470–479.

Wachinger, C., Salat, D.H., Weiner, M., Reuter, M., 2016. Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. Brain 139 (12), 3253–3266.

Wang, J., Li, W., Miao, W., Dai, D., Hua, J., He, H., 2014. Age estimation using cortical surface pattern combining thickness with curvatures. Med. Biol. Eng. Comput. 52 (4), 331–341.

Ziegler, G., Dahnke, R., Gaser, C., Initiative, A.D.N., et al., 2012. Models of the aging brain structure and individual decline. Front. Neuroinf. 6 (3).

Medical Image Analysis 41 (2017) 2–17

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Guiding multimodal registration with learned optimization updates

Benjamin Gutierrez-Becker[a,c,*], Diana Mateus[a], Loic Peter[a], Nassir Navab[a,b]

[a] Computer Aided Medical Procedures (CAMP), Technische Universität München, Boltzmanstr. 3 Garching, 85748, Germany
[b] Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA
[c] Department of Child and Adolescent Psychiatry, Psychosomatic and Psychotherapy, Ludwig-Maximilian-University, Waltherstr. 23. Munich, Germany

ARTICLE INFO

Article history:
Received 22 February 2017
Revised 1 May 2017
Accepted 3 May 2017
Available online 6 May 2017

Keywords:
Image registration
Machine learning
Multimodal registration
Motion estimation

ABSTRACT

In this paper, we address the multimodal registration problem from a novel perspective, aiming to predict the transformation aligning images directly from their visual appearance. We formulate the prediction as a supervised regression task, with joint image descriptors as input and the output are the parameters of the transformation that guide the moving image towards alignment. We model the joint local appearance with *context aware descriptors* that capture both local and global cues simultaneously in the two modalities, while the regression function is based on the gradient boosted trees method capable of handling the very large contextual feature space. The good properties of our predictions allow us to couple them with a simple gradient-based optimization for the final registration. Our approach can be applied to any transformation parametrization as well as a broad range of modality pairs. Our method learns the relationship between the intensity distributions of a pair of modalities by using prior knowledge in the form of a small training set of aligned image pairs (in the order of 1–5 in our experiments). We demonstrate the flexibility and generality of our method by evaluating its performance on a variety of multimodal imaging pairs obtained from two publicly available datasets, RIRE (brain MR, CT and PET) and IXI (brain MR). We also show results for the very challenging deformable registration of Intravascular Ultrasound and Histology images. In these experiments, our approach has a larger capture range when compared to other state-of-the-art methods, while improving registration accuracy in complex cases.

1. Introduction

Multimodal image registration is a fundamental task in medical image analysis, consisting in the alignment of two images of a given anatomical location acquired with different modalities. Multimodal registration is an important tool in clinical diagnosis, image-guided interventions, medical augmented reality, as well as in the validation of new imaging modalities (Markelj et al., 2012; Navab et al., 2016). In all these applications multimodal registration plays the key role of bringing and presenting complementary information in a spatially consistent way. In addition to the challenges of the monomodal case, multimodal registration has to deal with the potentially large appearance differences that result from each modality's acquisition principles. As the relation between the intensities from the two modalities is unknown and often neither linear nor bijective, an open question is the definition of a general energy function capable of relating the two modalities and guiding a multi-modal registration algorithm.

For instance, one common approach is to define similarity energy functions that map the appearance of both images to a scalar value (Fig. 2. left). Once the function is defined, the optimal spatial transformation between the images is computed maximizing the similarity. Under well-behaved energies (convex, smooth, *etc.*), the optimal transformation can be reached with simple gradient-based optimization algorithms, which compute iterative updates based on the energy gradient with respect to the transformation parameters (Fig. 2. right).

Unfortunately, explicitly defining a general and well-behaved energy function that models the unknown intensity relationship between the two modalities is not straightforward. Current multi-modal similarity standards based on information theory (Pluim et al., 2004), structural information (Heinrich et al., 2013; Wachinger and Navab, 2012) or metric learning (Michel et al., 2011; Simonovsky et al., 2016) rely on the strong assumption that the same structures are visible in both modalities (Fig. 3). In the latter case, such similarities do not have an analytical gradient nor guarantee the desired properties for an optimization energy. Therefore, their gradient-based optimization calls for local gradient

[*] Corresponding author at: Computer Aided Medical Procedures (CAMP), Technische Universität München, Boltzmanstr. 3 85748, Garching, Germany.

E-mail addresses: gutierrez.becker@tum.de (B. Gutierrez-Becker), mateus@in.tum.de (D. Mateus), peter@in.tum.de (L. Peter), navab@in.tum.de (N. Navab).

http://dx.doi.org/10.1016/j.media.2017.05.002

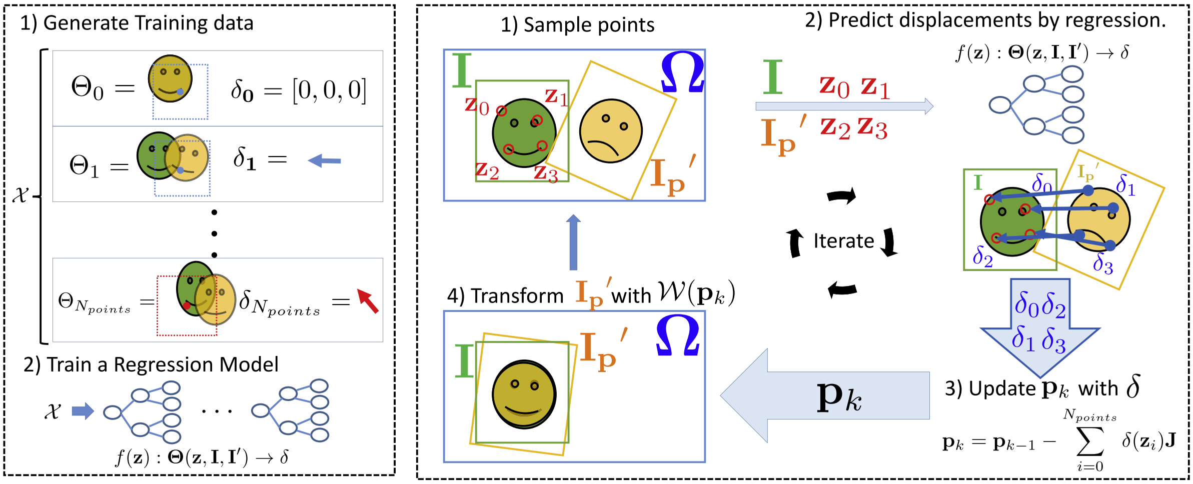

1) Generate Training data
$\Theta_0 =$
$\delta_0 = [0,0,0]$
$\Theta_1 =$
$\delta_1 =$
$\mathcal{X}$
$\Theta_{N_{points}} =$
$\delta_{N_{points}} =$
2) Train a Regression Model
$\mathcal{X}$
$f(\mathbf{z}) : \boldsymbol{\Theta}(\mathbf{z}, \mathbf{I}, \mathbf{I}') \rightarrow \delta$
1) Sample points
I
Ω
$\mathbf{z}_0$ $\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$
$\mathbf{I_p}'$
2) Predict displacements by regression.
$f(\mathbf{z}) : \boldsymbol{\Theta}(\mathbf{z}, \mathbf{I}, \mathbf{I}') \rightarrow \delta$
I $\mathbf{z}_0$ $\mathbf{z}_1$
$\mathbf{I_p}'$ $\mathbf{z}_2$ $\mathbf{z}_3$
Iterate
δ_0 δ_1 δ_2 δ_3
$\delta_0 \delta_2$
$\delta_1 \delta_3$
4) Transform $\mathbf{I_p}'$ with $\mathcal{W}(\mathbf{p}_k)$
$\mathbf{p}_k$
3) Update $\mathbf{p}_k$ with δ
$\mathbf{p}_k = \mathbf{p}_{k-1} - \sum_{i=0}^{N_{points}} \delta(\mathbf{z}_i)\mathbf{J}$

Fig. 1. Method Overview. **Training stage (left)**: A set of aligned multimodal images is used to generate a training set of images with known transformations. From this training set we train an ensemble of trees mapping the joint appearance of the images to displacement vectors. **Testing stage (right)**: We register a pair of multimodal images by predicting with our trained ensemble the required displacements δ for alignment at different locations **z**. The predicted displacements are then used to devise the updates of the transformation parameters to be applied to the moving image. The procedure is repeated until convergence is achieved.

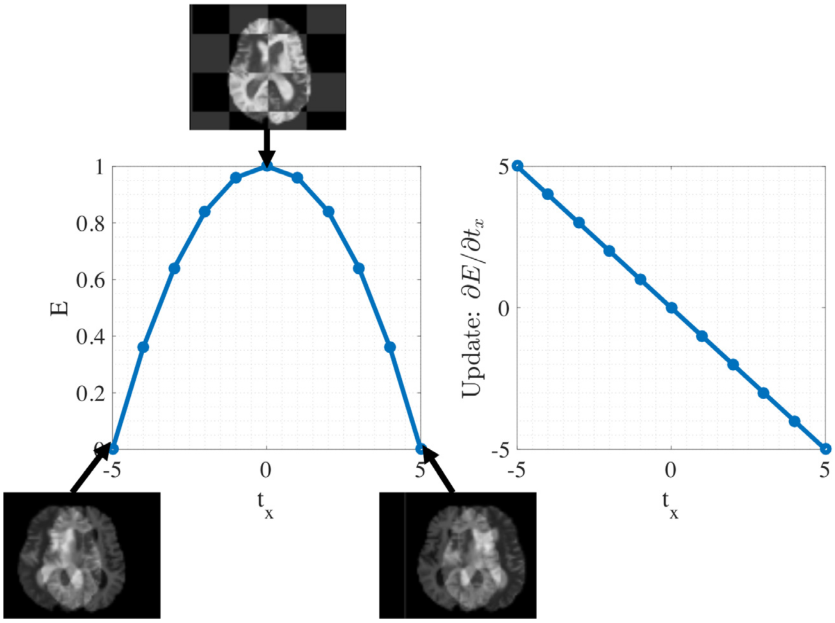


Fig. 2. Exemplary energy function *E*. **Left**: Continuous, convex and smooth behavior of *E* w.r.t. a transformation parameter. **Right**: Parameter update obtained by obtaining the derivative of the energy function with respect to a transformation parameter.

approximations or gradient free methods, which require advanced updates rules and an increased number of evaluations of the similarity metric.

In this work, we design a multimodal energy function that: i) is general, since it can create models capturing complex relationships between a wide range of modality pairs by using a small set of aligned examples, ii) can model such relationships based on global *and* local appearance, iii) can be easily optimized using a gradient-based method, and iv) that adapts to different transformation parameterizations. We model multimodal registration as a supervised regression problem, where given a pair of misaligned images we predict updates of the transformation parameters towards the correct alignment (*c.f* Fig. 1).

The joint appearance of the images is represented via a multimodal version of the Haar-like features (Criminisi et al., 2009) extracted from a sampling grid, which allows describing both the local and global-range context of each point. The regression task is formalized with gradient boosted trees, capable of handling the very high-dimensional Haar-like feature space, as well as of accurately approximating the transformation updates.

Our work is to the best of our knowledge, the first approach aiming at learning functions that map multimodal appearance to motion predictions, and showing how to effectively integrate them into a simple optimization scheme.

This paper is based on our previous work (Gutiérrez-Becker et al., 2016) but includes several extensions. First, we modify the method in order to predict not only the optimal update direction, but also the magnitude of the update vector in each iteration of the gradient-based optimizer. Second, we replace the regression model from random forest to gradient boosted trees (Friedman, 2001). We show how these two modifications lead to faster convergence times during testing as well as to an accurate registration. In addition, we include an evaluation of the improved properties of our method in terms of convergence and its training requirements using the IXI dataset. To demonstrate the generality of our method, we also extended our experiments to the publicly avail-

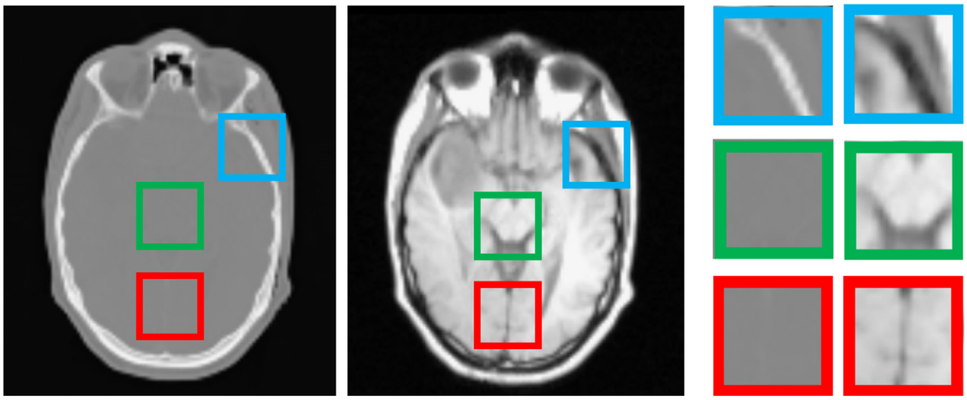


Fig. 3. Corresponding CT (**left**) and MR-T1 (**middle**) images of the brain obtained from the RIRE dataset. The highlighted regions are corresponding areas between both images (**right**). Some multimodal similarity metrics rely on structural similarities between images obtained using different modalities, like the ones inside the blue boxes. However in many cases structures which are clearly visible in one imaging modality correspond to regions with homogeneous voxel values in the other modality (red and green boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

able and widely used RIRE dataset (West, et al., 1997) for the rigid registration of three additional modalities: Computed Tomography (CT), Magnetic Resonance (MR) as well as Positron Emission Tomography (PET). We performed quantitative comparisons on the convergence of the proposed energy with a baseline (Mattes et al., 2001) and a state-of-the-art method (Heinrich et al., 2013).

2. Related work

Borrowing the classification of Sotiras et al. (2013), previous approaches to multimodal-registration fall in one among three categories. The first category comprises the *information-theoretic (IT)* methods like mutual information (Maes et al., 1997) and its variations (Wells et al., 1996; Mattes et al., 2001), which are arguably the most widely used methods given their simple implementation and their effectiveness to register different modalities (Pluim et al., 2003). Assuming that a global mapping between the intensities of the two modalities exists, such methods look for the transformation that maximizes the information of the intensity distributions. However, they are typically non-convex and suffer from the discrete approximations of the densities. Furthermore, IT methods suffer from a limited capture range and thus require a good initial transformation in order to converge.

A second family of approaches seeks to reduce the multimodal registration problem to a monomodal one. This can be done by synthesizing one modality from the other (Wein et al., 2007; Coupé et al., 2012) or by building an intermediate representation common to the two modalities (Wachinger and Navab, 2012; Heinrich et al., 2013; 2012).

