Multi-atlas Segmentation and Analysis of the Fetal Brain in Ventriculomegaly

Mohamed Oualid Benkarim

DOCTORAL THESIS UPF / 2018

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Directors of the thesis

Dr. Gerard Sanroma and Prof. Gemma Piella
Department of Information and Communication Technologies

Universitat Pompeu Fabra Barcelona

Directors:

Dr. Gerard Sanroma	Postdoctoral Researcher German Center for Neurodegenerative Diseases Bonn, Germany
Prof. Gemma Piella	Associate Professor Universitat Pompeu Fabra Barcelona, Spain

Review Committee (Main):

Prof. Julia Schnabel	King's College London
Prof. Federico Sukno	Universitat Pompeu Fabra
Prof. Pierrick Coupé	University of Bordeaux

Review Committee (Reserve):

Prof. Petia Radeva	Universitat de Barcelona
Prof. Constantine Butakoff	Barcelona Supercomputing Center

This work was carried out in the research group *Simulation, Imaging and Modelling for Biomedical Systems (SIMBIOsys)*, at the Department of Information and Communication Technologies of *Universitat Pompeu Fabra*, Barcelona, Spain. This thesis was supported by a scholarship from the Department of Information and Communication Technologies at *Universitat Pompeu Fabra* (DTIC-UPF).

To my mom

Abstract / Resumen

Abstract. Nowadays, imaging of the human brain is vastly used in clinical settings and by the neuroscientific research community. There is an ever-increasing demand for novel biomedical image analysis approaches and tools to study the brain from its early intrauterine stage through adolescence to adulthood. The intrauterine period, in particular, is a crucial stage for the study of early neurodevelopmental processes. The idiosyncratic nature of the fetal brain poses numerous challenges and asks for the development of new techniques that take into consideration the peculiarities of *in utero* neurodevelopment. Although still in its infancy, medical image analysis techniques are progressively landing on the study of fetal brains. The purpose of this thesis is to develop automatic segmentation approaches that can be applied to brains at different life stages, including the gestational period, and investigate *in utero* brain development under abnormal conditions. The main contributions of this thesis are twofold:

1. Brain segmentation from MR images: we proposed two novel segmentation frameworks that are independent of target dataset and, therefore, can be used in the segmentation of images of brains at different life stages. These methods approach the segmentation problem from two different but complementary perspectives. The first approach attempts to mitigate systematic errors caused by registration failures. Each training atlas produces a *confidence map* based on the registration quality that weighs the contribution of its labelmap in the final segmentation. The second approach focuses on reducing the impact of interpolation errors, which are inherent to most traditional segmentation approaches. Registration is only used to establish spatial correspondences, while atlas information is extracted from its native

space, respecting, therefore, the manual segmentation protocol used by the expert.

2. Analysis of fetuses with ventriculomegaly: ventriculomegaly is a condition in which one or both lateral ventricles are dilated. Postnatal neurodevelopmental impairment is observed in some fetuses with the aforementioned condition. To better understand *in utero* neurodevelopment, we analyzed the relationship of ventricular dilation with alterations in cortical folding. First, we studied this association using either lateral ventricular volume or diagnosis as a descriptor of ventriculomegaly. Then, statistical analysis and sparse linear regression were used to assess the relationship from two different approaches. In our second study, we adopt a holistic approach by incorporating the ventricular shapes instead of using a single scalar value (i.e., ventricular volume) to characterize ventricular enlargement. We proposed a novel approach to jointly analyze cortical and ventricular shapes based on their growth patterns, which allowed us to find fine-scaled associations between both shapes.

Resumen. En la actualidad, las imágenes del cerebro humano son ampliamente utilizadas en entornos clínicos y por la comunidad neurocientífica. Existe una demanda, cada vez mayor, de herramientas y enfoques de análisis de imágenes biomédicas novedosos para estudiar el cerebro desde su temprana etapa intrauterina hasta la adolescencia y la edad adulta. El periodo intrauterino, en particular, es una etapa crucial para el estudio de los procesos iniciales del neurodesarrollo. La naturaleza idiosincrásica del cerebro fetal plantea numerosos desafíos y requiere el desarrollo de nuevas técnicas que contemplen las peculiaridades del neurodesarrollo fetal. Aunque todavía está en su infancia, las técnicas de análisis de imágenes médicas están llegando progresivamente al estudio de los cerebros fetales. El objetivo de esta tesis es desarrollar métodos automáticos de segmentación que puedan aplicarse a cerebros en distintas etapas de la vida, incluyendo el periodo gestacional, e investigar el desarrollo del cerebro fetal bajo condiciones anormales. Las principales contribuciones de esta tesis son dos:

1. Segmentación del cerebro en imágenes de resonancia magnética: hemos propuesto dos nuevos métodos de segmentación que pueden utilizarse para segmentar cerebros en cualquier etapa vital. Estos métodos abordan el problema de la segmentación desde dos perspectivas diferentes pero complementarias. El primero intenta mitigar los errores sistemáticos causados por fallos en el registro. Cada atlas de entrenamiento produce un *confidence map* basado en la calidad del registro que pondera la contribución de su labelmap en la segmentación final. El segundo método se centra en reducir el impacto de errores de interpolación, que son inherentes a la mayoría de métodos de segmentación tradicionales. El registro sólo se usa para establecer correspondencias espaciales, mientras que la información del atlas se extrae de su espacio nativo, respetando, por lo tanto, el protocolo de segmentación manual utilizado por el experto.

2. Análisis de fetos con ventriculomegalia: la ventriculomegalia es una condición en la cual uno o ambos ventrículos laterales están dilatados. En algunos fetos con ventriculomegalia, se observa un deterioro en el neurodesarrollo posnatal. Para entender mejor el neurodesarrollo intrauterino, hemos analizado la relación entre la dilatación ventricular con alteraciones en el desarrollo cortical. Primero, hemos estudiado esta asociación usando el volumen de los ventrículos laterales o el diagnóstico como descriptor de la ventriculomegalia. Después, hemos empleado análisis estadísticos y regresión lineal sparse para evaluar la

relación desde dos enfoques diferentes. En el segundo estudio, hemos adoptado un enfoque holístico al incorporar las superficies ventriculares, en vez de usar un único valor escalar (p.ej., volumen ventricular) para caracterizar la dilatación ventricular. Para ello, hemos propuesto un enfoque novedoso que analiza conjuntamente las superficies corticales y ventriculares, basado en sus patrones de crecimiento, lo que permite descubrir asociaciones más detalladas entre ambas estructuras.

Preface

This thesis would not have been possible without the encouragement and support received from many people. First and foremost, I would like to thank my supervisors, Dr. Gerard Sanroma and Dr. Gemma Piella, for their unceasing help, valuable advice and inspirational guidance. It has been a privilege to work under your supervision. I am also grateful to Dr. Miguel Ángel González Ballester for offering me the opportunity to get involved in such an interesting project, and to Dr. Elisenda Eixarch and Nadine Hahner for all the discussions and their support on the clinical aspects of my research. I would also like to thank Dr. Gang Li for giving me the chance to be part of his research group during my stay in the University of North Carolina at Chapel Hill.

Many thanks to my colleagues from SIMBIOsys and Physense research groups for contributing to improving my research and making these years a pleasant journey. Thanks to my colleagues from the AI group for all the roaming through our existential doubts about the PhD in the breaks. I would also like to thank my colleagues from the University of North Carolina at Chapel Hill for making my stay such a wonderful and productive experience.

Finally, I am very grateful to my family and friends for standing next to me and making me laugh during the difficult times.

Contents

List of Figures

List of Tables

Acronyms

AD	Atrial Diameter
CHD	Congenital Heart Disease
CP	Cortical Plate
CSF	Cerebrospinal Fluid
EM	Expectation-Maximization
GA	Gestational Age
GM	Gray Matter
GMAT	Germinal Matrix
GW	Gestational Week
IMVM	Isolated Mild Ventriculomegaly
INSVM	Isolated Non-Severe Ventriculomegaly
IUGR	Intrauterine Growth Restriction
IZ	Intermediate Zone
LF	Label Fusion
MAS	Multi-Atlas Segmentation
MNI	Montreal Neurological Institute
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MV	Majority Voting
PVE	Partial Volume Effect
SCM	Spatial Confidence Map
SP	Subplate
SVM	Support Vector Machine
T1w	T1-weighted
T2w	T2-weighted
US	Ultrasound
VM	Ventriculomegaly
WM	White Matter

1

Introduction

The aim of this thesis is to explore and develop novel segmentation and analysis approaches to study the human brain and expand our knowledge base on *in utero* brain development.

1.1 Research context

Neuroimaging offers an unprecedented means to access and study the human brain. Nevertheless, acquired raw brain images must first undergo a set of preprocessing steps before analysis. One important task among this preprocessing pipeline is segmentation. Although a wide range of methods for automatic brain segmentation has been proposed in the literature, there is still room for improvement. Most existing methods addressed or were only evaluated on adult brain datasets. In recent years, the field has witnessed an increasing interest in developing segmentation techniques for neonates and fetuses. In this context, the development of generic, instead of population-specific (e.g., adults or fetuses), segmentation methods is challenging. Note that, population-specific, here, does not refer to the use of atlases from the same population (e.g., using adult brain atlases to segment adult brains), but rather to other forms of prior knowledge such as the characteristic features of the fetal brain to improve segmentation. For instance, to accurately delineate the main tissues of the fetal brain, most existing methods adopt an age-specific approach in order to take into consideration the dynamic cerebral changes occurring in very short timeframes at such early stages. In this thesis, we tackle the segmentation problem using a multi-atlas approach, which is composed of two main stages: registration and label fusion. Registration is used to align the atlases and propagate their labelmaps to the target image, while label fusion seeks an optimal combination of the labelmaps to find a consensus segmentation. Our aim is to develop segmentation approaches that are general enough to be used with different lifespan brain databases.

After preprocessing, we are in position to perform our analyses. The second part of this thesis focuses on the study of intrauterine neurodevelopment. Although this research area is still in its infancy, the study of the fetal brain is recently attracting the interest of the research community. Understanding the maturational cerebral processes occurring during the gestational period in the fetal brain is of great importance to the neuroscientific community and of added value to clinical decision-making. Important cerebral growth processes take place during this period that have a striking impact on the

shape, size, and appearance of the fetal brain. The study of the normal fetal brain serves to gain more insight into *in utero* neurodevelopment and establish the normative course of such processes. Hence, this can help in identifying and characterizing *in utero* cerebral maldevelopment in the presence of malformations. Conversely to the adult brain, for example, there are still no well-established tools to study *in utero* neurodevelopment. Delving into the fetal brain requires the development of novel tools and approaches that are challenged by its idiosyncratic nature. This thesis aims to contribute to this trend with methodological approaches and shed light on the putative maturational deviations under *in utero* maldevelopment, with the purpose of filling important research and knowledge gaps in the analysis of the fetal brain.

1.2 Contributions

The work carried out during my PhD can be grouped in two main research areas: segmentation of brain MR images and analysis of *in utero* neurodevelopment.

1.2.1 Brain MRI segmentation

In this research area, we contributed with two multi-atlas segmentation frameworks that approach the segmentation problem from a general perspective, without explicitly addressing the fetal brain, but rather tackling important methodological and design choices inherent to most conventional segmentation approaches in order to improve automatic segmentation performance, regardless of the target population.

1. We developed a multi-atlas segmentation framework that is robust to registration failures. For example, when working with the fetal brain, registration failures may come from different sources, most of which are caused by the rapid *in utero* growth processes in case of normal neurodevelopment. Cortical folding, for instance, is governed by an intensive sulcogyral formation that changes the shape of the fetal brain. For segmentation, registering a highly gyrencephalic fetal brain to a fetus with a smoother cortical sheet might propagate the registration errors to the label fusion process and, thereafter, produce a poor

segmentation. The proposed framework equips each training atlas with an online *confidence map* estimator to weigh the contribution of its labelmap in the label fusion process based on the spatially-varying quality of the registration between the atlas and the target image. For a given training atlas, registration quality is assessed and learned offline from the rest of atlases using a supervised approach. Continuing with the previous scenario, the contribution of the atlas corresponding to the highly gyrencephalic brain to the final segmentation would be decreased in cortical regions to account for its high dissimilarity to the smooth target brain, expressed in terms of registration errors.

2. We developed a multi-atlas segmentation framework that is robust to interpolation artifacts. Besides registration errors, conventional segmentation approaches suffer from interpolation artifacts when warping the atlas intensity images and labelmaps to the target image. The main reason is that different interpolation strategies are used when warping the intensity images and labelmaps, using linear and nearest neighbor interpolations, respectively, for example. The proposed framework is based on the patch-based label fusion approach and uses registration to only establish spatial correspondences, without warping the atlases to the target image. In this way, intensity patches and their labels are extracted from the images native spaces, which correspond to the true images employed and manual segmentations created by the expert. Furthermore, the benefit of using intensity and label information extracted from the atlas native spaces is that intensity patches are not distorted, due to interpolation, to have a similar appearance to the target patch, respecting, therefore, the appearance changes between different brains, such as the changes occurring in the fetal brain due to myelination.

1.2.2 Analysis of *in utero* neurodevelopment

This part is focused on the analysis of *in utero* neurodevelopment under lateral ventricular dilation, particularly, in fetuses diagnosed with isolated non-severe ventriculomegaly, which is characterized by a mild enlargement of the lateral ventricles with no other anomalies. Prognosis in such cases is predominantly positive, but few cases will have an adverse outcome. In the search for prognostic biomarkers that could be suitable as risk indicators of poor outcome, we specifically explore the effects of isolated non-

severe lateral ventricular enlargement in cortical folding from two different methodological perspectives:

1. We analyzed the associations of isolated non-severe ventriculomegaly with alterations in cortical folding in a regional basis using curvature-based folding measures to describe gyrification. In this study, ventriculomegaly was characterized using a single scalar: diagnosis or ventricular volume. Statistical analysis and sparse linear regression approaches were used to investigate the relationships of lateral ventricular enlargement with different cortical regions. Our findings reveal that curvature-based descriptors were able to capture alterations in cortical folding, lateral volume offered greater sensitivity and, most importantly, several cortical regions showed deviations in curvature from normative development in the abnormal cohort, with ipsilateral associations being more prominent than global associations.

2. We proposed a novel approach to study the relationship between isolated non-severe ventriculomegaly and alterations in cortical folding that characterizes ventriculomegaly by incorporating the lateral ventricular shapes rather than using a single scalar value. This approach allows us to analyze the associations at a more fine-grained level. Correlations between the growth patterns of both sets of cortical and ventricular shapes were used to find a low-dimensional space where correlated regions from both shapes would lie close to each other and, therefore, reveal the associations between lateral ventricular enlargement and altered cortical folding. This is the first approach in the literature that explores the effect of fetal ventriculomegaly in cortical convolutions using ventricular shapes.

1.3 Outline of the thesis

The thesis is organized into 6 chapters. Chapters 2-6 are self-contained and each of them corresponds to a published or under review paper, while Chapter 7 contains conclusions and future work.

Chapter 2. This chapter corresponds to a review paper on the state of the art of fetal brain MRI. In this chapter, we introduce the area of research and establish the scientific context in which the work of this

thesis has been carried out. We provide a comprehensive description of the peculiarities and challenges of working with the fetal brain and present an in-depth overview of the ongoing research on *in utero* neurodevelopment from both methodological and clinical perspectives.

Chapter 3. This chapter presents a multi-atlas segmentation framework. The main objective of this framework is to calibrate the contribution of each atlas labelmap to the final labeling of a given target image by considering the registration quality between this atlas and the target image. Experiments are carried out on the segmentation of subcortical structures from 2 well-known adult brain MRI datasets and on tissue segmentation of fetal brain MR images.

Chapter 4. This chapter presents a multi-atlas segmentation framework that improves the widely-used patch-based label fusion framework. In this chapter, we revisit the patch-based label fusion framework, identify its main design choices and propose important changes that result in performance gains and considerably reduced computational complexity. Experiments are carried out on subcortical structure segmentation of adult brain MRI and fetal brain MRI tissue segmentation.

Chapter 5. In this chapter, we analyze the associations of ventriculomegaly with deviations in cortical folding from normative development by characterizing this condition using the diagnosis as a binary indicator or lateral ventricular volume. Cortical folding was characterized using several curvature-based descriptors to capture different changes in curvature. Global and ipsilateral analyses were employed to study the relationships of ventriculomegaly with the different folding measures on a regional basis. Findings obtained from these analyses were further assessed using sparse linear regression. With both approaches, relevant associations of ventricular enlargement with alterations in cortical folding were found to be in great overlap.

Chapter 6. While following the same purpose of Chapter 5, this chapter undertakes an entirely novel approach to investigate these associations, based on jointly analyzing different anatomical shapes using their growth patterns. In this chapter, we study the role of ventricular surfaces and their association with alterations in cortical folding, moving from characterizing ventriculomegaly with a single scalar value based on lateral ventricular volume to a more sophisticated analysis

based on joint spectral embedding that integrates the ventricular surfaces.

Chapter 7. This chapter summarizes the most important ideas and contributions of this thesis in both brain segmentation and *in utero* neurodevelopment analysis. We highlight the strengths and shortcomings of the works carried out within this thesis, discuss the potential gains and advancements in fetal brain analysis and propose promising directions for future work.

2

Toward the automatic quantification of *in utero* brain development in 3D structural MRI

Abstract – Investigating the human brain *in utero* is important for researchers and clinicians seeking to understand early neurodevelopmental processes. With the advent of fast magnetic resonance imaging (MRI) techniques and the development of motion correction algorithms to obtain high-quality 3D images of the fetal brain, it i1s now possible to gain more insight into the ongoing maturational processes in the brain. In this paper we present a review of the major building blocks of the pipeline towards performing quantitative analysis of *in vivo* MRI of the developing brain and its potential applications in clinical settings. The review focuses on T1- and T2-weighted modalities, and covers state of the art methodologies involved in each step of the pipeline, in particular, 3D volume reconstruction, spatio-temporal modeling of the developing brain, segmentation, quantification techniques and clinical applications.

This chapter is adapted from:
Benkarim O. M., Sanroma G., Zimmer V. A., Muñoz-Moreno E., Hahner N., Eixarch E., Camara O., González Ballester M. A., and Piella G. (2017). Toward the automatic quantification of in utero brain development in 3D structural MRI: A review. *Human Brain Mapping*, 38:2772-2787. https://doi.org/10.1002/hbm.23536

2.1 Introduction

Quantitative image analysis of the *in vivo* fetal brain plays a crucial role in clinical decision-making and neuroscience research. During the last several years, a growing trend in using magnetic resonance imaging (MRI) for such studies is observed. MRI is the most common medical imaging modality for the diagnosis and follow-up of patients with brain abnormalities, and the understanding of normal neurodevelopment in adult brains. For fetuses, although ultrasound (US) is widely accepted as the primary technology for *in utero* imaging of the brain (Garel, 2008), US examination is often hampered by some limitations, including reduced amniotic fluid volume, maternal obesity, inappropriate fetal head position, multiple pregnancy, and bony reverberation artifacts from the skull (Glastonbury and Kennedy, 2002; Twickler et al., 2003). On the other hand, MRI offers superior contrast in soft tissues and an increased field of view compared to US. Prenatal diagnosis can therefore benefit from fetal MRI by complementing the findings in US. Indeed, in recent years, *in utero* MRI has shown to be of important added value in the study of disorders (Clouchoux et al., 2013; Kyriakopoulou et al., 2014; Scott et al., 2013) and early brain development (Clouchoux et al., 2012; Wright et al., 2014).

Quantification of fetal brains from MRI is more challenging than that of the adult brain since it requires additional processing techniques and makes some of the widespread techniques for adult brain MRI not applicable. In the pipeline to perform fetal brain MRI quantitative studies (see Figure 2.1), the first challenges come from motion artifacts during image acquisition. While advances in fast MRI sequences help decrease acquisition times, the development of motion correction techniques allows obtaining high-resolution 3D images of the fetal brain from the several motion-corrupted acquired stacks. These studies also require the delineation of tissues and structures of interest in the 3D reconstructed volumes. For this purpose, automatic segmentation techniques are desirable over manual labeling since the latter is very time-consuming and subject to inter- and intra-rater variability. Therefore, quantitative approaches often rely on automated segmentation algorithms to achieve accurate and reproducible measurements. However, in this third stage of the pipeline (i.e., segmentation), researchers have to face new difficulties concerning the nature of the fetal brain. The rapid and complex cerebral changes in shape and appearance (e.g., transient laminar pattern and myelination) that occur during intrauterine growth make existing techniques for adult brains unfeasible and

advocate for the development of novel approaches. This has increased the need to build spatio-temporal atlases of the fetal brain in order to capture these dynamic changes. Spatio-temporal atlases have shown to be useful for automatic segmentation, which is notably challenged by the low resolution of fetal brain images, the excessive amount of partial volume effects (PVEs), and the substantial dissimilarities in shape and MRI contrast between brains at different gestational ages (GAs).

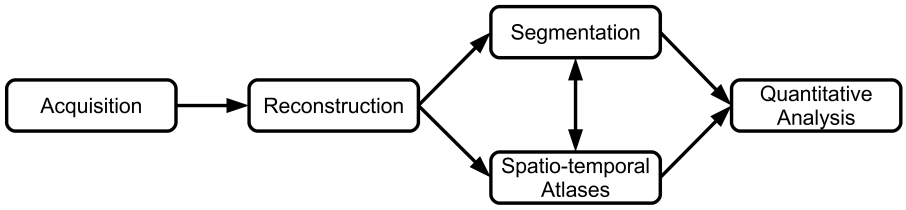

Figure 2.1: Pipeline to perform quantitative analysis of fetal brain MRI. Fast MRI sequences are used in the first stage to acquire several motion-corrupted stacks of the fetal brain, which are then used to obtain the final motion-corrected 3D reconstruction (second stage). The third stage is for approaches to build spatio-temporal atlases in order to capture the dynamic changes of the fetal brain, which can serve as spatial priors in the segmentation of brain tissues and other structures of interest. The fourth stage is dedicated to quantitative studies (e.g., volumetry, gyrification).

There is a growing body of literature on fetal brain MRI. This work aims to provide an overview of the automated pipeline for performing quantitative analysis of fetal brain from structural MRI. Acquisition and reconstruction stages of the pipeline are not the main purpose of this review and are briefly discussed in Section 2.2. The reader is referred to the work by (Studholme, 2011, and references therein) for a comprehensive review of these techniques. For each pipeline stage, the motivation and the difficulties that arise when working with fetal brain MRI are outlined from both clinical and methodological perspectives, and a literature review of the methodological advances is presented. Although this work mainly focuses on methods targeting the fetal brain in structural MRI, methods proposed for neonates are also discussed given the shared similarities once fetal-exclusive limitations are overcome. The rest of the paper is organized as follows. Section 2.2 overviews some of the challenges that arise in fetal brain MRI. Section 2.3 introduces the construction and use of spatio-temporal atlases. Section 2.4 is devoted to segmentation techniques of tissues and other anatomical structures in fetal and neonatal brains. In Section 2.5, a broad view of quantitative studies of normal brain development is provided. Section 2.6 describes potential clinical applications of fetal MRI in the study of early brain abnormalities.

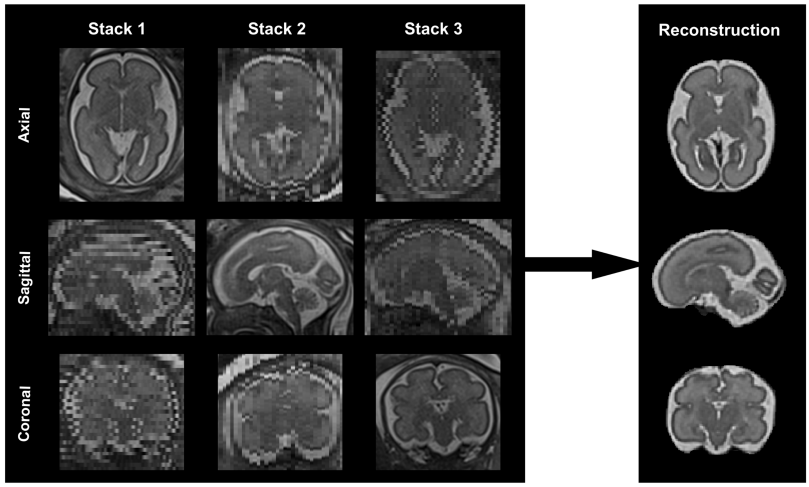

Figure 2.2: Example of reconstruction from motion-corrupted stacks of 2D MRI slices. First 3 columns (i.e., stacks 1, 2, and 3) correspond to axial, sagittal, and coronal acquisitions, respectively. Fourth column is the final reconstruction. Rows, from top to bottom, show axial, sagittal and coronal views, respectively, for each image.

A discussion about the state of the art is presented in Section 2.7. Finally, Section 2.8 concludes the paper.

2.2 Challenges of fetal brain MRI

Difficulties with fetal brain MRI start to appear as early as in the imaging process. Acquisition of full 3D MRI of the fetal brain is still impractical due mainly to the thick slice acquisition necessary to achieve good signal-to-noise ratio and the presence of motion artifacts caused by spontaneous movement of the fetus and maternal breathing. Shortening the acquisition time would help decrease the likelihood of motion artifacts (Malamateniou et al., 2013). Advances in fast MRI sequences, such as single shot fast spin-echo, in conjunction with post-processing techniques (i.e., motion correction and super-resolution) have granted the means to gain broad insight into the *in utero* fetal brain by providing high-resolution 3D volumes. Typically, several motion-corrupted stacks of thick 2D slices are acquired in orthogonal orientations and then used to reconstruct a high-resolution motion-free 3D volume of the brain. Existing methods for motion correction and reconstruction (Kim et al., 2010; Murgasova et al., 2012) require the delineation of the fetal brain (at least in one slice) from maternal tissue.

However, Keraudren et al. (2014) proposed a fetal brain extraction method from 2D stacks, that combined with the reconstruction approach in (Murgasova et al., 2012) provides a fully automated pipeline to obtain the final 3D volumes. A volume reconstruction example using this pipeline is shown in Figure 2.2. Recently, a GPU-accelerated slice-to-volume reconstruction method similar to (Murgasova et al., 2012) was proposed in (Kainz et al., 2015).

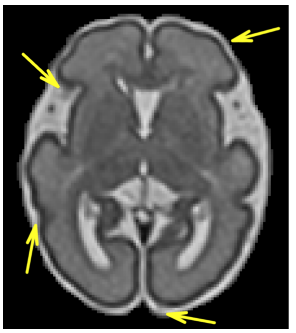

Figure 2.3: Examples of partial volume effect (arrows) in the boundaries between gray matter and cerebrospinal fluid, and background and cerebrospinal fluid in an axial MRI slice of a 29 GWs fetus.

In quantitative MRI studies, one of the requirements is the labeling of the different regions of the brain. Nevertheless, segmentation is not forthrightly applicable as MRI suffers from noise, intensity inhomogeneity and PVE that may negatively impact the performance of image processing techniques. PVE is present when multiple tissues contribute to a single voxel producing a blurring effect, for instance, in the boundary between gray matter (GM) and cerebrospinal fluid (CSF). PVEs can lead to volume measurement errors in the range of $20\% - 60\%$ (González Ballester et al., 2002) and are more recurrent in fetal brain MRI, as illustrated in Figure 2.3, due to its lower resolution when compared to adult brains images. In fetuses, these problems are accentuated because early brain development involves a rapid and complex sequence of morphological, functional, and appearance changes. Hence, in addition to the aforementioned obstacles, large tissue intensity variations are present in fetal (and neonatal) brains due to myelination and cell migration (Rutherford, 2001). Myelination is the last stage of white matter (WM) development and takes place from the second half of gestation to the end of adolescence (Dubois et al., 2014), making the intensities of WM similar to those of both cortical and subcortical GM in T1-weighted (T1w) and T2-weighted (T2w) MR images. Neuronal migration occurs from

the germinal matrix (GMAT), which is comprised of ventricular and subventricular zones, towards the cerebral wall to form the prospective neocortex. During this process of corticogenesis, the fetal brain goes through an intensive laminar organization, where the cortical plate (CP), the subplate (SP) and the intermediate zone (IZ) are formed, as shown in Figure 2.4. By the 27th gestational week (GW), however, intensity contrast between the IZ and SP begins to overlap and eventually transforms into neonatal WM (Prayer et al., 2006). Also late in the 2nd trimester, the GMAT commences to gradually regress until it completely disappears by term (Girard and Chaumoitre, 2012). These developmental processes make the segmentation of the developing brain in MRI more challenging as compared to adult brains.

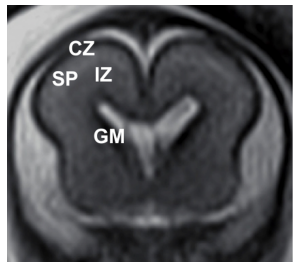

Figure 2.4: Coronal T2w slice of a fetus at 23 GWs illustrating the laminar organization of the brain: the hypointense CP is the outermost layer. The SP is hyperintense relative to the CP and IZ. Immediately beneath the SP, the hypointense IZ appears. The GMAT represents the innermost layer and is isointense to the CP.

2.3 Atlases of the developing brain

The core *raison d'être* of brain atlases is to (1) serve as a common reference coordinate system for spatial normalization of a group of individuals to study intra- and inter-group variability, and (2) act as an atlas for segmentation of brain regions (Fonov et al., 2011, and references therein). The study of the developing brain ought to be age-specific given transient laminar pattern and the evident vast differences in shape and appearance across age (see Figure 2.5) that occur during the maturational process. This age-specific character of segmentation and processing techniques of the developing brain has driven research (e.g., Habas et al., 2010; Serag et al., 2012b) towards the use of spatio-temporal atlases instead of a single atlas at a particular time-point. Compared to conventional atlases, spatio-temporal

atlases encode spatial as well as temporal variability. This allows to better retain the anatomical variability across age. Several atlases can be found in the literature for adult (Evans et al., 1993), pediatric (Wilke et al., 2003), infant (Joshi et al., 2004), neonatal (Murgasova et al., 2011; Serag et al., 2012a) and fetal (Habas et al., 2010) brains.

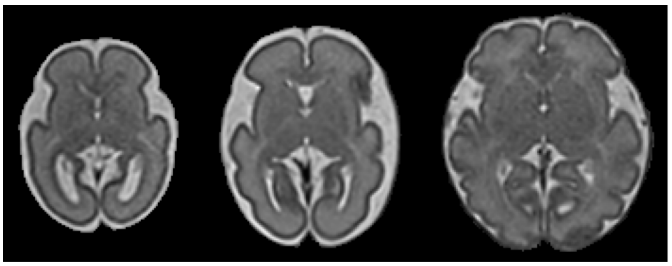

Figure 2.5: Rapid brain maturation. From left to right: 26, 29 and 34 GWs fetal brains.

2.3.1 Population-specific atlases

An ideal atlas of the human brain should have the desirable features of being (1) representative of the population, (2) unbiased and (3) sharp (i.e., with high contrast). Using a single anatomy as an atlas precludes fair representation since the arbitrary choice of the reference template does not encompass the neuroanatomical variability of the entire population. To better accommodate this variability, population average atlases, such as the Montreal Neurological Institute (MNI) template (Mazziotta et al., 1995) for adult brains, were constructed. These atlases are built by averaging the anatomical images from a particular population, based on distinctive criteria such as age, gender or ethnicity. The representative bias introduced when using a single-subject atlas can then, to some extent, be avoided by using the MNI template. Still, when targeting pediatric brains, Wilke et al. (2003) found considerable differences in tissue distribution between pediatric and adult data, substantially appreciated in GM. Thus, spatial normalization of pediatric brain images to the MNI or other adult templates is less accurate (Shi et al., 2011) and might introduce a strong bias in anatomical quantification. This problem is especially important in younger brains due to their continuous development throughout childhood and adolescence (Paus et al., 1999). Several researchers have therefore developed population-specific brain atlases for children and infants. Joshi et al. (2004) constructed a probabilistic atlas of anatomical structures from 8 T1w MR scans of 2-year-old children.

Wilke et al. (2008) created reference pediatric templates from 404 healthy subjects aged 5-18 years. Recently, Fonov et al. (2011) built age-specific MRI atlases for 6 age groups in the range of 4.5-18.5 years. Along the same lines, Altaye et al. (2008) proposed a method to create an infant (9 to 15 months) probabilistic atlas, and demonstrated that the use of their atlas performed better than a default adult or pediatric template in segmenting the infant brain. Kazemi et al. (2007) built a neonatal brain template based on T1w MR images of 7 individuals with GA between 39 and 42 weeks.

The major requirements for the construction of these adult and pediatric atlases are twofold: (1) the definition of a reference space, and (2) the transformation model that maps each brain to this common space. A potential source of bias comes forth in the selection or creation of the initial reference template during atlas construction. For example, Park et al. (2005) selected the initial template to be the closest brain to the geometrical mean in a low dimensional space. However, to reduce the dependence on any particular anatomy during normalization, the conventional approach is based on iterative strategies where the initial template is derived from the multiple anatomies in the database and successively updated in each iteration (Avants et al., 2010; Fonov et al., 2011). Sharpness of atlases, on the other hand, is closely related to the transformation model. A single-subject based atlas has well defined anatomical boundaries. For the average atlases to exhibit sharp anatomical details, non-rigid registration seems to be more appropriate for building such atlases (e.g., Fillmore et al., 2015; Fonov et al., 2011). In order to further enhance the anatomical details, Shi et al. (2014) proposed a patch-based sparse representation approach to fuse the information from the individual images after registration.

2.3.2 Spatio-temporal atlases

Methods for building atlases of the human brain have evolved in parallel to the emergence of new imaging techniques, being closely linked to the age of the individuals under study because of the substantial structural changes existing between age groups. Atlases described above either were produced for infant and older populations, or have a sparse age coverage (i.e., temporal variability is covered by only a few discrete temporal points). The study of neonates and fetuses at a precise developmental period becomes difficult because changes in the developing brain occur in the order of weeks or even days. One way to circumvent this limitation is to provide atlases with a fine-grained temporal resolution. Beyond static population-specific atlases,

the solution relies on the building of spatio-temporal atlases, whose purpose is to encode both longitudinal and inter-subject variability. For the temporal domain, brains along a certain age range need to be taken into account in the atlas creation, whilst, for the spatial domain, a sufficient number of subjects at a particular time-point is needed. Due to ethical and practical issues, building spatio-temporal atlases directly from repeated longitudinal imaging of the same subject is difficult, and they are therefore constructed from many individuals, scanned at different ages. Several methods exist in the literature to build this kind of atlases for neonates (Murgasova et al., 2011; Serag et al., 2012a,b; Zhang et al., 2016) and fetuses (Dittrich et al., 2014; Gholipour et al., 2014; Habas et al., 2010; Serag et al., 2012b). Table 2.1 summarizes the main features and datasets used for each of these methods. The time-varying dimension of these atlases is an advantageous characteristic in that it allows dynamic generation of average intensity images and corresponding prior tissue probability maps at any arbitrary time-point confined within the age range of the brain scans used to build the atlas.

Most of the aforementioned spatio-temporal atlases were created from T2w images using non-rigid (Dittrich et al., 2014; Gholipour et al., 2014; Habas et al., 2010; Serag et al., 2012a,b; Zhang et al., 2016) rather than rigid registration (Dittrich et al., 2014; Murgasova et al., 2011). Rigid registration is better suited for the spatio-temporal atlases to capture the variability across subjects, and this was used in (Dittrich et al., 2014) for age estimation. For segmentation purposes, non-rigid registration provides more accurate results. Although, as pointed out in (Murgasova et al., 2011), building the atlas with affine transformations may turn out to be advantageous for segmentation applications when the to-be-segmented image is non-rigidly registered to the newly created template. Furthermore, the spatio-temporal latent atlas in (Dittrich et al., 2014) was built in a semi-supervised manner, where segmentations of ventricles and cortex were only available for a reduced set of images. During the construction stage, these segmentations are transferred to the remaining images to create prior probability maps for the generated atlas. With regard to image modality, the neonatal spatio-temporal atlas in (Serag et al., 2012b) was created from both T1w and T2w modalities. Although this multi-channel atlas uses T1w and T2w images, registrations were only performed on T2w images and resulting transformations were used to deform the T1w modality. A neonatal spatio-temporal atlas for both T1w and T2w modalities was also created in (Zhang et al., 2016).

	Modality	Fetal/ Neonatal	Subjects	GA	Structures	Type of segmentation	Registration	
							Spatial	Temporal
(Habas et al., 2010)	T2w	Fetal	20	21-25	GM, WM GMAT VENT	Manual	Groupwise Non-rigid Segmentations	LS
(Murgasova et al., 2011)	T2w	Neonatal	142	29-44	CoGM CeGM WM, CSF BS, CB	Automatic	Pairwise Affine	KR
(Serag et al., 2012a)	T1w and T2w	Neonatal	204	29-44	CoGM CeGM WM, CSF BS, CB	Automatic	Pairwise Non-rigid For each GA (only T2w)	AKR
(Serag et al., 2012b)	T1w and T2w	Both	204 80	29-44 (neonatal) 22-39 (fetal)	CoGM CeGM WM, CSF BS, CB	Automatic (neonatal) Manual (fetal)	Pairwise Non-rigid For each GA (only T2w)	AKR
(Dittrich et al., 2014)	T2w	Fetal	12	21-25	CoGM VENT	Manual (only few)	Groupwise Rigid or Non-rigid	KR
(Gholipour et al., 2014)	T2w	Fetal	40	26-36	CoGM CeGM GMAT WM, CSF BS, CB	–	Groupwise Non-rigid	AKR
(Zhang et al., 2016)	T1w and T2w	Neonatal	35 (150 scans)	0-12 (months)	GM, WM CSF	Automatic	Groupwise Non-rigid	AKR

Table 2.1: Spatio-temporal atlases of the developing brain. Abbreviations: ventricles (VENT), central GM (CeGM), cortical GM (CoGM), brainstem (BS), cerebellum (CB), LS (least squares fitting), KR (kernel regression) and AKR (adaptive kernel regression). Unavailable information is marked as –.

A distinguished feature that needs to be considered in the construction of spatio-temporal atlases is the modeling of the temporal dimension. The first work on building a spatio-temporal atlas of the fetal brain (Habas et al., 2010) used non-linear modeling of the temporal variations in a reduced age range of 21-25 GWs. However, in larger age ranges, there may be no subjects at the exact age of interest. Therefore, the works in (Dittrich et al., 2014; Murgasova et al., 2011) adopted a temporal kernel regression method to compute the weighted contribution of the temporal neighbors in the creation of the average brain templates. In (Gholipour et al., 2014; Serag et al., 2012a,b), adaptive kernel regression was used to create their spatio-temporal atlases. Here, the adaptive kernel accounts for a sufficient number of subjects in the contribution to the atlas creation. This ensures a consistent level of detail for a synthesized atlas at any time-point when the brains are not uniformly distributed over the age range. (Zhang et al., 2016) used a patch-based approach similar to (Shi et al., 2014) for creating the atlas. However, the key improvement of this work over previous approaches is the temporal consistency of their atlas, since it was built based on subject-specific longitudinal information (i.e., using 4.3 scans per subject in average).

2.4 Segmentation of brain images

Automatic tissue segmentation of the human brain in MRI has been incentivized by its many clinical applications. Cortex delineation, for instance, is a prerequisite for the study of cortical thickness and gyrification (see Section 2.5). The success of quantitative analysis is heavily sustained by the accuracy of image segmentation algorithms. While there is a plethora of MRI segmentation techniques for the adult brain (e.g., Greenspan et al., 2006; Ortiz et al., 2013; Pham and Prince, 1999), a limited number of works to segment brain MRI of neonates and fetuses exists in the literature. The poor spatial resolution of images, and the varying intensity distribution between tissues, dynamic shape changes and reduced contrast in the brain at such early ages render more challenging the automatic brain segmentation of this age group compared to adults.

2.4.1 Segmentation of neonatal brain MRI

In the neonatal brain, WM exhibits substantial intensity variation due to the ongoing process of myelination that gradually reverses the WM-GM contrast, reaching a point around the ninth month when both tissues appear isointense (Barkovich, 2005). This may mislead intensity-based segmentation algorithms to identify PVEs at the boundary between CSF and GM as the yet unmyelinated WM (Xue et al., 2007). In order to cope with these systematic segmentation errors in neonates, several authors have proposed brain tissue segmentation methods combining intensity information with spatial priors and contextual information.

Among these works, an atlas-free approach (Xue et al., 2007) adopted the Expectation-Maximization (EM) algorithm with Markov Random Field (MRF) regularization for tissue classification combined with a knowledge-based strategy to correctly classify mislabeled PVEs at the CSF-GM boundary. Recently, another approach that does not require atlas priors was presented in (Gui et al., 2012), where a watershed technique was used to segment the brain MRI of neonates based on information about tissue connectivity, structure and relative positions. Most of published works, however, employed atlas-based approaches to guide the segmentation. Prastawa et al. (2005) developed an atlas-based automatic algorithm based on graph clustering and outlier removal. It used the EM scheme followed by a non-parametric kernel density estimation to obtain the final segmentation. Shi et al. (2010) proposed a joint registration-segmentation framework that used subject-specific tissue probabilistic atlases generated with adaptive fuzzy C-means (Pham and Prince, 1999) from follow-up data of the same subject. With the same purpose, an atlas-based approach interleaving registration and segmentation was employed in (Shi et al., 2011), which combined subject-specific cortical GM distribution with a data-driven neonatal atlas. Weisenfeld and Warfield (2009) presented an algorithm that iteratively performs sample refinement, segmentation, and fusion using STAPLE (Warfield et al., 2004). Wang et al. (2011b) proposed a level set segmentation framework with a thickness constraint in the cortical area, and combined local intensity information and atlas priors. This work was further improved in (Wang et al., 2014c) by using a subject-specific atlas and incorporating non-local (i.e., patch) information. In (Ledig et al., 2012), the spatio-temporal atlas in (Murgasova et al., 2011) was used with an extended version of the EM algorithm incorporating a second-order MRF to penalize inconsistent labeling caused by PVEs. In a recent work, Anbeek et al. (2013)

segmented eight different tissue classes in a supervised manner using intensity and voxel coordinates with K-nearest neighbors. Wang et al. (2015) also proposed an iterative supervised approach to segment the main tissues using random forest (Breiman, 2001) and features from several modalities (i.e., T1w, T2w and fractional anisotropy images), including the estimated probability maps. On premature neonates, brain tissue segmentation was carried out in (Beare et al., 2016; Sanroma et al., 2016a) on the dataset provided by the NeoBrainS12 challenge (Išgum et al., 2015). Sanroma et al. (2016a) proposed an ensemble approach that optimally combines the outputs of two complementary segmentation approaches (i.e., intensity-based and multi-atlas label fusion). In (Beare et al., 2016), a supervised approach was presented using a combination of unified segmentation, template adaptation and topological filtering.

Beyond the segmentation of the main tissues, methods for the parcellation of anatomical structures in the neonatal brain have been proposed in (Gousias et al., 2013; Makropoulos et al., 2014). There is a notable overlap between tissue segmentation and parcellation in that both approaches make use of image intensities. Although in the latter case, intensity-based features are not sufficient as in the case of tissue segmentation, and spatial priors derived from atlases become also necessary. Therefore, Gousias et al. (2013) presented a multi-atlas approach to segment the brain MRI in 50 regions using ALBERTs (Gousias et al., 2012), a dataset of manually annotated neonatal atlases. These atlases were also used in a hierarchical approach to label multiple brain structures using EM-MRF and knowledge-based rules for misclassified voxels (Makropoulos et al., 2014).

