Imperial College of Science, Technology and Medicine
Department of Computing

Reconstruction, Localisation, and Segmentation in Medical Images

Amir Alansary

A dissertation submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy and Diploma
of
Imperial College London

October, 2018

Abstract

Medical imaging plays a crucial role for identifying a medical condition, therapeutic response, disease level or the chance of developing a future complication. This thesis proposes robust and fully-automatic medical imaging methods that can assist doctors in their clinical decision making. In particular, it contributes to the main components of the medical imaging pipeline: image acquisition, reconstruction, and analysis. A novel approach is proposed for motion correction and reconstruction in foetal MRI. Whereas scanned images suffer from inter-slice motion artefacts because of the unconstrained foetal movements. The proposed method explores a new reconstruction paradigm that splits the 3D input image into overlapping patches.This enables conventional reconstruction methods to compensate for non-rigid deformations and reconstruct the whole uterus. Another new method is proposed for automatic foetal brain localisation, which is essential for assessing the brain development and maturation in foetal MRI.The proposed method uses superpixel graphs to train a random forest classifier, which allows to efficiently extract 3D global features instead of 2D local features. Results show high brain detection accuracy, without the need of any landmarks or prior information such as the gestational age of the foetus. Another method is also presented to automatically segment the placenta, which is typically used for the evaluation of the placenta during both normal and high-risk pregnancies. Automatic segmentation of the placenta is more challenging than brain segmentation because of its variability in position, shape and appearance. The proposed method adopts a multi-pathway convolutional neural network to efficiently captures both local and global contextual information. The predicted output is then refined using a condition random field to regularise outliers. This thesis also proposes and evaluates reinforcement learning based methods for the detection of anatomical landmarks, which is an important step for a wide range of applications in medical image analysis. Manual annotation of such landmarks is a tedious task and prone to observer errors. Fixed-scale and multi-scale search strategies are evaluated on three different datasets: foetal ultrasound, brain and cardiac MRI. The evaluation results outperform current supervised and other reinforcement learning state-of-the-art methods. Finally, a novel approach is proposed for automatic view planning using intelligent artificial agents to mimic the navigation process towards the target plane. Similar to the landmark agents, view planning agents learn to find the target plane iteratively using reinforcement learning. The experiments show that view planning agents can detect different standard planes from brain and cardiac MRI with state-of-the-art accuracy in a very short time.

To my beloved family

©The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Declaration of Originality

I, Amir Alansary, hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Acknowledgements

I would like to express my sincere gratitude and deep appreciation to my supervisors and colleagues from the biomedical imaging analysis group (BioMedIA) at Imperial College London. Their incredible support and continuous encouragement have greatly influenced and inspired me during my PhD journey.

In particular, I want to thank my esteemed supervisor, Daniel Rueckert, for his invaluable guidance and endorsement, and giving me the opportunity to work under his supervision in the first place. His constant support and precious suggestions, that are based on his broad and deep knowledge in medical imaging analysis, have helped me solve difficult research problems and publish high quality research. I would like also to thank Ben Glocker and Bernhard Kainz for their extraordinary helpful discussions and exceptional contribution to my research. I really appreciate all the collaborators and co-authors, who enabled and supported my work presented in this thesis. Many thanks to Miguel A. Gonzàlez Ballester and Anil Bharath for their time and feedback during the examination process. I want also to thank Mohamed Elmikaty, Salim Arslan, and Jonathan Passerat-Palmbach for their time, comments, and feedback during the preparation of this thesis. Special thanks to Amani El-Kholy for the continuous administration support during the whole PhD lifetime. I would like also to express my gratitude to the President's PhD Scholarship at Imperial College London for funding my research.

Last but not least, I am extremely grateful to my family, my parents Mohamed Yehia and Soheer Anwar, and my siblings Aya, Ahmed and Omay, for their support during my whole journey since my first steps tackling the research career. I am deeply grateful for having them in my life, and without them I would not be what I am.

Contents

Abstract **i**

Declaration of Originality **iv**

Acknowledgements **vii**

1 Introduction **1**

1.1 Motivation and Objectives 1

1.1.1 Foetal MRI Reconstruction, Localisation, and Segmentation 2

1.1.2 Landmark Detection and View Planning 3

1.2 Thesis Outline and Summary of Contributions 4

1.3 Publications 8

2 Background and Related Work **11**

2.1 Super-resolution Reconstruction in Foetal MRI 11

2.1.1 Background 11

2.1.2 Image Super-resolution 14

2.1.3 Related Work 15

2.2 Foetal Brain Localisation and Placenta Segmentation 19

2.2.1 Background 19

2.2.2 Machine Learning in Medical Images 21

2.2.3 Medical Image Classification 25

2.2.4 Deep Learning in Medical Images 32

2.2.5 Related Work 39

2.3 Anatomical Landmarks Detection and View Planning 44

2.3.1 Background 44

2.3.2 Reinforcement Learning for Medical Image Analysis 46

2.3.3 Related Work 53

3 Motion Correction and Reconstruction 56

3.1 Introduction 57

3.2 Method 58

3.2.1 Input Data and Initialisation 58

3.2.2 Patch-to-Volume Registration 63

3.2.3 Super-resolution 65

3.2.4 EM-Based Outlier Removal 65

3.3 Implementation 67

3.4 Evaluation and Experiments 69

3.4.1 Evaluation of Global Image Quality 69

3.4.2 Evaluation of Local Deformations 73

3.5 Results 74

3.5.1 Global Image Quality 74

3.5.2 Reconstruction of Foetal Organs 74

3.5.3 Foetal Brain Reconstruction at Late Gestation 79

3.5.4 Multi-scale Reconstruction 82

3.5.5 Local Deformations 83

3.5.6 Performance Analysis 84

3.6 Discussion 85

3.7 Conclusion 86

4 Foetal Brain Localisation and Placenta Segmentation 88

4.1 Foetal Brain Localisation 90

4.1.1 Introduction 90

4.1.2 Method 91

4.1.3 Experiments and Results 96

4.2 Placenta Segmentation 102

4.2.1 Introduction 102

4.2.2 Method 103

4.2.3 Experiments and Results 106

4.3 Conclusion 110

5 Anatomical Landmark Detection and View Planning 112

5.1 Introduction . . . 113

5.2 Anatomical Landmark Detection . . . 114

5.2.1 Method . . . 115

5.2.2 Experiments and Results . . . 119

5.3 Automatic View Planning . . . 127

5.3.1 Method . . . 128

5.3.2 Experiments and Results . . . 130

5.4 Discussion and Conclusion . . . 134

6 Conclusions 136

6.1 Thesis Summary . . . 136

6.1.1 Motion Correction and Reconstruction in Foetal MRI . . . 136

6.1.2 Foetal Brain Localisation . . . 137

6.1.3 Automatic Placenta Segmentation . . . 138

6.1.4 Anatomical Landmark Detection . . . 139

6.1.5 Automatic View Planning . . . 140

6.2 Challenges and Future Directions . . . 141

6.2.1 Deep Learning for Image Reconstruction . . . 141

6.2.2 Learning from Small Data . . . 141

6.2.3 Interpretability and Uncertainty . . . 142

6.2.4 Active Learning . . . 143

Bibliography **143**

List of Tables

3.1 Average PSNR results of different PVR variants 79

3.2 Average SSIM results of different PVR variants 79

3.3 Comparison between fixed-scale and multi-scale superpixels reconstruction . . . 82

3.4 Distance error results based on anatomical landmarks after PVR 84

3.5 Displacement errors resulting from non-rigid registration between different reconstructions and the motion-free adult brain phantom ground truth 84

4.1 Comparison between template matching and feature learning methods for foetal MRI brain localisation 91

4.2 Automatic foetal brain localisation evaluation results 100

5.1 Comparison between the proposed RL-agents and stat-of-the-art RL and supervised methods for detecting the cavum septum pellucidum in foetal US 121

5.2 Performance of different fixed and multi scale DQN agents for landmark detection 123

5.3 Performance of different fixed and multi scale DQN agents for detecting anterior and posterior commissure landmarks in brain MR images 124

5.4 Results of inter-observer errors, stratified decision forests, and RL-agents for the detection of apex and centre of mitral valve landmarks in cardiac MRI 126

5.5 The performance of multi-scale RL agents for automatic view planning from brain and cardiac MR images 132

5.6 The performance of multi-scale RL agents for automatic view planning from brain MRI scans 134

List of Figures

1.2 Typical medical imaging analysis pipeline 5

2.1 Motion artefacts of the whole uterus in a foetal MRI scan 13

2.2 Schematic diagram of image superresolution 14

2.4 A 3D visualisation of the placenta from foetal MRI 21

2.5 The human placenta from an MRI scan 21

2.7 Conventional image classification pipeline 25

2.8 Scale-invariant features (SIFT) 27

2.9 Support vector machine (SVM) 30

2.10 Decision trees and random forests 31

2.17 A template-to-slice matching for foetal brain localisation 42

2.18 Published works based on reinforcement learning for medical image analysis . . 47

2.20 DQN architecture for playing Atari games 50

2.21 Duel DQN architecture 52

3.1 Schematic overview of the PVR framework 59

3.2 Square and superpixel patches for PVR 61

3.3 Reconstructed foetal brain with approximated sinc function using Taylor series . 64

3.4 An overview of the main components of the PVR software 68

3.5 Synthetic non-rigid motion artefacts caused by skewing an adult brain phantom 75

3.6 Comparative reconstruction performance of SVR and PVR methods on synthetically corrupted MRI brain images 76

3.7 Example reconstructions of consecutive MR scans of a moving foetus 77

3.8 Three viewing planes through a motion corrupted scan of a moving twin before and after PVR 78

3.9 Qualitative results of PVR through a motion-corrupted foetal brain and placenta 80

3.10 Evaluation results of the computational overhead of different PVR variants . . . 81

4.1 Foetal brain from MRI with motion artefacts 90

4.2 Foetal brain localisation framework 92

4.3 Segmenting 2D in-plane foetal MRI into superpixels 93

4.5 Superpixel graphs for foetal brain MRI 96

4.7 A plot diagram showing the accuracy of the proposed localisation approach at different dictionary sizes 99

4.8 Foetal MRI brain automatically segmented using superpixel graphs and random forests 101

4.10 Placenta segmentation using superpixel graphs of DSIFT features and random forest 102

4.14 Dice score and Hausdorff distance errors measured for the automatically segmented placenta 108

4.16 Evaluating the volumetric results from the automatic segmentation of the placenta compared to the values from the clinical practice 109

4.18 Comparisons between the 3D reconstructed placenta using superpixel-based PVR and SVR 110

5.1 Schematic diagram of an RL scenario for landmark detection 116

5.2 DQN architecture for landmark detection 117

5.3 Annotated landmarks on foetal head ultrasound image scan 122

5.4 Annotated anterior and posterior commissure landmarks from an adult MRI brain image 123

5.5 The centre of apex and mitral valve landmarks annotated on cardiac MRI 125

5.8 RL agents for automatic view planning 129

5.9 Standard 4-chamber view in cardiac MR images annotated with ground truth anatomical landmarks 132

5.10 Standardised view planes from brain annotated with ground truth anatomical landmarks 133

Chapter 1

Introduction

1.1 Motivation and Objectives

Recent advances in medical imaging techniques and hardware platforms have provided radiologists and physicians with high-resolution visual representations of the human anatomy in terms of structure and function. This has improved several clinical tasks such as disease diagnosis, therapy planning and treatment decisions. Moreover, medical image analysis methods seek solving problems pertaining to interpretation of medical images with the purpose of improving the clinical assessment. Chapter 3 and 4 focus on foetal MRI as a challenging dataset because of the motion artefacts caused by the foetal random movements. Methods for correcting the motion, localising and segmenting foetal organs are presented and evaluated thoroughly in these chapters. Chapter 5 explores new ways in learning tasks that can be more difficult to learn using supervised labels. This can come from few positive class labels such in landmarks detection, where very few points are marked as positive samples while the majority are negative or background in a scanned image. Reinforcement learning is applied to learn from such tasks. Trained models are capable of finding target landmarks and standard planes by learning to navigate in the 3D images.

This thesis also contributes to the typical medical imaging pipeline, which can further improve and accelerate the process of extracting clinically relevant information from medi-

cal images, see Figure **??**. Such a pipeline consists of three main steps: (*i*) acquisition, (*ii*) reconstruction, and (*iii*) analysis.

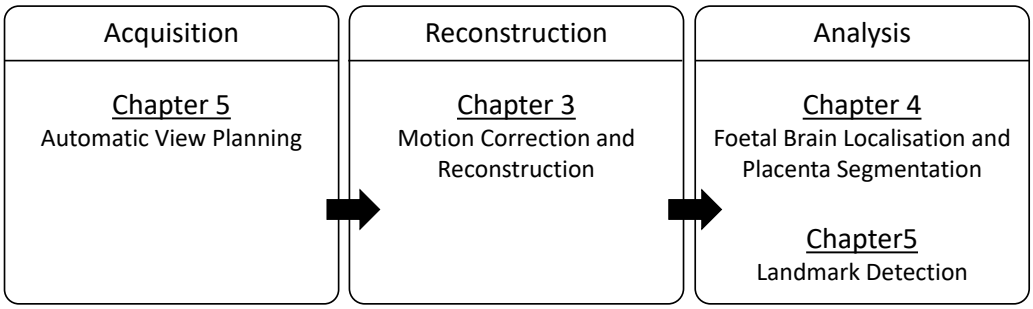


Figure 1.1: Schematic diagram of a typical medical imaging pipeline, which is composed of three main components: (*i*) acquisition; (*ii*) reconstruction. These three components are covered in the next chapters of this thesis.

1.1.1 Foetal MRI Reconstruction, Localisation, and Segmentation

The advent of single shot fast spin echo (ssFSE) T2-weighted sequences enabled spin echo image formation principles [Mansfield and Maudsley, 1977] in magnetic resonance imaging (MRI) to play an essential role in foetal diagnosis [Ertl-Wagner et al., 2002] and medical research [Levine et al., 1998]. As a result, foetal MRI is a growingly available technique used to evaluate and characterise abnormalities in the foetal brain Glenn and Barkovich [2006b]. It also provides a comprehensive evaluation of the developing brain with a detailed visualisation for the assessment of the developing cortex and sulcation pattern, which is extremely difficult with sonography Girard et al. [1993]. Therefore, the localisation of the foetal brain is essential for assessing its development and maturation. It is also a primary step for most of the current automatic motion correction and reconstruction methods for foetal MRI [Keraudren et al., 2014]. In return, foetal reconstruction methods [Kainz et al., 2014a, Kuklisova-Murgasova et al., 2012] have facilitated advanced image-based diagnostics and lead to novel insights about the human development [Malamateniou et al., 2013].

In addition, recent work [Routledge et al., 2015] has shown that MRI can be used for the evaluation of the placenta during both normal and high-risk pregnancies. Particularly, quantitative measurements such as placental volume and surface attachment to the uterine wall, are

required for identifying abnormalities. Also recording the structural appearance (*e.g.*, placental cotyledons and shape) is essential for clinical qualitative analysis. The placenta is usually examined after birth, on a flat surface providing a standard representation for obstetricians. Flat cutting planes, as common in radiology, show only a small part of the placenta. Medical imaging analysis tools including a 3D visualization are considered useful in particular for cases that require preoperative planning or surgical navigation (e.g. treatment of twin-to-twin transfusion syndrome). Hence, fully automatic 3D segmentation, correction of motion artefacts, and visualization is highly desirable for an efficient prenatal examination of the placenta in the clinical practice.

1.1.2 Landmark Detection and View Planning

The detection (localisation) of anatomical landmarks from medical images is an essential step for many image analysis and interpretation methods, such as (*i*) extracting biometric measurements of anatomical structures, (*ii*) view planning during acquisition and (*iii*) initialisation of tasks such as image registration and segmentation. In foetal imaging, anatomical landmarks are used for estimating qualitative scores of foetal biometric measurements, such as foetal growth rate, gestational age, and to identify abnormalities [Rahmatullah et al., 2012]. They are also required in order to identify standardized views such as transventricular and transcerebellar planes, which are commonly used in clinical practice for foetal health screening [Li et al., 2018a]. In brain imaging, the localisation of the anterior commissure (AC) and posterior commissure (PC) points in brain images is required to obtain the optimal view of the mid-sagittal plane. This can be used for the identification of pathological anatomy [Stegmann et al., 2005] or an initial step for image registration [Ardekani et al., 1997]. In cardiac MRI acquisition, the localisation of standard views, e.g. 2- and 4-chamber views, frequently employs a multi-step approach that involves automatic landmark detection [Le et al., 2017, Lu et al., 2011]. Landmark localisation can also be used to initialise deformable models and atlas-based approaches for the evaluation of cardiac ventricular health [Bai et al., 2013]. Since manual landmark annotation is time consuming and error prone, automatic methods [Criminisi et al., 2013, Ghesu et al., 2016, Li

et al., 2018a, Oktay et al., 2017, Payer et al., 2016, Urschler et al., 2018, Xu et al., 2017] have been developed to tackle this problem.

Furthermore, in medical imaging, obtaining accurate biometric measurements that are comparable across populations is essential for diagnosis and supporting critical decision making. For this purpose, standard view planes through a defined anatomy are commonly used in clinical practice to establish comparable metrics for assessing anomalies [Lu et al., 2011]. Finding these planes in an imaging examination through a 3D volume can be time-consuming and suffers from inter-observer variability. Automatic view planning can enable accelerated image acquisition, and facilitate the comparison between different imaging examinations using anatomically standardized biometric measurements.

1.2 Thesis Outline and Summary of Contributions

The aim of this thesis is to develop robust and reliable medical image analysis methods that can assist doctors in making diagnoses and treatment decisions. Figure 1.2 shows the main elements of the typical design of such methods for performing a medical imaging task. The methods proposed in this thesis can be categorized into: (*i*) motion correction and reconstruction from foetal MRI, (*ii*) foetal brain localisation, (*iii*) placenta segmentation, (*iv*) landmark detection, and (*v*) view planning. This thesis is organised as follows:

Chapter 2 provides a brief overview of the background of the relevant medical imaging methods covered in this thesis.

Chapter 3 introduces an automatic method for motion correction and reconstruction super-resolution in foetal MRI scans of the whole uterus. The proposed approach is able to reconstruct a large field of view of non-rigidly deforming structures using parallelised patch-wise optimization, super-resolution, and automatic outlier rejection. This chapter also presents an extensive qualitative and quantitative evaluation of the proposed method compared with state-of-the-art slice-to-volume reconstruction (SVR) methods taking into account: (*i*) computational overhead, and (*ii*) reconstruction accuracy using different metrics such as peak-signal-to-noise ratio

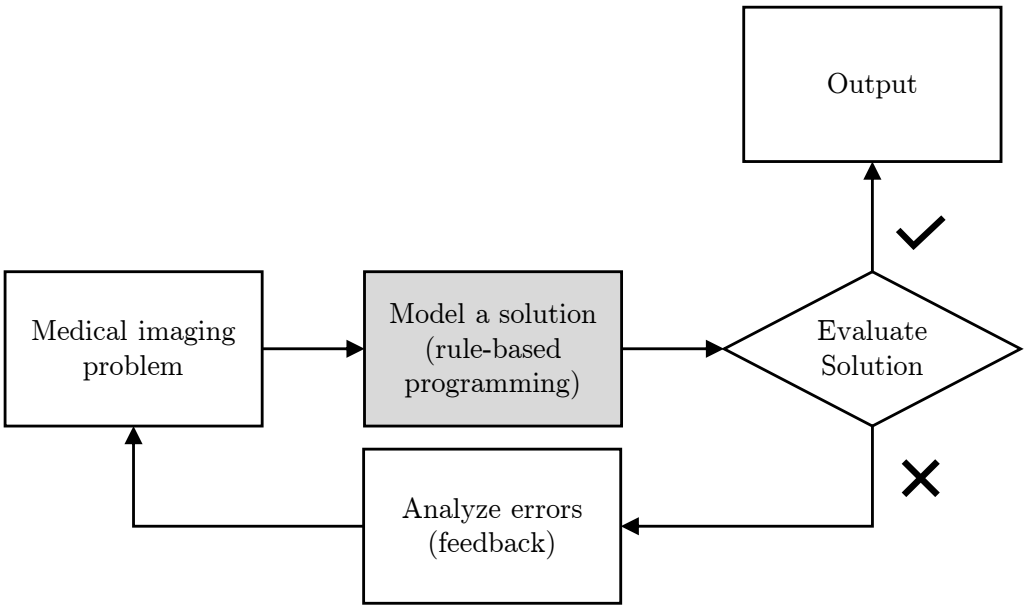


Figure 1.2: Schematic diagram shows the typical design of a medical image analysis method.

(PSNR), structural similarity index (SSIM), and cross correlation (CC). These experiments are performed using real foetal MRI data subject to maternal breathing and sudden foetal movements, and synthetic data with affine motion artefacts. Further evaluations are conducted by measuring the distance error for selected anatomical landmarks in the foetal head, as well as calculating the mean and maximum displacements resulting from automatic non-rigid registration to a motion-free ground truth image. These experiments demonstrate a successful application of the proposed motion correction method to the whole foetal body, uterus and placenta.

Chapter 4 is divided into two parts. The first part proposes a method for foetal brain localisation, which is a structure of interest for many foetal MRI studies. First, it extracts superpixels followed by the computation of a histogram of features for each superpixel using bag of words based on dense scale invariant feature transform (DSIFT) descriptors. Then a graph of superpixels is constructed for training a random forest classifier to distinguish between brain and non-brain superpixels. The localisation framework is evaluated on 55 MR datasets at gestational ages between $20 - 38$ weeks using 5-fold cross validation and achieves a 94.55% brain detection accuracy rate.

The second part of this chapter proposes a fully automatic segmentation framework of

the placenta from structural T2-weighted scans of the whole uterus, as well as an extension in order to provide an intuitive pre-natal view into this vital organ. Foetal MRI has revealed to be important for the evaluation of placenta's health during pregnancy. Quantitative assessment of the placenta requires segmentation, which proves to be challenging because of the high variability of its position, orientation, shape and appearance. Image acquisition is also corrupted by motion artefacts from both foetal and maternal movements. The proposed method adopts a 3D multi-scale convolutional neural network to automatically identify placental candidate pixels. The resulting classification is subsequently refined by a 3D dense conditional random field, so that a high resolution placental volume can be reconstructed from multiple overlapping stacks of slices. The proposed segmentation framework is tested on 66 subjects at gestational ages $20 - 38$ weeks achieving a Dice score of $71.95 \pm 19.79\%$ for healthy foetuses with a fixed scan sequence and $66.89 \pm 15.35\%$ for a cohort mixed with cases of intrauterine foetal growth restriction using varying scan parameters.

Chapter 5 is divided into two parts. The first part evaluates novel deep reinforcement learning (RL) strategies to train agents that can precisely and robustly localize anatomical landmarks in medical scans. Automatic detection of anatomical landmarks is an important step for a wide range of applications in medical image analysis. Manual annotation of landmarks is a tedious task and prone to observer errors. An artificial RL agent learns to identify the optimal path to the target landmark by interacting with an environment, in this case 3D images. Fixed- and multi-scale search strategies with novel hierarchical action steps in a coarse-to-fine manner are also investigated in this chapter. Several deep Q-network (DQN) architectures are evaluated for detecting multiple landmarks using three different medical imaging datasets: foetal head ultrasound, adult brain and cardiac MRI. The performance of the artificial agents surpasses state-of-the-art supervised and RL methods for landmark detection. The experiments also show that multi-scale search strategies perform significantly better than fixed-scale agents in images with large field of view and noisy background such as in cardiac MRI.

The second part of this chapter proposes a fully automatic method to find standardized view planes in 3D image acquisitions. Standard view images are important in clinical practice as they provide a means to perform biometric measurements from similar anatomical regions. These

views are often constrained to the native orientation of a 3D image acquisition. Navigating through target anatomy to find the required view plane is tedious and operator-dependent. RL enables a natural learning paradigm by interaction with the environment, which can be used to mimic experienced operators. Thus, several multi-scale DQN based agents are employed for this task, and extensively evaluated using: (*i*) the distance between anatomical landmarks and detected planes; and (*ii*) the angles between their normal vector and target. The proposed algorithm is assessed on the mid-sagittal and anterior-posterior commissure planes of brain MRI, and the 4-chamber long-axis plane commonly used in cardiac MRI, achieving accuracy of 1.53mm, 1.98mm and 4.84mm, respectively.

Chapter 6 concludes the whole work presented in this thesis, and discusses potential future directions.

1.3 Publications

Some of the results of this research work have been reported or are currently under review in top-tier conferences and journals in the fields of medical images. Below is the list of these publications:

1. **A. Alansary**, M. Rajchl, S. McDonagh, M. Murgasova, M. Damodaram, D. FA Lloyd, A. Davidson, M. Rutherford, J. V Hajnal, D. Rueckert, and B. Kainz. **PVR: patch-to-volume reconstruction for large area motion correction of fetal MRI.** *IEEE Transactions on Medical Imaging (TMI)*, 36(10), pages 2031–2044, 2017. [Alansary et al., 2017]

2. **A. Alansary**, M. Lee, K. Keraudren, B. Kainz, C. Malamateniou, M. Rutherford, J. V Hajnal, B. Glocker, and D. Rueckert. **Automatic brain localization in fetal MRI using superpixel graphs.** *In International Workshop on Medical Learning Meets Medical Imaging (MLMMI)*, pages 13–22. Springer, 2015. [Alansary et al., 2015]

3. **A. Alansary**, K. Kamnitsas, A. Davidson, R. Khlebnikov, M. Rajchl, C. Malamateniou, M. Rutherford, J. V Hajnal, B. Glocker, D. Rueckert, and B. Kainz. **Fast fully automatic segmentation of the human placenta from motion corrupted MRI.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 589–597. Springer, 2016. [Alansary et al., 2016]

4. **A. Alansary**, O. Oktay, Y. Li, L. Le Folgoc, B. Hou, G. Vaillant, K. Kamnitsas, A. Vlontzos, B. Glocker, B. Kainz, and D. Rueckert. **Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection.** *Medical Image Analysis*, 2019. [Alansary et al., 2019]

5. **A. Alansary**, L. Le Folgoc, G. Vaillant, O. Oktay, Y. Li, W. Bai, J. Passerat-Palmbach, R. Guerrero, K. Kamnitsas, B. Hou, S. McDonagh, B. Glocker, B. Kainz, and D. Rueckert. **Automatic View Planning with Multi-scale Deep Reinforcement Learning Agents.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018. [Alansary et al., 2018]

Here is an extended list of publications, which are not included in this thesis but are based on collaborations that contribute to medical image analysis:

1. B. Kainz, **A. Alansary**, C. Malamateniou, K. Keraudren, M. Rutherford, J. V. Hajnal, and D. Rueckert. **Flexible reconstruction and correction of unpredictable motion from stacks of 2D images.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 555–562. Springer, 2015. [Kainz et al., 2015b]

2. B. Kainz, D. Lloyd, **A. Alansary**, M. Murgasova, R. Khlebnikov, D. Rueckert, M. Rutherford, R. Razavi, and J. V. Hajnal. **High-performance motion correction of fetal MRI.** *Eurographics Digital Library*, 2016. [Kainz et al., 2016]

3. M. Rajchl, M. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, **A. Alansary**, O. Oktay, B. Kainz, and D. Rueckert. **Learning under distributed weak supervision.** *arXiv preprint arXiv:1606.01100*, 2016. [Rajchl et al., 2016]

4. H. Miao, G. Mistelbauer, A. Karimov, **A. Alansary**, A. Davidson, D. FA Lloyd, M. Damodaram et al. **Placenta Maps: In Utero Placental Health Assessment of the Human Fetus.** *IEEE Transactions on Visualization and Computer Graphics* 23, no. 6: 1612-1623, 2017. [Miao et al., 2017]

5. B. Hou, **A. Alansary**, S. McDonagh, A. Davidson, M. Rutherford, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz. **Predicting slice-to-volume transformation in presence of arbitrary subject motion.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 296–304. Springer, 2017. [Hou et al., 2017]

6. S. McDonagh, B. Hou, **A. Alansary**, O. Oktay, K. Kamnitsas, M. Rutherford, J. V. Hajnal, and B. Kainz. **Context-sensitive super-resolution for fast fetal magnetic resonance imaging.** *In Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pp. 116-126. Springer, 2017. [McDonagh et al., 2017]

7. B. Hou, B. Khanal, **A. Alansary**, S. McDonagh, A. Davidson, M. Rutherford, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz. **3D Reconstruction in Canonical Co-ordinate Space from Arbitrarily Oriented 2D Images.** *IEEE Transactions on Medical Imaging (TMI)*, 37(8), pages 1737–1750, 2018. [Hou et al., 2018a]

8. B. Hou, N. Miolane, B. Khanal, M. Lee, **A. Alansary**, S. McDonagh, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz. **Computing CNN Loss and Gradients for Pose Estimation with Riemannian Geometry.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018. [Hou et al., 2018b]

9. Y. Li, **A. Alansary**, J. Cerrolaza, B. Khanal, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert. **Fast Multiple Landmark Localisation Using a Patch-based Iterative Network.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018. [Li et al., 2018a]

10. Y. Li, B. Khanal, B. Hou, **A. Alansary**, J. Cerrolaza, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert. **Standard Plane Detection in 3D Fetal Ultrasound Using an Iterative Transformation Network.** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).* Springer, 2018. [Li et al., 2018b]

Chapter 2

Background and Related Work

The work presented in this thesis aims to automate the medical imaging pipeline from image acquisition to extracting useful information that can accelerate and improve the diagnosis process. This chapter outlines the background theory and notable existing literature for the methods related to the presented works in chapters 3, 4, and 5. Every section explains the background theory with a focus on state-of-the-art machine learning methods that are used in medical images. It provides a review of the literature for the presented methods, namely foetal MRI reconstruction, foetal brain localisation, placenta segmentation, landmark detection, and automatic view planning.

2.1 Super-resolution Reconstruction in Foetal MRI

This section discusses briefly the application of a typical super-resolution framework to foetal MRI reconstruction.

2.1.1 Background

The advent of single shot fast spin echo ssFSE T2-weighted sequences has enabled spin echo image formation principles [Mansfield and Maudsley, 1977] in MRI to play an essential role in

foetal diagnosis [Ertl-Wagner et al., 2002] and research [Levine et al., 1998]. In particular, cases for which US fails to acquire conclusive image data, *e.g.* in obese patients, require a complementary MRI scan [Garel, 2008, Huppert et al., 1999]. Recent advances in motion compensation facilitate advanced image-based diagnostics and lead to novel insights about the human development [Malamateniou et al., 2013]. Foetal MRI enables an ability to distinguish between individual structures such as brain, lung, kidney and liver, as well as pregnancy structures such as the placenta, umbilical cord and amniotic sac. It provides improved visualisation and structural information of the foetal anatomy, which enables to study abnormalities during pregnancy, *e.g.* neuro-developmental disorders [Glenn and Barkovich, 2006a], placental pathologies [Linduska et al., 2009], foetuses with congenital lung masses [Zamora et al., 2014], and conjoined twins [Pierro et al., 2015]. MRI is considered to be safe after the first trimester [Frates et al., 2004] for 1.5T [Patenaude et al., 2014] and 3T [Cannie et al., 2015] without the use of contrast agents, which may have teratogenic effects. Furthermore, this technology paves the way for researchers and clinicians to analyse correlations between childhood development and prenatal abnormalities.

During MR acquisition the foetus is not sedated and moves freely, while the mother also breathes normally. As a result, movements are likely to corrupt the scans, hiding pathology and causing overlap between different anatomical regions. In order to limit these artefacts, fast scanning sequences such as ssFSE [Levine, 2004] allow for the rapid acquisition of single slices at high *in-plane* resolution in a large field of view and good tissue contrast of the uterus. However, when acquiring a 3D volume through a stack of slices, inter-slice artefacts in the *out-of-plane* views are highly likely. Consequently, this restricts reliable diagnostics to individual 2D slices in the current clinical practice. Figure 2.1 depicts a typical example of motion related artefacts in a foetal ssFSE scan. The observed motion (*c.f.* Figure 2.1 b & c) is of unpredictable nature and consists of a combination of maternal respiration, bowel and foetal movements.