Learning has also been used for both synthesis (Van Nguyen et al., 2015) or to build intermediate representations (Oktay et al., 2015). These methods have been shown to achieve lower registration errors compared to information theoretic approaches in a variety of applications (Sotiras et al., 2013). However, they are usually designed to register a specific pair of modalities or rely on strong structural similarities between the modalities to be registered.

The third category corresponds to *similarity learning* approaches that leverage on *a priori* information in the form of a training set of aligned examples. Among these, *Generative* approaches approximate the joint intensity distribution of the images and minimize the difference of a new test pair of images to the learned distribution (Sabuncu and Ramadge, 2008), possibly in a Bayesian Framework (Zöllei and Wells, 2006). *Discriminative* methods, on the other hand, model the similarity learning problem as the classification of positive (aligned) and negative (misaligned) examples, discrimination typically done at the patch level (Jiang et al., 2008; Lee et al., 2009; Michel et al., 2011). Different strategies have been explored to approximate such similarities, including margin-based approaches (Lee et al., 2009), boosting (Michel et al., 2011) and most recently, deep learning (Simonovsky et al., 2016; Cheng et al., 2016). In contrast to the discriminative approaches above which aim at discerning between aligned and misaligned patches, we focus on regressing a motion predictor that guides the registration process towards alignment.

In the Computer Vision community, prior work has used motion predictions for monomodal tracking and pose-estimation. Jurie and Dhome (2002) proposed a linear predictor for template tracking, which relates the difference between the compared images to variations in template position. Dollár et al. (2010) introduced a cascaded regression approach to learn a mapping from image features to object pose parameters. The cascaded approach reduces the parameter error progressively by means of an ensemble of boosted regressors (ferns) that re-computes the features at each iteration. Xiong and De la Torre (2013) provides a generalization of the cascaded method of Dollar *et al.* to solving non-linear least squared problems via a supervised descent method in the context of face alignment. In practice, Xiong and De la Torre (2013) implements the supervised descent approach as a sequence of linear regressors that link the differences in appearance (SIFT descriptors) to the distance between landmarks. In this paper, we formulate the multimodal registration problem in terms of a quadratic alignment error between the two images. We optimize this function iteratively using a gradient-based scheme. Similar to Xiong and De la Torre (2013), we learn to predict the parameter updates at each iteration, although with gradient-boosting trees instead of linear regressors in order to be able to handle the higher dimensionality and larger complexity of the multi-modal task. This means that a boosted sequence of prediction takes place at each iteration of the gradient optimization approach. Such two level regression approach also bears some similarities to the work of Cao et al. (2014), who use two levels of gradient boosting regression together with feature selection and sparse coding to regress the whole facial shape in a non-parametric manner.

In the context of registration of medical images, Chou et al. (2013) presented an approach for learning updates of the transformation parameters in the context of 2D-3D monomodal registration. Similarly, Kim et al. (2012), proposed the prediction of a deformation field for registration initialization, achieved by modeling the statistical correlation between image appearances and deformation fields with Support Vector Regression. Hu et al. (2016) proposed a regression model which can predict a deformation field given changes of appearance on monomodal images of the fetal brain. Similarly to Hu et al., our work uses motion prediction for registration but does it for the multimodal case. To the best of our knowledge, our work is the first approach aiming at predicting motion for the registration of medical multimodal images.

From a higher-level perspective, our work is also related to contemporary methods combining learning motion predictions with optimization methods, such as the approach of Ghesu et al. (2016) to predict the next search direction towards an anatomical landmark based on reinforcement learning or the approach by Yang et al. (2016) using a deep patch-wise network to predict mono-modal image deformations.

3. Background: gradient-based optimization

The simplest form of optimization for a smooth and unconstrained continuous energy is an iterative gradient-based search (Nocedal and Wright, 2006). Starting at iteration $k = 0$ and from an initial estimate $\mathbf{p}_k$, gradient-based optimization algorithms follow the next steps:

(i) Convergence test: if conditions satisfied stop and use $\mathbf{p}_k$ as the solution.
(ii) Compute the search direction vector $\mathbf{\Delta}_k \in \mathbb{R}^{N_p}$.
(iii) Compute the step length, a positive scalar $\boldsymbol{\alpha}_k$ such that $E(\mathbf{p}_k - \alpha_k \mathbf{\Delta}_k) < E(\mathbf{p}_k)$.
(iv) Update $k = k + 1$ and $\mathbf{p}_k = \mathbf{p}_{k-1} - \alpha \mathbf{\Delta}_k$, and go back to step (i).

The various algorithms mainly differ in the way to compute the search direction $\mathbf{\Delta}_k$ and the step size $\boldsymbol{\alpha}_k$. Usually, the update direction is set as the gradient of the energy function $\mathbf{\Delta}_k = \frac{\partial E}{\partial \mathbf{p}}$ at the current point $\mathbf{p}_k$.[1] The step size may be considered as a hyper-parameter or estimated with the help of heuristics or approximate optimizations. One such approximation known as Newton approach, takes into account the energy's second derivatives (the Hessian $\mathbf{H}(E)$) to determine the update:

$$\mathbf{p}_k = \mathbf{p}_{k-1} - \mathbf{H}(E)^{-1} \frac{\partial E}{\partial \mathbf{p}}. \tag{1}$$

Newton's improved convergence rates comes at the price of an increased computational cost, as the estimation of the inverse of the Hessian can be expensive and ill-conditioned, in particular, for high dimensional problems. Computing the Hessian can be avoided by using quasi-Newton methods which approximate the Hessian matrix using an update rule leading to a faster computation. However such approximations can require a higher number of iterations when compared to the full Newton method if the approximation of the Hessian is not accurate.

4. Method

Multimodal registration is the problem of finding the optimal transformation $\mathcal{W}(\mathbf{p})$ that brings into alignment a fixed image $\mathbf{I_f} : \Omega_f \subset \mathbb{R}^3 \to \mathbb{R}$ and a moving image $\mathbf{I_m}$: $\Omega_m \subset \mathbb{R}^3 \to \mathbb{R}$, each of a different modality. Let the transformation be described by a vector of parameters $\mathbf{p} \in \mathbb{R}^{N_p}$. Then, the problem is formalized as that of finding the optimal displacement vector $\mathbf{p}^*$ such that:

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} E(\mathbf{I}, \mathbf{I}'_{\mathbf{p}}), \tag{2}$$

where $\mathbf{I}'_{\mathbf{p}}$ stands for the moving image transformed to a joint domain $\Omega \subset \mathbb{R}^3$ by $\mathcal{W}(\mathbf{p})$, $\mathbf{I}$ is the fixed image also resampled in Ω, and E is an energy measuring the similarity between $\mathbf{I}$ and $\mathbf{I}'_{\mathbf{p}}$.

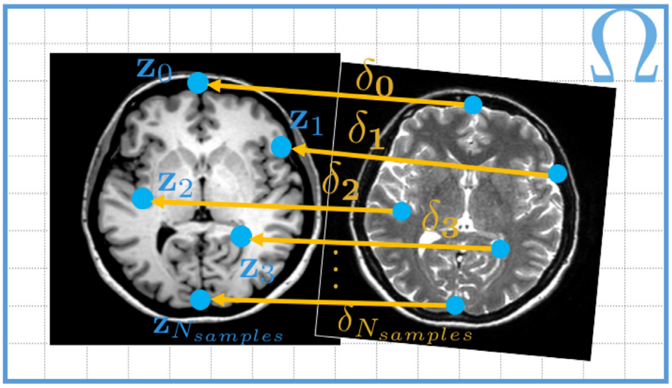


Fig. 4. The transformation $\mathcal{W}$ can be defined by the local displacements $\overrightarrow{\delta}(\mathbf{z}_i)$ at the grid positions $\mathbf{z}$. These local displacement vectors $\overrightarrow{\delta}(\mathbf{z}_i)$ point towards corresponding locations in the fixed and moving image.

In this work, we describe a multimodal energy E compatible with simple gradient-based optimization algorithms. The resultant updates, including search direction and step-size are effectively learned from a training set of aligned images. The problem is modeled as a supervised regression task. During the training phase, we learn to predict the search direction and step size given the local joint appearance of the two images. During the test phase, we aggregate local predictions towards a global parameter update. An overview of the method is presented in Fig. 1.

4.1. An optimization-aware energy for registration.

Without loss of generality,[2] we consider the transformation between the two multi-modal images as a discrete deformation field anchored to the elements of a set of control points $\{\mathbf{z}_i\}_{i=1}^{N_{\text{samples}}}$ on a joint domain Ω (see Fig 4). Formally, the deformation field is described by parameters $\mathbf{p} = [\overrightarrow{\delta}(\mathbf{z}_1), \ldots, \overrightarrow{\delta}(\mathbf{z}_i), \ldots, \overrightarrow{\delta}(\mathbf{z}_{N_{\text{samples}}})]^\top$, where each displacement $\overrightarrow{\delta}(\mathbf{z}_i)$ is a vector in $\mathbb{R}^3$ and the number of parameters equates that of the "control" points times three, *i.e.* $N_p = N_{\text{samples}} \times 3$.

The displacement field connects anatomical corresponding points $\mathbf{z}_i \sim \mathbf{z}'_i$ in the two images such that $\mathbf{z}'_i = \mathbf{z}_i + \overrightarrow{\delta}(\mathbf{z}_i)$, with $\{\mathbf{z}'_i\}_{i=1}^{N_{\text{samples}}}$. We then define the registration energy function as the sum of distances between the corresponding points $\sum_{\mathbf{z}_i \in \Omega} ||\mathbf{z}_i - \mathbf{z}'_i||^2$, or equivalently as the L2-norm of the displacement field:

$$E(\mathbf{I}, \mathbf{I}'_{\mathbf{p}}) = \frac{1}{2} \sum_{\mathbf{z}_i \in \Omega} ||\overrightarrow{\delta}(\mathbf{z}_i)||^2 \tag{3}$$

The energy in Eq. (3) is convex, has a smooth gradient, and leads to gradient-based parameter updates pointing towards the global minimum, and thus favors fast convergence. The global minimum is located at the transformation for which all the displacement updates $\overrightarrow{\delta}(\mathbf{z}_i) = 0$, which corresponds to a perfect alignment.

We can easily compute the energy gradient $\frac{\partial E}{\partial \mathbf{p}_i} = \frac{\partial E}{\partial \overrightarrow{\delta}} = \overrightarrow{\delta}(\mathbf{z}_i)$ as well as the Hessian given by the 3 × 3 identity matrix $\mathbf{H}(\mathbf{z}) = \mathbb{I}_3$. Leading to a Newton-like update (Eq. (1)):

$$\mathbf{p}_k = \mathbf{p}_{k-1} - \sum_{\mathbf{z}_i \in \Omega} \overrightarrow{\delta}(\mathbf{z}_i) \tag{4}$$

Our definition of E is so far based on the assumption that correspondences $\mathbf{z}_i \sim \mathbf{z}'_i$ are given. In the real registration setting,

[1] This is the update direction to find the minimum of the energy function. The same formulation can be used to maximize an energy function by using the update rule $\mathbf{p}_k + \alpha_k \mathbf{\Delta}_k$ minimizing E.

[2] In Section 4.2.4 we explain how to generalize the method to other parameterizations.

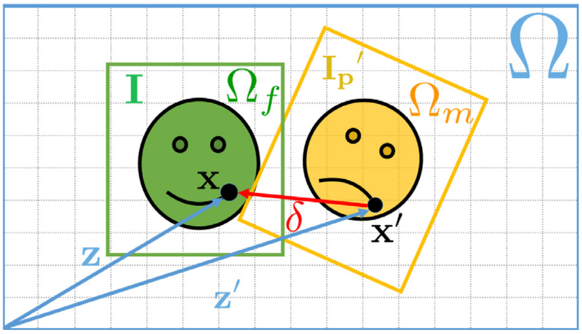


Fig. 5. Generating a displacement field. The vector $\vec{\delta}$ relates corresponding points $\mathbf{x} \sim \mathbf{x}'$ after they have been moved to locations $\mathbf{z} \sim \mathbf{z}'$ by applying transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}$. During training, vector $\vec{\delta}$ becomes the regression target required at location $\mathbf{z}$ to bring $\mathbf{I}'_\mathbf{p}$ into alignment with $\mathbf{I}$.

correspondences are unknown. However, instead of addressing the correspondence problem and explicitly defining E, we focus on predicting directly from the images the displacement field $\{\vec{\delta}(\mathbf{z}_i)\}$. We interpret this field in the context of a gradient-based optimization, as the next search direction and step-size of the parameter updates towards alignment. Such a formulation is independent of the image intensities of each modality and their relationship, while allowing for an iterative refinement of the predictions. Furthermore and as we show later, given an appropriate predictor, the behavior of the updates will be close to that of an ideal energy function.

4.2. Learning multimodal motion predictors

In Eq. (3), we model the registration energy as the squared sum of local offsets $\vec{\delta}$ between corresponding points in both images. In practice, since for a new pair of images these offsets are unknown, we estimate them by learning a regression function $f(\mathbf{z}) : \Theta(\mathbf{z}, \mathbf{I}, \mathbf{I}'_\mathbf{p}) \mapsto \vec{\delta}(\mathbf{z})$. The input to f is a feature vector $\Theta(\mathbf{z}, \mathbf{I}, \mathbf{I}'_\mathbf{p})$ describing the joint appearance of the point $\mathbf{z}$ in both modalities. Hereafter, we denote it $\Theta(\mathbf{z})$ for simplicity. In the following subsections we describe in details the different steps of our method:

(i) Creating a training dataset $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_\text{points}}$ from multimodal images under known misalignments (Section 4.2.1).
(ii) Defining a descriptor for the joint appearance features $\Theta(\mathbf{z})$ (Section 4.2.2).
(iii) Modeling and fitting the regression function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \vec{\delta}(\mathbf{z})$ (Section 4.2.3).
(iv) Generalizing the motion predictions to other transformation parameterizations (Section 4.2.4).
(v) Using predicted parameter updates to solve the multi-modal registration problem during test time (Section 4.2.5).

4.2.1. Generating the training set

We assume we are given prior knowledge about the relationship between the intensity distributions of the two modalities in the form of aligned image pairs. To generate the training set $\mathcal{X}$, we apply multiple known transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}_{j=1}^{N_\text{transfo}}$ to the aligned images, mapping the coordinates of two originally superposed points $\mathbf{x} \in \Omega_f$ and $\mathbf{x}' \in \Omega_m$ to distinct locations in a common image domain $\mathbf{z}, \mathbf{z}' \in \Omega \subset \mathbb{R}^3$ (see Fig. 5).

Because the applied transformations are known, we can determine the ground truth displacement $\vec{\delta}_n \in \mathbb{R}^3$ needed to find the originally corresponding point $\mathbf{z}'_n$ in the moving image, and bring it into alignment with $\mathbf{z}$, *i.e.* $\vec{\delta}_n = \mathbf{z}'_n - \mathbf{z}_n$. With this information and sampling N_points from the transformed images, we build a training set consisting of pairs of feature vectors Θ and their corresponding offset vector $\vec{\delta}$, *i.e.* $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_\text{points}}$. The process is illustrated in Fig. 6.