2.4.2 Segmentation of fetal brain MRI

Automatic segmentation of fetal brain MRI is even more intricate since in addition to the existing difficulties in neonates, fetal brains have a transient laminar pattern (see Section 2.2) and MRI quality is highly affected by fetal and maternal movements during acquisition. There is much less literature on fetal brain MRI segmentation. Table 2.2 lists the main characteristics of existing methods. One of the first works (Claude et al., 2004) proposed a semi-automated method to segment 3 regions (i.e., posterior fossa, brainstem and vermis) on a single MRI slice of the fetal brain, using region growing based on intensity and gradient features. Also in 2D MRI, Bach Cuadra et al. (2009) developed an automatic labeling method for the main tissues that was performed independently in each slice using EM-MRF

and anatomical priors in form of a cortical distance map. Bayesian segmentation was first performed to segment the fetal brain in 7 different classes. GM and WM were modeled as a mixture of 2 Gaussians each, and 2 classes for mislabeled partial volume voxels, which were to be correctly classified during the MRF stage.

Among pioneering works in addressing the segmentation of fetal brains in 3D reconstructed MRI, Gholipour et al. (2011) used a semi-automated method with little user interaction based on intensity information, level sets, and morphological operations to deal with PVEs. Given that intensity information is insufficient to isolate cortical GM from WM, Caldairou et al. (2011) included anatomical priors through a topological K-means clustering algorithm to segment the cortex in an atlas-free approach. In (Habas et al., 2008), the GMAT was segmented in fetuses at a reduced age range (20.5-22.5 GWs) using the EM framework with a single probabilistic atlas. However, with the advent of spatio-temporal atlases of the fetal brain, age-specific atlases can be generated at any GA to serve as priors in the segmentation process (Habas et al., 2010; Serag et al., 2012b). Multi-atlas segmentation approaches were also used in (Gholipour et al., 2012; Koch et al., 2014). Gholipour et al. (2012) developed a method built over multi-atlas segmentations that incorporates a shape model of structures and regional intensity values within a probabilistic framework to achieve automatic segmentation of multiple shapes with similar intensities. The main purpose was to accurately segment the lateral ventricles in subjects with normal, dilated, or fused ventricles. In (Koch et al., 2014), a semi-supervised graph-based method was proposed to overcome the unavailability of subjects within certain GA ranges. After a first labeling stage where only atlases are able to propagate information, label probabilities of test images in subsequent iterations are also used to improve the labeling of images whose GA is not available in the training set. It is worth mentioning that several of the aforementioned segmentation methods for neonates and fetuses can be used interchangeably as long as the priors and the spatial constraints are consistent with the dataset under study. In fact, the method proposed in (Ledig et al., 2012) for neonates was used by Wright et al. (2014) to segment fetal brain MRI.

	Method	MRI	Spatial priors/constraints	Structures
(Claude et al., 2004)	Region growing Semi-automatic	2D (one slice)	–	Posterior fossa Brainstem Vermis
(Bach Cuadra et al., 2009)	EM-MRF	2D	Cortical distance map	Cortical GM Central GM WM CSF
(Gholipour et al., 2011)	Level sets Morphological op.	3D	–	Brain CSF
(Caldairou et al., 2011)	Topological K-means	3D	Anatomical priors	Cortical GM
(Habas et al., 2008)	EM-MRF	3D	Probabilistic atlas	GMAT
(Habas et al., 2010)	EM-MRF	3D	Probabilistic atlas	Cortical GM WM GMAT VENT
(Serag et al., 2012b)	–	3D	Probabilistic atlas	Cortical GM Lateral VENT Hemispheres
(Gholipour et al., 2012)	Multiatlas multishape	3D	Multiple atlases	Lateral VENT
(Koch et al., 2014)	Graph-based label propagation	3D	Multiple atlases	Lateral VENT

Table 2.2: Fetal brain MRI segmentation methods. Unavailable information or not used is marked as –.

2.5 Quantification of early normal brain development

Quantifying the patterns of normal gyrification and the underlying growth processes in the developing fetal brain from *in utero* MRI may offer insights into the changes that occur during normal fetal brain development, and provide a baseline for comparison to abnormal development. During early stages of fetal cerebral development, the brain is lissencephalic in appearance (Rutherford, 2001). However, as growth proceeds, the brain undergoes drastic changes in its morphology. In the latter half of gestation, the normal process of human brain maturation is manifested by substantial increases in volume without equivalent changes in thickness, and an increasing complexity of the CP following a highly orchestrated sequence of gyral-sulcal formation (Clouchoux et al., 2012; Rajagopalan et al., 2011a; Wright et al., 2014). In particular, this sequence occurs in a hierarchical manner in which primary and secondary sulci form in a consistent spatio-temporal pattern during normal gestation, followed by tertiary sulci that show increasing variability across individuals (Bendersky et al., 2006; Studholme, 2011). The timing of this has been considered an accurate marker of brain development (Garel et al., 2003), with any divergence from this pattern being conceived as a potential stable biomarker for abnormal functional development.

Several works inspect the global and regional patterns of tissue maturation in the developing brain to provide a comprehensive understanding of the maturational process the human brain embarks on from early weeks of gestation (Gholipour et al., 2011; Scott et al., 2013) up to approximately the second postnatal year (Aljabar et al., 2008), period where the majority of brain growth occurs (Rutherford, 2001). Other studies reported the patterns of cortical convolutions (Clouchoux et al., 2012; Wright et al., 2014) that take place in the human brain during this period. This section offers an overview of the role MRI plays in the quantitative study and assessment of *in utero* brain development.

2.5.1 Cortical folding

The study of cortical folding is important because will aid clinicians in the understanding of normal gyrification and the detection of cortical maldevelopment. Evidence from MRI studies of preterm neonates (Dubois et al.,

2008) may not be consistent with fetal brain development since prematurity *per se* may be considered a limitation in the representation of normal *in utero* neurodevelopment (Rutherford et al., 2008). Indeed, preterm-born infants compared with term-born controls at term equivalent age showed alterations in cortical volume (Padilla et al., 2014) and folding (Melbourne et al., 2014). When compared to fetal brains, prematurity also showed an impact on cortical folding (Lefèvre et al., 2015). In fetuses, conclusions drawn from *ex vivo* studies (e.g., Bendersky et al., 2006) about sulcal emergence might be influenced by deformations inherent to brain fixation and the substantial fluid loss during histological processing (Dubois et al., 2008; Habas et al., 2012), which may slightly affect the measurements. With the possibility to perform *in vivo* fetal MRI, mapping of cortical folding has been initially restricted to analysis of 2D slices (Garel et al., 2003; Prayer et al., 2006). This approach exhibits significant limitations in that 2D measurements are very dependent on the use of consistent 2D planes, which is susceptible to motion during acquisition. Furthermore, exact tissue boundaries may be difficult to find on thick 2D MRI slices due to PVEs. Volumetric reconstructions of the fetal brain, on the other hand, ensure selection of the appropriate planes and enable measurements that take advantage of the 3D anatomy of the brain.

To provide a reliable timeline of the normal *in utero* brain development, researchers have taken benefit of motion-corrected 3D reconstruction techniques, which can be easily integrated into an automated pipeline for the assessment and quantification of the timing of cortical folding. Using only cohorts of healthy fetuses, (Clouchoux et al., 2012) delineated sulcal fundi in 12 fetuses between 25 and 35 GWs, Habas et al. (2012) presented a temporal mapping of the emergence of individual sulci in 40 MRI scans from 38 fetuses with age ranging 20-28 GWs, and Wright et al. (2014) studied global and regional gyrification measures in 80 fetuses over a wider age range of 22-39 GWs. All of these works used a surface-based approach, where curvature information was estimated on the inner CP surfaces extracted from WM segmentation. Following the pipeline illustrated in Figure 2.1, segmentations in (Clouchoux et al., 2012) were obtained after 3D volume reconstruction (Gholipour et al., 2010) using an atlas-based approach with manual correction. In (Habas et al., 2012), images were first reconstructed with the method in (Kim et al., 2010) and automatically segmented using priors from a spatio-temporal atlas (Habas et al., 2010). Also in (Wright et al., 2014), images went through 3D volume reconstruction (Jiang et al., 2007) and automatic segmentation (Ledig et al., 2012).

Among their findings, (Habas et al., 2012) showed that the increase in surface area related to gyrification is linear from 20 to 28 GWs. After the 28th GW, gyrification accelerates and becomes more complex (Clouchoux et al., 2012), following a non-linear growth model. In particular, a Gompertz model showed to best fit the folding measures studied in (Wright et al., 2014). Furthermore, they found a positive correlation of these folding measures with GA, which was stronger than that of GA and volume. (Clouchoux et al., 2012) also created 4 average cortical templates evenly distributed along the age range of the subjects that showed the major changes in gyrification occurring during normal fetal brain development. A more precise timetable of early sulcation was reported in (Habas et al., 2012), demonstrating that 3D MRI provided more sensitivity than 2D MRI in the detection of sulcal emergence. Establishing normative timing for gyrification in the fetal brain will allow identification of deviations in development and early treatment.

2.5.2 Patterns of tissue maturation

Similarly as in the case of cortical folding, there exists a vast literature from *postmortem* (Huang et al., 2009), *ex utero* (Aljabar et al., 2008; Murgasova et al., 2011; Xue et al., 2007) and, *in vivo* fetal MRI (Kazan-Tannus et al., 2007; Limperopoulos et al., 2010; Prayer et al., 2006) and US (Endres and Cohen, 2001; Roelfsema et al., 2004) on tissue growth and laminar organization of the developing brain. However, as discussed in 2.5.1, generalizing the evidence from these studies to normal *in utero* brain maturation might render inadequate.

Based on reconstructed 3D MRI of the fetal brain, Gholipour et al. (2011) carried out a volumetric study using automated segmentations of brain tissue in a cohort of 25 fetuses ranging from 19 to 39 GWs. In (Corbett-Detig et al., 2011), global and local patterns of SP growth were studied through quantitative analyses of temporal changes in SP volume and thickness in 21 fetuses within the age of 20 to 26 GWs. From manual segmentations of the SP and supratentorial brain volume (as a sum of CP, WM, and GMAT), they also analyzed the relationship between volume and GA. From automatic segmentation of 48 scans of 39 fetuses with age between 21 and 31 GWs, Scott et al. (2011) presented volumetric growth trajectories of CP, SP and IZ, GMAT, deep gray nuclei, and ventricles. Morphometry was also used to study the complex series of local tissue volume changes the developing brain undergoes in its normal course toward acquiring its gyrencephalic

adult aspect. Aljabar et al. (2008) had previously used Tensor-Based Morphometry (TBM) (Davatzikos et al., 1995) in a longitudinal study to provide global and local growth factor estimates of GM and WM for a cohort of 25 preterm subjects scanned at 1 and 2 years. TBM uses accurate spatial normalization of brain anatomy into a common reference space to capture the pattern of regional structural differences across a set of anatomies by computing the derivatives (i.e., Jacobian map) of the deformation fields required to bring each anatomy to the same stereotaxic space (Studholme, 2011). For the *in utero* fetal brain, Rajagopalan et al. (2011a) provided a mapping of growth patterns by quantifying tissue locations that were expanding at a different growth rate than the overall cerebral tissue. TBM analysis combined with a linear model of age was used to create these maps from fetuses between 20 and 28 GWs. However, besides modeling magnitude of local tissue volume increase with scalar TBM (e.g., Dubois et al., 2008; Habas et al., 2012; Rajagopalan et al., 2011a), directional growth information is valuable in order to acquire more knowledge about fetal brain development. Rajagopalan et al. (2011b) extended the study beyond volume increase by incorporating its normal and tangential components on either side of the SP-CP interface, which permitted to model the variational growth patterns that underlie the mechanism of sulcation between both tissues. Furthermore, Rajagopalan et al. (2012) quantified brain development as a combination of volume and direction change patterns, and provided the principal growth direction at a particular location.

According to these studies, supratentorial brain volume increased quadratically with GA (Gholipour et al., 2011). From 20 to 26 GWs, the increase in SP tissue was proportional with the increase in supratentorial volume, although at different rates among brain regions (Corbett-Detig et al., 2011). Tissue-dependent growth rates where found in (Scott et al., 2011), with CP growing faster than all other tissue zones, especially along the midline surface of the frontal and parietal lobes (Rajagopalan et al., 2011a). Also, significant changes in direction of growth were found to occur primarily in the CP at locations corresponding to the formation of primary sulci. When the direction of cortical growth at any sulcus changes rapidly, it occurs in conjunction with change in direction of growth in the underlying cerebral mantle (Rajagopalan et al., 2012). Finally, slower growth was found in the ventricular regions adjacent to the CP, and the GMAT, which begins to regress after 25 GWs (Corbett-Detig et al., 2011; Rajagopalan et al., 2011a).

2.5.3 Interhemispheric structural asymmetries

Interhemispheric asymmetries have also been studied in fetal brain MRI to establish a precise timing for their *in utero* emergence. These asymmetries may be indicators of cortical functional specialization (Dubois et al., 2008), and alterations of this pattern may be useful as an early biomarker for abnormal neurodevelopment (Studholme and Rousseau, 2014). Using fractal dimension analysis in outer cortical surfaces reconstructed from 2D slices, Shyu et al. (2010) found earlier development of cortical complexity in the right hemisphere than in the left in a cohort of 32 fetuses with GA between 27 and 37 weeks. This was also reported in preterm (Dubois et al., 2008) and term (Hill et al., 2010) neonates, particularly evident at the level of the superior temporal sulcus. Interhemispheric asymmetries were also analyzed in 3D reconstructed fetal brain MRI by (Rajagopalan et al., 2011a), where the emergence of asymmetries was detected using TBM analysis based on symmetric groupwise registration of tissue maps of 40 fetal brains and their reflected versions along the sagittal midline. A similar approach was carried out in (Rajagopalan et al., 2012) to detect directional asymmetries in the same cohort with GA between 20 and 28 weeks. They found significant local asymmetries in volume and growth direction in the periSylvian fissure, showing that asymmetries in this area start around 20 GWs. Among their findings, (Habas et al., 2012) reported statistical significance of interhemispheric asymmetries in the periSylvian region by 23 GWs and in the parieto-occipital sulcus after 26 GWs.

In accordance with *ex vivo* studies and in conformity with asymmetries reported in adult and neonatal brains (Hill et al., 2010), these findings confirm that gyrogenesis occur earlier in the right hemisphere than in the left, and that cerebral interhemispheric asymmetries start during the intrauterine period.

2.6 Examples of clinical applications

Brain malformations account for one third of fetal anomalies and 60% have no identifiable etiology (Rodríguez et al., 2010). Although US is the standard imaging modality for fetal evaluation, it is well demonstrated that fetal MRI has a greater sensitivity to detect specific brain abnormalities that could be occult on prenatal US (Banović et al., 2014). In this section, several contributions of fetal MRI to the diagnosis of brain abnormalities *in*

utero are described. We will focus on quantification studies of Intrauterine growth restriction (IUGR), congenital heart disease (CHD), and ventriculomegaly (VM).

2.6.1 Intrauterine growth restriction

IUGR refers to a condition in which the weight of the fetus is below the 10th percentile for GA. This condition affects 10-15% of the population (Gardosi, 2011) and it is associated with a wide range of short- and long-term neurodevelopmental disorders. The effect of late-onset IUGR in the *in utero* development of the brain was studied in (Egaña-Ugrinovic et al., 2014, 2013) from 2D MRI slices. In (Egaña-Ugrinovic et al., 2013), differences in cortical development were assessed in 52 late-onset IUGR and 50 control fetuses. Late-onset IUGR fetuses, compared with controls, presented deeper fissures, more pronounced right asymmetry and smaller brain volumes. Corpus callosum development was analyzed in (Egaña-Ugrinovic et al., 2014) using 117 late-onset IUGR and 73 control fetuses. The area of the corpus callosum was significantly smaller in IUGR fetuses compared to the control group. Further, they found that these morphometric differences were in correlation with worse neurobehavioral performance. These findings reflect a perturbation in normal fetal brain development and can be used as potential biomarkers to predict abnormal neurodevelopment in pregnancies at risk.

2.6.2 Congenital heart disease

CHD refers to a structural abnormality of the heart present at birth. A wide spectrum of brain abnormalities has been identified with CHD in preterm and term neonates before they undergo cardiac surgery, which may suggest the occurrence of abnormal brain development *in utero* (Donofrio and Massaro, 2010). To better understand the impact of CHD in impaired neurodevelopment outcome, Limperopoulos et al. (2010) performed the first *in vivo* quantitative MRI study of 55 fetuses with CHD and 50 healthy fetuses between 25 and 37 GWs. Volumetric MRI analysis and spectroscopy showed a progressive deceleration in global brain growth (i.e., intracranial cavity and total brain volumes) and metabolism in the cohort with CHD over the 3rd trimester of gestation. Using *in utero* MRI, Mlczoch et al. (2013) reported a 39% incidence of brain abnormalities in a cohort of 53 fetuses with age

between 20 and 37 GWs with CHD. In (Brossard-Racine et al., 2014), MRI scans of 144 fetuses with CHD and 194 controls of age ranging between 18 and 39 GWs were also studied for brain anomalies. Their findings showed a significantly higher frequency (23%) of structural brain abnormalities in the CHD group compared with less than 2% of recurrence in fetuses from the control group. This highlights the close relationship between heart and brain development. Based on segmentations of 3D high-resolution reconstructed volumes of 30 control fetuses and 18 fetuses diagnosed with hypoplastic left heart syndrome, Clouchoux et al. (2013) demonstrated a progressive third-trimester decline in volumetric growth of cortical GM, subcortical GM and WM, in addition to significant region-specific cortical development delays in the hypoplastic left heart syndrome group. These findings are consistent with postnatal data demonstrating that delayed fetal brain maturation and development *in utero* appears to begin in the 3rd trimester (McQuillen et al., 2010).

2.6.3 Ventriculomegaly

VM is one of the fetal brain anomalies that is frequently diagnosed during the gestational period. When no other anomalies are present, it is called isolated VM. VM is defined as a ventricular atrial diameter greater than 10 mm at any GA (Cardoza et al., 1988), while the width in normal subjects lies between 6 and 9 mm (Scott et al., 2013). Atrial diameter measurements larger than 15 mm constitute severe VM, whilst measurements between 10-15 mm are classified into mild VM (10-12 mm) and moderate VM (12.1-15 mm) (Kyriakopoulou et al., 2014). In case of isolated VM, this latter dichotomy is typically used as a prognostic biomarker of the neurodevelopmental outcome of the fetuses. That is, cases with isolated mild VM are generally associated with good outcomes, though some will have abnormal outcomes. A robust method to clearly distinguish between both cohorts is therefore critical in counseling pregnancies (Scott et al., 2013).

Whether in antenatal US or fetal MRI, routine assessment of VM relies on a simple 2D measurement of the atrial diameter (see Figure 2.6) on a particular plane at the level of the atrium (Cardoza et al., 1988). Nonetheless, reproducibility of these measurements is known to be variable, especially in US. The possibility of 3D fetal MRI simplifies volumetric measurements, and promotes computation of other features such as shape measurements. To the best of our knowledge, VM has been studied in 3 works in the literature using motion-corrected 3D reconstructions of the fetal brain (Gholipour

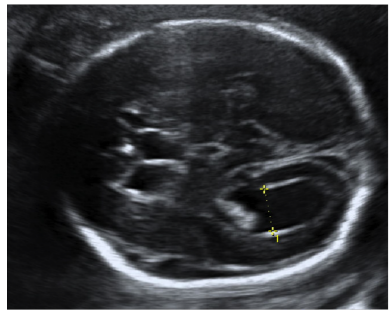

Figure 2.6: Measurement of the atrial diameter in US.

et al., 2012; Kyriakopoulou et al., 2014; Scott et al., 2013). The volume of the ventricles has been shown to be more distinctive than the atrial diameter in diagnosing VM (Gholipour et al., 2012). In (Scott et al., 2013), volumetric and curvature analyses were performed to compare a group of isolated mild ventriculomegaly (IMVM) to a cohort of healthy fetuses between 22 and 25.5 GWs by identifying potential IMVM-specific deviations in tissue volume, and cortical and ventricular local surface curvature during fetal brain development. Except enlarged ventricular volume in IMVM, no significant difference was found in brain tissue or cortical volume between groups. However, evidence of cortical GM enlargement in IMVM fetuses was found in (Kyriakopoulou et al., 2014) after analyzing the differences between 60 normal fetuses and 65 with IMVM across a wider age range of 22-38 GWs.

2.7 Discussion

The emergence of fast-sequence MRI combined with advanced techniques for motion correction (e.g., Jiang et al., 2007; Murgasova et al., 2012) has enabled the formation of 3D volumes of the *in vivo* fetal brain. This has supposed an immense step towards the understanding of the early cerebral maturational processes compared to conventional prenatal US, and the discovery of new 3D biomarkers associated with fetal brain anomalies that are more distinctive and reproducible. However, in contrast to the widespread techniques existing in the literature for adult brains, fetal brain MRI needs to be approached in a different manner due to the complex and rapid changes that occur in the brain. Methods working with MR images of the fetal brain must take into consideration the transient nature of several tissues (e.g., GMAT), the inverted contrast between tissues, and the sub-

stantial shape variation of the brain as growth proceeds. These challenges have stimulated the development of new methodological approaches that permit the study of the developing brain.

To better capture the dynamics of the laminar pattern and the changes in cortical folding, literature in fetal brain MRI encourages the creation of spatio-temporal atlases (e.g., Gholipour et al., 2014; Serag et al., 2012b), which shed light onto the fetal brain growth patterns by encoding temporal and inter-subject variability, and provide a common reference space for the study of the developmental process in the fetus. In addition, image processing techniques such as registration and segmentation may take advantage of these atlases to achieve better accuracy. Concerning segmentation methods of the fetal brain in MRI, the frequent PVE problem present in the boundary between CSF and GM, and the existing tissues at a particular GA have to be considered. Hence, atlas-based segmentation is highly useful in these scenarios to aid in segmentation, and spatio-temporal atlases have proved to be of great benefit.

State of the art segmentation methods allow performing accurate volumetric analyses, providing a broader view than the standard 2D measurements used in clinical settings. Furthermore, studies concerning sulcation and gyrification of the cerebral cortex in the fetus are now easily viable through MRI, allowing for both qualitative and quantitative inspection. Segmentation of CP and SP is necessary for the study of cortical folding. Mapping and ordering the normal patterns of cortical folding (e.g., Clouchoux et al., 2012; Habas et al., 2012) can be used as a baseline to help detect regions of abnormal or delayed folding correlated with possible neurological disorders. Recently, spatio-temporal cortical surface atlases of the developing brain were created in (Li et al., 2015; Wright et al., 2015), providing patterns of cortical developmental trajectories at every point in the cortical surface and, therefore, establishing an accurate normative timing for gyrification.

In clinical settings, fetal brain MRI has become an important tool in confirming and complementing prior findings in US. Furthermore, now it is possible to develop automatic methods to facilitate the diagnosis of brain abnormalities *in utero*, and provide scalability to study large populations.

There are still many open directions to explore for researchers in fetal brain MRI. Reconstruction algorithms from motion-corrupted stacks rely on the segmentation of the brain from 2D slices. However, there exist no well-established segmentation approaches. In 3D reconstructed images of the fetal brain, the posterior medial part near the ventricles, for example, is

a complicated region for the segmentation of the cortex. The availability of public databases with ground truth annotations and segmentation challenges, such as the NeoBrains12 (Išgum et al., 2015), for the fetal brain may boost research advances in this area. Regarding spatio-temporal modeling, existing atlases were created from healthy subjects and, therefore, only capture the morphological changes of the *in utero* brain in its normal course. A promising direction of future work could be the construction of disease-specific spatio-temporal atlases that show the dynamic disease-related changes in the brain and help understand disease progression. Another interesting direction of future work is to analyze the impact of congenital diseases in neurodevelopmental outcome. Longitudinal studies will allow neuroscientists to assess the effect of *in utero* brain anomalies in cognitive development and link the findings in structural MRI with brain connectivity and measurements from other modalities.

This work was restricted to T1w and T2w MRI. However, other imaging modalities, such as fMRI and DTI, can also be used to study the *in vivo* fetal brain. In (Huang et al., 2009), DTI allowed both macro- and microscopic characterization of brain development, while fMRI was also used to capture the emerging connectivity patterns in the fetal brain (Jakab et al., 2014). Computational growth models constitute another type of approach to understand the physical forces behind the formation of the cortical convolutions in the developing brain (Nie et al., 2010, 2011; Tallinen et al., 2016). Combining the heterogeneous findings from all these modalities and approaches could synergistically improve the understanding of *in utero* brain development.

2.8 Conclusions

In this paper, we presented a thorough review of methodological advances to study early brain development from *in utero* structural MRI. The review outlined the challenging context for neuroscience research to achieve an improved understanding of *in vivo* fetal brain maturational mechanisms, motivated the need for the implementation of novel processing approaches, and reported the potential gains resulting from quantitative fetal MRI studies that can be realized in clinical practice.

3

Discriminative Confidence Estimation for Probabilistic Multi-atlas Label Fusion

Abstract – Quantitative neuroimaging analyses often rely on the accurate segmentation of anatomical brain structures. In contrast to manual segmentation, automatic methods offer reproducible outputs and provide scalability to study large databases. Among existing approaches, multi-atlas segmentation has recently shown to yield state-of-the-art performance in automatic segmentation of brain images. It consists in propagating the labelmaps from a set of atlases to the anatomy of a target image using image registration, and then fusing these multiple warped labelmaps into a consensus segmentation on the target image. Accurately estimating the contribution of each atlas labelmap to the final segmentation is a critical step for the success of multi-atlas segmentation. Common approaches to label fusion either rely on local patch similarity, probabilistic statistical frameworks or a combination of both. In this work, we propose a probabilistic label fusion framework based on atlas label confidences computed at each voxel of the structure of interest. Maximum likelihood atlas confidences are estimated using a supervised approach, explicitly modeling the relationship between local image appearances and segmentation errors produced by each of the atlases. We evaluate different spatial pooling strategies for modeling local segmentation errors. We also present a novel type of label-dependent appearance features based on atlas labelmaps that are used during confidence estimation to increase the accuracy of our label fusion. Our approach is evaluated on subcortical structure segmentation in adults brains from the MICCAI 2013 SATA Challenge and the ADNI datasets, and on tissue segmentation of a fetal brain dataset. Overall, our results indicate that the proposed label fusion framework achieves superior performance to state-of-the-art approaches in the majority of the evaluated brain structures and tissues, and shows more robustness to registration errors.

This chapter is adapted from:
Benkarim, O. M., Piella, G., González Ballester, M. A., and Sanroma, G. (2017). Discriminative confidence estimation for probabilistic multi-atlas label fusion. *Medical Image Analysis*, 42:274-287. https://doi.org/10.1016/j.media.2017.08.008
The content of this chapter is updated to further include experiments on fetal brain MRI segmentation, which were not present in the original paper.

3.1 Introduction

Brain segmentation from magnetic resonance imaging (MRI) is an important preprocessing step for many neuroimaging studies, e.g., volumetry, cortical thickness, *etc.* For this task, automatic methods are desirable over manual segmentation since the latter is very time-consuming and subject to inter- and intra-rater variability. Although good outcomes can be achieved for the segmentation of the main tissues based only on image intensities (Ashburner and Friston, 2005; Leemput et al., 1999; Shattuck et al., 2001), segmentation of anatomical structures (e.g., defined by their functional properties) renders intensity information insufficient and atlas priors become an imperative resource in order to accurately delineate such structures. In this setting, single-atlas based segmentation uses a single atlas that is registered to the to-be-segmented image and then propagates its labelmap to the target using the resulting warp from the registration step. Single-atlas based segmentation, however, suffers from (1) representative bias in that a single atlas may not capture the neuroanatomical variability of the general population, and (2) high sensitivity to registration errors since only one atlas is used. To address these drawbacks, multi-atlas segmentation (MAS) makes use of multiple atlases to segment a given target image (Aljabar et al., 2009; Heckemann et al., 2006; Lötjönen et al., 2010). In this way, it better adapts to the anatomical variability of the population and highly mitigates the effect of registration failures in the final segmentation.

Indeed, MAS has recently shown to be a promising technique for brain structural segmentation (Iglesias and Sabuncu, 2015; Sanroma et al., 2016b). It consists in fusing the propagated labelmaps from a set of training atlases to a target image. There are two main steps: 1) image registration, where the spatial transformations are computed to warp the atlas labelmaps to the target image, and 2) label fusion, where these candidate segmentations (i.e., warped labelmaps) are fused into a consensus segmentation. The focus of this paper is on improving label fusion, the second step of MAS. Label fusion is a rather challenging problem that consists in finding the optimal combination of the propagated atlas labelmaps at each region of the target image to obtain the best segmentation. The most straightforward way to approach this problem is to use majority voting (MV) (Klein et al., 2005; Rohlfing et al., 2004), which assigns to each target the most frequent label occurring among the training atlases. This method has shown superior performance over single-atlas based label propagation. However, since all the atlases are combined with equal weight, having atlases too dissimilar

to the target will push the resulting segmentation away from the true target anatomy. In order to solve this problem, several works have proposed more robust label fusion strategies that weigh each atlas vote contribution based on its similarity to the target image (e.g., Artaechevarria et al., 2009; Coupé et al., 2011; Sabuncu et al., 2010). STAPLE (Warfield et al., 2004) and similar methods use a statistical approach to label fusion. Although STAPLE was initially conceived to globally assess the performance of different raters, many works build on STAPLE to provide spatially varying statistical label fusion approaches for MAS. Non-Local STAPLE (Asman and Landman, 2013) and STEPS (Cardoso et al., 2013) extend STAPLE by including appearance information from the images and integrating the non-local means approach (Buades et al., 2005) into the statistical framework of STAPLE. Other works tackle label fusion in MAS from the machine learning perspective, using classification-based approaches (Powell et al., 2008; Sdika, 2015; Zikic et al., 2013), reconstruction-based approaches (Benkarim et al., 2014; Zhang et al., 2012), or a combination of both (Sanroma et al., 2016a, 2015b).

In this work, we propose a probabilistic framework with the following contributions:

- We estimate spatially varying confidences for each training atlas in an offline way to reduce computation burden at test time.
- We formulate our method in a probabilistic framework and obtain maximum likelihood confidence parameters through discriminative learning.
- We explore different spatial pooling strategies for modeling local segmentation errors.
- We propose novel label-dependent features to be used together with appearance features to estimate the confidences in the proposed framework.

This paper is an extension of a recently published conference paper (Benkarim et al., 2016). In this current work, we implement more sophisticated spatial pooling strategies to make our method more accurate and computationally efficient, present a more extensive description of the proposed label fusion framework, evaluate the performance of our approach in a fetal brain MRI dataset and 2 adult brain MRI datasets, assess the robustness of our

method to registration failures by using several registration settings (affine and 2 different non-rigid registrations), provide an in-depth review of the literature and a comparison of our approach with the state-of-the-art, and include a thorough discussion of the results.

The outline of the paper is as follows. Section 3.2 is devoted to state-of-the-art label fusion approaches in MAS. Section 3.3 presents the details of our proposed method. In Section 3.4 we describe the experimental setting and present the results. In Section 3.5 we discuss the advantages and limitations of our approach. Section 3.6 concludes the paper.

3.2 Related work

The selection of the label fusion strategy is a crucial step in MAS and has been extensively studied in the literature. Label fusion approaches can be grouped in 3 categories according to the strategy used for fusing the different atlas labelmaps to produce the final segmentation: similarity-based, statistical-based and learning-based approaches (González-Villà et al., 2016; Iglesias and Sabuncu, 2015).

3.2.1 Similarity-based approaches

One major trend is to assign weights (or confidences) to each warped atlas labelmap based on the similarity of its intensity image with the target image. Label fusion with MV can be viewed as a trivial case of these approaches with atlas labelmaps combined with uniform weights. The main assumption of similarity-based approaches is that regions with similar intensities have similar labeling. Several works used this heuristic to perform label fusion. Global weighted voting assigns to each registered atlas a global weight based on its overall similarity with the target image (Artaechevarria et al., 2009). This approach, however, does not consider the spatially varying accuracy of registration, and subsequently, of the confidences. Among approaches that tackle this problem, we can find works using local (Artaechevarria et al., 2009; Isgum et al., 2009; Sabuncu et al., 2010) and non-local (Coupé et al., 2011; Rousseau et al., 2011) weighted voting. Local weighted voting uses one-to-one correspondences of the atlases and the target image, which compensates for the potential misalignments by increasing the weights of the locally well-aligned atlases (and reducing the weights of the rest). In

non-local weighted voting, the one-to-one correspondence constraint is relaxed by adopting the non-local means approach proposed in (Buades et al., 2005), offering even more flexibility to compensate for registration errors. Moreover, coarsely warped atlases are enough to achieve satisfactory results, thus leading to moderate computational requirements during image registration. In this last approach, the confidence of each atlas is measured using the most similar or all patches from a small search neighborhood around a given voxel. This non-local patch-based strategy has been widely adopted by subsequent methods. The method proposed in (Wang et al., 2013), for example, searches for the most similar patch from each atlas and models pairwise dependencies between atlases to reduce the weights of correlated atlases during label fusion.

In similarity-based approaches, performance is sensitive to the choice of the similarity measure, and more importantly, image similarity does not always correlate well with atlas confidence (Sanroma et al., 2014).

3.2.2 Statistical approaches

Another kind of weighting schemes alleviates the bias induced by similarity-based label fusion by estimating atlas confidences through a more direct measure of the anatomical overlap (Warfield et al., 2004). These approaches alternate the segmentation of the target anatomy and confidence estimation for each of the competing candidate labelmaps by comparing to a consensus segmentation in an iterative fashion. STAPLE is the most representative work and defines a principled statistical framework based on the Expectation-Maximization (EM) algorithm to perform such estimation. STAPLE, however, was initially conceived to assess the performance of different raters and its performance in MAS is not significantly better than MV (Artaechevarria et al., 2009; Asman and Landman, 2013). Furthermore, STAPLE does not take into consideration the intensity information available from the images during the confidence estimation process. Many extensions build on STAPLE to provide statistical label fusion approaches for MAS. Gorthi et al. (2014) proposed an approach that incorporates the versatility of local similarity-based approaches into the estimation of the confidences. STEPS (Cardoso et al., 2013) proposes a local ranking strategy based on image similarity to improve the confidence estimation in STAPLE on a voxel-by-voxel basis. The Non-Local STAPLE (Asman and Landman, 2013) integrates the non-local means approach (Buades et al., 2005) and includes appearance information into the statistical framework of STAPLE.

Nonetheless, as pointed out earlier, using image similarity can induce a bias in the estimation of the confidence.

3.2.3 Learning-based approaches

Learning-based methods constitute a different approach to MAS. They attempt to learn, from a set of examples extracted from the training atlases, a function that maps local image appearances to the correct label. A global classifier per atlas using Random Forest (Breiman, 2001) was proposed in (Morra et al., 2010; Zikic et al., 2013). Compared to patch-based approaches, the use of global classifiers further relaxes the one-to-one correspondence constraint. However, global classifiers are usually limited in capturing the complex appearance patterns associated with structural segmentation. This can be circumvented to some extent by using region- or structure-wise classifiers (Powell et al., 2008; Wang et al., 2011a), and/or the feature vectors can be augmented to include spatial information for the classifier. Voxel-wise classifiers were also used in the literature for MAS. The work presented in (Hao et al., 2014) proposed a MAS approach to estimate the target image's label that learns voxel-wise support vector machine (SVM) classifiers based on the voxel's k nearest positive and negative training samples. Sdika (2015) also used voxel-wise SVM classifiers in a single-atlas based segmentation framework. This approach can be extended to the MAS framework by learning such classifiers in each of the atlas spaces.

The advantage of supervised approaches to MAS is that they can incorporate additional features (e.g., texture, shape, spatial location, *etc.*) (Bai et al., 2015; Hao et al., 2014), which may benefit the classifiers. Furthermore, learning can be performed offline (Sdika, 2015; Zikic et al., 2013), reducing the computational burden of training the classifiers for each target image.

3.3 Methodology

In this section, we provide a description of the proposed probabilistic label fusion framework. Figure 3.1 shows the pipeline of the method, which is composed of two phases:

- Training phase: for each atlas, we compute its confidence model by maximum likelihood estimation. For this task, registration of each

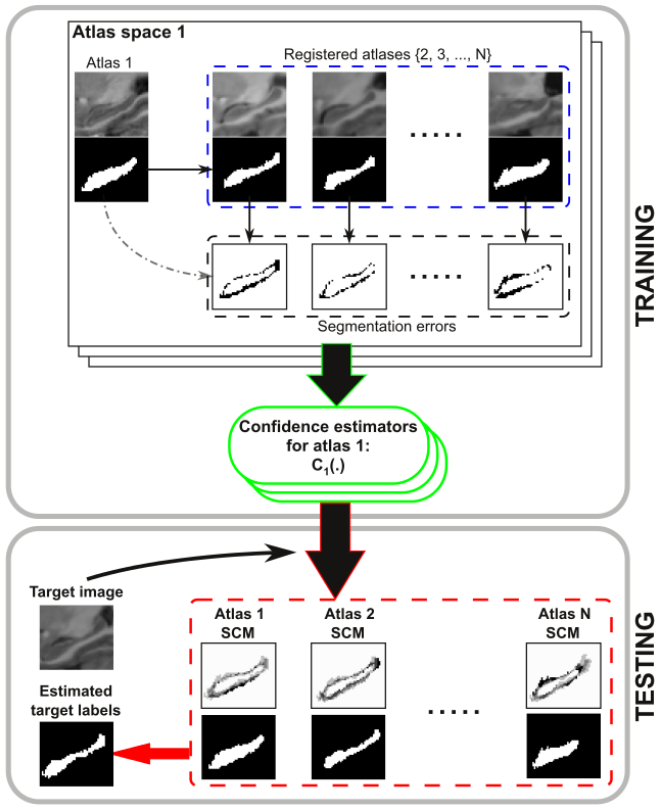

Figure 3.1: Pipeline of the proposed label fusion approach. **Training**: for each atlas, the remaining atlases are registered onto the atlas space and confidence models are computed. **Testing**: given a novel to-be-segmented image, SCMs from each atlas are obtained using the confidence models. Target labels are then estimated according to the proposed label fusion framework.

atlas to the spaces of the remaining training atlases is first carried out. Then, confidence models in each atlas space are estimated in an offline manner. We propose two ways of estimating the confidence models: 1) a naive approach depending only on local label statistics, and 2) a learning-based approach modeling the relationship between local image appearances and segmentation errors. Confidence estimation in the space of each atlas is important in order to cope with systematic segmentation errors caused by registration failures.

- Testing phase: for a given target image, *spatial confidence maps* (SCMs) are obtained after supplying the target image to the confidence models computed in the training phase. The target's final segmentation is

then estimated in a voxel-by-voxel basis with the proposed framework using the SCMs in conjunction with the atlas labelmaps.

3.3.1 Probabilistic label fusion

In the MAS setting, we have a set of atlas images $\mathbf{A}$ along with their labelmaps $\mathbf{D}$, where $D_{ij} \in \mathbf{D}$ and $D_{ij} = \{1, \ldots, p\}$, indicates which one of the p structures is present at voxel i of the j-th atlas. Now consider a novel target image T, where T_i denotes the intensity value at voxel i, we denote the to-be-estimated target labelmap as F.

Our proposed label fusion follows the derivation of a spatially varying version of STAPLE proposed in (Asman and Landman, 2012). The goal is to find the target labels that maximize the following posterior probability:

$$f\left(F|\mathbf{D},\mathbf{C}\right) = \prod_i f\left(F_i|\mathbf{D}_i,\mathbf{C}_i\right) = \prod_i \frac{f\left(\mathbf{D}_i|F_i,\mathbf{C}_i\right) f\left(F_i\right)}{f\left(\mathbf{D}_i,|\mathbf{C}_i\right)}. \tag{3.1}$$

where $\mathbf{D}_i$ denotes the set of atlas decisions for voxel i and $\mathbf{C}_i$ denotes their respective confidences (or weights). Note that we assume conditional independence in the target voxels. Further assuming independence among the atlas decisions, we obtain the following expression:

$$f\left(F_i|\mathbf{D}_i,\mathbf{C}_i\right) = \frac{\prod_j f\left(D_{ij}|F_i,C_{ij}\right) f\left(F_i\right)}{\sum_{s\in\{1,0\}} \prod_j f\left(D_{ij},|F_i=s,C_{ij}\right) f\left(F_i=s\right)}. \tag{3.2}$$

The binary segmentation case is considered in Eq. (3.2), i.e., we have only two labels denoted $\{0,1\}$. For multiple structures, a one-versus-rest approach can be used.

Accordingly, the probability of the target label F_i being foreground (i.e., label 1) is defined as:

$$f\left(F_i=1|\mathbf{D}_i,\mathbf{C}_i\right) = \frac{a_i}{a_i+b_i}, \tag{3.3}$$

where

$$a_i = f\left(F_i=1\right) \prod_j f\left(D_{ij}|F_i=1,C_{ij}\right), \tag{3.4}$$

$$b_i = f\left(F_i = 0\right) \prod_j f\left(D_{ij} | F_i = 0, C_{ij}\right). \tag{3.5}$$

Here, we are interested in $f\left(D_{ij} | F_i = s, C_{ij}\right)$, which is the probability of observing the decision of j-th atlas on voxel i, given that the target label is s and the atlas confidence at that voxel is C_{ij}. This term expresses the likelihood that the atlas and target labels coincide, and is defined as:

$$f\left(D_{ij} | F_i = s, C_{ij}\right) = \begin{cases} C_{ij} & \text{if } \; D_{ij} = s \\ 1 - C_{ij} & \text{otherwise.} \end{cases} \tag{3.6}$$

In the EM framework used by STAPLE-based approaches, Eq. (3.3) corresponds to the estimation of the hidden reference segmentation (i.e., E-step) given the rater performance parameters or confidences, C_{ij}. These confidences are then updated during the M-step based on the previous E-step, and this process is repeated interleaving both steps until reaching convergence. The main difference of our approach with (Asman and Landman, 2012) lies in the computation of the C_{ij} confidences in Eq. (3.6), which is the central part of our work. We propose to estimate spatially varying confidences (i.e., for each voxel) in an offline manner using the atlases in the training set instead of the iterative EM-based approach used in (Asman and Landman, 2012).

3.3.2 Confidence estimation

Let us focus on the computation of the confidence for a single voxel i of a single atlas j, denoted as $c \equiv C_{ij}$ for brevity (the same procedure is repeated for the rest of the voxels on the rest of atlases). Similarly, let us denote as $d \equiv D_{ij}$ the label at voxel i in the j-th atlas. We denote as $\mathbf{f} = \left\{\tilde{D}_{ik}, k \neq j\right\}$, the training set of target observations for the voxel i in the j-th atlas composed of the registered labelmaps of the rest of atlases. This is indicated by the blue panel in Figure 3.1. We compute the confidence at each voxel by maximizing the following joint likelihood:

$$\begin{aligned} \hat{c} &= \arg\max_c f\left(\mathbf{f}, d | c\right) \\ &= \arg\max_c \prod_k f\left(d | f_k, c\right) f\left(f_k | c\right), \end{aligned} \tag{3.7}$$

where $f_k \in \mathbf{f}$. We discard the second term in the product since we assume that target labels are only affected by the confidence parameters in the presence of an atlas. Taking the logarithm and substituting the atlas likelihood term by its expression in Eq. (3.6) yields:

$$\begin{aligned}\hat{c} &= \arg\max_{c} \sum_{k} \log f\left(d|f_k, c\right) \\ &= \arg\max_{c} \sum_{f_k=d} \log c + \sum_{f_k \neq d} \log\left(1-c\right). \end{aligned} \tag{3.8}$$

Taking derivatives, the optimal confidence is

$$c = \frac{n_h}{n_h + n_m}, \tag{3.9}$$

where n_h and n_m are the number of coincident target labels (*hits*) and different target labels (*misses*), respectively, from the atlas label. This defines our naive approach. When all atlases are used to compute the confidences, this approach yields similar results to MV. Note that Eq. (3.7) is the analogue of the M-step in STAPLE-based approaches. However, we are using solely the training atlases and no estimation of the true hidden segmentation is considered, as opposed to (Asman and Landman, 2012).