Slice-to-volume registration (SVR) combined with super-resolution image reconstruction techniques [Park et al., 2003] can be applied to compensate motion between single slices, assuming there is minimal motion in the in-plane direction (*i.e.* for a single slice of k-space, there are no significant motion artefacts that occur during the acquisition of that slice). This is done



Figure 2.1: Three view-planes for raw 3D data acquired through stacks of ssFSE images covering the whole uterus. The transverse (a) is the *in-plane* view, *i.e.*, native 2D slice scan orientation. Motion causes streaky artefacts for multi-planar reconstructions (MPR) in orthogonal views (b) and (c) caused by both maternal and foetal movements between the acquisition of individual slices.

by reconstructing a high-resolution (HR) image from multiple and overlapping low-resolution (LR) images, as shown in Figure 2.2. To provide a sufficiently high number of samples for such an approach, multiple stacks of 2D-slices need to be acquired, ideally in orthogonal orientations.

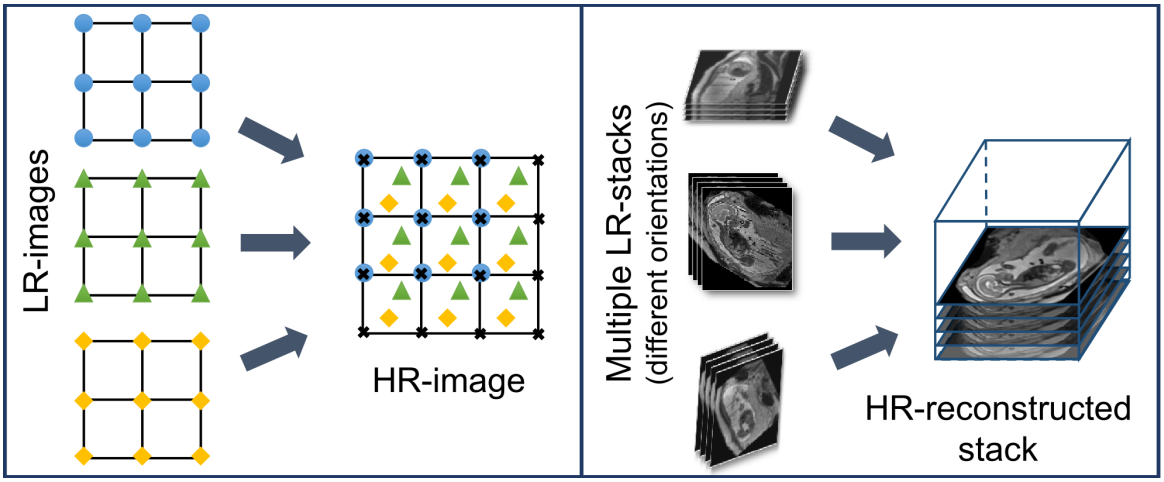


Figure 2.2: Illustration of the basic ideas behind reconstruction [Park et al., 2003]: A simplified example of a 2D 4x4 HR grid sampling from a 2D 3x3 LR grids (left) and a practical example of 3D foetal MRI using multiple overlapping stacks of slices, by reconstructing a 3D HR image with an isotropic voxel size from LR images with anisotropic voxel size.

2.1.2 Image Super-resolution

A simple LR $\rightarrow$ HR reconstruction model [Park et al., 2003] can be formalized as:

$$Y_i = W_i X + n_i \quad \text{for} \quad 1 \leq i \leq N, \tag{2.1}$$

where Y_i denotes the i-th LR image of total N images, and X is the HR image. The matrix W_i combines motion, sub-sampling and degradation effects: $W_i = DBT_i$, where D is the sub-sampling matrix, B is the blurring matrix, and T_i is the transformation matrix of observation i. The noise of observation i is represented by n_i. LR images can be considered as a down-sampled, motion corrupted, blurred, and noisy version of the HR image. The resulting reconstruction depends on two main parts: motion correction (estimating W_i) and super-resolution (estimating X). Image registration can be used to estimate motion, interpolation to obtain a uniformly spaced HR image, and regularized super resolution with automatic outlier rejection to remove

blur and noise. Volumetric foetal MR image reconstruction is more challenging than typical image reconstruction problems due to the unconstrained random motion of the foetus during slice acquisitions. Slice misalignments can lead to a loss of spatial coherence and typically introduce anisotropic voxel sizes and intensity inhomogeneities.

2.1.3 Related Work

The majority of motion compensation approaches for foetal MRI are based on SVR techniques that aim to obtain a motion-free and high resolution volume of a foetal target region. Registration of individual 2D slices with a higher resolution 3D volume [Fei et al., 2003] is the core approach of these algorithms. SVR methods assume that all acquired images are centred at a specific organ (*e.g.*, brain, thorax) and cover three orthogonal image directions. Figure **??** shows the core elements of SVR and the contribution of previous frameworks from the literature.

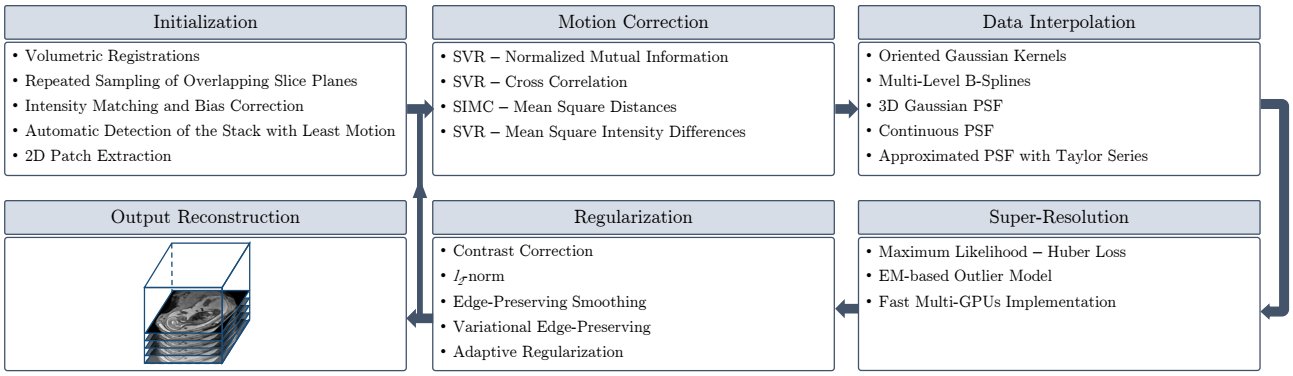


Figure 2.3: Overview of the required modules of state-of-the-art SVR methods and main components introduced by previous works.

The first SVR-based reconstruction framework for foetal MRI was introduced by Rousseau et al. [2006]. It includes steps to correct 2D slice misalignments, intensity inhomogeneity distortions, and reconstructs an isotropic HR foetal brain from sets of LR images. Motion correction is done by applying a global 3D rigid alignment between the LR images using one image as a reference to define the global coordinate system. Then every slice is aligned to the initial reconstructed HR volume. Normalized mutual information is maximised using the gradient ascent method for both registration steps. A narrow Gaussian kernel is applied as a point spread function (PSF) for volume reconstruction and empty voxels are filled using the

mean of the surrounding voxels. The image contrast is corrected using one LR image as a reference.

Jiang et al. [2007] introduced the acquisition of many thin slices to provide sufficient sampling of the region of interest. Cross correlation is used as a cost function for the SVR steps assuming that the data have consistent contrast properties. After that, multilevel B-splines are applied to the volumetric reconstruction for data interpolation, which has the advantage of reducing blurring of the reconstructed image supported by including the thin slices.

Kim et al. [2010] proposed a method for slice intersection motion correction (SIMC) of multi-slice MRI for 3D foetal brain image formation. The method is based on slice-to-slice registration using spatially weighted mean square intensity differences (MSD) of the signal between slices as an energy, assuming that the MRI contrasts are identical. Maternal tissues are excluded from the energy computations using a windowing function of a parametric ellipsoid model. Similar to [Rousseau et al., 2006], temporally adjacent slices are grouped together then divided into half iteratively. The splitting process is performed using discrete cosine basis functions.

Gholipour et al. [2010] were the first to introduce a mathematical model for super-resolution (SR) volume reconstruction from slice acquisitions of foetal brains. The main difference to previous methods is that it includes knowledge of the slice acquisition model and the SR reconstruction is performed based on maximum likelihood and a robust M-estimation minimisation for an error norm function. A Tikhonov regularization term is added to the cost function in order to enforce a solution when the number of acquired samples is not high enough for solving the reconstruction problem.

Rousseau et al. [2010, 2013] proposed to use a variational regularization including an approximation of total variation (TV) to better preserve edges. Later, Tourbier et al. [2015] introduced an adaptive regularization by applying novel fast convex optimization techniques to design an efficient optimization algorithm for the super-resolution problem using edge-preserving TV regularization.

Kuklisova-Murgasova et al. [2012] were able to reconstruct the foetal brain using intensity

matching and complete outlier removal. The main steps of their reconstruction method are: (*i*) 3D registration of the acquired stacks using a template stack; (*ii*) extracting region of interest (the foetal head) from all the stacks; (*iii*) intensity matching and bias correction between the slices based on an EM framework, where the differential bias fields and slice-dependent scaling factors are estimated during the reconstruction; (*iv*) motion correction using [Rousseau et al., 2006] based on the normalized cross correlation as a similarity measure and an approximated 3D Gaussian PSF similar to [Jiang et al., 2007]. A posterior probability is used to define the inlier and outlier voxels within the EM framework in order to remove the motion-corrupted artefacts and misaligned data. Blurring in reconstructed images is reduced by integrating edge-preserving regularization based on anisotropic diffusion within the SR reconstruction framework.

Kainz et al. [2015a] developed a fast multi-GPU accelerated implementation for the method presented in [Kuklisova-Murgasova et al., 2012], which is based on 2D-3D registration, SR with automatic outlier rejection and an optional intensity bias correction. They extended the reconstruction framework by automatically selecting the stack with least motion as the reference stack and using a fully flexible and accurate PSF instead of approximated functions. Using a multi-GPU framework enabled the SR reconstruction process to be approximately five to ten times faster than using a multi-CPU framework.

Modelling non-rigid transformations with multiple rigid transformations and without an initial registration target has only been preliminarily explored for foetal MRI in [Kainz et al., 2015b]. Other works for non-rigid 3D-3D registration between an outlier-free, regularly sampled source and target volume, constrain parts of an image to move independently and rigidly (e.g. bones through manual segmentation and landmark annotation), while allowing other parts to deform [Little et al., 1996]. In 3D-3D registration settings globally continuous 3D deformation fields, can be approximated by fusing multiple local rigid transformations as shown in [Arsigny et al., 2009, Commowick et al., 2008].

Practical Limitations of SVR

SVR methods have been successfully employed to address these problems in foetal MR and are typically applied to small regions and organs with rigid body characteristics that are identified by manual annotations [Gholipour et al., 2010, Kim et al., 2010, Kuklisova-Murgasova et al., 2012, Rousseau et al., 2006] or less precise, automated segmentations [Alansary et al., 2015, Keraudren et al., 2014, Tourbier et al., 2017]. Such approaches cannot be applied to the whole body and uterus reconstruction because of the assumption of rigid motion in the 2D to 3D registration step of SVR. As a result, different areas in each slice that are likely to move in different directions will break this assumption, *e.g.*, the head and thorax. Further, an extension of 2D-3D registration to include non-rigid deformations is only well defined with each slice and not well-constrained in 3D. Current SVR approaches will fail in the presence of non-rigid deformations and unpredictable organ shapes. This restricts the application of SVR to regions that are manually or automatically annotated. Thus, most of the previous SVR methods are limited to the foetal brain as the main region of interest for foetal reconstruction due to the high incidence of neuro-developmental disability in premature infants. Only recently, Kainz et al. [2014b], Lloyd et al. [2016] proposed a motion corrected 3D reconstruction of foetal thoracic structures from prenatal MRI. Moreover, SVR is computationally expensive due to the exponential increase of computation with the size of the target area. This leads to prohibitive post-processing times in the clinical practice. Parallelized implementations [Kainz et al., 2015a] can address run-time problems, however, methodologically SVR is still restricted to small, rigid body areas.

Reconstruction of Large-scale Anatomy

MRI has further been shown to be useful for the evaluation of the whole uterus and structures like the placenta. During both normal and high-risk pregnancies, the whole uterine appearance and the condition of the placenta are considered to be an indicator for foetal health after birth [Routledge et al., 2015]. Placental functions affect the birth weight as it controls the transmission of nutrients from the maternal to the foetal circulation [Salafia et al., 2009].

However, the whole foetal body and secondary uterine parts can be inherently inconsistent. Different foetal body parts can move independently from the uterus. This makes the application of SVR and 2D-3D registration to the full uterus erroneous in the presence of foetal motion and maternal respiration.

Besides, multiple births is a case where classical SVR pipelines, based on preprocessing steps to identify consistent rigid regions, usually fails. The presence of multiple instances of the same foetal structure was not considered in previous methods. Therefore, a fully automatic motion correction method for the whole uterus is very desirable and will enable the application of standard 3D image analysis techniques, *e.g.*, [Alansary et al., 2016, Keraudren et al., 2015].

2.2 Foetal Brain Localisation and Placenta Segmentation

This section discusses briefly the application of machine learning (ML) methods that have been used in medical images (*i.e.* image classification and segmentation). This is then extended to include state-of-the-art deep learning methods, followed by a summary of the related work in foetal brain localisation and placenta segmentation.

2.2.1 Background

Foetal Brain Localisation

As discussed in the previous section, foetal MRI has significantly improved in the last two decades, and is emerging as a novel, non-invasive tool for diagnosis and planing of surgical interventions. It provides higher contrast and larger field-of-view than ultrasound. Thus, it provides better structural information of the different foetal organs such as the brain, spine and body. Foetal brain localisation is important for assessing the foetal brain development and maturation. It is also a primary step for many of the automatic motion correction techniques for

foetal MRI [Keraudren et al., 2014]. It has been also used as an essential landmark to extract other foetal organs [Keraudren et al., 2015]. Problems that hinder the design of automated image analysis tools for foetal MRI usually arise from: (*i*) the high variability in shape, size, orientation, and anatomical configuration of the foetus; (*ii*) intensity non-uniformities (bias artefacts); (*iii*) partial volume effects; and (*iv*) motion artefacts caused by the unconstrained foetal motion (see Figure 2.1). In order to tackle these challenges, previous methods for foetal brain localisation [Anquez et al., 2009, Ison et al., 2012, Keraudren et al., 2013] rely on machine learning (ML) to train a model from labelled images that can later be applied for localising the brain from unseen test data.

Placenta Segmentation

The functions of the placenta affect the foetal birth weight, growth, prematurity, and neurodevelopment since it controls the transmission of nutrients from the maternal to the foetal circulatory system. Figure 2.4 shows a 3D visualisation of the placenta from foetal MRI. Recent work [Routledge et al., 2015] has shown that MRI can be used for the evaluation of the placenta during both normal and high-risk pregnancies. Particularly, quantitative measurements such as placental volume and surface attachment to the uterine wall, are required for identifying abnormalities. In addition, recording the structural appearance (*e.g.* placental cotyledons and shape) is essential for clinical qualitative analysis. Moreover, the placenta is usually examined after birth, on a flat surface providing a standard representation for obstetricians. Flat cutting planes, as common in radiology, show only a small part of the placenta. A 3D visualisation is considered useful in particular for cases that require preoperative planning or surgical navigation (*e.g.* treatment of twin-to-twin transfusion syndrome). Hence, fully automatic 3D segmentation, correction of motion artefacts, and visualisation is highly desirable for an efficient pre-natal examination of the placenta in the clinical practice.

As discussed earlier, fast MRI acquisition techniques allow acquiring single 2D images of the moving uterus and foetus fast enough, so that to limit the in-plane motion artefacts. However, 3D data acquisition and subsequent automatic segmentation is challenging because maternal

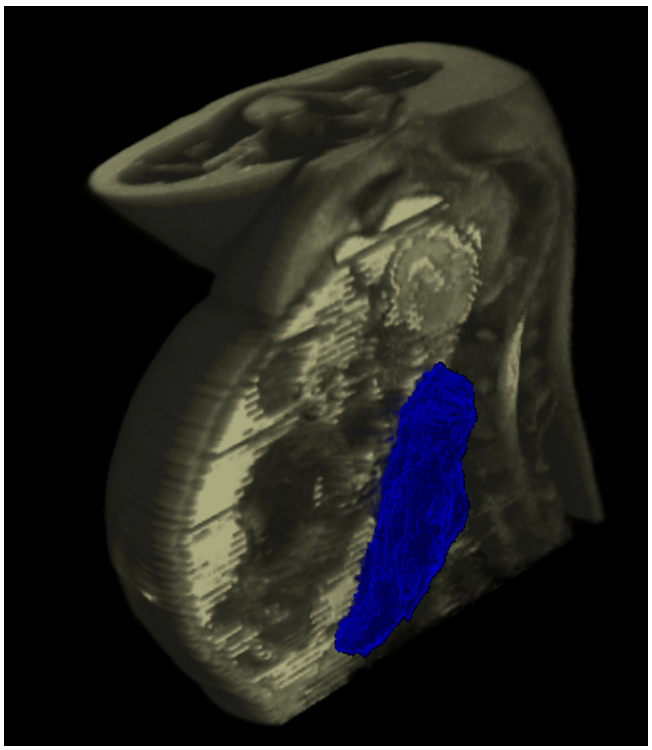
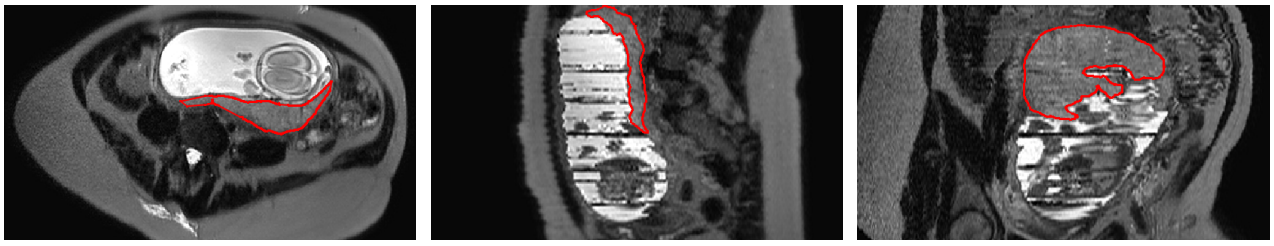


Figure 2.4: A 3D visualisation of the placenta from foetal MRI.

respiratory motion and foetal movements displace the overall anatomy, which causes motion artefacts between individual slices. Furthermore, a high variability of the placenta's position, orientation, thickness, shape and appearance inhibits conventional image analysis approaches to be successful, see Figure 2.5.



(a) Axial (b) Sagittal (c) Coronal

Figure 2.5: Three orthogonal 2D planes from a motion corrupted 3D stack of slices showing a delineated placenta. The native scan orientation (a) shows no motion artefacts, while (b) and (c) do.

2.2.2 Machine Learning in Medical Images

Machine learning (ML) is concerned with designing broad algorithms for learning from data to make a decision [Samuel, 1959]. It can also provide insights about complex tasks and data by revealing unexpected correlations in the data. Figure **??** shows a typical ML framework for a medical imaging task. For example, in foetal brain localisation, the data are MRI brain images and a decision has to be made to classify a region (bounding box) into brain or non-brain.

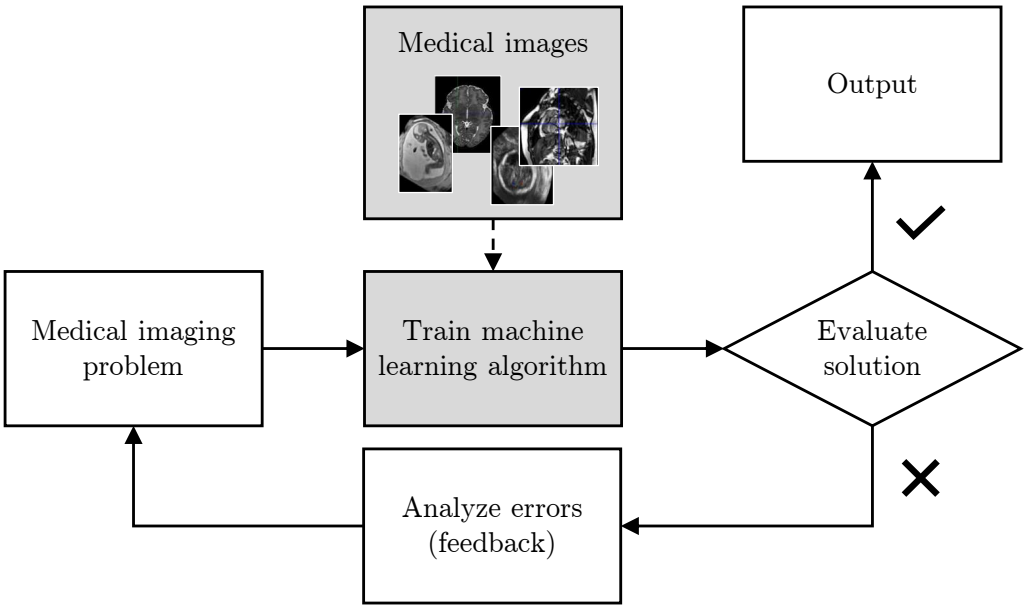


Figure 2.6: Typical machine learning framework for solving a medical imaging problem.

A good ML model can be considered as a simplified version of the training data, which can discard details and generalize well on new instances. The steps to find such a model can be summarised as: (*i*) data pre-processing, (*ii*) model selection, (*iii*) model training by finding parameters that minimise a cost function, and (*iv*) applying the trained model on new test cases (inference). Ideally this model should achieve low generalisation error (GE) or out-of-sample error, where the GE can be defined as the difference between training and test errors. ML systems can be characterised based on three main factors: (*i*) the amount and type of supervision during training, (*ii*) the continuity of learning from instances, and (*iii*) the evaluation during training.

Machine learning can be classified according to the amount and type of supervision during training:

- **Supervised**: Target or ground truth labels are provided with the training dataset $D = \{x, y\}_{n=1}^{N}$, where x and y represent input data features and ground truth labels. The main goal is to find the best model parameters θ that minimises the output of a loss function $L(y, \hat{y})$, where $\hat{y}$ is the predicted output of the model. In medical imaging analysis task such as classification, an ML algorithm can be trained to classify healthy and pathological cases by providing many training examples. This can be extended to a regression task as well, whereas the algorithm predicts a value of how severe is the pathological case (*e.g.* k-nearest neighbours (KNN) [Altman, 1992], linear regression, logistic regression, support vector machines (SVM) [Cortes and Vapnik, 1995], decision trees and random forests (RF) [Ho, 1995], and artificial neural networks (ANN) [McCulloch and Pitts, 1943]).

- **Unsupervised**: External supervision with ground truth labels is not provided during training. This is can be applied to detect similar patterns for different classes (*e.g.* healthy and pathological subjects). There are several loss functions that can be applied for unsupervised learning. One common example is the reconstruction loss function $L(x, \hat{x})$, where the model learns to reconstruct the input image x to $\hat{x}$ through lower dimensional representations. There are many methods for unsupervised learning such as: (*i*) clustering (*e.g.* k-means, hierarchical cluster analysis (HCA) [Sibson, 1973], expectation maximisation (EM)), (*ii*) visualisation (*e.g.* output 2D or 3D representations of complex data), and (*iii*) dimensionality reduction by simplifying the data without loosing much information (*e.g.* principal component analysis (PCA) [Jolliffe, 2011], locally-linear embedding (LLE) [Roweis and Saul, 2000], t-distributed stochastic neighbour embedding (t-SNE) [Maaten and Hinton, 2008]).

- **Semi-supervised**: The algorithm is given partially labelled training data, which can be considered as a combination of supervised and unsupervised algorithms. For example, deep belief networks (DBNs) [Hinton, 2009] consist of unsupervised component of stacked restricted Boltzmann machines (RBMs) [Smolensky, 1986], and then the whole system is trained in supervised manner.

- **Reinforcement**: An artificial agent learns interactively by sampling sequential states

from an environment. At each step the agents takes an action and receives a reward signal that indicates if it moves to the target solution or not. The main goal is to learn the best strategy or policy, so as to maximise the accumulated rewards over the agent's lifetime (*e.g.* Q-learning [Watkins and Dayan, 1992] and actor-critic methods [Sutton and Barto, 1998a])

Another way of categorising learning methods is based on the continuity of learning from new instances. This is can be classified into two main groups:

- **Batch or offline learning:** The algorithm learns from all available data during training. Such methods can be time consuming and memory expensive, as they learn the whole system before applying to testing (offline learning).
- **Online or incremental learning**: The algorithm learns incrementally by adding examples of data sequentially (*e.g.* individual or small groups (mini-batches) of instances). This is can be faster and cheaper memory-wise compared to offline learning. It is also a good option for limited resources or big data (out-of-core learning). It introduces a new parameter to the learning algorithm, namely, the learning rate or step size. It controls how fast the algorithm adapts the change in the data. High learning rate values adapt quicker than small values, however, it forgets old instances of data quickly. Thus, small learning rate values can be less sensitive to noise in data.

Furthermore, an ML model can be classified based on how it generalizes into two main categories:

- **Instance-based learning**: The algorithm learns to recognise the training examples and generalizes to new instances using similarity measures (*e.g.* Euclidean distance).
- **Model-based learning**: The algorithm builds a model from examples that is applied for testing. This is done by evaluating the model using a utility (fitness) function, which measures how good the model is (it is also called a cost function if it measures how bad the model is, *e.g.* linear regression).

The performance of the trained model can be affected by the size of the training data, especially if there are insufficient data for learning. In fact, Banko and Brill [2001] proved that different ML algorithms would perform identically on a complex problem, when *enough data* are used for training. Thus, medium and small datasets are more challenging for ML algorithms. The size of the data is not the only concern, when it comes to selecting the training examples. It is also essential to select representative training data to avoid sampling errors, such as (*i*) sampling noise from using very small unrepresentative samples, and (*ii*) sampling bias from using very large samples with a flawed sampling method. Preprocessing the data, removing errors (*e.g.* noise and outliers), is a vital step for any ML pipeline. Care must be taken as well while selecting a model for specific training data, whereas choosing the wrong model may result in either overfitting or underfitting. The former happens when the model works well on the training data but fails to generalise on the testing data. This can happen if the training data are very small or noisy, or the model is very complex and learns the details in training dataset. While underfitting occurs when the model does not perform well on both training and testing data. This is caused mainly by selecting a simple model that is not able to learn underlying structures in the data.

2.2.3 Medical Image Classification

A conventional ML-based image classification pipeline includes three main steps: (*i*) feature extraction, (*ii*) image descriptors, and (*iii*) training a classifier, see Figure 2.7.



Figure 2.7: Conventional image classification pipeline.

Image Features

An image attribute is commonly known as the data type, *e.g.* intensity colour, and a feature can be simply defined as an attribute plus its value. In ML, **image features** should characterise

the image content in a meaningful way. Suitable image features usually are discriminative and invariant, so that the same features can be detected in multiple images. *Feature detection* (selection) is the process of finding pixels with features of interest (*e.g.* edges, lines and corners). These features are computed for every pixel or for certain parts of the image if it is computationally expensive. *Feature descriptors* are then extracted by applying a local patch around the detected feature, this process is called *feature engineering*. Descriptors that describe visual features in the image can be divided into two main types, namely, (*i*) general visual descriptors that contain low level features such as the colour, shape and texture; and (*ii*) specific domain visual descriptors that contain high level features about an object such as face detection.

Viola and Jones [2001] developed a rapid object localisation method using a boosted cascade of simple features. Their object localisation procedure classified images using simple *rectangular features* based on Haar basis functions [Papageorgiou et al., 1998]. They were able to compute these features very rapidly using integral image [Lewis, 1995]. Ison et al. [2012] were able to develop a fully automated foetal brain extraction and orientation detection approach using similar 2D-Haar like features. The main advantages of Viola's localisation method are the extremely fast feature computation and the scale and location invariant detection. However, it is sensitive to intensity variations and can hardly deal with object rotations. In contrast to Haar features, *steerable features* [Freeman and Adelson, 1991] can deal with object rotations using a small set of rotated versions of the convolution kernel. Ojala et al. [2002] proposed another approach, local binary patterns (LBP), that can cope with intensity variations using a binary pattern computed from comparing each pixel with its neighbours.

Instead of extracting features for every pixel in the image for object detection, Lowe [1999] developed an approach that can detect and extract features for interesting points on the object in the image using scale-invariant features (SIFT). This method is based on transforming the image into feature vectors that are invariant to image scaling, translation and rotation. First, Gaussian filters are applied to the image at different scales. Then the maxima and minima of the differences of Gaussians (DoG) are computed to identify keypoint locations in the image. Keypoints with low-contrast or located on an edge are discarded or eliminated from the keypoints list for more generalization. Consistent orientations are assigned to the extracted key-

points based on local image gradient directions. The final keypoint descriptors are constructed using spatial histograms of the image gradients over a window centred on each keypoint, see Figure 2.8.

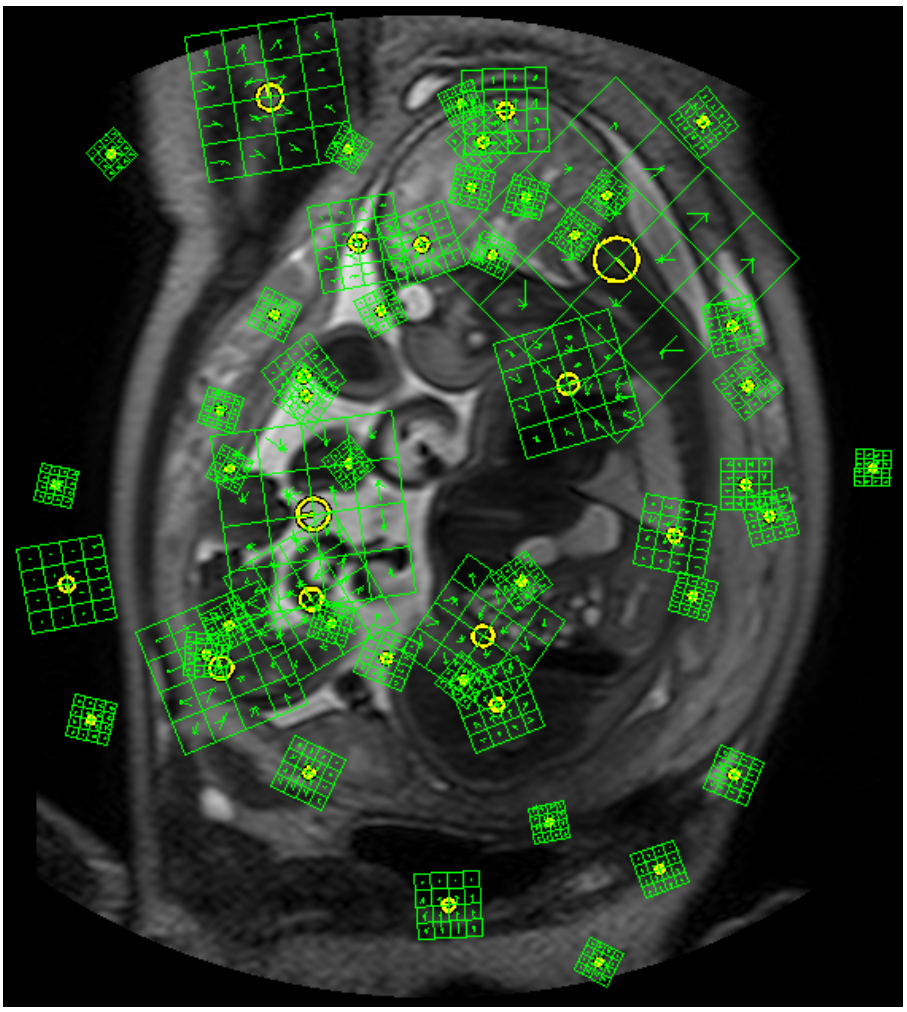


Figure 2.8: Selected SIFT keypoint descriptors applied to a 2D fetal MR image.

Feature learning techniques are usually performed to transform the extracted features into an effective representation for machine learning tasks, e.g. numerical vectors. Fei-Fei and Perona [2005] proposed a bag-of-words (BoW) model that can represent the image into a bag of codewords by treating image features as words. A bag of visual words is a sparse vector of occurrence counts (histogram) of a codebook or vocabulary of local image features (words). The codebook is generated so that each codeword represents similar objects or patches. Leung and Malik [2001] presented a simple method for constructing a codebook by applying k-means clustering to all the vectors, whilst the number of the clusters defines the codebook size and their centres represent the codewords. Since each patch or object can be mapped to a specific

codeword using the generated codebook, the image can be represented using a histogram of the codewords.