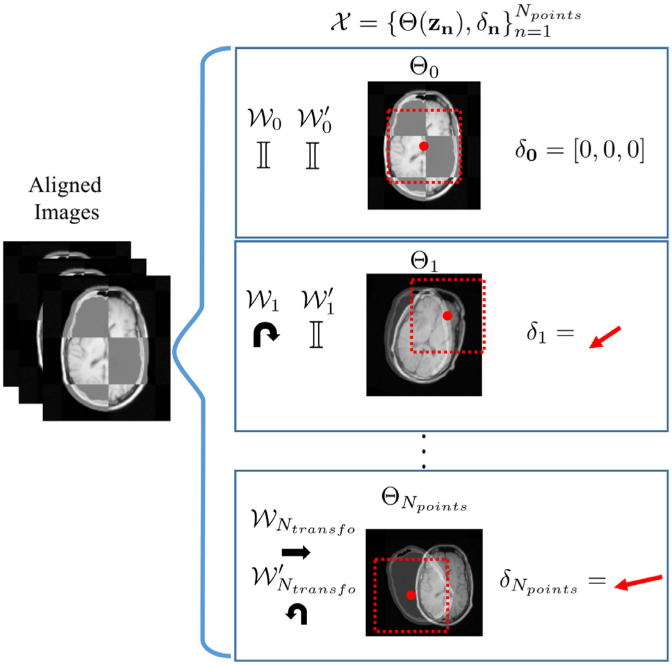

$\mathcal{X} = \{\Theta(\mathbf{z_n}), \delta_\mathbf{n}\}_{n=1}^{N_{points}}$
Aligned Images
Θ_0
$\mathcal{W}_0$ $\mathcal{W}'_0$
$\delta_\mathbf{0} = [0, 0, 0]$
Θ_1
$\mathcal{W}_1$ $\mathcal{W}'_1$
$\delta_1 =$
$\Theta_{N_{points}}$
$\mathcal{W}_{N_{transfo}}$
$\mathcal{W}'_{N_{transfo}}$
$\delta_{N_{points}} =$

Fig. 6. Generating the training dataset. Pairs of aligned images are generated and arbitrary known transformations are applied to them. The region surrounding a point, here depicted with dotted lines, is characterized using the feature vector Θ described in Section 4.2.2. To each feature vector we assign the displacement $\vec{\delta}$ required to bring the image at location $\mathbf{z}$ into alignment. The training set $\mathcal{X}$ is built from the collection of features and their corresponding displacements, *i.e.* $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_\text{points}}$.

4.2.2. Describing joint appearance with context-aware multimodal features.

We characterize the joint appearance of a pair of images around a location $\mathbf{z}$ by means of a feature vector $\Theta(\mathbf{z}) \in \mathbb{R}^H$. We model $\Theta(\mathbf{z})$ with a multi-modal adaptation to the context-aware Haar-like features (Criminisi et al., 2009). We use such rich high-dimensional descriptors to be able to encode the very large input space consisting of the joint local appearance of all image regions under all considered transformations.

The feature descriptor $\Theta(\mathbf{z})$ is built as a collection of H scalar features $[\theta_1, \ldots, \theta_h, \ldots, \theta_H]^\top$, where each θ_h is computed as an operation on a pair of boxes located at given offset locations relative to the point $\mathbf{z}$. More formally, θ_h is characterized by two boxes $\mathbf{b_1}$, $\mathbf{b_2}$ (*c.f* Fig. 7-left), parametrized by:

- Their *relative position* and *size* ($\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3$, $w_1, h_1, w_2, h_2, d_1, d_2 \in \mathbb{R}$) (*c.f* Fig. 7, top right). The position and size of the boxes are allowed to range from a couple of pixels to half of the size of the image. Using small boxes close to the sample location $\mathbf{z}$ allows the feature vector to accurately describe the local joint appearance around the point. Larger boxes and further positions instead capture the global context, which is important to perform registration when little or no overlap between images exist or when ambiguities can not be resolved using local appearance.
- The *modality* where the box operates $m = \{0, 1\}$ (*c.f* Fig. 7-middle-right). If m has the same value for both boxes we can capture the spatial context of each point within an image. If the value of m is different for each box, the feature is able to capture the functional relation across modalities.

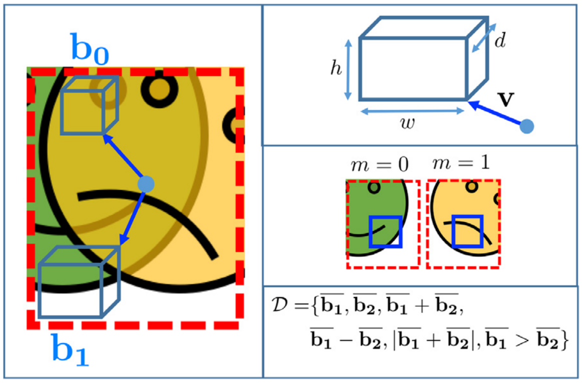

$\mathbf{b_0}$
$\mathbf{b_1}$
h
d
w
$\mathbf{v}$
$m = 0$
$m = 1$
$\mathcal{D} = \{\overline{\mathbf{b_1}}, \overline{\mathbf{b_2}}, \overline{\mathbf{b_1}} + \overline{\mathbf{b_2}},$
$\overline{\mathbf{b_1}} - \overline{\mathbf{b_2}}, |\overline{\mathbf{b_1}} + \overline{\mathbf{b_2}}|, \overline{\mathbf{b_1}} > \overline{\mathbf{b_2}}\}$

Fig. 7. Context-aware features. Each element of the feature vector $\Theta(\mathbf{z})$ is constructed by obtaining a pair of boxes at different positions relative to a location $\mathbf{z}$ (left). These boxes are described by its position and size (right, top), the modality that they describe (right, middle) and an operation between boxes (right, bottom).

- An operation *0* between boxes taken from the set : $\mathcal{D} = \{ \vec{\mu}\mathbf{b_1}, \vec{\mu}\mathbf{b_2},\ \vec{\mu}\mathbf{b_1} + \vec{\mu}\mathbf{b_2},\ \vec{\mu}\mathbf{b_1} - \vec{\mu}\mathbf{b_2}, |\vec{\mu}\mathbf{b_1} - \vec{\mu}\mathbf{b_2}|, \vec{\mu}\mathbf{b_1} > \vec{\mu}\mathbf{b_2}\}$, where the overline denotes the mean over the box intensities.

Considering the combinatorial nature of the parameters above, we face an infinite-dimensional feature space $\mathbb{R}^H$, which could be inefficient for learning. However, such high-dimensional feature spaces can be naturally handled by ensemble trees with axis-aligned splits, which enable individual features θ_h to be computed *on the fly* during training instead of precomputing the full vectors $\Theta(\mathbf{z})$. In addition, feature calculation is sped up by using precomputed integral volumes.

4.2.3. Displacement prediction with ensemble methods

In this subsection, we explain how to predict the offsets $\overrightarrow{\delta}(\mathbf{z}) \in \mathbb{R}^3$ from the features $\Theta(\mathbf{z})$ by approximating a function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \overrightarrow{\delta}(\mathbf{z})$. We model *f*(**z**) with an ensemble of regression trees, given their ability to handle high-dimensional feature spaces. A regression tree is a binary tree consisting of a set of nodes and leaves (Criminisi and Shotton, 2013). Each internal node splits the feature space into two parts according to an axis-aligned test function $g(\Theta(\mathbf{z}), h, T)$, where θ_h designates one of the dimensions of the feature vector $\Theta(\mathbf{z})$ and $T \in \mathbb{R}$ is a threshold. Given a subset of training samples $S \subset \mathcal{X}$ arriving to a given node, the split function creates a partition $S = \{S_L, S_R\}$, where S_L corresponds to the set $\theta_h < T$ and conversely, S_R to the set of features for which $\theta_h > T$. Finally, nodes without children are called leaves, and in the case of regression trees store a continuous value, *i.e.* $\overrightarrow{\delta}(\mathbf{z}) \in \mathbb{R}^3$.

During *training*, a set of labeled examples $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_{\text{points}}}$ is passed through all of the trees, and the parameters of the node splitting functions *h*, *T* are optimized to minimize the prediction error. The criteria used to determine the best split parameters is the minimization of the sample covariance:

$$\theta_h^*, T^* = \arg\min_{\theta_h, T} \operatorname{trace}(\mathbf{\Sigma}_{S_L}) + \operatorname{trace}(\mathbf{\Sigma}_{S_R}), \tag{5}$$

where $\mathbf{\Sigma}_{|S_{L,R}|}$ stands for the covariance matrix of the training offsets $\{\vec{\delta}_k\}_{k=1}^{|S_{L,R}|}$ of the features falling in each subset. Computing the trace instead of the full covariance matrix allows a faster computation of the splitting criteria. To preserve the generalization benefits of randomized splits over the forest the parameters are usually obtained through randomized node optimization. However, given the high dimensionality of $\Theta(\mathbf{z})$, we opt instead for the automatic scale selection strategy proposed by Peter et al. (2015), which enables us to optimize the value for the box parameters responsible for the position and scale. This choice has a positive impact on the performance of our method.

During *testing*, the prediction of the displacement $\vec{\delta}$ at a given location $\mathbf{z}$ is computed by passing the feature vector through the ensemble, and summing the individual tree predictions. The prediction at each node is performed independently, without explicitly taking into account the spatial position of each grid point.

We considered two approaches for ensemble tree regression. The first approach is a regression forest (RF), as presented in our previous work (Gutiérrez-Becker et al., 2016), where the predictions of the individual trees are combined through a simple average $f(\mathbf{z}) = \sum_{t=1}^{N_{\text{trees}}} \frac{1}{N_{\text{trees}}} \mathcal{T}_t(\mathbf{z})$. Here, each tree is independent of each other, allowing for their parallelization during both training and prediction.

The second regression approach is based on Gradient Boosted Trees (GBT), introduced by Friedman (2001). GBT has shown to have lower prediction errors when compared to general random forests when tuned correctly in a variety of scenarios (Caruana and Niculescu-Mizil, 2006).

The predictor *f* in GBT is a weighted sum of functions:

$$f(\mathbf{z}) = \sum_{t=1}^{N_{\text{trees}}} \beta \mathcal{T}_t(\mathbf{z}), \tag{6}$$

where each $\mathcal{T}_t$ corresponds to a regression tree and β is a scalar weighting each regression tree. However, Different to regression forests, where the training of each tree is independent, in GBT the function *f*(**z**) is built sequentially as:

$$f_t(\mathbf{z}) = f_{t-1}(\mathbf{z}) + \beta_t \mathcal{T}_t(\mathbf{z}). \tag{7}$$

At each stage *t*, a tree in GBT $\mathcal{T}_t$ minimizes the squared loss between the currently predicted displacement and the ground truth $||f_{t-1}(\mathbf{z}_n) - \vec{\delta}_n||^2$, instead of trying to recover the $\vec{\delta}_n$ directly. Apart from the change in the target value, each regression tree is trained as before finding the splits that reduce the sample covariance trace (*c.f* Eq. (5)). Even though GBT requires a sequential training and therefore individual trees can not be trained in parallel, the sequential aggregation allows for shallower trees when compared to the forests, leading also to comparable training times with lower prediction errors.

4.2.4. Generalizing to arbitrary transformations

Notice that so far we have chosen $\vec{\delta}_n$ as the regression targets instead of the transformation parameters. This choice is compatible with having the transformation parametrized as a displacement field. However, we now show that by the simple chain rule of derivatives, the results of Eq. (4) can be generalized to other types of transformation while keeping the learning stage independent of the parametrization. Indeed, using the chain rule the gradient of the energy may be split as $\frac{\partial E}{\partial \mathbf{p}} = \frac{\partial E}{\partial \vec{\delta}} \frac{\partial \vec{\delta}}{\partial \mathbf{p}}$, where $\frac{\partial E}{\partial \vec{\delta}}$ are the spatial derivatives and $\frac{\partial \vec{\delta}}{\partial \mathbf{p}}$ corresponds to a Jacobian relating the displacement to the transformation parameters which we denote hereafter $\mathbf{J}(\mathbf{z})$ for simplicity. The Jacobian is only dependent on the chosen parametrization and therefore does not change during the optimization. This means that we only require computing $\frac{\partial E}{\partial \vec{\delta}}$ at each iteration in order to retrieve the update direction. In the same way the Hessian of *E* will be computed as :

$$\mathbf{H} = \frac{\partial^2 E}{\partial \overrightarrow{\delta}^2} \frac{\partial^2 \overrightarrow{\delta}}{\partial \mathbf{p}^2} \tag{8}$$

4.2.5. Using multimodal motion predictors for registration

Once the regression function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \overrightarrow{\delta}(\mathbf{z})$ is trained, we use it to perform multimodal registration on a pair of previously unseen images $\mathbf{I_f}$ and $\mathbf{I_m}$. We follow a standard gradient-based optimization (*c.f* Section 3), where we calculate the search direction vector $\mathbf{\Delta}$ and the optimal step size $\boldsymbol{\alpha}$ at every iteration *k*. The iterative procedure is illustrated in Fig. 1. First, a set of testing points $\{\mathbf{z}_m\}_{m=1}^{N_{\text{test}}} \in \Omega$ is randomly sampled from the fixed image. We then extract the feature vectors for the point set $\{\Theta(\mathbf{z}_m)\}_{m=1}^{N_{\text{test}}}$ and pass them through the tree ensemble. The output of the ensemble are the predicted local displacement estimates $\{\hat{\vec{\delta}}_m\}_{m=1}^{N_{\text{test}}}$. We then compute the global update (*c.f* Eq. (4)) by adding the contribution of each local displacement to the transformation parameters $\hat{\mathbf{\Delta}} = \sum_{m=1}^{N_{\text{test}}} \hat{\vec{\delta}}_m J(z)$. Finally, a convergence test is performed and if necessary the procedure is repeated. In our case we stop the optimization when the difference of the energy function between iterations E falls below a threshold ϵ.

We have seen in Section 4.1 that with perfect displacement predictions, the Hessian estimate of the step length for our method is $\alpha = 1$. However, as we expect the predictions to have some error we reduce the step size by a factor λ, which we empirically evaluate.

5. Experiments and results

To evaluate the performance of our method under different scenarios we perform three series of experiments on different multimodal datasets.

Our *first* experiments rely on pairs of multi-modal MR images of the brain from the IXI dataset.[3] The focus is on studying the amount of data required for training the regression of the displacement field as well as on demonstrating the fast convergence of our registration approach.