Nevertheless, we further believe that local image appearances provide valuable clues for estimating this confidence. Therefore, we extend the previous naive method by substituting the constant confidence in Eq. (3.6) by a more complex function informed by the image appearances, as follows:

$$f\left(D_{ij}|F_i = s, \mathcal{C}_{ij}\right) = \begin{cases} \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{if } \; D_{ij} = s \\ 1 - \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{otherwise,} \end{cases} \tag{3.10}$$

where $\mathbf{t}_i$ and $\mathbf{a}_{ij}$ are image appearance features extracted around voxel i from the target atlas image and the j-th atlas respectively. $\mathcal{C}_{ij}\left(\cdot\right)$ is a function denoting the confidence we have that the atlas label is correct given the target and atlas image appearances (as shown in the green panel in Figure 3.1). By using image appearances, we can effectively capture the effects of registration errors on modeling such confidence. Again, our goal is to compute such function as to maximize the joint probability of each atlas observation given the training set. Using a similar development as in the naive case, we arrive at the following expression:

$$\hat{\mathcal{C}} = \arg\max_{\mathcal{C}} \sum_{f_k=d} \mathcal{C}\left(\mathbf{t}_k, \mathbf{a}\right) - \sum_{f_k \neq d} \mathcal{C}\left(\mathbf{t}_k, \mathbf{a}\right), \tag{3.11}$$

where $\mathbf{t}_k$ and $\mathbf{a}$ denote the local image appearances of the k-th target training sample and atlas in the training set, respectively.

In the testing stage, given a new target image T, it is first warped to each of the atlases in the training set. Then, SCMs are computed using the confidence functions of Eq. (3.11) based on intensity information from both the target, T, and the atlases, $\mathbf{A}$. Next, SCMs and their corresponding atlas labelmaps, $\mathbf{D}$, are transformed back to the target space. Finally, we compute the label fusion using Eq. (3.1), as shown in the red panel of Figure 3.1.

3.3.2.1 Training

Expression (3.11) corresponds to the minimization of an empirical error subject to the constraint that the computed function must be a probability density function. For this purpose, we consider a learning-based approach to build voxel-wise classifiers as our confidence estimators. Note that we segment each structure separately, thus using binary classifiers. In order to explain the procedure to create the samples used to train each voxel-wise classifier, let us assume the simple case of one-to-one correspondences. For each training atlas (in its native space), classifiers are built for each voxel. Consider an atlas $A \in \mathbf{A}$ in its native space and a target atlas $W \in \mathbf{A} \backslash \{A\}$ warped to A. For the i-th voxel, let a_i and w_i respectively represent the patches of atlas A and the warped target atlas W, with corresponding labels d_i^a and d_i^w. That is, the pair:

1. (a_i, d_i^a) represents the patch and label of the i-th voxel in atlas A.
2. (w_i, d_i^w) represents the patch and label of the i-th voxel in the target atlas W.

These 2 pairs are used to create a single training sample (x_i, y_i) corresponding to atlas W for the i-th classifier of atlas A as follows:

- For the features, we use a patch-based approach. The feature vector, x_i, consists on the intensity difference between the atlas patch and the patch from the target:

$$x_i = a_i - w_i.$$

- The class label (i.e., the label used to train our classifiers), y_i, is built from the atlas labels (i.e., the voxel labels, d_i^a and d_i^w) and corresponds to the segmentation error produced by the atlas (i.e., A) when segmenting the target (i.e., W), and is defined as:

$$y_i = \delta(d_i^a, d_i^w),$$

  where $\delta(\cdot,\cdot)$ is the Kronecker delta function. If d_i^a and d_i^w are equal, then y_i is 1, and 0 otherwise. In other words, the class label (or the ground truth during training) tells us if atlas A correctly segments atlas W ($y_i = 1$) or not ($y_i = 0$) at the i-th voxel.

Given N training atlases in our database, in this simple case of one-to-one correspondences, the number of training samples used to train the i-th classifier of atlas A is $N-1$, where each sample is built from atlas A and each of the remaining $N-1$ warped atlases W. Therefore, the i-th classifier of atlas A attempts to learn from all the x_i what are the patterns of intensity differences that lead atlas A to produce correct or erroneous labels, based on the rest of training atlases.

In the test stage, when a novel to-be-segmented image arrives, it is transformed to the spaces of all training atlases. Given the test image T warped to the space of atlas A and the patch t_i of the test image at the i-th voxel. The patch difference a_i - t_i is fed to the classifier, which will predict how likely is the label of atlas A at voxel i (i.e., d_i^a) to be the correct label for t_i. The higher the predicted probability by the classifier, the more likely is the test patch to have a similar label to atlas A. This is what we interpret as confidence in our label fusion. Once this has been done for all N training atlases, we will have N label candidates for t_i along with their predicted confidences.

This is how confidences are estimated with Eq. (3.11). Note that in the naive case, according to Eq. (3.9), the optimal confidence is equal to the proportion of correct labels in the training set. We do not use the feature vectors x_i, just generate the class labels, y_i, to compute this confidence, which tells us how good is the label d_i^a of atlas A in segmenting the rest of training atlases W.

Instead of using simple one-to-one correspondences, we adopt two different spatial pooling strategies to build the training set for each voxel-wise classifier: 1) non-local means approach in the target space and 2) non-local means approach in both atlas and target spaces. In the following we describe both approaches in detail.

3.3.2.2 Non-local means approach in target space

Here we use one-to-many correspondences. Given voxel i in the atlas space A, whose patch a_i is represented as the blue box in Figure 3.2, we used its label, d_i^a, to segment all voxels within a neighborhood window in the warped atlas W, $S_w(i)$. This is illustrated in Figure 3.2 as a red box in the target atlas. The features and labels for A and a given warped atlas W are extracted as follows:

$$x_j = a_i - w_j, \forall j \in S_w(i),$$

$$y_j = \delta(d_i^a, d_j^w), \forall j \in S_w(i).$$

The advantage of this approach is twofold: 1) there are more samples to learn the voxel-wise classifiers than in the one-to-one correspondences case, and 2) confidence estimators are more robust as they are trained to take into account larger registration errors (i.e., all patches in the neighborhood window of W at the i-th voxel). The number of samples created to train the classifier is $(N-1) \times |S_w(i)|$, where $|\cdot|$ denotes the size of the neighborhood window (e.g., for a $3 \times 3 \times 3$ neighborhood window, each warped atlas W contributes with 27 samples.)

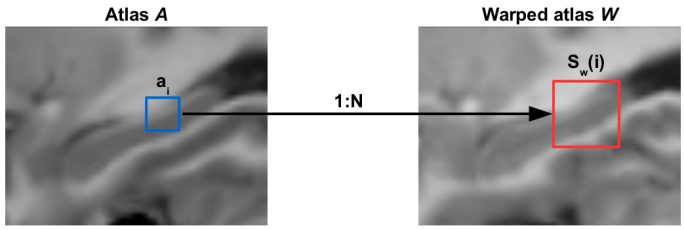

Figure 3.2: Non-local means approach in target space. The blue box represents the patch around the i-th voxel in the atlas space A. The red box in the target space W represents the window search from which we extract all patches.

3.3.2.3 Non-local means approach in target and atlas spaces

This is an extension of the previous point to have many-to-many correspondences. Here, instead of using a single voxel i in the atlas space A to

segment the target W, we take into consideration all voxels in its neighborhood, $S_a(i)$ (depicted in Figure 3.3 as a green box), and then use k-nearest neighbors to select the voxel label of the most similar atlas patch $a_i^{\hat{k}}$, corresponding to a given target patch, (see arrow labeled *best:N* in Figure 3.3). The samples created from A and a particular target atlas W are defined as:

$$x_j = a_i^{\hat{k}} - w_j, \forall j \in S_w(i),$$

$$y_j = \delta(d_i^{a\hat{k}}, d_j^w), \forall j \in S_w(i),$$

with $\hat{k}$ indexing the atlas patch most similar to w_j:

$$\hat{k} = \arg\max_{k \in S_a(i)} sim(a_k, w_j),$$

where *sim* denotes a similarity measure (e.g., cosine similarity). In this way, segmentation errors produced by atlas A are based on appearance information. In fact, the voxel-wise classifiers in this case are built upon a similarity-based approach, therefore, learning not only the appearance patterns that lead the current atlas A to mislabel the remaining target atlases, but also the behavior of the similarity measure in segmenting the target patches. Furthermore, using many-to-many correspondences without k-nearest neighbors to construct the datasets for our confidences estimators will make training computationally expensive given that we learn voxel-wise classifiers. The number of samples used to train the classifiers is the same as in the one-to-many correspondences, being the only difference that the atlas patch $a_i^{\hat{k}}$ in this case is not fixed but the most similar among all patches in $S_a(i)$ to the target patch, w_j. At test time and for the i-th voxel, given a test patch t_i, the most similar patch and its label from atlas A to t_i are

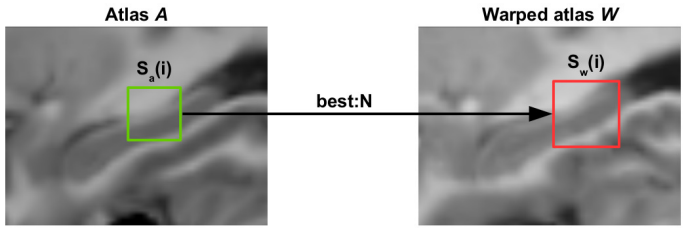

Figure 3.3: Non-local means in both spaces. The green box represents the window search from which the best (i.e., most similar) patch is selected for each patch in the target atlas W (red box in target atlas).

selected. The classifier then predicts the confidence of this label (i.e., the label corresponding to the most similar patch in A) in correctly segmenting t_i based on the patch difference.

3.3.2.4 Label-dependent feature extraction

In patch-based approaches, the simplest way to represent local features is to use a cubic patch around the voxel of interest, as stated in subsection 3.3.2.1. Here, to fully take advantage of our learning-based confidence estimators, we propose to use additional features based on the atlas labelmap. This contribution uses the label patch of atlas A to extract label-dependent features from the warped images W. As illustrated in Figure 3.4, given the label patch of the atlas A around the i-th voxel, we identify the target voxels corresponding to foreground and background regions (in the case of binary segmentation) and compute different summary statistics. Finally, the difference between foreground and background features is calculated and the resulting features are appended to the intensity patch.

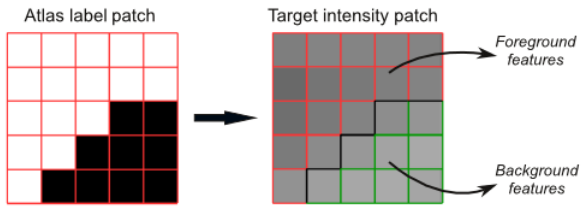

Figure 3.4: Label-dependent feature extraction.

With our label-dependent features, we attempt to characterize the intensity distributions of the target patches according to a given atlas label patch. In principle, it is expected that background voxels would exhibit a different intensity distribution when compared to foreground voxels since they do belong to different structures. Therefore, the more accurate the atlas label patch is in segmenting the target patches, the larger the features difference would be between these regions.

3.4 Experiments

In this section, we present the evaluation of our proposed approach and provide a comparison of its performance with state-of-the-art MAS methods

on the segmentation of adult subcortical brain structures and fetal brain tissues.

3.4.1 Data and preprocessing

The proposed approach was evaluated on 3 brain MRI datasets:

1. MICCAI 2013 SATA Challenge dataset[1]: This dataset is composed of 35 T1-weighted MR images of control subjects with age ranging from 19 to 90 years (32.4 years old in average). The size of the images is $256 \times 256 \times 287$ with a spatial resolution of 1 mm isotropic. Ground-truth segmentations are available for 7 subcortical structures: accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus proper.

2. ADNI dataset[2]: We used a subset of 135 T1-weighted MR images (44 normal controls, 46 subjects with mild cognitive impairment and 45 with Alzheimer's disease). The age distribution is: 40 between 60-70 years, 55 between 70-80 years and 40 with more than 80 years. The size of the images is $197 \times 233 \times 189$ with a voxel size of $1 \times 1 \times 1$ mm. In this dataset, ground-truth segmentations are only available for the hippocampi.

3. Fetal brain dataset: We used 32 subjects from a cohort within a research project on congenital isolated ventriculomegaly. Ages of the included subjects range between 26 to 29.3 gestational weeks. T2-weighted MR imaging was performed on a 1.5-T scanner (SIEMENS 105 MAGNETOM Aera syngo MR D13; Munich, Germany) with a 8-channel body coil. All images were acquired without sedation and following the American college of radiology guidelines for pregnancy and lactation. Half Fourier acquisition single shot turbo spin echo (HASTE) sequences were used with the following parameters: echo time of 82 ms, repetition time of 1500 ms, number of averaging = 1, 2.5 mm of slice thickness, 280×280 mm field of view and voxel size of $0.5 \times 0.5 \times 2.5$ mm. For each subject, multiple orthogonal acquisitions were performed: 4 axial, 2 coronal and 2 sagittal stacks. Brain location and extraction from 2D slices was carried out in an

[1] https://masi.vuse.vanderbilt.edu/workshop2013
[2] adni.loni.usc.edu

automatic manner using the approach by Keraudren et al. (2014), followed by high-resolution 3D volume reconstruction using the method by Murgasova et al. (2012). Ground-truth segmentations were obtained for the following tissues and structures: cerebro-spinal fluid (CSF), cortical gray matter (CoGM), white matter (WM), lateral ventricles (LV), cerebellum (CB) and brainstem (BS). To obtain the ground-truth structures, first, 4 subjects were manually segmented by two expert raters. Then, the remaining subjects were segmented using the automatic method by Sanroma et al. (2016a) and the automatic segmentations were manually corrected by the same expert raters.

Our method requires pairwise registrations since it needs to have each atlas in the rest of the training spaces. However, to save computational time, all images were registered to a common reference space (i.e., MNI152 template for both SATA and ADNI datasets, and for the fetal brains, a custom template built using the images in our dataset). Pairwise mappings were then obtained by composing the transformation of the source atlas to the template space and the inverse transformation from the template to the target atlas. Furthermore, for image intensity to be consistent across atlases, all images were normalized using histogram matching (Nyul et al., 2000).

3.4.2 Experimental Setup

We evaluated our method using the following configurations:

- Naive: the naive approach, where segmentation is based only on local label statistics (i.e., voxel-wise label errors as priors).
- SCMNF: the SCM approach using one-to-many correspondences with only patch intensities as features.
- SCMWF: similar to SCMNF but including label-dependent features.
- SCMNF2: the SCM approach based on many-to-many correspondences with only patch intensities as features.
- SCMWF2: similar to SCMNF2 but including label-dependent features.

For comparison, we considered the following state-of-the-art methods: MV, local weighted voting with inverse similarity metric (LWV) (Artaechevarria et al., 2009), STAPLE, STEPS and joint label fusion (JOINT) (Wang et al., 2013).

The summary statistics we used as label-dependent features for SCMWF and SCMWF2 in the experiments were: mean, maximum and minimum intensities, and the center of mass of each region. Regarding the classifiers used for our confidence estimators, we used logistic regression. For SCMNF2 and SCMWF2, the similarity measure used to select the best atlas patch is cosine similarity:

$$cos(x_i, x_j) = \frac{x_i \cdot x_j}{||x_i|| \cdot ||x_j||}.$$

No parameter tuning was performed for the experiments. We used the default values for all methods, except for the radius of the patch and window search that was set to 1 (i.e., a patch and window search size of $3 \times 3 \times 3$). For logistic regression, the penalty parameter C was set to 1.

For SATA and fetal datasets, a 3-fold cross-validation procedure was used in our evaluation strategy, and for ADNI, 35 atlases were selected for training and the remaining 100 for test. The 35 training atlases were selected in order to span the space of all images using spectral clustering based on normalized correlation. For quantitative comparison, we used the Dice similarity coefficient (Dice, 1945), determined as follows:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|},$$

where A and B are the reference and automatic segmentations, respectively, and the modified Hausdorff distance (MHD) (Dubuisson and Jain, 1994), defined as:

$$MHD(S, T) = \max(d(S, T), d(T, S)),$$

where S and T are the sets of voxels in the boundary of A and B respectively, and d is a directed distance measure between the first and the second sets based on Euclidean distance. MHD is reported in mm throughout the whole article. Statistical significance is measured using the Wilcoxon signed rank test and is reported at $p < 0.05$.

Finally, in order to assess the robustness of our approach to registration failures, all experiments were replicated using three different registrations settings:

1. AF: Affine registration,

2. NR1: Affine followed by a non-rigid registration at a coarse scale using the symmetric diffeomorphic mapping (SyN) proposed by Avants et al. (2008). Non-rigid registration was done in a multi-resolution fashion using a regular grid with control point spacings of 8 and 4 mm, and

3. NR2: Affine followed by a finer non-rigid registration (NR2) using SyN. Non-rigid registration was done in a multi-resolution fashion using a regular grid with control point spacings of 8, 4, 2 and 1 mm.

3.4.3 Implementation and computational complexity

Our method was implemented in Python using the logistic regression Python wrapper provided by *Scikit-learn* (Pedregosa et al., 2011) for the *liblinear* library (Fan et al., 2008). For STAPLE and STEPS, we used the implementations distributed in the NiftySeg[3] software package. For JOINT, the implementation shipped with the ANTs[4] package was used.

Experiments were executed on a PC running 64bit Ubuntu Linux 14.04 LTS with a system configuration Intel(R) Core(TM) i7-4790 CPU (3.60GHz) x 8 with 32 GB of RAM.

Execution times required by our offline learning vary depending on the size of the structure (i.e, number of voxel-wise classifiers) and the use of label-dependent features. To reduce the runtimes, learning was not performed for voxels where all the atlases were in consensus (i.e., same label). Training our confidence estimators for the accumbens, for instance, took around 3 and 12 minutes for SCMNF2 and SCMWF2, respectively. For SCMNF and SCMWF, learning took approximately 3 and 10 minutes. At test time, all versions of our method produced segmentations for the accumbens in less than 2 seconds, similarly to the rest of the methods, except JOINT that took around 10 seconds. For one of the largest structures, the hippocampus in the ADNI dataset, segmentation took around 4 seconds for SCMNF, 6 seconds for SCMNF2, and 12-15 seconds for SCMWF and SCMWF2. For MV, the Naive approach and STAPLE, segmentation took around 2 seconds. Segmentation times for STEPS and LWV were less than 5 seconds and for JOINT took around 20 seconds.

[3] http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftySeg
[4] https://stnava.github.io/ANTs

	SATA							ADNI
	Acc	Amy	Cau	Hip	Pal	Put	Tha	Hip
MV	0.777±0.052	0.799±0.038	0.826±0.096	0.831±0.037	0.882±0.027	0.920±0.019	0.908±0.026	0.767±0.049
Naive	0.779±0.052	0.799±0.038	0.828±0.096	0.830±0.037	0.886±0.027	0.920±0.019	0.912±0.026	0.768±0.049
STAPLE	0.767±0.064	0.797±0.041	0.819±0.103	0.828±0.036	0.877±0.027	0.915±0.018	0.904±0.027	0.768±0.058
STEPS	0.768±0.075	0.797±0.044	0.822±0.105	0.832±0.042	0.882±0.029	0.919±0.018	0.908±0.028	0.799±0.043
LWV	0.784±0.053	0.802±0.037	0.863±0.075	0.843±0.030	0.881±0.027	0.919±0.018	0.914±0.022	0.796±0.045
JOINT	0.799±0.039	0.827±0.024	0.888±0.068	0.871±0.021	**0.888±0.027**	**0.926±0.018***	0.923±0.014	0.860±0.037
SCMNF	0.792±0.049	0.812±0.030	0.902±0.051	0.867±0.016	0.885±0.024	0.923±0.026	0.925±0.011	0.844±0.035
SCMWF	0.805±0.047	0.818±0.033	0.905±0.049	0.871±0.016	0.886±0.026	0.923±0.024	0.924±0.010	0.850±0.039
SCMNF2	0.808±0.047	0.825±0.032	0.906±0.042	0.872±0.018	0.885±0.028	0.922±0.021	0.923±0.013	0.853±0.038
SCMWF2	**0.811±0.044***	**0.830±0.028**	**0.907±0.040**†	**0.877±0.015***	0.886±0.027	0.924±0.019	**0.925±0.011**	**0.866±0.026**
MV	0.255±0.078	0.234±0.061	0.235±0.229	0.224±0.107	0.125±0.032	0.085±0.026	0.109±0.046	0.329±0.139
Naive	0.310±0.080	0.256±0.062	0.249±0.225	0.238±0.099	0.134±0.031	0.091±0.028	0.110±0.043	0.361±0.139
STAPLE	0.267±0.094	0.236±0.062	0.248±0.250	0.228±0.110	0.129±0.032	0.090±0.025	0.115±0.050	0.279±0.098
STEPS	0.255±0.099	0.230±0.064	0.245±0.256	0.223±0.124	0.123±0.034	0.085±0.025	0.110±0.051	0.243±0.083
LWV	0.226±0.062	0.224±0.056	0.199±0.284	0.183±0.060	0.124±0.032	0.085±0.026	0.099±0.044	0.249±0.139
JOINT	0.208±0.055	0.194±0.036	0.153±0.201	0.154±0.047	0.118±0.033	0.077±0.026	0.087±0.024	0.178±0.094
SCMNF	0.213±0.058	0.198±0.045	0.151±0.232	0.150±0.037	0.119±0.032	0.077±0.026	0.085±0.026	0.184±0.101
SCMWF	0.212±0.057	0.197±0.045	0.149±0.228	0.148±0.036	0.119±0.032	0.077±0.025	**0.084±0.024**	0.180±0.098
SCMNF2	0.206±0.056	0.192±0.044	0.150±0.223	0.149±0.038	0.117±0.031	0.078±0.024	0.086±0.030	0.182±0.093
SCMWF2	**0.204±0.055***	**0.190±0.043**	**0.142±0.202**	**0.145±0.034***	**0.117±0.031**	**0.077±0.023**	0.084±0.027	**0.171±0.083***

Table 3.1: Subcortical structure segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per structure, averaged left and right. Results obtained using the non-rigid registration NR2. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2 or SCMWF2. Abbreviations: accumbens (Acc), amygdala (Amy), caudate (Cau), hippocampus (Hip), pallidum (Pal), putamen (Put) and thalamus proper (Tha).

	Fetal					
	BS	CB	CSF	CoGM	LV	WM
MV	0.940±0.006	0.955±0.004	0.939±0.024	0.874±0.037	0.910±0.020	0.967±0.013
Naive	0.945±0.005	0.958±0.004	0.941±0.023	0.875±0.035	0.912±0.019	0.967±0.013
STAPLE	**0.950±0.005**	0.961±0.005	0.944±0.023	0.887±0.037	0.922±0.031	0.970±0.013
STEPS	0.949±0.005	0.960±0.005	0.942±0.023	0.885±0.037	0.922±0.025	0.970±0.013
LWV	0.950±0.006	0.963±0.005	0.949±0.020	0.894±0.031	0.932±0.022	0.973±0.011
JOINT	0.943±0.005	0.961±0.005	0.948±0.020	0.898±0.022	0.930±0.029	0.974±0.010
SCMNF	0.949±0.006	0.963±0.005	0.952±0.019	0.902±0.027	0.936±0.024	0.976±0.010
SCMWF	0.948±0.006	0.963±0.005	0.952±0.019	0.903±0.026	0.935±0.024	0.976±0.010
SCMNF2	0.946±0.006	0.964±0.005	0.952±0.019	0.903±0.024	**0.940±0.024**	0.975±0.010
SCMWF2	0.946±0.005	**0.964±0.005**	**0.953±0.019**	**0.905±0.025***	0.935±0.027	**0.976±0.010**
MV	0.046±0.005	0.036±0.012	0.065±0.055	0.127±0.092	0.074±0.019	0.033±0.025
Naive	0.042±0.004	0.034±0.012	0.065±0.054	0.127±0.090	0.076±0.018	0.033±0.025
STAPLE	**0.038±0.005**	0.032±0.012	0.060±0.055	0.118±0.093	0.065±0.029	0.031±0.026
STEPS	0.039±0.005	0.032±0.011	0.062±0.055	0.118±0.093	0.065±0.022	0.031±0.026
LWV	0.038±0.005	0.031±0.012	0.049±0.041	0.098±0.066	0.054±0.019	0.026±0.018
JOINT	0.044±0.005	0.032±0.012	0.050±0.042	0.090±0.048	0.055±0.024	0.024±0.016
SCMNF	0.039±0.005	0.030±0.012	0.046±0.039	0.089±0.057	0.051±0.022	0.023±0.017
SCMWF	0.040±0.005	0.030±0.012	**0.046±0.039**	0.089±0.055	0.052±0.022	**0.023±0.017**
SCMNF2	0.041±0.005	0.030±0.012	0.047±0.040	0.088±0.056	**0.048±0.021**	0.024±0.018
SCMWF2	0.041±0.005	**0.030±0.012**	0.046±0.041	**0.087±0.055***	0.052±0.026	0.023±0.018

Table 3.2: Fetal brain tissue segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per tissue. Results obtained using the non-rigid registration NR2. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods.

3.4.4 Results

Tables 3.1 and 3.2 report segmentation performance of all tested approaches for each structure or tissue, in terms of Dice and MHD, in both adult (i.e., SATA and ADNI) and fetal MRI databases, respectively. These results correspond to NR2 registration (i.e., finer non-rigid registration), with which all methods provided the best segmentations. Segmentation results corresponding to AF and NR1 registrations are included in Appendix A.

The performance of our Naive approach is similar in both Dice overlap and MHD to MV. In fact, when all atlas labelmaps are used to compute the constant confidence in Eq. (3.6), it is equivalent to MV, being the additional transformations between atlas spaces the only difference. We should expect better segmentation results when using pairwise registration, because, in the conducted experiments, target images were only registered to the common space. STAPLE did not show superior performance in the segmentation of adult subcortical structures over our Naive approach and MV with this registration setting, although small improvements are observed on the fetal dataset, where STAPLE produced the best segmentations for the brainstem. The other STAPLE-based approach used in this comparison (i.e., STEPS) showed practically no difference in performance with STAPLE in the fetal brains. In the adult datasets, STEPS yielded slightly better results than the aforementioned methods, especially in the hippocampi from the ADNI dataset. Furthermore, STEPS was outperformed by LWV in the segmentation results of all structures except the pallidum and the hippocampus from ADNI.

With the exception of STEPS in the SATA dataset and LWV in the ADNI dataset, we can observe a clear dichotomy in performance across all subcortical structures, as reported in Table 3.1, between approaches ignoring image intensity (i.e., MV, Naive and STAPLE) and the rest of methods using appearance information. We can already see a quantitative increase in Dice overlap and a decrease in MHD with LWV and STEPS. However, JOINT and the four versions of our approach provided the most accurate segmentations, with statistically significant improvement in all structures over MV, Naive, STAPLE and STEPS. In the fetal dataset, although this dichotomy is less notable with this registration setting, a considerable increase in performance is observed when using AF and NR1 registration configurations (see Tables A.3 and A.4).

Comparing the different intensity-based configurations of our approach, in

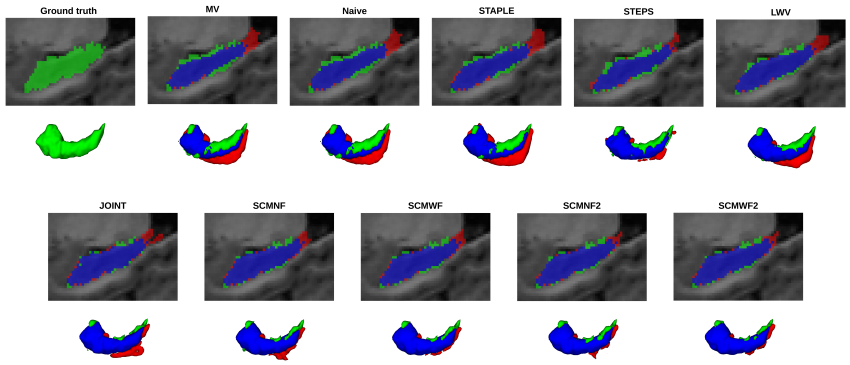

Figure 3.5: Sagittal view and 3D rendering of right hippocampus segmentations for a randomly chosen image from the database. Green and red depict manual and automatic segmentations respectively. Overlap between automatic segmentation and ground truth is shown in blue.

the adult datasets, SCMNF outperformed LWV and STEPS, but the considerable boost in performance was due to the inclusion of the label-dependent features in SCMWF, which reached an overall Dice score and MHD comparable to JOINT (see Table 3.3). Still, when adopting the many-to-many correspondences to learn the confidence estimators, segmentation results with our novel approaches (i.e., SCMNF2 and SCMWF2) were better than their original analogue versions. In fact, SCMNF2 produced similar Dice overlaps and MHD to SCMWF without using label-dependent features. In SCMWF2, the inclusion of these has further improved the segmentation results in all structures according to Dice overlap, as shown in Table 3.1. In terms of MHD, SCMWF2 was outperformed by SCMWF in the thalamus proper, though the difference is minuscule. Dice overlaps achieved by SCMWF2 were statistically higher than the rest of methods in the accumbens and hippocampus, and in the caudate results were statistically significant except for SCMNF2. For MHD, SCMWF2 was statistically better than all methods in the accumbens and the hippocampi from both adult datasets. Figure 3.5 shows an example of caudate segmentations for all the different approaches. Regarding the fetal dataset, we can observe no important performance differences between the four intensity-based versions of our proposed approach.

Table 3.3 summarizes the global performance for all three registrations settings in the segmentation of adult subcortical structures. In terms of overall mean Dice overlap for NR2 registration, our Naive approach (0.850) and MV (0.849) outperformed STAPLE (0.844), while our Naive approach

	SATA			ADNI		
	AF	NR1	NR2	AF	NR1	NR2
MV	0.737±0.083	0.808±0.056	0.849±0.042	0.635±0.068	0.693±0.055	0.767±0.049
Naive	0.737±0.083	0.808±0.056	0.850±0.042	0.632±0.068	0.694±0.055	0.768±0.049
STAPLE	0.738±0.084	0.815±0.062	0.844±0.045	0.670±0.072	0.711±0.065	0.768±0.058
STEPS	0.746±0.086	0.827±0.033	0.847±0.049	0.733±0.058	0.763±0.048	0.799±0.043
LWV	0.767±0.078	0.847±0.047	0.858±0.037	0.710±0.065	0.748±0.051	0.796±0.045
JOINT	0.855±0.046	0.872±0.034	0.875±0.030	0.835±0.043	0.853±0.031	0.860±0.037
SCMNF	0.844±0.048	0.866±0.032	0.872±0.030	0.811±0.039	0.833±0.034	0.844±0.035
SCMWF	0.849±0.046	0.869±0.030	0.876±0.029	0.818±0.040	0.838±0.037	0.850±0.039
SCMNF2	0.857±0.042	0.871±0.030	0.877±0.028	0.832±0.039	0.848±0.037	0.853±0.038
SCMWF2	**0.865±0.037***	**0.874±0.029**	**0.880±0.026**	**0.843±0.038***	**0.856±0.036**	**0.866±0.026**
MV	0.383±0.207	0.238±0.102	0.181±0.083	0.573±0.247	0.439±0.166	0.329±0.139
Naive	0.410±0.207	0.255±0.096	0.198±0.081	0.622±0.257	0.476±0.167	0.361±0.139
STAPLE	0.383±0.213	0.245±0.112	0.188±0.089	0.475±0.177	0.379±0.124	0.279±0.098
STEPS	0.357±0.209	0.235±0.119	0.182±0.093	0.358±0.130	0.300±0.098	0.243±0.083
LWV	0.263±0.167	0.190±0.092	0.163±0.080	0.368±0.209	0.299±0.161	0.249±0.139
JOINT	0.185±0.118	0.145±0.070	0.142±0.060	0.237±0.121	0.188±0.099	0.178±0.094
SCMNF	0.189±0.112	0.148±0.064	0.142±0.065	0.244±0.119	0.205±0.104	0.184±0.101
SCMWF	0.185±0.109	0.146±0.062	0.141±0.064	0.237±0.114	0.200±0.100	0.180±0.098
SCMNF2	0.180±0.103	0.145±0.063	0.140±0.064	0.235±0.113	0.200±0.095	0.182±0.093
SCMWF2	**0.172±0.093***	**0.141±0.056**	**0.137±0.059**	**0.212±0.093***	**0.184±0.082***	**0.171±0.083***

Table 3.3: Subcortical structure segmentation. Overall mean Dice scores (top entries) and MHD (bottom entries) per database for each registration setting. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods.

offered higher MHD than STAPLE (0.198 and 0.188 mm for Naive and STAPLE). Still, with AF and NR1 registrations, STAPLE segmentations were better according to the evaluation measures. In STEPS, the benefit of using intensity information to drive label fusion was manifested in a superior performance over STAPLE for all registration settings, with 0.8%, 1.2% and 0.3% improvements in Dice overlap for AF, NR1 and NR2 registrations, respectively. Furthermore, STEPS also outperformed LWV in the segmentation of the hippocampi from ADNI, although LWV segmentations in the SATA dataset were better than segmentations from STEPS for all registration configurations. In the fetal dataset, as shown in Table 3.4, STEPS outperformed MV and our Naive approach in all registration settings, while STAPLE showed a 0.1% improvement in Dice overlap over STEPS only with NR2. LWV produced considerably better segmentations than STEPS in both Dice and MHD, with 4.9%, 4.4% and 0.6% improvements in Dice overlap for AF, NR1 and NR2 registrations, respectively.

Regardless of the registration setting, we have to emphasize the results of four methods: JOINT, SCMWF, SCMNF2 and SCMWF2. They all produced very robust segmentations, although with some distinctions. For the NR2 registration, as illustrated in Table 3.1, our SCMWF2 approach yielded the highest Dice scores with 1.2%, 1.9%, 0.6% increase over JOINT in the accumbens, caudate and hippocampus (in both adult datasets) respectively. In the fetal dataset, SCMWF2 provided improvements of 0.5% and 0.7% over JOINT in CSF and CoGM. In fact, JOINT outperformed none of our methods in the segmentation of fetal brain tissues, whereas, in the adult brains, it showed superior performance in pallidum and putamen according to Dice, although the improvements are minor (i.e., 0.2% in both structures) and not manifested in terms of MHD. SCMWF2 achieved the best MHD in all adult subcortical structures except in the thalamus proper, where SCMWF provided the lowest distances. In the fetal dataset, with the exception of the brainstem, the lowest MHD were achieved, depending on the tissue of interest, by SCMWF, SCMNF2 or SCMWF2. Overall, SCMWF2 showed the best performance in the adult datasets for all registration settings, as shown in Table 3.3, while for the fetal dataset, the best results were provided by SCMNF2 or SCMWF2 (see Table 3.4). In all datasets, SCMWF2 was statistically superior in both Dice and MHD to JOINT when using the AF registration setting. In ADNI, SCMWF2 also showed significantly lower MHD when using NR1 and NR2.

Most noteworthy is that when using coarse registrations, such as AF, Dice scores of SCMWF2 were 1%, 0.8% and 1.5% superior to JOINT in SATA,

	Fetal		
	AF	NR1	NR2
MV	0.804±0.042	0.851±0.028	0.931±0.017
Naive	0.803±0.042	0.850±0.028	0.933±0.017
STAPLE	0.797±0.047	0.850±0.035	0.939±0.019
STEPS	0.830±0.037	0.864±0.031	0.938±0.018
LWV	0.879±0.033	0.908±0.020	0.944±0.016
JOINT	0.919±0.023	0.933±0.013	0.942±0.015
SCMNF	0.926±0.022	0.937±0.014	0.946±0.015
SCMWF	0.928±0.021	0.938±0.013	0.946±0.015
SCMNF2	0.929±0.022	**0.941±0.013**	0.947±0.015
SCMWF2	**0.934±0.019***	0.940±0.014	**0.947±0.015**
MV	0.193±0.059	0.138±0.036	0.064±0.035
Naive	0.194±0.059	0.140±0.036	0.063±0.034
STAPLE	0.205±0.062	0.141±0.043	0.057±0.037
STEPS	0.154±0.044	0.120±0.037	0.058±0.035
LWV	0.111±0.043	0.078±0.024	0.049±0.027
JOINT	0.071±0.032	0.055±0.016	0.049±0.025
SCMNF	0.064±0.030	0.052±0.017	0.047±0.025
SCMWF	0.063±0.029	0.051±0.016	0.047±0.025
SCMNF2	0.064±0.031	**0.049±0.016**	**0.046±0.025**
SCMWF2	**0.056±0.023***	0.049±0.017	0.046±0.026

Table 3.4: Fetal brain tissue segmentation. Overall mean Dice scores (top entries) and MHD (bottom entries) for each registration setting. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods.

ADNI and fetal databases, respectively. While when using finer registrations, as is the case of NR2, the improvement reduced to 0.5%, 0.6% and 0.5%. This demonstrates that our approach is more robust against registration failures when compared to the rest of methods. To better illustrate the robustness of our approach, Figures 3.6 and 3.7 displays boxplots of Dice and MHD for each structure or tissue and each registration setting, comparing SCMWF2 and JOINT. As we can observe, the more accurate the registration, the better the segmentations, and generally, the lower the performance gap between both approaches.

3.5 Discussion

Our approach relies on the assumption that systematic segmentation errors caused by registration failures are atlas-dependent, and therefore can be diminished if the appearance patterns that lead to such errors are learned

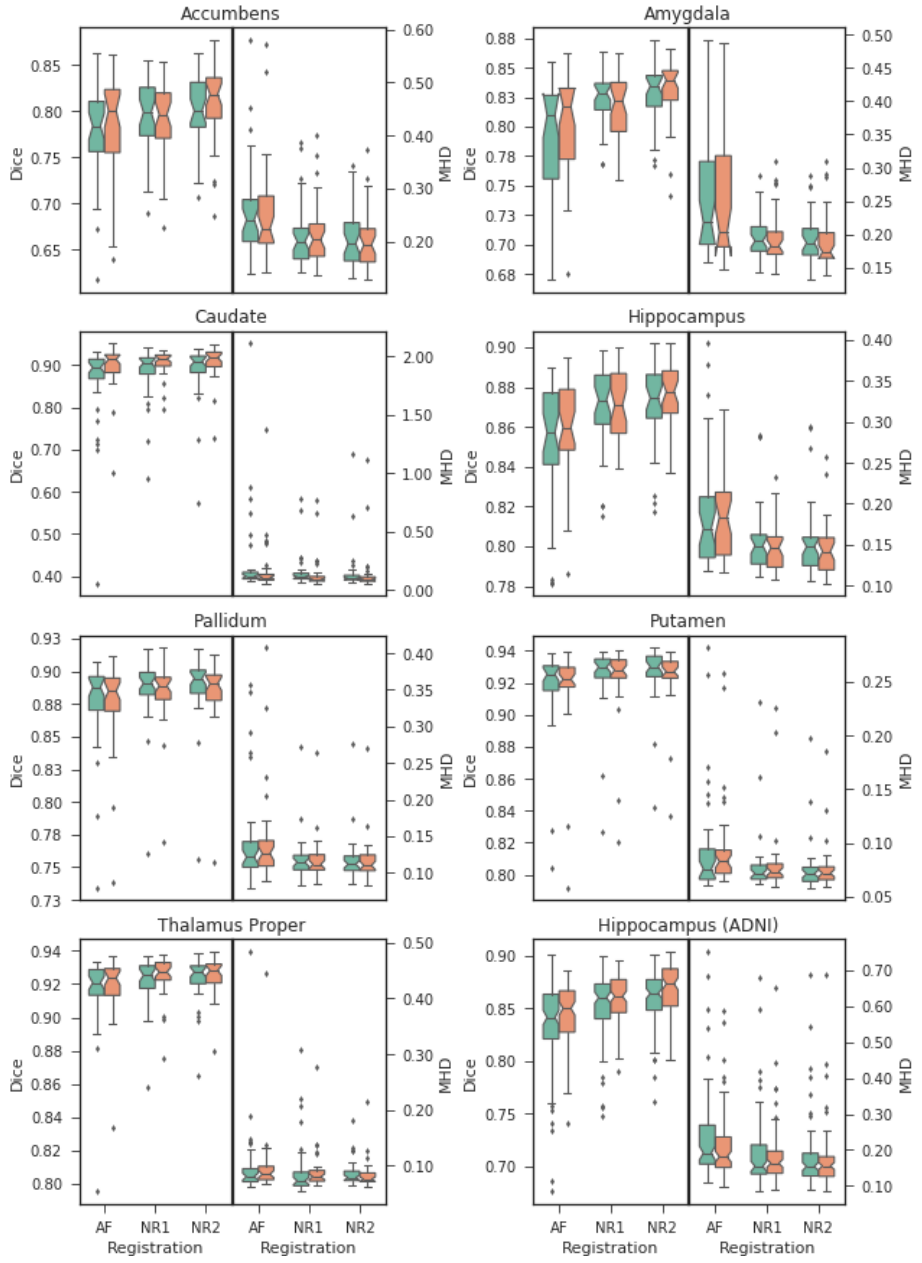

Figure 3.6: Boxplots of Dice and MHD for each subcortical structure from SATA and ADNI datasets comparing SCMWF2 (orange) and JOINT (green).