Image Classification

Image classification can be defined in machine learning as the process of identifying images or parts of them based on image features. The classification algorithm that maps the input data into classes is called a *classifier*. It can be modelled as: $\hat{Y} = h_\theta(X)$, where $\hat{Y}$ is a vector of n classes, X is a vector of k features, and h_θ represents the model with θ parameters. Typically, h_θ has many different θ and configurations to adjust before it achieves optimal performance on distinguishing between different classes. θ can be estimated by minimising the error of a cost function $J(\theta)$. A regularization part can also be added to reduce overfitting (*e.g* l_p-norm, Tikhonov [Tikhonov, 1963] and lasso [Tibshirani, 1996] regularisation). Every learning model tends to perform better for some problem types than others. The following are examples of different classifiers commonly used in the literature.

Boosting methods usually refer to learning algorithm that combines several *weak* classifiers into a *strong* predictor using:

$$H_T(x) = \sum_{i=1}^{T} \alpha_i h_i(x), \tag{2.2}$$

where h_i is the i-th weak classifier weighted with α_i and x is the input attribute. For example, Viola and Jones [2001] proposed an adaptive boosting (AdaBoost) [Freund and Schapire, 1997] learning algorithm for both selecting best Haar features and training classifiers. The algorithm trains several weak classifiers sequentially, where each corrects the output of its predecessor. This model is invariant to both location and scale. However, its accuracy may decline with object rotations and intensity variations. Another drawback of these sequential methods is their scalability, where it is not possible to train each weak learner in parallel.

Support vector machine (SVM) [Cortes and Vapnik, 1995] is another popular model that is used for classification or regression especially for medium and small datasets. SVM learns with supervision to build a hyperplane or set of hyperplanes in order to separate the

training-data points. There might be many hyperplanes that can separate the training data but the best hyperplane is chosen based on the largest separation or margin between two classes, see Figure 2.9. For non-linear SVMs, the input data are mapped to a higher dimensional space than the original one where the classes are linearly separable. Thus new data can be mapped to the same space and then classified based on which side of the hyperplane they fall on. Some classification problems can not be solved linearly, thus non-linear classification can be performed with SVM using a non-linear kernel function (*e.g.* polynomial, Gaussian, hyperbolic tangent).

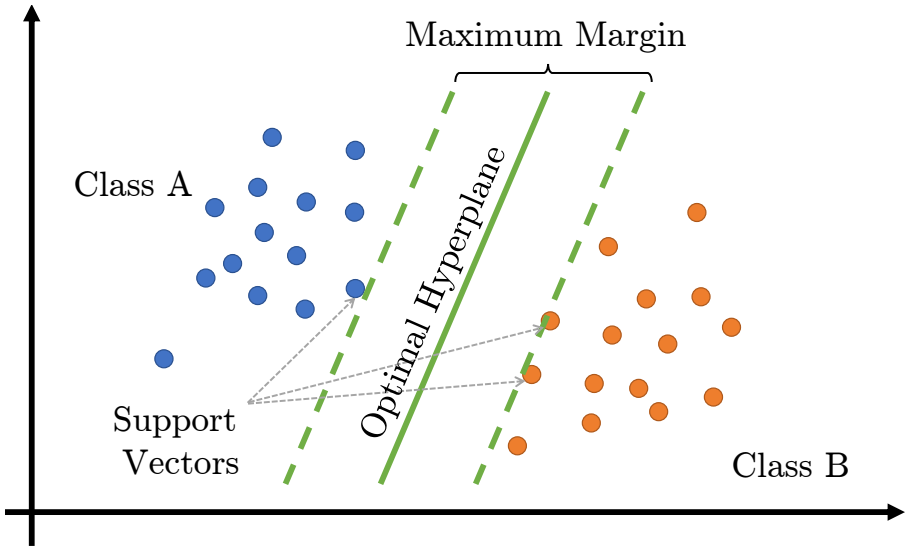
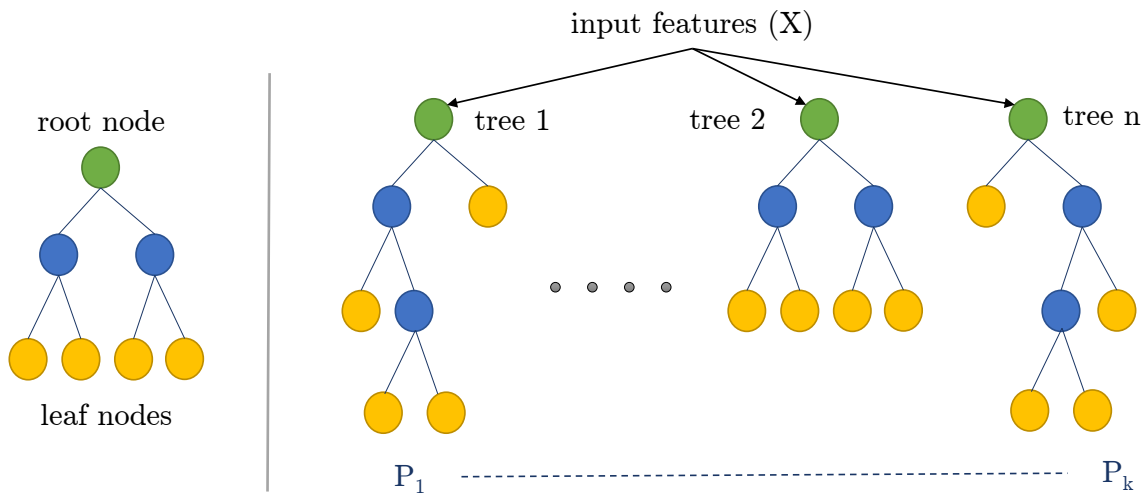


Figure 2.9: An illustrative diagram showing support vector machine (SVM) input space. The optimal hyperplane is selected using the largest margin between two classes based on their support vectors.

Decision trees (DT) [Quinlan, 1986] are another powerful versatile algorithms that can be used for both classification and regression. DT result in multiple outcomes based on a series of decisions. It consists of three main components: root and leaf nodes, and branches, see Fig 2.10 (a). A tree starts with the root node, which is then split in two or more branches based on specific criteria or questions to be answered. This decision making process can continue until the whole tree is built. The outcomes represents the leaf nodes in the tree. The main advantage of DT is that their structure is fairly intuitive and interpretable (*white box model*), which can easily provide explanations why a decision is made.

Random forests (RF) [Breiman, 2001] is an Ensemble learning method that aggregates the predictions of a group of DT classifiers. Combining the output of different classifier often



(a) Decision tree (b) Random forest

Figure 2.10: An illustrative diagram showing: (a) a decision tree components (root and leaf nodes, and branches), and (b) a random forest of n-trees. The final decision between k-classes is made by majority voting in classification or averaging in regression.

achieves better accuracy than the best classifier (*wisdom of the crowd*). The algorithm is based on training the individual decision trees with random samples of the training data (*bagging*) at each candidate split node. The split is usually computed based on the information gain. The final decision is made by averaging the predictions from all the trees for regression, or taking the majority vote for classification. This can be computationally efficient by training each tree, as well as predictions, in parallel via different CPU cores. During training, the sampled data can be divided into two parts (e.g. 63% and 37%), where the first part is used for training each tree. The remaining samples can then be used for evaluation without the need for a validation set, which are called *out-of-bag* (OOB) instances. The OOB error can further be used to rank important features, which are likely to appear closer to the root nodes. Geurts et al. [2006] extended the random forest in [Breiman, 2001] by selecting a random value for each feature under consideration instead of using the information gain for computing the optimal feature. Their method is called extremely randomized trees or *ExtraTrees* as it adds extra randomness step to the typical random forest. This extra randomizations can make the decision boundaries smoother and decrease the overfitting in some cases. In medical imaging, Criminisi et al. [2011] developed a detection and localisation approach in CT studies using multivariate regression forests based on the offset displacements of the current voxel with respect to bounding box of

the object.

2.2.4 Deep Learning in Medical Images

In Section 2.2.2, different conventional ML methods were discussed with the focus of foetal brain localisation. The majority of these methods rely on engineering visual features for training a classifier. For organs with significant variability in their appearance (*e.g.* placenta), more complex features are required in order to localise or segment the target organ robustly. Besides, typical hand-crafted features often require expert knowledge that is usually domain-specific and differs from one application to another. *Deep learning* allows different models to learn complex feature representations using artificial neural networks (ANNs) composed of several layers. Deep convolutional neural networks (CNNs) [Fukushima, 1980] contain many layers that transform a raw input image with convolution filters to complex but useful features or representations. An early work [Lo et al., 1995] applied CNN techniques for lung nodule detection. However, DL methods require a big amount of training data, which prevented CNNs from scaling up at that time because of the lack of computational resources. Later, Krizhevsky et al. [2012] proposed a new efficient implementation of deep CNN architectures with the advance of computing power through (*e.g.* GPUs and multi-core CPU systems), and the availability of large data sets. This has enabled deep CNNs to achieve state-of-the-art performance for various medical image tasks in the last few years, such as: classification, detection, segmentation, and registration [Litjens et al., 2017]. Another advantage of such deep models is that they do not require any engineering steps in their pipeline as in traditional ML methods; thus, they can be trained *end-to-end* [Bengio et al., 2013]. This section gives a brief overview of both ANNs and CNNs, and illustrates the typical building blocks.

Artificial neural networks (ANN)

ANNs are versatile, powerful and scalable machine learning models. This has made them a key component of the core of current DL methods for tackling large and complex learning

tasks (*e.g.* classifying millions of images and speech recognition systems). McCulloch and Pitts [1943] introduced the first ANNs based on a very simplified computational model of biological neurons to perform complex computations using propositional logic. Their model has one or more binary inputs $\mathbf{x}$ and one binary output$y$, which activates when a certain number of inputs are *True*. A network of these simple computational neurons were capable of computing any logical proposition. After few years, Rosenblatt [1958] presented a slightly modified version of the artificial neuron based on a linear threshold unit (LTU), which is called a *Perceptron*. Compared to [McCulloch and Pitts, 1943], the inputs and output are numbers instead of binary values, where inputs $\mathbf{x}$ are associated with weights $\mathbf{w}$. The LTU computes the output of a *step function* applied to the weighted sum of the inputs. This can be defined as: $h_w(\mathbf{x}) = \sigma\left(\mathbf{w}^T \cdot \mathbf{x}\right)$, where the σ is a step function (*e.g.* Heaviside) that outputs the positive class if the result of the linear summation is above a threshold, or outputs the negative class otherwise, see Figure **??**.

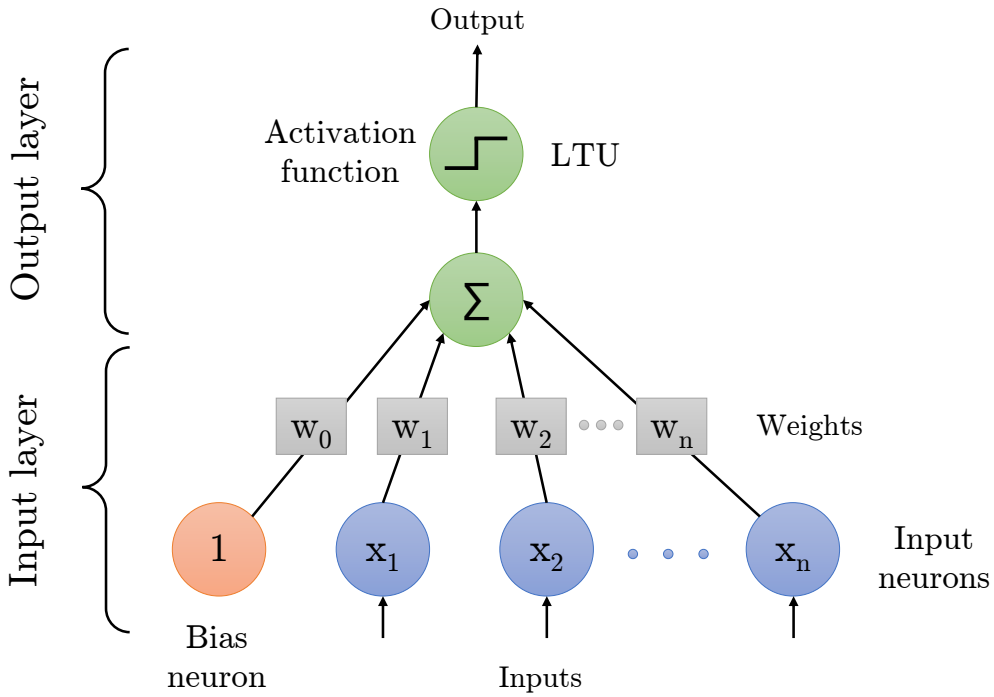


Figure 2.11: An illustrative diagram of an artificial Perceptron [Rosenblatt, 1958]. The output of the Perceptron equals the weighted sum of the inputs (and a general bias feature $b = 1$) through a linear threshold unit. A training algorithm is then can be used to find the best values for $\mathbf{w}$ that minimise a loss function L.

Rosenblatt [1958] proposed to apply *Hebbian learning* [Hebbs, 1949] for training a Perceptron in order to find the optimal weights $\mathbf{w}^*$. The learning algorithm increases the connection between two neurons whenever they output the same value. So the new connection weight

value $w_{i,j}^{new}$ can be calculated as:

$$w_{i,j}^{t+1} = w_{i,j}^{t} + \alpha \left(\hat{y}_j - y_j\right) x, \tag{2.3}$$

where $w_{i,j}^{t}$ is the connection weight between the i-th input x_i neuron and the j-th output $\hat{y}_j$ neuron at the current step t. y_j is the output neuron at the current step, and α is the learning rate. Multiple Perceptrons can be stacked together in layers producing composite linear combinations of their inputs, which is called mulit-layer Perceptron (MLP). Figure **??** shows an MLP of an input layer of M neurons, output layer of K neurons, and L hidden layers each of size N_l neurons. Where each layer is fully connected to the next layer (except the output layer) and contains by default a bias neuron. A deep neural network (DNN) contains at least two hidden layers.

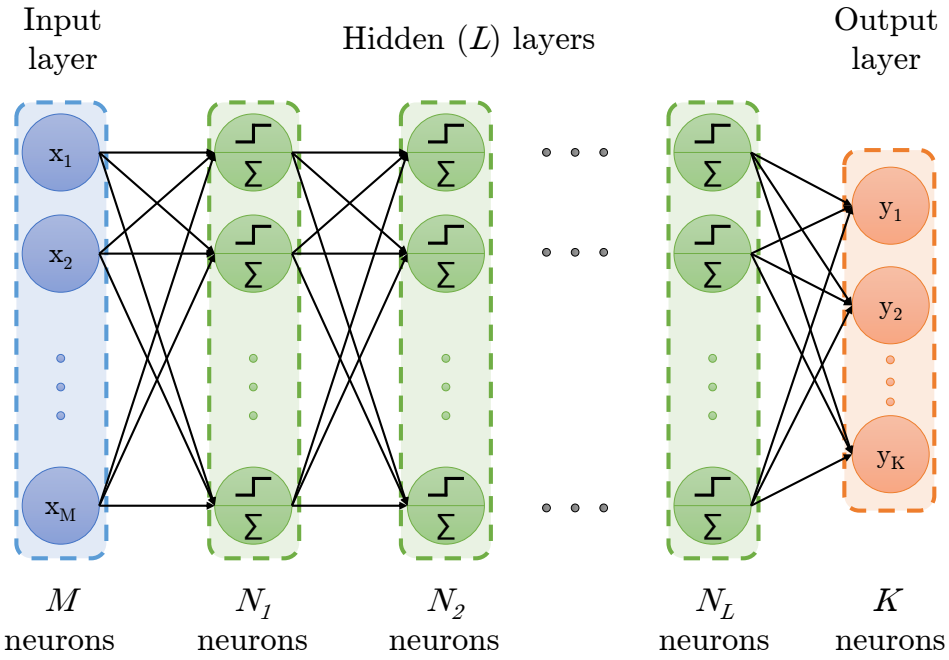


Figure 2.12: An illustrative diagram of an MLP of an input layer $\mathbf{x}$ of M neurons, output layer $\mathbf{y}$ of K neurons, and L hidden layers each of size N_l neurons.

Rumelhart et al. [1985] proposed a *backpropagation* training algorithm to train MLPs using gradient descent. During training, the output of each neuron at every layer is calculated for each training instance (forward pass). Then the error between the ground truth and network's prediction is measured ($\hat{y}-y$). After that, the contributed error from each neuron of the previous

hidden layer to the previous measured error is calculated. This is done to all the hidden layers until reaching the input layer (backward pass). The last step is applying a gradient descent step to all the connection weights in the network using the measured error gradients. This training algorithm can be summarised into four main steps: (*i*) calculate predictions through the forward pass; (*ii*) measure the prediction error; (*iii*) propagate the error to the previous layers through the backward pass; and (*iv*) update the weights using gradient descent. Also to insure that the gradient descent works, the step function is replaced with the logistic function:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}. \tag{2.4}$$

Nowadays, different *activation functions* are being used for the backpropagation algorithm such as: the hyperbolic tangent function, $\tanh(z) = 2\sigma(2z)$; and rectified linear unit function, $ReLU(z) = \max(0, z)$ [Dahl et al., 2013]. The typical MLP architectures modify the output layer by replacing the individual activation functions by a shared *softmax* function, which is called a *feed forward network* (FNN). The softmax function is defined as:

$$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{j=1} k \exp(z_j)}. \tag{2.5}$$

The output of a softmax function is squashed between $(0, 1]$, which can be used to represent probability distribution over the predicted K outputs of the network.

Convolutional neural networks (CNN)

CNNs are the most popular DNN architecture currently used in medical image analysis [Litjens et al., 2017]. They are very similar to ANNs and consist of learnable weights and biases. The standard architecture of CNNs was inspired from the visual system and the hierarchical connectivity pattern between neurons of the animal visual system Fukushima and Miyake [1982]. Where each cortical neuron responds to a restricted region of the receptive field. The whole visual field is covered by partially overlapping the different receptive fields of every neuron. The hierarchical structure of layers allows the network to extract low-level features in the first

hidden layer, then assembles them into higher-level features in the next layers. The main distinctive characteristics of CNNs compared to regular ANNs can be summarised into three points:

1. **Local receptive fields**: The architecture of an ANN does not consider the spatial structure of the images. Each neuron treats all input pixels exactly the same, including very far and close pixels. In contrast, CNNs use a spatial architecture by connecting each neuron with a small, localised regions of the input image (*local receptive field*). The output of each layer is then constructed by sliding the local receptive field across the entire previous (input) layer.

2. **Shared weights**: ANNs transform a single vector input through layers of fully connected neurons, and the neurons in each single layer do not share any connections and completely independent. While in CNNs, the same weights are shared for each hidden neuron during the convolution operation. This means that all the neurons for each layer detect exactly the same feature at different locations in the input. This makes the network translation-invariant, where all neurons shared the same features. Another advantage of sharing weights is that it significantly reduces the number of used parameters in the network. A complete convolutional layer consists of several feature maps.

3. **Pooling layers**: CNNs offer the use of another type of layers to simplify the output of the convolutional layers by subsampling feature map outputs. For example, *max-pooling* is a common pooling layer that outputs the maximum activation in a 2×2 input region. The pooling operation is then applied to each feature map separately. Pooling layers also reduces the number of parameters needed in the subsequent layers. However, it may decrease the accuracy of segmentation networks because of throwing away the exact positional information.

The typical architecture of CNNs is usually composed of convolutional and pooling layers, while the exact details of different models vary greatly. For example, LeCun et al. [1998] proposed one of the earliest CNN architectures *LeNet-5*, inspired by [Fukushima and Miyake, 1982].

Figure **??** shows the architecture of a LeNet model. The lower layers consists of convolution and max-pooling layers, while the higher layers represent a traditional fully connected MLP.

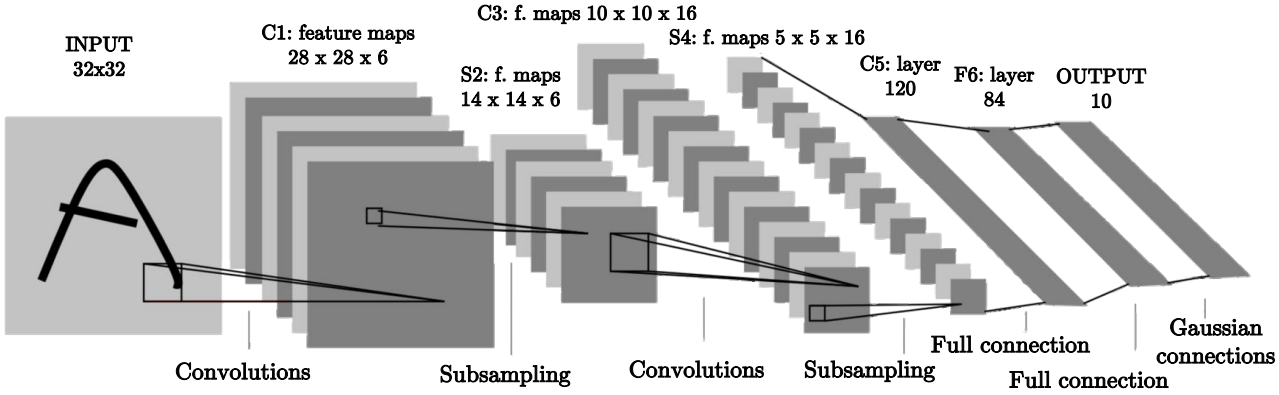


Figure 2.13: The architecture of LeNet-5 for handwritten digits recognition using the MNIST dataset ($60,000$ training and $10,000$ testing images) [LeCun and Cortes, 2010]. Input: grey scale image of size 32×32. C1: convolutional layers of 6 feature maps and kernel size 5×5 and stride 1. S2: subsampling (pooling) layer with kernel size 2×2 and stride 2. C3: convolutional layers of 16 feature maps and kernel size 5×5 and stride 1. S4: pooling layer with kernel size 2×2 and stride 2. C5: convolutional layers of 120 feature maps and kernel size 5×5 and stride 1. FC6: fully connected layer of 84 units. Output: FC layer of 10 units with softmax activation. Figure courtesy [LeCun et al., 1998].

The LeNet model was used for digit classification, and applied commercially to recognise hand-written numbers on greyscale images of cheques (digitized to 32×32 pixels) [LeCun et al., 1998]. The ability to classify more complex natural images with higher resolution requires more and larger convolutional layers. Krizhevsky et al. [2012] proposed a deeper architecture (*AlexNet*) based on LeNet achieving the best results for ImageNet LSVRC-2012 [Russakovsky et al., 2015] classification with 15.3% error rate, and by large margin compared to the second place 26.2% error rate. The efficient implementation of AlexNet on GPU allowed deeper CNNs to work on bigger images. It makes use of the dropout [Srivastava et al., 2014] technique for regularization to avoid overfitting, and ReLU [Dahl et al., 2013] as an activation function. It uses also an ensemble of seven CNN in order to reduce the error rate from 16.4% to 15.3%. With the increase of the computing power and availability of more data, bigger and deeper architecture have been proposed in the recent years, such as VGG-16 [Simonyan and Zisserman, 2014], GoogLeNet (Inception) [Szegedy et al., 2015], and ResNet [He et al., 2016].

CNNs for Semantic Segmentation

In medical image analysis, a popular task is segmenting an image into different labels by assigning each pixel to an object class (*e.g.* white and grey matter in brain MRI). This process is called *semantic segmentation*, which provides an understanding of images at the pixel level. Typical semantic segmentation methods before deep learning relied on engineering features and strong classifiers such as random forests [Shotton et al., 2011], see Section 2.2.2. As with the success of CNNs for image classification, they have also achieved success on segmentation problems. Classification CNNs usually have fully connected layers that require a fixed-size input. Ciresan et al. [2012] proposed a patch-based classification approach for segmentation using CNNs, where each pixel was separately classified using a patch of image centred around it. However, this approach is computationally slow, as the network has to extract and classify a patch around each pixel from the input image. Another common approach is to apply encoder-decoder networks, whereas the encoder represents a typical CNN and the decoder learns to up-sample semantically lower resolution features from the encoder output to the pixel space (*dense predictions*). Long et al. [2015] proposed a fully convolutional networks (FCN) for semantic segmentation by generating dense predictions using up-convolution layers and without any fully connected layers. FCNs can predict segmentations for images of any size, and are computationally faster than patch-based approaches. This work inspired the majority of the current CNN-based approaches for semantic segmentation. Sub-sampling layers or strided convolutions increase the field of view and aggregate bigger context from the input image. However, they discard the exact spatial information, while the precise alignment of class maps to input images is essential for accurate segmentation results. Long et al. [2015] introduced applying skip connections to combine high layer information with low layer information, which results in predicting finer details while reserving coarse semantic information, see Figure **??**.

Ronneberger et al. [2015] proposed an encoder-decoder architecture (*U-Net*) based on a bigger FCN [Long et al., 2015]. Another main difference between U-Net and FCN is the ability to learn up-sampling feature maps instead of up-sampling predictions in the decoder network, as shown in Figure **??**. Later, Çiçek et al. [2016] extended the 2D U-Net for 3D data.

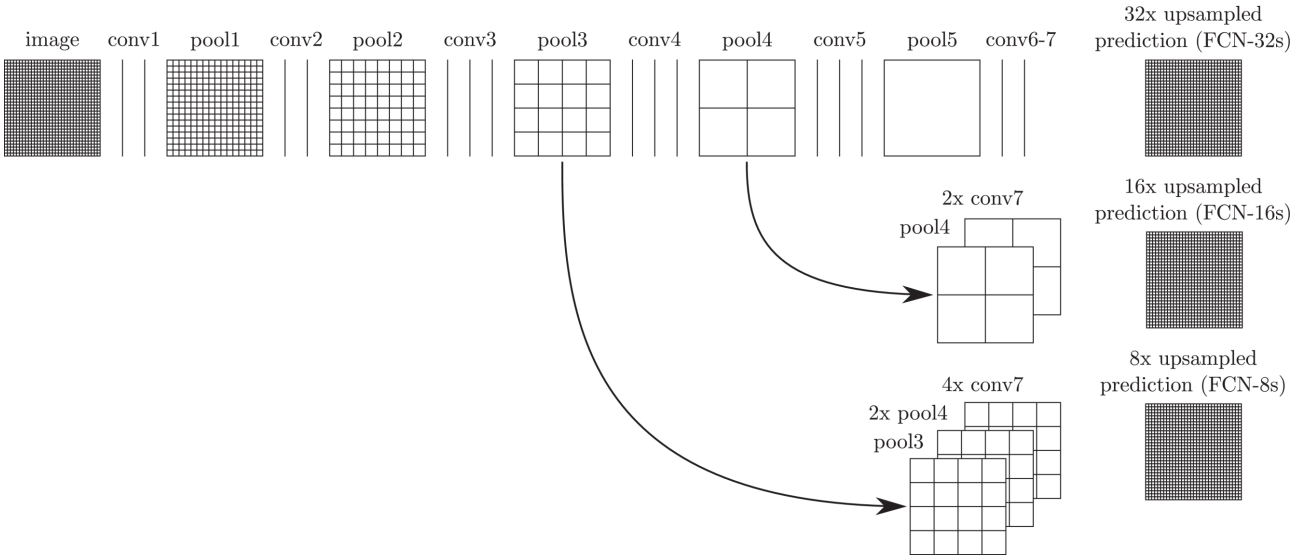


Figure 2.14: FCN with a skip architecture of a three-stream network. The first row (FCN-32s) shows a typical CNN, while its output predictions are upsampled with a stride 32. The second row (FCN-16s) combines the predictions of *pool4* upsampled with a stride 16. The third row (FCN-8) adds the predictions from *pool3* layer upsampled with a stride 16. Figure courtesy [Long et al., 2015].

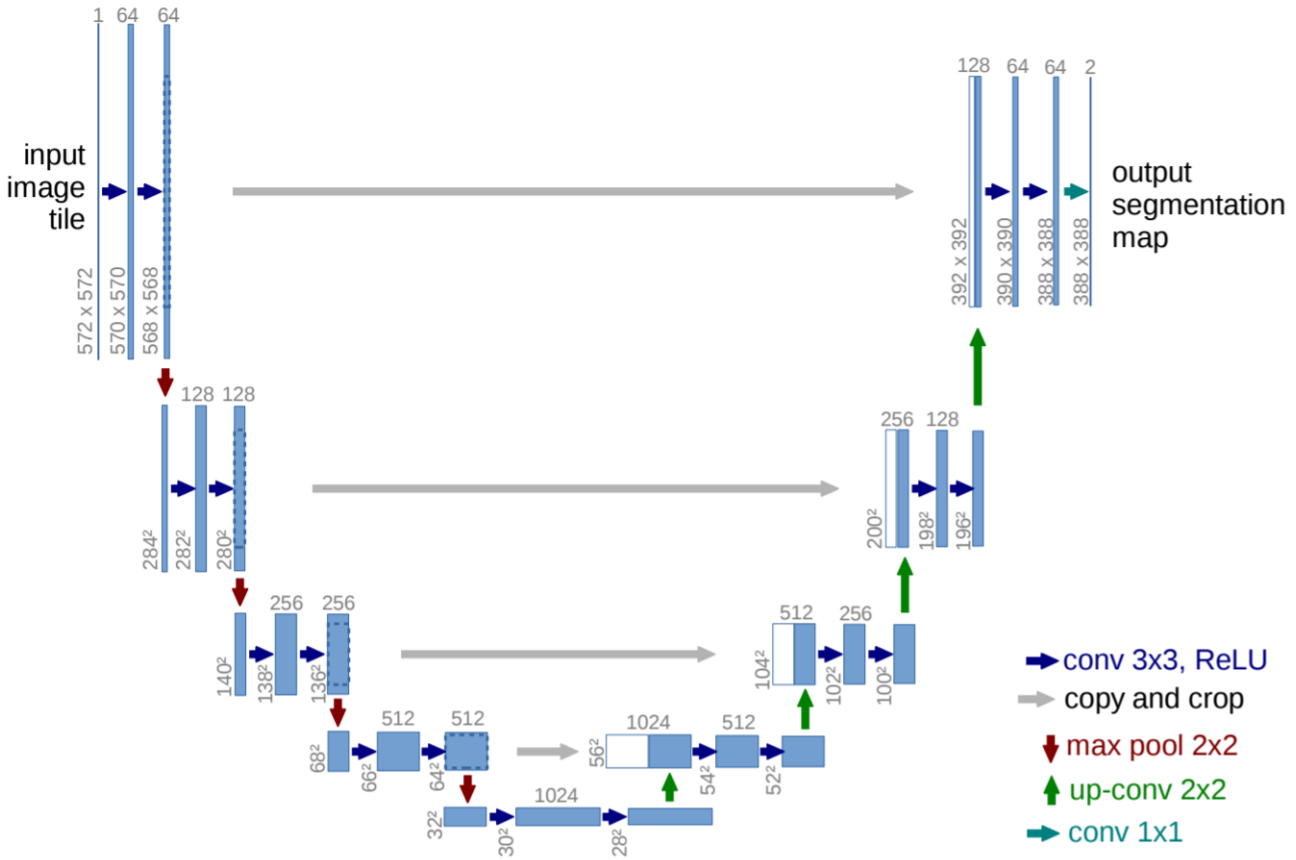


Figure 2.15: The architecture of U-Net showing the encoder-decoder networks. Skip connections here share feature maps at different layers. Figure courtesy [Ronneberger et al., 2015].

For medical images with larger field of view, large image context is necessary for accurate predictions. The naïve approach to increase the input context is to use larger patches, however, this will significantly increase the number of parameters of the network, and hence will require a bigger memory. Kamnitsas et al. [2017] proposed a multi-scale efficient 3D CNN (*deepMedic*) to increase the context of the input image. Inspired by [Farabet et al., 2013], deepMedic feeds the network down-sampled along with the original sampled high-resolution representations of the input, as shown in Figure **??**.

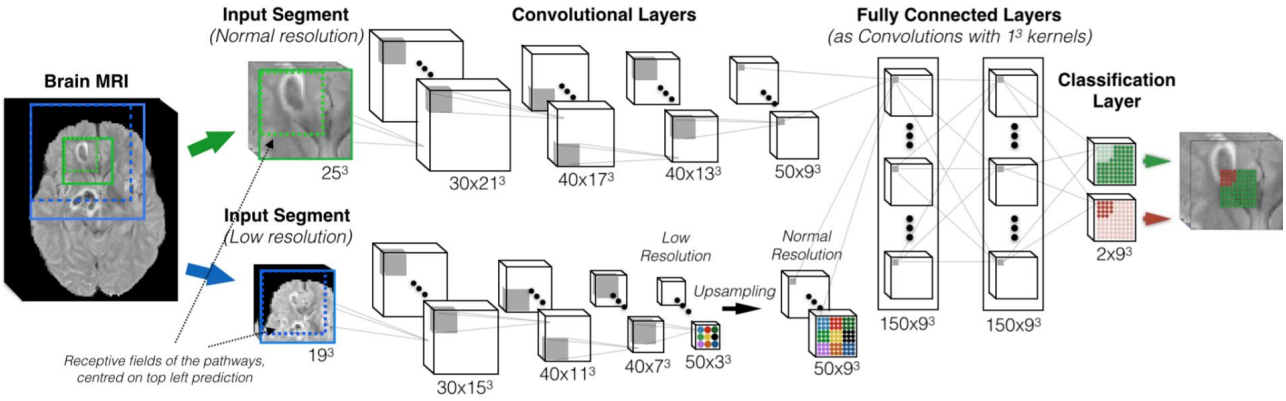


Figure 2.16: The architecture of deepMedic showing two-stream networks. The first network is applied to the high resolution image patch. While the second network is applied to the down-sampled low resolution image for capturing bigger context. The output of the second pathway is then sampled and concatenated with the output of the first pathway. Figure courtesy [Kamnitsas et al., 2017].