The *second series of experiments* evaluates the performance of our method both in terms of registration accuracy and capture range for a variety of imaging modality pairs. To this end, we performed rigid registration on the publicly available RIRE dataset[4] (West, et al., 1997), consisting of images of adult brains obtained using different MR protocols as well as CT and PET. We evaluated our algorithm using *all* the modality pairs available in the RIRE database, which includes: CT-T1, CT-T2, CT-PD, PET-T1, PET-T2 and PET-PD pairs, showing the generality of our approach.

In the *third experiment*, we use our method for the deformable registration of two complex modalities: Intravascular Ultrasound images and histological slices (Katouzian et al., 2007) (see Fig. 16). This dataset is particularly challenging, first, because the images are noisy and have acquisition artifacts, and second, because the underlying assumptions of most similarity metrics, like local structural similarities between statistics on the intensities of the images, are not valid. During these experiments, we do a comparative evaluation of our method with respect to two other similarity metrics, namely Normalized Mutual Information (NMI) and the Self-Similarity Context descriptor (SSC) (Heinrich et al., 2013). We show that our learning based approach improves the results of multimodal registration in terms of accuracy and capture range. We also provide a detailed analysis of the properties of our method in terms of smoothness of the optimization updates and fast convergence. Finally, for all experiments, we also compare the behavior of our initial regression using random forest (LOU) and the new one based on gradient boosted trees (LOU2), where LOU stands for Learning Optimization Updates.

5.1. Implementation details

Our registration framework was implemented using the Insight Segmentation and Registration Toolkit (ITK).[5] For all our experiments we performed optimization using a simple gradient descent optimizer and the same parameterizations were used for all methods. In the case of NMI we used the Mattes Mutual Information Metric included on the ITK framework and in the case of SSC we adapted the implementation provided by the authors to our framework. In all cases the control points to evaluate the similarity metrics were sampled randomly, taking approximately a proportion of 0.1 of the total voxels in the image. All metrics were evaluated using the same number of control points to ensure a fair comparison. Interpolation between control points was performed with a b-spline interpolation. The size of the boxes and offsets for the Haar-like features was limited to a maximum of half the size of the image in each dimension.

Images were processed by performing histogram matching to a reference image. This was done in order to reduce the amount of possible intensity variations observed during training and testing.

5.2. Evaluation on the IXI dataset: convergence and amount of training data

Our first experimental setup is based on the IXI dataset, which contains T1, T2 and PD-weighted images of the brain from healthy subjects. We perform two different types of experiments. First, we evaluate the registration accuracy of our method given different training dataset sizes (*c.f* Section 5.2.1). Second, we study the convergence of our algorithm in terms of number of iterations required for convergence as well as for different step sizes $\boldsymbol{\alpha}$ (*c.f* Section 5.2.2).

For the purpose of these experiments, we extract a dataset consisting of pairs of corresponding T1-T2 images from 10 subjects. We pre-processed the images with skull-stripping and performed histogram matching to a reference image. We carefully selected pairs of images with little or no misalignment between the T1-T2 images. We further removed any residual alignment error by aligning manually placed landmarks in both images.

5.2.1. Dataset size

Here, we evaluate the number of aligned images required to build a regression model capable of performing accurate registrations.

Training: We split our dataset into two groups: 5 image pairs for training and 5 for testing. We then train 5 different regression models, each with an increasing number of training images. To each image pair, we apply a random transformation, sampled from a uniform distribution in the range of ± *size* for translations and ± 1 rad for rotations, where *size* corresponds to the size of the image. In total 1250 image pairs are generated for each modality pair and 10% of their voxels are taken at random for training.

Testing: We perform rigid registration using the 5 different regression models on the 5 images left out for testing. In order to assess the robustness of our algorithm to different initializations, we perform 30 registrations per image pair, each one at a different initial position for the moving image in a range between ± *size* for translations and ± 1 rad for rotations. We evaluate models created using both our previously presented method using random forests (LOU) and our new approach based on gradient boosted trees (LOU2). The results are shown in the box plots in Fig. 8. For reference, we perform registration on the same set of images using NMI and SSC as a similarity metric and using the same simple gradient descent optimizer and we plot the median registration error as a dotted line.

The box plots show that our method is able to accurately register the test image pairs under a large range of initializations. The median registration error was comparable to the error obtained using NMI and SSC, even when the number of training images was reduced to a single pair of aligned images. Including additional images into the training dataset helps our regression model to

[3] http://brain-development.org/ixi-dataset/.

[4] http://www.insight-journal.org/rire/.

[5] https://itk.org/.

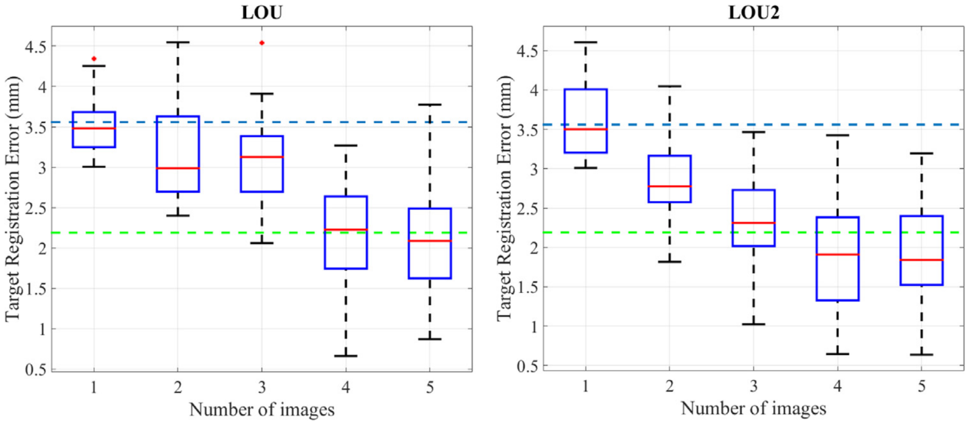


Fig. 8. Comparison of registration error for models trained using different dataset sizes for LOU and LOU2. The dotted blue line indicates the median registration error for registration in the same images using Normalized Mutual Information and the green line the median registration error for registration using SSC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reduce the final registration error even further. We also observe that LOU2 produces slightly lower registration errors when compared to LOU, and that, LOU2 is able to reduce the registration error using a lower number of training images. This can be attributed to the lower prediction error obtained using Gradient Boosted Trees as an ensemble technique when compared to Random Forests.

5.2.2. Convergence

For our second experiment with IXI, we take the models trained on 5 images (last box plot in Fig. 8-right) and perform rigid registration for a test pair of images with an arbitrary initial transformation. For clarity, we illustrate the behavior of the updates independently for different parameters. To this end, we perform two separate experiments. First, we initialize the moving image with a 50 mm translation offset in the axial direction. In the second experiment, we rotate the moving image around the axial axis by 0.5 rad for LOU and LOU2, but only by 0.25 rad for NMI given its smaller capture range. We show in Fig. 9 the evolution of the error over the iterations in each case and for both our methods and NMI using the same gradient-based optimizer.

In order to assess the influence of the step size α for each method, we present three curves with different step sizes. The step size shown in the red corresponds to the step size that presented the lowest final error after a line search in the parameter α.

We can draw some interesting observations from the results in Fig. 9. The path across the energy function for LOU and LOU2 is smoother and reaches a transformation close the global optimum in just a few iterations. In general LOU and LOU2 were able to find an accurate solution after no more than 10 iterations compared to the hundreds required using NMI.

The longer optimization time can be explained by the noisy approximations of the gradient of NMI. This noisy gradient forces the optimization algorithm to use a very small step size in order to ensure that the optimization converges to a solution close to the global optimum. For this reason, the convergence times for LOU ($\sim$ 10 s) and LOU2 ($\sim$ 5 s) were an order of magnitude faster when compared to NMI ($\sim$ 100 s).

5.3. Evaluation on the public dataset (RIRE)

In order to demonstrate the flexibility and generality of our multi-modal registration dataset, we train an independent regression model for each available modality pair in the publicly available RIRE dataset (CT-T1, CT-T2, CT-PD, PET-T1, PET-T2 and PET-PD). Only the single pre-aligned image pairs provided in the RIRE dataset are used for training.

We report the average Target Registration Error (TRE) as obtained from the online RIRE evaluation platform and compare the results for both our LOU and LOU2 methods with respect to Normalized Mutual Information (NMI) and the Self-Similarity Context descriptor (SSC) (Heinrich et al., 2013).

Training: We follow a similar procedure as for the IXI experiments. We again generate transformations using translations ranging from ± *size* and rotations from ± 1 rad. However, this time we consider the raw images without skull stripping. The only preprocessing step is histogram matching between the test images and the training image in order to account for differences in the dynamic range of the images.

Testing: For each testing image pair (in total 33 image pairs) in the dataset we perform rigid registration 30 times, each starting from a different initial misalignment of the moving image. This initial transformation is sampled at random from a uniform distribution in the range of ± 0.5**size* for translation in each of the axis and rotations of ± 0.5 rad.

Results of our evaluation are shown in Fig. 11. The box plots indicate the final registration error after convergence for the four compared methods and all combinations of image modality pairs. In the case of **CT-MR**, we observe that the final median registration error is comparable across the different methods. When the initialization is close to the optimum solution, SSC, LOU and LOU2 lead to comparable low registration error. However, NMI and SSC result more often in higher registration errors when the initial transformation is large. As such transformations are not covered by capture range of the algorithm, the optimizer converges to a local optimum. We performed Mann–Whitney *U* statistical tests (Mann and Whitney, 1947) between each pair of methods for all modality pairs. Almost all of these tests resulted on a significant difference between methods ($p < 0.05$) with the exception of the test between LOU and LOU2 on the CT-PD data where the null hypothesis was not rejected.

LOU and LOU2 are more robust to the initial alignment between the images and converged to a low registration error for a broader range of initial transformations. In the case of registering **PET- MR** images, the registration error of SSC is higher, which can be attributed to the poor structural information in PET images. LOU and LOU2 result in lower errors for all the PET experiments. Among our two methods, LOU2 has a broader capture range resulting in lower registration errors.

In order to assess the accuracy of each one of the methods for a standard initialization we also performed registration for each pair of images with the initial position given by the RIRE database. Our results are shown in Table 1 and are also available online on the

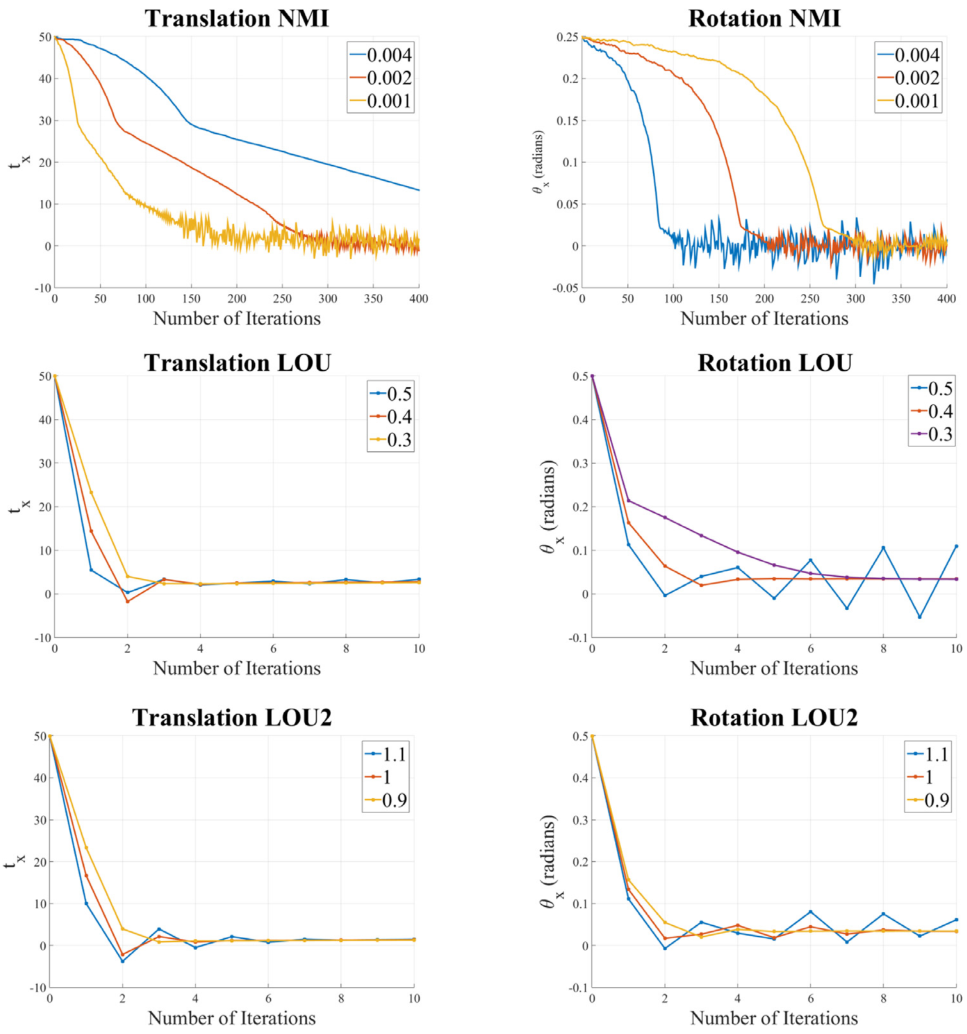


Fig. 9. Behavior of a gradient descent optimizer using different strategies to calculate parameter updates. (**Top**) Gradient of NMI. (**Middle**) Updates calculated using LOU. (**Bottom**) Updates from LOU2. Each plot line corresponds to a different step size α. Our methods have a faster convergence ($\sim$ 10 iterations) and a smoother behavior when compared to NMI. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

Table 1
Median TRE obtained after registering the image pairs of the RIRE dataset without initialization.

	NMI	SSC	LOU	LOU2
CT-T1	1.06 ± 1.16	2.06 ± 1.17	0.95 ± 1.30	0.89 ± 0.90
CT-T2	4.07 ± 7.74	5.53 ± 11.67	3.80 ± 4.17	3.07 ± 3.62
CT-PD	1.07 ± 12.46	1.58 ± 1.43	3.49 ± 1.91	3.07 ± 1.89
PET-T1	4.10 ± 1.83	8.11 ± 3.21	5.22 ± 3.48	5.29 ± 3.70
PET-T2	3.72 ± 2.66	7.29 ± 7.75	3.27 ± 2.91	4.17 ± 1.91
PET-PD	4.07 ± 2.11	2.62 ± 2.81	3.46 ± 2.39	2.98 ± 2.24

Table 2
Experiment ID in the RIRE database for the experiments in Table 1.