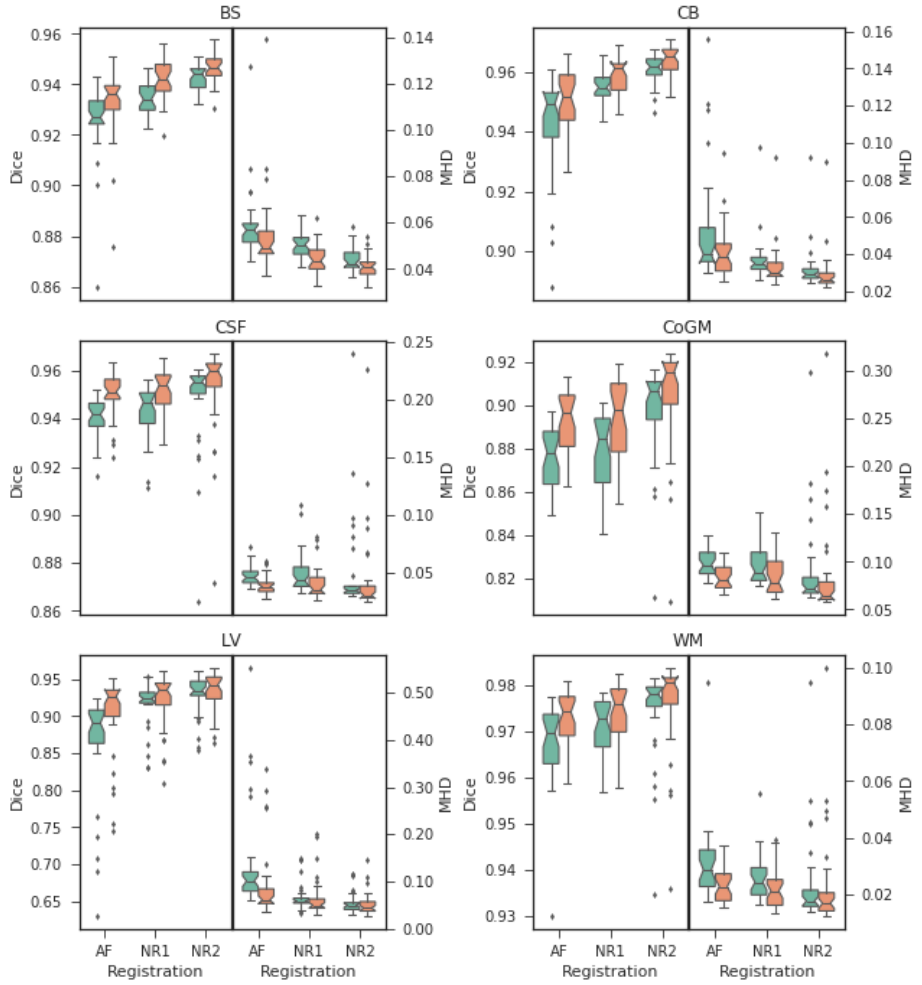

Figure 3.7: Boxplots of Dice and MHD for each tissue from our fetal dataset comparing SCMWF2 (orange) and JOINT (green).

in each atlas space, taking into consideration the registration model. In what follows, we present a methodological comparison with the state of the art, and discuss the strengths and weaknesses of the proposed approach.

3.5.1 Learning from segmentation errors

There is a few number of works in the MAS literature that approach label fusion considering segmentation errors. The concepts of atlas accuracy map proposed in (Sdika, 2010) and reliability map proposed in (Wan et al.,

2008) were computed by co-registering the training atlases. However, both approaches ignored intensity information. Moreover, Wan et al. (2008) performed label fusion only for voxels where the corresponding confidence was superior to a predefined threshold, leaving unlabeled the rest of ambiguous voxels. The main drawback of these approaches is that these maps are static and may incur poor generalization if the target images are considerably different from the atlases in the training set. The risk of overfitting is also present when using offline learning (Iglesias and Sabuncu, 2015). Nonetheless, our confidence estimators do take into account the target patch appearance to compute the local confidences, and in the SCMNF2 and SCMWF2 versions we further select the most similar patch from the training atlases prior to feeding the difference between the atlas and target patches to the confidence estimator.

Supervised learning segmentation approaches existing in the literature (Bai et al., 2015; Hao et al., 2014; Sdika, 2015) learn directly from the labels, gathering the patches from the different atlases to train their classifiers. To the best of our knowledge, the method proposed by Wang et al. (2011a) is the only work that shares similarities with our approach in that both methods learn from segmentation errors instead of labels. However, the wrapper method learns the disagreements between the segmentation produced by a particular host method and the ground truth segmentation in the space of the target images. Whereas in our case, we learn local confidence parameters for each atlas individually as part of a probabilistic label fusion framework.

3.5.2 The benefit of intensity in segmentation accuracy

As already shown in subsection 3.4.4, incorporating intensity information in the label fusion process grants superior performance with regard to the rest of methods (e.g., MV and STAPLE). This is further emphasized by the substantial performance gain of our SCM-based approaches over the Naive method. Nevertheless, using solely intensity-based similarity, without accounting for other factors, is not enough to provide the best segmentations, as illustrated by the performance gap between LWV and, for instance, JOINT. This latter method uses a patch-based weighted voting approach where weight assignment is based on modeling dependencies between pairs of atlas patches and the target image, with the purpose of reducing the confidence of correlated erroneous atlas votes. In our approach, weight assignment accounts for the segmentation errors produced by each atlas

after co-registering the remaining atlases. Besides, intensity samples used for training the confidence estimators are augmented with label-dependent features in the case of SCMWF and SCMWF2.

3.5.3 Similarity-based confidence estimation

Similarity-based approaches employ heuristic measures that may not be directly related to segmentation accuracy. Yet, these approaches (e.g., Coupé et al., 2011) have demonstrated excellent results in MAS. Therefore, in SCMNF2 and SCMWF2, a similarity-based approach is used in combination with supervised learning to build the confidence estimators. In SCMNF and SCMWF, an atlas patch had a single static label (i.e., the corresponding label of the central voxel from the expert segmentation). Segmentation errors were then obtained by comparing this label to the labels of the target atlases, disregarding any clue from the intensity patches. By adopting the many-to-many correspondences scheme, we equipped the atlas with information to decide what label from its surrounding neighborhood corresponds to a particular target patch. Strictly speaking, segmentation errors here are based on a specific similarity measure. Hence, what our confidence estimators try to learn is not the segmentation errors as known in SCMNF (and SCMWF), but the segmentation errors produced by an atlas through employing this specific similarity measure.

This may seem computationally more costly than the procedure used in SCMNF since we introduce an additional intermediate stage (i.e., k nearest neighbors). Nonetheless, k nearest neighbors did not suppose an important overhead and learning times were similar as mentioned in subsection 3.4.3. Additionally, SCMNF2 demonstrated the benefits of this approach by yielding segmentation results comparable to SCMWF, with SCMWF2 outperforming SCMWF in most structures. The proposed approach uses a simple and fast classifier (i.e., logistic regression). In the state of the art, existing MAS approaches used SVM (Bai et al., 2015; Hao et al., 2014; Sdika, 2015) and random forest (Wang et al., 2014b) to learn their local classifiers. For example, the learning-based approach proposed by Bai et al. (2015) used SVM with the radial basis function kernel. Thus, using a simpler model, such as logistic regression in our case, can lead to reduced training times, especially when thousands of local classifiers are to be learned.

The choice of the classifier, however, is not straightforward and is application dependent. Its performance may depend on several factors including:

image modalities, nature of the features, number of samples, *etc.* Given the modularity of the proposed method, other supervised learning approaches can be used to learn our confidence estimators. Deep learning, for instance, is gaining an increasing interest in medical image analysis (Litjens et al., 2017). In our case, with deep learning, we can take advantage of the 3D nature of the image patches rather than representing the patches as feature vectors (Cireşan et al., 2012; de Brébisson and Montana, 2015). Beyond patch-based approaches, architectures such as the U-net proposed by Ronneberger et al. (2015), and its 3D extensions (Çiçek et al., 2016), can also be used as global confidence estimators since they can take whole images as input and output a classification for each pixel/voxel, which permits to take into consideration larger contextual information for each voxel by analyzing the images at multiple scales. Although more time-consuming than logistic regression, this is a promising line of future work.

3.5.4 The effect of label-dependent features

By incorporating label-dependent features, results showed that both SCMWF and SCMWF2 can produce state-of-the-art results. The benefits of using additional features beyond patch intensity is well-demonstrated in (Asman et al., 2015; Bai et al., 2015; Hao et al., 2014; Wang et al., 2011a). In (Asman et al., 2015), 1009 dimensional feature vectors were used. Similarly, high-dimensional vectors of 1003 features were used by Wang et al. (2011a). For voxel-wise classifiers, Hao et al. (2014) used feature vectors of 379 elements and SVM with l1-regularization to select the sparsest solution from all the possibly redundant features. In (Bai et al., 2015), 260-element samples were built from intensity, gradient and contextual features. In our case, SCMNF and SCMNF2 used feature vectors of only 27 dimensions (i.e., intensity patch of $3 \times 3 \times 3$), whereas augmented feature vectors of 33 elements (i.e., 6 additional label-dependent features) were used in SCMWF and SCMWF2. The informative power of the label-dependent features can be observed in the considerable increase in performance of SCMWF with respect to SCMNF, especially in the adult brain dataset. However, the boost achieved by SCMWF2 with regard to SCMNF2 is not that large, except in the segmentation of the hippocampus from the ADNI dataset, with 1.3% increase in Dice. This is possibly due to the fact that we are reaching inter-rater variability is some structures (e.g., thalamus proper). Moreover, more compact segmentations were obtained by SCMNF2 than SCMWF2 for the accumbens, pallidum, putamen and thalamus proper in terms of MHD for

the NR2 registration, as shown in Table 3.1. In the fetal dataset, there is no consistent difference in performance of SCMNF and SCMNF2 with their label-dependent versions.

The idea of using information from the label patch was already used in (Wang et al., 2011a) by appending to the feature vectors the segmentation labels produced by the host method in the neighborhood of each voxel. However, they just used the raw label patch as features. Our label-dependent feature extraction procedure is different from all the aforementioned approaches. Instead of predefined filters, we use the atlas label patch as a mask to compute the difference between the features extracted from each region of the intensity patch. This reduced number of features seemed to provide the local classifiers with potential information to better discriminate between correct and erroneous atlas patches.

Feature extraction is computationally expensive, especially when using the non-local approach with voxel-wise classifiers. As reported in subsection 3.4.3, offline learning took 3 minutes for SCMNF, with a runtime increase of 7 minutes for SCMWF when incorporating the label-dependent features. Considering sample selection strategies may turn out advantageous to decrease computational cost. Bai et al. (2015), for instance, performed patch selection to reduce such computations.

3.5.5 The influence of outliers

Target images that highly deviate from the anatomies in the training set have a negative impact in registration, giving rise to misaligned structures (i.e., outliers). However, our approach has proven its robustness against outliers since our proposed approach achieved the best segmentations with the lowest Dice and MHD standard deviations for all registration settings in both adult (see Table 3.3) and fetal (see Table 3.4) datasets. Moreover, the similarity-based label fusion approaches used in the conducted experiments (i.e., LWV and JOINT) seem to perform poorly in the segmentation of outliers, as illustrated in Figure 3.8. This figure shows the caudate from SATA and lateral ventricles from our fetal dataset with the worst segmentations provided by all methods. For the lateral ventricles, Dice overlaps produced by LWV, JOINT, SCMNF2 and SCMWF2 using AF registration are 0.542, 0.630, 0.632 and 0.754 (0.779, 0.831, 0.842 and 0.841 for NR1 registration), respectively. For the caudate, the Dice scores achieved by LWV, JOINT, SCMNF2 and SCMWF2 using AF registration are 0.150,

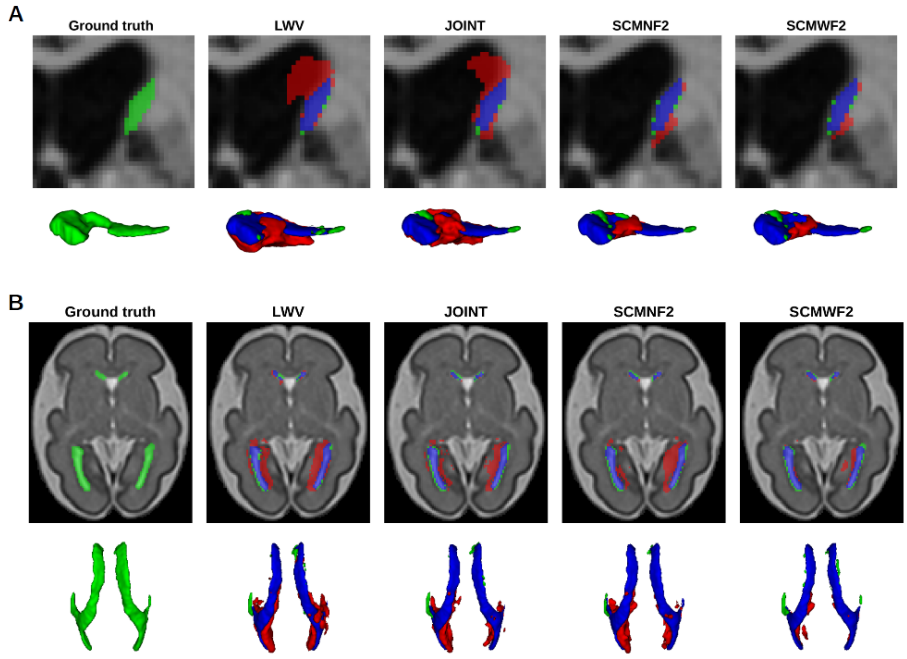

Figure 3.8: Segmentation of outliers. Illustration of A) left caudate and B) lateral ventricles automatic segmentations with the lowest Dice scores and largest MHD using NR1 and AF registrations, respectively. Green and red depict manual and automatic segmentations respectively. Overlap is depicted in blue. For each structure, the first row shows ground truth and automatic segmentations in coronal (caudate) and axial (lateral ventricles) views. The second row shows the corresponding 3D renderings.

0.217, 0.396 and 0.417 (0.415, 0.486, 0.692 and 0.701 for NR1 registration), respectively. Note that overlaps provided by all methods when using the affine registration are below 0.5, and for NR1 registration both LWV and JOINT Dice overlaps are below 0.5. We can observe that this is a clear failure in registration caused by the enlarged ventricles next to the caudate.

In both cases, segmentation errors produced by the aforementioned similarity-based approaches are mainly due to over-segmentation (i.e., red color). Especially important is the over-segmentation produced by identifying part of the left ventricle as caudate, although there is great difference in intensity because the ventricles have a lower intensity range. In case on the lateral ventricles, the fetal dataset is composed of healthy controls and fetuses with dilated ventricles, which might have misguided the fusion process to label voxels in the periventricular regions as ventricle. This is further illustrated by the Dice and MHD boxplots of the caudate and lateral ventricles in Figures 3.6 and 3.7, respectively. Note, however, that results shown in the

boxplots corresponding to the subcortical structures do not coincide with the Dice scores of this outlier because the boxplots were created by averaging the results achieved for the left and right caudate. Our confidence estimators seem to show more robustness in the presence of outlier patches. Another clear impact of the presence of outliers in performance of JOINT can be seen in the boxplots of the cerebellum, thalamus proper and the hippocampus from ADNI, notably with the affine registration.

3.5.6 Robustness to registration failures

The claim of our work is that systematic segmentation errors due to registration can be substantially mitigated using the proposed approach. Thus, we studied the effect of registration in segmentation results in order to asses how segmentation performance evolves from using more coarse (i.e., AF) to finer registrations (i.e., NR2). From the overall segmentation performance reported in Tables 3.3 and 3.4 and boxplots in Figures 3.6 and 3.7, we can conclude that our method is more robust to registration errors. With coarse registrations, which are more prone to failures, our approach achieved the largest performance increment compared to the rest of approaches. Therefore, demonstrating to be robust to registration failures. Overall mean Dice scores obtained by SCMWF2 with AF registration were lower (0.865 overlap in SATA, 0.843 in ADNI and 0.934 in our fetal dataset) than the ones obtained with NR2 (0.880, 0.866 and 0.947). The same occurs with MHD, with SCMWF2 providing larger distances (0.172 mm in SATA and 0.212 in ADNI and 0.056 in our fetal dataset) when using AF than the distances achieved with NR2 (0.137, 0.171 and 0.046). This difference in the performance of our approach between AF and NR2, could be substantially reduced by taking advantage of the many-to-many correspondences scheme and using larger patch and window search sizes, rather than the $3 \times 3 \times 3$ size used in this work.

3.5.7 Limitations and future directions

The main limitation of our approach is that confidence learning is performed for each training atlas, which makes it computationally expensive. One possible way to lessen this computational burden is to restrict the learning process to the most representative atlas spaces, for example, by clustering the atlases and only learning in the centroid spaces. To further

reduce computational time, although at the expense of sacrificing segmentation accuracy, clustering can also be applied to learn considering groups of neighboring voxels instead of using voxel-wise classifiers. On the other hand, in this work, the size of intensity patches and search neighborhoods used in both target an atlas spaces was set to $3 \times 3 \times 3$ (i.e., radius of 1). As future work, the impact of this parameter in segmentation accuracy can be studied. Likewise, parameter tuning can be performed to find the best value for the penalty parameter used in logistic regression. This parameter can be optimized in a local manner for each of the voxel-wise classifiers or for the whole structure.

The methods based on the many-to-many correspondences scheme yielded very accurate results. Still, more work is required here to asses the effect of the similarity metric and the strategy used to predict the label, considering only the most similar patch or some heuristic weighting the patches contribution to the final label. In the fetal dataset, our label-dependent features did not provide a comparable boost in performance as in the segmentation of the adult datasets. Thus, another direction of future work is to explore different strategies to extract our label-dependent features or even adopt a supervised approach to learn such features rather than using feature engineering. Moreover, additional features as the ones used in (Bai et al., 2015; Hao et al., 2014) could be considered. Finally, a very promising direction of future work is to consider correlations between voxels and/or the votes of the training atlases (Wang et al., 2014b). Voxel correlations are especially important to produce compact segmentations and remove isolated regions, as shown in Figure 3.8.

3.6 Conclusions

Registration failures constitute a potential source of systematic errors in MAS. In this manuscript, we have proposed a probabilistic label fusion framework that takes into consideration local atlas confidences at each point by the estimation of the so-called spatial confidence maps. Given the nature of our approach, we have also proposed a novel label-dependent feature extraction that provided valuable information in the prediction of the confidences. Systematic errors due to registration are accounted for during label fusion since confidence learning is performed in atlas space. As opposed to STAPLE-like approaches, this learning process is performed in an offline manner using the available training atlases. Therefore, computational

complexity at test time is comparable to the simplest approaches. Furthermore, incorporating neighborhood information in atlas space to compute the segmentation errors rendered our approach more robust to registration errors. Experimental results have shown that our approach yields superior performance to state-of-the-art approaches in the segmentation of fetal brain tissues and the majority of subcortical brain structures.

4

Patch Spaces and Fusion Strategies in Patch-based Label Fusion

Abstract – In the field of multi-atlas segmentation, patch-based approaches have shown promising results in the segmentation of biomedical images. In the most common approach, registration is used to warp the atlases to the target space and then the warped atlas labelmaps are fused into a consensus segmentation based on local appearance information encoded in form of patches. The registration step establishes spatial correspondence, which is important to obtain anatomical priors. Patch-based label fusion in the target space has shown to produce very accurate segmentations although at the expense of registering all atlases to each target image. Moreover, appearance (i.e., patches) and label information used by label fusion is extracted from the warped atlases, which are subject to interpolation errors. In this work, we revisit and extend the patch-based label fusion framework, exploring the role of extracting this information from the native space of both atlases and target images, thus avoiding interpolation artifacts, but at the same time, we do it in a way that it does not sacrifice the anatomical priors derived by registration. We further propose a common formulation for two widely-used label fusion strategies, i.e., similarity-based and a particular type of learning-based label fusion. The proposed framework is evaluated on subcortical structure segmentation in adult brains and tissue segmentation in fetal brain MRI. Our results indicate that using atlas patches in their native space yields superior performance than warping the atlases to the target image. The learning-based approach tends to outperform the similarity-based approach, with the particularity that using patches in native space lessens the computational requirements of learning. As conclusion, the combination of learning-based label fusion and native atlas patches yields the best performance with reduced test times than conventional similarity-based approaches.

This chapter is adapted from:
Benkarim O. M., Piella G., Hahner N., Eixarch E., González Ballester M. A., and Sanroma G. Patch spaces and fusion strategies in patch-based label fusion. [Under review.]

4.1 Introduction

Segmentation of brain magnetic resonance imaging (MRI) is an important step for many neuroimaging studies. The strive for reproducibility and reasonable runtimes of segmentations renders automatic labeling approaches more appealing than manual segmentation. There is an important body of literature concerning automatic segmentation of biomedical images that can be categorized into supervised and unsupervised approaches. Atlas-based segmentation falls in the former category in that it uses a set of atlases (i.e., images with their corresponding predefined labelmaps) to segment new images. Single atlas-based approaches employ a single atlas, which is registered to the target image in order to find a mapping that is used to propagate its labelmap to the target image. Nonetheless, single atlas-based segmentation exhibits poor performance under the occurrence of registrations errors and/or when the target image is different from the atlas. Multi-atlas segmentation (MAS) approaches overcome this limitation by using several atlases (Aljabar et al., 2009; Heckemann et al., 2006; Lötjönen et al., 2010), and consequently, better encompass the neuroanatomical variability of the target population and alleviate the impact of registration failures.

MAS has recently shown promising results in the segmentation of biomedical images (González-Villà et al., 2016; Iglesias and Sabuncu, 2015; Sanroma et al., 2016b). The major steps common to all MAS approaches consist on registration and label fusion (LF). During registration, the atlases and the target image are spatially transformed to the same space in order to establish spatial correspondences. Then, LF optimally combines all atlas labelmaps into a consensus segmentation on the target image. One of the simplest approaches to perform LF is majority voting (MV) (Klein et al., 2005; Rohlfing et al., 2004), which labels a target voxel using the most frequent label among the training atlases. Although MV provides better results than single atlas-based segmentation, its performance is still sensitive to registration errors because atlas labels are uniformly combined regardless of how accurate is their registration to the target. More sophisticated fusion strategies can be used. We can identify two main fusion strategies in the literature: similarity-based and learning-based LF strategies. Similarity-based LF strategies assign a weight to each atlas vote based on its similarity to the target image (e.g., Artaechevarria et al., 2009; Coupé et al., 2011; Sabuncu et al., 2010), whereas learning-based strategies tackle LF from the machine learning perspective by learning a labeling function from a set of training examples that maps image appearances to the correct anatomical label (Bai

et al., 2015; Benkarim et al., 2017a; Hao et al., 2014). As a limitation of learning-based approaches compared to similarity-based approaches is that the former require an additional step to learn the labeling function at test time for each point.

Independently of the fusion strategy, in order to account for the spatially varying registration quality, several authors have proposed local (Artaechevarria et al., 2009; Isgum et al., 2009; Sabuncu et al., 2010) and non-local (Coupé et al., 2011; Hao et al., 2014; Wang et al., 2013) LF approaches. Among these methods, patch-based LF estimates the weights or learns the labeling function based on 3D intensity patches around the voxels of interest. Inspired by the non-local means approach (Buades et al., 2005), patch-based LF is further extended to the one-to-many correspondences scheme, considering all atlas patches in the neighborhood of the voxel of interest (Benkarim et al., 2017a; Coupé et al., 2011; Rousseau et al., 2011).

In this paper, we focus on the non-local patch-based LF framework. Although registration quality is not critical thanks to the non-local means strategy, approaches within this framework still require establishing correspondences between the atlases and the test image in order to locally estimate the weights or learn the labeling function. However, these approaches do not discern between establishing correspondences and atlas warping, treating them as a unique and transparent process. In this work, we propose a patch-based LF framework that distinguishes between establishment of correspondences and image warping compared to traditional approaches (i.e., based on warped atlases) that consider both steps as indivisible. To do so, we revisit the patch-based LF framework to identify its main steps, namely: 1) the fusion space where LF is carried out, 2) the fusion strategy used to label test voxels, and 3) the patch space used to extract atlas and test patches. Furthermore, we adapt two widely-used LF strategies (i.e., similarity-based and learning-based LF) to our proposed framework and analyze their impact on runtime and performance. Our proposed LF framework allows us to sample atlas and test patches in native space with the spatial priors encoded in the correspondences. This turns out to be especially advantageous for learning-based fusion strategies as learning is performed offline, regardless of the target image.

This paper builds on our recently published work (Benkarim et al., 2017b). In this current work, we present a more in-depth description of the design choices that conform the patch-based LF framework; we propose a common formulation for similarity-based and a particular type of learning-based LF

(i.e., SVM with RBF kernel), which may be useful for their common understanding; we evaluate the performance of our proposed LF framework in 2 brain MRI datasets: adult brain subcortical structure segmentation and tissue segmentation of fetal brain MR images; and we provide a comparison to traditional approaches.

The outline of the paper is as follows. Section 4.2 reviews the patch-based label fusion framework. Section 4.3 presents the details of our proposed framework. In Section 4.4 we describe the experimental setting and present the results. In Section 4.5 we discuss the advantages and limitations of our approach. Section 4.6 concludes the paper.

4.2 The patch-based label fusion framework

In the patch-based LF framework, registration is crucial for establishing accurate voxel-wise correspondences between the target image and each of the training atlases. In this section, we revisit the patch-based LF framework to identify the main design choices that define any patch-based segmentation method and analyze their strengths and weaknesses. There are several choices in patch-based LF that need to be considered in the fusion process, namely: fusion space, fusion strategy and patch space.

4.2.1 Fusion space

This is the space where the estimated segmentation of the target image is computed via LF. It is typically done in one of the following spaces:

1. Target fusion space: the target labelmap is computed directly in the target space. To that end, spatial transformations between the atlases and the target image are computed to warp the atlases to the target space, where LF takes place based on the target image T and the warped atlases $\tilde{A}_i$, as illustrated in Figure 4.1.

2. Template fusion space: the target labelmap is computed in a template space. To that end, both atlases and target image are warped to a template space (e.g., MNI152) and LF is performed in this common space using the warped atlases $\tilde{A}_i$ and the warped target $\tilde{T}$. The estimated segmentation is then warped back to the target space, as shown in Figure 4.2.

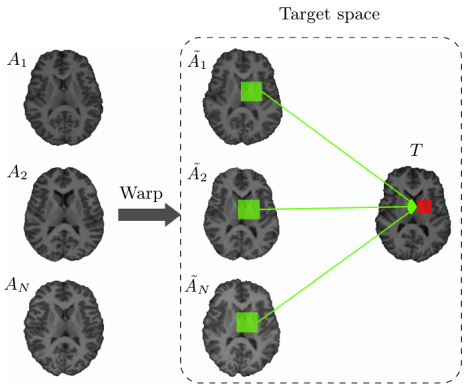

Figure 4.1: Target fusion space. The target labelmap is computed directly in the target space via LF using the warped atlases $\tilde{A}_i$.

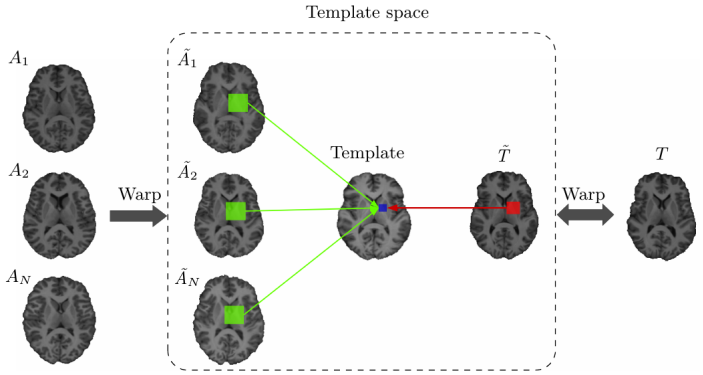

Figure 4.2: Template fusion space: target labelmap is computed in the template space using the atlases and target warped to the template space. Then, the estimated segmentation is warped back to the target space.

In the vast literature on MAS, target fusion space is the most common. Its advantages are twofold: 1) the estimated target labelmap does not need to be warped back to the target space, thus incurring in no label interpolation errors; and 2) the target image suffers no distortion due to registration-based interpolations. However, given N_a training atlases and N_t target images, the computational burden of target fusion space is high because it requires registering all training atlases to each target image (i.e., $N_a \times N_t$ registrations). To lower the computational requirements, researchers proposed non-local patch-based LF strategies that only use coarse registrations (i.e., affine) (Coupé et al., 2011; Rousseau et al., 2011). Although such an approach is faster, we still need to perform $N_a \times N_t$ atlas warpings (intensity

images and labelmaps). In order to reduce the number of registrations, template fusion space turns out to be appealing since only $N_a + N_t$ registrations are needed to segment N_t target images. Nevertheless, in template fusion space, we do not work with the original target image but with a deformed version in the template space. What is commonly done is to concatenate registrations through the template. This still allows working in target space but with the same number of registrations as in template space.

With regard to the fusion strategy, target fusion space is adopted especially for methods using similarity-based fusion strategies (Artaechevarria et al., 2009; Coupé et al., 2011; Wang et al., 2013), whereas template fusion space is usually adopted by methods using a learning-based fusion strategy (Manjón and Coupé, 2017; Sanroma et al., 2015a). The main reason behind this choice is that template fusion space allows voxel-wise classifiers to be trained offline. On the other hand, using the target fusion space implies learning the classifiers online for each new target image, which is computationally demanding (Hao et al., 2014; Tong et al., 2013). Classifiers are trained on the warped atlases to segment a single target image and cannot be reused to segment other target images. Here, atlas selection (Aljabar et al., 2009; van Rikxoort et al., 2010) can be used to reduce the size of the training set and hence learning times, though atlas selection introduces another free parameter (e.g., the number of most similar atlases) into the MAS setting, to be chosen during an intermediate validation step, for instance.

4.2.2 Fusion strategy

The strategy used to fuse the multiple atlas labelmaps is an important step in MAS. In this work, we focus on two widely-known strategies: similarity-based and learning-based LF. Given a set of atlas patches x_i and their corresponding labels y_i, the target patch x_t is segmented as follows:

1. Similarity-based LF: atlas labels, y_i, are weighted according to the similarity of their intensity patches with the target patch (Coupé et al., 2011; Rousseau et al., 2011). The more similar x_i is to the target patch x_t, the higher the contribution of its label, y_i, to the final labeling of the target patch. If we use the exponential of the negative sum of square differences to compute the weights (Coupé et al., 2011), for the binary case, we have:

$$\hat{y}_t = sign\Big(\sum_i \exp(-\gamma \|x_k - x_t\|^2) y_i\Big), \tag{4.1}$$

where $\gamma = \min \|x_k - x_t\|_2$ and $\hat{y}_t \in \{-1, +1\}$ is the estimated label for the target patch, indicating foreground (i.e., $+1$) or background (i.e., -1).

2. Learning-based LF: these are supervised approaches that learn from the atlas patches x_i a function that maps local image appearances to the corresponding label, y_i (Bai et al., 2015; Hao et al., 2014). When a target image arrives, this function is used to predict a label for each target voxel based on its patch, x_t, i.e.,:

$$\hat{y}_t = f(x_t), \tag{4.2}$$

where $f()$ is the labeling function learned from the training pairs (x_i, y_i).

Note that, for both label fusion strategies, we assume that patches are already in correspondence, which is established during the registration process. Furthermore, correspondences can be established using 1) a one-to-one scheme to only capture local information (i.e., patch) or one-to-many scheme that considers all atlas patches in the neighborhood of the voxel of interest. In this paper, we consider the latter scheme.

In our setting, where voxel-wise classifiers are required, learning-based approaches are, generally, more time-consuming since we have to learn a model for each voxel. This burden is especially accentuated when adopting an online strategy, learning different classifiers for each target image. Offline strategies, on the other hand, considerably reduce the computational requirements by learning a single classifier per voxel, independently from the target image. Similarity-based LF can also be seen as a learning-based approach, falling within the category of lazy learning. These approaches tend to be faster and do not need to learn a model from the training patches.

4.2.3 Patch space

Patch space refers to the space used to extract the appearance (e.g. patches) and label information used by the LF process. In most approaches, the patch space coincides with the fusion space. Particularly, LF in target fusion space

is based on patches extracted from the atlases warped to the target, $\tilde{A}_i$ (see green arrows in Figure 4.1), not from the atlases in their native spaces (i.e., A_i). Similarly, in template fusion space, patches are extracted from the atlases and the target image warped to the template space (i.e., $\tilde{A}_i$ and $\tilde{T}$ respectively, as shown in Figure 4.2). Irrespective of the fusion space, the warped atlases are subject to interpolation errors. This is an important drawback because the interpolation strategy used for the atlas intensity images (e.g., linear) is different from the one used for the labelmaps (e.g., nearest neighbors). In this way, a deformed labelmap might no longer be faithful to its corresponding deformed intensity image under the manual segmentation protocol followed by the expert. In case of using the template as fusion space, there is an additional source of error due to interpolation of the estimated labelmap, which occurs when warping it back to the target space.

4.3 Proposed framework

So far, researchers have paid little attention to the space from where the patches are extracted. In this work, we propose a LF approach that uses appearance and label information from the native spaces of both atlases and target images, the so-called *native patch space*. Thus, registration is used only to find spatial correspondences between the atlases and the target image, without deforming any of them. The advantages of native patch space are threefold: 1) there is no need to warp the atlases, thus avoiding any inaccuracies between atlas images and corresponding labelmaps due to interpolation artifacts. This allows LF to be driven by the true appearance patterns used by the expert to create the ground truth, 2) avoiding warping the atlases also implies a higher storage efficiency, since there is no need to keep two instances of each atlas (i.e., the original and the warped one), and 3) learning-based fusion strategies can be applied directly in the target fusion space, without the need to train the classifiers online for each target.

The proposed approach is based on the observation that LF in template or target fusion spaces relies on different (i.e., deformed) versions of the atlases and not the original ones, and therefore, there exists a risk of introducing noise in the segmentation process due to interpolation errors. In contrast, in our approach, the mappings computed during registration are only used to find spatial correspondences between the atlases and the target image, but the images are not deformed. Moreover, in order to avoid registering

the atlases each time a novel target image arrives, atlases are registered to a common reference space. Figure 4.3 illustrates how LF is carried out in our novel approach. Throughout the rest of the paper, we will refer to it as LF in native patch space.

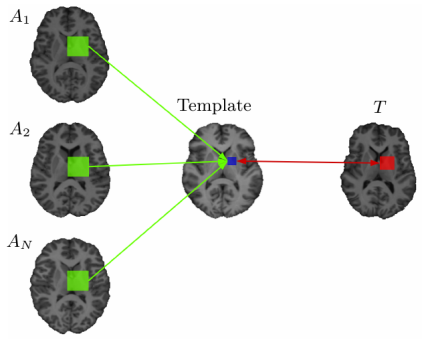

Figure 4.3: LF in native patch space. Correspondences between the atlases A_i and the target image T are used to extract the patches from their respective native spaces. Target labelmaps are directly computed in the target fusion space.

4.3.1 Extraction of patches in native space

We adopt the one-to-many correspondences scheme by using the non-local patch-based LF (Coupé et al., 2011). Let ϕ_i be the mapping of the i-th atlas, A_i, to the template space and ϕ_t the mapping of the target image, which were obtained from the registration step. For each voxel v_t and corresponding patch x_t (see red box in Figure 4.3) on the target image, we proceed as follows:

1. Find the corresponding voxel in template space (red arrow in Figure 4.3):
$$\tilde{v} = \phi_t(v_t). \tag{4.3}$$

2. Find corresponding voxels in each training atlas space (green arrows in Figure 4.3):
$$v_i = \phi_i^{-1}(\phi_t(v_t)) = \phi_i^{-1}(\tilde{v}),\ i = 1, \ldots, N_a, \tag{4.4}$$
with i indexing the atlases in the database.

3. Extract patches and corresponding labels in the neighborhood of v_i (see green boxes in Figure 4.3):

$$D = \{(x_{ij}, y_{ij}) \,|\, \forall j \in S(v_i),\ i = 1, \ldots, N_a\}, \tag{4.5}$$

where y_{ij} is the label of the j-th voxel in the neighborhood of v_i, denoted as $S(v_i)$.

4. Label target patch x_t using the set of patches D and some LF strategy as explained in the following.

4.3.2 Fusion strategies in native space

To perform LF for the binary case, we define our labeling function as:

$$\hat{y}_t = sign\Big(\sum_{k=1}^{|D|} K(x_k, x_t) y_k + b\Big), \tag{4.6}$$

where $K(x_k, x_t)$ is the similarity between the k-th patch in D and x_t, $y_k \in \{-1, +1\}$ is its corresponding label, $|\cdot|$ denotes cardinality, and b is a bias term.

Within this framework, we propose to estimate the weights using a similarity-based approach (SimLF) (Artaechevarria et al., 2009; Coupé et al., 2011). In SimLF, the weight of each atlas label is based on the intensity similarity between its patch and the target patch. Note that setting $b = 0$ and $K(x_k, x_t) = \exp(-\gamma \|x_k - x_t\|^2)$, with $\gamma = \min \|x_k - x_t\|_2$ (Coupé et al., 2011), SimLF is equal to the similarity weighted voting in Eq. (4.1), which estimates the label for the target patch x_t as a weighted combination of the labels in D.

This first LF approach uses all patches in D to segment a given target patch. This can negatively impact the performance of LF since there may be many noisy or irrelevant patches, especially when using the one-to-many correspondences scheme to extract atlas patches. To prevent this, some approaches have been proposed to only use the most reliable patches in the LF process. For example, structural similarity (Wang et al., 2004) was used by Coupé et al. (2011), as an intermediate step, to discard the most dissimilar patches prior to label fusion. For the same purpose, other existing approaches employ sparse patch-based reconstruction techniques (e.g., Wu

et al., 2015a). In this work, we pre-select the most reliable patches by extending our labeling function in Eq. (4.6) as follows:

$$\hat{y}_t = sign\Big(\sum_{k=1}^{|D|} K(x_k, x_t) y_k \alpha_k + b\Big) \tag{4.7}$$

where $\alpha_k \geq 0$ is the pre-selection coefficient of the k-th training patch, such that $\alpha_k > 0$ only if patch x_t is found to be relevant for label fusion. In order to find the α_k coefficients and the intercept b, we propose to use a supervised approach. This is our second fusion strategy, LearnLF, which consists on learning a labeling function that relates the appearance information of patches and their corresponding anatomical labels (x_k, y_k) using as training set the atlas patches and labels in D. Learning-based fusion strategies have already been proposed in the literature (Hao et al., 2014; Tong et al., 2013). Noteworthy is the fact that Eq. (4.7) corresponds to the dual decision function used in the well-known kernel-SVM (Cortes and Vapnik, 1995) to perform predictions. When $K(x_k, x_t) = \exp(-\gamma \|x_k - x_t\|^2)$, it is called the radial basis function (RBF) kernel, with γ being the inverse kernel width. Hence, the labeling function for our LearnLF strategy is obtained by optimizing the following cost function corresponding to the dual problem of kernel-SVM:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{|D|} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{|D|} \alpha_i \\ &\text{subject to} \quad \sum_{i=1}^{|D|} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C, \end{aligned} \tag{4.8}$$

where C is a regularization parameter that controls misclassification of training samples.

Similarly to SimLF, LearnLF also weighs the contribution of the atlas labels based on their similarity. Note the striking similarity between SimLF and LearnLF when using the RBF kernel in LearnLF. Key differences between both approaches are that LearnLF includes the learned coefficients α_k, which, by definition of kernel-SVM, are only different than zero for the atlas patches playing the role of so-called support vectors. This can be interpreted as a form of patch selection. Moreover, the scale parameter γ was

defined in SimLF as the distance of the target patch to the most similar atlas patch, $\gamma = \min \|x_k - x_t\|_2$ (Coupé et al., 2011), while in LearnLF γ was set to the patch size. Finally, note that since the proposed approach uses the native patch space, the atlas training patches for each classifier are invariant to the to-be-labeled target image, and therefore the classifiers for each point can be learned offline.

4.4 Experiments

In this section, we evaluate our proposed LF in native patch space in the segmentation of subcortical brain structures of 35 adult brain MR images from the MICCAI 2013 SATA Challenge dataset and in tissue segmentation of 32 fetal brain MRI. The proposed approach is compared to traditional LF approaches. More specifically, we compared the LF performance using all the patch spaces, namely, template, target and the proposed native patch space. Both SimLF and LearnLF fusion strategies were used in each of the patch spaces. For the latter, the regularization hyperparameter of the SVM was set to $C = 1$.

Different patch and neighborhood sizes were used, with a radius of 1 and 2. For validation, a 3-fold cross-validation procedure was used with Dice similarity coefficient between the ground truth and the estimated segmentations to assess performance. Finally, all experiments were replicated using two registration settings: affine and non-rigid using the symmetric diffeomorphic mapping of ANTs (Avants et al., 2008).

As preprocessing, all images in the adult dataset were registered to the MNI152 template (Fonov et al., 2009). For the fetal dataset, images were registered to a template built using the images in our dataset. To establish correspondences for LF in target fusion space, pairwise mappings between the images were obtained by composing the transformations through the template. Furthermore, for image intensity to be consistent across atlases, histogram matching was used (Nyul et al., 2000) for both datasets.

4.4.1 Segmentation of subcortical structures

The MICCAI 2013 SATA Challenge dataset[5] is composed of 35 T1-weighted MR images of control subjects with age ranging from 19 to 90 years (32.4

[5] https://masi.vuse.vanderbilt.edu/workshop2013

	Affine			Non-rigid		
	Template	Target	Native	Template	Target	Native
Acc	0.758±0.046	0.749±0.057	**0.783±0.052**	0.717±0.058	0.786±0.037	**0.797±0.035**
Amy	0.778±0.041	0.765±0.062	**0.790±0.044**	0.773±0.031	0.817±0.029	**0.818±0.029**
Cau	0.865±0.070	0.856±0.101	**0.884±0.083**	0.840±0.071	0.883±0.071	**0.892±0.070**
Hip	0.836±0.028	0.828±0.037	**0.846±0.032**	0.820±0.021	0.857±0.024	**0.862±0.023**
Pal	0.857±0.041	0.855±0.049	**0.869±0.043**	0.852±0.028	0.880±0.027	**0.882±0.027**
Put	0.900±0.029	0.907±0.030	**0.914±0.027**	0.882±0.022	**0.922±0.018**	0.921±0.019
Tha	0.909±0.026	0.908±0.027	**0.914±0.028**	0.902±0.023	**0.920±0.014**	0.920±0.016

Table 4.1: Adult dataset: Average Dice overlaps (and standard deviation) of automatic segmentations produced by SimLF using template, target and native patch spaces. Bold type indicates the best average segmentation performance.

	Affine			Non-rigid		
	Template	Target	Native	Template	Target	Native
Acc	0.770±0.039	0.788±0.040	**0.797±0.044**	0.715±0.059	0.796±0.042	**0.801±0.039**
Amy	0.794±0.026	0.803±0.027	**0.815±0.023**	0.764±0.033	0.812±0.027	**0.824±0.023**
Cau	0.876±0.058	0.898±0.056	**0.903±0.054**	0.838±0.073	0.902±0.051	**0.904±0.055**
Hip	0.855±0.018	0.867±0.020	**0.874±0.017**	0.822±0.023	0.876±0.015	**0.880±0.014**
Pal	0.866±0.028	0.878±0.032	**0.884±0.028**	0.849±0.029	0.884±0.025	**0.887±0.025**
Put	0.905±0.028	0.919±0.029	**0.924±0.025**	0.882±0.026	0.924±0.020	**0.927±0.020**
Tha	0.915±0.019	0.921±0.018	**0.923±0.017**	0.903±0.020	0.924±0.011	**0.926±0.011**

Table 4.2: Adult dataset: Average Dice overlaps (and standard deviation) of automatic segmentations produced by LearnLF using template, target and native patch spaces. Bold type indicates the best average segmentation performance.

years old in average). The size of the images is $256 \times 256 \times 287$ with a spatial resolution of 1 mm isotropic. Ground-truth segmentations are provided for seven subcortical structures: accumbens (Acc), amygdala (Amy), caudate (Cau), hippocampus (Hip), pallidum (Pal), putamen (Put) and thalamus proper (Tha).