2.2.5 Related Work

Foetal Brain Localisation

Anquez et al. [2009] was able to segment the foetal head based on: (*i*) a 3D template registration (matching) to detect the foetal eyeballs (*ii*) followed by a subsequent 2D/3D graph-cut segmentation to finally extract the foetal skull bone content. The motivation behind using the foetal eyes as an initial structure was due to their very small size that can be considered insensitive to image inhomogeneity; their spherical shape is rotationally invariant; and their high contrast with their neighbours. Biometric priors, shape and contrast are used to guide the eyes localisation process. Both biometric priors (e.g. the eyes diameter and the distance between

their centres) and shape information are extracted from manual tracing on the image database. The eyes could be localised successfully using the appearance model, however, it could not be used for segmenting the foetal head. This is because of the high variability of the visual characteristics of the foetal brain in addition to the surrounding maternal tissue. Therefore, a mean shape model is constructed by averaging registered binary shapes. Then the mid-sagittal slice is exctracted using the extracted eyes in the first step to define the orientation of the head. Finally, the shape model is precisely aligned based on the mid-sagittal slice location.

Kim et al. [2010] used a semi-automated brain extraction method as an initial step for their motion correction algorithm of multi-slice foetal MRI scans. Initially, a boundary box around the brain is manually cropped in one reference stack (the first axial stack). Then, this manually extracted rectangular region is used as a template for finding the corresponding boundary boxes in other stacks. The latter step is done by matching the stack against the predefined region of interest in the reference stack using a global rigid body transformation.

Ison et al. [2012] developed a fully automated brain extraction and orientation framework in raw foetal MRI, which combined a random forest classifier and an approximate high-order Markov random field (MRF) solution. Firstly, the maternal tissue is suppressed using a binary random forest classifier based on random offset 3D Haar-like descriptors [Criminisi et al., 2009]. Secondly, a maximal sensitivity filter is applied to the coarse random forest predictions in order to compute the prior brain probability. Thirdly, the final landmark estimates are optimized using a rotation and scale invariant shape model based on a sparse MRF appearance model [Donner et al., 2011] with triangular relationships. Finally, the brain centre, orientation, and bounding box are estimated using a confidence probability summation.

Taleb et al. [2013] proposed an automatic template-based foetal brain extraction from three (at least) orthogonal anatomical images in each direction (axial, coronal and sagittal). Their approach used the intersection of the set of images in order to find a 3D region of interest. Then this 3D ROI is extracted for each original image and registered using FSL-FLIRT [Jenkinson et al., 2002] to an age-specific template [Serag et al., 2012] to obtain an estimation of the brain mask of each orthogonal image. Lastly, the estimated brain masks are fused using 3D

consistency measures based on the normalized correlation between them.

Keraudren et al. [2013] proposed a method for the localisation of the foetal brain from a stack of 2D misaligned slices because of the foetal motion. Their localisation framework consists of four main steps: (*i*) maximally stable extremal elliptical regions are extracted for each 2D slice; (*ii*) estimated brain regions are filtered using prior information of the foetal brain size and aspect ratio; (*iii*) SIFT [Lowe, 1999] are extracted for each selected region (bundled SIFT); (*iv*) histograms of bag-of-words (BOW) [Csurka et al., 2004] are calculated for the selected regions; (*v*) the final fitting 3D cube surrounding the brain is selected using RANdom SAmple Consensus procedure (RANSAC) [Fischler and Bolles, 1981] based on prior knowledge of the brain dimensions.

Keraudren et al. [2014] later extended their foetal brain localisation framework to segment the whole brain for motion correction. They converted the probability map to a binary segmentation using CRF [Lafferty et al., 2001] based on the formulation presented in [Boykov and Funka-Lea, 2006] for the CRF energy function minimisation. They relied on the 3D CRF for refining the output of the random forest classifier by extracting the blob-like structures. Finally, they integrated the automated masking of the foetal brain with the motion correction method presented in [Kuklisova-Murgasova et al., 2012]

Kainz et al. [2014a] developed an automatic brain detection framework using dense rotation invariant feature descriptors based on Spherical Harmonics. First, input images are preprocessed by normalizing the image intensities and applying a non-local median filter [Li et al., 2009] for denoising. Then, rotation invariant features are defined using the angular power spectrum of the expansion coefficients of spherical harmonic functions, which are similar to [Skibbe et al., 2012] but using feature vector that describe structures as small patches of the organ not as a whole. The features extraction function is implemented on graphical processing units (GPU) achieving around 60 times faster feature space transformation. After that, a classification forest [Criminisi and Shotton, 2013] is used for computing a brain probability map using the spherical harmonics features. The generated probability map is thresholded followed by selecting the largest 3D component in order to extract the brain mask. Finally, the thresh-

olded map is used to initialise a 2D level-set [Bradski, 2000], which generate the final accurate segmentation of the foetal brain after a few hundred iterations.

Taimouri et al. [2015] were able to localise the foetal brain using a template-to-slice block matching technique. Firstly, multiple blocks are extracted from the query image using a 2D template block image by computing the sum of square distances (SSD) between them. Since the SSD matrix is rotation variant, the template image is rotated by applying a rotation matrix. The dimension is reduced by breaking the rigid transformation into rotation and translation parts then the parameters of each part are estimated separately. The extracted blocks with highest similarity are then matched using expectation maximisation (EM) in order to estimate the translation value in the next step of the EM process. The final transformation is chosen based on the highest similarity value using the normalized cross correlation, see Figure 2.17. The main drawback of template-matching approaches is that using a rigid template cannot be generalized to a larger testing dataset with high variability in the gestational age range. Another limitation is the lack of the flexibility necessary to deal with motion artefacts as well as foetal abnormalities.

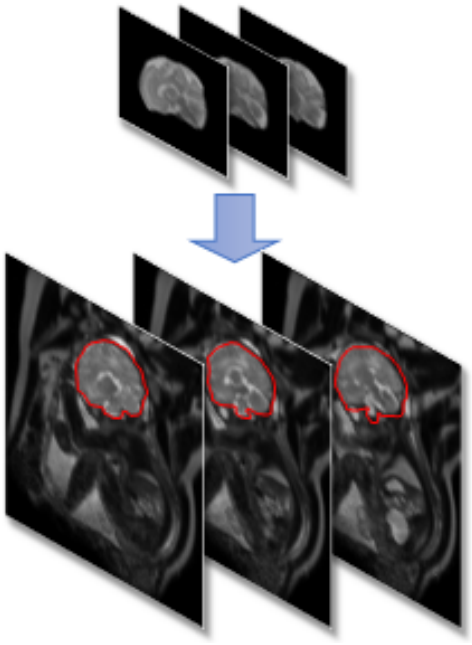


Figure 2.17: A template-to-slice block matching technique of the foetal brain in a 2D MR image.

Foetal Organs and Placenta Segmentation

Previous works for automatic segmentation of foetal MRI were focused on brain segmentation, as shown in the previous section. Keraudren et al. [2015] have recently extended their foetal localisation pipeline to localise other foetal organs. Previous works for the segmentation of the placenta are either semi-automatic [Stevenson et al., 2015] or interactive [Wang et al., 2015a, 2016, 2018] approaches, which require user's interactions. Chapter 4 presents the first work that addresses the fully automatic segmentation of the placenta from MRI.

Stevenson et al. [2015] present a semi-automatic approach for measuring the placental volume from motion free 3D ultrasound with a random walker (RW) algorithm. Their method shows a good inter-observer reproducibility but requires extensive user interaction and several minutes per segmentation. Even though ultrasound is fast enough to acquire a motion free volume, the lack of structural information and weak tissue gradients make it only useful for volume measurements. Wang et al. [2015a, 2016] present an interactive method for the segmentation of the placenta from MR images, which requires user interaction to initialise the localisation of the placenta. Their approach performs well on a small cohort of six subjects but shows a user-dependent variability in segmentation accuracy. Recently, Wang et al. [2018] have proposed a deep learning-based interactive segmentation work for multiple organs from foetal MRI, including the placenta. Their framework incorporates CNNs into a bounding box and scribble-based segmentation pipeline. A weighted loss is used for fine tuning based on the network and interaction-based uncertainties.

Keraudren et al. [2015] developed an automated framework for the localisation of the foetal heart, lungs, and liver using random forests with steerable features [Zheng et al., 2007]. The localisation framework is divided into four main steps: (*i*) preprocessing, (*ii*) detecting candidate locations for the heart, (*iii*) detecting candidate locations for the lungs and liver, and (*iv*) spatial optimization of candidate organ centres. During training all the images are mapped into a standard coordinate system using anatomical landmarks. However, during testing they make use of the efficiency of the steerable features by training local features under any rotating coordinate system without rotating the whole 3D volume. These images are then resampled

to an isotropic voxel size according to the gestational age followed by median filtering while preserving edges to attenuate the motion artefacts. In order to detect the candidate locations of the heart, each voxel is oriented to the standard coordinate system; where axis-1 points form the heart to the brain, axis-2 points from the left lung to the right lung, and axis-3 is orthogonal to the first two axes and points from back to front. Assuming that the brain is already detected using [Keraudren et al., 2014], the search region is defined between two sphere radius $R1$ and $R2$ from the brain centre. Then a classification forest is used to calculate the initial probability for each voxel to be located in the heart followed by a regression forest to predict their offset to the heart. They used four types of features for the random forests, namely, image intensity, gradient magnitude, gradient orientation, and distance to the brain centre. The final candidates of the centre of the heart are selected using a voting process with the highest probabilities. After selecting the heart candidates, the candidate locations for the lungs and liver are detected using the same steps. However, the random forest classifies each voxel as a heart, left lung, right lung, liver, or background. The same features are used for classification but replacing the distance to the brain with the distance to the heart, and adding the projection of the voxel on the heart-brain axis as a new feature to locate the current voxel below or above the heart relative to the brain. Then the voting results from the last regression random forest classifier are smoothed with a Gaussian filter. Finally, candidate organs centres are spatially optimized using local maxima based on their individual voting scores and their spatial configuration in order to obtain a rough segmentation.

2.3 Anatomical Landmarks Detection and View Planning

This section explains the recent advances in reinforcement learning and their application to medical imaging analysis. This is followed by summarising the related work for landmarks detection and view planning.

2.3.1 Background

Anatomical Landmarks Detection

Accurate detection of anatomical landmarks from medical images is an essential step for many image analysis and interpretation methods. For instance, the localisation of the anterior commissure (AC) and posterior commissure (PC) points in brain images is required to obtain the optimal view of the mid-sagittal plane. This can be used as an initial step for image registration [Ardekani et al., 1997] or for the identification of pathological anatomy [Stegmann et al., 2005]. Another example is the automated localisation of standard views such as 2- and 4-chamber views in cardiac MRI examinations. This usually requires automatic landmark detection [Le et al., 2017, Lu et al., 2011] as a key step. Landmark localisation can also be used to initialise deformable models and atlas-based approaches for the evaluation of cardiac ventricular health [Bai et al., 2013]. Furthermore, in foetal imaging, anatomical landmarks are used for estimating qualitative scores of foetal biometric measurements, such as: foetal growth rate, gestational age, and to identify abnormalities [Rahmatullah et al., 2012]. They are also required in order to identify standardized views such as transventricular and transcerebellar planes, which are commonly used in clinical practice for foetal health screening [Li et al., 2018a].

Since manual landmark annotation is time consuming and error prone, automatic methods were developed to tackle this problem. The design of such methods is challenging due to varying organ morphology, orientation, pathology, and image quality. Ghesu et al. [2016] showed that an RL agent is capable of finding accurately landmarks in unseen images. One of the main advantages of applying RL to the landmark detection problem is the ability to learn simultaneously both a search strategy and the appearance of the object of interest as a unified behavioural task for an artificial agent. This approach does not require hand-crafted features and can be trained end-to-end. Moreover, RL has the power to perform in a partial field-of-view or with incomplete data, which can be useful for real-time applications.

Automatic View Planning

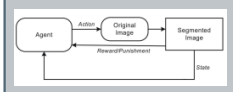
In medical imaging, obtaining accurate biometric measurements that are comparable across populations is essential for diagnosis and supporting critical decision making. For this purpose, standard view planes through a defined anatomy are commonly used in clinical practice to establish comparable metrics. Finding these planes in an imaging examination through a 3D volume is slow and suffers from inter-observer variability. The neuro-imaging community defines a standard (axial) image plane by adopting the anterior-posterior commissure (ACPC) line. Transforming an image to the ACPC coordinate system includes a number of steps: (*i*) marking the AC point, (*ii*) obtaining the optimal view of the ACPC and the mid-sagittal plane, and (*iii*) marking the PC point. Accurate detection of the mid-sagittal plane is useful for the initial step in image registration [Ardekani et al., 1997]. It is also used in evaluation of pathological brains by estimating the departures from bilateral symmetry in the cerebrum [Stegmann et al., 2005]. Similarly, in cardiac MRI standard views are used to assess anomalies. Because of the complexity of cardiac anatomy, the appearance of relevant structures can exhibit large variance according to the positioning of the imaging plane. During conventional cardiac MRI acquisition, the localisation of short and long-axis of the heart requires a multi-step approach that involves double-oblique slices, exhibiting both inter and intra-observer variance [Lu et al., 2011]. These steps include: (*i*) whole 3D pilot image acquisition, (*ii*) left ventricle (LV) localisation, (*iii*) short axis orientation, (*iv*) 3-chamber view calculation, (*v*) landmark detection in mid-ventricular slices, and (*vi*) 4- and 2-chamber view calculation.

2.3.2 Reinforcement Learning for Medical Image Analysis

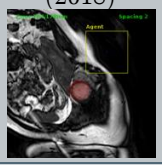
Machine learning enables automatic methods to learn from data to either make a decision or take an action. As discussed earlier in Section 2.2.2, ML algorithms can be classified into three main categories: unsupervised, supervised and reinforcement learning. Unsupervised learning methods rely on exploring and inferring hidden structures from unlabelled data. In a supervised manner, the learning is done from a training set of labelled examples provided by an expert. Reinforcement learning (RL) involves learning by interaction with an environment,

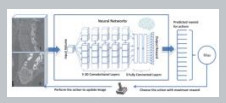
which allows artificial agents to learn complex tasks that may require several steps to reach a solution [Sutton and Barto, 1998b]. RL has been applied to several medical imaging applications such as landmark detection [Ghesu et al., 2016, 2017a,b], tissue localisation [Maicas et al., 2017] and segmentation [Sahba et al., 2006, Shokri and Tizhoosh, 2003], image registration [Krebs et al., 2017, Liao et al., 2017], and view planning [Alansary et al., 2018], see Figure 2.18. The following is a brief overview of the theory behind RL followed by the application of deep learning to approximate its solution.

Image Segmentation	Image Localization	Landmark Detection	Image Registration	View Planning
RL for image thresholding and segmentation Shokri et al. (2003) Sahba et al. (2006)	Deep RL for Active Breast Lesion Detection from DCE-MRI Maicas et al. (2017)	Artificial agent for anatomical landmark detection in medical images Ghesu et al. (2016, 2017) Alansary et al. (2018)	Artificial Agent for Robust Image Registration (rigid, non-rigid, 2D/3D) Liao, R. et al., Krebs J. et al., Miao, S. et al. (2017)	Automatic view planning using deep RL agents Alansary et al. (2018)









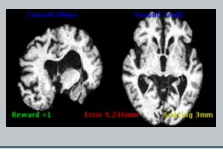


Figure 2.18: Previously published RL works with application to medical imaging analysis.

Reinforcement Learning

Inspired by behavioural psychology, reinforcement learning (RL) can be defined as a computational approach for learning by interacting with an environment so as to maximise cumulative reward signals [Sutton and Barto, 1998b]. A learning agent interacts with an environment E at every state s, as shown in Figure **??**. A single decision is made to choose an action a from a set of multiple discrete actions A. Each valid action choice results in an associated scalar reward, defining the reward signal, R. This sequential decision making can be formulated as a Markov decision process (MDP), where each s_t and a_t are conditionally independent of all previous states and actions holding the Markov assumption. The main goal is to learn an optimal policy that maximises not only the immediate reward but also subsequent future rewards. The

optimal function can be computed directly given the whole MDP using dynamic programming. However, in many applications (including in medical imaging) the MDP is usually incomplete, where the agent cannot directly observe all states. RL approximates the optimal function iteratively by sampling states and actions from the MDP, and learning from experience. There are several algorithms to solve an RL problem such as certainty equivalence, temporal difference (TD) and Q-learning. Recently, Q-learning have shown success in several medical image analysis tasks [Ghesu et al., 2016, 2017a,b, Krebs et al., 2017, Liao et al., 2017, Maicas et al., 2017, Sahba et al., 2006]. Similarly, in this work, Q-learning methods are used as a solution for the RL problem formulation of landmark detection.

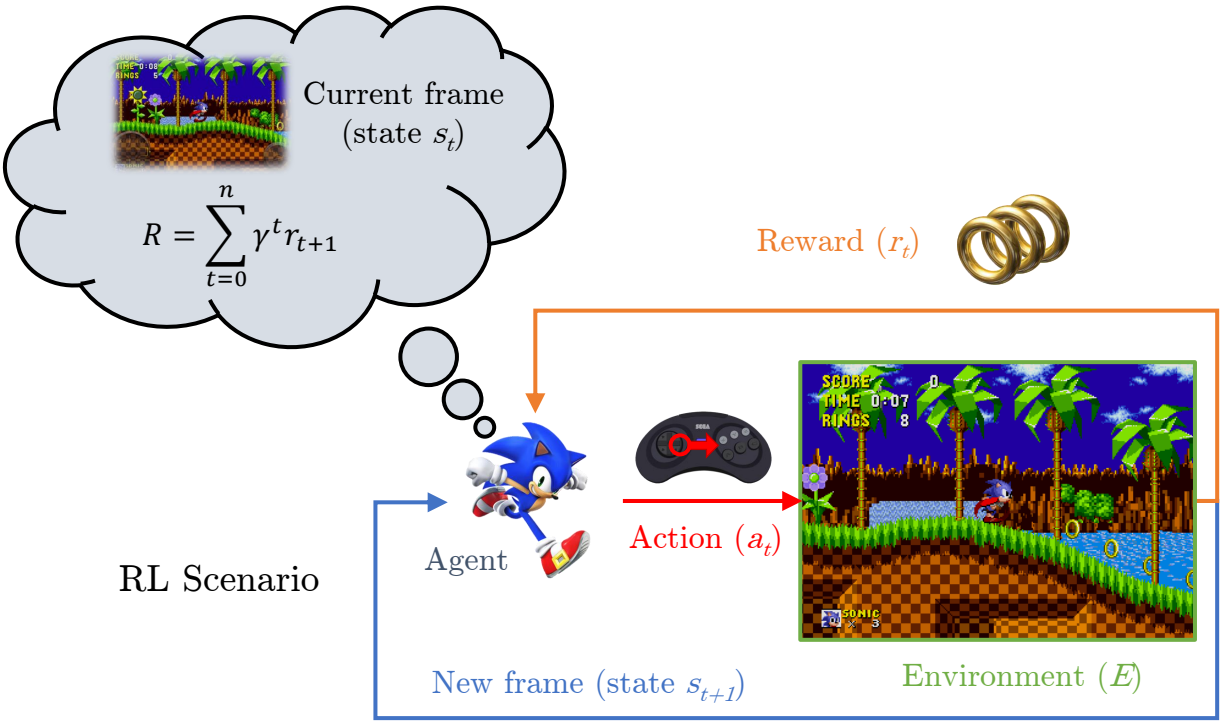


Figure 2.19: Illustrative diagram of a reinforcement learning scenario on video games. The agent interacts with the environment by taking actions and receiving states (frames). The optimal performance can be achieved by maximising the accumulated reward seen during the agent's lifetime.

Q-learning

Learning an optimal RL policy is defined as learning to map a given state to an action by maximising the sum of numerical rewards r seen over the agent's lifetime. The optimal action-selection policy can be identified by learning a state-action value function $Q(s, a)$ [Watkins

and Dayan, 1992], which measures the quality of taking a certain action a_t in a given state s_t. The Q-function is defined as the expected value of the accumulated discounted future rewards $E[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} | s, a]$. $\gamma \in [0, 1]$ is a discount factor that is used to weight future rewards accordingly. It can represent the uncertainty in the agent's environment by providing a probability of living to see the next state. This value function can be unrolled recursively (using the Bellman Equation Bellman [2013]) and can thus be solved iteratively:

$$Q_{i+1}(s, a) = E\left[r + \gamma \max_{a'} Q_i(s', a')\right], \tag{2.6}$$

where s' and a' are the next state and action. We can find the optimal action for each state by solving Equation 2.6. The optimal action will have the highest long-term reward $Q^*(s, a)$.

Deep Q-learning (DQN)

The advent of deep learning has fuelled the current highly active RL research field. Mnih et al. [2015] proposed using a deep CNN to approximate $Q(s, a) \approx Q(s, a; \omega)$, where ω represents the network parameters. This is known as deep Q-network (DQN), and achieved human-level performance in a suite of Atari games. Approximating the Q-value function in this manner allows the network to learn from large data sets using mini-batches, see Figure 2.20.

A naïve implementation of DQN suffers from instability and divergence issues because of: *(i)* the correlation between sequential samples, *(ii)* rapid changes in Q-values and the distribution of the data, and *(iii)* unknown reward and Q-values range that may cause large and unstable gradients during backpropagation. This can be tackled [Mnih et al., 2015] by using a target $Q(\omega^-)$ network that is periodically updated with the current $Q(\omega)$ every n iterations, where ω^- represents the frozen weights of the target network. Freezing the target network during training stabilizes rapid policy changes. To avoid the problem of successive data sampling, an experience replay memory (D) [Lin, 1993] can be used to store transitions of $[s, a, r, s']$ and randomly sampling mini-batches for training. The approximation of best parameters ω^* can be learned end-to-end using stochastic gradient descent (SGD) of the derivative of the DQN loss

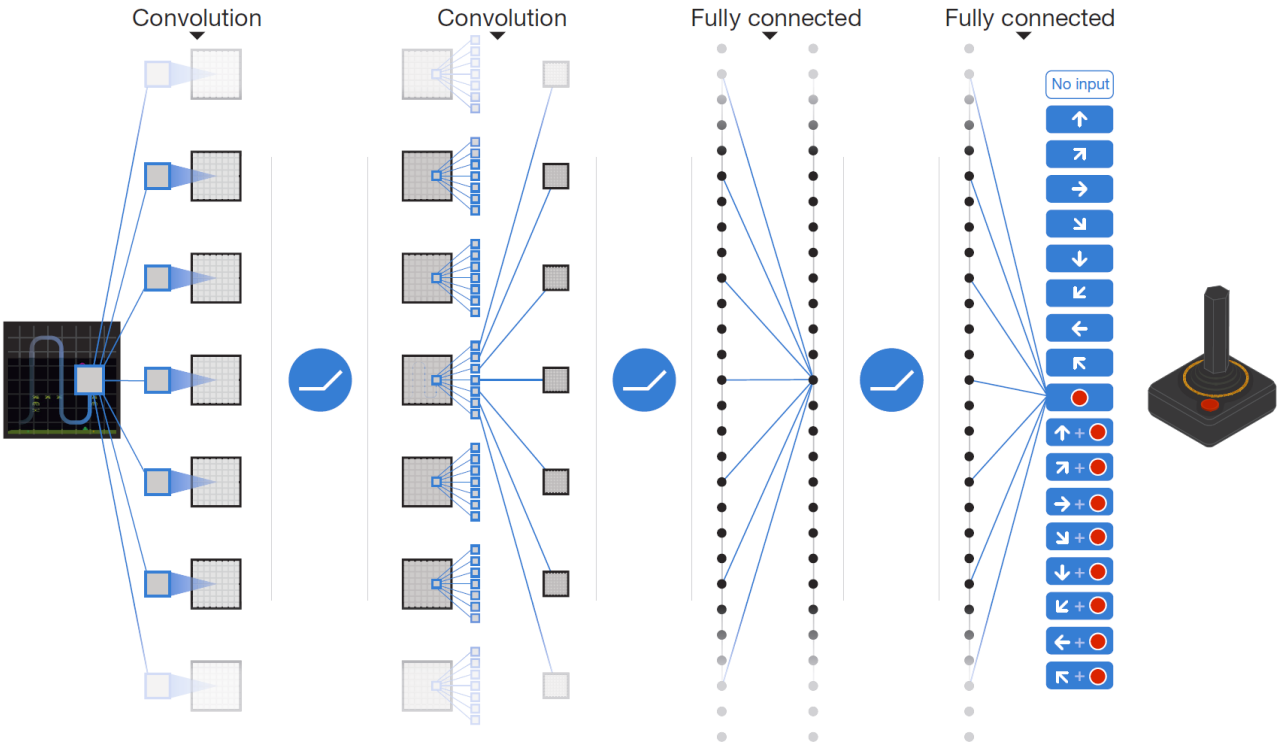


Figure 2.20: The original DQN architecture used for playing Atari games. The network is composed of three convolutional layers and two fully connected layers, where each hidden layer is followed by a ReLU activation function. The input of the network is the last 4 image frames of size 84×84. Figure courtesy [Mnih et al., 2015].

function $\frac{\delta L(\omega)}{\delta \omega}$, where:

$$L_{DQN}(\omega) = E_{s,r,a,s' \sim D}\left[\left(r + \gamma \max_{a'} Q(s', a'; \omega^{-}) - Q(s, a; \omega)\right)^2\right]. \tag{2.7}$$

In order to prevent Q-values from becoming too large, also to ensure that gradients are well-conditioned, rewards r are clipped between $[-1,+1]$. This trick works for most of the applications in practice, however, it may have the drawback of not differentiating between small and large rewards. Below two recent state-of-the-art improvements to the standard DQN are outlined, and evaluated experimentally for landmark detection and view planning in Chapter 5.

Double DQN (DDQN)

In noisy stochastic environments, DQN [Mnih et al., 2015] sometimes significantly overestimates the values of actions [Hasselt, 2010]. This is caused by a bias introduced from using the maximum action value as an approximation for the maximum expected value. The max operator, $\max Q(s', a'; \omega^{-})$, uses the same values to select and evaluate an action resulting in selecting overestimated (overoptimistic) values. Hasselt [2010], Van Hasselt et al. [2016] proposed a solution, a double DQN (DDQN), to mitigate bias by decoupling the selected action from the target network. Thus, the current network is used for the action selection resulting in a modified loss function:

$$L_{DDQN}(\omega) = E_{s,r,a,s' \sim D}\left[\left(r + \gamma \max_{a'} Q(s', Q(s', a; \omega); \omega^{-}) - Q(s, a; \omega)\right)^2\right]. \tag{2.8}$$

DDQN improves the stability of learning and may translate to the ability to learn more complex tasks. The results of DDQN [Van Hasselt et al., 2016] illustrate reduction in the observed overestimation and better performance than DQN [Mnih et al., 2015] on several Atari games.

Duel DQN

Q-values correspond to the quality of taking a certain action given a certain state $Q(s, a)$. Wang et al. [2015b] proposed to decompose this action-state value function into two more fundamental notions of *value*. The first is an action-independent value function $V(s)$, which provides an estimate for the value of each state without having to learn the effect of each action. The second is an action-dependent advantage function $A(s, a)$, which calculates potential benefits of each action. Intuitively, the Q-function learns separately how good a certain state is and how much better taking a certain action would be compared to the others. The new combined dueling DQN function is defined as:

$$Q(s, a) = A(s, a) + V(s). \tag{2.9}$$

This can be implemented by splitting the fully connected layers in the DQN architecture to compute the advantage and state value functions separately, then combining them back into a single Q-function only at the final layer with no extra supervision, see Figure 2.21.

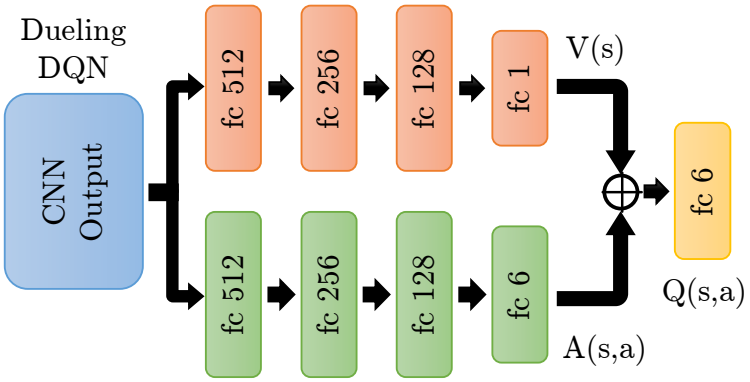


Figure 2.21: Duel DQN architecture, which splits the fully connected (FC) layers into two paths: the state value $V(s)$ and action advantage $A(s, a)$ functions.

Duel DQN can achieve more robust estimates of the state value by decoupling it from specific actions. s is more explicitly modelled, which yields higher performance in general. Duel DQN [Wang et al., 2015b] shows better results than the previous baselines of DQN [Mnih et al., 2015] and DDQN [Van Hasselt et al., 2016] on several Atari games. In summary, duel DQN and DDQN introduced vast improvements in performance compared to DQN, yet it does

not necessarily result in better performance in all environments.

2.3.3 Related Work

Anatomical Landmarks Detection

Typical landmark localisation methods can be categorized into three approaches: registration, appearance-based and image-based. The first category depends on robust rigid or non-rigid image registration techniques to match corresponding points of interest between target and reference images [Potesil et al., 2010, Rueckert et al., 2003]. Appearance-based methods rely on spatial priors that capture the location of different landmarks by learning an appearance model [Milborrow and Nicolls, 2014, Potesil et al., 2015, Zhou et al., 2009]. Image-based methods learn a set of image features located around the anatomical landmarks [Betke et al., 2003].

In the literature, most of the published works have adopted machine learning algorithms for landmark detection by learning a combined appearance and image based model. For example, Criminisi et al. [2013] proposed a regression forest-based landmark detection framework to locate organs in full-body CT scans, which uses Haar-like appearance features. Despite being fast and robust, this approach achieves less accurate localisation results for larger organ structures. Gauriau et al. [2015] extended the work of [Criminisi et al., 2013] by incorporating statistical shape priors derived from segmentation masks with cascaded regression. Oktay et al. [2017] used a stratification-based training model for a decision forest, where the latent variables within the stratified trees are probabilistic. Urschler et al. [2018] proposed a unified random forest framework combining appearance information with geometrical distribution of landmark points. These methods achieve robust results for locally similar structures by learning particular hand-crafted features, extracted from training data. However, the design of such features requires prior knowledge about the points of interest.

With the success of deep learning in different image-based applications, Zheng et al. [2015] proposed a two-stage approach for landmark detection using CNNs. The first stage comprises

of a shallow network with one hidden layer that is used to extract a number of 3D point candidates using a sliding window. This is followed by a deeper network, which is applied on image patches extracted around the selected points. Zhang et al. [2017] proposed a similar approach utilizing two CNNs to learn 3D displacements to a common template, which is followed by another convolutional layer for predicting the coordinates of multiple landmarks jointly. The first network is trained using image patches, whereas the second network shares the same weights from the first network with extra layers. The second network is trained using the whole image instead of patches to learn global information on top of the local information learned by the first network. Payer et al. [2016] adopted a CNN to model spatial configurations to detect multiple landmarks. The first block of their architecture generates local appearance heatmaps for individual landmark locations. Subsequently, the relative position of a single point with respect to the rest of the landmarks is learned through another convolutional kernel. The final heatmap combines both local appearance and spatial configuration between all landmarks. In order to capture global as well as local information, Andermatt et al. [2017] presented a method based on multi-dimensional gated recurrent units combining two recurrent neural networks. The first network detects a candidate region around the point of interest followed by a second network for more accurate localisation. All previous methods rely on learning the search strategy and localisation in two stages. This may increase the possibility that the second stage misclassifying multiple candidates from the output of the first stage as positive.