	NMI	SSC	LOU	LOU2
CT-T1	185750	185753	185752	185751
CT-T2	185754	185755	185808	185756
CT-PD	185762	185763	185764	185759
PET-T1	185769	185771	185804	185777
PET-T2	185773	185772	185784	185785
PET-PD	185767	185766	185801	185800

RIRE website using the ids shown in Table 2. We observe that given a good initialization all methods presented a similar final accuracy. The main advantage of our method in this dataset lies therefore the increased capture range as observed in Fig. 11 and the fastest convergence times due to the reduced number of required iterations for convergence. Qualitative results of these experiment are also shown on Fig. 10 where a pair of images from the RIRE database are shown before and after registration using LOU2

One of our driving hypothesis is that the prior knowledge used in our learning-based approach should serve to increase the capture range for multimodal registration. To demonstrate this hypothesis is verified, we compute our predicted updates for different initial misalignments and compare them with the gradient-based updates of NMI. The update plots are shown in Fig. 12 for one pair of PET-MR images and in Fig. 13 for one CT-MR pair.

For NMI, the updates based on the similarity gradient tend to be smooth for the range of parameters close to the optimal transformation, but noisy when the transformation parameters are far from the optimal alignment. This behavior causes the NMI gradient-based registration algorithm to fail when the initialization is far from the optimal solution. Furthermore, the optimal step size for NMI is small,[6] leading the gradient-ascent algorithm to converge in a larger number of iterations when compared to our method.

[6] Found through line search seeking to maximize the capture range of NMI while keeping a comparable error to our method.

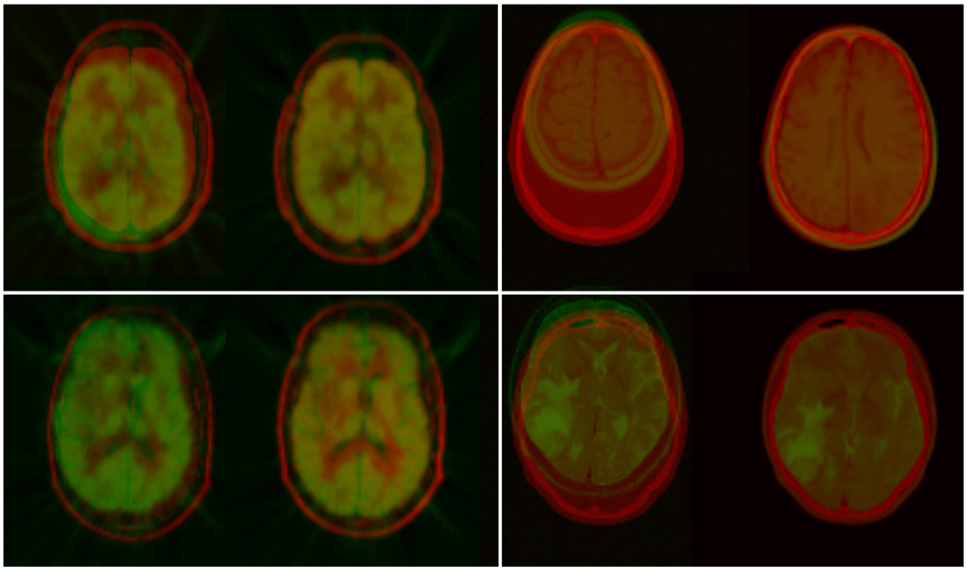


Fig. 10. Exemplary results of registered images of the RIRE database using our method (LOU2). In each box the image of the left corresponds to the initial position given by the RIRE database and the image of the right corresponds to the image after registration. Top left: PET-PD; Bottom Left: PET-T2; Top left: CT-T2; Bottom left: CT-T1.

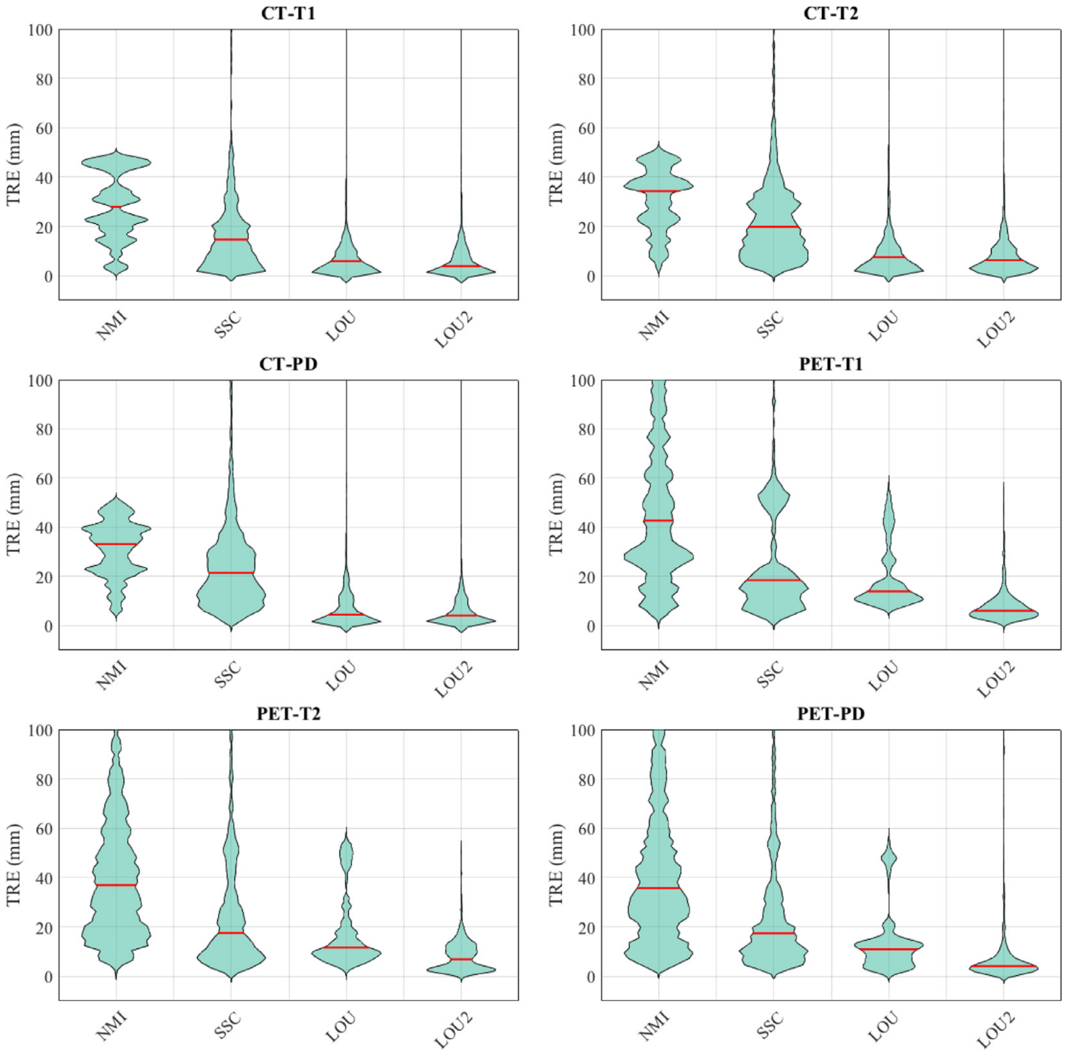


Fig. 11. Final registration error for the RIRE dataset. The width of the violin plot represents the distribution of the TRE, the red line indicates the median and the black line the mean. Results are shown for registration on PET - MR image pairs and CT- MR image pairs. The plot summarizes the results for all patients from 30 different initializations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

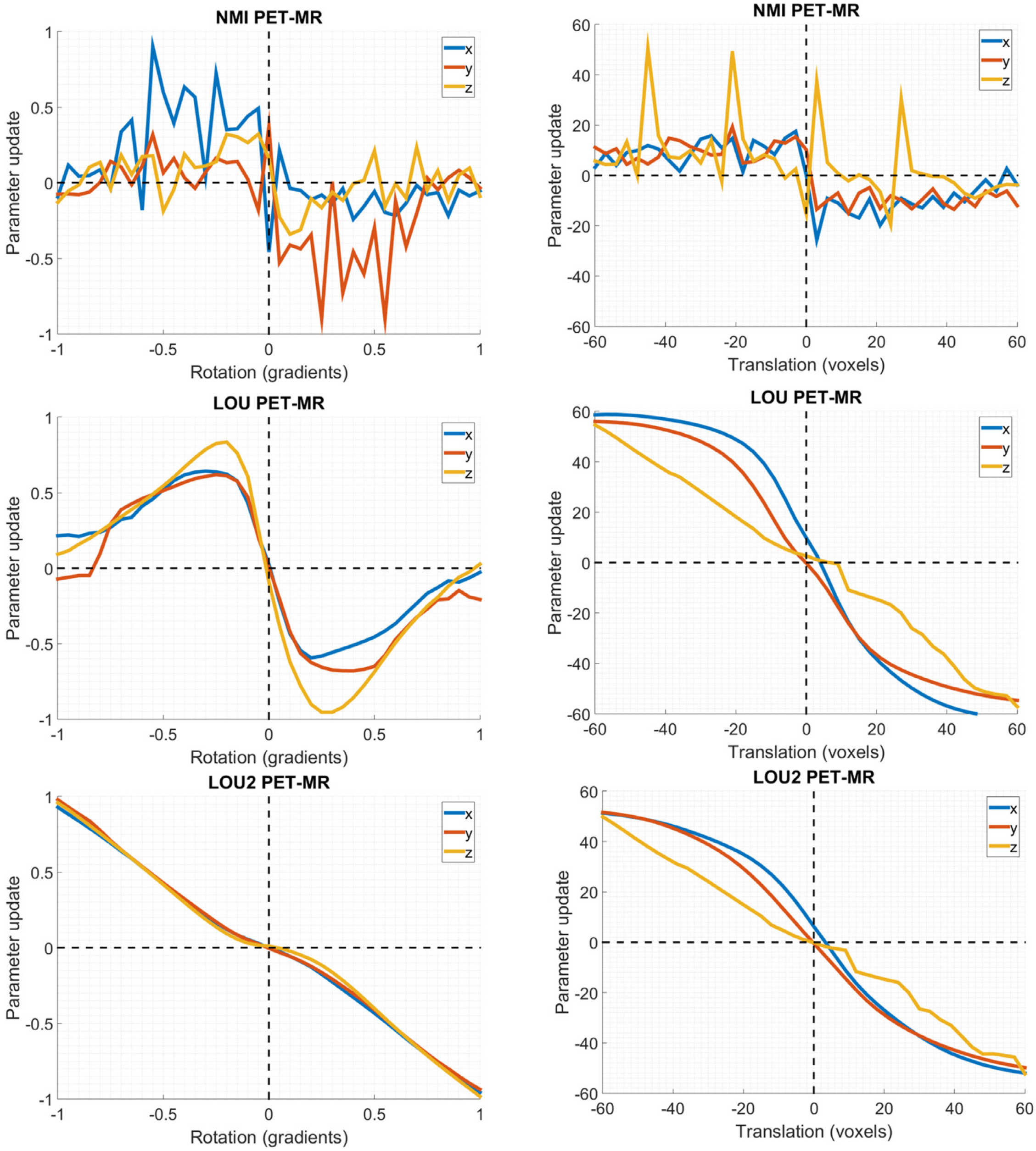


Fig. 12. Comparison of estimated parameter updates using different methods for **PET-MR** pairs: NMI, updates calculated using the gradient of NMI with respect to transformation parameters; LOU, updates calculated using our approach presented on Gutiérrez-Becker et al. (2016); LOU2: updates calculated using the presented method. The updates with our method are smoother and the estimated step size is close to the optimal.

For the learning-based methods, LOU and LOU2 , the optimization updates generated by our metric are smoother for all modality pairs. Additionally, in the case of LOU2, the predicted step size conforms to the ideal optimal step (*c.f* Fig. 2) for a wide capture range. The fact that the update is proportional to the distance to the optimal solution, allows the gradient-based algorithm to converge with the fewest iterations. The differences among the updates of different are most notable for the PET-MR pair (Fig. 12), most probably given the lack of structural similarity between the modalities. Similar behavior was observed for all converged instances of the algorithm given different image pairs and initializations.

5.3.1. *Feature relevance*

We analyze which features are more relevant for the registration task by analyzing how many times each feature type is selected in the training process. In Fig. 14 we observe the frequency at which different types of features were selected at different trees in a gradient boosting ensemble trained for the registration of CT and T1 images of the brain. The histograms of the first column correspond to the first trained tree of the ensemble, while the second and third columns correspond to the 30th and 100th tree respectively.

We can observe that in general features with short offsets and small boxes are favored by the ensemble of regression trees. However, features corresponding to long range appearance are still useful and are considered by the trees. We observed that in general, early trees tended to select a broader range of scales, while trees corresponding to later stages of the boosted ensemble selected mostly short range features. This behavior occurs because long range features are useful to perform a rough initialization for images with large initial misalignments but are less useful for the posterior fine alignment of the images. The first few trees of the

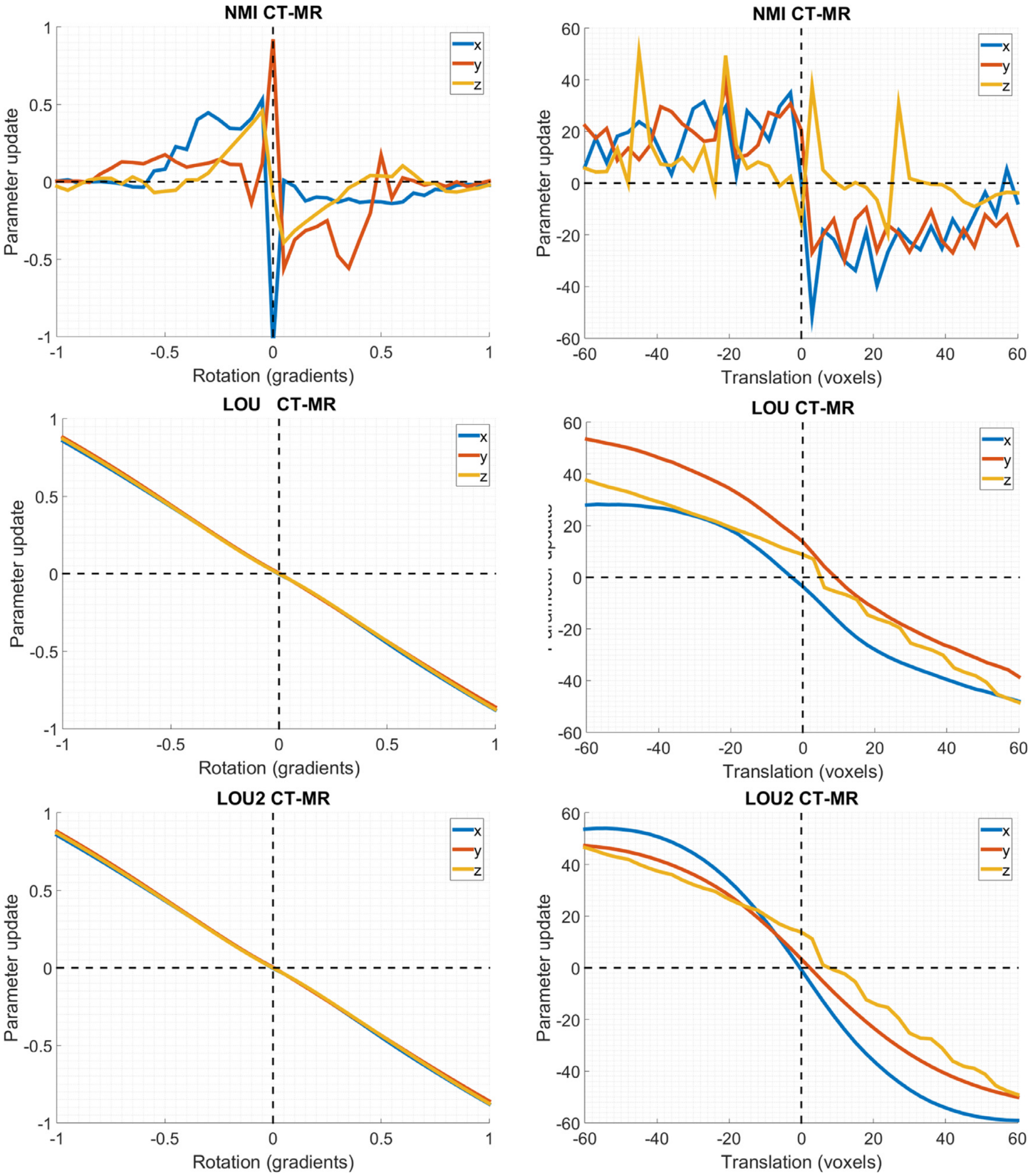


Fig. 13. Comparison of estimated parameter updates using different methods for **CT-MR** pairs: NMI, updates calculated using the gradient of NMI with respect to transformation parameters; LOU, updates calculated using our approach presented on Gutiérrez-Becker et al. (2016); LOU2: updates calculated using the presented method. The updates with our method are smoother and the estimated step size is close to the optimal.

ensemble are therefore able to perform a rough alignment of the images and later trees added to the ensemble reduce the final registration error.