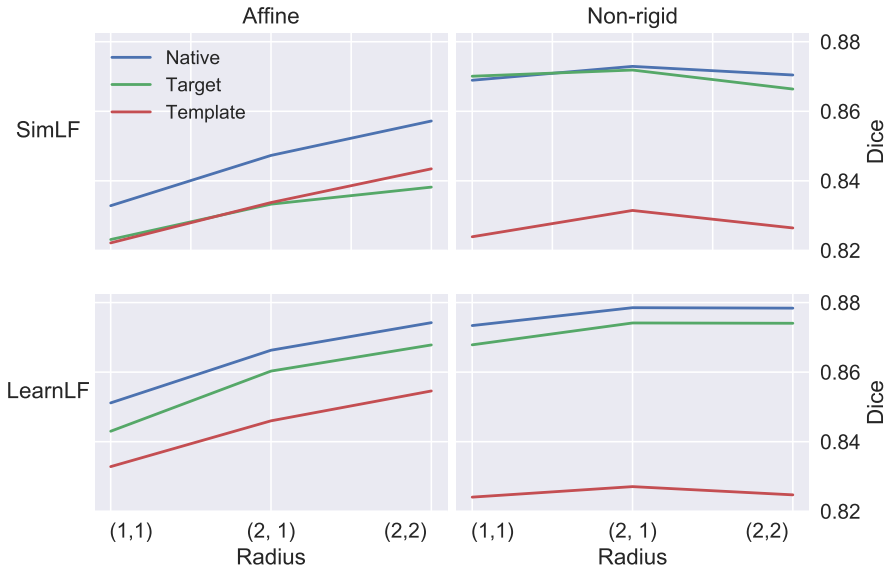

Figure 4.4: Adult dataset: Comparison of different patch spaces (colored lines), different fusion strategies and different registration settings. From top to bottom: SimLF and LearnLF fusion strategies. From left to right: affine and non-rigid registrations. Vertical axis represents performance (Dice score) and horizontal axis indicates different patch and neighborhood radii.

Dice overlap scores for all seven subcortical structures are compiled in Tables 4.1 and 4.2 for SimLF and LearnLF, respectively. Both tables report overlaps corresponding to a radius of 2 (i.e., size of $5 \times 5 \times 5$) for both patch and neighborhood search, with which all methods reached their best performance. The highest Dice overlap is highlighted in bold for each registration setting. When using affine registration, regardless of the fusion strategy, LF in the proposed native space (i.e., native patch space) outperformed the rest of approaches in all structures. The same occurred when using non-rigid registration, with the exceptions of putamen and thalamus proper in SimLF, although the difference is negligible (i.e., around 0.01).

In template patch space, both SimLF and LearnLF provided better segmentations with affine than non-rigid registration. Moreover, SimLF in target

patch space showed reduced performance than in template patch space (except for putamen), as shown in the two first columns of Table 4.1. SimLF in target patch space, however, yielded considerably better segmentations when using non-rigid registration. With LearnLF, segmentations produced in target patch space were more accurate, with both registration settings, than those in template patch space.

	Affine		Non-rigid	
	SimLF	LearnLF	SimLF	LearnLF
Template	0.843±0.040	0.855±0.031	0.826±0.036	0.825±0.037
Target	0.838±0.052	0.868±0.032	0.866±0.032	0.874±0.027
Native	**0.857±0.044**	**0.874±0.030**	**0.870±0.031**	**0.878±0.027**

Table 4.3: Adult dataset: Overall performance in terms of average Dice overlap (and standard deviation) for the different patch spaces using patch and neighborhood search radii of 2. Bold type indicates the best average segmentation performance.

Figure 4.4 shows mean overall Dice scores achieved by the tested LF strategies for different registration settings, and patch and neighborhood radii, comparing their performance depending on patch space. Overall Dice scores corresponding to a radius of 2 for both patch and neighborhood search are reported in Table 4.3. We can observe that performance in native patch space is higher than in other spaces. Moreover, for affine registration, the difference in performance when compared with LF in target space (i.e., target patch space) is the largest. Finally, it is worth noting that the LearnLF strategy achieved better performance than SimLF, which highlights the importance of the learning approach versus the similarity-based approach.

4.4.2 Fetal brain tissue segmentation

In this second experiment, we included 32 subjects from a cohort within a research project on congenital isolated ventriculomegaly. Approval was obtained for the study protocol from the Ethics Committee of the Hospital Clínic in Barcelona - Spain (HCB/2014/0484) and all patients gave written informed consent. Ages of the included subjects range between 26 to 29.3 gestational weeks. T2-weighted MR imaging was performed on a 1.5-T scanner (SIEMENS 105 MAGNETOM Aera syngo MR D13; Munich, Germany) with a 8-channel body coil. All images were acquired without sedation and following the American college of radiology guidelines for pregnancy and lactation. Half Fourier acquisition single shot turbo spin echo

	Affine			Non-rigid		
	Template	Target	Native	Template	Target	Native
BS	0.918±0.014	0.931±0.018	**0.931±0.014**	0.927±0.005	0.940±0.007	**0.945±0.005**
CB	0.941±0.010	0.948±0.014	**0.950±0.011**	0.950±0.005	0.959±0.006	**0.963±0.006**
CSF	0.929±0.007	**0.948±0.007**	0.948±0.008	0.930±0.017	**0.953±0.017**	0.950±0.016
CoGM	0.838±0.016	0.876±0.021	**0.882±0.016**	0.850±0.023	0.889±0.022	**0.903±0.022**
LV	0.867±0.067	**0.887±0.067**	0.881±0.075	0.909±0.029	**0.932±0.033**	0.929±0.031
WM	0.962±0.007	**0.969±0.007**	0.967±0.008	0.966±0.008	**0.976±0.008**	0.973±0.008

Table 4.4: Fetal dataset: Average Dice overlaps (and standard deviation) of automatic segmentations produced by SimLF using template, target and native patch spaces. Bold type indicates the best average segmentation performance.

	Affine			Non-rigid		
	Template	Target	Native	Template	Target	Native
BS	0.926±0.009	**0.941±0.010**	0.940±0.009	0.930±0.005	0.946±0.007	**0.949±0.005**
CB	0.946±0.008	**0.958±0.010**	0.957±0.009	0.952±0.005	0.964±0.006	**0.966±0.005**
CSF	0.933±0.007	**0.954±0.006**	0.954±0.007	0.932±0.016	**0.956±0.016**	0.954±0.015
CoGM	0.859±0.009	0.905±0.011	**0.905±0.010**	0.859±0.021	0.905±0.020	**0.912±0.022**
LV	0.896±0.039	0.918±0.038	**0.920±0.040**	0.915±0.026	0.941±0.024	**0.941±0.022**
WM	0.968±0.005	0.977±0.005	**0.977±0.005**	0.969±0.007	**0.979±0.007**	0.977±0.007

Table 4.5: Fetal dataset: Average Dice overlaps (and standard deviation) of automatic segmentations produced by LearnLF using template, target and native patch spaces. Bold type indicates the best average segmentation performance.

(HASTE) sequences were used with the following parameters: echo time of 82 ms, repetition time of 1500 ms, number of averaging = 1, 2.5 mm of slice thickness, 280 × 280 mm field of view and voxel size of 0.5 × 0.5 × 2.5 mm. For each subject, multiple orthogonal acquisitions were performed: 4 axial, 2 coronal and 2 sagittal stacks. Brain location and extraction from 2D slices was carried out in an automatic manner using the approach by Keraudren et al. (2014), followed by high-resolution 3D volume reconstruction using the method by Murgasova et al. (2012). Ground-truth segmentations were obtained for the following tissues and structures: cerebro-spinal fluid (CSF), cortical gray matter (CoGM), white matter (WM), lateral ventricles (LV), cerebellum (CB) and brainstem (BS). To obtain the ground-truth structures, first, 4 subjects were manually segmented by two expert raters. Then, the remaining subjects were segmented using the automatic method by Sanroma et al. (2016a) and the automatic segmentations were manually corrected by the same expert raters.

Similarly to the adult brain segmentation, Dice overlap scores for tissue segmentation of our fetal brain dataset are reported in Tables 4.4 and 4.5 for

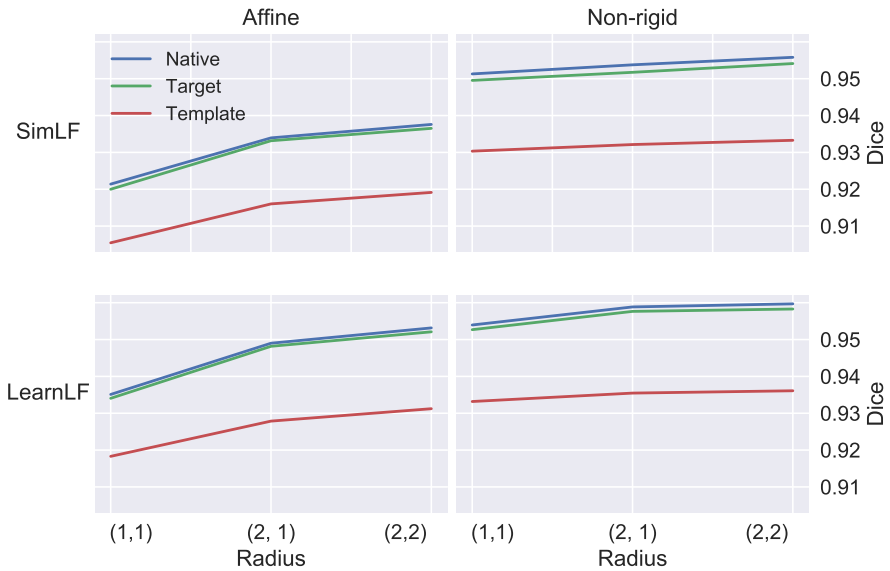

Figure 4.5: Fetal dataset: Comparison of different patch spaces (colored lines), different fusion strategies and different registration settings. From top to bottom: SimLF and LearnLF fusion strategies. From left to right: affine and non-rigid registrations. Vertical axis represents performance (Dice score) and horizontal axis indicates different patch and neighborhood radii.

SimLF and LearnLF, respectively, with patch and neighborhood search radii of 2. Overall, we can observe that LearnLF in conjunction with non-rigid registration produced the best segmentations, as shown in Table 4.5. LF in template patch space provided the lowest overlaps with both registration settings. On the other hand, with this database, the improvement of LF in native patch space over target patch space is not consistent since, for some tissues (e.g., CSF), the latter offered better segmentations. Nevertheless, no notable differences can be found between both patch spaces, with the largest Dice overlap difference being 0.007 in CoGM when using LearnLF and non-rigid registration.

Figure 4.5 shows mean Dice scores of all tissues for different patch and neighborhood radii, grouped by registration setting and fusion strategy. Overall, there is a slight improvement in performance (i.e., 0.002) of LF in native patch space over target patch space when using non-rigid registration, with LearnLF achieving the highest mean overlap, as reported in Table 4.6. Moreover, as mentioned above, LearnLF in our proposed framework is computationally cheaper than in target patch space. These results are in great concordance with the results obtained in the previous dataset, highlighting the advantages of our proposed framework.

	Affine		Non-rigid	
	SimLF	LearnLF	SimLF	LearnLF
Template	0.909±0.047	0.921±0.039	0.922±0.040	0.926±0.038
Target	0.927±0.037	0.942±0.027	0.942±0.030	0.948±0.025
Native	**0.927±0.037**	**0.942±0.026**	**0.944±0.025**	**0.950±0.022**

Table 4.6: Fetal dataset: Overall performance in terms of average Dice overlap (and standard deviation) for the different patch spaces using patch and neighborhood search radii of 2. Bold type indicates the best average segmentation performance.

4.5 Discussion

4.5.1 The impact of interpolation

LF in native patch space is not subject to interpolation errors. It is worth noting that in our proposed approach correspondences are also interpolated to valid voxel coordinates. This, however, does not involve interpolation

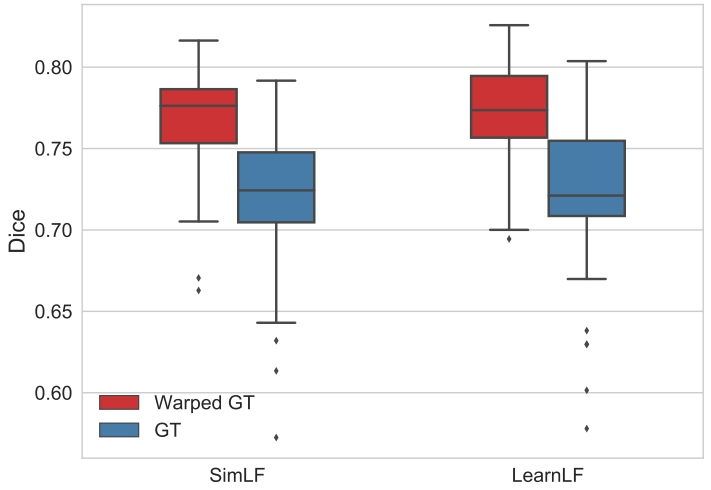

Figure 4.6: Accumbens segmentation using template patch space. Comparison of average Dice scores computed in the original space (blue) and in template space (red) using the ground truth segmentations warped to the template space.

errors related to warping the intensity images and the labelmaps. The impact of interpolation on segmentation accuracy can be illustrated using LF in template patch space. In this setting, the segmentation is estimated in template space and warped back to the target space. If we warp the ground truth of the target image to the template space, we can compute the Dice overlap in template space and compare with the overlap in target space after warping the estimated segmentation to its original space. Note that Dice overlap in template space serves no other purpose than illustrating the negative impact of interpolation, in this case, of the estimated labelmap.

Figure 4.6 shows boxplots of Dice overlap in template patch space using the warped and original ground truth segmentations. Results correspond to the segmentation of accumbens structure. The mean Dice overlap in template space (i.e., using the warped ground truth segmentations) obtained with SimLF/LearnLF is 0.767/0.771, decreasing to 0.720/0.719 after warping back the estimated labelmaps. Segmentation quality is notably degraded after warping the estimated labelmaps from the template to their original spaces, which emphasizes the negative impact of interpolation.

4.5.2 The role of patch pre-selection

Patch pre-selection in our LearnLF strategy has shown to outperform SimLF, where all patches are used. The benefits of patch pre-selection are twofold: a) faster segmentation runtimes and b) higher accuracy. First, by removing irrelevant patches, the target patch is compared with a considerably small number of training patches. In this context, Coupé et al. (2011) used a heuristic approach to discard patches from the LF based on their structural similarity with the target patch, Hao et al. (2014) selected the same number of positive and negative patches for each voxel-wise classifier, and Tong et al. (2013) summarized the training patches building a small dictionary using sparse-based representation. Nevertheless, all these approaches perform LF in target patch space, which requires patch pre-selection to be carried out online, for each target image. In our LearnLF strategy, patch pre-selection is performed offline, regardless of the target patch.

4.5.3 Computational complexity

The proposed approach can be seen as a combination of LF in both template and target space, exploiting two important aspects: a) a common template and b) atlas native spaces. First, using a common space allows registering the atlases to a template, independently of the target images. Thus, atlas registration is only performed once, as opposed to LF in target patch space. Second, since voxel-wise correspondences are established offline, computational complexity of learning-based LF strategies is considerably reduced. Only one model has to be learned for each voxel in the template, although using patches extracted from the atlases' native space. When a new target image arrives, it is registered to the template to find the spatial correspondences. Then, we iterate over the target voxels to find the corresponding voxels in the template, which contain the learned models. When using similarity-based LF strategies, the advantage of native patch space comes from evading interpolation errors, with the same computational cost as SimLF in target patch space.

4.5.4 Generalization of native patch space

In terms of performance, the focus of traditional approaches has lied, mainly, in the fusion strategy, while registration (and, indistinguishably, warping)

has been relegated to a secondary plane, with the major concern related to reducing runtimes using affine rather than non-rigid registration. In this work, by identifying the crucial building blocks of the patch-based LF approaches, we proposed an improvement in the patch-based LF framework that is transparent to most existing approaches. Since our improvement does not affect the fusion strategy, it can be easily integrated into existing (e.g., Bai et al., 2015; Wang et al., 2013) and adopted by novel patch-based LF approaches.

4.6 Conclusions

In this work, we revisited the well-known patch-based LF framework, identified the critical decisions that have notable impact on its performance, and proposed an improvement that leads to superior performance and considerably reduced runtimes in terms of registration and learning. With the exception of global LF approaches, patch-based methods require establishing voxel-wise correspondences between the training atlases and the target image. This has driven existing patch-based LF approaches to perform label fusion in target or template spaces. Because of its high performance due to more accurate registration, most patch-based LF approaches carry out fusion in the target space, which requires all training atlases to be spatially transformed to the space of each target image. Consequently, segmentation of a given target patch is performed based on training patches extracted from the atlases warped to the target space. However, although ignored by these approaches, establishing correspondences does not imply using the warped atlases, not even performing their warping. In our proposed framework, we disentangle the process of establishing correspondences from atlas warping and skip the latter. Extracting patches and their corresponding labels from the atlases' native space instead of using some deformed version after warping them, for instance, to the target space, has shown to be better. For learning-based approaches, classifiers can be learned offline using the available training atlases and reused in the segmentation of novel target images. Finally, our experiments showed that learning-based LF outperforms similarity-based LF, which reinforces the advantage of using the native patch space due to the added computational advantages that it implies for learning-based fusion strategies.

5

Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly

Abstract – Neuroimaging of brain diseases plays a crucial role in understanding brain abnormalities and early diagnosis. Of great importance is the study of brain abnormalities *in utero* and the assessment of deviations in case of maldevelopment. In this work, brain magnetic resonance images from 23 isolated non-severe ventriculomegaly (INSVM) fetuses and 25 healthy controls between 26 and 29 gestational weeks were used to identify INSVM-related cortical folding deviations from normative development. Since these alterations may reflect abnormal neurodevelopment, our working hypothesis is that markers of cortical folding can provide cues to improve the prediction of later neurodevelopmental problems in INSVM subjects. We analyzed the relationship of ventricular enlargement with cortical folding alterations in a regional basis using several curvature-based measures describing the folding of each cortical region. Statistical analysis (global and hemispheric) and sparse linear regression approaches were then used to find the cortical regions whose folding is associated with ventricular dilation. Results from both approaches were in great accordance, showing a significant cortical folding decrease in the insula, posterior part of the temporal lobe and occipital lobe. Moreover, compared to the global analysis, stronger ipsilateral associations of ventricular enlargement with reduced cortical folding were encountered by the hemispheric analysis. Our findings confirm and extend previous studies by identifying various cortical regions and emphasizing ipsilateral effects of ventricular enlargement in altered folding. This suggests that INSVM is an indicator of altered cortical development, and moreover, cortical regions with reduced folding constitute potential prognostic biomarkers to be used in follow-up studies to decipher the outcome of INSVM fetuses.

This chapter is adapted from:
Benkarim, O. M., Hahner, N., Piella, G., Gratacos, E., González Ballester, M. A., Eixarch, E., and Sanroma, G. (2018). Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly. *NeuroImage: Clinical*, 18:103-114. https://doi.org/10.1016/j.nicl.2018.01.006

5.1 Introduction

Cortical folding is a major developmental process the human brain embarks on during the intrauterine period to acquire its highly gyrencephalic adult morphology. In early gestation, the cortex is a smooth sheet that becomes intensively convoluted following an ordered sequence of sulcogyral formation, with primary and secondary sulci obeying stable spatio-temporal patterns, while more irregular patterns govern the emergence of tertiary sulci. These cortical convolutions are intrinsically related to the functional organization of the cortex. Consequently, alterations in the degree and pattern of cortical folding might have a profound impact on brain function (Fernández et al., 2016). In adults, several studies have revealed associations of altered folding with functional disabilities in a wide spectrum of disorders such as schizophrenia (Jou et al., 2005) and attention-deficit/hyperactivity disorder (Wolosin et al., 2009). These functional disturbances might involve early cortical folding malformations and manifest in adulthood as symptomatic consequences of said maldevelopment (Batty et al., 2015; Powell, 2010; Rehn and Rees, 2005; Wolosin et al., 2009).

Since gyrification commences early in pregnancy, gestation constitutes a vulnerable period for cortical folding, where prenatal diagnosis of cerebral abnormalities is of paramount importance. In the fetus, ventriculomegaly (VM) is the most frequent abnormal finding in prenatal ultrasound examination and occurs in around 1% of fetuses (Huisman et al., 2012; Salomon et al., 2007). Fetal VM is a condition in which the lateral ventricles are dilated, and is defined as an atrial diameter of $\geq$10 mm of one or both lateral ventricles at any gestational age (GA) from 14 weeks onwards (Cardoza et al., 1988), being 6-8 mm the width in normal fetuses. These measurements remain stable in the second and third trimesters (ISUOG Guidelines, 2007). In case of ventricular enlargement, an atrial diameter in the range of 10-15 mm constitutes non-severe VM, whilst a measurement larger than 15 mm refers to severe VM. Non-severe VM is further classified into mild (10-12 mm) and moderate (12-15 mm). In case of no other anomalies, it is called isolated VM.

Though studies have found associations of ventricular enlargement with attention-deficit/hyperactivity disorder (Lyoo et al., 1996) and schizophrenia (Vita et al., 2000; Wright et al., 2000), the implications of fetal VM in such disorders remain largely unclear due to scarce long-term follow-up studies and the appearance of confounding factors during development. Never-

theless, isolated non-severe ventriculomegaly (INSVM)-associated neurodevelopmental deficits have been observed in neonates and infants (Gómez-Arriaga et al., 2012; Leitner et al., 2009; Sadan et al., 2007). When VM is diagnosed, postnatal prognosis is highly dependent on the presence of other abnormalities and the degree of ventricular dilation (Griffiths et al., 2010). There is a high risk of poor neurodevelopmental outcome when other abnormalities are diagnosed and/or the ventricles are severely dilated. However, INSVM fetuses are not so prone to have neurodevelopmental problems (Griffiths et al., 2010; Melchiorre et al., 2009), and the ones that will have unfavorable outcome cannot be characterized solely by the atrial diameter (Beeghly et al., 2010). With altered cortical folding found in fetuses with INSVM (Scott et al., 2013), the assessment of cortical folding can play an important role in prognosis reliability (Li et al., 2011).

Although ultrasound is the most used imaging modality for evaluating pregnancies, VM is a common indication for fetal magnetic resonance imaging (MRI) (Rutherford, 2001). Indeed, MRI of the *in vivo* fetal brain has recently attracted increasing attention from the neuroscientific community and is becoming an important tool in the study of *in utero* brain development (Benkarim et al., 2017c; Studholme and Rousseau, 2014). There are several works in the literature that use 3D MRI to investigate the intrauterine cerebral growth in healthy populations (Clouchoux et al., 2012; Habas et al., 2012; Wright et al., 2014, 2015). These neuroimaging studies attempt to identify and set the normative morphological and functional changes the fetal brain undergoes during its maturational course. On the other hand, neuroimaging of diseased brains provides the means to find disease-specific deviations from the aforementioned normative development and the discovery of stable biomarkers that accurately discriminate such diseases. Using 3D reconstructed MRI, isolated VM has been previously studied in (Kyriakopoulou et al., 2014; Scott et al., 2013). Scott et al. (2013) analyzed volumetric and cortical folding differences between 16 cases and 16 controls in the age range of 22-25.5 gestational week (GW). Volumetric analysis was carried out by Kyriakopoulou et al. (2014) in 60 controls and 65 cases within the GA range of 22-38 weeks. Among their findings, Kyriakopoulou et al. (2014) showed increased cortical volume in fetuses with VM, and Scott et al. (2013) found reduced cortical folding in both hemispheres, although in a narrow area near the parieto-occipital sulcus.

In this work, 3D reconstructed fetal brain MR images were used to investigate the relationship of INSVM with alterations in gyrification between a cohort of 25 healthy controls and 23 INSVM fetuses within the age range

of 26-29 GWs. There are no studies that investigated cortical folding under VM in the third trimester of gestation. During this period, numerous cortical landmarks are prominently developed in the normal fetal brain (e.g., superior temporal sulcus and calcarine fissure) (Clouchoux et al., 2012), which allow susceptible deviations in gyrification to be reliably detected. Cortical folding was quantified using several curvature-based folding measures (e.g., mean curvature, shape index and curvedness). These descriptors offer a different perspective into intrauterine neurodevelopment than brain volumetry. Curvedness, for instance, was shown to provide different information and be more accurate in the prediction of GA than brain volume (Hu et al., 2013; Wu et al., 2015b). With several descriptors we can, furthermore, capture different shape characteristics of the cortex. Positive and negative versions of some folding measures (e.g., positive and negative mean curvature) were further incorporated to respectively account for folding confined in gyral and sulcal areas, which affords a separate inspection of cortical folding. The cortex was parcellated in several regions to study cortical folding differences in a regional basis. Statistical analysis and sparse regression approaches were adopted to analyze folding differences related to ventricular enlargement and identify cortical regions with altered folding. The present study seeks to add to previous studies by providing insights into the gyrification alterations potentially associated with INSVM at mid-third trimester of gestation (where the majority of primary sulci are formed), assessing the relationship from different methodological approaches, and characterizing the implication of ventricular enlargement laterality in cortical alterations.

5.2 Materials and methods

5.2.1 Subjects

For our study, we included 25 healthy controls and 23 subjects diagnosed with INSVM from a larger prospective cohort of 81 subjects within a research project on congenital isolated VM. INSVM was defined as unilateral or bilateral ventricular width between 10-14.9 mm. All fetuses were from singleton pregnancies and met the inclusion criteria of having no abnormal karyotype, infections or malformations with risk of abnormal neurodevelopment. Approval was obtained for the study protocol from the Ethics Committee of the Hospital Clínic in Barcelona - Spain (HCB/2014/0484) and all patients gave written informed consent. Fetal MRI was performed

between 26-29 GW. Pregnancies were dated according to the first-trimester crown-rump length measurements (Robinson and Fleming, 1975). Table 5.1 presents the number of subjects and mean GA per cohort, with INSVM cases grouped by left, right or bilateral ventricular enlargement.

	N	GW	M/F
Control	25	27.6±0.9	14/11
INSVM			
Bilateral	5	27.3±0.9	4/1
Left	8	28.1±0.8	8/0
Right	10	27.2±1.0	9/1
Total	23	27.5±1.0	21/2

Table 5.1: Demographics. Number of subjects (N), mean GA and standard deviation expressed in GW, and gender (M/F, where M and F stand for male and female, respectively) per cohort. The INSVM cohort is further divided in 3 subgroups (left, right and bilateral) according to unilateral or bilateral VM diagnosis.

5.2.2 MRI acquisition and reconstruction

T2-weighted MR imaging was performed on a 1.5-T scanner (SIEMENS MAGNETOM Aera syngo MR D13; Munich, Germany) with a 8-channel body coil. All images were acquired without sedation and following the American college of radiology guidelines for pregnancy and lactation. Half Fourier acquisition single shot turbo spin echo (HASTE) sequences were used with the following parameters: echo time of 82 ms, repetition time of 1500 ms, number of averaging = 1, 2.5 mm of slice thickness, 280×280 mm field of view and voxel size of $0.5 \times 0.5 \times 2.5$ mm^3. For each subject, multiple orthogonal acquisitions were performed: 4 axial, 2 coronal and 2 sagittal stacks. Final 3D motion-corrected reconstructions were obtained from these 8 stacks of thick 2D slices. Brain location and extraction from 2D slices was carried out in an automatic manner using the approach proposed by Keraudren et al. (2014), followed by high-resolution 3D volume reconstruction using the method presented in (Murgasova et al., 2012). Figure 5.1 shows an example of 2 raw acquisitions and the final reconstructed volume.

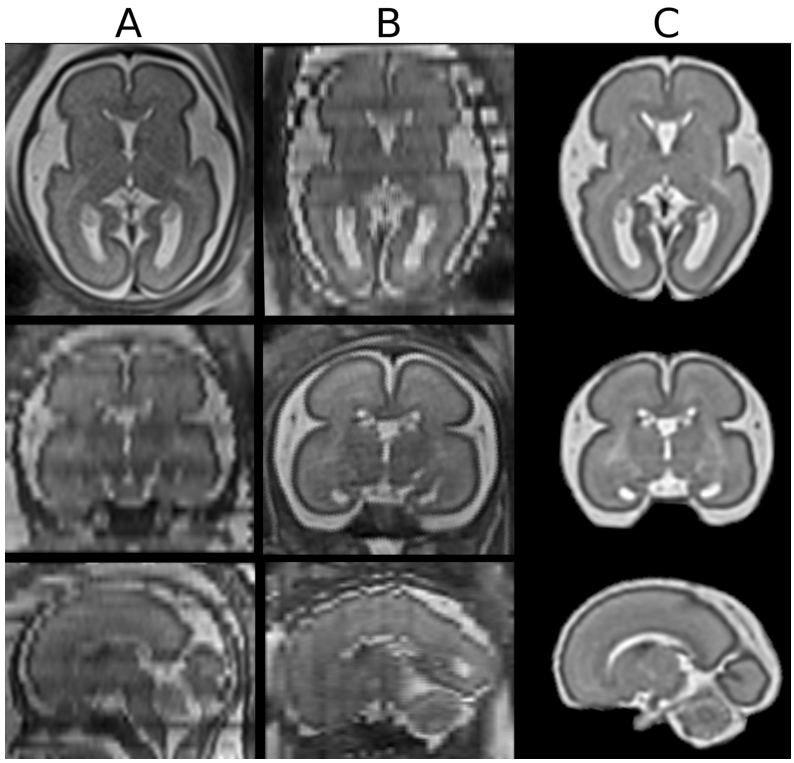

Figure 5.1: Brain MRI of a 26 GW-old healthy control reconstructed from 8 stacks of 2.5 mm slice thickness. From top to bottom: axial, coronal and sagittal views of axial (A) and coronal (B) raw stacks, and final reconstruction (C).

5.2.3 Tissue segmentation

Reconstructed MR images were segmented with the method proposed in (Sanroma et al., 2016a) into white matter (WM), cortex, cerebrospinal fluid (CSF), ventricles, cerebellum and brainstem. Briefly, this is an ensemble method that learns the optimal spatial combination of a set of base methods. It is based on the hypothesis that different segmentation methods complement each other in different regions of the brain. As base segmentation methods we used joint label fusion (Wang et al., 2013) and *Atropos* (Avants et al., 2011), which were then spatially combined as proposed by Sanroma et al. (2016a). Segmentations for 4 subjects, manually delineated by an expert, were used as atlases for the segmentation method. Figure 5.2 displays an example of a reconstructed brain MRI and its corresponding automatic labeling.

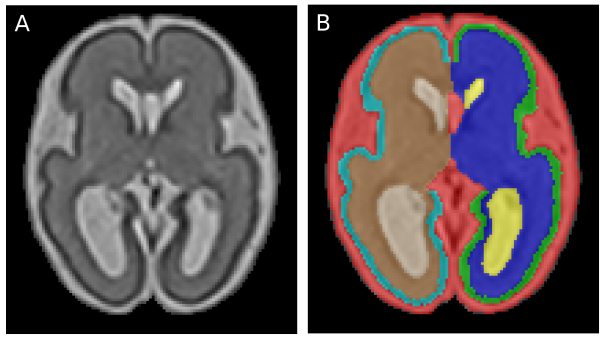

Figure 5.2: Fetal brain MRI segmentation: 26.4 GW-old fetus with right INSVM (A) and corresponding segmentation (B), with different labels for left and right WM, cortex and lateral ventricles.

Cortical region	Part	Abbreviation
Anterior temporal lobe	lateral	ATLlp
	medial	ATLmp
Cingulate gyrus	anterior	CGap
	posterior	CGpp
Frontal lobe		FL
Gyri parahippocampalis et ambiens	anterior	GPAap
	posterior	GPApp
Insula		Ins
Lateral occipitotemporal gyrus, gyrus fusiformis	anterior	LOGFap
	posterior	LOGFpp
Medial and inferior temporal gyri	anterior	MITGap
	posterior	MITGpp
Occipital lobe		OL
Parietal lobe		PL
Superior temporal gyrus	middle	STGmp
	posterior	STGpp

Table 5.2: Cortical regions from the neonatal atlas provided by Makropoulos et al. (2014) and their abbreviations. Each hemisphere is parcellated into these 16 cortical regions.

5.2.4 Cortical surface extraction

To study gyrification, the inner cortical surface was chosen instead of the outer surface since the interface between WM and cortex is more stable and less prone to segmentation errors due to partial volume effects than the cortex-CSF interface. The WM binary masks were smoothed using a 2 mm full width at half-maximum Gaussian kernel and cortical surface meshes for each hemisphere were then reconstructed with the marching cubes algorithm (Lorensen and Cline, 1987).

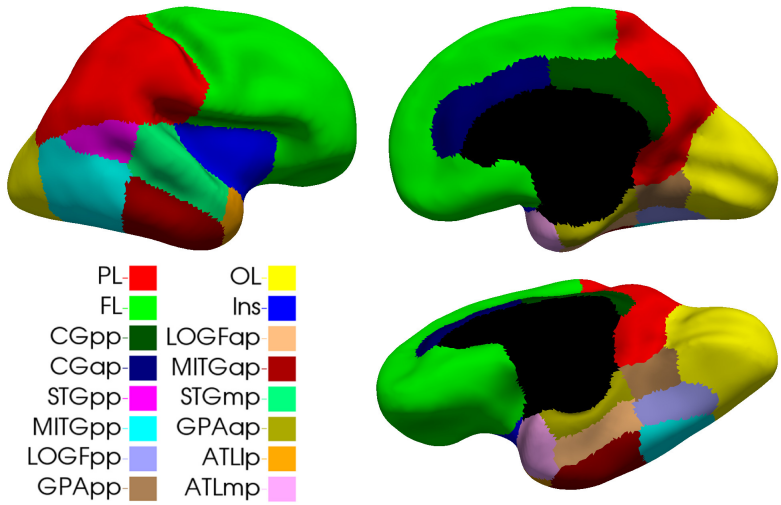

Figure 5.3: Inner cortical surface parcellation.

Furthermore, a neonatal atlas with cortical regions (Makropoulos et al., 2014) was used to parcellate each hemisphere into 16 cortical regions. For each fetal brain, a neonatal atlas with a similar age was registered to it using the symmetric diffeomorphic mapping (SyN) proposed by (Avants et al., 2008). Registration was performed based on intensity images and cortical masks extracted from both the fetal brain (based on the initial tissue segmentation) and the neonatal atlas (as a union of all cortical labels based on the segmentations provided by (Makropoulos et al., 2014)). In this way, registration is driven to provide a more accurate parcellation of the cortex. Parcellations were then visually inspected and corrected for possible mislabeling. Table 5.2 lists the different cortical regions and Figure 5.3 shows them propagated onto the surface mesh.

5.2.5 Regional cortical folding

Gyrification changes related to ventricular enlargement were investigated at both hemispheric and regional levels using a curvature-based approach. This approach has been previously used to study cortical folding in premature neonates (Rodriguez-Carranza et al., 2008; Shimony et al., 2016) and healthy fetuses (Wright et al., 2014; Wu et al., 2015b). For each fetus, principal curvatures (i.e., k_1 and k_2) were computed at each vertex on the inner cortical surface mesh and 15 curvature-based folding measures were derived as described in (Rodriguez-Carranza et al., 2008; Shimony et al., 2016; Wright et al., 2014). Several folding measures were used in this work since a single curvature descriptor (e.g., mean curvature) may not be sufficient to fully characterize cortical folding. These measures are summarized in Table 5.3.

For each vertex on a surface mesh, 11 local curvature descriptors (see first 11 entries in Table 5.3: from MC to NSI) were computed from the principal curvatures, as shown in the third column. Then, folding for a given cortical region was characterized by a weighted average of all points belonging to such region (Eq. (5.1) in Table 5.3) for each curvature descriptor. These descriptors included MC, which is an extrinsic property of a surface created by distance preserving folding (without distortion), with positive MC values located on gyri and negative values on sulci (Schaer et al., 2008). For a given cortical region, MC denotes the overall folding of the region, while PMC only focuses on convexity (i.e., positive MC) and NMC on concavity (i.e., negative MC). On the other hand, GC is an intrinsic property of the surface that captures curvature changes created by distortion (Pienaar et al., 2008). A cortical region is ellipsoidal when the average GC is positive and saddle-shaped when negative (Van Essen and Drury, 1997). PGC and NGC summarize surface points with positive and negative GC, respectively. While jointly using the signs of MC and GC allows identifying different shapes, the SI proposed by Koenderink and van Doorn (1992) is able to capture continuous surface shapes ranging from concavity (i.e., -1) to convexity (i.e., +1) (Hu et al., 2013). PSI and NSI summarize region points located on gyri and sulci, respectively. Complementary to SI, Koenderink and van Doorn (1992) also proposed the CI that quantifies size by the degree of deviation of a surface from a flat plane. We also consider FI, defined as a measure of cylindrical parts of a surface (Van Essen and Drury, 1997). Four more folding measures were additionally incorporated into the analysis of gyrification (see last 4 entries in Table 5.3). MCL2 and

	Abbreviation	Definition
Mean curvature	MC	$H = \frac{k_1+k_2}{2}$
Positive mean curvature	PMC	H^+
Negative mean curvature	NMC	H^-
Gaussian curvature	GC	$K = k_1 k_2$
Positive Gaussian curvature	PGC	K^+
Negative Gaussian curvature	NGC	K^-
Curvedness index	CI	$C_i = \sqrt{\frac{k_1^2+k_2^2}{2}}$
Folding index	FI	$F_i = \lvert k_1 \rvert \left(\lvert k_1 \rvert - \lvert k_2 \rvert \right)$
Shape index	SI	$S_i = -\frac{2}{\pi} \arctan\left(\frac{k_2+k_1}{k_2-k_1}\right)$
Positive shape index	PSI	S_i^+
Negative shape index	NSI	S_i^-
		$\sum_i *w / \sum_i w$ (5.1)
Mean curvature L_2 norm	MCL2	$\sqrt{\sum_i H^2 w / \sum_i w}$
Gaussian curvature L_2 norm	GCL2	$\sqrt[4]{\sum_i K^2 w / \sum_i w}$
Mean curvature norm ratio	MCNR	$\sum_i H^2 w / \sum_i \lvert H \rvert w$
Gaussian curvature norm ratio	GCNR	$\sum_i K^2 w / \sum_i \lvert K \rvert w$

Table 5.3: Curvature-based folding measures used in our study. The $*$ symbol in Eq. (5.1) can be substituted by any of the above curvature measures to compute the corresponding folding measure for a given cortical region. Note that superscripts $^+$ and $^-$ denote positive and negative curvature, respectively. For example, PMC is 0 in all points where $H < 0$ and H otherwise. Conversely, NMC is 0 when $H > 0$ and H otherwise. The same applies for all remaining curvatures with $^+$ and $^-$ superscripts.

GCL2 are invariant to translation, rotation and scaling, and measure bending of the surface and the amount of surface with constant Gauss curvature, respectively (Batchelor et al., 2002). MCNR and GCNR are further area independent and quantify gyrification as a ratio of curvatures (Rodriguez-Carranza et al., 2008).

The contribution of a point in the surface mesh to regional folding measures was weighted by the mean area of all cells sharing this point (w in Table 5.3). For all cortical folding measures used in our analyses, the higher their values, the more folded is the cortical region, with the exceptions of the negative measures (i.e., NMC, NGC and NSI).

5.2.6 Statistical analysis

The aim of our study is to analyze the effect of INSVM in gyrification. Thus, to test the effect of ventricular enlargement in cortical folding, we used a general linear model (GLM) with 2 versions to characterize the changes in the aforementioned folding measures. The first model, F_D, was used to test the effect of diagnosis, as a group factor (0 for control and 1 for INSVM), and GA and supratentorial volume as covariates:

$$F_D = \beta_0 + \beta_1 GW + \beta_2 STV + \beta_3 DG, \tag{5.2}$$

where GW is GA in weeks, DG refers to the diagnosis, and STV represents supratentorial volume. The second model, F_V, uses ventricular volume as a covariate instead of diagnosis:

$$F_V = \beta_0 + \beta_1 GW + \beta_2 STV + \beta_3 VV, \tag{5.3}$$

where VV represents ventricular volume. Since the *in utero* brain becomes continuously more convoluted as growth proceeds, GA was incorporated into both models to control for the effect of age on the analyzed folding measures. Furthermore, curvature-based folding measures depend on brain size (Rodriguez-Carranza et al., 2008) and supratentorial volume (composed of WM, gray matter (GM) and ventricles) was considered as a covariate in both models to account for scaling differences (Shimony et al., 2016).

In the second model, F_V, considering ventricular volume rather than diagnosis can improve the accuracy of the model. First, because ventricular volume was found to be more distinctive in the analysis of VM than atrial diameter (Gholipour et al., 2012), and therefore more properly representative of such condition. Second, and most importantly, in fetuses with VM, ventricular dilation is not restricted to the atria (Scott et al., 2013). Instead of a dichotomous variable carrying information solely about the width of the atrium (in fact, diagnosis only tells us if the atrial width surpasses the threshold of 10 mm or not), ventricular volume is able to capture the extent of enlargement.

Besides examining folding measures at a regional level with the 2 proposed models, we further analyzed the association of lateral ventricular enlargement (using Eq. (5.3)) per hemisphere instead of the association of total ventricular volume with the whole cortex.

For each cortical region, 15 folding measures were extracted and tested. Results were then considered statistically significant after Bonferroni correction at $p < 0.05/15 \approx 0.0033$.

5.2.7 Sparse linear models for ventricular volume prediction

Regression models with sparsity-inducing regularization, such as the well-known Lasso (Tibshirani, 1994), are widely used in neuroimaging data analysis for diagnosis prediction and prognostic biomarker discovery (Klöppel et al., 2012; Mwangi et al., 2014; Shimizu et al., 2015). Furthermore, these approaches are well-suited for high-dimensional data, especially when the number of predictors (i.e., features) greatly exceeds the number of training samples (i.e., subjects). Our purpose is to find the most relevant cortical regions (each region is composed of 15 curvature features) in the prediction of total ventricular volume using sparse regression models. We use Lasso to enforce that the regression model will be sparse and only a portion of the predictor variables will be used in the model. This allows finding the most relevant features for the prediction of ventricular volume. By comparing these results with the ones obtained with GLM, we can draw conclusions about the cortical regions that are related to ventricular enlargement.

In our setting, each subject is represented by a 482-dimensional feature vector with 15 folding measures extracted from each of the 32 cortical regions and 2 additional features corresponding to GA and STV. Formally, let $Y \in \mathbb{R}^n$ denote a vector of total ventricular volume for all n subjects and an $n \times p$ design matrix X with the p folding measures from each subject arranged in rows. The Lasso estimator is defined as:

$$\min_{\beta} \left(||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{5.4}$$

where $\beta \in \mathbb{R}^p$ is a coefficient vector, $||\cdot||_2^2$ is the squared Euclidean norm, λ is the penalization parameter and the second term corresponds to the l_1-norm. The first term aims at fitting the data (as in conventional regression) and the second term is the regularization for improving generalization, controlled by the λ parameter. A larger λ will force most of the features to be 0 (i.e., not selected).

Lasso is designed for selecting individual variables. In this way, we can indirectly conclude that regions with more selected features are more related to INSVM. However, our purpose is to inspect relevance at region level rather than at measure or variable level. In order to directly extract conclusions about the important regions, we can enforce Lasso to select all or none of the features from a given region. Group Lasso is an extension of Lasso that incorporates such constraint by performing selection on predefined groups of variables rather than individual variables (Yuan and Lin, 2006). Thus, all 15 folding measures corresponding to a given cortical region are considered as a group in Group Lasso regression. GA and STV are also considered separately as single feature groups. The Group Lasso estimator is defined as:

$$\min_{\beta} \left(||Y - X\beta||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}\, ||\beta_l||_2 \right), \tag{5.5}$$

where L is the number of groups (i.e., 34 groups: 32 cortical regions, one group for GA and another for STV), p_l is the number of variables in the l-th group and $||\cdot||_2$ is the Euclidean norm. When L is equal to the number of variables p, this optimization problem is equivalent to Lasso (Friedman et al., 2010).