Ghesu et al. [2016] adopted a deep RL-agent to navigate in a 3D image with fixed step actions for automatic landmark detection. The artificial agent tries to learn the optimal path from any location to the target point by maximising the accumulated rewards of taking sequential action steps. Xu et al. [2017], inspired by [Ghesu et al., 2016], proposed a supervised method for action classification using image partitioning. Their model learns to extract an action map for each pixel of the input image across the whole image into directional classes towards the target point. They use an FCN with a large receptive field to capture rich contextual information from the whole image. Their method achieves better results than using an RL agent, however, it is restricted to 2D or small sized 3D images due to the computational complexity of 3D CNNs. In order to overcome this additional computational cost, Li et al. [2018a] presented

a 2.5D patch-based iterative CNN to detect individual or multiple landmarks simultaneously. Furthermore, Ghesu et al. [2017a,b] extended their RL-based landmark detection approach to exploit multi-scale image representations.

View Planning

Ardekani et al. [1997] proposed a method to automatically detect the mid-sagittal plane in 3D brain images by maximising the cross-correlation between the two image sections on either side of the sought plane. Stegmann et al. [2005] proposed to use a sparse set of profiles in the plane normal direction and maximise the local symmetry around them. In [Le et al., 2017, Lu et al., 2011], they proposed an automatic view planning algorithm for cardiac MRI acquisition. Their methods are based on learning the anatomy segmentation and detecting anchor landmarks in order to calculate standard cardiac views. These methods require prior knowledge of the whole 3D image for the purpose of plane detection. This involves manual annotation of anatomical landmarks, which is a tedious and time-consuming task. In our method, we use the acquired standardized views for cardiac scans in training without any manual labelling.

Chapter 3

Motion Correction and Reconstruction

This chapter is based on:

Amir Alansary, Martin Rajchl, Steven G McDonagh, Maria Murgasova, Mellisa Damodaram, David FA Lloyd, Alice Davidson, Mary Rutherford, Joseph V Hajnal, Daniel Rueckert, and Bernhard Kainz. **PVR: patch-to-volume reconstruction for large area motion correction of fetal MRI**. *IEEE transactions on medical imaging (TMI)*, 36(10):2031–2044, 2017. [Alansary et al., 2017]

Abstract

This chapter presents a novel method for the correction of motion artefacts that are present in foetal Magnetic Resonance Imaging (MRI) scans of the whole uterus. Contrary to previous slice-to-volume registration (SVR) methods, requiring an inflexible anatomical enclosure of a single *investigated organ, the proposed patch-to-volume reconstruction (PVR) approach is able to reconstruct a large field of view of non-rigidly deforming structures. It relaxes rigid motion assumptions by introducing a specific amount of redundant information that is exploited with parallelised patch-wise optimization, super-resolution, and automatic outlier rejection. An efficient parallel implementation of PVR is also provided for allowing its execution within reasonable time on commercially available graphics processing units (GPU), enabling its use in the*

clinical practice. PVR improves the reconstruction accuracy in the presence of affine motion artefacts compared to conventional SVR in synthetic experiments. Furthermore, the performance of PVR is evaluated qualitatively and quantitatively on real foetal MRI data subject to maternal breathing and sudden foetal movements. The evaluation is done with respect to the originally acquired data using threee quality metrics: peak-signal-to-noise ratio (PSNR), structural similarity index (SSIM), and cross correlation (CC). Dissimilarity maps based on SSIM are shown to provide a method for visual inspection of reconstruction uncertainty. Furthermore, distance errors were measured for selected anatomical landmarks in the foetal head, as well as the mean and maximum displacements resulting from automatic non-rigid registration to a motion-free ground truth image. These experiments demonstrate a successful application of PVR motion compensation to the whole foetal body, uterus and placenta.

3.1 Introduction

The main aim of the proposed method in this chapter is to facilitate the automatic reconstruction of whole collections of motion corrupted stacks without the need of corresponding image segmentations. By treating rigid image patches as piecewise constant segments of organs further allows limited correction of non-rigid tissue motion. This patch-to-volume reconstruction (PVR) finds automatically rigidly connected areas, which can be used as segmentation prior for further refinement using conventional SVR in small regions of interest. A multi-scale patch approach is also introduced in Section 3.2.1, and the reconstruction quality thoroughly evaluated on the whole uterus including the foetal brain and placenta. The breaking points of SVR and variations of PVR on synthetically motion corrupted brain phantom data is tested in Section 3.5.1. The presented approach is the first method that is able to reconstruct foetal organs and detailed 3D volumes of secondary, non-rigidly moving structures such as the placenta. It also evaluates the non-rigid deformations that may result from PVR using anatomical landmarks annotated in the foetal head and non-rigid registration.

SVR-based motion compensation methods make use of the assumption that rigid regions,

e.g., brain and thorax, of 2D input slices deforms rigidly, where a global 3D volume is reconstructed by iteratively registering these 2D input slices. In order to increase the granularity of the input data, 2D image patches of arbitrary shape are used instead of whole slices for SVR reconstruction. Section 3.2.1 explores square patches and dilated superpixels [Achanta et al., 2012] for the definition of the patch shape. Superpixels provide a method to define semantically meaningful regions while reducing the required data redundancy and computational overhead.

3.2 Method

PVR relies on the fact that certain regions of the scanned anatomy are rigid and can be reconstructed with SVR algorithms. However, unlike SVR, it is fully automatic and provides a full field of view reconstruction. Data consistency is obtained by oversampling a region of interest at different scan orientations. Robust statistics can be used to identify mis-registered or heavily corrupted data [Gholipour et al., 2010, Kuklisova-Murgasova et al., 2012]. Figure 3.1 depicts a schematic overview of the proposed PVR framework.

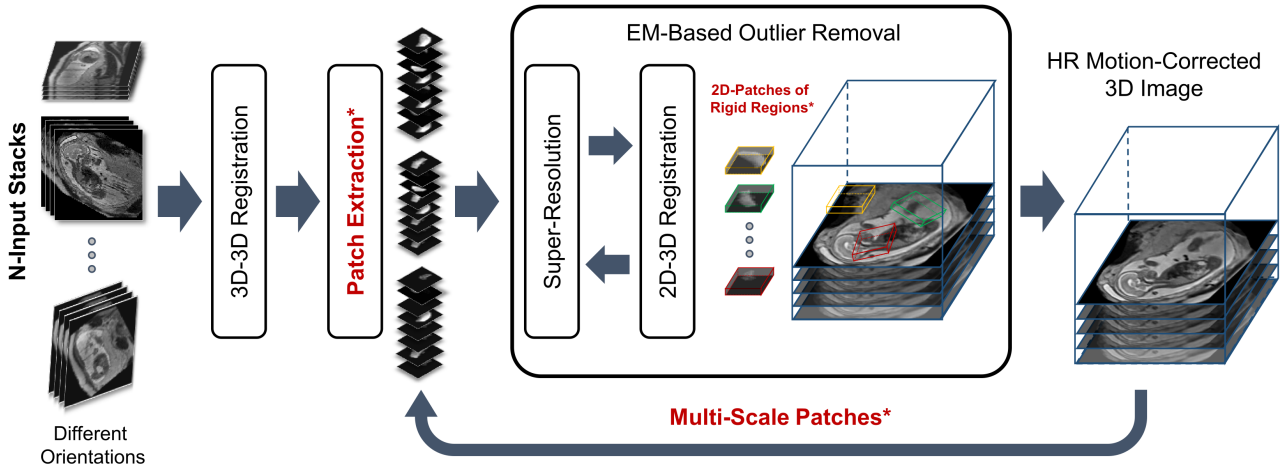


Figure 3.1: Schematic and modular overview of the proposed patch-to-volume reconstruction (PVR) framework. The key parts are 3D-3D registration, patch extraction, 2D-3D registration, super-resolution, and EM-based outlier removal. Core contributions of PVR are written in red and marked with asterisk.

3.2.1 Input Data and Initialisation

A template stack is either manually or automatically chosen from available input stacks by detecting the stack with fewest motion artefacts [Kainz et al., 2015a]. Global intensity matching is applied to normalise intensity values of all input images around the mean of the reference image. This is followed by global 3D-3D alignments to spatially initialise the reconstruction target. Input data can be represented as stacks of 2D images (patches) consisting of:

$$Y = \{y_s | s \in S\}, \tag{3.1}$$

where y_s is a patch of arbitrary 2D-shape and indexed by the location s. S is the set of all locations in all p stacks, $S = \{s_1, s_2, \ldots, s_M\}$, and M is total number of patches. Each y_s is assigned an initial transformation θ_s from the result of the 3D registration to the reference image. The final transformation of voxel location p_r in X to a pixel location p_k in a patch y_s can be represented as:

$$T = W_k^{-1}(s) \cdot \theta^{-1}(s) \cdot W_r, \tag{3.2}$$

where W_k and W_r represent image-to-world and world-to-image transformations. The transformed voxel $p_s(x, y, z, 1)^T$ from X to a pixel $p_k(i, j, 0, 1)^T$ in y_s is then defined as:

$$p_k = T \cdot p_r. \tag{3.3}$$

Patch Extraction

In the simplest naïve case the shape of y_s is square defined via its edge length a and stride ω. This definition is generally applicable to any kind of oversampled motion corrupted data. If a and ω are fixed, no prior knowledge about the data is assumed. However, ideally each y_s corresponds to a meaningful subregion of the volume in which motion can be characterized as rigid. Typically, square patches are overlapping to provide redundant representations of the same locations. The computational expense of such an approach increases with patch number and size a. Additional consideration must include the inherent trade-off between a and the

assumption of it containing rigid motion.

An alternative to naïve shape definitions of y_s is to find correlation between voxel locations and their neighbors, which can be found by unsupervised image segmentation techniques such as *superpixels* (SP) [Achanta et al., 2012]. These techniques allow similar-sized segments to be obtained from local intensity information (see Figure 3.2) instead of employing dense sampling of overlapping patches, enabling the image reconstruction with fewer but more useful data blocks. Further, reducing the total amount of required data blocks for reconstruction lowers the computational overhead, positively impacting the overall run-time. Additionally, larger rigid areas require less computational effort for image registration and super-resolution, and more importantly less dependency on inherent image data parameters (*e.g.*, voxel spacing, organ size, subject size).

There are several techniques for generating SP in the literature [Achanta et al., 2012, Felzenszwalb and Huttenlocher, 2004, Vedaldi and Soatto, 2008], while a fast and efficient SP approach is desirable for the clinical practice. Simple linear iterative clustering (SLIC) [Achanta et al., 2012] obtains regular SP based on minimising the distance D between the centroids of SP with an initial step size a. D is defined as:

$$D = \sqrt{{d_c}^2 + \left(\frac{d_s}{a}\right)^2 t^2}, \tag{3.4}$$

where d_c and d_s are the intensity and spatial Euclidean distances that are controlled by the adaptive compactness parameter t for each SP. A defined number of SP centers can be approximated using: $N_{sp} \approx k\sqrt{(n/2)}$, where n is the number of voxels in an image of size d_x, d_y and k a constant factor Kainz et al. [2015b]. Here, a regular grid is used for initialisation with a constant distance a between SP centroids so that $N_{sp} = \frac{d_x}{a} + \frac{d_y}{a}$ and dilate by $\gamma\%$. Thus, instead of providing a target number of SP, N_{sp} is defined by the image size and a controls the relative size of the SP.

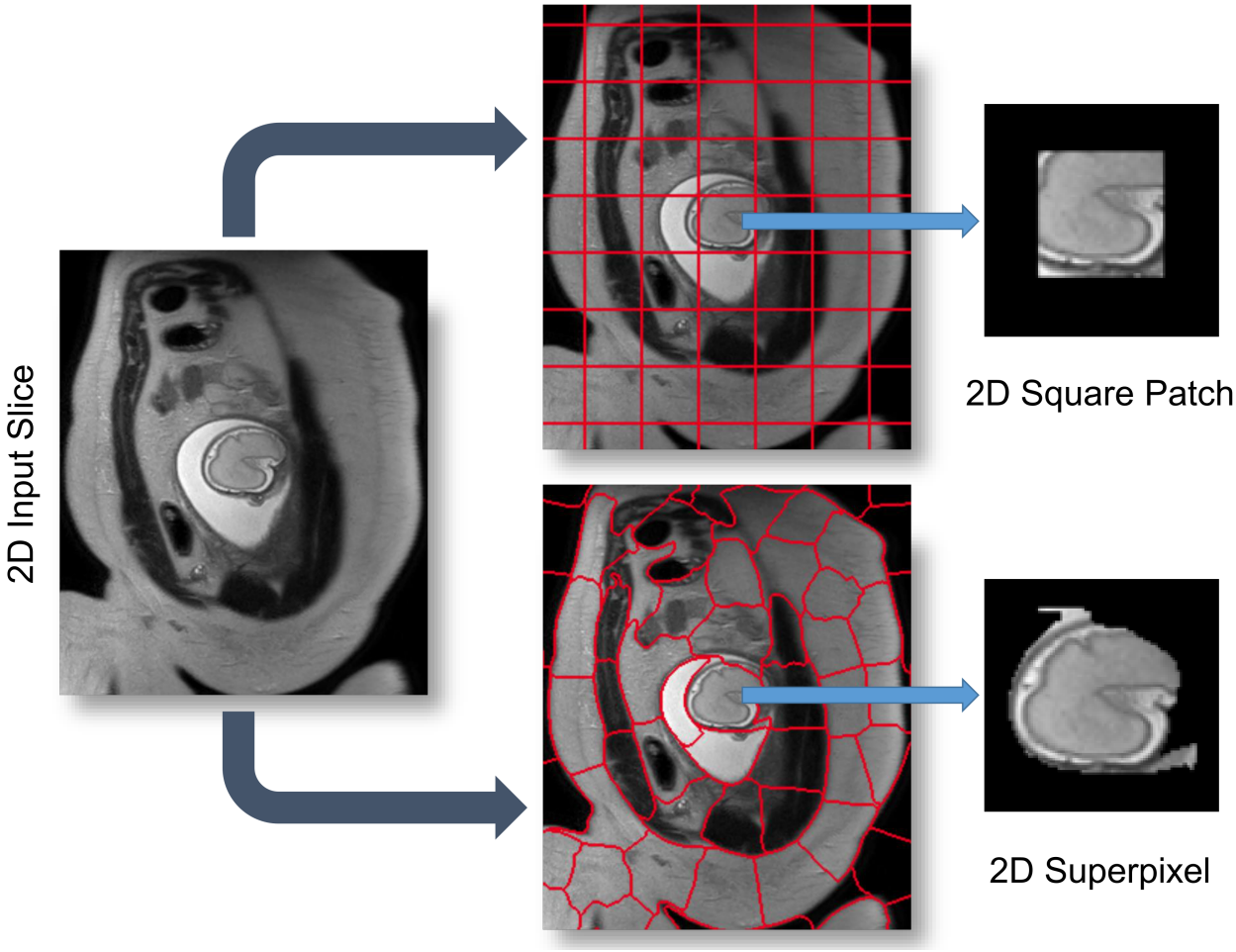


Figure 3.2: An illustrative figure showing both square patches and superpixels methods for the patch extraction step. A 2D superpixel shows more flexibility than a square patch in extracting rigid regions or similar voxels. In practice, superpixels are dilated with few pixels to include some contextual information in order to increase the accuracy of the patch to volume registration step.

Multi-scale Patches

Although larger patch regions are less likely to include rigidly connected regions, they may perform better during 2D-3D registration due to the additional contextual information of each patch. In contrast, smaller patch sizes are more likely to represent rigidly deformed regions, but provide less contextual information, potentially affecting the 2D-3D registration. A good trade-off between the size of the patch region and the likelihood of rigid motion needs to be found. Here, multi-scale patches are used for reconstruction to exploit the advantages of different patch sizes. Input data are represented as stacks of 2D patches:

$$Y_i = \{y_s | s \in S_i\}, \tag{3.5}$$

where, instead of using the same Y as a unique input, a different scale of Y_i is used for each iteration i at $\gamma\%$ of its original size. Similar to Equation 3.1, S_i is the set of all locations in all p stacks but with different size for each iteration i. This is done by re-calling the patch extraction module with a different patch size (see Figure 3.1). Additionally, to increase contextual information for estimating the transformations, y_s superpixel patches are overlapped by dilating each with γ pixels using a flat structuring element b of a fixed neighbourhood (*e.g.* 26 pixels), hence $\bar{y}_s = y_s \oplus b$. Smaller γ values result in a faster reconstruction. Ideally γ is $> 50\%$ of a, ensuring that every pixel is covered by multiple samples.

3.2.2 Patch-to-Volume Registration

An HR-image X is reconstructed from a number of motion corrupted patches y_s using 2D-3D registration-based super-resolution similar to [Kainz et al., 2015a, Kuklisova-Murgasova et al., 2012], where an accurate PSF calculation is used to generate a gradually improving approximation of X and further employed to initialise the 2D-3D registration and computation of robust statistics. The PSF is applied to update the intensity values during the patch-to-

volume registration by updating X iteratively using:

$$X(p_r^{n+1}) = PSF(p_k - p_k^*) \cdot I_s(p_k) + X(p_r^n), \tag{3.6}$$

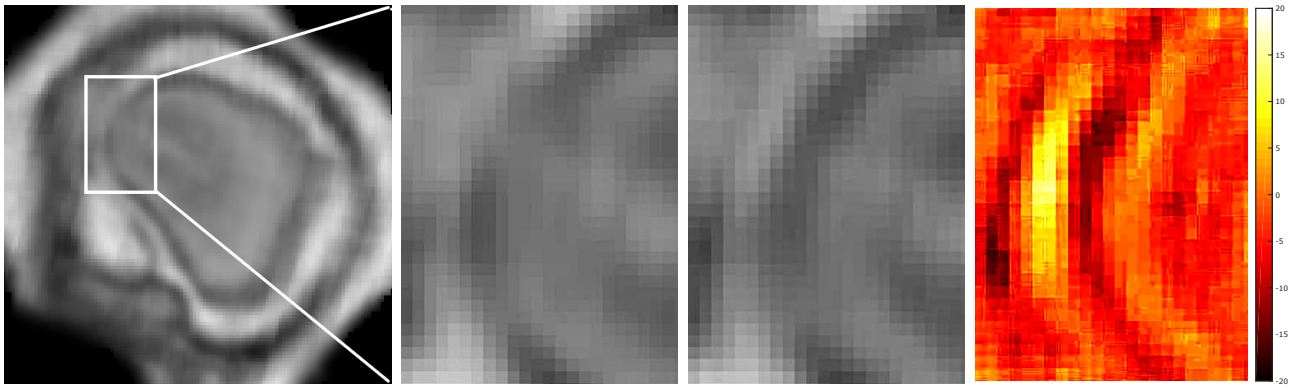
where p_k^* is the simulated point from the current X to patch y_s. $I_s(p_k)$ is the intensity value at location k in patch y_s.

In [Kainz et al., 2015a, Kuklisova-Murgasova et al., 2012], the PSF is a sinc function for the in-plane and is the slice profile for the through-plane, measured for the employed MRI sequence (ssFSE), according to [Jiang et al., 2007]. For a better approximation of small values close to 0, the PSF's implementation is based on the Taylor series approximation of the sinc function. The series is truncated after several terms and bound the remainder based on relative error ϵ. The Taylor series approximation of the sinc function is defined as:

$$sinc(R) = 1 - \frac{R^2}{3!} + \frac{R^4}{5!} - \frac{R^6}{7!} + \cdots . \tag{3.7}$$

This approximation of the PSF achieves qualitative improvement of the reconstructed image compared to the sinc implementation. An example from the first iteration of a foetal brain reconstruction is shown in Figure 3.3. This is expected since Taylor series approximation is common for computations requiring a high level of accuracy close to the machine precision limits of floating point computations [Goldberg, 1991].

During the optimization process, individual 2D patches are continuously rigidly registered to the current 3D reconstruction of X and reintegrated into X using iterative super-resolution with gradient descent optimization. Any similarity metric can be used as a cost function for the registration step such as mutual information [Kuklisova-Murgasova et al., 2012, Rousseau et al., 2006], cross correlation (CC) [Jiang et al., 2007, Kainz et al., 2015a], or mean square intensity differences [Gholipour et al., 2010, Kim et al., 2010]. Choosing the best similarity metric for reconstruction depends on the input data. CC has been found to be effective for input data with similar intensity distribution [Penney et al., 1998]. In these experiments, CC is employed as the similarity metric for 2D-3D registration, after rescaling the intensities between



(a) (b) (c) (d)

Figure 3.3: Example for the observed differences in the first iteration of a foetal brain MRI reconstruction (a). (b) shows a magnified region using a sinc function for the PSF similar to [Kainz et al., 2015b] and (c) shows the result from using a Taylor series approximation of the sinc function as used in this work. Taylor series approximation allows a better approximating of small values close to zero. (d) shows the difference between both images.

the input stacks.

3.2.3 Super-resolution

Given that the position and orientation of each patch y_s relative to X is known at any point k, patches y_s^* can be simulated from sampling the current approximation of X using the PSF. X can subsequently be iteratively improved by minimising the error between y_s and y_s^*. Gradient decent is used for the optimization. The optimization has to be regularised to prevent amplification of registration error, noise and under-sampling artefacts. Adaptive regularisation is applied to reduce smoothing effects. This results in an objective function defined as:

$$\sum_s PSF(p_{s_k} - p^*_{s_k})^2 + \lambda \sum_d \phi \left(\frac{p^n_{r+d} - p^n_r}{\delta \left| d \right|^2} \right), \tag{3.8}$$

with $\phi(t) = 2\sqrt{1+t^2} - 2$, voxel locations $p_r \in X$, d a vector between the index of a voxel and one of its 26 neighbours, and δ a parameter that controls when the gradient between neighbouring voxels is considered to be an edge. For the data used in these experiments, a gradient magnitude between 100 and 400 yields good results. The regularisation term represents edge preserving filtering [Charbonnier et al., 1997] and is used similarly as presented in [Kainz et al., 2015a]. λ

is decreased after each iteration. Following the experimental findings in [Kuklisova-Murgasova et al., 2012] a suitable starting value for λ is $0.8\delta^2$, halved after every iteration.

3.2.4 EM-Based Outlier Removal

Correctly registered patches $\hat{y}_s$ should provide a higher contribution to the final reconstruction, presenting a low error e when compared to the original image data. Gholipour et al. [2010] initially introduced an approach to account for outliers during super-resolution based on Huber function statistics. Similar to [Kuklisova-Murgasova et al., 2012], the expectation maximisation algorithm is employed for outlier removal by classifying $\hat{y}_s$ and the included pixels into an inlier and outlier class. A zero-mean Gaussian distribution $G_\sigma(e)$ with variance σ^2 is used for the inliers and a uniform distribution with constant density:

$$m = \frac{1}{max(e) - min(e)} \tag{3.9}$$

for the outliers. This makes use of available, highly redundant information (*i.e.*, overlapping $\hat{y}_s$), to find partly matching patches and to depreciate or fully reject erroneous voxels. The log-likelihood for each patch is maximised using:

$$y_s|logP(Y, \Phi) = \sum logP(e|\sigma, c), \tag{3.10}$$

to be part of a region of rigid motion. Φ is the current estimate of the reconstructed volume X, the variance σ^2 of the errors e, and the proportion of correctly matched voxels c. The posterior probability for a pixel $\in \hat{y}_s$ being identified as inlier is

$$p = \frac{G_\sigma(e)c}{G_\sigma(e)c + m(1-c)}. \tag{3.11}$$

The updates of c and σ^2 are computed as:

$$\hat{p} = \sqrt{(\sum_{\hat{y}_s} p^2)/N}, \tag{3.12}$$

where N is the number of pixels in $\hat{y}_s$ [Kuklisova-Murgasova et al., 2012]. An inlier and outlier probability is then defined for each $\hat{y}_s$ and excluded from processing if classified as an outlier (*e.g.*, if it contains structures moving in opposite directions during scanning, such as the foetal head and thorax). Only if information in $\hat{y}_s$ is consistent with the originally acquired data, will the registered patch contribute to the SR reconstruction of X.

Identification of rigid regions and SVR refinement: The rigidity of regions is measured by keeping track of the probability p of each pixel of every $\hat{y}_s$. This allows the identification of locations best fitting the rigid 2D-3D registration constraints. Candidate regions, that contain rigid motion components, can be identified by combining p and $\hat{p}$ into a 3D volume utilizing the same PSF as used in the reconstruction. This can further initialise the rigid SVR reconstruction and visualize the data uncertainty during reconstruction.

3.3 Implementation

Parallelization:

The high data redundancy required for the proposed approach makes conventional single threaded implementation practically not feasible. Computational complexity of PVR is exponentially higher than SVR, depending on the employed patch overlap. For optimal performance PVR is implemented via general-purpose programming on graphics processing units (GPGPU) using the compute unified device architecture (CUDA, NVIDIA, Santa Clara, CA) language [Nickolls et al., 2008, Sanders and Kandrot, 2010]. CUDA is a highly evolved single instruction multiple data (SIMD) programming language, which allows a large part of the proposed framework to be mapped onto GPU hardware. Currently, CUDA is the only high-level general purpose GPU language that provides bi-directional texture access via surfaces in a kernel, which is essential for the efficient implementation of certain parts of the reconstruction framework. This section discusses PVR's key implementation details.

A modular design is used to allow experimentation with the separate components of the algorithm. An overview of this design is shown in Figure 3.4. The modules are encapsulated

in a CUDA library, which can be used independently from the instantiating framework. The successor of IRTK[1] is used for interfacing with medical image data.

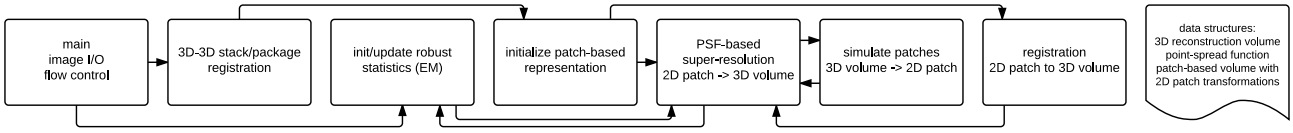


Figure 3.4: The software modules defined for the implementation of the proposed approach. For implementation details, please refer to the provided source code.

PVR exploits three major opportunities for parallel processing:

I. *Patch-level:* Individual patches are mapped to blocks of a CUDA computing grid and the contained voxels are mapped to individual threads. Depending on the used GPU hardware, patch processing can also be mapped directly to the computing grid, such that each thread works on a complete patch (limited by the employed patch size). The resulting thread divergence provides opportunities for advanced GPU scheduling strategies [Steinberger et al., 2012] and for a direct translation of optimization strategies for image registration, for example patch-wise gradient descent.

II. *Voxel-level:* For the parallelization of PSF-based super-resolution and robust statistics, a similar three-fold procedure definition is followed as used in [Kainz et al., 2015a]. The voxels within each patch are processed using kernel level parallelization and parallel pixel-volume, volume-pixel, and volume-volume procedures are applied.

III. *Patch-batch:* PVR scales to multiple GPUs through distributing independent subsets of patches over the desired number of devices. Synchronization is done through averaging of the resulting sub reconstruction volumes on the master GPU. Initial 3D-3D registration is performed on a single master GPU, which allows optimal coalesced memory access.

Availability of Source Code

The source code of a C++/CUDA implementation of the proposed method, including parallelization strategies, is publicly available on GitHub repository [2]. The source code for the

[1] `https://github.com/BioMedIA/MIRTK`
[2] `https://github.com/bkainz/foetalReconstruction`

implementation of PVR is licensed under MIT license.

3.4 Evaluation and Experiments

3.4.1 Evaluation of Global Image Quality

Evaluation of adult brain MRI reconstruction: The performance and limitations of PVR are evaluated in terms of accuracy and robustness with synthetic deformations of *adult* brain data. Similar to [Gholipour et al., 2010], an isotropic 1 mm^3 T2-weighted adult brain phantom with no noise obtained from the *Brainweb* database [Cocosco et al., 1997] is used for this experiment. Three experiments with different deformations are performed, namely:

I. *Rigid Transformation:* Synthetic rigid motion artefacts are generated by translating the axial, sagittal and coronal images with $\pm d_x mm$ in the x-direction.

$$T_{d_x} = \begin{bmatrix} 1 & 0 & 0 & d_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

II. *Bulk Transformation:* Synthetic bulk motion artefacts are generated by rotating the upper half of the original image with $\pm\theta_z^\circ$ around the z-axis.

$$T_{\theta_x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) & 0 \\ 0 & sin(\theta_x) & cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$T_{\theta_y} = \begin{bmatrix} cos(\theta_y) & 0 & sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(\theta_y) & 0 & cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$T_{\theta_z} = \begin{bmatrix} cos(\theta_z) & -sin(\theta_z) & 0 & 0 \\ sin(\theta_z) & cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

III. *Non-rigid deformation:* Synthetic non-rigid motion artefacts are generated by skewing (shearing) the axial, sagittal, coronal images using:

$$T_{S_{xyz}} = \begin{bmatrix} 1 & S_{xy} & S_{xz} & 0 \\ S_{yx} & 1 & S_{yz} & 0 \\ S_{zx} & S_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where one combined skewing value is used as: $S_{xyz} = tan(\pm\theta^{\circ}_{xyz})$ in the xyz-direction.

Following image deformation, a motion-corrupted 3D stack is constructed by sampling 2D images from both motion-free and motion-corrupted stacks in an interleaved manner similar to foetal MRI acquisition [Glenn and Barkovich, 2006b]. Where for the bulk motion experiment a 3D stack is constructed by only rotating the upper half of the brain. Three stacks are used for the reconstructions where each stack is sampled with a voxel size of 1.25x1.25x2.5 mm^3. Standard axial, sagittal, and coronal orientations are shown in Figure 3.5. An HR image with isotropic voxel size 1.25 mm^3 is reconstructed using SVR [Kainz et al., 2015a], square patch- and superpixel-based PVR.

Evaluation of foetal organ MRI reconstructions: Evaluating the quality of reconstructed foetal MRI is challenging due to the absence of motion-free ground truth data. For this purpose, a novel approach for the evaluation problem is proposed based on the originally

acquired slice images. Assuming that 2D in-plane patches extracted from the original stacks contain no motion artefacts, they are used as gold standard and compare with corresponding simulated patches from the reconstructed volume. Evaluation metrics (see Section 3.4.1) are computed between the reconstructed input stacks and the final motion corrected image and averaged over the whole volume. The foetal brain is typically used to assess the quality of reconstruction as it moves rigidly, fulfilling the rigid motion assumption for SVR-based methods in the 2D-3D registration step. However, soft tissue organs such as the placenta deform non-rigidly. For this reason, the placenta and whole uterus are selected as challenging test cases for PVR and SVR.

Evaluation metrics: The following metrics are employed for measuring the quality of the reconstructed image:

Cross-correlation (CC) to measure the similarity between the intensities of input $I(i,j)$ and reconstructed image $\tilde{I}(i,j)$ at the location (i,j), which is defined as:

$$CC = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{(I(i,j) - I_\mu)(\tilde{I}(i,j) - \tilde{I}_\mu)}{\sigma_I \sigma_{\tilde{I}}}, \tag{3.13}$$

where N and M are the dimensions of a 2D slice.

The peak signal-to-noise ratio (PSNR) is used to measure the error introduced by motion and is based on the mean squared error (MSE) between the original 2D in-plane patch and the reconstructed image, which is defined as:

$$MSE = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[I(i,j) - \tilde{I}(i,j) \right]^2 \tag{3.14}$$

where N and M are the dimensions of a 2D slice. $I(i,j)$ is the value of the pixel (i,j) from the original 2D plane I, while $\tilde{I}$ represents the reconstructed image. PSNR is defined as:

$$PSNR = 10 \log \frac{I_{max}^2}{MSE} \tag{3.15}$$

where I_{max} is the maximum intensity in the original image. An improved reconstruction quality

usually results in higher PSNR. However, PSNR does not reflect well subjective human perception of image quality as it is mainly based on estimating absolute errors between individual pixels.

The structural similarity index (SSIM) accounts for image degradation as perceived changes in structural information [Wang et al., 2004]. It measures the structural similarity by comparing normalised local patterns of pixel intensities, which is similar to the human visual system's abilities to extract information based on structure. The SSIM is defined as:

$$SSIM = \frac{(2\mu_I\mu_{\tilde{I}} + c_1)(2\sigma_{I\tilde{I}} + c_2)}{(\mu_I^2\mu_{\tilde{I}}^2 + c_1)(\sigma_I^2 + \sigma_{\tilde{I}}^2 + c_2)}, \tag{3.16}$$

where μ_I, $\mu_{\tilde{I}}$, σ_I^2 and $\sigma_{\tilde{I}}^2$ are the average and variance of the intensities of the original 2D in-plane slice and the reconstructed slice respectively. $\sigma_{I\tilde{I}}$ is the covariance of I and $\tilde{I}$. c_1 and c_2 are defined as $(k_1L)^2$ and $(k_2L)^2$ in order to balance the division with weak denominator, where L is the dynamic range of the intensities in image I. Similar to [Wang et al., 2004], k_1 and k_2 are equal to 0.01 and 0.03 respectively.