In the third row of Fig. 14 we also show the proportion of times that trees select boxes from either the fixed or the moving image, or both images simultaneously. By observing the histograms we can conclude that our method extracts information within a single image to determine the relative position of each control point in the image and simultaneously obtains information from both images to determine the relationship of the intensities of both modalities. Interestingly while the first trees tend to select all features in an even distribution, further trees tend to rely more on features that take both modalities into account at the same time.

In the bottom row of Fig. 14 we show which operations between boxes are selected. All operations seem to have equal importance on the first trained trees with the exception of the binary operation between boxes. The low importance of binary operations between boxes can be explained by the high importance of the relationship between the intensity values of both modalities. Later trees tend to prefer operations between boxes instead of operations using a single box. This is related to the previous observation that trees trained on the later stages of the ensemble require more information on the local intensity relationship between images in order to reduce registration error.

5.4. IVUS-Histology deformable registration

For our third set of experiments, we perform deformable multi-modal registration in a dataset of IVUS and histology image pairs. Registration of IVUS and histology pairs is important for the

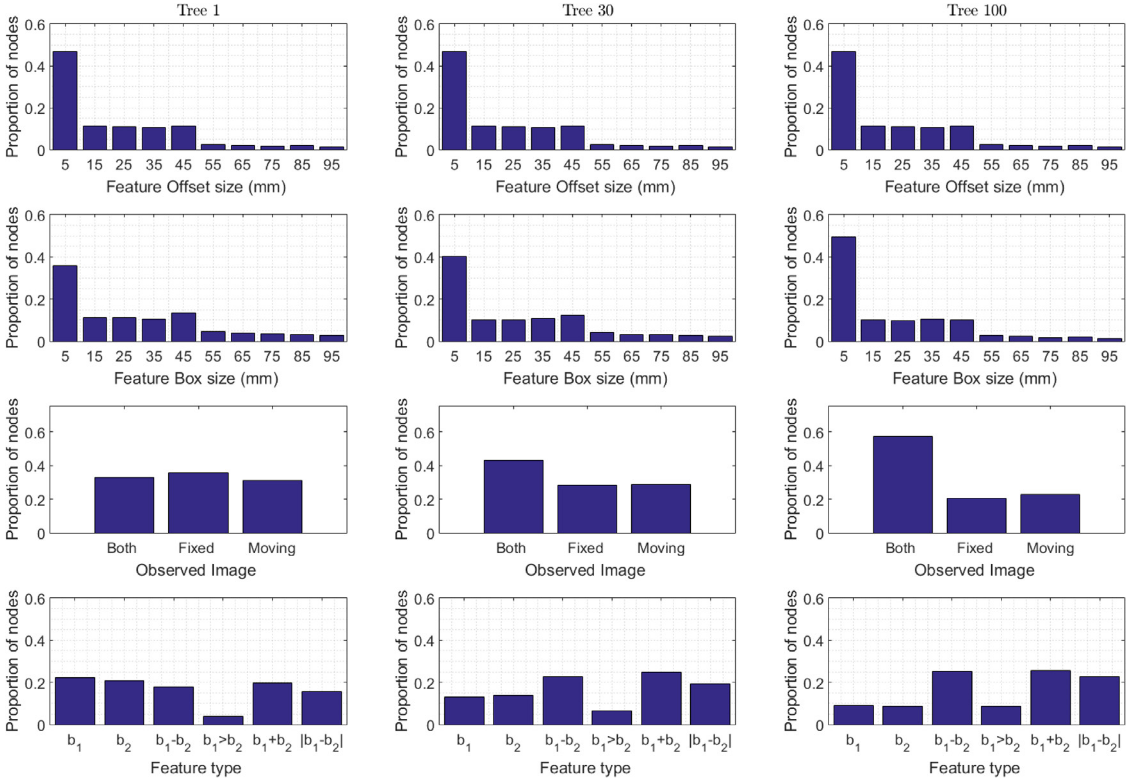


Fig. 14. Histograms showing the distribution of the selected features during the training of a boosting ensemble for CT-T1 registration. The top row corresponds to the size of the offsets, the second row to the size of the boxes, the third row to the image where the features are extracted from and the fourth row to the operation between boxes.

characterization of artherosclerotic tissues (Katouzian et al., 2012). To build the training set we align the IVUS and histological slices using the method presented in Katouzian et al. (2012), which is based on the alignment of manually-segmented structures. We then generate 100 transformations of each aligned pair by deforming the images using a B-spline with random parameters in a range between ± 20 for the B-spline coefficients , leading to a training set of 5 initially aligned images and 500 transformed images. We train our model following the same settings as for the RIRE experiments, but this time using a mask around the region actually containing tissue in the histological image. Please note that the synthetic transformations have only been used for training purposes, but testing was performed on pairs of histology and IVUS images without any additional transformations applied to them.

We quantitatively compare our approach against other methods by measuring the overlap (DICE) of segmented stenosis regions both in IVUS and the histology images. Even though using overlap measures is not the ideal measure to assess registration accuracy, it is still reliable for distinguishing between reasonable from inaccurate registrations (Rohlfing, 2012). For testing, we use again gradient-based optimization and we parametrize the transformation with a 3rd-order B-spline with 5 nodes per dimension distributed uniformly along the image. We do a 2-fold cross-validation evaluation with the 10 image pairs. The DICE scores after registration are shown in Fig. 15. Here, NMI and SSC present, in general, lower overlap measures when compared to our two supervised methods. Reasons for the comparably lower scores are the complex relationship between the intensities of both modalities which is difficult to capture by the joint histogram of mutual information, and the lack of structure which can be leveraged on by SSC. Our supervised approaches, on the other hand, result in much larger DICE values, indicating more accurate registration. After performing a Mann–Whitney U test between our method, SSC and NMI, our approach proved to yield a statistical significant improvement in the registration error ($p < 0.05$). The median registration error was similar between our two approaches, but the Gradient Boosted trees of LOU2 reduces the maximum registration error. Visual examples of the experiment are illustrated in Fig. 16, where we show overlays of IVUS and histology pairs before and after registration as well as the generated deformation fields.

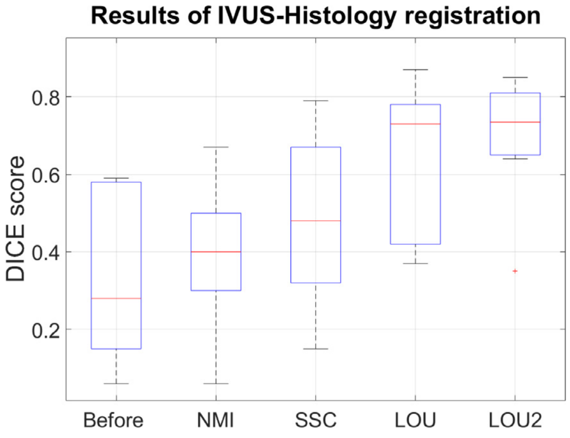


Fig. 15. Results of the multi-modal deformable registration of IVUS-Histology images. The registration success of the different methods is measured by means of the DICE score, indicating the overlap between the IVUS and Histology tissue masks after registration. The box plot shows the results of a 2-fold cross-validation experiment on ten images.

5.5. Amount of training data

Similar to the experiments performed on the IXI dataset, we evaluated the differences on the performance of our trained models depending on the amount of aligned images used to generate

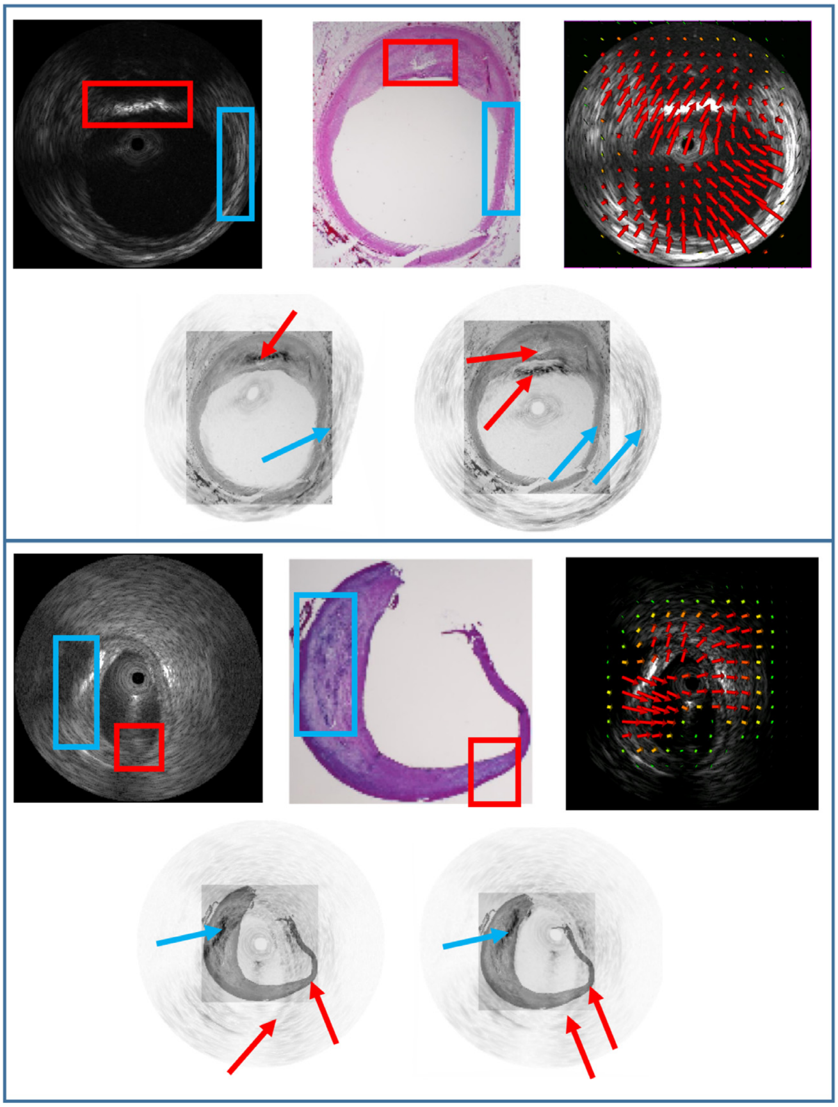


Fig. 16. Results of the IVUS-Histology deformable registration experiment. On the top box the example with the highest DICE score after registration using LOU2 (DICE = 0.81), and in the bottom box an example with a low DICE score after registration (DICE = 0.65). On top of each box the IVUS and histology images prior to registration and the deformation field obtained after the first iteration of LOU2. The red and blue boxes show corresponding regions in both images. In the bottom left part of the box, an overlay of the images before registration is shown. The arrows point towards the same regions in the boxes on the top. On the bottom right of each box, we show the overlay of the registered images using our method (LOU2), with the arrows indicating how the previously mismatching regions overlap after registration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the training set. As the IVUS-Histology dataset presents a bigger variation both in terms of appearance and deformations, we expect that increasing the dataset size has a bigger impact in registration accuracy when compared to previous experiments. We therefore evaluated 5 different models, each one trained using different number of aligned images ranging from 1 to 5. In Fig. 17 we can observe the results of this evaluation. Similar to our previous experiments, increasing the number of images improved the registration accuracy in the IVUS-Histology dataset. We can observe that our method is able to perform a more accurate registration when compared to NMI and SSC even after a single training image is included. However adding extra images on the training set allowed our method both to reduce the number of registration outliers as well as to reduce the final registration error.

6. Conclusions

In this work, we have a presented a novel approach to solving the multimodal registration problem based on a supervised regression and a gradient-based optimizer. Different to prior methods based on similarity design or learning, we directly target the prediction of the optimizer updates. To this end, we first show how the updates are related to the displacement field aligning the two images. We then demonstrate that using a training set of image pairs under known misalignments it is possible to train a regressor predicting the displacement fields from changes in the joint visual appearance of the images. Finally, we described how the predicted displacements can be generalized to other transformation parameterizations, and how the transformation updates can be inscribed within a simple gradient-based optimizer.

In the experimental evaluation, we have shown the flexibility and generality of our method to work on scenarios with very different modality pairs. Our method achieves comparable registration accuracy for several modality pairs were other methods have proven to be successful (for example, CT to MR). However, we have also shown that the same method is able to accurately register difficult pairs of modalities, such as IVUS to histology, for which other multimodal registration methods tend to fail (IVUS to Histology).