The optimal λ values for Lasso and Group Lasso based on leave-one-out cross-validation are used to identify the most relevant cortical regions. In terms of performance in predicting the ventricular volume, it is expected that Lasso would outperform Group Lasso since Group Lasso is more constrained than Lasso in that it is forced to include or drop all measures representing a particular cortical region.

One problem with Lasso and Group Lasso is that selected (i.e., relevant) features can vary depending on the data used to compute the model. To that end, Randomized Lasso (Meinshausen and Bühlmann, 2010) determines the most stable features by randomly selecting the data used to compute the models multiple times. The Randomized Lasso estimator is defined as:

$$\min_{\beta} \left(||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{W_j} \right), \tag{5.6}$$

where W_j is an independent and identically distributed random variable in $[\alpha, 1]$ and $\alpha \in (0, 1]$ is the weakness parameter that is used to change

the value of λ in Eq. (5.6) to be randomly chosen in the range $[\lambda, \lambda/\alpha]$. For a more detailed description, the reader is referred to the original work of Meinshausen and Bühlmann (2010).

5.3 Results

5.3.1 Ventricular volume

Total ventricular volumes for control and INSVM fetuses are shown in Figure 5.4. Average total ventricular volumes were lower in healthy controls (3.65 ± 1.12 cm^3) than in INSVM fetuses (8.91 ± 2.70 cm^3). We used a GLM to analyze the association of ventricular volume with GA and diagnosis. There was a slight increase of 0.76 cm^3 ($p = 0.01$) in ventricular volume with increasing GA. In association with diagnosis, ventricular volume was on average 5.34 cm^3 larger ($p < 10^{-20}$) in the INSVM cohort than in healthy controls, which is obvious since INSVM was diagnosed based on ventricular enlargement.

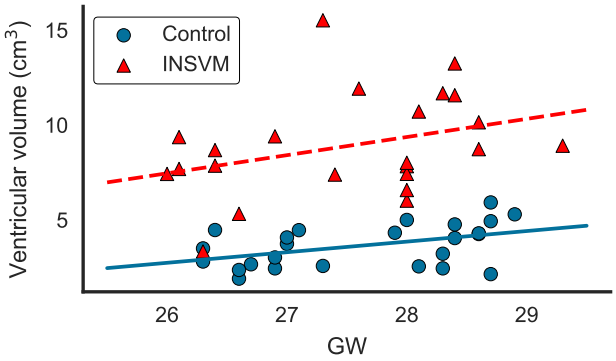

Figure 5.4: Total ventricular volume with increasing GA.

5.3.2 Statistical analysis

Two types of analysis were conducted to assess the relationship between ventricular enlargement and cortical folding. First, a global analysis, using both F_D and F_V, was performed to investigate the association of INSVM with cortical folding. Second, a hemispheric analysis was conducted, using the F_V model, where the effect of ventricular dilation was studied per

hemisphere. In this last analysis, the relationship between ventricular dilation and cortical folding was studied per hemisphere. Therefore, the VV covariate represents the volume of the corresponding lateral ventricle.

Table 5.4 lists the cortical regions where significant differences in cortical folding related to INSVM diagnosis (F_D) or total ventricular enlargement (F_V) were found in at least one of the folding measures analyzed. We found that ventricular volume in Eq. (5.3) was more sensitive than diagnosis as shown in Table 5.4, reaching significance in more folding measures and cortical regions (i.e., occipital and parietal lobes). Moreover, when analyzing ipsilateral associations, ventricular volume showed a stronger effect on the cortical folding in the same hemisphere (see Table 5.5) when compared to the global model using total ventricular volume. In general, the cortical regions most associated with VM were insula, occipital and parietal lobes, and the regions located in the posterior part of the temporal lobe.

As shown in the last column of Table 5.4, whole hemispheres were included as single regions to check for global effects. In the whole hemisphere, diagnosis was not associated with any folding measure, while significantly lower values of PMC and MCL2 (all $p = 0.002$) associated with ventricular enlargement were found in the right hemisphere, and bilaterally higher values of NGC (left: $p = 0.002$, right: $p = 0.001$). At a regional basis, significant differences in PMC (left: $p < 10^{-4}$, right: $p < 10^{-5}$) and CI (left: $p < 0.0002$, right $p < 10^{-5}$), among others, were found bilaterally with both F_D and F_V models in GPApp and MITGpp, with diagnosis and ventricular dilation associated with a decreased cortical folding in all measures reaching significance. In the insula, significant effects of diagnosis were found in NMC ($p < 0.002$) and NGC ($p < 0.001$) solely in the right hemisphere. Ventricular volume was, however, associated with several folding measures in the right hemisphere and bilaterally with NGC (left: $p < 0.002$, right: $p < 10^{-5}$) and FI (both $p < 0.001$). Significant changes were also found by both F_D and F_V in GPAap and LOGFpp, although no curvature measure showed bilateral differences, with the exceptions of GPAap, where MC (both $p < 0.001$) and PMC (both $p < 0.001$) were bilaterally associated with ventricular enlargement. In association with abnormality, GPAap showed increased curvature. In LOGFpp, we found variable effects depending on the hemisphere and folding measure. Reduced cortical folding in the left SGTpp (i.e., increased NMC and NGC) was only found by the F_D. The second model (F_V), however, seemed to be more sensitive since significant folding differences in the occipital and parietal lobes were not detected in the first model (i.e., using diagnosis as factor, F_D). Increasing ventricular

	GPAap	GPApp	Ins	LOGFpp	MITGpp	OL	PL	STGpp	Hem
MC	R/B	B/B	-/R	-/R	R/R				
PMC	R/B	B/B			B/B	-/B	-/B		-/R
NMC			R/R	L/-	-/R	-/B		L/-	
GC	-/L			L/L		-/B			
PGC		B/B	-/R	L/-	L/B	-/B	-/L		
NGC		R/R	R/B		L/B	-/B	-/L	L/-	-/B
CI		B/B	-/R		B/B	-/B	-/L		
FI	-/L	B/B	-/B		B/B				
SI	-/L	-/R		R/R					
PSI	-/L	L/B		R/R					
NSI				L/R					
MCL2	-/L	B/B	-/R	L/-	B/B	-/B	-/L		-/R
GCL2		L/B		L/-	L/L	-/R			
MCNR		B/B		L/-	L/-	-/R			
GCNR		L/B				-/R			

Table 5.4: Regions with statistically significant differences in cortical folding associated with diagnosis/ventricular enlargement. Abbreviations: left hemisphere (L), right hemisphere (R), bilateral (B), and hemisphere (Hem). The term before the forward slash indicates statistical significance found using the F_D model, while the one after the slash indicates significance when using F_V (i.e., F_D/F_V). Abbreviations L, R, and B represent that significance is found only in the left, right hemisphere, and bilaterally, respectively. When there is no statistical significance for one model, '-' is used. For example, the entry corresponding to GPAap and MC shows that the effect of diagnosis (i.e. F_D model) in the MC folding measure of the right GPAap is statistically significant, whereas it is not in the left GPAap. However, the effect of ventricular dilation (i.e., F_V) in both right and left GPAap is statistically significant.

	GPAap	GPApp	Ins	LOGFpp	MITGpp	OL	PL	STGpp	Hem
MC	B	B	B	R	B				
PMC	R	B			B	B	B		B
NMC	L		B	L	R	B	L		L
GC				L		B			
PGC		B	B	L	B	B	L		B
NGC		B	B		B	B	L		B
CI	R	B	B	L	B	B	B	R	B
FI		B	L		B	R			
SI	L	B		R					
PSI	L	B		R					
NSI	L	L		R		L			
MCL2		B	B	L	B	B	L		B
GCL2		B		L	R	R			
MCNR		B		L	R	B			
GCNR		B		L		R			

Table 5.5: Regions with statistically significant differences in cortical folding associated with hemispheric ventricular enlargement. Abbreviations: left hemisphere (L), right hemisphere (R), bilateral (B), and hemisphere (Hem). Abbreviations R/L indicate that statistical significance is achieved only in the region of right/left hemisphere by the F_V model that only uses the right/left ventricular volume. When statistical significance is found by both hemispheric models on their respective sides, B is used.

volume was significantly associated with bilateral gyrification decrease of the occipital lobe in multiple folding measures (e.g., NMC left: $p < 0.0002$, right: $p < 10^{-4}$). In the parietal lobe, ventricular volume was associated with reduced curvature in the left hemisphere and bilaterally only with PMC (left: $p < 0.0001$, right: $p = 0.002$).

Note that in this first analysis (using both F_D and F_V models), we assumed that diagnosis and ventricular enlargement might be associated with gyrification differences over the whole cortical plate, irrespective of its side. That is, the diagnosis used as group factor in F_D and the total ventricular volume in F_V could not distinguish the laterality. However, since there are cases diagnosed with left, right or bilateral INSVM in our abnormal cohort, we further conducted a second analysis to assess the association of ventricular dilation with reduced cortical folding for each hemisphere separately.

Results from this second hemispheric analysis are shown in Table 5.5. Cortical regions and folding measures statistically associated with ventricular volume show great overlap with the previous analysis using total ventricular volume (i.e., F_V), though it is important to emphasize the differences concerning the insula and whole hemisphere. In this second analysis, lateral ventricular dilation had a significant effect on reduced folding in both insula and whole hemisphere in the majority of curvature measures. Reductions in folding related bilaterally to NMC, PGC, NGC, CI and MCL2 were found in the insula, with increased curvature only observed in MC. Regarding the whole hemispheres, lateral ventricular volume was consistently correlated with reduced folding in all statistically significant curvature measures, with NGC rendering the lowest p-values (left: $p < 10^{-4}$ and right: $p = 0.0002$). Some of these measures are shown per hemisphere in Figure 5.5. This suggests that albeit cortical folding might be bilaterally altered in the presence of unilateral ventricular enlargement, ipsilateral association is higher.

5.3.3 Ventricular volume prediction

Sparse linear models were used to predict total ventricular volume and identify the most relevant cortical regions and curvature-based features related to ventricular enlargement. Before regression, all features were scaled to a mean of 0 and a standard deviation of 1. Average mean absolute errors (MAE) using a leave-one-out cross-validation strategy for each value of λ in the range $[10^1, 10^{-4}]$ are shown in Figure 5.6. The goal of this experiment is to choose the optimal value for the regularization parameter λ. This

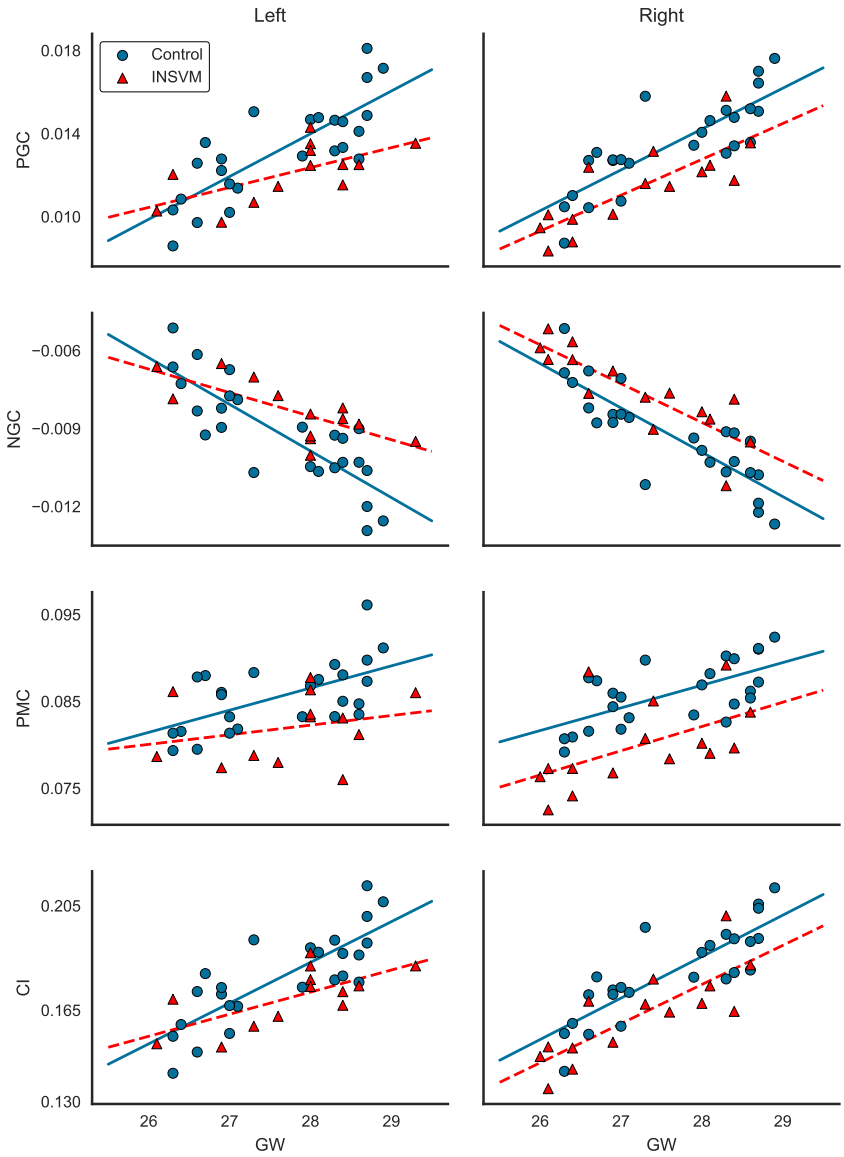

Figure 5.5: Scatter plots of PGC, NGC, PMC and CI versus GA for each hemisphere, with linear fits for controls and subjects with left or right INSVM, depending on the hemisphere. Note that fetuses with bilateral ventricular enlargement appear as INSVM in both hemispheres.

is, the optimal amount of regularization (i.e., sparsity) is chosen to be the one that leads to the lowest ventricular volume prediction error. We can see that Lasso yields more accurate volume prediction than Group Lasso.

Lasso is more free to choose the most relevant features from any cortical region, while Group Lasso is constrained to select groups of features (i.e., a group of features corresponds to a cortical region), except for GA and STV, which are single feature groups.

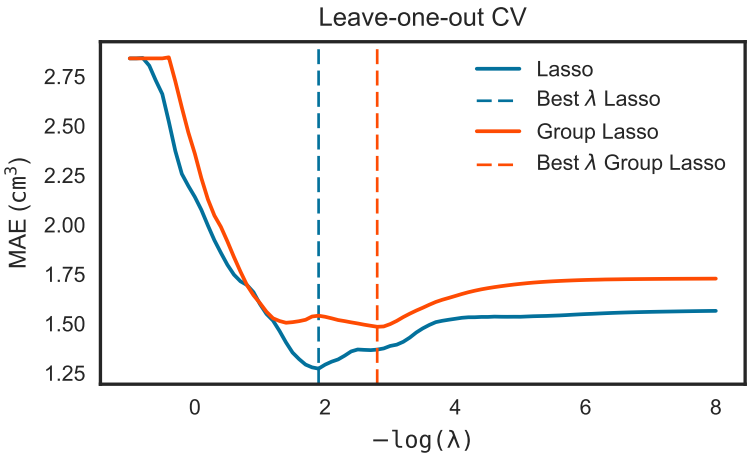

Figure 5.6: Leave-one-out cross-validated mean absolute error evolution using different regularization values. Comparison of Lasso with Group Lasso.

Figure 5.7 shows the MAE for the optimal λ values corresponding to Lasso and Group Lasso, respectively, separated by cohort. For all fetuses, the MAE obtained is 1.27 cm^3 (control: 0.98, INSVM: 1.59) and 1.48 cm^3 (control: 1.20, INSVM: 1.79) for Lasso and Group Lasso, respectively. With 3.65 ± 1.12 cm^3 and 8.91 ± 2.70 cm^3 being the ventricular volumes in normal and INSVM cohorts, respectively. Note that both methods provided higher mean error in the INSVM cohort than in normal fetuses. This may be due to the large volume variance and heterogeneity (see Figure 5.4) existing in the abnormal cohort (including left, right and bilateral INSVM).

The advantage of using sparse regularization is that most feature coefficients will be driven towards zero and only few relevant features will be selected. However, there may be some instability in the features selected by Lasso since they may vary depending on the data used to compute the model. To further assess the relevance of selected regions, stability selection was considered using Randomized Lasso. After training Lasso and Group Lasso with their respective best λ values on the complete dataset, and running Randomized Lasso, the percentage of features selected was: 6.02% (28 folding measures and STV), 50.21% (240 folding measures representing 16

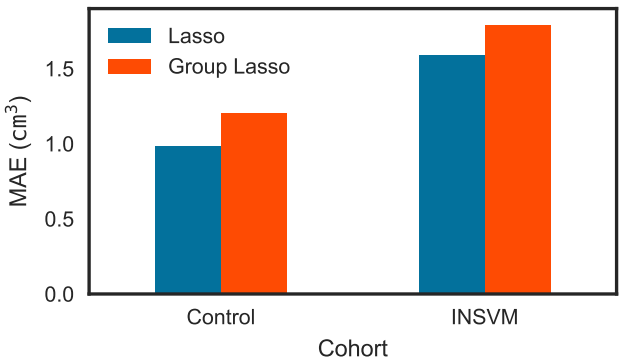

Figure 5.7: Total ventricular volume mean absolute error in normal and INSVM cohorts: Comparison between Lasso and Group Lasso. Results are obtained with the optimal cross-validated λ for each method.

cortical regions, GA and STV) and 6.02% (28 folding measures and STV) for Lasso, Group Lasso and Randomized Lasso, respectively. This shows that a few folding measures are relevant for predicting ventricular volume, according to the number of features selected by both Lasso and Randomized Lasso. Noteworthy is the fact that STV was selected by all sparse models, whereas GA was relevant for predicting ventricular volume only in Group Lasso. Although Figure 5.4 shows a correlation between ventricular volume and GA, STV appears to better capture this association. This is because STV carries information about GA ($p < 10^{-12}$) and, most importantly, is highly sensitive to ventricular volume ($p < 10^{-6}$).

Figure 5.8 shows the most relevant regions, in terms of number of features selected. It is important to note that MITGpp, GPAap, GPApp, occipital lobe and insula were selected bilaterally by all sparse models, which is in great accordance with our previous statistical analyses. However, in the parietal lobe, that showed significant association with ventricular volume enlargement, no curvature feature was selected by any regression model. This suggests that, although this region is correlated with ventricular enlargement, is not part of the predictive pattern obtained when considering all measures in conjunction. Curvature features selected by Lasso and Randomized Lasso are shown in Figure 5.9. The most recurrent features (selected in more than 3 cortical regions by Lasso or Randomized Lasso) were MC, PMC, GC, SI and PSI. MC and PMC denote extrinsic folding differences between cortices of healthy fetuses and those of the INSVM cohort in the cortical regions were these folding measures were found relevant by our sparse estimators. MC was found relevant in the right insula, for example,

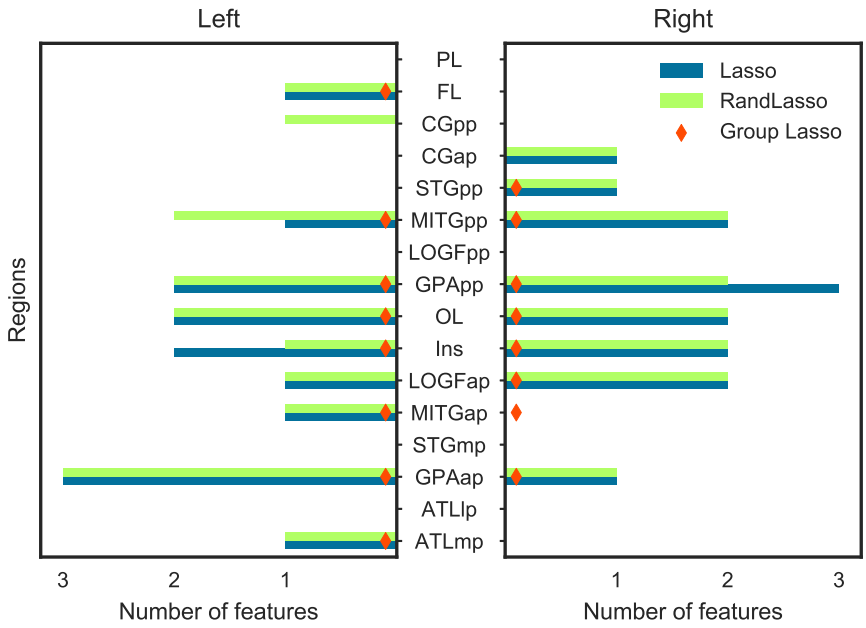

Figure 5.8: Number of features selected from cortical regions considered by left and right hemisphere. Note that Group Lasso selects all the curvature features from a particular region.

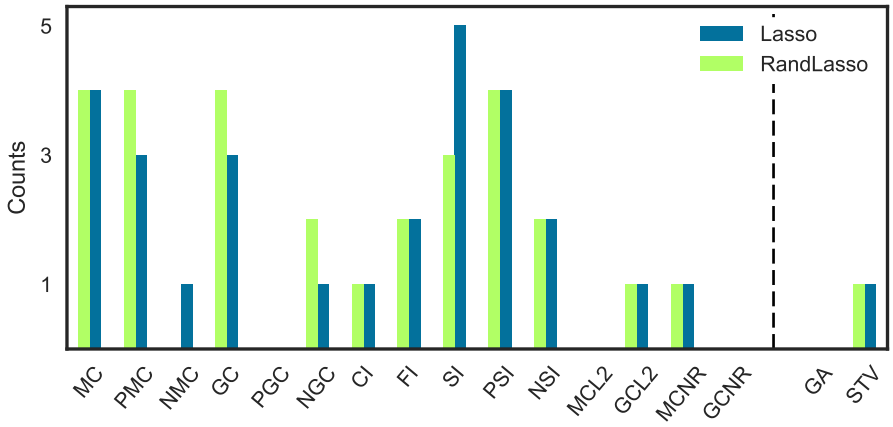

Figure 5.9: Feature selection counts corresponding to the number of cortical regions where a particular curvature feature was selected.

which is predominantly more concave in the normal cohort than in INSVM fetuses with larger ventricles (as shown in Figure 5.10). The left GPAap, however, was extrinsically less folded (lower MC) in INSVM subjects. In

the right MITGpp, PMC was selected by both Lasso and Randomized Lasso because of its negative correlation with ventricular volume. The gyrus is more prominent in the normal cohort, as clearly illustrated in Figure 5.10. The relevance of GC implies that there exists a difference in intrinsic folding (i.e., with distortion) between both cohorts in regions such as the left occipital lobe (where GC was found relevant). SI was found to be relevant, for example, in the right LOGFap and PSI in left GPAap. The former expresses that right LOGFap is, in general, more convex in healthy fetuses, while the latter points out to the existence of a less prominent convexity within the left GPAap. This, however, does not mean that other selected measures were not important. It simply indicates that the most recurrent measures were found to be strongly correlated with ventricular volume in several cortical regions, which emphasizes their predictive power and higher sensitivity over the rest of measures to capture the putative effects of INSVM in such regions. On the other hand, PGC, MCL2 and GCNR were not relevant for any cortical region, which may reflect that when used together with the rest of folding measures, they provided no additional information for our sparse models in the prediction of ventricular volume or were not able to capture folding differences between both cohorts. Despite some differences, results obtained with GLM analyses and sparse regression are concordant.

5.4 Discussion

In this work, we studied the association of INSVM with altered cortical folding *in utero*. Using high-quality 3D reconstructions of fetal brain MR images, inner cortical surfaces were extracted and several folding measures were computed to characterize the curvature and degree of gyrification of each fetal brain. Two different approaches based on statistical analysis and sparse linear regression were employed to identify cortical regions associated with ventricular dilation. In the first approach we probed the capability of diagnosis or ventricular volume to explain the observed folding measure for each cortical region individually. We also corrected for other covariates (i.e., GA and supratentorial volume) that could affect the relationship to make sure that the differences are only due to diagnosis or ventricular volume. In the second approach we aimed to find a pattern among the observed regions that is predictive of total ventricular volume. Such pattern is composed of several cortical regions and we considered that a region is relevant if it is included in such pattern. Both approaches offer complementary information

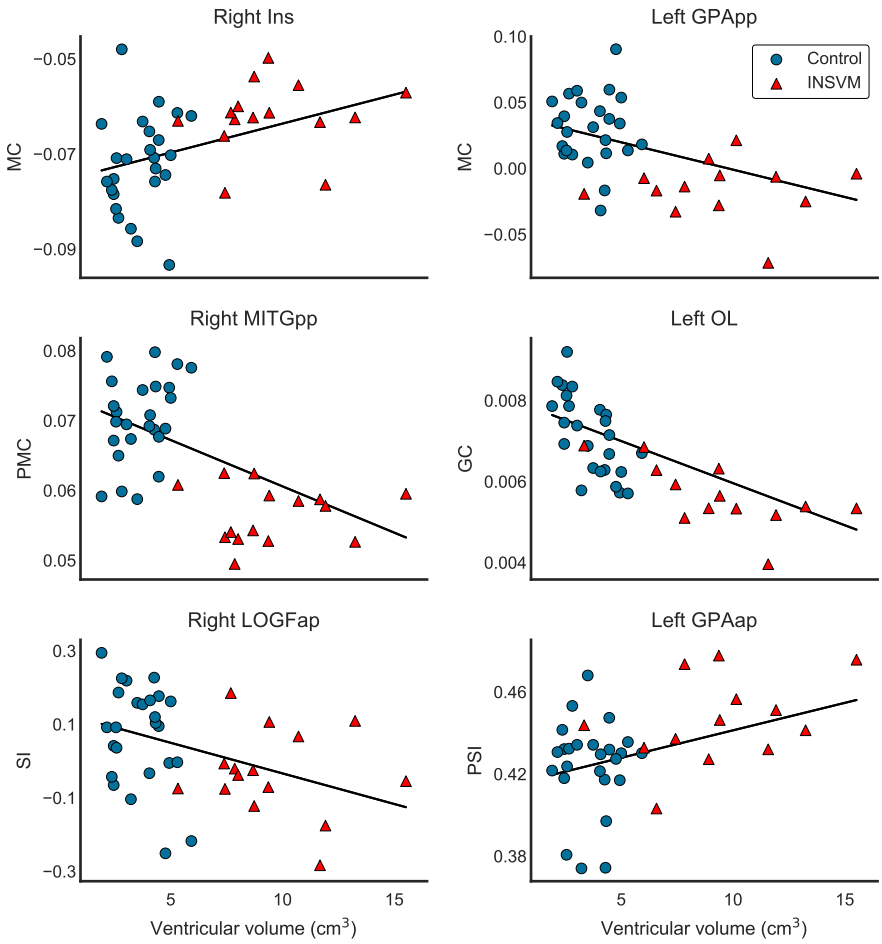

Figure 5.10: Examples of folding measures (selected by our sparse models) against ventricular volume. Linear fit is included to show relationship of measures with increasing ventricular volume.

since regions are either considered separately or in groups. We conclude that the regions that are significant in the statistical analysis and also selected by the sparse linear models can be considered to play an important role in gyrification in fetuses with INSVM. Both approaches showed evidence of disturbances in cortical folding in the presence of VM. Their results were in great accordance, with reduced folding in insula, MITGpp, GPApp and occipital lobe demonstrating consistent bilateral relationship with increasing ventricular volume.

5.4.1 Alterations in cortical folding

In the statistical analyses, ventricular volume was more sensitive to changes in folding than diagnosis, being able to capture differences in additional regions (i.e., occipital and parietal lobes) and in the whole hemisphere. When cortical folding was analyzed separately for each hemisphere, there was an increased number of folding measures that significantly correlated with ventricular volume in the insula and occipital lobe in both hemispheres. This shows stronger associations of altered cortical folding with ipsilateral ventricular enlargement and underlines the importance of using hemispheric analyses. We argue this is because our dataset is composed of INSVM subjects with unilateral and bilateral ventricular dilation, and global analysis is not able to discern the laterality. Although this might suggest stronger ipsilateral effects of ventricular enlargement in cortical folding, a contralateral analysis might reveal interactions across different hemispheres. Cortical regions found relevant by the sparse linear models, emphasizing regions selected by Randomized Lasso, were in great overlap with regions significantly associated with ventricular volume in the statistical analyses, with the important exception of the parietal lobe. Although Randomized Lasso selected new cortical regions (e.g., left: CGpp, ATLmp and LOGFap, right: LOGFap and MITGap), it is worth mentioning that these regions might have been selected to account for the amount of ventricular volume not explained by the regions with reduced folding.

Based on 3D reconstructed MR images, Scott et al. (2013) used a vertex-wise approach to investigate the association of isolated mild ventriculomegaly (IMVM) with alterations in cortical folding. Significant deviations in mean curvature were only found in the anterior aspect of the parieto-occipital sulcus. On the contrary, in the current study, several cortical regions showed significantly lower values associated with INSVM in multiple folding measures (e.g., mean curvature, Gaussian curvature and curvedness). Our findings encompass the narrow area of the parieto-occipital sulcus that distinguished fetuses with IMVM from the normal cohort. Scott et al. (2013) analyzed folding alterations in fetuses with mild VM (i.e., atrial diameter of 10-12 mm) instead of INSVM that includes both mild and moderate VM. The inclusion of fetuses with larger lateral ventricular volumes in our cohort than the ones with IMVM may explain the additional regions with altered folding found in our study. Moreover, Scott et al. (2013) compared folding between 16 healthy controls and 16 IMVM fetuses, while our database is composed of 23 INSVM subjects and 25 controls.

According to sulcogenesis studies based on cortical surfaces (Clouchoux et al., 2012; Habas et al., 2012), the parieto-occipital sulcus, separating the parietal and occipital lobes, undergoes significant changes starting from the 23rd GW. In the occipital lobe, significant changes in curvature are found in the calcarine sulcus and lingual gyrus around the 24th GW and continue until they are well formed around 28-29 GWs. This might have precluded finding folding differences in these areas between controls and cases in (Scott et al., 2013). The same occurs in the insula, for instance, where first significant folding changes are detected around 24-25 GWs. Between 22 and 25.5 GWs, the age range studied in (Scott et al., 2013), most primary sulci are still emerging, which may difficult the detection of cortical alterations. The GA of our cohort is one of the factors that allowed us to identify gyrification differences in more cortical areas than only in the parieto-occipital sulcus. In normal *in utero* brain development, the age range of 26-29 GWs of our cohort constitutes a period of cortical folding that occurs immediately before gyrification reaches its peak of growth rate around the 30th GW in all cortical regions (Wright et al., 2014). By the end of this period, most primary gyri and sulci are formed in the developing brain.

Although folding differences in our work were analyzed at a regional scale, reduced folding in the aforementioned cortical regions hints to the existence of gyrification alterations that can be found in more localized areas confined within these regions.

5.4.2 Prognosis in INSVM fetuses

We know from several published studies assessing perinatal and long-term outcomes that the incidence of unfavorable neurodevelopmental outcome is low in INSVM fetuses, and even lower in mild than moderate VM (Gaglioti et al., 2005; Griffiths et al., 2010; Leitner et al., 2009). Within isolated VM, ventricular size and progressive ventricular enlargement are the main factors showing association with adverse outcome and postnatal cerebral abnormalities not detected *in utero* (Baffero et al., 2015; Gilmore et al., 2008; Kelly et al., 2001; Scala et al., 2017).

Neurodevelopmental impairment observed with postnatal assessment in fetuses diagnosed with INSVM include delays in cognitive, psychomotor and language skills (Beeghly et al., 2010; Leitner et al., 2009; Lyall et al., 2012; Ouahba et al., 2006). Risk of attention-deficit/hyperactivity disorder was

also observed in children with prenatal non-severe VM (Ball et al., 2013; Gilmore et al., 2001). These functional deficits were also associated with abnormal cortical descriptors in a vast variety of disorders such as mental retardation (Zhang et al., 2010), attention-deficit/hyperactivity disorder (Wolosin et al., 2009) and schizophrenia (Jou et al., 2005; Powell, 2010). Fetuses with prenatally diagnosed INSVM showed larger cortical GM volume (Kyriakopoulou et al., 2014), which was also present in neonates along with alterations in WM (Gilmore et al., 2008; Lyall et al., 2012). The usefulness of cortical folding in predicting adverse outcome of fetuses with INSVM remains speculative and is yet to be evidenced by longitudinal and follow-up studies. Nonetheless, cortical folding has proved to be a potential predictor in preterm neonates (Moeskops et al., 2015), schizophrenia (Guo et al., 2015) and Alzheimer's disease (Cash et al., 2012).

Fetal MRI grants the means to obtain more detailed information of the *in utero* brain. With the recent advancements on motion-correction and super-resolution techniques, fetal MRI is becoming an important neuroimaging modality that offers the possibility to use a wide variety of approaches to study malformations during *in utero* neurodevelopment. For INSVM, using fetal MRI for the search for novel prognostic biomarkers that are able to elucidate the neurodevelopmental outcome of fetuses with non-severely dilated ventricles is essential to identify fetuses at higher risk of neurodevelopmental deficits. Cortical regions with significant convolutional alterations found in this study constitute potential candidate biomarkers that must be assessed in future follow-up studies to validate their prognostic or predictive power in distinguishing between the two subgroups of INSVM fetuses (i.e., INSVM fetuses with favorable and adverse outcome).

Still, the association of VM with decreased cortical folding and its effects in postnatal neurodevelopment remain poorly understood. Moreover, drawing conclusions about the direct influences of VM on gyrification, with the possible exception of cortical areas surrounding the dilated ventricles, is not straightforward since the etiology of VM remains unknown (Kelly et al., 2001) and there may be other common hidden factors that cause both lateral ventricular dilation and decreasing gyrification.

5.4.3 Limitations and future work

One limitation of the current study is the large extension of the cortical regions identified to be related to INSVM. Although cortical folding al-

terations were found in the whole hemispheres, reduced cortical folding in these regions (e.g., occipital lobe) could be concentrated in more focalized areas. Therefore, in order to find such specific areas, a vertex-wise analysis needs to be conducted. Regarding the cohorts, the study group was composed of fetuses diagnosed with mild or moderate VM and no distinction was made to assess the alterations in gyrification associated with each subgroup. However, this may not be considered a limitation since we analyzed the correlation with lateral ventricular volume instead of atrial diameter. Furthermore, from the neurodevelopmental perspective, there is no consensus in the literature on the differences between mild and moderate VM. Another limitation is the gender distribution in our cohorts. Our dataset is composed of 25 healthy controls (14 males, 11 females) and 23 INSVM subjects (21 males, 2 females), as shown in Table 5.1. This imbalance between groups (i.e., male over-representation in the INSVM cohort, with only 2 females) has precluded studying the effect of gender. A further limitation comes from the automatic tissue segmentation. Only 4 images were manually delineated and used as atlases to segment the remaining 44 brain MR images. This might have affected our analyses, especially in the anterior part of the parahippocampal gyri, where segmentation could be improved.

In the current study, although only curvature-based folding measures were analyzed, other measures related with cortical folding such as gyrification index (Zilles et al., 1988) and sulcal depth can be studied. Given the existence of cortical folding alterations associated with INSVM, cortical thickness constitutes another measurement worth investigating. Future work also involves testing all folding measures and cortical regions associated with INSVM with follow-up data, using neonatal images and other information such as neurodevelopmental test scores. This may serve to select, among all these candidate cortical regions, the ones with altered folding that might be used as biomarkers to identify the fetuses with INSVM that will have poor outcome.

5.5 Conclusions

In this work, we studied *in utero* gyrification in the presence of non-severe ventricular enlargement. Our findings demonstrate a relationship of VM with reduced folding in several cortical regions, not only restricted to the parieto-occipital sulcus. Delving deeper into other *in utero* maturational processes that occur in the fetal brain, such as cortical folding, can shed

light on other possible malformations that might lead to adverse neurodevelopment of INSVM fetuses. To the best of our knowledge, this is the first work to analyze INSVM-specific cortical folding alterations within the GA range of 26-29 weeks using 3D reconstructed fetal MRI.

6

Revealing Regional Associations of Cortical Folding Alterations with In Utero Ventricular Dilation Using Joint Spectral Embedding

Abstract – Fetal ventriculomegaly (VM) is a condition with dilation of one or both lateral ventricles, and is diagnosed as an atrial diameter larger than 10 mm. Evidence of altered cortical folding associated with VM has been shown in the literature. However, existing studies use a holistic approach (i.e., ventricle as a whole) based on diagnosis or ventricular volume, thus failing to reveal the spatially-heterogeneous association patterns between cortex and ventricle. To address this issue, we develop a novel method to identify spatially fine-scaled association maps between cortical development and VM by leveraging vertex-wise correlations between the growth patterns of both ventricular and cortical surfaces in terms of area expansion and curvature information. Our approach comprises multiple steps. In the first step, we define a joint graph Laplacian matrix using cortex-to-ventricle correlations. Next, we propose a spectral embedding of the cortex-to-ventricle graph into a common underlying space where their joint growth patterns are projected. More importantly, in the joint ventricle-cortex space, the vertices of associated regions from both cortical and ventricular surfaces would lie close to each other. In the final step, we perform clustering in the joint embedded space to identify associated sub-regions between cortex and ventricle. Using a dataset of 25 healthy fetuses and 23 fetuses with isolated non-severe VM within the age range of 26-29 gestational weeks, our results show that the proposed approach is able to reveal clinically relevant and meaningful regional associations.

This chapter is adapted from:
Benkarim O. M., Sanroma G., Piella G., Rekik I., Hahner N., Eixarch E., González Ballester M. A., Shen D., and Li G. Revealing regional associations of cortical folding alterations with in utero ventricular dilation using joint spectral embedding. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. [Accepted.]

6.1 Introduction

During the intrauterine life, the fetal brain undergoes drastic maturational changes. Cortical folding is one of the major processes that occur during this period, and any deviation from its normal developmental course might lead to adverse postnatal outcome (Benkarim et al., 2017c). In prenatal ultrasound examination, ventriculomegaly (VM) is the most frequent abnormal finding in the fetal brain. VM is a condition with dilation of one or both lateral ventricles, as shown in Figure 6.1A. It is diagnosed as an atrial diameter larger than 10 mm at any gestational age (Cardoza et al., 1988). Evidence of altered cortical folding associated with *in utero* VM has been shown by studies in the literature. Among their findings, cortical gray matter was significantly enlarged in fetuses with isolated VM (Kyriakopoulou et al., 2014). Using curvature-based analysis, studies also found reduced cortical folding in the insula, the occipital lobe and the posterior part of the temporal lobe (Benkarim et al., 2018; Scott et al., 2013).

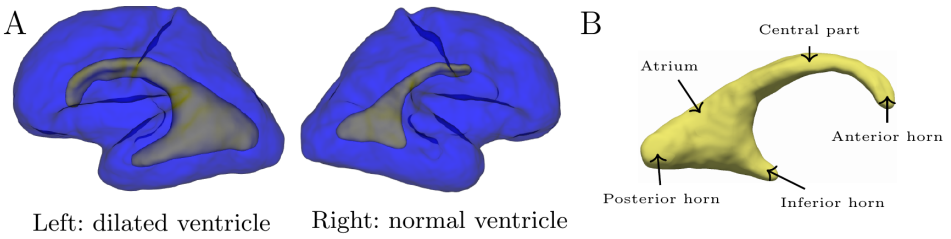

Figure 6.1: A: Cortical and ventricular surfaces of a 28 gestational weeks fetus with left VM. B: Regions of the lateral ventricle.

To study the association between VM and the morphology of cortical folding, existing works either use diagnosis or ventricular volume to characterize this condition. Although ventricular volume captures the extent of enlargement and is more distinctive than the dichotomous information offered by diagnosis, a single scalar value might not be sufficient to provide all the information related to ventricular enlargement (e.g., spatial information about the dilated ventricular regions). In this work, we aim to find associations between ventricular regions (see Figure 6.1B) and cortical folding by incorporating into our analysis the ventricular surfaces. For this purpose, we propose a novel approach to jointly analyze the cortical and ventricular shapes based on their growth patterns. The motivation for using growth patterns is their ability to reflect the underlying micro-structural brain changes. The main idea of our approach is to find a common underlying representation

of the vertex-wise growth patterns of both cortical and ventricular surfaces such that vertices with associated patterns from both anatomical surfaces can lie close to each other. In this way, regional associations can be conveniently found by identifying clusters containing vertices from both surfaces in the new latent space. The contributions of our work are threefold:

- We propose a novel approach for joint analysis of different anatomical shapes based on their growth patterns.
- We identify, for the first time, spatially fine-scaled associations related to *in utero* VM between ventricular surfaces and alterations in cortical folding.
- We use fusion of similarity matrices to capture associations based on multiple cortical features.

6.2 Method

Given P subjects and their corresponding cortical and ventricular surfaces with N_c and N_v vertices respectively, for each subject, the growth patterns $\mathbf{x}_i$ for each vertex are represented by:

$$\mathbf{x}_i = [x_i^1, x_i^2, \cdots, x_i^P], \tag{6.1}$$

where x_i^k is the feature (e.g., local surface area) of the k-th subject at vertex i. In this study, growth patterns were built using a cross-sectional dataset. Although it is preferable to use longitudinal data, repeated *in utero* imaging is difficult due to ethical and practical issues.

Cortical and ventricular growth patterns are not necessarily to be represented using the same feature (e.g., we can use area for ventricles while curvature for cortices). We assume that there exists a common underlying representation for these heterogeneous growth patterns, $\mathbf{x}_i$, such that vertices of associated regions from both surfaces can lie close to each other and, most likely, form dense clusters. Thus, we propose to find a shared representation of cortical and ventricular growth patterns using joint projection onto a common space:

$$Y = \underset{Y}{\operatorname{argmin}} \sum_{i,j} \|Y_i^c - Y_j^c\|^2 S_c(i,j) + \sum_{i,j} \|Y_i^v - Y_j^v\|^2 S_v(i,j) + \mu \sum_{i,j} \|Y_i^c - Y_j^v\|^2 S_{cv}(i,j), \tag{6.2}$$

where $Y = [Y^c, Y^v]^T$ is the common latent representation with $N = (N_c + N_v)$ rows such that the first N_c rows correspond to the embedded cortical growth patterns (i.e., Y^c) and the remaining N_v rows belong to the ventricle (i.e., Y^v), T stands for matrix transpose, S_c and S_v are the intra-structure similarity matrices, S_{cv} is the similarity matrix between cortical and ventricular growth patterns, and μ is a trade-off parameter. Given two similar (i.e., high $S_{cv}(i,j)$) growth patterns, $\mathbf{x}_i^c$ and $\mathbf{x}_j^v$ from cortex and ventricle respectively, the third term in Eq. (6.2) enforces their projections (i.e., Y_i^c and Y_j^v) to fall close to each other. This also occurs for similar growth patterns from the same surface (enforced by the first and second terms).