Structural dissimilarity (DSSIM) heat maps are calculated in order to visualize the dissimilarities between original and reconstructed images. DSSIM is calculated as a distance metric derived from SSIM:

$$DSSIM = \frac{(1 - SSIM)}{2} \tag{3.17}$$

3.4.2 Evaluation of Local Deformations

Compared to SVR, PVR relaxes rigidity constraints and may therefore introduce deformation artefacts like shrinking of anatomical structures. Hence, the following two experiments are performed to assess the geometric integrity of the 3D image reconstruction in terms of *mm*:

Evaluation with anatomical landmarks: A set of 3D anatomical landmarks, that can be clearly defined in 3D reconstructions of real foetal MRI head scans, are used to assess the accuracy of nonrigid deformations resulting from PVR compared to SVR. In particular,

10 clinically relevant anatomical landmarks are selected in the eyes and sub-cortical region of the foetal brain. These landmarks define stable anatomical structures of the corpus callosum and the lateral ventricles.A clinical expert has chosen a set of 11 foetal MRI scans with the best reconstruction quality of the foetal head using SVR. The reconstruction of this set from the raw scan data is repeated using PVR with patches of size 32 and stride 16 pixels. Two different observers have annotated the selected anatomical landmarks independently across the 11 subjects. The accuracy is assessed by calculating the average 3D Euclidean distance errors between PVR and SVR (mean annotation points of the two observers), measured for each landmark for all 11 subjects.

Evaluation of displacement: This experiment evaluates the quality of PVR in cases of non-rigid motion between the scanned images using an approximated displacement error measured on the simulated skewing motion experiment of the brain phantom. The reconstructed images are first rigidly and then non-rigidly (b-spline [Rueckert et al., 1999]) registered to the motion-free ground truth image. The transformations resulting from non-rigid registration are used to transform a set of points sampled uniformly on a 3D grid (5% of the image size) inside the brain area. To ensure a fair comparison, reconstructed images from high motion skewing angles $> 1.5°$ are excluded, which may affect registration to the reference image if the reconstructed image is blurred or distorted.

3.5 Results

3.5.1 Global Image Quality

Reconstruction of adult brain MRI: Experiments on adult brain MR data using the Brainweb database [Cocosco et al., 1997] includes introducing synthetic non-rigid motion artefacts as described in Section 3.4.1. Example slices of standard planes of original and corrupted data are depicted in Figure 3.5. Comparative experimental results of SVR and PVR reconstruction methods are shown in Figure 3.6 for PSNR, SSIM and CC. For all metrics, PVR

shows an improved performance over SVR, particularly in the presence of deformations with higher skewing angles. Furthermore, superpixel-based PVR achieves performance similar to PVR using arbitrary square patches, while requiring a lower number of input patches.

3.5.2 Reconstruction of Foetal Organs

Exemplary PVR and SVR reconstructions under motion introduced by kicking of the foetus are shown in Figure 3.7. PVR reconstruction results show an improved visual appearance and less blurring in the region with severe motion artefacts (arrow).

An example of a challenging clinical case with a kidney malformation in one of twin foetuses, is shown in Figure 3.8. Clinical partners confirmed that such complications are easier to examine and to quantify after PVR-based reconstruction.

Comparative experiments of PVR variants were carried out on 32 foetal MR scans at gestational ages of approximately 20 weeks with voxel size around 1.00x1.00x2.50 mm^3 and 1.25x1.25x1.25 mm^3, constituting challenging image corruption samples. The ssFSE T2-weighted sequences data have been acquired on a Philips Achieva 1.5T, the mother lying 20° tilt on the left side to avoid pressure on the inferior vena cava and aligned to the main axes of the foetus. Tables 3.1 and 3.2 show numerical results of evaluating individual stacks before reconstruction (baseline), and the final reconstructed image from 3-6 stacks using square patches ($a = 32$ and $\omega = 16$), superpixels ($a = 16, \gamma = 60\%$) and multi-scale superpixels (with initial $a = 16, \gamma = 60\%$ and $\times 1.5$ scaling factor for every new iteration). Statistical testing between baseline and PVR variants was carried out using paired T-Tests and differences between using fixed or multi-scale and between square patches or superpixels were assessed via Two-factor ANOVA with repeated measures. In Tables 3.1 and 3.2 the names of PVR variants are marked in bold if statistically significant differences have been found during analysis, *i.e.*, **FS** and **MS** and/or **Square Patches** and **Superpixel** pairs are bold if the results between them differ significantly.

The evaluation of the reconstruction quality of a whole 3D image into a single-valued metric

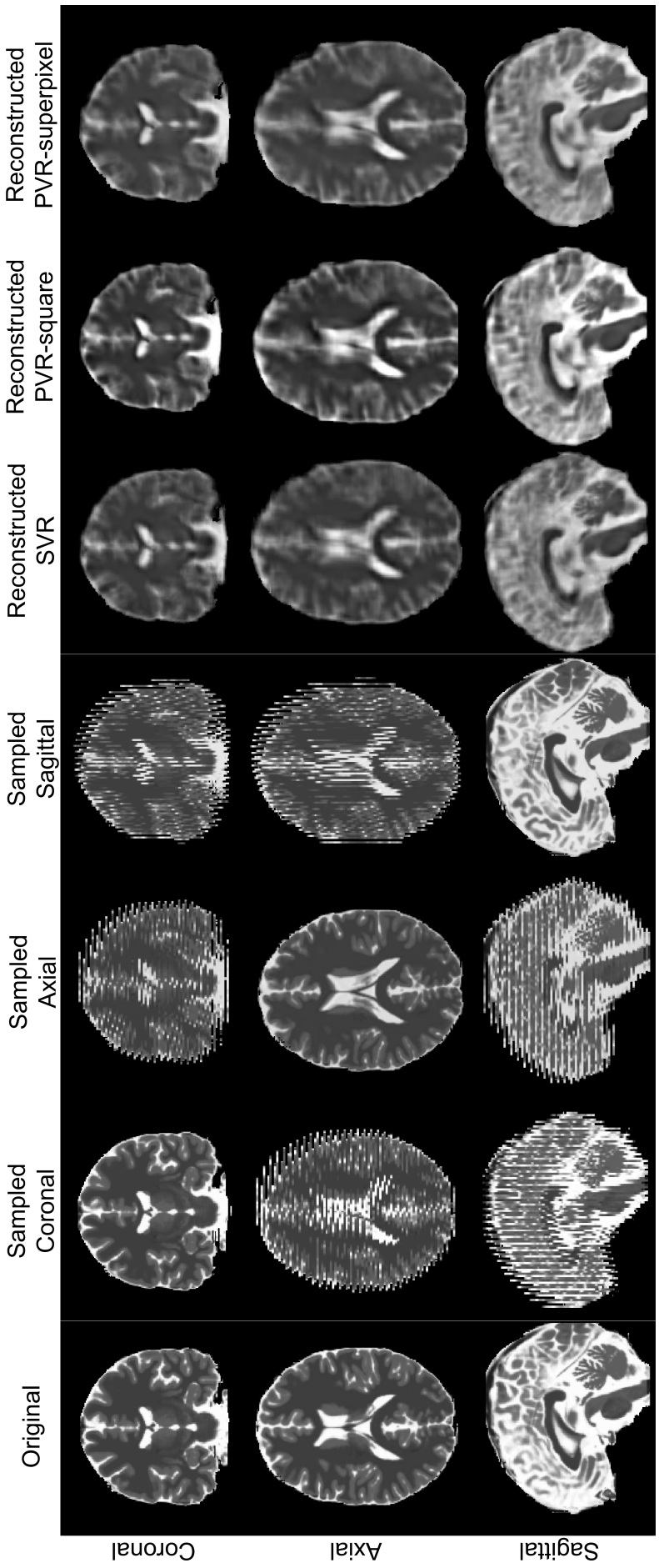


Figure 3.5: Strong synthetic non-rigid motion artefacts caused by skewing an adult brain phantom with an angle of ($\theta^{\circ}_{xyz} = \pm 1$). Rows: MRI in standard orientations: coronal, axial, and sagittal. Columns: original scan (1x1x1 mm^3), and sampled (1.25x1.25x2.5 mm^3) axially, coronally and sagittally, and SVR, PVR-square patches and PVR-superpixel reconstructed isotropic images (1.25x1.25x1.25 mm^3). It appears that PVR can reconstruct visually better images. However, it may allow reconstructions with slight deformations due to relaxing the rigidity assumption using small patches.

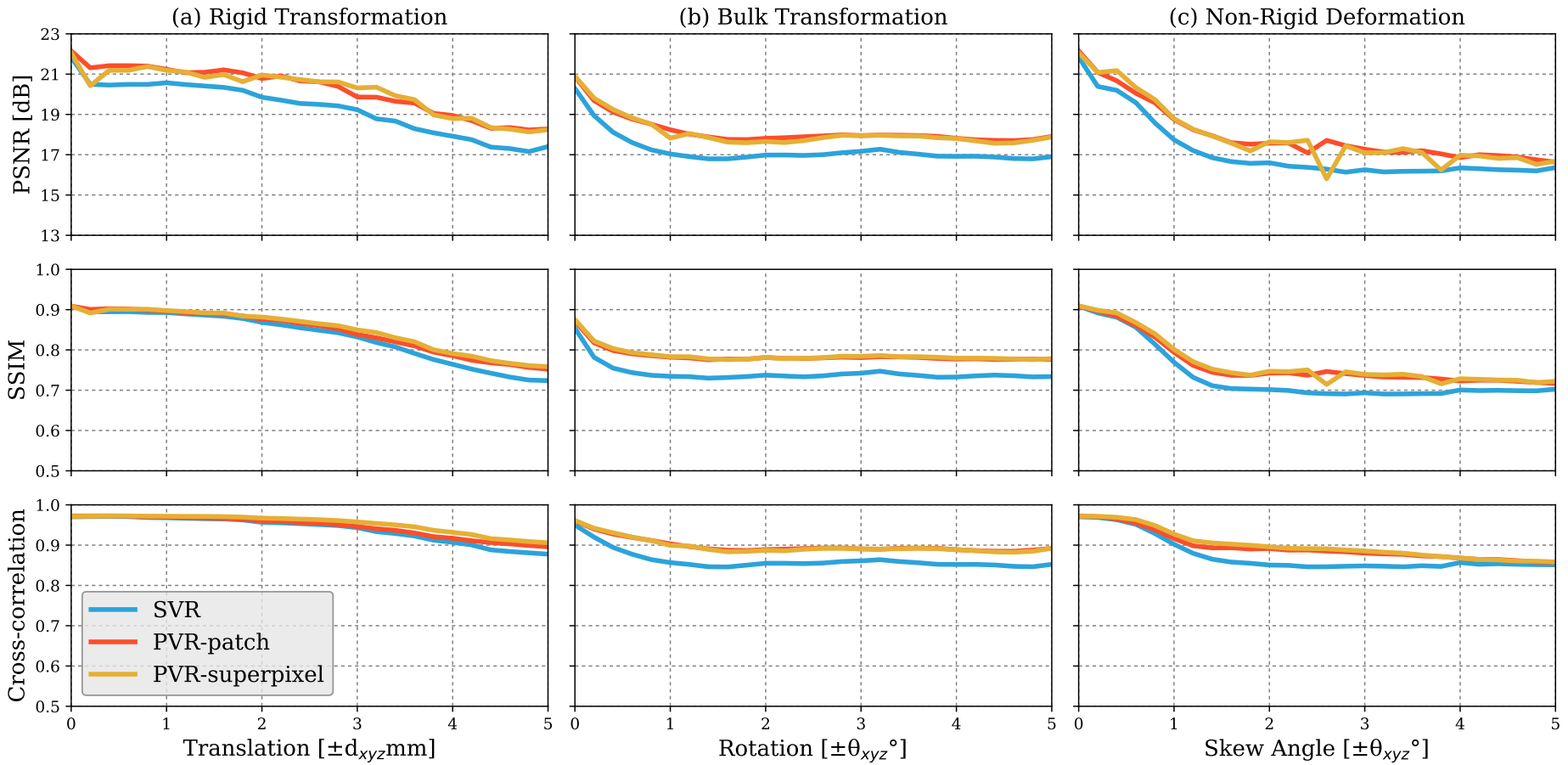


Figure 3.6: Comparative reconstruction performance of SVR and PVR methods on synthetically corrupted Brainweb [Cocosco et al., 1997] data using rigid translational transformations (1st column), bulk transformations (2nd column) and non-rigid skewing deformations (3rd column). Left to right: PSNR, SSIM and CC over skew angle in degrees and translations in mm for SVR (blue), superpixel-based PVR ($a = 16, \gamma = 60\%$, yellow) and PVR using square patches ($a = 32, \omega = 16$, red).

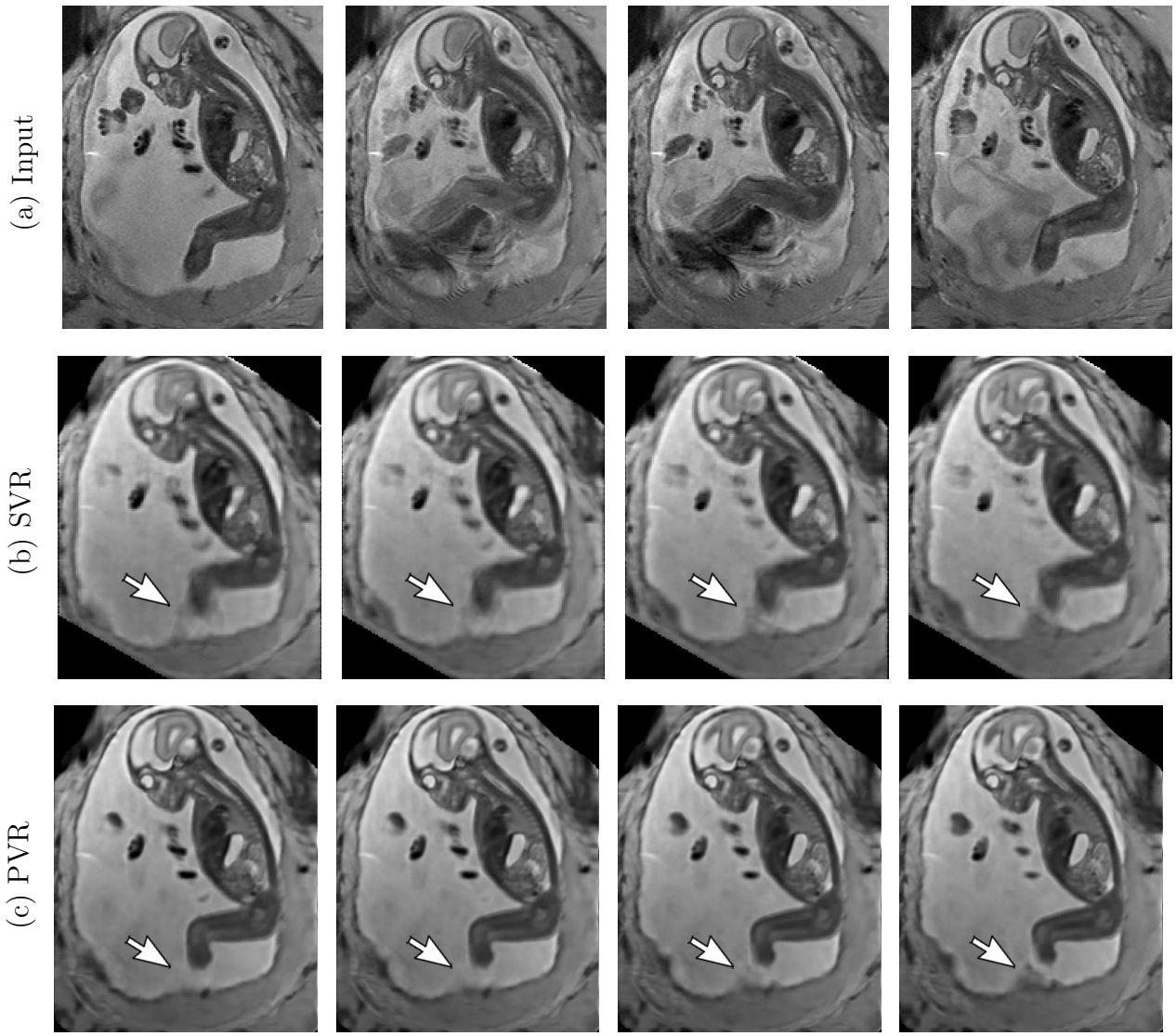


Figure 3.7: Example reconstructions of consecutive MR scans of a moving foetus (kicking): input data (a) and corresponding cutting planes through an SVR- (b) and PVR-reconstructed (c) volumes. SVR produces blurry but readable results because of high data redundancy and outlier rejection through robust statistics. PVR with square patches of $a = 32$ and $\omega = 16$ appears visually superior. The arrow points at an area of substantial quality differences caused by independent rapid movements of the leg.

	Brain		*Placenta*		*Uterus*	
Baseline	16.97 ± 3.77		19.95 ± 4.40		19.29 ± 3.82	
	Square Patches	Superpixels	**Square Patches**	**Superpixels**	Square Patches	Superpixels
FS	26.70 ± 1.45	26.60 ± 1.58	31.07 ± 1.35	31.00 ± 1.46	27.17 ± 1.43	27.35 ± 1.36
MS	27.03 ± 1.38	26.35 ± 1.54	30.85 ± 1.50	29.62 ± 1.30	26.87 ± 1.31	26.40 ± 1.24

Table 3.1: Average PSNR results ($N = 32$) for the input stack (baseline) and PVR variants with fixed (FS) and multi-scale (MS) variants of square patches and superpixels. All mean differences of PVR against baseline are statistically significant ($p < 0.05$). Names of all statistically significantly different PVR variants are stated in **bold**.

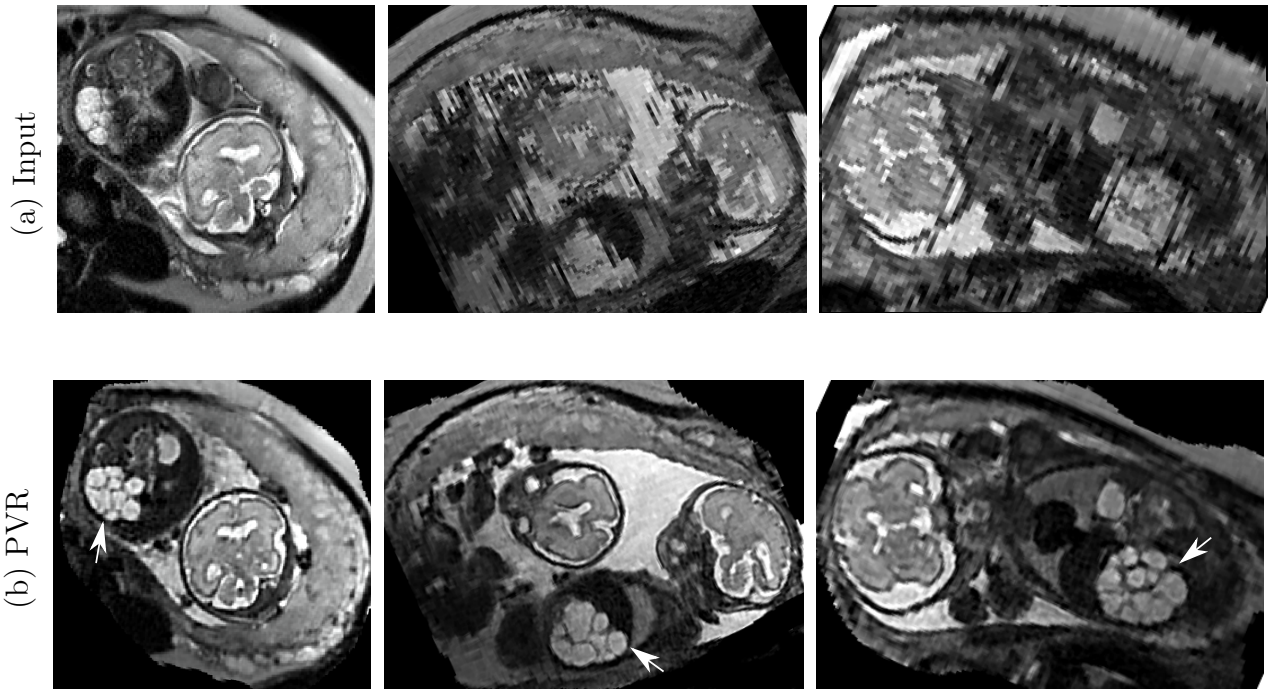


Figure 3.8: Three viewing planes through the original motion corrupted scan of a moving twin with a gestational age of 28 weeks (a) and PVR reconstruction using multi-scale superpixels (b). For this dataset, the uterus is masked to save unnecessary computation time in areas containing maternal tissue. The white arrow points at a unilateral multicystic kidney of one of the twins.

	Brain		*Placenta*		*Uterus*	
Baseline	0.00 ± 0.01		0.00 ± 0.02		0.01 ± 0.03	
	Square Patches	**Superpixels**	**Square Patches**	**Superpixels**	**Square Patches**	**Superpixels**
FS	0.51 ± 0.03	0.51 ± 0.03	0.58 ± 0.03	0.58 ± 0.03	0.53 ± 0.03	0.50 ± 0.03
MS	0.48 ± 0.04	0.45 ± 0.05	0.54 ± 0.04	0.49 ± 0.05	0.50 ± 0.03	0.47 ± 0.04

Table 3.2: SSIM results ($N = 32$) for the input stack (baseline) and PVR variants with fixed (FS) and multi-scale (MS) variants of square patches and superpixels. All mean differences of PVR against baseline are statistically significant ($p < 0.05$). Names of all statistically significantly different PVR variants are stated in **bold**.

may not properly reflect the performance differences, as it is based on averaging values of all the pixels of all the input stacks. Furthermore, Tables 3.1 and 3.2 indicates significant differences between variants of PVR but these differences have only minimal qualitative effect on reconstruction accuracy. Therefore, Figure 3.9 evaluates the reconstruction quality of PVR additionally using dissimilarity heat maps based on the measured DSSIM (see Section 3.4.1). This approach allows further qualitative evaluation and allows for uncertainty visualization of PVR reconstructions.

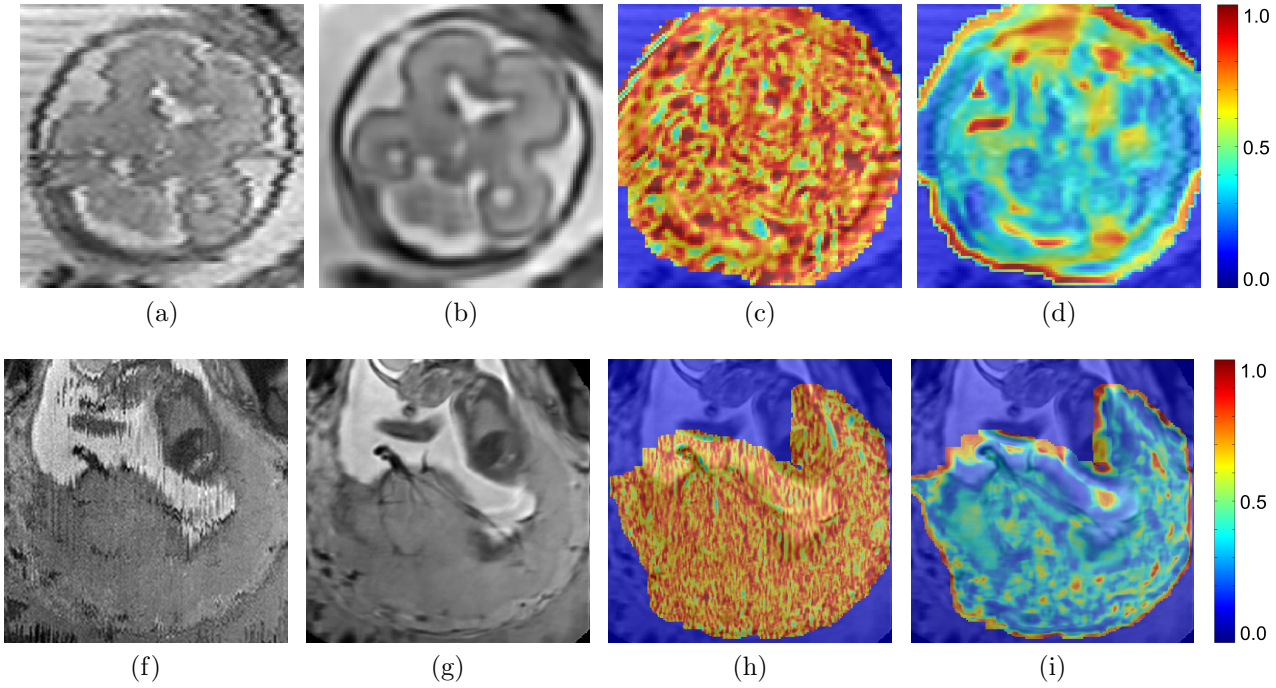
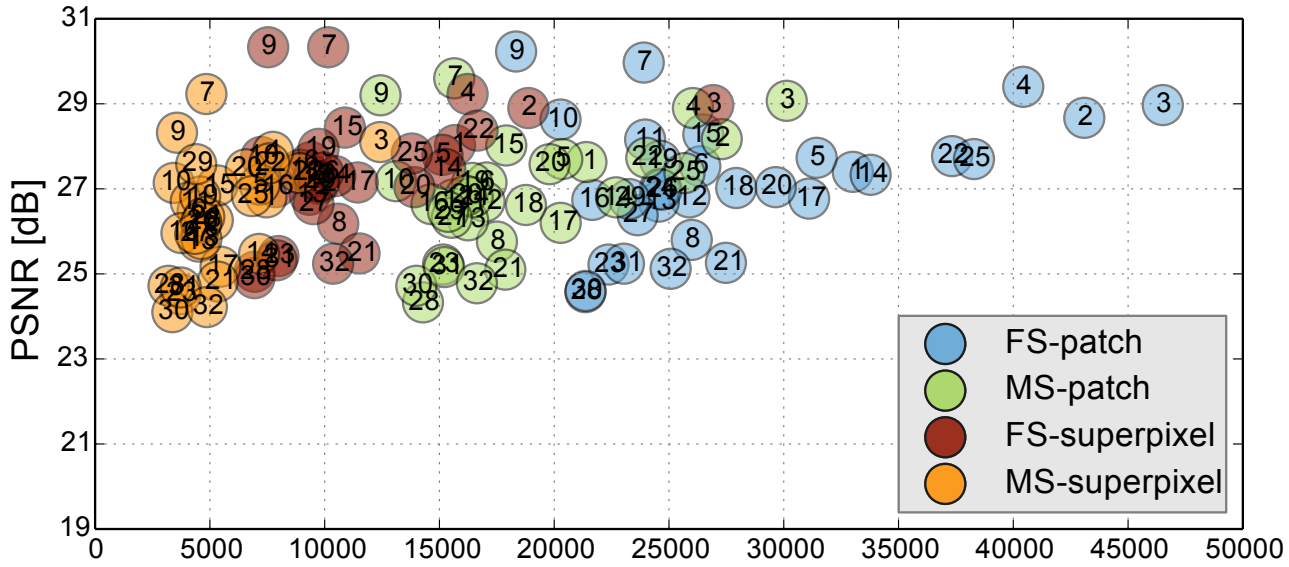


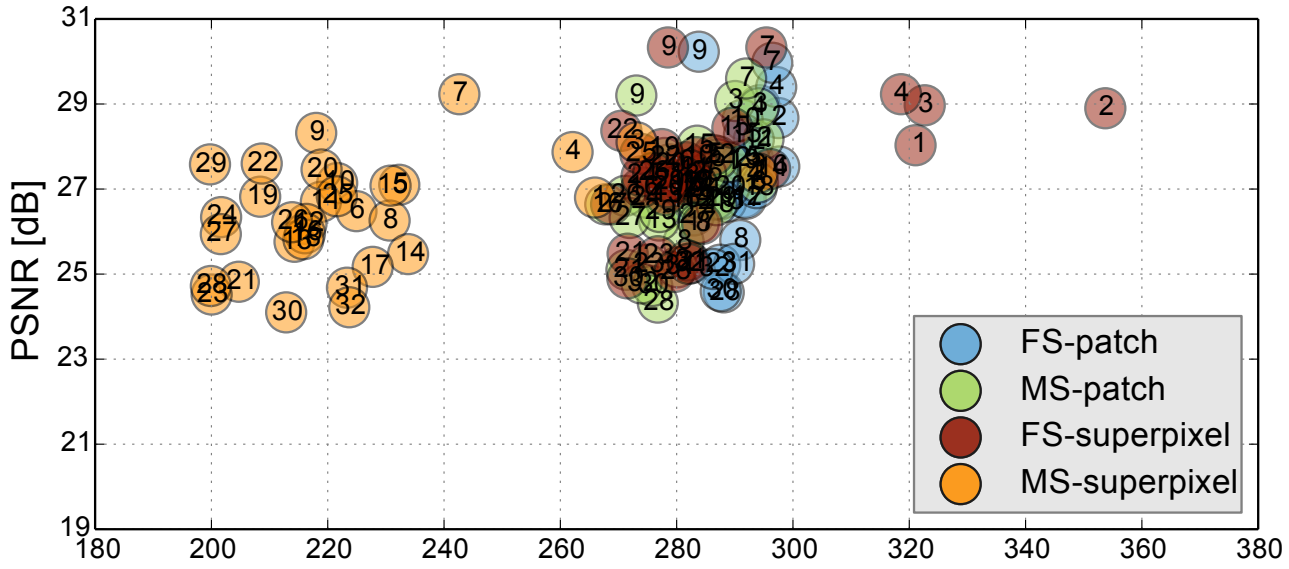
Figure 3.9: A sample 2D cutting plane through a motion-corrupted foetal brain (a) and placenta (f), after PVR using square patches with $a = 32$ and $\omega = 16$ (b) and (g). The DSSIM heat map for a baseline before reconstruction (c) and (h), and after PVR (d) and (i). The average DSSIM of the foetal brain equals 0.497 (c) and 0.248 (d), while for the placenta equals to 0.491 (h) and 0.214 in (i).

3.5.3 Foetal Brain Reconstruction at Late Gestation

Examination of the brain is one of the primary uses for foetal MRI. While PVR's primary aim is to reconstruct large fields of view and non-rigid motion, the performance of PVR is evaluated on the same conditions under which SVR is usually applied (accurate delineation of a brain mask followed by rigid full slice to volume registration). This experiment evaluates if PVR can



(a) Number of patches



(b) Overhead pixels (%)

Figure 3.10: (a) Number of generated patches and (b) necessary additional overhead pixels (%) of the different PVR variants versus their reconstruction PSNR quality of the whole uterus (see Tables 3.1 and 3.2. Optimal results are found in the upper left corner of the plots, *i.e.*, high reconstruction quality and low computational overhead. The subject number is highlighted inside each circle marker. Multi-scale superpixels (MS-superpixel) achieve similar reconstruction quality to fixed-size (FS-patch), multi-scale (MS-patch) square patches while clustering in the area of minimal computational overhead.

efficiently substitute SVR in pipelines where only the reconstruction of a single, masked rigidly moving organ is required, *i.e.*, for compensation of pure bulk motion.

Here, a second dataset comprising of ten mature foetuses (GA $>$ 33 weeks) with little motion) is collected. It contains expert brain segmentation and validated state-of-the-art SVR reconstructions [Kainz et al., 2015a]. PVR is applied only in the delineated brain area and compared to clinical SVR reconstructions in terms of PSNR, SSIM, CC and computational overhead. In contrast to [Kainz et al., 2015b], where only a small area of the brain was evaluated, the full brain mask is evaluated in this experiment. Patch parameters are chosen similar to [Kainz et al., 2015b] ($a = 32$, $\omega = 16$), which performed best for square patches according to the evaluation in [Kainz et al., 2015b] using the multi-scale superpixel approach. Figure **??** shows that PVR provides similar reconstruction and motion correction quality for the brain as would be the case if a tight expert mask [Keraudren et al., 2014] for a region of interest would have been used for SVR. However, the computational overhead of processing patches is always larger than if SVR is used. In this experiment SVR is considered to be the baseline processing 100% of the available pixels. The required runtime of any PVR variant will always be longer than if SVR is used for bulk motion compensation of a defined region of interest.