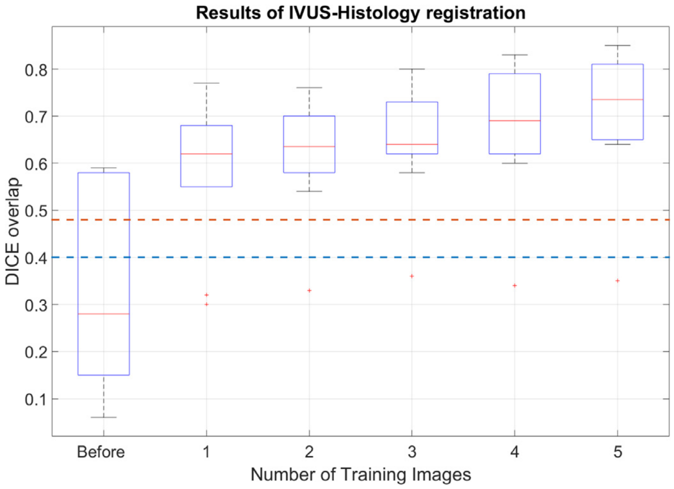


Fig. 17. Results of the multi-modal deformable registration of IVUS-Histology images using LOU2 as a function of the number of aligned images used for training. We can observe an increase of registration accuracy. The blue dotted line represents the median DICE score obtained after registration using NMI and the red line the DICE score after registration with SSC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Indeed, the supervised regression allows our method to deal with modalities displaying very different appearances and with weak structural similarities.

Our method requires sets of aligned images to train our supervised method. Nevertheless, we have observed that LOU and LOU2 are able to perform reliable registrations even when the training set size is very small. For example, the low reported errors in the RIRE experiments resulted from our model being trained on a single pair of aligned images transformed 100 times. The extra effort of generating such training sets can be justified in large scale studies, or when the ability of our method to perform registration on complex modality pairs is required (*i.e.* when other metrics fail). Adding extra pairs of aligned images to the training set enhances the accuracy of our method over other multimodal registration approaches, but it is not a requirement if focusing on our method's large capture range for comparable registration errors.

We have also shown that our method can easily be adapted to work with different parametrizations. By modeling our transformation using a displacement field we were able to easily integrate both a rigid registration parametrization and a deformable b-spline parametrization. Although this has not been thoroughly explored on this work, an additional advantage to parametrize our transformation as a displacement field is that a spatial regularization term could be easily applied to the displacement fields estimated by our regression model. Such a regularization term could prove important for the success of our method in scenarios where larger deformations are expected. At this point, it is important to mention that generating training sets for a highly deformable registration setting is not trivial and is an issue that has not yet been thoroughly explored in the supervised learning of similarity metrics. Generating training sets for deformable registration that are both realistic and extensive is an area to be addressed in order to extend our method to other scenarios and is an interesting area for future research.

The experiments have also shown that our method has an increased capture range and a faster convergence than the compared approaches. This is the result of modeling our metric as a motion prediction problem which takes the optimization into account. In the case of rigid registration, our method was able to converge in a maximum of 10 iterations, while 50 iterations were enough for an accurate deformable registration of the IVUS-Histology database. In both cases, the registration was successful even when the initial transformations were far from the optimal solution.

Our experiments mainly focused on registration on imaging settings on which acquisition protocols can be controlled and remain fairly homogeneous. However an open challenge still to be addressed by our method and other learning based approaches is that of highly variable environments, such as deformable registration of multimodal images in an intraoperative setting or registration of US images acquired at arbitrary positions and acquisition angles. The main reasons we have not yet tackled this challenge are the requirement to generate ground truth data which can be used to train our regression models and the difficulty of modeling the large difference in appearance which occur in an intra operative scenario. However our experiments so far have shown that our method is able to handle multiple modalities as well as different parametrization, which encourage us to further explore solutions which can tackle more challenging cases.

In the future, we plan to test our method in other scenarios where prior knowledge is required to improve registration accuracy, for instance, for the registration of intra-operative 2D ultrasound images to pre-operative MR for surgical navigation. We also believe the approach can contribute to mono-modal and volume-to-slice registration problems.

Acknowledgments

Benjamin Gutierrez is under the financial support of CONACYT-DAAD with scholarship number A/12/72597. We thank Amin Katouzian and Andrew Laine for allowing us to use the IVUS-Histology dataset. We also thank Mattias Heinrich for kindly providing the implementation of SSC.

References

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. IJCV 107 (2), 177–190.
Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 161–168.
Cheng, X., Zhang, L., Zheng, Y., 2016. Deep similarity learning for multimodal medical images. Comput. Methods Biomech. Biomed. Eng. 1–5.
Chou, C.-R., Frederick, B., Mageras, G., Chang, S., Pizer, S., 2013. 2d/3d image registration using regression learning. Comput. Vision Image Understanding 117 (9), 1095–1106.
Coupé, P., Hellier, P., Morandi, X., Barillot, C., 2012. 3d rigid registration of intraoperative ultrasound and preoperative MR brain images based on hyperechogenic structures. J. Biomed. Imaging 2012, 1.
Criminisi, A., Shotton, J. (Eds.), 2013. Decision forests for computer vision and medical image analysis, [Advances in Computer Vision and Pattern Recognition]. Springer London, pp. 245–260.
Criminisi, A., Shotton, J., Bucciarelli, S., 2009. Decision forests with long-range spatial context for organ localization in ct volumes. In: MICCAI. Citeseer, pp. 69–80.
Dollár, P., Welinder, P., Perona, P., 2010. Cascaded pose regression. In: CVPR. IEEE, pp. 1078–1085.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.
Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D., 2016. An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III. Springer International Publishing, Cham, pp. 229–237.
Gutiérrez-Becker, B., Mateus, D., Peter, L., Navab, N., 2016. Learning optimization updates for multimodal registration. In: MICCAI. Springer, pp. 19–27.
Heinrich, M., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J., 2012. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal 16 (7).
Heinrich, M.P., Jenkinson, M., Papiez, B., Brady, M., Schnabel, J., 2013. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: MICCAI, pp. 187–194.
Hu, S., Wei, L., Gao, Y., Guo, Y., Wu, G., Shen, D., 2016. Learning-based deformable image registration for infant MR images in the first year of life. Med. Phys. 44, 158,170.
Jiang, J., Zheng, S., Toga, A., T., Z., 2008. Learning based coarse-to-fine image registration. In: CVPR, pp. 1–7.
Jurie, F., Dhome, M., 2002. Hyperplane approximation for template matching. TPAMI 24 (7), 996–1000.

Katouzian, A., Karamalis, A., Lisauskas, J., Eslami, A., Navab, N., 2012. Ivus-histology image registration. In: International Workshop on Biomedical Image Registration. Springer, pp. 141–149.

Katouzian, A., Sathyanarayana, S., Li, W., Thomas, T., Carlier, S.G., 2007. Challenges in tissue characterization from backscattered intravascular ultrasound signals. Medical Imaging. International Society for Optics and Photonics. 651300–651300

Kim, M., Wu, G., Yap, P.T., Shen, D., 2012. A general fast registration framework by learning deformation appearance correlation. IEEE Trans. Image Process. 21 (4), 1823–1833.

Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N., Scholkopf, B., 2009. Learning similarity measure for multi-modal 3d image registration. CVPR, pp. 186–193.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging 16 (2), 187–198.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18 (1), 50–60.

Markelj, P., Tomaževič, D., Likar, B., Pernuš, F., 2012. A review of 3d/2d registration methods for image-guided interventions. Med. Image Anal. 16 (3), 642–661.

Mattes, D., Haynor, D.R., Vesselle, H., Lewellyn, T.K., Eubank, W., 2001. Nonrigid multimodality image registration. In: Medical Imaging 2001. International Society for Optics and Photonics, pp. 1609–1620.

Michel, F., Bronstein, M., Bronstein, A., Paragios, N., 2011. Boosted metric learning for 3d multi-modal deformable registration. In: ISBI. IEEE, pp. 1209–1214.

Navab, N., Hennersperger, C., Frisch, B., Fürst, B., 2016. Personalized, relevance-based multimodal robotic imaging and augmented reality for computer assisted interventions. Med. Image Anal. 33, 64–71.

Nocedal, J., Wright, S.J., 2006. Numerical Optimization, second ed. Springer, New York.

Oktay, O., Schuh, A., Rajchl, M., Keraudren, K., Gomez, A., Heinrich, M.P., Penney, G., Rueckert, D., 2015. Structured decision forests for multi-modal ultrasound image registration. In: MICCAI. Springer, pp. 363–371.

Peter, L., Pauly, O., Chatelain, P., Mateus, D., Navab, N., 2015. Scale-adaptive forest training via an efficient feature sampling scheme. MICCAI. Springer.

Pluim, J., Maintz, J., Viergever, M., 2004. f-Information measures in medical image registration. TMI 23 (12), 1508–1516.

Pluim, J.P., Maintz, J.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. TMI 22 (8), 986–1004.

Rohlfing, T., 2012. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans. Med. Imaging 31 (2), 153–163.

Sabuncu, M., Ramadge, P., 2008. Using spanning graphs for efficient image registration. TMI 17 (5), 788–797.

Simonovsky, M., Gutierrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration. In: MICCAI, pp. 10–18.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. TMI 32 (7), 1153–1190.

Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 677–684.

Wachinger, C., Navab, N., 2012. Entropy and Laplacian images: structural representations for multi-modal registration. Med. Image Anal. 16 (1), 1–17.

Wein, W., Khamene, A., Clevert, D.-A., Kutter, O., Navab, N., 2007. Simulation and fully automatic multimodal registration of medical ultrasound. In: MICCAI. Springer, pp. 136–143.

Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1 (1), 35–51.

West, J., Fitzpatrick, J.M., Wang, M.Y., Dawant, B.M., Maurer Jr, C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., et al., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. J. Comput. Assist. Tomogr. 21 (4), 554–568.

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment. In: CVPR, pp. 532–539.

Yang, X., Kwitt, R., Niethammer, M., 2016. Fast predictive image registration. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings. Springer International Publishing, Cham, pp. 48–57. doi:10.1007/978-3-319-46976-8_6.

Zöllei, L., Wells, W., 2006. Multi-modal image registration using dirichlet-encoded prior information. In: Biomedical Image Registration, Third International Workshop, WBIR 2006, Utrecht, The Netherlands, July 9–11, 2006, Proceedings, pp. 34–42. doi:10.1007/11784012_5.

A Multi-armed Bandit to Smartly Select a Training Set from Big Medical Data

Benjamín Gutiérrez[1,2](✉), Loïc Peter[2,3], Tassilo Klein[4], and Christian Wachinger[1]

[1] Artificial Intelligence in Medical Imaging (AI-Med),
KJP, LMU München, Munich, Germany
gutierrez.becker@tum.de
[2] CAMP, Technische Universität München, Munich, Germany
[3] Translational Imaging Group, University College London, London, UK
[4] SAP SE Berlin, Berlin, Germany

Abstract. With the availability of big medical image data, the selection of an adequate training set is becoming more important to address the heterogeneity of different datasets. Simply including all the data does not only incur high processing costs but can even harm the prediction. We formulate the smart and efficient selection of a training dataset from big medical image data as a multi-armed bandit problem, solved by Thompson sampling. Our method assumes that image features are not available at the time of the selection of the samples, and therefore relies only on meta information associated with the images. Our strategy simultaneously exploits data sources with high chances of yielding useful samples and explores new data regions. For our evaluation, we focus on the application of estimating the age from a brain MRI. Our results on 7,250 subjects from 10 datasets show that our approach leads to higher accuracy while only requiring a fraction of the training data.

1 Introduction

Machine learning has been one of the driving forces for the huge progress in medical imaging analysis over the last years. Of key importance for learning-based techniques is the training dataset that is used for estimating the model parameters. Including all available data in a training set is becoming increasingly impractical, since processing the data to create training models can be very time consuming on huge datasets. In addition, most processing may be unnecessary because it does not help the model estimation for a given task. In this work, we propose a method to select a subset of the data for training that is most relevant for a specific task. Foreshadowing some of our results, such a guided selection of a subset for training can lead to a higher performance than using all the available data while requiring only a fraction of the processing time.

The task of selecting a subset of the data for training is challenging because at the time of making the decision, we do not yet have processed the data and we do therefore not know how the inclusion of the sample would affect the prediction. On the other hand, in many scenarios each image is assigned *metadata* about

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 38–45, 2017.
DOI: 10.1007/978-3-319-66179-7_5

the subject (sex, diagnosis, age, etc.) or the image acquisition (dataset of origin, location, imaging device, etc.). We hypothetize that some of this information can be useful to guide the selection of samples but it is a priori not clear which information is most relevant and how it should be used to guide the selection process. To address this, we formulate the selection of the samples to be included in a training set as reinforcement learning problem, where a trade-off must be reached between the *exploration* of new sources of data and the *exploitation* of sources that have been shown to lead to informative data points in the past. More specifically, we model this as a multi-armed bandit problem solved with Thompson sampling, where each arm of the bandit corresponds to a cluster of samples generated using meta information.

In this paper, we apply our sample selection method to brain age estimation [7] from MRI T1 images. The estimated age serves as a proxy for biological age, whose difference to the chronological age can be used as indicator of disease [6]. The age estimation is a well-suited application for testing our algorithm as it allows us to work with a large number of datasets, since the subject's age is one of the few variables that is included in every neuroimaging dataset.

1.1 Related Work

Our work is mostly related to active learning approaches, whose aim is to select samples to be labeled out of a pool of unlabeled data. Examples of active learning approaches applied to medical imaging tasks include the work by Hoi *et al.* [9], where a batch mode active learning approach was presented for selecting medical images for manually labeling the image category. Another active learning approach was proposed for the selection of histopathological slices for manual annotation in [21]. The problem was formulated as constrained submodular optimization problem and solved with a greedy algorithm. To select a diverse set of slices, the patient identity was used as meta information. From a methodological point of view, our work relates to the work of Bouneffouf *et al.* [1], where an active learning strategy based on contextual multi-armed bandits is proposed. The main difference between all these active learning approaches and our method is that image features are not available a priori in our application, and therefore can not be used in the sample selection process. Our work also relates to domain adaptation [15,20]. In instance weighting, the training samples are assigned weights according to the distribution of the labels (class imbalance) [10] and the distribution of the observations (covariate shift) [16]. Again these methods are not directly applicable in our scenario because the distribution of the metadata is not always defined on the target dataset.

2 Method

2.1 Incremental Sample Selection

In supervised learning, we model a predictive function $f : (\mathbf{x}, \mathbf{p}) \mapsto y$ depending on a parameter vector $\mathbf{p}$, relating an observation $\mathbf{x}$ to its label y. In our application, $\mathbf{x} \in \mathbb{R}^m$ is a vector with m quantitative brain measurements from the image and $y \in \mathbb{R}$ is the age of the subject. The parameters $\mathbf{p}$ are estimated by using a training set $S^T = \{s_1, s_2, \ldots, s_{N_{train}}\}$, where each sample $s = (\mathbf{x}, y)$ is a pair of a feature vector and its associated true label. Once the parameters are estimated, we can predict the label $\tilde{y}$ for a new observation $\tilde{\mathbf{x}}$ with $\tilde{y} = f(\tilde{\mathbf{x}}, \mathbf{p}^*)$, where the prediction depends on the estimated parameters and therefore the training dataset. In our scenario, the samples to be included in the training set S^T are selected from a large source set $S = \{h_1, h_2, .., h_{N_{total}}\}$ containing *hidden* samples of the form $h = \{\hat{\mathbf{x}}, \hat{y}, \mathbf{m}\}$. Each h contains hidden features $\hat{\mathbf{x}}$ and label $\hat{y}$ that can only be revealed after processing the sample. In addition, each hidden sample possesses a d-dimensional vector of metadata $\mathbf{m} \in \mathbb{Z}^d$ that encodes characteristics of the patient or the image such as sex, diagnosis, and dataset of origin. In contrast to $\hat{\mathbf{x}}$ and $\hat{y}$, $\mathbf{m}$ is known a priori and can be observed at no cost. To include a sample h from set S into S^T, first its features and labels have to be revealed, which comes at a high cost. Consequently, we would like to find a sampling strategy that minimizes the cost by selecting only the most relevant samples according to the metadata $\mathbf{m}$.