Since we are interested in identifying associations between the growth patterns of both structures, similarity between the growth patterns is defined in terms of correlation. First, we build the inter-structure similarity matrix based on the correlations between the growth patterns of both surfaces as follows:

$$S_{cv}(i,j) = \frac{1 + \rho(\mathbf{x}_i^c, \mathbf{x}_j^v)}{2}, \tag{6.3}$$

where ρ is Pearson's correlation coefficient. Similarly, intra-structure similarity matrices (S_c and S_v) are built to capture within surface correlations. The joint similarity matrix is constructed by filling its block-diagonal with the intra-structure similarity matrices and the off-diagonal with the inter-structure similarity matrix:

$$S = \begin{pmatrix} S_c & \mu S_{cv} \\ \mu S_{cv}^T & S_v \end{pmatrix}. \tag{6.4}$$

Then, we compute the normalized Laplacian of the joint similarity matrix:

$$L = I - D^{-1/2} S D^{-1/2}, \tag{6.5}$$

where D is the degree matrix of S (i.e., a diagonal matrix such that $D(i,i) = \sum_j S(i,j)$), and I is the identity matrix. Laplacian eigenmaps (Belkin and

Niyogi, 2003) can then be used to solve Eq. (6.2) based on the joint Laplacian and find the common underlying space Y.

To discover the regional relationships induced by ventricular enlargement, we cluster the embedded growth patterns using hierarchical clustering. Associated regions are identified by clusters containing vertices from both shapes.

Features for Cortical Growth Patterns

The area of ventricular surfaces increases dramatically with the enlargement and can be considered reliable in capturing the ventricular dilation. However, alterations in cortical folding can be characterized by multiple distinct features. Therefore, we extend our approach to include both area and curvedness (derived from curvature) as features for cortical surfaces by fusing the similarity matrices created for each of them with ventricular area: S_1 built using area for both structures, and S_2 using curvedness for cortices. For each similarity matrix, S_m, $m \in \{1, 2\}$, we derive two matrices (Wang et al., 2014a):

$$P_m(i,j) = \begin{cases} \frac{S_m(i,j)}{2\sum_{k \neq i} S_m(i,k)} & i \neq j \\ 1/2 & \text{otherwise.} \end{cases} \tag{6.6}$$

$$W_m(i,j) = \begin{cases} \frac{S_m(i,j)}{2\sum_{k \in \mathcal{N}_i} S_m(i,k)} & j \in \mathcal{N}_i \\ 0 & \text{otherwise,} \end{cases} \tag{6.7}$$

where $\mathcal{N}_i$ denotes the neighborhood of the i-th vertex in terms of the vertices with most correlated growth patterns. Fusion is then iteratively conducted:

$$P_1^{t+1} = W_1 P_2^t W_1^T, \qquad P_2^{t+1} = W_2 P_1^t W_2^T, \tag{6.8}$$

where P_m and W_m are the dense and sparse similarity matrices derived from S_m (i.e., S_1 and S_2). In this way, the reliable similarity information is diffused across similarity matrices. Finally, the dense matrices are averaged to obtain the fused matrix:

$$P_f = (P_1 + P_2)/2. \tag{6.9}$$

The fused similarity matrix, P_f, is able to capture common and complementary associations, and remove spurious and isolated correlations. We use P_f (rather than S) to compute the joint Laplacian and project the growth patterns.

6.3 Experiments

6.3.1 Data and Preprocessing

In our experiments, we used a fetal brain MRI dataset of 25 healthy controls and 23 subjects with isolated non-severe ventriculomegaly (INSVM) between 26 and 29 gestational weeks. The INSVM cohort was composed of fetuses with unilateral or bilateral ventricular width between 10-14.9 mm, with no abnormal karyotype, infections or malformations with risk of abnormal neurodevelopment. T2-weighted MR images were acquired on a 1.5 T scanner (SIEMENS MAGNETOM Aera) with an 8-channel body coil. For each subject, multiple orthogonal 2D scans (i.e., 4 axial, 2 coronal, and 2 sagittal stacks) were collected, from which a high-resolution 3D image was reconstructed using the method in (Murgasova et al., 2012).

Tissue segmentation was performed on the reconstructed images using a learning-based method (Sanroma et al., 2016a). Then, cortical and ventricular surfaces were extracted for each hemisphere. In order to establish vertex-wise correspondences, for each structure, surfaces were co-registered and resampled to the same number of vertices (Xia et al., 2018).

6.3.2 Experimental Setup

Ventricular growth patterns were built with area information from each vertex, which was computed as one third of the total area of adjacent triangles (Li et al., 2013). For cortices, we used both area and curvedness. Thus, we conducted 3 different experiments, using: 1) correlations between ventricular area and cortical area, 2) ventricular area and cortical curvedness, and 3) fusing both similarity matrices (i.e., using both area and curvedness from the cortices and area for the ventricles). For clustering, we used 2 to 25 clusters to illustrate the number of correlated regions identified with different clusters. The optimal associations between ventricles and cortices were determined by finding the most appropriate number of cluster using

the silhouette coefficient:

$$s(i) = (b(i) - a(i))/max(a(i), b(i)),$$

where i indexes vertices in the embedded space, $a(i)$ is the mean distance between the i-th vertex to the rest of vertices in its cluster, and $b(i)$ is the minimum average distance computed with the vertices in the rest of clusters.

6.3.3 Results

Although there may exist contralateral associations and unilateral ventricular enlargement may be associated with alterations in the opposite hemisphere, for this work, associations were only studied for each hemisphere independently. Figure 6.2 shows the associations identified by our approach in the left hemisphere between ventricular dilation and cortical folding. Surfaces are displayed such that cortical and ventricular regions found to be associated are depicted with the same color code. From these results, regardless of using area or curvedness to characterize the cortical growth patterns, we can observe that, with 3 clusters, the posterior part of the ventricular surface and the posterior part of the cortical surface fall into the same cluster (blue for area and pink for curvedness). This pattern is replicated for the anterior part (green for area and cyan for curvedness) and

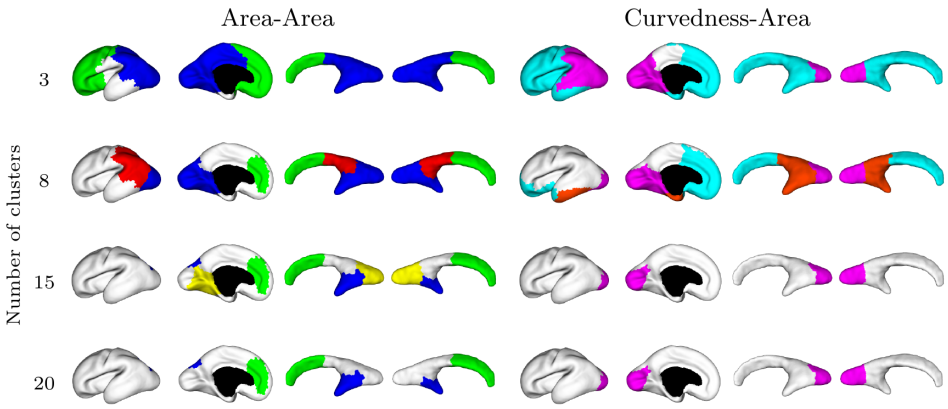

Figure 6.2: Comparison of regional associations identified in the left hemisphere using correlations between: a) cortical area and ventricular area, and b) cortical curvedness and ventricular area. Associations between cortex and ventricle are color-coded, with white depicting regions with no associations.

further preserved with increasing clusters, as clusters emerge in the posterior/anterior parts of both surfaces. As the number of clusters increases, we obtain more localized and fine-grained associations (i.e., shared clusters), which emphasizes the strength of the maintained associations.

Comparing the associations found when using area expansion and curvature information for the cortex, we can see that, with 8 clusters, the anterior horn and part of the ventricular body are associated with a region nearby the anterior cingulate gyrus. This association is captured with cortical area (green) for a larger number of clusters than curvedness (cyan). The most important association found by curvedness is between the posterior (i.e., occipital) horn and the occipital lobe (pink). With 8 clusters, the association includes the calcarine and the parieto-occipital fissures, although only a small part of the latter fissure is preserved with 20 clusters. This association is also found by cortical area with 8 (blue) and 15 clusters (yellow).

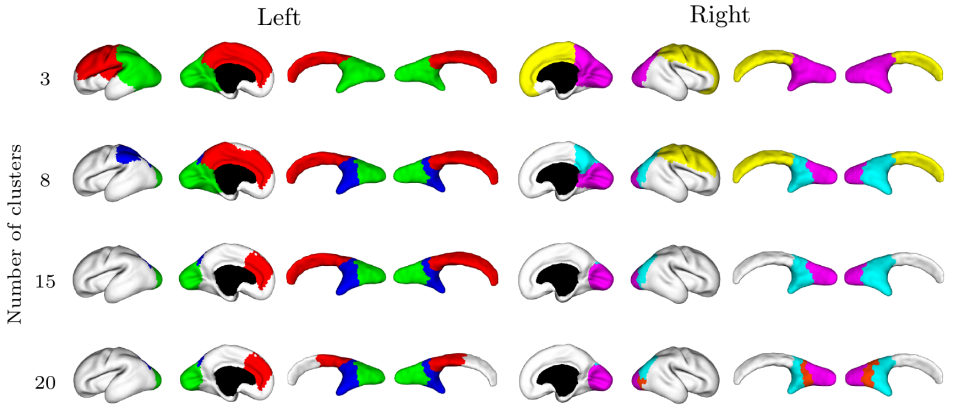

Figure 6.3: Regional associations identified using fused similarity matrix for left and right hemisphere separately. Associations between cortex and ventricle are color-coded, with white depicting regions with no associations.

Still, using a single feature to describe the growth patterns might not be able to capture all putative associations or give rise to spurious ones. Results using the fused similarity matrix for different number of clusters are shown for both hemispheres in Figure 6.3. Noteworthy is that associations found in both hemispheres are in large overlap, with the only difference being the correlation between the ventricular body and the anterior horn with the anterior cingulate gyrus (red) in the left cortex. Nonetheless, in both hemispheres, the posterior horn and the occipital lobe belonged in the same cluster (green in left hemisphere), and the inferior horn and the atrium (blue

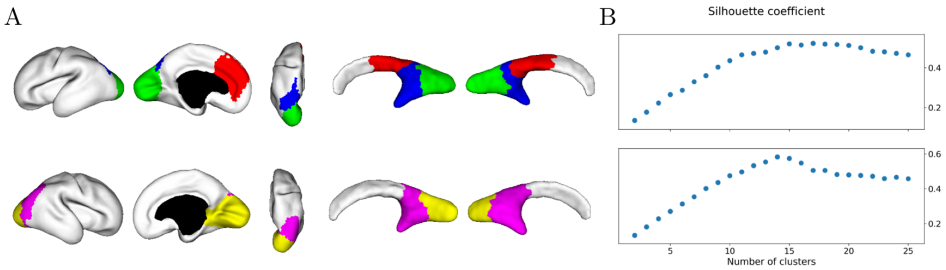

Figure 6.4: A: Associations between ventricular enlargement and cortical folding using the fused similarity matrix corresponding to the optimal number of clusters in terms of silhouette coefficient for left (top) and right (bottom) hemispheres. B: Silhouette scores for different number of clusters for each hemisphere.

and cyan in left and right hemispheres, respectively) showed to be correlated with the superior part of the parietal lobe. Associations corresponding to the best clustering in terms of silhouette score are shown in Figure 6.4. The highest values of silhouette coefficient were found with 17 and 14 clusters for the left and right hemispheres, respectively. The atrium was identified bilaterally (blue and pink for left and right hemispheres, respectively). Since the atrium is the ventricular region used in clinical practice to diagnose VM, this highlights the clinic relevance of our results. In the cortex, the occipital lobe was found to be associated with the posterior horn (green and pink) in both hemispheres, which is consistent with findings in the literature (Benkarim et al., 2018; Scott et al., 2013). In association with VM, our approach is able to identify meaningful cortical and ventricular regions. Furthermore, it provides the means to establish relationships between these regions and gain more insight into the fine-grained associations between ventricular enlargement and cortical development.

6.4 Conclusions

In this work, we have presented a novel approach to identify spatially fine-grained correlations between different shapes based on their growth patterns. This is the first work that approaches the study of associations between fetal VM and cortical folding alterations by jointly analyzing cortical and ventricular shapes. Our results demonstrate that the proposed approach is able to identify clinically relevant regions (e.g., atrium in the ventricle and occipital lobe in the cortex) and further establish relationships between these regions. For future work, instead of fusing similarity matrices

from different features prior to performing the embedding, multi-view approaches can be explored. Also, additional features (e.g., local gyrification index) can be used to identify other correlated regions.

7

Conclusions

After a thorough review in Chapter 2 of the literature on neurodevelopment in fetuses, this thesis has developed novel approaches to study the fetal brain. Within the automatic pipeline to study *in utero* neurodevelopment, our work was carried out in two important stages of this pipeline. First, we proposed two segmentation frameworks, which are nonspecific to the fetal brain albeit rather generic enough to be applied to any brain dataset, and were evaluated on the segmentation of adult and fetal brains. Second, we analyzed the deviations of *in utero* brain development from its normative course in the presence of ventriculomegaly, which is the most common abnormal finding in the fetal brain.

7.1 Research summary

7.1.1 Segmentation of brain MRI

Beyond the well-known limitations inherent to MR imaging (e.g., intensity inhomogeneity, noise and partial volume effects), the fetal brain poses additional challenges due to its dynamic rapid growth that might degrade segmentation quality and need to be carefully addressed to produce accurate results. In the state of the art, these challenges are dealt with using same or near-same gestational age atlases to the target subject to perform the segmentation, which emphasizes the age-specific design of existing approaches. If we pay close attention to the target fetal brain, we can surmise that this age-specific design is adopted to circumvent two important challenges: 1) changes in tissue intensity due to myelination and the laminar organization occurring in the fetal brain throughout gestation, and 2) changes in size and shape caused by brain growth and cortical folding with which the brain acquires its highly convoluted morphology. In this thesis, we approached these challenges from different methodological perspectives: registration accuracy and intensity similarity, which correspond to our two proposed segmentation frameworks.

In Chapter 3, we focused on systematic errors caused by registration failures. We proposed a patch-based multi-atlas segmentation framework that learns voxel-wise confidence estimators for each atlas to weigh its contribution in the final segmentation according to the spatially-varying quality of registration between such atlas and the target image. In the case of fetuses, using atlases with considerably different gestational ages to segment a given subject will initially result in notable registration failures. Thus, in our

framework, their contribution to the label fusion would be reduced. Even if our framework does not target exclusively the fetal brain, the age-specific character of the aforementioned approaches is implicitly incorporated into our framework. Noteworthy is the fact that these spatially-varying confidence maps prevent atlases with large anatomical differences with the target image from contributing to the label fusion only in regions where registration potentially fails, as opposed to age-specific approaches that completely drop atlases that are from different ages, irrespective of their anatomical similarity. Furthermore, we proposed to additionally augment the intensity information fed in the form of cubic patches to the voxel-wise estimators with label-dependent features that have shown to provide superior discriminative power to our supervised estimators, as reported in the results.

Fetuses with dissimilar gestational ages manifest dissimilar brain tissue intensities in magnetic resonance imaging (MRI). After registration, age-specific approaches tend to warp the atlases to the target image, forcing atlas intensity images to acquire a similar appearance to the target image as a result of interpolation. Warping the atlases, however, is an important source of errors. First, because interpolation strategies used to warp the atlas intensity images and labelmaps are different. Second, because atlas appearance is distorted to best match the target's, even though they could be substantially different. The latter point is even more critical when working with fetal brains if we consider myelination and the transient nature of some tissues such as the subplate and the intermediate zone. In Chapter 4, we proposed a patch-based multi-atlas segmentation framework where intensity and label information is extracted from the native spaces of the atlases and target images, thus avoiding interpolation errors. In this framework, training atlases are registered to a common template space, which is used to establish correspondences between atlases and target images. However, no warping is performed. In this way, atlases, both intensity images and labelmaps, do not suffer from interpolation errors and we avoid duplicating atlases due to warping. When using a fetal atlas with a notably different gestational age from the image to be segmented, we are using the intensity and label patches in the native space, whose contribution is to be calibrated during label fusion, instead of some distorted version produced after warping that tends to be more similar to the target patch.

In this framework, we further explore the similarities between weighted-voting and a particular case of supervised label fusion using SVM with RBF kernel. The support vectors found by the SVM during learning can be related to patch pre-selection, and weighted-voting is carried out solely

based on these relevant patches (i.e., support vectors). Our results highlight the importance of patch pre-selection as the latter approach outperforms the former.

7.1.2 Analysis of *in utero* neurodevelopment

In Chapters 5 and 6, we studied cortical folding alterations in the presence of isolated non-severe ventriculomegaly. The dataset used in the analysis of fetal ventriculomegaly is composed of fetuses within the age range of 26-29 gestational weeks, which corresponds to the mid-third trimester of gestation, an intrauterine period that, to the best of our knowledge, has not been covered in the literature. Thus, our work improves and extends our knowledge base about *in utero* maldevelopment.

In Chapter 5, we analyzed the relationships of ventricular enlargement with altered cortical folding on a regional basis by subdividing the cortical surfaces into several regions. Cortical folding for each region was characterized with several curvature-based descriptors, while diagnosis and lateral ventricular volume were used to characterize ventriculomegaly. This study was carried out using two different approaches. First, global and hemispheric analyses were performed to asses the association of ventriculomegaly with regional folding measures. General linear models were used to analyze the significance of diagnosis or ventricular volume in predicting each of the folding measures, accounting for gestational age and supratentorial volume as covariates. Second, sparse linear regression approaches were used to predict total lateral ventricular volume using all folding measures from every cortical region. Our purpose was to focus on regional associations rather than a particular folding measure.

From the findings of our first approach, we can conclude that ventriculomegaly is better characterized by ventricular volume, which captures the extent of dilation, than diagnosis. Moreover, these findings also suggest stronger ipsilateral than global relationships, as our hemispheric analyses showed to capture more significant associations. The second approach further confirmed the findings of our statistical analyses, highlighting relevant associations found between ventriculomegaly and altered cortical folding in the insula, the parietal lobe and the posterior part of the temporal lobe.

In Chapter 6, we analyzed these associations from a completely different perspective. We proposed a novel approach to study fine-grained associations between ventriculomegaly and cortical folding based on joint spectral

embedding. In this work, ventriculomegaly is no longer characterized by a single scalar value, but rather using the lateral ventricular shapes. Local area was used to describe the ventricular surfaces, whereas curvedness and local area were used to describe gyrification in the cortical surfaces. After surface alignment, cortical and ventricular growth patterns for each vertex were built to find the correlations between both sets of shapes. Our proposed approach proceeds by first constructing the joint similarity matrix based on vertex-wise growth patterns correlations, computing the graph joint Laplacian and projecting the data to a common low-dimensional space where associations between groups of vertices from both sets of shapes can be identified using hierarchical clustering. In this underlying representation, we assume that growth patterns would lie close to each other if they are in strong correlation. Our findings suggest that the relationships between ventricular dilation and alterations in cortical folding may obey a regional pattern, although more research is needed in this direction.

Although a proof of concept study, our work, in Chapter 6, on incorporating the lateral ventricular shapes, along with the cortical shapes, into the study of ventriculomegaly represents an unprecedented novel approach to the joint analysis of different anatomical shapes, which is not restricted to our case study. Furthermore, instead of the cross-sectional dataset used in our work due to the ethical and practical issues related to fetal brain imaging, our approach can be used with longitudinal data. Note that, for fetuses, using cross-sectional instead of longitudinal data is a common practice in the literature to build, for instance, spatio-temporal atlases.

7.2 Future research directions

Most differences in morphology and appearance between fetal brains would translate into registration errors. However, although our first segmentation framework mainly targets these errors, it does not consider the correlations between training atlases nor does it enforce spatial smoothness. In the case of fetal brain segmentation, it is very clear that we do not need to include young fetuses into the label fusion for segmenting very old fetal brains. Therefore, atlas selection can play an important role. Also in targeting the general brain population, atlas selection would considerably reduce the computational complexity of our segmentation framework. Registration errors are learned offline offering very fast segmentation runtimes. Still, learning is performed in each of the atlases, thus learning times increase

with the number of atlases in the training set. This can be reduced by identifying the most representative atlases of the population and restricting learning to these spaces. A potential line of future research could be to explore clustering techniques to select such atlases. The first segmentation framework also introduced a novel set of features to augment the intensity information of the patches, the so-called label-dependent features. In this work, a small number of hand-crafted features were computed based on atlas label patches. Given their proven added discriminative value, another direction of research would involve adopting a supervised approach to learn these features instead of using feature engineering.

The computational burden required by this first segmentation framework is vastly reduced in the second approach proposed in Chapter 4. This second approach exploits the use of a common template space to perform offline learning in its supervised version. Therefore, only one classifier is learned regardless of the number of training atlases in the database. A promising direction of work would be to integrate the ideas of both approaches into a common framework, therefore, learning registration errors based on the native patch and label information. This new framework would jointly lessen the impact of registration errors and completely remove interpolation artifacts from affecting the label fusion.

With regard to fetal brain segmentation, the short age range of our dataset did not allow us to test our proposed frameworks with very difficult cases. Although around half of the subjects in the dataset were diagnosed with unilateral or bilateral ventriculomegaly, which makes their segmentation challenging, these frameworks could be tested using subjects from a larger age range database.

In the analysis of fetal ventriculomegaly, our first work investigated associations of ventricular dilation with gyrification for each cortical region. A given folding measure describing large cortical regions such as the frontal or parietal lobes might not capture the subtle folding changes occurring along the surface of such lobes. A vertex-wise analysis could be further performed to reveal more localized areas within the aforementioned regions found by this study. Furthermore, all folding measures were curvature-based. Measures of cortical folding such as sulcal depth and local gyrification index could offer a different insight into cortical folding and help find other possible associations. This is also worth investigating as a continuation of our second work on ventriculomegaly presented in Chapter 6. In this approach, gyrification was characterized in terms of curvedness and local area, then

correlations with ventricular vertex-wise area were computed. This produced two similarity matrices that we fused prior to projecting the data. However, more sophisticated approaches could be proposed to merge both the fusion and the projection steps, such as multiview manifold learning.

Hemispheric statistical analyses revealed more associations in terms of the number of folding measures than global analysis. In the second work, only ipsilateral associations were analyzed, working with each hemisphere separately. Two directions of work with important clinical significance could be explored in this scenario. First, extending our second work to include both hemispheres. Second, the overlap between hemispheric and global findings encourages the research on contralateral associations. The clinical significance of our findings requires more study to assess the prognostic power of the folding measures and the cortical regions to be used as potential biomarkers to discern and identify those subjects that will have poor postnatal neurodevelopment. Longitudinal analysis may shed more light on this, by incorporating follow-up postnatal imaging into the study of fetal ventriculomegaly.

A

Discriminative Confidence Estimation for Probabilistic Multi-atlas Label Fusion (Supplementary Tables)

In support of the overall results reported in Chapter 3, we further include Tables A.1 and A.2 to show segmentation performance (in both SATA and ADNI datasets) when using the affine (AF) and the coarse non-rigid registrations (NR1) for each subcortical structure separately. Analogously, Tables A.3 and A.4 report segmentation results on our fetal dataset for AF and NR1 registrations, respectively. Performance is reported in terms of Dice overlap (top rows) and MHD (bottom rows) for each table.

	SATA							ADNI
	Acc	Amy	Cau	Hip	Pal	Put	Tha	Hip
MV	0.608±0.087	0.670±0.086	0.715±0.164	0.720±0.056	0.770±0.078	0.820±0.063	0.858±0.046	0.635±0.068
Naive	0.609±0.087	0.669±0.086	0.714±0.164	0.721±0.056	0.767±0.078	0.821±0.063	0.858±0.046	0.632±0.068
STAPLE	0.605±0.095	0.681±0.084	0.712±0.175	0.724±0.057	0.770±0.077	0.821±0.057	0.854±0.045	0.670±0.072
STEPS	0.629±0.105	0.688±0.081	0.723±0.161	0.729±0.064	0.773±0.080	0.822±0.061	0.859±0.049	0.733±0.058
LWV	0.642±0.098	0.695±0.081	0.784±0.134	0.760±0.053	0.780±0.079	0.833±0.063	0.875±0.040	0.710±0.065
JOINT	0.776±0.054	0.791±0.048	0.860±0.103	0.851±0.032	0.876±0.034	0.917±0.028	0.916±0.024	0.835±0.043
SCMNF	0.751±0.077	0.759±0.056	0.878±0.068	0.831±0.036	0.868±0.040	0.911±0.038	0.914±0.019	0.811±0.039
SCMWF	0.767±0.072	0.770±0.055	0.881±0.060	0.836±0.037	0.869±0.040	0.911±0.034	0.913±0.021	0.818±0.040
SCMNF2	0.777±0.063	0.787±0.053	0.889±0.057	0.847±0.031	0.872±0.039	0.913±0.030	0.915±0.020	0.832±0.039
SCMWF2	**0.788±0.056***	**0.802±0.043***	**0.896±0.054***	**0.860±0.024***	**0.877±0.033**	**0.917±0.028**	**0.919±0.018**	**0.843±0.038***
MV	0.547±0.188	0.476±0.189	0.481±0.509	0.456±0.214	0.299±0.144	0.230±0.112	0.193±0.095	0.573±0.247
Naive	0.623±0.191	0.500±0.194	0.515±0.503	0.485±0.216	0.316±0.140	0.239±0.113	0.194±0.091	0.622±0.257
STAPLE	0.553±0.202	0.462±0.189	0.494±0.546	0.448±0.214	0.300±0.146	0.225±0.101	0.197±0.096	0.475±0.177
STEPS	0.487±0.204	0.431±0.172	0.457±0.498	0.425±0.230	0.287±0.145	0.224±0.109	0.190±0.103	0.358±0.130
LWV	0.371±0.169	0.345±0.144	0.317±0.451	0.283±0.110	0.224±0.120	0.159±0.084	0.140±0.090	0.368±0.209
JOINT	0.260±0.092	0.256±0.086	0.244±0.381	0.190±0.071	0.146±0.070	0.095±0.052	0.102±0.073	0.237±0.121
SCMNF	0.283±0.136	0.272±0.099	0.207±0.286	0.206±0.068	0.149±0.074	0.099±0.051	0.105±0.068	0.244±0.119
SCMWF	0.278±0.132	0.267±0.096	0.202±0.280	0.202±0.066	0.147±0.072	0.098±0.050	0.103±0.067	0.237±0.114
SCMNF2	0.263±0.108	0.260±0.095	0.194±0.271	0.195±0.065	0.145±0.064	0.098±0.047	0.104±0.068	0.235±0.113
SCMWF2	**0.253±0.097**	**0.247±0.086**	**0.177±0.235***	**0.186±0.058**	**0.143±0.064**	**0.095±0.044***	**0.100±0.063**	**0.212±0.093***

Table A.1: Subcortical structure segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per structure, averaged left and right. Results obtained using affine registration (i.e., AF). Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods. Abbreviations: accumbens (Acc), amygdala (Amy), caudate (Cau), hippocampus (Hip), pallidum (Pal), putamen (Put) and thalamus proper (Tha).

				SATA				ADNI
	Acc	Amy	Cau	Hip	Pal	Put	Tha	Hip
MV	0.697±0.085	0.760±0.052	0.776±0.103	0.789±0.045	0.860±0.037	0.888±0.035	0.885±0.039	0.693±0.055
Naive	0.693±0.081	0.759±0.053	0.777±0.101	0.789±0.044	0.860±0.037	0.888±0.035	0.886±0.038	0.694±0.055
STAPLE	0.712±0.101	0.771±0.054	0.784±0.114	0.796±0.050	0.864±0.038	0.889±0.036	0.886±0.044	0.711±0.065
STEPS	0.727±0.053	0.753±0.036	0.856±0.043	0.819±0.025	0.844±0.031	0.889±0.027	0.900±0.018	0.763±0.048
LWV	0.758±0.057	0.790±0.044	0.847±0.099	0.830±0.038	0.881±0.031	0.916±0.030	0.906±0.032	0.748±0.051
JOINT	**0.794±0.046**	**0.823±0.029***	0.885±0.070	0.869±0.023	**0.887±0.031**	**0.924±0.024**	0.922±0.016	0.853±0.031
SCMNF	0.777±0.064	0.802±0.031	0.900±0.039	0.857±0.027	0.882±0.027	0.920±0.030	0.922±0.015	0.833±0.034
SCMWF	0.784±0.052	0.809±0.035	0.903±0.038	0.863±0.021	0.882±0.030	0.920±0.028	0.923±0.015	0.838±0.037
SCMNF2	0.785±0.048	0.811±0.035	0.902±0.036	0.867±0.021	0.884±0.029	0.921±0.026	0.925±0.014	0.848±0.037
SCMWF2	0.791±0.047	0.815±0.035	**0.905±0.035**†	**0.871±0.020**	0.885±0.029	0.923±0.026	**0.925±0.013**†	**0.856±0.036**
MV	0.359±0.131	0.290±0.069	0.307±0.224	0.288±0.123	0.150±0.038	0.124±0.049	0.148±0.080	0.439±0.166
Naive	0.415±0.117	0.311±0.075	0.321±0.209	0.301±0.110	0.160±0.037	0.129±0.050	0.148±0.073	0.476±0.167
STAPLE	0.367±0.148	0.290±0.069	0.327±0.261	0.289±0.125	0.155±0.038	0.128±0.049	0.159±0.093	0.379±0.124
STEPS	0.350±0.159	0.278±0.066	0.313±0.272	0.285±0.147	0.147±0.042	0.123±0.049	0.152±0.097	0.300±0.098
LWV	0.271±0.098	0.247±0.054	0.225±0.264	0.210±0.070	0.141±0.037	0.108±0.048	0.125±0.074	0.299±0.161
JOINT	**0.212±0.061**	0.199±0.038	0.159±0.228	0.156±0.045	0.119±0.031	**0.080±0.035**	0.092±0.055	0.188±0.099
SCMNF	0.229±0.070	0.204±0.044	0.150±0.186	0.158±0.037	0.121±0.033	0.083±0.038	0.092±0.038	0.205±0.104
SCMWF	0.225±0.068	0.202±0.043	0.148±0.180	0.156±0.036	0.120±0.032	0.083±0.038	0.091±0.036	0.200±0.100
SCMNF2	0.217±0.065	0.199±0.045	0.150±0.185	0.153±0.035	0.118±0.032	0.083±0.035	0.093±0.043	0.200±0.095
SCMWF2	0.214±0.060	**0.195±0.042**	**0.137±0.153***	**0.149±0.031**	**0.118±0.030**	0.082±0.034	**0.090±0.038**	**0.184±0.082***

Table A.2: Subcortical structure segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per structure, averaged left and right. Results obtained using the non-rigid registration NR1. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2 or SCMWF2. Abbreviations: accumbens (Acc), amygdala (Amy), caudate (Cau), hippocampus (Hip), pallidum (Pal), putamen (Put) and thalamus proper (Tha).

	Fetal					
	BS	CB	CSF	CoGM	LV	WM
MV	0.874±0.031	0.878±0.036	0.836±0.025	0.642±0.048	0.679±0.096	0.917±0.016
Naive	0.874±0.031	0.878±0.036	0.836±0.025	0.639±0.049	0.678±0.097	0.916±0.016
STAPLE	0.866±0.031	0.868±0.043	0.824±0.024	0.664±0.042	0.657±0.127	0.906±0.013
STEPS	0.877±0.027	0.878±0.036	0.828±0.018	0.716±0.042	0.750±0.083	0.932±0.012
LWV	0.903±0.025	0.916±0.027	0.911±0.015	0.791±0.033	0.802±0.085	0.950±0.012
JOINT	0.925±0.015	0.943±0.017	0.941±0.008	0.875±0.014	0.865±0.074	0.967±0.009
SCMNF	0.930±0.016	0.947±0.016	0.944±0.008	0.883±0.016	0.884±0.068	0.970±0.008
SCMWF	0.931±0.015	0.948±0.015	0.944±0.008	0.885±0.015	0.887±0.066	0.971±0.008
SCMNF2	0.932±0.015	0.949±0.016	0.946±0.008	0.892±0.011	0.883±0.074	0.970±0.007
SCMWF2	**0.932±0.015**†	**0.951±0.011**†	**0.950±0.009***	**0.893±0.015**†	**0.902±0.057***	**0.973±0.006***
MV	0.110±0.040	0.117±0.053	0.156±0.032	0.342±0.068	0.347±0.136	0.083±0.023
Naive	0.111±0.039	0.118±0.053	0.157±0.032	0.345±0.070	0.352±0.137	0.083±0.023
STAPLE	0.115±0.042	0.128±0.066	0.168±0.032	0.358±0.061	0.366±0.153	0.096±0.021
STEPS	0.104±0.034	0.117±0.054	0.151±0.023	0.257±0.050	0.236±0.088	0.062±0.016
LWV	0.081±0.030	0.079±0.039	0.076±0.017	0.178±0.038	0.205±0.119	0.047±0.016
JOINT	0.059±0.015	0.053±0.030	0.048±0.008	0.099±0.014	0.139±0.110	0.031±0.014
SCMNF	0.056±0.016	0.048±0.026	0.045±0.008	0.093±0.014	0.117±0.102	0.027±0.012
SCMWF	0.055±0.015	0.047±0.025	0.045±0.008	0.091±0.014	0.116±0.102	0.027±0.011
SCMNF2	**0.054±0.018**	0.047±0.027	0.044±0.007	0.086±0.011	0.125±0.114	0.027±0.011
SCMWF2	0.055±0.018	**0.041±0.014***	**0.040±0.008***	**0.084±0.013***	**0.092±0.078***	**0.023±0.006***

Table A.3: Fetal brain tissue segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per tissue. Results obtained using affine registration (i.e., AF). Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2 or SCMWF2.

	Fetal					
	BS	CB	CSF	CoGM	LV	WM
MV	0.914±0.012	0.918±0.017	0.879±0.023	0.654±0.054	0.816±0.048	0.927±0.015
Naive	0.914±0.012	0.918±0.017	0.879±0.023	0.650±0.054	0.814±0.048	0.926±0.015
STAPLE	0.914±0.014	0.913±0.021	0.874±0.021	0.675±0.053	0.801±0.090	0.921±0.014
STEPS	0.916±0.012	0.917±0.018	0.881±0.024	0.715±0.056	0.821±0.062	0.938±0.013
LWV	0.930±0.010	0.941±0.011	0.924±0.016	0.807±0.035	0.887±0.037	0.958±0.010
JOINT	0.935±0.006	0.955±0.005	0.943±0.011	0.880±0.017	0.916±0.032	0.971±0.006
SCMNF	0.940±0.007	0.957±0.006	0.946±0.011	0.887±0.020	0.921±0.033	0.973±0.006
SCMWF	0.940±0.007	0.957±0.006	0.947±0.010	0.888±0.020	0.922±0.032	0.973±0.006
SCMNF2	0.941±0.007	0.959±0.006	0.948±0.010	0.893±0.015	**0.928±0.032**	0.974±0.006
SCMWF2	**0.942±0.007**†	**0.959±0.006**†	**0.951±0.010***	**0.894±0.018**†	0.922±0.038	**0.974±0.006**
MV	0.067±0.011	0.069±0.020	0.112±0.030	0.345±0.094	0.165±0.044	0.072±0.021
Naive	0.067±0.011	0.069±0.020	0.112±0.029	0.350±0.093	0.170±0.045	0.073±0.020
STAPLE	0.066±0.011	0.074±0.025	0.114±0.027	0.340±0.092	0.172±0.079	0.079±0.023
STEPS	0.065±0.010	0.071±0.022	0.104±0.029	0.269±0.091	0.152±0.050	0.058±0.020
LWV	0.054±0.008	0.049±0.015	0.066±0.022	0.167±0.049	0.094±0.034	0.039±0.013
JOINT	0.050±0.006	0.037±0.012	0.049±0.018	0.097±0.020	0.068±0.030	0.026±0.009
SCMNF	0.046±0.006	0.035±0.012	0.046±0.016	0.093±0.025	0.065±0.032	0.024±0.008
SCMWF	0.046±0.006	0.035±0.012	0.046±0.016	0.091±0.024	0.064±0.031	0.023±0.008
SCMNF2	0.045±0.006	0.033±0.012	0.045±0.017	0.087±0.021	**0.059±0.031**	0.023±0.008
SCMWF2	**0.044±0.006***	**0.033±0.012**†	**0.041±0.014***	**0.085±0.020***	0.066±0.042	**0.022±0.007***

Table A.4: Fetal brain tissue segmentation. Mean Dice scores (top entries) and MHD (bottom entries) per tissue. Results obtained using the non-rigid registration NR1. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The $*$ symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2 or SCMWF2.

Bibliography

Aljabar, P., Bhatia, K., Murgasova, M., Hajnal, J., Boardman, J., Srinivasan, L., Rutherford, M., Dyet, L., Edwards, A., and Rueckert, D. (2008). Assessment of brain growth in early childhood using deformation-based morphometry. *NeuroImage*, 39(1):348–358.

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738.

Altaye, M., Holland, S., Wilke, M., and Gaser, C. (2008). Infant brain probability templates for MRI segmentation and normalization. *NeuroImage*, 43(4):721–730.

Anbeek, P., Išgum, I., van Kooij, B., Mol, C., Kersbergen, K., Groenendaal, F., Viergever, M., de Vries, L., and Benders, M. (2013). Automatic segmentation of eight tissue classes in neonatal brain MRI. *PLOS ONE*, 8(12):e81895.

Artaechevarria, X., Muñoz Barrutia, A., and Ortiz-de Solórzano, C. (2009). Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE Transactions on Medical Imaging*, 28(8):1266–1277.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851.

Asman, A. J., Huo, Y., Plassard, A. J., and Landman, B. A. (2015). Multi-atlas learner fusion: An efficient segmentation approach for large-scale data. *Medical Image Analysis*, 26(1):82–91.

Asman, A. J. and Landman, B. A. (2012). Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging*, 31(6):1326–1336.

Asman, A. J. and Landman, B. A. (2013). Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis*, 17(2):194–208.

Avants, B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., and Gee, J. (2010). The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49(3):2457–2466.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating auto-

mated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., and Gee, J. C. (2011). An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4):381–400.

Bach Cuadra, M., Schaer, M., Andre, A., Guibaud, L., Eliez, S., and Thiran, J.-P. (2009). Brain tissue segmentation of fetal MR images. In *Proc. Workshop on Image Analysis for Developing Brain, MICCAI'09*, pages 1–9.

Baffero, G. M., Crovetto, F., Fabietti, I., Boito, S., Fogliani, R., Fumagalli, M., Triulzi, F., Mosca, F., Fedele, L., and Persico, N. (2015). Prenatal ultrasound predictors of postnatal major cerebral abnormalities in fetuses with apparently isolated mild ventriculomegaly. *Prenatal Diagnosis*, 35(8):783–788.

Bai, W., Shi, W., Ledig, C., and Rueckert, D. (2015). Multi-atlas segmentation with augmented features for cardiac MR images. *Medical Image Analysis*, 19(1):98–109.

Ball, J. D., Abuhamad, A. Z., Mason, J. L., Burket, J., Katz, E., and Deutsch, S. I. (2013). Clinical outcomes of mild isolated cerebral ventriculomegaly in the presence of other neurodevelopmental risk factors. *Journal of Ultrasound in Medicine*, 32(11):1933–1938.

Banović, V., Škrablin, S., Banović, M., Radoš, M., Gverić-Ahmetašević, S., and Babić, I. (2014). Fetal brain magnetic resonance imaging and long-term neurodevelopmental impairment. *International Journal of Gynecology & Obstetrics*, 125(3):237–240.

Barkovich, A. J. (2005). Magnetic resonance techniques in the assessment of myelin and myelination. *Journal of Inherited Metabolic Disease*, 28(3):311–343.

Batchelor, P. G., Smith, A. D. C., Hill, D. L. G., Hawkes, D. J., Cox, T. C. S., and Dean, A. F. (2002). Measures of folding applied to the development of the human fetal brain. *IEEE Transactions on Medical Imaging*, 21(8):953–965.

Batty, M. J., Palaniyappan, L., Scerif, G., Groom, M. J., Liddle, E. B., Liddle, P. F., and Hollis, C. (2015). Morphological abnormalities in prefrontal surface area and thalamic volume in attention deficit/hyperactivity disorder. *Psychiatry Research: Neuroimaging*, 233(2):225–232.

Beare, R. J., Chen, J., Kelly, C. E., Alexopoulos, D., Smyser, C. D., Rogers, C. E., Loh, W. Y., Matthews, L. G., Cheong, J. L. Y., Spittle, A. J., Anderson, P. J., Doyle, L. W., Inder, T. E., Seal, M. L., and Thompson, D. K. (2016). Neonatal brain tissue classification with morphological adaptation and unified segmentation. *Frontiers in Neuroinformatics*, 10:12.

Beeghly, M., Ware, J., Soul, J., du Plessis, A., Khwaja, O., Senapati, G. M., Robson, C. D., Robertson, R. L., Poussaint, T. Y., Barnewolt, C. E., Feldman, H. A., Estroff, J. A., and Levine, D. (2010). Neurodevelopmental outcome of fetuses referred for ventriculomegaly. *Ultrasound in Obstetrics and Gynecology*, 35(4):405–416.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Bendersky, M., Musolino, P., Rugilo, C., Schuster, G., and Sica, R. (2006). Normal anatomy of the developing fetal brain. Ex vivo anatomical-magnetic resonance imaging correlation. *Journal of the Neurological Sciences*, 250(1-2):20–26.

Benkarim, O. M., Hahner, N., Piella, G., Gratacos, E., González Ballester, M. A., Eixarch, E., and Sanroma, G. (2018). Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly. *NeuroImage: Clinical*, 18:103–114.

Benkarim, O. M., Piella, G., González Ballester, M. A., and Sanroma, G. (2016). Enhanced probabilistic label fusion by estimating label confidences through discriminative learning. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 9901*, pages 505–512.

Benkarim, O. M., Piella, G., González Ballester, M. A., and Sanroma, G. (2017a). Discriminative confidence estimation for probabilistic multi-atlas label fusion. *Medical Image Analysis*, 42:274–287.

Benkarim, O. M., Piella, G., González Ballester, M. A., and Sanroma, G. (2017b). On the role of patch spaces in patch-based label fusion. In *Proc. International Workshop on Patch-based Techniques in Medical Imaging (Patch-MI), MICCAI'17, LNCS vol. 10530*, pages 37–44.

Benkarim, O. M., Radeva, P., and Igual, L. (2014). Label consistent multiclass discriminative dictionary learning for MRI segmentation. In *Proc. International Conference on Articulated Motion and Deformable Objects (AMDO), LNCS vol. 8563*, pages 138–147.

Benkarim, O. M., Sanroma, G., Zimmer, V. A., Muñoz Moreno, E., Hahner, N., Eixarch, E., Camara, O., González Ballester, M. A., and Piella, G. (2017c). Toward the automatic quantification of in utero brain development in 3D structural MRI: A review. *Human Brain Mapping*, 38(5):2772–2787.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brossard-Racine, M., du Plessis, A., Vezina, G., Robertson, R., Bulas, D., Evangelou, I., Donofrio, M., Freeman, D., and Limperopoulos, C. (2014). Prevalence and spectrum of in utero structural brain abnormalities in fetuses with complex congenital heart disease. *American Journal of Neuroradiology*, 35(8):1593–1599.

Buades, A., Coll, B., and Morel, J. M. (2005). A non-local algorithm for image denoising. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2*, pages 60–65.

Caldairou, B., Passat, N., Habas, P., Studholme, C., Koob, M., Dietemann, J., and Rousseau, F. (2011). Data-driven cortex segmentation in reconstructed fetal MRI by using structural constraints. In *Proc. International Conference on Computer Analysis of Images and Patterns (CAIP), LNCS, vol. 6854*, pages 503–511.