	FS-16	**FS-24**	**FS-32**	**FS-40**	**FS-48**	**MS-16**
PSNR	17.1	17.30	17.60	**17.9**	17.2	17.3
SSIM	0.441	**0.473**	0.464	0.435	0.391	0.420
CC	0.682	0.691	0.715	**0.721**	0.659	0.678

Table 3.3: Comparison between fixed-scale (FS) and multi-scale (MS) superpixels reconstruction using 3 input images phantom brain deformed with skewing angle $\theta_{xyz} = \pm 1^{\circ}$.

3.5.4 Multi-scale Reconstruction

This section evaluates different scale parameters for the multi-scale superpixel reconstruction to further assess the influence of the number of scales on the reconstruction quality. Similar to the phantom brain experiment, the brain image is reconstructed from 3-stacks deformed

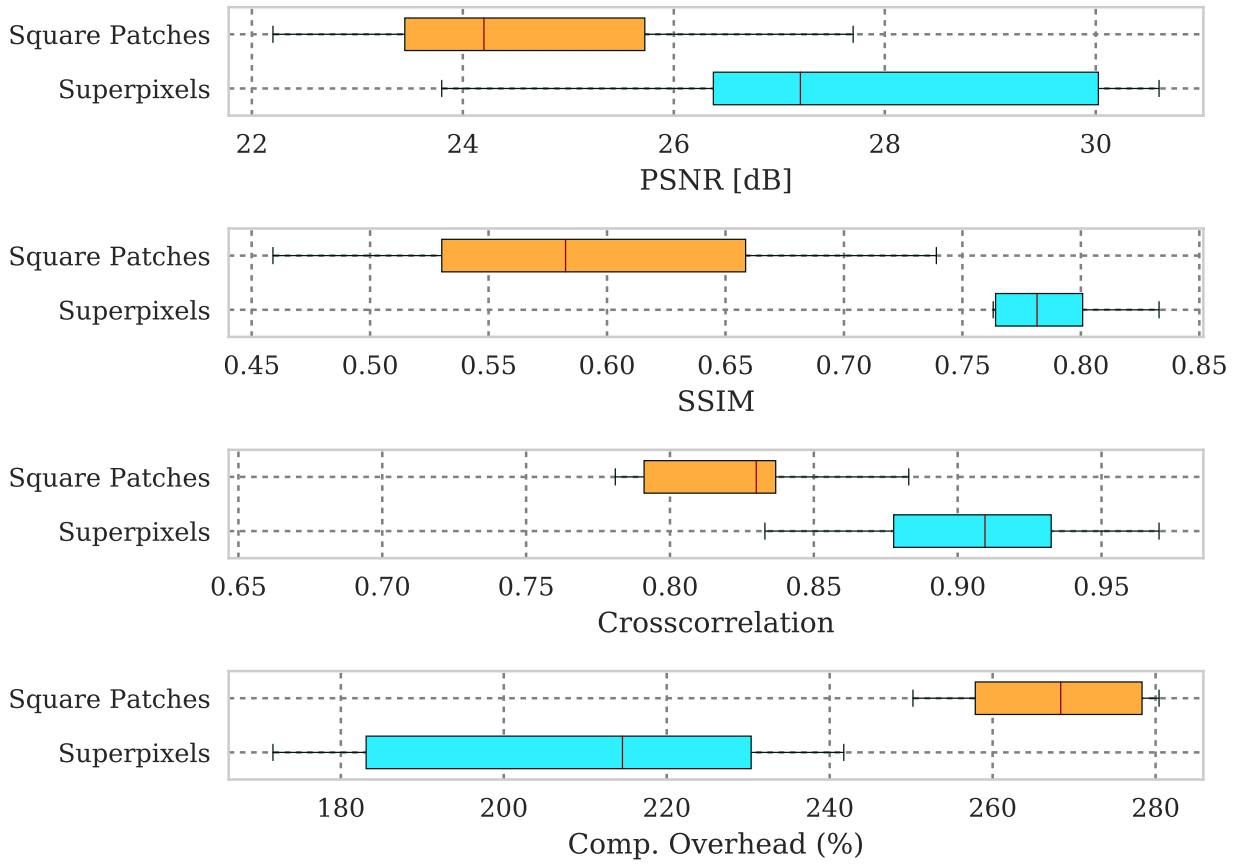


Figure 3.11: Comparison of the best performing PVR parameters from [Kainz et al., 2015b] for square patches and multi-scale superpixels with SVR for motion compensation of the brain in mature (GA $>$ 33 weeks) foetuses with little motion. PVR's image quality is similar to the small-area SVR method currently used in the clinical practice. The necessary patch overlap for PVR requires more computation, which results generally in longer runtime for any PVR variant.

with skewing angle $\theta_{xyz} = \pm 1^{\circ}$. The ground truth image is used for evaluating quantitatively each reconstructed image with five different scales as shown in Table 3.3. Multi-scale (MS) reconstruction is configured to increase the scale of the patch by a factor of $\times 1.5$, four times starting from $a = 16$ to 40 (2-iteration per each scale). Fixed-scale (FS) reconstruction is configured to fix the size of the patch for every reconstruction iteration. Figure 3.10 and Table 3.3 show that MS-PVR achieves similar performance as FS-PVR, while larger scale patches result in a lower number of input patches and overhead pixels.

3.5.5 Local Deformations

Landmarks: The results for each landmark are summarized in Table 3.4. The inter-observer error is 0.92 ± 0.41 *mm*. The average distance between all landmarks in PVR reconstructions and the mean landmark location in SVR reconstructions between the two observers is 1.13 ± 0.87 *mm*. In order to provide an indication of the scale of rigid motion being addressed in the foetal brain landmark images, the estimated displacement distance of all 3D landmarks during imaging is measured by calculating the distance between each annotated landmark and its inverse-transformed point from the correspondent slice of an input stack, assuming rigid slice motion for SVR. The maximum of the calculated displacements equals to 79.85 *mm*. The total average equals to 13.58 ± 15.23 *mm* for all landmarks, see Table 3.4. The p-values from a paired T-test show that the difference between the error introduced between observers and PVR can be considered to be not statistically significant ($p = 0.0166$)). Overall, this evaluation demonstrates that PVR does not introduce notable structural distortion compared to rigid SVR.

Displacement: The mean and maximum displacement of 9 reconstructed images are calculated for SVR and PVR with different parameters. The results in Table 3.5 shows that PVR with patch size of 64 achieves the smallest mean error, while this error increases with the decrease of the patch size.

Landmarks	Slice-to-volume motion [mm]	Inter-observer error [mm]	PVR-SVR distance [mm]	p-value
Left center of the eye	15.10 ± 15.90	0.95 ± 0.29	1.00 ± 0.43	0.7983
Right center of the eye	11.87 ± 11.01	0.88 ± 0.40	0.96 ± 0.41	0.6381
Genu of corpus callosum outer aspect	13.86 ± 15.16	0.69 ± 0.36	0.83 ± 0.24	0.7670
Splenium of corpus callosum outer aspect	13.02 ± 15.06	1.08 ± 0.56	1.11 ± 0.75	0.8901
Left Frontal horn of lateral ventricle	14.93 ± 16.83	1.03 ± 0.40	1.35 ± 1.00	0.3296
Right Frontal horn of lateral ventricle	17.04 ± 18.83	0.90 ± 0.22	1.25 ± 0.89	0.2511
Left occipital horn of lateral ventricle	13.40 ± 14.59	1.04 ± 0.34	0.98 ± 0.48	0.7483
Right occipital horn of lateral ventricle	11.41 ± 13.20	1.05 ± 0.38	1.05 ± 0.56	0.9906
Left temporal horn of lateral ventricle	12.89 ± 15.88	0.69 ± 0.36	1.28 ± 1.43	0.2278
Right temporal horn of lateral ventricle	11.83 ± 12.76	0.80 ± 0.46	1.47 ± 1.33	0.1396
All landmarks	13.58 ± 15.23	0.92 ± 0.41	1.13 ± 0.87	0.0166

Table 3.4: Motion in terms of the distance the 3D landmarks have been estimated (SVR) to have been displaced during imaging, and the displacement of these landmarks between PVR and clinically validated SVR reconstructions.

Displacement	SVR	PVR-Superpixel 16	PVR-Patch 128	PVR-Patch 64	PVR-Patch 32	PVR-Patch 16
Mean	1.83 ± 1.55	1.58 ± 1.05	1.70 ± 1.10	1.58 ± 1.03	1.67 ± 1.07	2.40 ± 1.71
Max	4.91 ± 4.26	4.75 ± 3.66	4.86 ± 3.54	4.87 ± 3.92	5.02 ± 3.92	5.72 ± 4.09

Table 3.5: Mean and maximum displacement error in mm resulting from non-rigid registration between different reconstructions and the motion-free adult brain phantom ground truth.

3.5.6 Performance Analysis

This section evaluates the computational performance of each PVR variant. Measuring the overall runtime is not meaningful because this would be highly machine specific and would include data transfer overhead and non optimized functions. The runtime varied between 2000–4000 seconds on the testing machines, depending on the system configuration. Instead the computational overhead introduced by PVR is analysed in comparison to SVR. The overhead can be measured by counting the number of processed patches and the number of additionally processed voxels. These values are compared to the achieved reconstruction quality in Figure 3.10. Multi-scale superpixels show significantly better performance than other PVR variants and introduce only minimal necessary overhead while gaining similar image quality to more naïve PVR variants. Multi-scale superpixels are potentially five times faster than other variants.

3.6 Discussion

PVR surpasses reconstruction performance of the state-of-the-art SVR method for cases with considerable non-rigid deformations and inconsistent bulk motion. Different variants of PVR have been evaluated using fixed-size and multi-size square patches, and fixed-size and multi-size superpixels. ANOVA analysis has shown significant differences between these approaches for different areas of the uterus. However, evaluation of motion compensation methods is difficult especially due to the lack of ground truth in foetal MRI. Mapping the reconstruction quality of a whole 3D volume into a single-valued metric may not properly reflect qualitative differences, if based on averaging all measured values of all input stacks. Therefore, additional extensive qualitative analyses are performed to present examples and evaluation based on structural dissimilarity (DSSIM) heat maps.

While an SVR reconstruction is constrained by using whole orthogonal slices assuming rigidity, PVR may introduce geometric deformations due to the relaxation of the rigidity assumption to smaller patches instead of whole slices. For this reason, local deformations were evaluated by selecting anatomical landmarks from the foetal head and compare PVR with SVR, which shows that the difference is not statistically significant for individual landmarks. Also dense, non-rigid registration are employed to further validate the geometric accuracy of the reconstruction. Reconstructed images from SVR and PVR with different variants are non-rigidly aligned to a motion-free brain phantom. The mean and maximum displacements of the deformations are then calculated to measure the accuracy. Both experiments evaluate local deformations, although the latter may include errors introduced by the imperfect non-rigid registration process. Maximum displacement errors are mainly observed in feature-poor regions like the amniotic fluid, where automatic non-rigid registration does not perform well. Despite showing strong evidence that PVR works well for the whole uterus, it should be noted that the experiments to assess the reconstruction quality on brain anatomy do not necessarily extend to the achieved reconstruction accuracy of other areas (e.g. moving limbs). Focused validation of individual regions will be required for other, domain specific applications.

In addition to reconstruction and motion correction of the whole uterus, this chapter have

demonstrated that the proposed method works for multiple births cases with multiple foetal instances present in the volumetric image. These cases are more likely to have complications and to undergo MRI during pregnancy but would require extensive manual effort to be successfully reconstructed with state-of-the-art methods.

Although the proposed method is able to reconstruct the whole uterus automatically, small parts like limbs that move rapidly between the acquisition of individual slices are more difficult to recover. This is especially problematic for very young foetuses that have more space to move inside the womb. In cases of rapid limb movements (>2 cm between individual slices) PVR is not able to find structural consensus between overlapping patches and blurry image regions will be reconstructed. This is a general problem of automatic intensity-based optimization methods and methods that are able to understand the semantic content of each patch will likely be required for future improvements.

PVR introduces a considerable computational overhead to the reconstruction stage of foetal MR image processing pipelines. The amount of necessary additional redundant information is evaluated to give a general idea about the expected runtime of different PVR variants. Patches based on multi-scale superpixels are significantly more efficient than a naïve implementation of overlapping square patches, while maintaining a similar reconstruction accuracy. Quantitatively, square patches perform slightly better for the brain, which is most likely due to the rigid nature of the enclosing skull. Superpixel-based patches achieve better results for regions that are likely affected by non-rigid movements like the placenta and the whole uterus.

3.7 Conclusion

This chapter have introduced and evaluated the concept of patch-to-volume reconstruction (PVR) in order to compensate non-rigid motion artefacts from foetal MRI scans without requiring a defined region of interest. PVR splits the 3D input image into overlapping square patches and superpixels and employs automatic EM-based outlier rejection to find consistent data.

PVR is able to automatically reconstruct whole collections of motion corrupted stacks without the need for image segmentation and manual identification of rigid regions. Extensive experiments have shown that PVR can reconstruct the whole uterus, selected foetal organs, and secondary, non-rigidly moving pregnancy structures such as the placenta.

PVR's reconstruction quality and computational performance has been evaluated quantitatively and qualitatively on an adult phantom T2-weighted brain with synthetic non-rigid motion artefacts, as well as on the whole uterus from motion corrupted foetal MRI data including foetal brain, placenta and cases with multiple births.

Chapter 4

Foetal Brain Localisation and Placenta Segmentation

This chapter is based on:

Amir Alansary, Matthew Lee, Kevin Keraudren, Bernhard Kainz, Christina Malamateniou, Mary Rutherford, Joseph V Hajnal, Ben Glocker, and Daniel Rueckert. **Automatic brain localisation in fetal MRI using superpixel graphs**. *In Medical Learning Meets Medical Imaging (MLMMI)*, pages 13–22. Springer, 2015. [Alansary et al., 2015]

Amir Alansary, Konstantinos Kamnitsas, Alice Davidson, Rostislav Khlebnikov, Martin Rajchl, Christina Malamateniou, Mary Rutherford, Joseph V Hajnal, Ben Glocker, Daniel Rueckert, and Bernhard Kainz. **Fast fully automatic segmentation of the human placenta from motion corrupted MRI**. *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 589–597. Springer, 2016. [Alansary et al., 2016]

Abstract

This chapter addresses this issue by automatic localisation and segmentation of the foetal anatomy, in particular, the brain and placenta which are structures of interest for many foetal MRI studies. Foetal MRI is emerging as an effective, non-invasive tool in prenatal diagnosis and pregnancy follow-up. However, there is a significant variability of the position and orientation of the foetus in the MR images. This makes these images more difficult to analyse and interpret compared to standard adult MR imaging, which standardized anatomical imaging aligned planes.

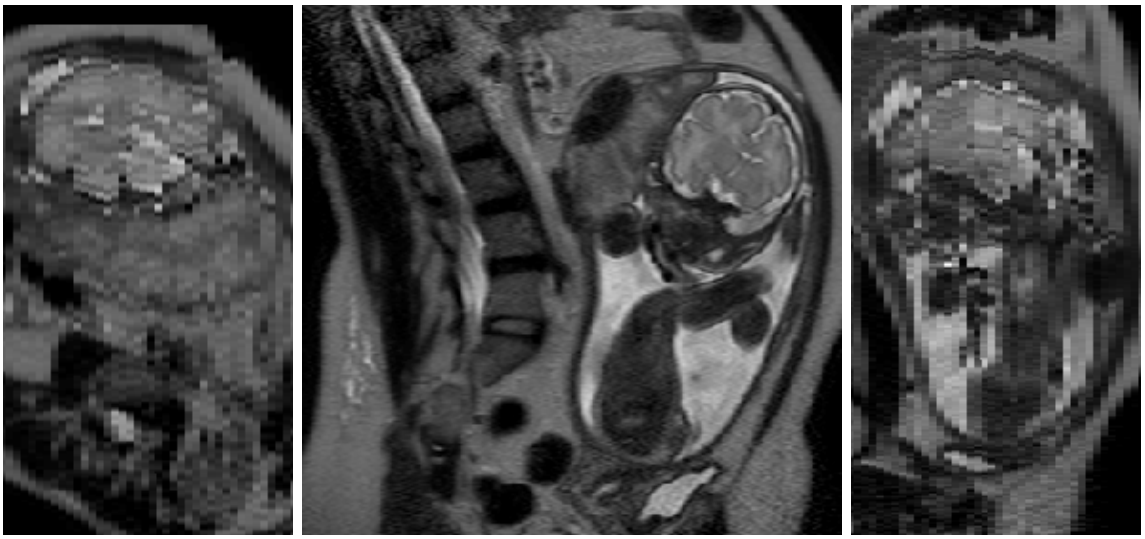
The first part of this chapter proposes an automatic method for foetal brain localisation. The method is based on extracting superpixels from the input image, then a histogram of features for each superpixel is computed using bag of words (BoW) of dense scale invariant feature transform (DSIFT) descriptors. This is followed by constructing a graph of superpixels for training a random forest classifier to distinguish between brain and non-brain superpixels. The localisation framework has been tested on 55 *MRI scans at gestational ages between* 20–38 *weeks. The proposed method is evaluated using* 5*-fold cross validation achieving a* 94.55% *brain detection accuracy rate.*

The second part proposes a fully automatic segmentation framework of the placenta from structural T2-weighted scans of the whole uterus, as well as an extension in order to provide an intuitive pre-natal view into this vital organ. The proposed method adopts a 3D multi-scale convolutional neural network to automatically identify placental candidate pixels. The resulting classification is subsequently refined by a 3D dense conditional random field, so that a high resolution placental volume can be reconstructed from multiple overlapping stacks of slices. The segmentation framework has been tested on 66 *subjects at gestational ages* 20–38 *weeks achieving a Dice score of* $71.95 \pm 19.79\%$ *for healthy foetuses with a fixed scan sequence and* $66.89 \pm 15.35\%$ *for a cohort mixed with cases of intrauterine foetal growth restriction (IUGR) using varying scan parameters.*

4.1 Foetal Brain Localisation

4.1.1 Introduction

Foetal MRI is an emerging field of research providing an important non-invasive tool for diagnosis and assessing foetal development. In particular, foetal brain localisation is important for assessing the brain development and maturation. Defining the brain from scan images provides useful measurements (*e.g.* reduced head circumference), which can be used for diagnosis foetal growth restriction (FGR). In addition, it is a primary step for the majority of the automatic motion correction techniques, as discussed in the previous chapter. This chapter proposes an automatic method for foetal brain localisation from motion-corrupted MRI scans. The design of such methods is challenging because of the high variability in the orientation, shape, and size of the brain with the gestational age. Moreover, the unconstrained foetal movements affects the image quality, especially in the out-of-plane direction (see Figure 4.1).



(a) axial (b) Coronal (c) Sagittal

Figure 4.1: Three orthogonal cutting planes through a stack of foetal MRI images. The quality of the in-plane (coronal) slices is not affected by motion, however, there are inter-slice artefacts appear in the out-of-plane views (axial and sagittal).

Previously, Anquez et al. [2009] presented an automatic method based on 3D template matching to detect the eyes, enabling a subsequent 2D/3D graph-cut segmentation that extracts the brain. This approach is based on 3D templates and lacks the flexibility necessary to deal

with motion artefacts as well as foetal abnormalities. The methods proposed in [Ison et al., 2012, Keraudren et al., 2013] address the variability of foetal MRI through machine learning. Ison et al. [2012] proposed a random forest classifier first distinguishes between maternal and foetal tissues before classifying different tissues of the foetal head, while [Keraudren et al., 2013] combines prior knowledge of the foetal size with maximally stable extremal regions (MSER) detection and a bag-of-words model. Table 4.1 summarises the limitation of the previous methods. The proposed framework in this chapter tackles these limitation by constructing superpixel graphs for training a random forest classifier. Where the outputs are the predictions for each pixel in the input image to brain or non-brain tissues. The following section describes the localisation framework in detail.

Template-based matching [Anquez et al., 2009, Taimouri et al., 2015, Taleb et al., 2013]	**Feature-based learning** [Ison et al., 2012, Kainz et al., 2014a, Keraudren et al., 2013]
Sensitive to noise, motion artefacts and foetal abnormalities that may change the appearance of the brain.	2D features are less sensitive to motion artefacts between slices. However, 3D context information are essential for slices that contain less pixels of the object of interest.
Sensitive to gestational age, as the appearance and size of the brain significantly change during the foetal growth. A different template per age is required.	Prior information (*e.g.* brain size and gestational age) could be required for more accurate results.
Difficult to generalize to a bigger datasets with large variations.	It can adapt to new testing datasets by training on new samples from the new data.

Table 4.1: Comparison between the methods used in the literature for foetal MRI brain localisation based on two many categories: (*i*) template-based matching, and (*ii*) feature-based learning.

4.1.2 Method

The proposed approach for the automatic localisation of the foetal brain consists of four main steps as shown in Figure 4.2. Input data are composed of 3D foetal MRI datasets. The first step is to divide each 2D slice into $\{N_1, ..., N_s\}$ superpixels, in order to minimise the local redundancy in the input data. By clustering and constructing a single descriptor for each superpixel, the impact of noise on each descriptor is reduced whilst preserving homogeneous

regions that most likely belong to the same anatomical region. 2D superpixels are extracted in the direction of the in-plane slices, whereas out-of-plane slices contain motion artefacts. The second step is to calculate image descriptors for each pixel, then aggregate them into one histogram h_i for each i-superpixel. The third step is to build superpixel graphs based on each superpixel's neighbours. Then, each superpixel's histogram is normalized with its neighbours in the graphical model. During the fourth step, a random forest generates a probability map of the brain for every superpixel. Finally, the output probability map is refined using another auto-context classifier followed by selecting the largest 3D component.

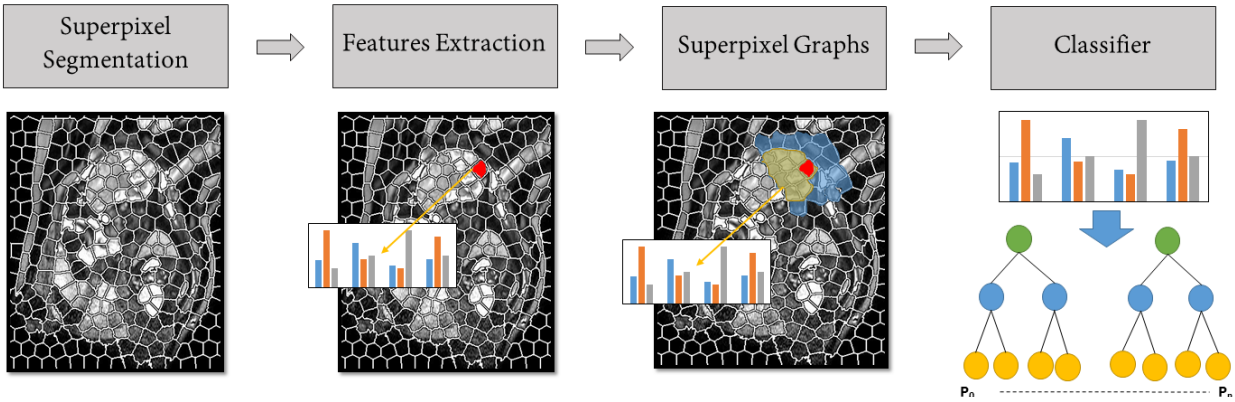


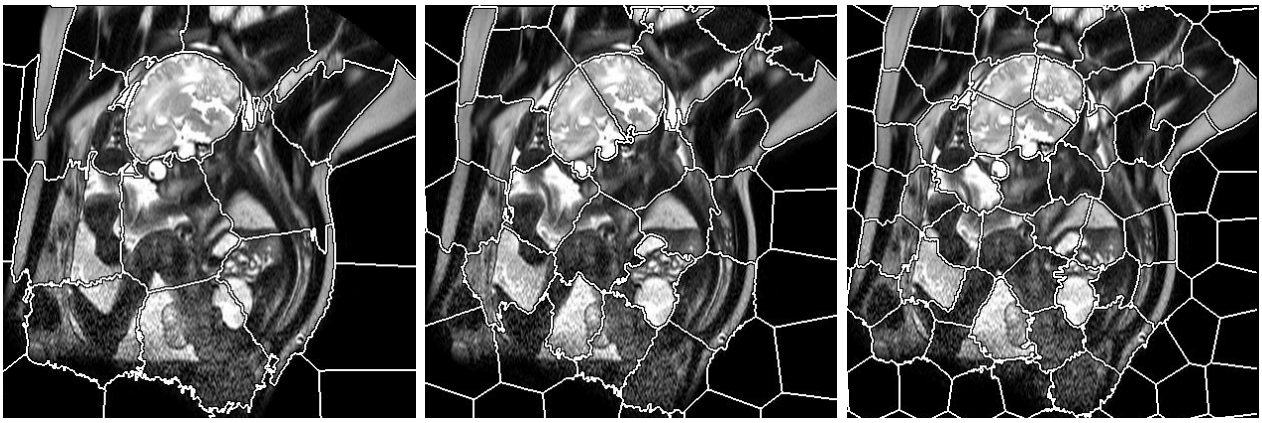
Figure 4.2: The automatic localisation of the foetal brain framework.

Superpixels

Extracting superpixels is a popular unsupervised image segmentation technique that clusters image pixels into groups of pixels based on the correlation between each pixel and its neighbours. In the literature [Achanta et al., 2012], many superpixel techniques have produced good segmentation results, where their performance may vary according to the speed and accuracy. Medical image analysis can be computationally expensive when compared to other image analysis domains because of the shape and size of the scanned images (*e.g.* 3D volumes as opposed to 2D images). Here, the proposed framework adopts the simple linear iterative clustering (SLIC) technique [Achanta et al., 2012], which is fast to compute while achieving a good segmentation quality (as shown in Figure 4.3) with lower computational cost, so that the method scales well when processing the many slices of a volume. SLIC segments pixels into compact and nearly uniform superpixels. These segmentations are applied to the acquired 2D in-plane slices, where

the foetal movements cause misalignment between these 2D slices. The number of superpixels for each 2D slice is a hyper-parameter that needs to be adjusted carefully. Here, the ad-hoc heuristics proposed by [Hair et al., 2006] is adopted to optimize superpixels for the foetal brain. Where, this rule is scaled with a constant factor $k \in \mathbb{R}_{>0}$. k is function of the total number of pixels in a 2D input image n. Total number of superpixels $s = |\{N_1, ..., N_s\}|$ in a 2D slice can be defined as:

$$s \approx k \cdot \sqrt{(n/2)}. \tag{4.1}$$



(a) $s \approx 25$ (b) $s \approx 50$ (c) $s \approx 100$

Figure 4.3: An illustrative example of segmenting a 2D in-plane foetal MRI slice into superpixels using 4.1. Where $s \approx 25$ (a), $s \approx 50$ (b), and $s \approx 100$ (c) superpixels.

Image Descriptors

A bag-of-features model is constructed by extracting dense scale invariant feature transform (DSIFT) [Vedaldi and Fulkerson, 2010] descriptors. This is done by first computing SIFT [Lowe, 1999] descriptors for each pixel in every 2D image in the training set at a fixed scale and orientation. A visual vocabulary is then learned using k-means clustering of the DSIFT descriptors, where their centres are used to quantise a dictionary of k words. The descriptors are extracted from pixels by finding their closest matching word from the learned dictionary, aggregating the frequency of words in each superpixel N_i into one histogram h_i with k-bins, as shown in Figure **??**. The k-dimensional histogram acts as the final feature descriptor for each superpixel. However, the descriptor is constructed in such a way that only contains local information about

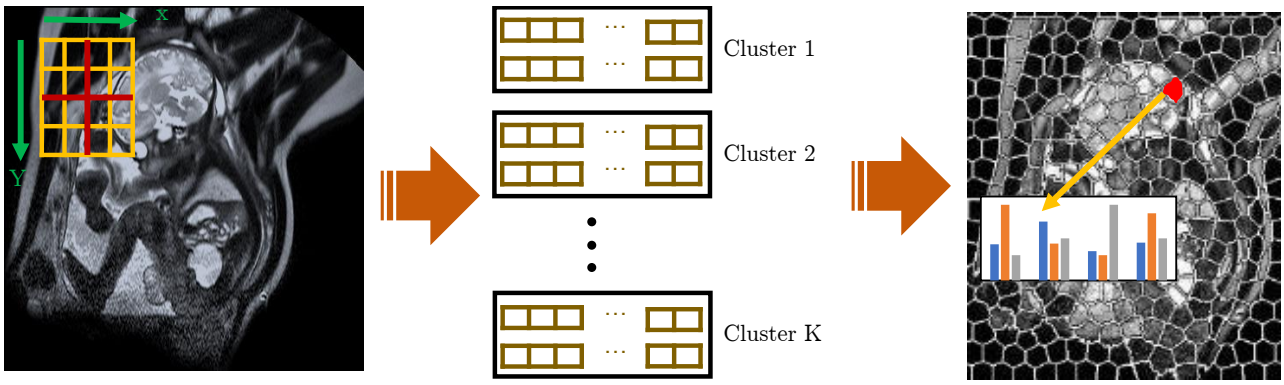


Figure 4.4: An illustrative diagram of the constructed superpixel-based image descriptors. DSIFT features are extracted for each pixel, then clustered using k-means buidling a dictionary of k word (BoW). Finally, a histogram h_i with k-bins is calculated using the frequency of words in each superpixel N_i.

the superpixel itself. This leads to a loss of the large-scale image context. Also, due to the nature of the superpixels, their histograms of features tend to be sparse. Most of the DSIFT descriptors within a superpixel are likely to be mapped to the same word.

Superpixel Graphs

To overcome the problem of sparse descriptors for superpixels, graphs of superpixels are constructed using the distance between the centroids of superpixels as edges [Fulkerson et al., 2009]. These edges are weighted based on both spatial and geodesic distances. Firstly, the centroid c_i of each N_i superpixel is computed. Then, the r neighbours are identified based on the similarity score between two superpixels N_i and N_o with centroids c_i and c_o. The similarity between two superpixels is defined as:

$$f_r(N_i, N_o) = 1 - \frac{d(c_i, c_o)}{D}, \tag{4.2}$$

where $d(\cdot, \cdot)$ is the Euclidean distance, and D is the length of the diagonal of the 2D image. This normalises the score to be between $[0, 1]$. The closer $f_r(c_i, c_o)$ is to 1, the smaller is the distance between these two centroids. This is followed by assigning weights w_j for the extracted r neighbours candidates based on the geodesic distances between their centroids. The geodesic

distance is estimated using:

$$d_r(N_i, N_o) = \sum_{j=1}^{m} |I(p_j) - I(p_{j-1})|\,, \tag{4.3}$$

where $I(p_j)$ is the intensity of the pixel p_j, and m equals to the number of pixels located on the straight line between c_i and c_o. Next, the weights w_j are calculated by normalising d_r, so that $w_j = 1$ when d_r is the lowest and $w_j = 0$ when d_r is the highest. Finally, the histograms of features for the extracted r superpixels are aggregated based on the calculated weights using:

$$\tilde{h}_i = \frac{\sum_{j=0}^{r} w_j h_j}{\left\| \sum_{j=0}^{r} w_j h_j \right\|_{\ell 1}} \tag{4.4}$$

where $h_0 = h_i$ or the histogram of the current superpixel in consideration. Learning-based classifiers usually do not perform well when the input features have very different scales. Hence, *feature scaling* is performed using l_1-norm. Using graphs of superpixels enables the proposed localisation method to overcome the motion artefacts between the 2D slices by extracting the superpixel neighbours in 3D, which would be more difficult using graphs of individual pixels. This is because of the nature of superpixels that most likely represent rigid regions, see Figure 4.5. By selecting both spatial and intensity neighbours, the final selected features for each extracted histogram are increased compared to the initial sparse features. Consequently, this extend the extracted information to include the image context instead of using only local information. Figure 4.5 shows an illustrative figure of the proposed superpixel-based graph in both 2D and 3D.