2.2 Multiple Partitions of the Source Data

In order to guide our sample selection algorithm, we create multiple partitions of the source dataset, where each one considers different information from the metadata $\mathbf{m}$. Considering the j-th meta information ($1 \leq j \leq d$), we create the j-th partition $S = \cup_{i=1}^{\eta_j} C_i^j$ with η_j a predefined number of bins for $\mathbf{m}[j]$. As a concrete example, sex could be used for partitioning the data, so $S = C_{\text{female}}^{\text{sex}} \cup C_{\text{male}}^{\text{sex}}$ and $\eta_{\text{sex}} = 2$. In the case of continuous variables such as age, partitions can be done by quantizing the variable into bins. All the clusters generated using different meta information are merged into a set of clusters $\mathcal{C} = \{C_l^j\}$. Since partitions can be done using different elements of $\mathbf{m}$ a sample can be assigned to more than one cluster.

We hypothesize that given this partitioning, there exist clusters $C_i \in \mathcal{C}$ that contain more relevant samples than others for a specific task. Intuitively, we would like to draw samples h from clusters with a higher probability of returning a relevant sample. However, since the relationship between the metadata and the task is uncertain, the utility of each cluster for a specific task is unknown beforehand. We will now describe a strategy that simultaneously *explores* the clusters to find out which ones contain more relevant information and *exploits* them by extracting as many samples from relevant clusters as possible.

2.3 Sample Selection as a Multi-armed Bandit Problem

We model the task of sequential sample selection as a multi-armed bandit problem. At each iteration t, a new sample is added to the training dataset S^T. For adding a sample, the algorithm decides which cluster $C_i \in \mathcal{C}$ to exploit and randomly draws a training sample s_t from cluster C_i. The corresponding feature vector $\mathbf{x}_t$ and label y_t are revealed and the usefulness of the sample s_t for the given task is evaluated, yielding a reward $r_t \in \{-1, 1\}$. A reward $r_t = 1$ is given if adding the sample improves the prediction accuracy of the model and $r_t = -1$ otherwise.

At $t = 0$, we do not possess knowledge about the utility of any cluster. This knowledge is incrementally built as more and more samples are drawn and their rewards are revealed. To this end, each cluster is assigned a distribution of rewards $\varPi_i$. With every sample the distribution better approximates the true expected reward of the cluster, but every new sample also incurs a cost. Therefore, a strategy needs to be designed that explores the distribution for each of the clusters, while at the same time exploiting as often as possible the most rewarding sources.

To solve the problem of selecting from which C_i to sample at every iteration t, we follow a strategy based on Thompson sampling [17] with binary rewards. In this setting, the expected rewards are modeled using a probability P_i following a Bernoulli distribution with parameter $\pi_i \in [0, 1]$. We maintain an estimate of the likelihood of each π_i given the number of successes α_i and failures β_i observed for the cluster C_i so far. Successes ($r = 1$) and failures ($r = -1$) are defined based on the reward of the current iteration. It can be shown that this likelihood follows the conjugate distribution of a Bernoulli law, i.e., a Beta distribution $Beta(\alpha_i, \beta_i)$ so that

$$P(\pi_i|\alpha_i, \beta_i) = \frac{\varGamma(\alpha_i + \beta_i)}{\varGamma(\alpha_i)\varGamma(\beta_i)}(1 - \pi_i)^{\beta_i - 1}\pi_i^{\alpha_i - 1}. \tag{1}$$

with the gamma function $\varGamma$. At each iteration, $\hat{\pi}_i$ is drawn from each cluster distribution P_i and the cluster with the maximum $\hat{\pi}_i$ is chosen. The procedure is summarized in Algorithm 1.

Algorithm 1. Thompson Sampling for Sample Selection

```
$\alpha_i = 1, \beta_i = 1, \forall i \in \{1, \dots, N\}$
for $t = 1, 2, ...$ do
    for $i = 1, \dots, N$ do
        Draw $\hat{\pi}_i$ from $Beta(\alpha_i, \beta_i)$.
    Reveal sample $h_t = \{\mathbf{x}_t, y_t, \mathbf{m}_t\}$ from cluster $C_j$ where $j := \arg\max_i \hat{\pi}_i$.
    Add sample $h_t$ to $S^T$ and remove from all clusters.
    Obtain new model parameters $\mathbf{p}^*$ from updated training set $S^T$.
    Compute reward $r_t$ based on new prediction $\tilde{y} = f(\mathbf{x}, \mathbf{p}^*)$.
    if $r_t == 1$ then $\alpha_j = \alpha_j + 1$
    else $\beta_j = \beta_j + 1$
```

3 Results

In order to showcase the advantages of the multi-armed bandit sampling algorithm (MABS), we evaluate our method in estimating the biological age of a subject given a set of volume and thickness features of the brain. We choose this task because of the big number of available brain scans in public databases and the relevance of age estimation as a diagnostic tool for neurodegenerative diseases [18]. For predicting the age, we reconstruct brain scans with FreeSurfer [5] and extract volume and thickness measurements to create feature vectors $\mathbf{x}$. Based on these features, we train a regression model for predicting the age of previously unseen subjects.

3.1 Data

We work on MRI T1 brain scans from 10 large-scale public datasets: ABIDE [3], ADHD200 [14], AIBL [4], COBRE [13], IXI[1], GSP [2], HCP [19], MCIC [8], PPMI [12] and OASIS [11]. From all of these datasets, we obtain a total number of 7,250 images, which is to the best of our knowledge the largest dataset ever used for brain age prediction. Since each one of these datasets is targeted towards different applications, the selected population is heterogeneous in terms of age, sex, and health status. For the extraction of thickness and volume measurements, we process the images with FreeSurfer. Even though this is a fully automatic tool, the feature extraction is a computationally intensive task, which is by far the bottleneck of our age prediction regression model.

3.2 Age Estimation

We perform age estimation on two different testing scenarios. In the first, we create a testing dataset by randomly selecting subsets from all the datasets. The aim of this experiment is to show that our method is capable of selecting samples that will create a model that can generalize well to a heterogeneous population. In the second scenario, the testing dataset corresponds to a single dataset. In this scenario, we show that the sample selection permits tailoring the training dataset to a specific target dataset.

Experiment 1. For the first experiment we take all the images in the dataset and we divide them randomly into three sets: (1) a small validation set of 2% of all samples to compute the rewards given to MABS, (2) a large testing set of 48% to measure the performance of our age regression task, and (3) a large hidden training set of 50%, from which samples are taken sequentially using MABS. We perform the sequential sample selection described in Algorithm 1 using the following metadata to construct the clusters $\mathcal{C}$: *age*, *dataset*, *diagnosis*, and *sex*. We experiment with considering all of the metadata separately, to investigate the importance of each one, and the joint modeling considering all partitions at once. We opted to use ridge regression as our learning algorithm because of its fast training and good performance for our task, but other regression models can be easily plugged into our method. Rewards r are given to each bandit by estimating and observing if the r^2 score of the prediction in the validation set increases. It is important to emphasize that the testing set is not observed by the bandits in the process of giving rewards. Every experiment is repeated 20 times using different

[1] http://brain-development.org/ixi-dataset/.

random splits and the mean results are shown. We compare with two baselines: the first one (RANDOM) consists of obtaining samples at random from the hidden set and adding them sequentially to the training set. As a second baseline (AGE PRIOR), we add samples sequentially by following the age distribution of the testing set. The results of this first experiment are shown in Fig. 1 (top left). In almost all of the cases, using MABS as a selection strategy performed better than the baselines. Notably, an increase in performance is obtained not only when the relationships between the metadata and the task are direct, like in the case of the clusters constructed by age, but also when this relationship is not clear, like in the case of clustering the images using only dataset or diagnostic information. Another important aspect is that even when the meta information is not informative, like in the case of the clusters generated by sex, the prediction using MABS is not affected.

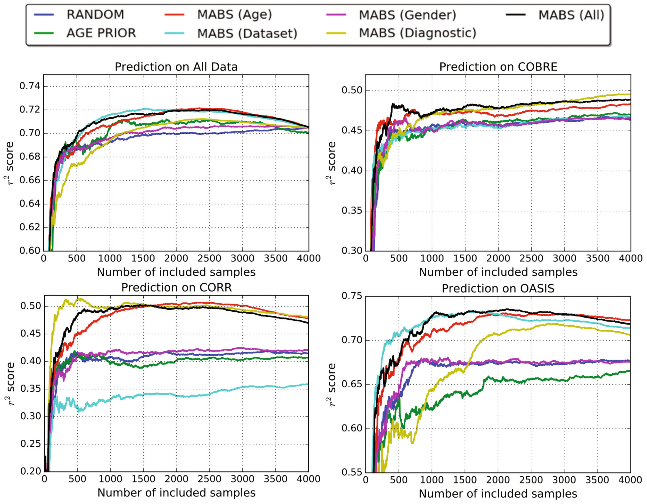


Fig. 1. Results of our age prediction experiments in terms of r^2 score. A comparison is made between MABS using different strategies to build the clusters $\mathcal{C}$, a random selection of samples, and a random selection based on the age distribution of the test data. To improve the presentation of the results, we limit the plot to 4,000 samples.

Experiment 2. For our second experiment, we perform age estimation with the test data being a specific dataset. This experiment follows the same methodology as the previous one with the important difference of how the datasets are split. This time the split is done by choosing: (1) a small validation set, taken only from the target dataset, (2) a testing set, which corresponds to the remaining samples in the target dataset not included in the validation set, and (3) a hidden dataset containing all the samples from the remaining datasets. The goal of this experiment is to show that our approach can be applied to selecting samples according to a specific population and prediction task. Figure 1 shows the results for three different target datasets. We observe that bandits operating on single metadata like diagnosis or dataset can perform very well for the sample selection. However, the best metadata is different for each of the presented datasets. We also observe that MABS using all available metadata extracts informative samples more efficiently than the baselines and always close to the best performing single metadata MABS. This strengthens our hypothesis that it is difficult to define an a priori relationship between the metadata and the task. Consequently, it is a better strategy to pass all the metadata from multiple sources to MABS and let it select the most relevant information.

4 Conclusion

We have proposed a method for efficiently and intelligently sampling a training dataset from a large pool of data. The problem was formulated as reinforcement learning, where the training dataset was sequentially built after evaluating a reward function at every step. Concretely, we used a multi-armed bandit model that was solved with Thompson sampling. The intelligent selection considered metadata of the scan to construct a distribution about the expected reward of a training sample. Our results showed that the selective sampling approach leads to higher accuracy than using all the data, while requiring less time for processing the data. We demonstrated that our technique can either be used to build a general model or to adapt to a specific target dataset, depending on the composition of the test dataset. Since our method does not require to observe the information contained in the images, it could also be applied to predict useful samples even before the images are acquired, guiding the recruitment of subjects.

Acknowledgements. This work was supported in part by SAP SE, the Faculty of Medicine at LMU (FöFoLe), and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

References

1. Bouneffouf, D., Laroche, R., Urvoy, T., Feraud, R., Allesiardo, R.: Contextual bandit for active learning: active thompson sampling. In: Loo, C.K., Yap, K.S., Wong, K.W., Teoh, A., Huang, K. (eds.) ICONIP 2014. LNCS, vol. 8834, pp. 405–412. Springer, Cham (2014). doi:10.1007/978-3-319-12637-1_51
2. Buckner, R., Hollinshead, M., Holmes, A., Brohawn, D., Fagerness, J., O'Keefe, T., Roffman, J.: The brain genomics superstruct project. Harvard Dataverse Network (2012)
3. Di Martino, A., Yan, C., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**(6), 659–667 (2014)
4. Ellis, K., Bush, A., Darby, D., et al.: The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. Int. Psychogeriatr. **21**(04), 672–687 (2009)

5. Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**(3), 341–355 (2002)
6. Franke, K., Luders, E., May, A., Wilke, M., Gaser, C.: Brain maturation: predicting individual brainage in children and adolescents using structural mri. Neuroimage **63**(3), 1305–1312 (2012)
7. Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Alzheimer's Disease Neuroimaging Initiative: Estimating the age of healthy subjects from t 1-weighted mri scans using kernel methods: Exploring the influence of various parameters. Neuroimage **50**(3), 883–892 (2010)
8. Gollub, R.L., Shoemaker, J., King, M., White, T., Ehrlich, S., Sponheim, S., Clark, V., Turner, J., Mueller, B., Magnotta, V., et al.: The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Neuroinformatics **11**(3), 367–388 (2013)
9. Hoi, S., Jin, R., Zhu, J., Lyu, M.: Batch mode active learning and its application to medical image classification. In: ICML, pp. 417–424. ACM (2006)
10. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intel. Data Anal. **6**(5), 429–449 (2002)
11. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. Cognitive Neurosci. **19**(9), 1498–1507 (2007)
12. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al.: The parkinson progression marker initiative (PPMI). Prog. Neurobiol. **95**(4), 629–635 (2011)
13. Mayer, A., Ruhl, D., Merideth, F., Ling, J., Hanlon, F., Bustillo, J., Cañive, J.: Functional imaging of the hemodynamic sensory gating response in schizophrenia. Hum. Brain Mapp. **34**(9), 2302–2312 (2013)
14. Milham, M.P., Fair, D., Mennes, M., Mostofsky, S.H., et al.: The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers Syst. Neurosci. **6**, 62 (2012)
15. Pan, S., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
16. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference **90**(2), 227–244 (2000)
17. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3/4), 285–294 (1933)
18. Valizadeh, S., Hänggi, J., Mérillat, S., Jäncke, L.: Age prediction on the basis of brain anatomical measures. Hum. Brain Mapp. **38**(2), 997–1008 (2017)
19. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T., Yacoub, E., Ugurbil, K., WU-Minn HCP Consortium, et al: The WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
20. Wachinger, C., Reuter, M.: Domain adaptation for alzheimer's disease diagnostics. Neuroimage **139**, 470–479 (2016)
21. Zhu, Y., Zhang, S., Liu, W., Metaxas, D.N.: Scalable histopathological image analysis via active learning. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8675, pp. 369–376. Springer, Cham (2014). doi:10.1007/978-3-319-10443-0_47