Cardoso, J. M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J.,

Fox, N. C., and Ourselin, S. (2013). STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. *Medical Image Analysis*, 17(6):671–684.

Cardoza, J., Goldstein, R., and Filly, R. (1988). Exclusion of fetal ventriculomegaly with a single measurement: The width of the lateral ventricular atrium. *Radiology*, 169(3):711–714.

Cash, D. M., Melbourne, A., Modat, M., Cardoso, M. J., Clarkson, M. J., Fox, N. C., and Ourselin, S. (2012). Cortical folding analysis on patients with Alzheimer's disease and mild cognitive impairment. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 7512*, pages 289–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 9901*, pages 424–432.

Cireşan, D. C., Gambardella, L. M., Giusti, A., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Proc. Advances in Neural Information Processing Systems 25 (NIPS)*, pages 2852–2860.

Claude, I., Daire, J.-L., and Sebag, G. (2004). Fetal brain MRI: segmentation and biometric analysis of the posterior fossa. *IEEE Transactions on Biomedical Engineering*, 51(4):617–626.

Clouchoux, C., du Plessis, A., Bouyssi-Kobar, M., Tworetzky, W., McElhinney, D., Brown, D., Gholipour, A., Kudelski, D., Warfield, S., McCarter, R., Robertson, R., Evans, A., Newburger, J., and Limperopoulos, C. (2013). Delayed cortical development in fetuses with complex congenital heart disease. *Cerebral Cortex*, 23(12):2932–2943.

Clouchoux, C., Kudelski, D., Gholipour, A., Warfield, S., Viseur, S., Bouyssi-Kobar, M., Mari, J., Evans, A., du Plessis, A., and Limperopoulos, C. (2012). Quantitative in vivo MRI measurement of cortical development in the fetus. *Brain Structure and Function*, 217(1):127–139.

Corbett-Detig, J., Habas, P., Scott, J., Kim, K., Rajagopalan, V., McQuillen, P., Barkovich, A., Glenn, O., and Studholme, C. (2011). 3D global and regional patterns of human fetal subplate growth determined in utero. *Brain Structure and Function*, 215(3-4):255–263.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954.

Davatzikos, C., Vaillant, M., Resnick, S., Prince, J., Letovsky, S., and Bryan, R. (1995). A computerized approach for morphological analysis of the corpus callosum. *Journal of Computer Assisted Tomography*, 20(1):88–97.

de Brébisson, A. and Montana, G. (2015). Deep neural networks for anatomical brain segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–28.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dittrich, E., Raviv, T., Kasprian, G., Donner, R., Brugger, P., Prayer, D., and Langs, G. (2014). A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation. *Medical Image Analysis*, 18(1):9–21.

Donofrio, M. and Massaro, A. (2010). Impact of congenital heart disease on brain development and neurodevelopmental outcome. *International Journal of Pediatrics*, 2010:12589–12596.

Dubois, J., Benders, M., Cachia, A., Lazeyras, F., Ha-Vinh Leuchter, R., Sizonenko, S., Borradori-Tolsa, C., Mangin, J., and Hüppi, P. (2008). Mapping the early cortical folding process in the preterm newborn brain. *Cerebral Cortex*, 18(6):1444–1454.

Dubois, J., Dehaene-Lambertz, G., Kulikova, S., Poupon, C., Hüppi, P., and Hertz-Pannier, L. (2014). The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants. *Neuroscience*, 276:48–71.

Dubuisson, M. P. and Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Proc. International Conference on Pattern Recognition, vol. 1*, pages 566–568.

Egaña-Ugrinovic, G., Sanz-Cortés, M., Couve-Pérez, C., Figueras, F., and Gratacós, E. (2014). Corpus callosum differences assessed by fetal MRI in late-onset intrauterine growth restriction and its association with neurobehavior. *Prenatal Diagnosis*, 34(9):843–849.

Egaña-Ugrinovic, G., Sanz-Cortes, M., Figueras, F., Bargalló, N., and Gratacós, E. (2013). Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction. *American Journal of Obstetrics and Gynecology*, 209(2):126.e1–126.e8.

Endres, L. and Cohen, L. (2001). Reliability and validity of three-dimensional fetal brain volumes. *Journal of Ultrasound in Medicine*, 20(12):1265–1269.

Evans, A., Collins, D., Mills, S., Brown, E., Kelly, R., and Peters, T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. *IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, 3:1813–1817.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Fernández, V., Llinares-Benadero, C., and Borrell, V. (2016). Cerebral cortex expansion and folding: what have we learned? *The EMBO Journal*, 35:1021–1044.

Fillmore, P., Richards, J., Phillips-Meek, M., Cryer, A., and Stevens, M. (2015). Stereotaxic MRI brain atlases for infants from 3 to 12 months. *Developmental Neuroscience*, 37(6):515–532.

Fonov, F., Evans, A., Botteron, K., Almli, C., McKinstry, R., and Collins, D. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-prints*.

Gaglioti, P., Danelon, D., Bontempo, S., Mombr, M., Cardaropoli, S., and Todros, T. (2005). Fetal cerebral ventriculomegaly: outcome in 176 cases. *Ultrasound in Obstetrics and Gynecology*, 25(4):372–377.

Gardosi, J. (2011). Clinical strategies for improving the detection of fetal growth restriction. *Clinics in Perinatology*, 38(1):21–31.

Garel, C. (2008). Imaging the fetus: When does MRI really help? *Pediatric Radiology*, 38(3):467–470.

Garel, C., Chantrel, E., Elmaleh, M., Brisse, H., and Sebag, G. (2003). Fetal MRI: Normal gestational landmarks for cerebral biometry, gyration and myelination. *Child's Nervous System*, 19(7-8):422–425.

Gholipour, A., Akhondi-Asl, A., Estroff, J., and Warfield, S. (2012). Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly. *NeuroImage*, 60(3):1819–1831.

Gholipour, A., Estroff, J., Barnewolt, C., Connolly, S., and Warfield, S. (2011). Fetal brain volumetry through MRI volumetric reconstruction and segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 6(3):329–339.

Gholipour, A., Estroff, J., and Warfield, S. (2010). Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain MRI. *IEEE Transactions on Medical Imaging*, 29(10):1739–1758.

Gholipour, A., Limperopoulos, C., Clancy, S., Clouchoux, C., Akhondi-Asl, A., Estroff, J., and Warfield, S. (2014). Construction of a deformable spatiotemporal MRI atlas of the fetal brain: Evaluation of similarity metrics and deformation models. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 8674*, pages 292–299.

Gilmore, J. H., Smith, L. C., Wolfe, H. M., Hertzberg, B. S., Smith, J. K., Chescheir, N. C., Evans, D. D., Kang, C., Hamer, R. M., Lin, W., and Gerig, G. (2008). Prenatal mild ventriculomegaly predicts abnormal development of the neonatal brain. *Biological Psychiatry*, 64(12):1069–1076.

Gilmore, J. H., van Tol, J. J., Streicher, H. L., Williamson, K., Cohen, S. B., Greenwood, R. S., Charles, H., Kliewer, M. A., Whitt, J., Silva, S. G., Hertzberg,

B. S., and Chescheir, N. C. (2001). Outcome in children with fetal mild ventriculomegaly: a case series. *Schizophrenia Research*, 48(2-3):219–226.

Girard, N. and Chaumoitre, K. (2012). The brain in the belly: What and how of fetal neuroimaging? *Journal of Magnetic Resonance Imaging*, 36(4):788–804.

Glastonbury, C. and Kennedy, A. (2002). Ultrafast MRI of the fetus. *Australasian Radiology*, 46(1):22–32.

Gómez-Arriaga, P., Herraiz, I., Puente, J. M., Zamora-Crespo, B., Núñez Enamorado, N., and Galindo, A. (2012). Mid-term neurodevelopmental outcome in isolated mild ventriculomegaly diagnosed in fetal life. *Fetal Diagnosis and Therapy*, 31(1):12–18.

González Ballester, M., Zisserman, A., and Brady, M. (2002). Estimation of the partial volume effect in MRI. *Medical Image Analysis*, 6(4):389–405.

González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., and Lladó, X. (2016). A review on brain structures segmentation in magnetic resonance imaging. *Artificial Intelligence in Medicine*, 73:45–69.

Gorthi, S., Akhondi-Asl, A., Thiran, J.-P., and Warfield, S. K. (2014). Optimal MAP parameters estimation in STAPLE - Learning from performance parameters versus image similarity information. In *Proc. International Workshop on Machine Learning in Medical Imaging (MLMI), MICCAI'14, LNCS vol. 8679*, pages 174–181.

Gousias, I., Edwards, A., Rutherford, M., Counsell, S., Hajnal, J., Rueckert, D., and Hammers, A. (2012). Magnetic resonance imaging of the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants. *NeuroImage*, 62(3):1499–1509.

Gousias, I., Hammers, A., Counsell, S., Srinivasan, L., Rutherford, M., Heckemann, R., and Edwards, A. (2013). Magnetic resonance imaging of the newborn brain: Automatic segmentation of brain images into 50 anatomical regions. *PLOS ONE*, 8(4):e59990.

Greenspan, H., Ruf, A., and Goldberger, J. (2006). Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1233–1245.

Griffiths, P., Reeves, M., Morris, J., Mason, G., Russell, S., Paley, M., and Whitby, E. (2010). A prospective study of fetuses with isolated ventriculomegaly investigated by antenatal sonography and in utero MR imaging. *American Journal of Neuroradiology*, 31(1):106–111.

Gui, L., Lisowski, R., Faundez, T., Hüppi, P., Lazeyras, F., and Kocher, M. (2012). Morphology-driven automatic segmentation of MR images of the neonatal brain. *Medical Image Analysis*, 16(8):1565–1579.

Guo, S., Iwabuchi, S., Balain, V., Feng, J., Liddle, P., and Palaniyappan, L. (2015). Cortical folding and the potential for prognostic neuroimaging in schizophrenia. *The British Journal of Psychiatry*, 207(5):458–459.

Habas, P., Kim, K., Corbett-Detig, J., Rousseau, F., Glenn, O., Barkovich, A., and

Studholme, C. (2010). A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain with application to segmentation. *NeuroImage*, 53(2):460–470.

Habas, P., Kim, K., Rousseau, F., Glenn, O. A., Barkovich, A. J., and Studholme, C. (2008). Atlas-based segmentation of the germinal matrix from in utero clinical MRI of the fetal brain. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 5241*, pages 351–358.

Habas, P., Scott, J., Roosta, A., Rajagopalan, V., Kim, K., Rousseau, F., Barkovich, J., Glenn, O., and Studholme, C. (2012). Early folding patterns and asymmetries of the normal human brain detected from in utero MRI. *Cerebral Cortex*, 22(1):13–25.

Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., Fan, Y., and for the Alzheimer's Disease Neuroimaging Initiative (2014). Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation. *Human Brain Mapping*, 35(6):2674–2697.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126.

Hill, J., Dierker, D., Neil, J., Inder, T., Knutsen, A., Harwell, J., Coalson, T., and Van Essen, D. (2010). A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. *Journal of Neuroscience*, 30(6):2268–2276.

Hu, H.-H., Chen, H.-Y., Hung, C.-I., Guo, W.-Y., and Wu, Y.-T. (2013). Shape and curvedness analysis of brain morphology using human fetal magnetic resonance images in utero. *Brain Structure and Function*, 218(6):1451–1462.

Huang, H., Xue, R., Zhang, J., Ren, T., Richards, L., Yarowsky, P., Miller, M., and Mori, S. (2009). Anatomical characterization of human fetal brain development with diffusion tensor magnetic resonance imaging. *Journal of Neuroscience*, 29(13):4263–4273.

Huisman, T. A. G. M., Tekes, A., and Poretti, A. (2012). Brain malformations and fetal ventriculomegaly: What to look for? *Journal of Pediatric Neuroradiology*, 1(3):185–195.

Iglesias, J. E. and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219.

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M. A., and van Ginneken, B. (2009). Multi-atlas-based segmentation with local decision fusion: Application to cardiac and aortic segmentation in ct scans. *IEEE Transactions on Medical Imaging*, 28(7):1000–1010.

ISUOG Guidelines (2007). Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. *Ultrasound in Obstetrics and Gynecology*, 29(1):109–116.

Išgum, I., Benders, M. J., Avants, B., Cardoso, M. J., Counsell, S. J., Gomez, E. F., Gui, L., Hppi, P. S., Kersbergen, K. J., Makropoulos, A., Melbourne, A., Moeskops, P., Mol, C. P., Kuklisova-Murgasova, M., Rueckert, D., Schnabel, J. A., Srhoj-Egekher, V., Wu, J., Wang, S., de Vries, L. S., and Viergever, M. A. (2015). Evaluation of automatic neonatal brain segmentation algorithms: The NeoBrainS12 challenge. *Medical Image Analysis*, 20(1):135–151.

Jakab, A., Schwartz, E., Kasprian, G., Gruber, G. M., Prayer, D., Schöpf, V., and Langs, G. (2014). Fetal functional imaging portrays heterogeneous development of emerging human brain networks. *Frontiers in Human Neuroscience*, 8:852.

Jiang, S., Xue, H., Glover, A., Rutherford, M., Rueckert, D., and Hajnal, J. (2007). MRI of moving subjects using multislice snapshot images with volume reconstruction (SVR): Application to fetal, neonatal, and adult brain studies. *IEEE Transactions on Medical Imaging*, 26(7):967–980.

Joshi, S., Davis, B., Jomier, M., and Gerig, G. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:151–160.

Jou, R. J., Hardan, A. Y., and Keshavan, M. S. (2005). Reduced cortical folding in individuals at high risk for schizophrenia: a pilot study. *Schizophrenia Research*, 75(2-3):309–313.

Kainz, B., Steinberger, M., Wein, W., Kuklisova-Murgasova, M., Malamateniou, C., Keraudren, K., Torsney-Weir, T., Rutherford, M., Aljabar, P., Hajnal, J., and Rueckert, D. (2015). Fast volume reconstruction from motion corrupted stacks of 2D slices. *IEEE Transactions on Medical Imaging*, 34(9):1901–1913.

Kazan-Tannus, J., Dialani, V., Kataoka, M., Chiang, G., Feldman, H., Brown, J., and Levine, D. (2007). MR volumetry of brain and CSF in fetuses referred for ventriculomegaly. *American Journal of Roentgenology*, 189(1):145–151.

Kazemi, K., Moghaddam, H., Grebe, R., Gondry-Jouet, C., and Wallois, F. (2007). A neonatal atlas template for spatial normalization of whole-brain magnetic resonance images of newborns: Preliminary results. *NeuroImage*, 37(2):463–473.

Kelly, E. N., Allen, V. M., Seaward, G., Windrim, R., and Ryan, G. (2001). Mild ventriculomegaly in the fetus, natural history, associated findings and outcome of isolated mild ventriculomegaly: a literature review. *Prenatal Diagnosis*, 21(8):697–700.

Keraudren, K., Kuklisova-Murgasova, M., Kyriakopoulou, V., Malamateniou, C., Rutherford, M., Kainz, B., Hajnal, J., and Rueckert, D. (2014). Automated fetal brain segmentation from 2D MRI slices for motion correction. *NeuroImage*, 101:633–643.

Kim, K., Habas, P., Rousseau, F., Glenn, O., Barkovich, A., and Studholme, C. (2010). Intersection based motion correction of multislice MRI for 3-D in utero fetal brain image formation. *IEEE Transactions on Medical Imaging*, 29(1):146–158.

Klein, A., Mensh, B., Ghosh, S., Tourville, J., and Hirsch, J. (2005). Mindboggle: Automated brain labeling with multiple atlases. *BMC Medical Imaging*, 5(1):7.

Klöppel, S., Abdulkadir, A., Jack, C. R., Koutsouleris, N., Mourão Miranda, J., and Vemuri, P. (2012). Diagnostic neuroimaging across diseases. *NeuroImage*, 61(2):457–463.

Koch, L., Wright, R., Vatansever, D., Kyriakopoulou, V., Malamateniou, C., Patkee, P., Rutherford, M., Hajnal, J., Aljabar, P., and Rueckert, D. (2014). Graph-based label propagation in fetal brain MR images. In *Proc. International Workshop on Machine Learning in Medical Imaging (MLMI), MICCAI'14, LNCS vol. 8679*, pages 9–16.

Koenderink, J. J. and van Doorn, A. J. (1992). Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564.

Kyriakopoulou, V., Vatansever, D., Elkommos, S., Dawson, S., McGuinness, A., Allsop, J., Molnár, Z., Hajnal, J., and Rutherford, M. (2014). Cortical overgrowth in fetuses with isolated ventriculomegaly. *Cerebral Cortex*, 24(8):2141–2150.

Ledig, C., Wright, R., Serag, A., Aljabar, P., and Rueckert, D. (2012). Neonatal brain segmentation using second order neighbourhood information. In *Workshop on Perinatal and Paediatric Imaging (PaPI), MICCAI'14*, pages 33–40.

Leemput, K. V., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based tissue classification of mr images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908.

Lefèvre, J., Germanaud, D., Dubois, J., Rousseau, F., de Macedo Santos, I., Angleys, H., Mangin, J.-F., Höppi, P. S., Girard, N., and De Guio, F. (2015). Are developmental trajectories of cortical folding comparable between cross-sectional datasets of fetuses and preterm newborns? *Cerebral Cortex*, 26(7):3023–3035.

Leitner, Y., Stolar, O., Rotstein, M., Toledano, H., Harel, S., Bitchonsky, O., Ben-Adani, L., Miller, E., and Ben-Sira, L. (2009). The neurocognitive outcome of mild isolated fetal ventriculomegaly verified by prenatal magnetic resonance imaging. *American Journal of Obstetrics and Gynecology*, 201(2):215.e1–215.e6.

Li, G., Nie, J., Wang, L., Shi, F., Lin, W., Gilmore, J. H., and Shen, D. (2013). Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age. *Cerebral Cortex*, 23(11):2724–2733.

Li, G., Wang, L., Shi, F., Gilmore, J. H., Lin, W., and Shen, D. (2015). Construction of 4D high-definition cortical surface atlases of infants: Methods and applications. *Medical Image Analysis*, 25(1):22–36.

Li, Y., Estroff, J. A., Mehta, T. S., Robertson, R. L., Robson, C. D., Poussaint, T. Y., Feldman, H. A., Ware, J., and Levine, D. (2011). Ultrasound and MRI of fetuses with ventriculomegaly: Can cortical development be used to predict postnatal outcome? *American Journal of Roentgenology*, 196(6):1457–1467.

Limperopoulos, C., Tworetzky, W., McElhinney, D., Newburger, J., Brown, D., Robertson, R., Guizard, N., McGrath, E., Geva, J., Annese, D., Dunbar-Masterson, C., Trainor, B., Laussen, P., and Du Plessis, A. (2010). Brain volume and metabolism in fetuses with congenital heart disease: Evaluation with quantitative magnetic resonance imaging and spectroscopy. *Circulation*,

121(1):26–33.

Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(4):163–169.

Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., and Rueckert, D. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365.

Lyall, A. E., Woolson, S., Wolfe, H. M., Goldman, B. D., Reznick, J. S., Hamer, R. M., Lin, W., Styner, M., Gerig, G., and Gilmore, J. H. (2012). Prenatal isolated mild ventriculomegaly is associated with persistent ventricle enlargement at ages 1 and 2. *Early human development*, 88(8):691–698.

Lyoo, I. K., Noam, G. G., Lee, C. K., Lee, H. K., Kennedy, B. P., and Renshaw, P. F. (1996). The corpus callosum and lateral ventricles in children with attention-deficit hyperactivity disorder: A brain magnetic resonance imaging study. *Biological Psychiatry*, 40(10):1060–1063.

Makropoulos, A., Gousias, I., Ledig, C., Aljabar, P., Serag, A., Hajnal, J., Edwards, A., Counsell, S., and Rueckert, D. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Transactions on Medical Imaging*, 33(9):1818–1831.

Malamateniou, C., Malik, S., Counsell, S., Allsop, J., McGuinness, A., Hayat, T., Broadhouse, K., Nunes, R., Ederies, A., Hajnal, J., and Rutherford, M. (2013). Motion-compensation techniques in neonatal and fetal MR imaging. *American Journal of Neuroradiology*, 34(6):1124–1136.

Manjón, J. V. and Coupé, P. (2017). Hippocampus subfield segmentation using a patch-based boosted ensemble of autocontext neural networks. In *Proc. International Workshop on Patch-based Techniques in Medical Imaging (Patch-MI), MICCAI'17, LNCS vol. 10530*, pages 29–36.

Mazziotta, J., Toga, A., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development: The International Consortium for Brain Mapping (ICBM). *NeuroImage*, 2(2, Part A):89–101.

McQuillen, P., Goff, D., and Licht, D. (2010). Effects of congenital heart disease on brain development. *Progress in Pediatric Cardiology*, 29(2):79–85.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Melbourne, A., Kendall, G., Cardoso, M., Gunny, R., Robertson, N., Marlow, N., and Ourselin, S. (2014). Preterm birth affects the developmental synergy between cortical folding and cortical connectivity observed on multimodal MRI. *NeuroImage*, 89(0):23–34.

Melchiorre, K., Bhide, A., Gika, A. D., Pilu, G., and Papageorghiou, A. T. (2009).

Counseling in isolated mild fetal ventriculomegaly. *Ultrasound in Obstetrics and Gynecology*, 34(2):212–224.

Mlczoch, E., Brugger, P., Ulm, B., Novak, A., Frantal, S., Prayer, D., and Salzer-Muhar, U. (2013). Structural congenital brain disease in congenital heart disease: Results from a fetal MRI program. *European Journal of Paediatric Neurology*, 17(2):153–160.

Moeskops, P., Benders, M. J. N. L., Kersbergen, K. J., Groenendaal, F., de Vries, L. S., Viergever, M. A., and Igum, I. (2015). Development of cortical morphology evaluated with longitudinal mr brain images of preterm infants. *PLOS ONE*, 10(7):1–22.

Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Toga, A. W., and Thompson, P. M. (2010). Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation. *IEEE Transactions on Medical Imaging*, 29(1):30–43.

Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S., Doria, V., Serag, A., Gousias, I., Boardman, J., Rutherford, M., Edwards, A., Hajnal, J., and Rueckert, D. (2011). A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763.

Murgasova, M., Quaghebeur, G., Rutherford, M., Hajnal, J., and Schnabel, J. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical Image Analysis*, 16(8):1550–1564.

Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244.

Nie, J., Guo, L., Li, G., Faraco, C., Miller, L. S., and Liu, T. (2010). A computational model of cerebral cortex folding. *Journal of Theoretical Biology*, 264(2):467–478.

Nie, J., Li, G., Wang, L., Gilmore, J. H., Lin, W., and Shen, D. (2011). A computational growth model for measuring dynamic cortical development in the first year of life. *Cerebral Cortex*, 22(10):2272–2284.

Nyul, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150.

Ortiz, A., Górriz, J., Ramírez, J., and Salas-González, D. (2013). Improving MRI segmentation with probabilistic GHSOM and multiobjective optimization. *Neurocomputing*, 114(0):118–131.

Ouahba, J., Luton, D., Vuillard, E., Garel, C., Gressens, P., Blanc, N., Elmaleh, M., Evrard, P., and Oury, J. (2006). Prenatal isolated mild ventriculomegaly: outcome in 167 cases. *BJOG: An International Journal of Obstetrics & Gynaecology*, 113(9):1072–1079.

Padilla, N., Alexandrou, G., Blennow, M., Lagercrantz, H., and Ådén, U. (2014). Brain growth gains and losses in extremely preterm infants at term. *Cerebral Cortex*, 25(7):1897–1905.

Park, H., Bland, P., Hero III, A., and Meyer, C. (2005). Least biased target

selection in probabilistic atlas construction. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 3750*, pages 419–426.

Paus, T., Zijdenbos, A., Worsley, K., Collins, D., Blumenthal, J., Giedd, J., Rapoport, J., and Evans, A. (1999). Structural maturation of neural pathways in children and adolescents: In vivo study. *Science*, 283(5409):1908–1911.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pham, D. and Prince, J. (1999). Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging*, 18(9):737–752.

Pienaar, R., Fischl, B., Caviness, V., Makris, N., and Grant, P. (2008). A methodology for analyzing curvature in the developing brain from preterm to adult. *International Journal of Imaging Systems and Technolog*, 18(1):42–68.

Powell, S., Magnotta, V. A., Johnson, H., Jammalamadaka, V. K., Pierson, R., and Andreasen, N. C. (2008). Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage*, 39(1):238–247.

Powell, S. B. (2010). Models of neurodevelopmental abnormalities in schizophrenia. In *Proc. Behavioral Neurobiology of Schizophrenia and Its Treatment, CTBN vol. 4*, pages 435–481.

Prastawa, M., Gilmore, J., Lin, W., and Gerig, G. (2005). Automatic segmentation of MR images of the developing newborn brain. *Medical Image Analysis*, 9(5):457–466.

Prayer, D., Kasprian, G., Krampl, E., Ulm, B., Witzani, L., Prayer, L., and Brugger, P. (2006). MRI of normal fetal brain development. *European Journal of Radiology*, 57(2):199–216.

Rajagopalan, V., Scott, J., Habas, P., Kim, K., Corbett-Detig, J., Rousseau, F., Barkovich, A., Glenn, O., and Studholme, C. (2011a). Local tissue growth patterns underlying normal fetal human brain gyrification quantified in utero. *Journal of Neuroscience*, 31(8):2878–2887.

Rajagopalan, V., Scott, J., Habas, P., Kim, K., Rousseau, F., Glenn, O., Barkovich, A., and Studholme, C. (2011b). Spatiotemporal morphometry of adjacent tissue layers with application to the study of sulcal formation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 6892*, pages 476–483.

Rajagopalan, V., Scott, J., Habas, P., Kim, K., Rousseau, F., Glenn, O., Barkovich, A., and Studholme, C. (2012). Mapping directionality specific volume changes using tensor based morphometry: An application to the study of gyrogenesis and lateralization of the human fetal brain. *NeuroImage*, 63(2):947–958.

Rehn, A. E. and Rees, S. M. (2005). Investigating the neurodevelopmental hypothesis of schizophrenia. *Clinical and Experimental Pharmacology and Physiology*, 32(9):687–696.

Robinson, H. P. and Fleming, J. E. E. (1975). A critical evaluation of sonar crown-rump length measurements. *BJOG: An International Journal of Obstetrics & Gynaecology*, 82(9):702–710.

Rodríguez, M., Martínez, F., Ten, P., Pedregosa, J., Fernández-Mayoralas, D., Jiménez, M., and de la Peña, M. (2010). Fetal MRI in CNS abnormalities. Relevant issues for obstetricians. *Revista Argentina de Radiología*, 74(4):385–396.

Rodriguez-Carranza, C., Mukherjee, P., Vigneron, D., Barkovich, J., and Studholme, C. (2008). A framework for in vivo quantification of regional brain folding in premature neonates. *NeuroImage*, 41(2):462–478.

Roelfsema, N., Hop, W., Boito, S., and Wladimiroff, J. (2004). Three-dimensional sonographic measurement of normal fetal brain volume during the second half of pregnancy. *American Journal of Obstetrics and Gynecology*, 190(1):275–280.

Rohlfing, T., Brandt, R., Menzel, R., and Maurer, C. R. (2004). Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 23(8):983–994.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Rousseau, F., Habas, P. A., and Studholme, C. (2011). A supervised patch-based approach for human brain labeling. *IEEE Transactions on Medical Imaging*, 30(10):1852–1862.

Rutherford, M. (2001). *MRI of the neonatal brain*. www.mrineonatalbrain.com.

Rutherford, M., Jiang, S., Allsop, J., Perkins, L., Srinivasan, L., Hayat, T., Kumar, S., and Hajnal, J. (2008). MR imaging methods for assessing fetal brain development. *Developmental Neurobiology*, 68(6):700–711.

Sabuncu, M. R., Yeo, B. T. T., Leemput, K. V., Fischl, B., and Golland, P. (2010). A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729.

Sadan, S., Malinger, G., Schweiger, A., Lev, D., and Lerman-Sagie, T. (2007). Neuropsychological outcome of children with asymmetric ventricles or unilateral mild ventriculomegaly identified in utero. *BJOG: An International Journal of Obstetrics & Gynaecology*, 114(5):596–602.

Salomon, L. J., Bernard, J. P., and Ville, Y. (2007). Reference ranges for fetal ventricular width: a non-normal approach. *Ultrasound in Obstetrics and Gynecology*, 30(1):61–66.

Sanroma, G., Benkarim, O. M., Piella, G., and González Ballester, M. Á. (2016a). Building an ensemble of complementary segmentation methods by exploiting probabilistic estimates. In *Proc. International Workshop on Machine Learning in Medical Imaging (MLMI), MICCAI'16, LNCS vol. 10019*, pages 27–35.

Sanroma, G., Benkarim, O. M., Piella, G., Wu, G., Zhu, X., Shen, D., and González-Ballester, M. A. (2015a). Discriminative dimensionality reduction for patch-based label fusion. In *Proc. International Workshop on Machine Learning Meets Medical Imaging (MLMMI), ICML'15, LNCS vol. 9487*, pages 94–103.

Sanroma, G., Wu, G., Gao, Y., and Shen, D. (2014). Learning to rank atlases for multiple-atlas segmentation. *IEEE Transactions on Medical Imaging*, 33(10):1939–1953.

Sanroma, G., Wu, G., Gao, Y., Thung, K.-H., Guo, Y., and Shen, D. (2015b). A transversal approach for patch-based label fusion via matrix completion. *Medical Image Analysis*, 24(1):135–148.

Sanroma, G., Wu, G., Kim, M., González Ballester, M. A., and Shen, D. (2016b). Chapter 11 - Multiple-Atlas Segmentation in Medical Imaging. In Zhou, S. K., editor, *Medical Image Recognition, Segmentation and Parsing*, pages 231–257. Academic Press.

Scala, C., Familiari, A., Pinas, A., Papageorghiou, A. T., Bhide, A., Thilaganathan, B., and Khalil, A. (2017). Perinatal and long-term outcomes in fetuses diagnosed with isolated unilateral ventriculomegaly: systematic review and meta-analysis. *Ultrasound in Obstetrics and Gynecology*, 49(4):450–459.

Schaer, M., Cuadra, M. B., Tamarit, L., Lazeyras, F., Eliez, S., and Thiran, J. P. (2008). A surface-based approach to quantify local cortical gyrification. *IEEE Transactions on Medical Imaging*, 27(2):161–170.

Scott, J., Habas, P., Kim, K., Rajagopalan, V., Hamzelou, K., Corbett-Detig, J., Barkovich, A., Glenn, O., and Studholme, C. (2011). Growth trajectories of the human fetal brain tissues estimated from 3D reconstructed in utero MRI. *International Journal of Developmental Neuroscience*, 29(5):529–536.

Scott, J., Habas, P., Rajagopalan, V., Kim, K., Barkovich, A., Glenn, O., and Studholme, C. (2013). Volumetric and surface-based 3D MRI analyses of fetal isolated mild ventriculomegaly. *Brain Structure and Function*, 218(3):645–655.

Sdika, M. (2010). Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. *Medical Image Analysis*, 14(2):219–226.

Sdika, M. (2015). Enhancing atlas based segmentation with multiclass linear classifiers. *Medical Physics*, 42(12):7169–7181.

Serag, A., Aljabar, P., Ball, G., Counsell, S., Boardman, J., Rutherford, M., Edwards, A., Hajnal, J., and Rueckert, D. (2012a). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3):2255–2265.

Serag, A., Kyriakopoulou, V., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., Aljabar, P., Counsell, S. J., Boardman, J. P., and Rueckert, D. (2012b). A multichannel 4D probabilistic atlas of the developing brain: Application to fetuses and neonates. *Annals of the BMVA*, 53(3):1–14.

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and

Leahy, R. M. (2001). Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876.

Shi, F., Fan, Y., Tang, S., Gilmore, J., Lin, W., and Shen, D. (2010). Neonatal brain image segmentation in longitudinal MRI studies. *NeuroImage*, 49(1):391–400.

Shi, F., Shen, D., Yap, P., Fan, Y., Cheng, J., An, H., Wald, L., Gerig, G., Gilmore, J., and Lin, W. (2011). CENTS: Cortical enhanced neonatal tissue segmentation. *Human Brain Mapping*, 32(3):382–396.

Shi, F., Wang, L., Wu, G., Li, G., Gilmore, J. H., Lin, W., and Shen, D. (2014). Neonatal atlas construction using sparse representation. *Human Brain Mapping*, 35(9):4663–4677.

Shimizu, Y., Yoshimoto, J., Toki, S., Takamura, M., Yoshimura, S., Okamoto, Y., Yamawaki, S., and Doya, K. (2015). Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group lasso. *PLOS ONE*, 10(5):1–23.

Shimony, J. S., Smyser, C. D., Wideman, G., Alexopoulos, D., Hill, J., Harwell, J., Dierker, D., Van Essen, D. C., Inder, T. E., and Neil, J. J. (2016). Comparison of cortical folding measures for evaluation of developing human brain. *NeuroImage*, 125:780–790.

Shyu, K., Wu, Y., Chen, T., Chen, H., Hu, H., and Guo, W. (2010). Analysis of fetal cortical complexity from MR images using 3D entropy based information fractal dimension. *Nonlinear Dynamics*, 61(3):363–372.

Studholme, C. (2011). Mapping fetal brain development in utero using MRI: The big bang of brain mapping. *Annual Review of Biomedical Engineering*, 13:345–368.

Studholme, C. and Rousseau, F. (2014). Quantifying and modelling tissue maturation in the living human fetal brain. *International Journal of Developmental Neuroscience*, 32(0):3–10.

Tallinen, T., Chung, J. Y., Rousseau, F., Girard, N., Lefevre, J., and Mahadevan, L. (2016). On the growth and form of cortical convolutions. *Nature Physics*, 12(6):588–593.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tong, T., Wolz, R., Coup, P., Hajnal, J. V., and Rueckert, D. (2013). Segmentation of mr images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. *NeuroImage*, 76:11–23.

Twickler, D., Magee, K., Caire, J., Zaretsky, M., Fleckenstein, J., and Ramus, R. (2003). Second-opinion magnetic resonance imaging for suspected fetal central nervous system abnormalities. *American Journal of Obstetrics and Gynecology*, 188(2):492–496.

Van Essen, D. C. and Drury, H. A. (1997). Structural and functional analyses of human cerebral cortex using a surface-based atlas. *Journal of Neuroscience*,

17(18):7079–7102.

van Rikxoort, E. M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M. A., Pluim, J. P., and van Ginneken, B. (2010). Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. *Medical Image Analysis*, 14(1):39–49.

Vita, A., Dieci, M., Silenzi, C., Tenconi, F., Giobbio, G. M., and Invernizzi, G. (2000). Cerebral ventricular enlargement as a generalized feature of schizophrenia: a distribution analysis on 502 subjects. *Schizophrenia Research*, 44(1):25–34.

Wan, J., Carass, A., Resnick, S. M., and Prince, J. L. (2008). Automated reliable labeling of the cortical surface. *Proceedings. IEEE International Symposium on Biomedical Imaging*, 2008:440.

Wang, B., Mezlini, A. M., Demir, F., and et al. (2014a). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333–337.

Wang, H., Cao, Y., and Syeda-Mahmood, T. (2014b). Multi-atlas segmentation with learning-based label fusion. In *Proc. International Workshop on Machine Learning in Medical Imaging (MLMI), MICCAI'14, LNCS vol. 8679*, pages 256–263.

Wang, H., Das, S. R., Suh, J. W., Altinay, M., Pluta, J., Craige, C., Avants, B., and Yushkevich, P. A. (2011a). A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55(3):968–985.

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623.

Wang, L., Gao, Y., Shi, F., Li, G., Gilmore, J. H., Lin, W., and Shen, D. (2015). Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage*, 108:160–172.

Wang, L., Shi, F., Li, G., Gao, Y., Lin, W., Gilmore, J. H., and Shen, D. (2014c). Segmentation of neonatal brain MR images using patch-driven level sets. *NeuroImage*, 84:141–158.

Wang, L., Shi, F., Lin, W., Gilmore, J., and Shen, D. (2011b). Automatic segmentation of neonatal images using convex optimization and coupled level sets. *NeuroImage*, 58(3):805–817.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921.

Weisenfeld, N. and Warfield, S. (2009). Automatic segmentation of newborn brain

MRI. *NeuroImage*, 47(2):564–572.

Wilke, M., Holland, S., Altaye, M., and Gaser, C. (2008). Template-O-Matic: A toolbox for creating customized pediatric templates. *NeuroImage*, 41(3):903–913.

Wilke, M., Schmithorst, V., and Holland, S. (2003). Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. *Magnetic Resonance in Medicine*, 50(4):749–757.

Wolosin, S. M., Richardson, M. E., Hennessey, J. G., Denckla, M. B., and Mostofsky, S. H. (2009). Abnormal cerebral cortex structure in children with ADHD. *Human Brain Mapping*, 30(1):175–184.

Wright, I. C., Rabe-Hesketh, S., Woodruff, P. W., David, A. S., Murray, R. M., and Bullmore, E. T. (2000). Meta-analysis of regional brain volumes in schizophrenia. *American Journal of Psychiatry*, 157(1):16–25.

Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M., Hajnal, J., Rueckert, D., and Aljabar, P. (2014). Automatic quantification of normal cortical folding patterns from fetal brain MRI. *NeuroImage*, 91(0):21–32.

Wright, R., Makropoulos, A., Kyriakopoulou, V., Patkee, P., Koch, L., Rutherford, M., Hajnal, J., Rueckert, D., and Aljabar, P. (2015). Construction of a fetal spatio-temporal cortical surface atlas from in utero MRI: Application of spectral surface matching. *NeuroImage*, 120:467–480.

Wu, G., Kim, M., Sanroma, G., Wang, Q., Munsell, B. C., and Shen, D. (2015a). Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. *NeuroImage*, 106:34–46.

Wu, J., Awate, S., Licht, D., Clouchoux, C., du Plessis, A., Avants, B., Vossough, A., Gee, J., and Limperopoulos, C. (2015b). Assessment of mri-based automated fetal cerebral cortical folding measures in prediction of gestational age in the third trimester. *American Journal of Neuroradiology*, 36(7):1369–1374.

Xia, J., Zhang, C., Wang, F., Benkarim, O. M., Sanroma, G., Piella, G., González Ballester, M. A., Hahner, N., Eixarch, E., Shen, D., and Li, G. (2018). Fetal cortical parcellation based on growth patterns. In *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 696–699.

Xue, H., Srinivasan, L., Jiang, S., Rutherford, M., Edwards, A., Rueckert, D., and Hajnal, J. (2007). Automatic segmentation and reconstruction of the cortex from neonatal MRI. *NeuroImage*, 38(3):461–477.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, D., Guo, Q., Wu, G., and Shen, D. (2012). Sparse patch-based label fusion for multi-atlas segmentation. In *Proc. International Workshop on Multimodal Brain Image Analysis (MBIA), MICCAI'12, LNCS vol. 7509*, pages 94–102.

Zhang, Y., Shi, F., Wu, G., Wang, L., Yap, P. T., and Shen, D. (2016). Consistent spatial-temporal longitudinal atlas construction for developing infant brains.

IEEE Transactions on Medical Imaging, 35(12):2568–2577.

Zhang, Y., Zhou, Y., Yu, C., Lin, L., Li, C., and Jiang, T. (2010). Reduced cortical folding in mental retardation. *American Journal of Neuroradiology*, 31(6):1063–1067.

Zikic, D., Glocker, B., and Criminisi, A. (2013). Atlas encoding by randomized forests for efficient label propagation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 8151*, pages 66–73.

Zilles, K., Armstrong, E., Schleicher, A., and Kretschmann, H.-J. (1988). The human pattern of gyrification in the cerebral cortex. *Anatomy and Embryology*, 179(2):173–179.

Publications

Journal papers

1. **Benkarim O. M.**, Piella G., Hahner N., Eixarch E., González Ballester M. A., and Sanroma G. Patch spaces and fusion strategies in patch-based label fusion. [Under review.]
2. Sanroma G., **Benkarim O. M.**, Piella G., Lekadir K., Hahner N., Eixarch E., and González Ballester M. A. Learning to combine complementary segmentation methods for fetal and 6-month infant brain MRI segmentation. [Under review.]
3. Sanroma G., **Benkarim O. M.**, Piella G., Camara O., Wu G., Shen D., Gispert J. D., Molinuevo J. L., and González Ballester M. A. (2018). Learning non-linear patch embeddings with neural networks for label fusion. *Medical Image Analysis*, 44:143-155.
4. **Benkarim O. M.**, Hahner N., Piella G., Gratacos E., González Ballester M. A., Eixarch E., and Sanroma G. (2018). Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly. *NeuroImage: Clinical*, 18:103-114.
5. **Benkarim O. M.**, Piella G., González Ballester M. A., and Sanroma G. (2017). Discriminative confidence estimation for probabilistic multi-atlas label fusion. *Medical Image Analysis*, 42:274-287.
6. **Benkarim O. M.**, Sanroma G., Zimmer V. A., Muñoz-Moreno E., Hahner N., Eixarch E., Camara O., González Ballester M. A., and Piella G. (2017). Toward the automatic quantification of in utero brain development in 3D structural MRI: A review. *Human Brain Mapping*, 38:2772-2787.

Conference papers

1. **Benkarim O. M.**, Sanroma G., Piella G., Rekik I., Hahner N., Eixarch E., González Ballester M. A., Shen D., and Li G. Revealing regional associations of cortical folding alterations with in utero ventricular dilation using joint spectral embedding. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. [Accepted.]

2. Xia J., Zhang C., Wang F., **Benkarim O. M.**, Sanroma G., Piella G., González Ballester M. A., Hahner N., Eixarch E., Shen D., and Li G. (2018). Fetal cortical parcellation based on growth patterns. In *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 696-699.

3. Sanroma G., Andrea V., **Benkarim O. M.**, Manjón J. V., Coupé P., Camara O., Piella G., and González Ballester M. A. (2017). Early prediction of Alzheimer's disease with non-local patch-based longitudinal descriptors. In *Proc. International Workshop on Patch-based Techniques in Medical Imaging (Patch-MI), MICCAI'17, LNCS vol. 10530*, pages 74-81.

4. **Benkarim O. M.**, Piella G., González Ballester M. A., and Sanroma G. (2017). On the role of patch spaces in patch-based label fusion. In *Proc. International Workshop on Patch-based Techniques in Medical Imaging (Patch-MI), MICCAI'17, LNCS vol. 10530*, pages 37-44.

5. Sanroma G., **Benkarim O. M.**, Piella G., and González Ballester M. A. (2016). Building an ensemble of complementary segmentation methods by exploiting probabilistic estimates. In *Proc. International Workshop on Machine Learning in Medical Imaging (MLMI), MICCAI16, LNCS vol. 10019*, pages 27-35.

6. **Benkarim O. M.**, Piella G., González Ballester M. A., and Sanroma G. (2016). Enhanced probabilistic label fusion by estimating label confidences through discriminative learning. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 9901*, pages 505-512.

7. Sanroma G., **Benkarim O. M.**, Piella G., Wu G., Zhu X., Shen D., and González Ballester M. A. (2015). Discriminative dimensionality reduction for patch-based label fusion. In *Proc. International Workshop on Machine Learning Meets Medical Imaging (MLMMI), ICML'15, LNCS vol. 9487*, pages 94-103.