Classification

A two class random forest (RF) [Breiman, 2001] is trained using the normalised feature vectors to classify each superpixel as brain or non-brain. Each input image is annotated pixel-wise, and each superpixel N_i is assigned the class label l_i that corresponds to the label with the highest frequency inside it. After the initial classification, the output prediction are applied to

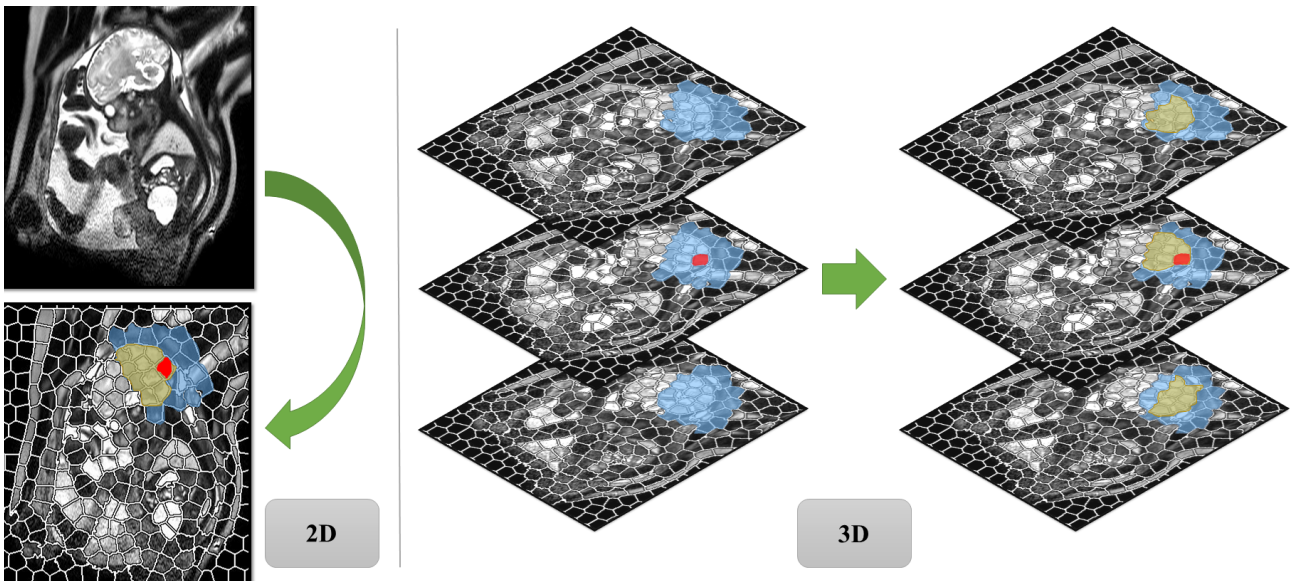


Figure 4.5: The proposed superpixel-based graph in 2D and 3D. The red superpixel represents the current superpixel in consideration, the r-spatial neighbours are coloured in blue, and the neighbours with higher weights (w_j) are coloured in green. w_j is calculated based on the geodesic distances between the centroids of each superpixel.

train a second random forest along with 10% of the most important features used for training the first classifier. The important features are computed using the contribution of each feature on the decrease of the weighted impurity of a tree, averaged and ranked for the whole forest. This results in adding an auto-context classifier [Tu and Bai, 2010] that can boost the classifier performance and the classification accuracy. The output volume is then filtered by finding the largest 3D component, which is the brain mask in this case, followed by a convex-hull extraction [Barber et al., 1996] to obtain a clean homogeneous segmentation.

4.1.3 Experiments and Results

Data

The proposed framework has been tested on 55 foetal scans at wide gestational age between 20–38 weeks. Thirty subjects of these datasets are from healthy foetuses and 25 datasets are from foetuses with intrauterine foetal growth restriction (IUGR). The data was acquired with a 1.5T Philips MRI system using single shot fast spin echo (ssFSE) sequences with voxel size $0.8398 \times 0.8398 \times 4\text{mm}^3$. Ground truth labels were obtained by manual segmentation of the

brain performed by expert observers.

Implementation

Pre-processing: The consistency of the input numerical attributes to any learning algorithm is essential, and these values must be at the same scale. There are two common ways for rescaling: (i) normalization using $(x - min)/max$ scaling, and (ii) standardisation using $(x - \mu)/\vartheta$ scaling. The former results in values that are shifted and rescaled between $[0 - 1]$, while the latter results in a distribution with a zero mean value. Standardisation does not bound the output values to specific range and less affected by outliers. Choosing the best rescaling method is application-dependant, and it is usually chosen experimentally. Here, the input scan intensities are standardised as a preprocessing step for the localisation pipeline.

Training: SLIC superpixel extraction [Achanta et al., 2012] is adopted for generating superpixels that are optimized for foetal MRI data. The vlfeat library [Vedaldi and Fulkerson, 2010] is used for generating DSIFT features and the scikit-learn library [Pedregosa et al., 2011] is used for the random forest classifier. The code is implemented using python and MatLab with the mex-c environment. To determine number of superpixels s per image, k is selected equal to 5 in Equation 4.1. The number of neighbours selected for superpixel graphs r were set to 25 calculated for each $xy-$plane in three slices. Number of sample pixels representing the brain tissue (positive samples) are less than the number of non-brain or background pixels (negative samples), *skewed datasets*. Hence, the positive and negative training samples are balanced for the classifier by sampling training superpixels generated from a cropped volume around the brain (+25% of the maximum brain diameter in the xy-plane).

Post-processing The prior knowledge that brains appear brighter in T2-scans allowed to suppress some of the background pixels by thresholding any pixel less than 10% of the maximum intensity value of the whole subject. All these parameters are chosen heuristically on a validation dataset.

Evaluation Metrics

In order to facilitate the comparison with previous published work [Ison et al., 2012, Keraudren et al., 2013], the performance of the proposed method is evaluated using common metrics from the literature, such as:

- **Brain detection accuracy** is defined as the extracted mask covering at least 70% of the brain, similar to the definition presented in [Ison et al., 2012] but calculated for the whole 3D brain for simplification.
- **Centre of the brain detection accuracy** is measured by calculating the percentage of centres that lie inside the ground truth of the manually labelled brains.
- **Brain coverage** is measured by the percentage of the manually-labelled brain that are covered by the segmented boundary box.
- **Dice coefficient** [Dice, 1945] is also applied to measure the segmentation accuracy of the proposed approach. It measures the agreement between the segmented and ground truth brain labels using:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{4.5}$$

where TP, FP and FN are the true and false positive, and the false negative pixels, respectively. This is equivalent to size of the overlap between segmented and ground truth labels divided by the total size of both. TP, FP and FN are visualised in Figure **??**.

Results

A 5-fold cross validation was used for evaluating foetal localisation approach (11 test patients 44 training patients per fold). The random forest classifier achieved an average accuracy score of 96.17% per a superpixel basis. Table 4.2 shows the accuracy of the proposed approach at different dictionary sizes $k = 50$, $k = 100$, $k = 400$ and $k = 800$. These results shows that increasing the dictionary size or the histogram bins (sparsity) increases the dice accuracy.

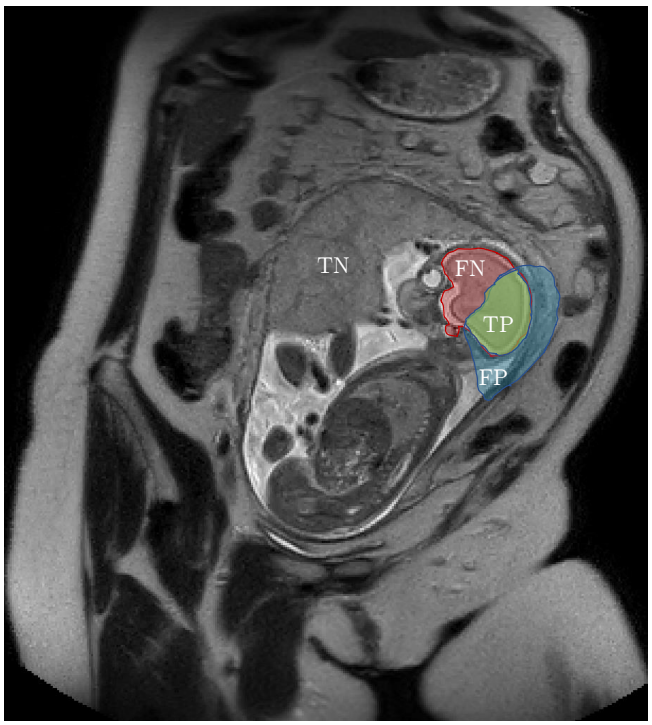


Figure 4.6: An illustrative figure for the dice score on a sample 2D MR image of the uterus. Where the ground truth label (TP+FN) is in *red*; the segmented brain (TP+FP) is in *blue*; the intersection is in *green*; and the background (TN+FP) is not coloured. The Dice score is computed using Equation 4.5.

However, it also increases significantly the processing time of training and testing. These experiments were done using parallel processing on a CPU with 32-cores and 128GB RAM.

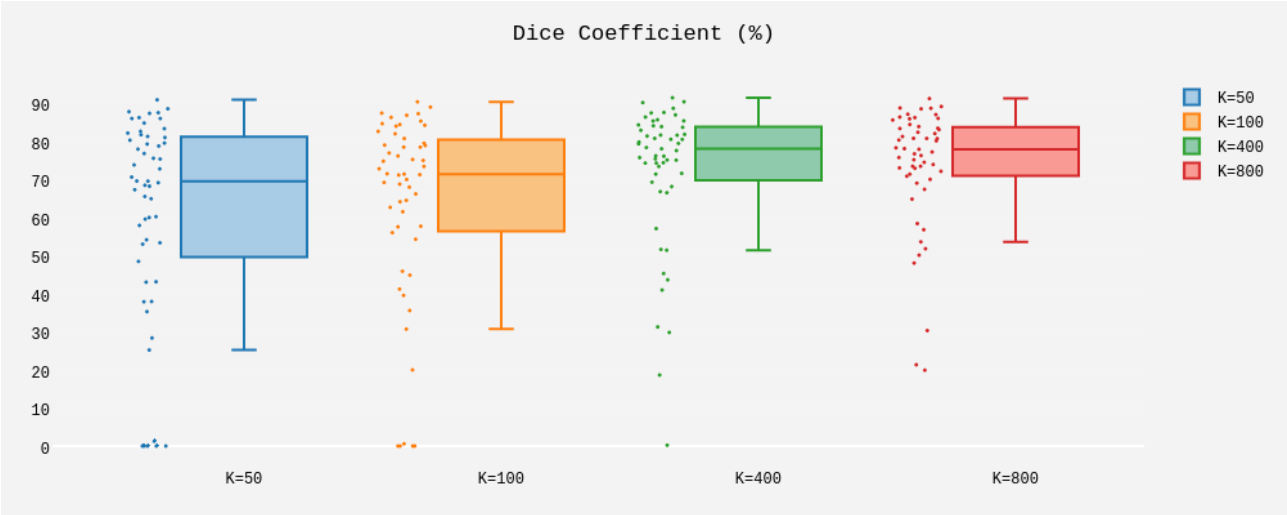


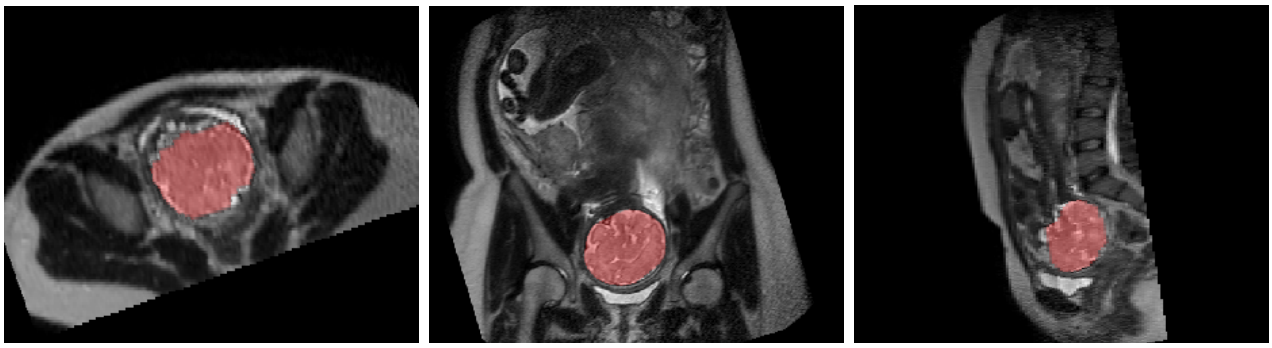
Figure 4.7: A plot diagram showing the accuracy of the proposed localisation approach at different dictionary sizes $k = 50$, $k = 100$, $k = 400$ and $k = 800$.

The brain localisation approach have achieved a 94.55% detection accuracy. It also could detect the centre of the brain with prediction accuracy 98.18% of the test subjects while in [Ison

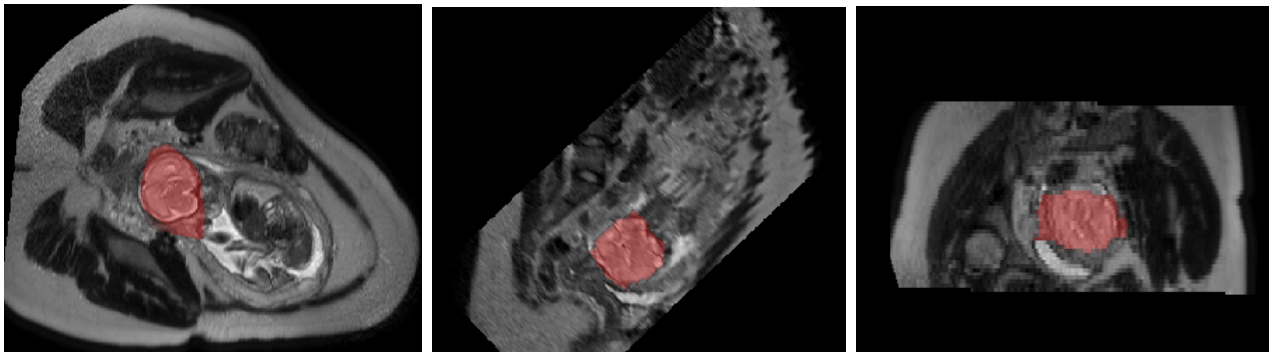
Evaluation metric	**[Alansary et al., 2015]**				**[Keraudren et al., 2013]**			**[Ison et al., 2012]**		
	k = 50	**k = 100**	**k = 400**	**k = 800**	**2D SIFT**	**3D SIFT**	**Bundled SIFT**	**Axial**	**Coronal**	**Sagittal**
Brain detection (% subjects)	87.27	89.09	94.55	90.91	98	85	100	91	98	28
Center in brain (% subjects)	80.00	81.82	98.18	98.18	-	-	-	67	81	53
Brain coverage ($\mu \pm \sigma\%$)	85.59 ± 30.51	85.9 ± 28.2	90.03 ± 16.63	90.54 ± 15.17	38	23	85	53	28	53
Dice coefficient ($\mu \pm \sigma\%$)	61.07 ± 26.94	63.22 ± 25.47	71.96 ± 19	73.62 ± 15.9	-	-	-	-	-	-
Average train time (minutes)	31.55	50.25	256.09	424.93	-	-	-	-	-	-
Average test time (minutes)	4.63	4.85	6.31	8.36	-	-	-	-	-	-

Table 4.2: A comparison between the proposed foetal brain localisation approach [Alansary et al., 2015] compared to [Ison et al., 2012, Keraudren et al., 2013]. The accuracy are measure based on the evaluation metrics explained in Section 4.1.3. Columns [2-5] show the results of using different BoW dictionary sizes: $k = 50$, $k = 100$, $k = 400$ and $k = 800$. Columns [6-8] show the results from [Keraudren et al., 2013] using: 2D, 3D and bundled SIFT features. Columns [9-11] show the results from [Ison et al., 2012] based on different orientation: axial, coronal and sagittal, while the other methods are evaluated on all orientation. The accuracy numbers are the ones reported in the published papers.

et al., 2012] they achieved only 81%, 78% and 60% using coronal, axial and sagittal training data. In addition, the proposed approach does not depend on the orientation of the acquired data and it does not use any previous landmarks as in [Ison et al., 2012]. Keraudren et al. [2013] achieved 100% detection accuracy of the brain; their method, however, requires previous information about the gestational age of the foetus to find the expected size of the brain. This information is later used to remove the outliers of the detected brain mask. The proposed method has an advantage to be generic and does not require any prior information. Figure 4.8 shows the segmented brain at different cross sections for three test subjects with different dice accuracy.



(a) S1-axial (b) S1-coronal (c) S1-sagittal



(d) S2-axial (e) S2-coronal (f) S2-sagittal



(g) S3-axial (h) S3-coronal (i) S3-sagittal

Figure 4.8: The segmented brain at different cross sections (axial, coronal, and sagittal) for three different test subjects S1, S2, and S3. The dice corresponding coefficients of S1=91.22%, S2=70.01%, and S3=56.79%.

4.2 Placenta Segmentation

4.2.1 Introduction

Recently, magnetic resonance imaging has revealed to be important for the evaluation of placenta's health during pregnancy [Routledge et al., 2015]. Quantitative assessment of the placenta requires a segmentation, which proves to be challenging because of motion artefacts from both foetal and maternal movements during image acquisition. Figure **??** shows an example of the placenta in foetal MRI from two patients, while the second patient is a twin.

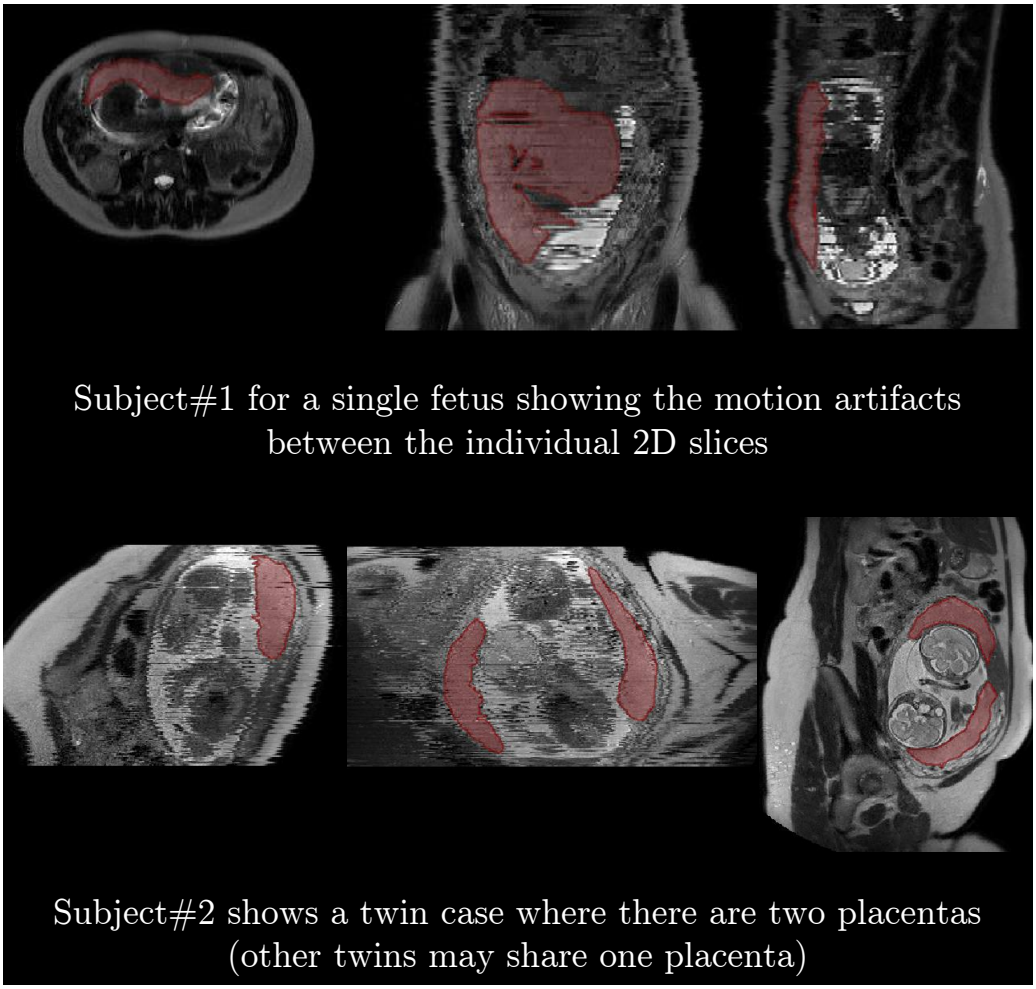
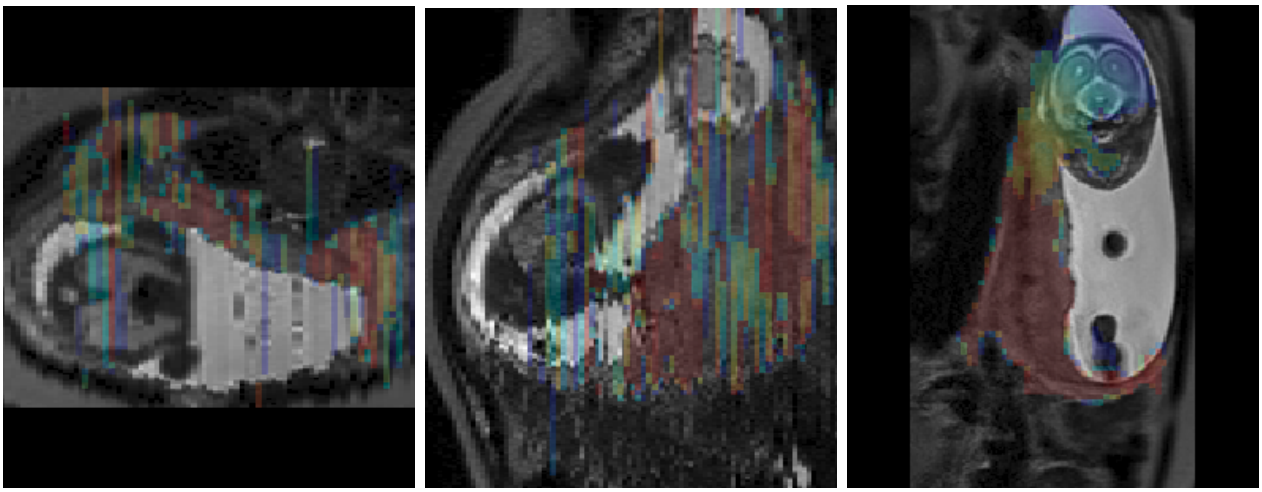


Figure 4.9: An example of the placenta in foetal MRI from two patients (while the second subject is a twin), showing the high variability of its position, orientation, shape and appearance.

In the previous Section 4.1, engineering visual features with a random forest classifier achieves good results for foetal brain localisation. However, the high variability of the placental position, orientation, shape and appearance requires more complex features for accurate

and robust segmentation results, see Figure 4.10.



(a) (b) (c)

Figure 4.10: Placenta segmentation using the method proposed in Section 4.1 superpixel graphs of DSIFT features and random forest. (a), (b) and (c) are three orthogonal cutting planes, where (c) represents the in-plane slice. (a) and (b) show over-segmentation of the placenta especially around the maternal side of the placenta and around the foetal brain. While (b) shows under-segmentation of the placenta especially around the middle part.

This section proposes the first work that addresses the fully automatic segmentation of the placenta from MRI. Previously, placenta segmentation was performed using semi-automatic [Stevenson et al., 2015] or interactive [Wang et al., 2015a, 2016, 2018] approaches, which require manual users interactions. Additionally, recording the structural appearance (*e.g.* placental cotyledons and shape) is essential for clinical qualitative analysis. Where the placenta is usually examined after birth, on a flat surface providing a standard representation for obstetricians. Thus, motion correction methods from Chapter 3 is applied on the segmented placenta to reconstruct a high-resolution motion-free image. Moreover, a 3D visualisation is considered useful in particular for cases that require preoperative planning or surgical navigation. For example, in twin-to-twin transfusion syndrome (TTTS), twins share a common monochorionic placenta that contains abnormal blood vessels. This may result in blood flow from one twin to the other, and hence a failure to grow normally or to survive in severe cases. The treatment of TTTS can be done using intrauterine laser ablation of placental vessels to coagulate vessels communicating between the twins. At the end of this section, an extension of the segmentation pipeline is provided to visualise the segmented and reconstructed placenta *in vivo* using flexible cutting planes.

4.2.2 Method

The proposed approach combines a 3D multi-scale CNN architecture for segmentation with a 3D dense CRF for segmentation refinement. This approach can be extended to compensate for motion and to provide a clinically useful visualisation. Figure **??** shows an overview of the proposed framework.

Multi-scale CNN for Placenta Segmentation

A 3D deep multi-scale CNN architecture [Kamnitsas et al., 2017] is deployed to segment the placenta from the whole uterus. This multi-scale architecture has the advantage of capturing larger 3D contextual information, which is essential for detecting highly variable organs. It is composed of 11-layers deep and two pathways. Both pathways are complementary as the main pathway extracts local features, whereas the second one extracts larger contextual features. Multi-scale features are integrated efficiently by down-sampling the input image and processing the two pathways in parallel. In order to deal with the variations of the placenta's appearance, data augmentation is used for training by flipping the image around the main 3D axes (maternal orientation).

Despite the fact that the multi-scale architecture can interpret contextual information, inference is subject to misclassification and errors. Hence, a CRF is applied to penalize inconsistencies of the segmentation by regularizing classification priors with the relational consistency of their neighbours. The CRF model is 3D fully connected by applying a linear combination of Gaussian kernels to define the pairwise edge potentials [Krähenbühl and Koltun, 2012], which is defined as:

$$E(\mathbf{x}) = \sum_{i \in N} U(x_i) + \sum_{i<j} V(x_i, x_j), \tag{4.6}$$

where i and j are pixel indexes. The unary potential U is given by the probabilistic predictions

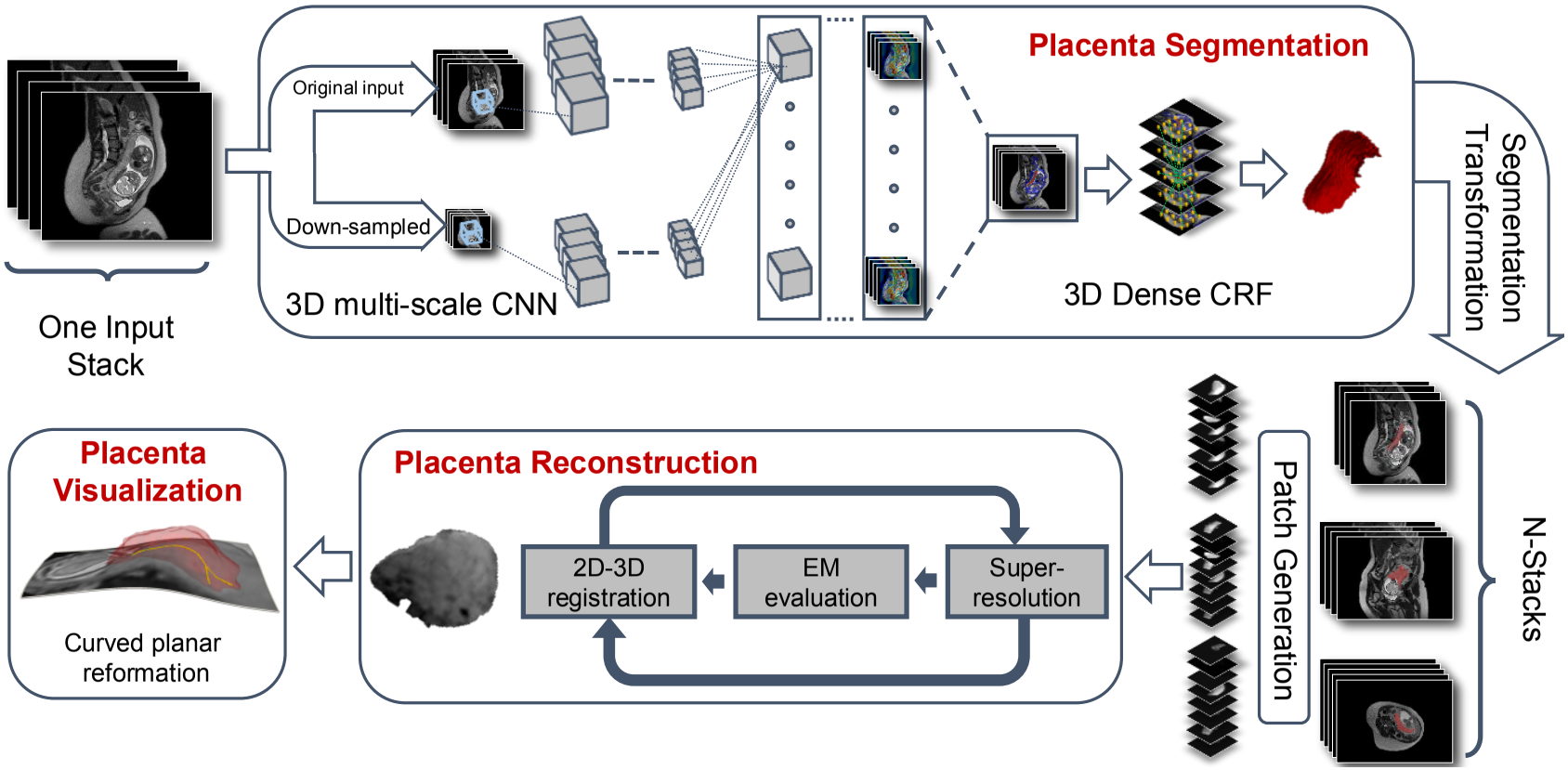


Figure 4.11: The proposed framework for automatic placenta segmentation with extensions for motion correction and visualisation. The pipeline consists of three main stages: (*i*) a multi-scale CNN for segmentation, (*ii*) PVR for motion correction, and (*iii*) CPR for visualisation.

of the CNN classification. Whereas the pairwise potential V is defined by

$$V(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} \left(\omega_1 e^{\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_\alpha^2} - \frac{\left|I_i - I_j\right|^2}{2\theta_\beta^2} \right)} + \omega_2 e^{\left(\frac{\left|p_i - p_j\right|^2}{2\theta_\gamma^2} \right)} \right),$$

where I and p are intensity and position values. $\mu(x_i, x_j)$ is a simple label compatibility function given by the Potts model $[x_i \neq x_j]$. Here, ω_1 controls the importance of the appearance of nearby pixels to have similar labels. ω_2 controls the size of the smoothness kernel for removing isolated regions. θ_α, θ_β and θ_γ are used to adjust the degree of similarity and proximity. The configuration parameters have been chosen heuristically on a validation dataset. Although this tissue classification approach is capable of segmenting the placenta robustly, the segmentation is still subject to inter-slice motion artefacts. Placental masks can be then applied for motion correction methods presented in Chapter 3.

Placenta Segmentation Recovery

To tackle these motion artefacts caused by foetal and maternal movements, the segmentation framework is combined with flexible motion compensation algorithm based on PVR [Alansary et al., 2017]. As shown in Chapter 3, this technique requires multiple orthogonal stacks of 2D slices to provide a better reconstruction quality. It is based on splitting the input data into overlapping square patches or superpixels [Achanta et al., 2012]. The motion-free 3D image is then reconstructed from the extracted patches using iterative super-resolution and 2D/3D registration steps. The motion-corrupted and misaligned patches are excluded during the reconstruction using an EM-based outliers rejection model. This process is extended to allow propagation of the placental masks to the final reconstruction through evaluating an MR specific point spread function, registration-based transformation, and the learned confidence weights.

Placenta Visualisation

This section presents an extension to the segmentation pipeline proposed earlier to visualise the reconstructed placenta based on a novel application of shape abstraction using a flexible cutting plane. It is supported by a mean-curvature flow skeleton [Tagliasacchi et al., 2012] generated from the triangulated polygonal mesh of the placenta segmentation and textured similar to curved planar reformation [Kanitsar et al., 2002], see Figure **??**. Clinicians revealed that such a representation is potentially desirable since it compares well to a flattened placenta after birth. This part is further evaluated in [Miao et al., 2017].

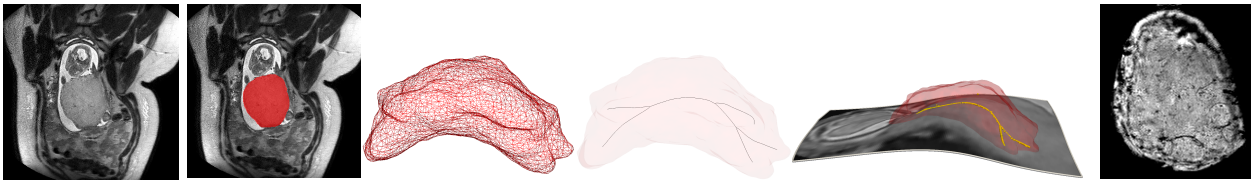


Figure 4.12: A native plane (a) cannot represent all structures of the placenta at once. Therefore, the placenta is segmented based on the multi-scale CNN (b); the motion is corrected using [Alansary et al., 2017]; project the placental mask into the the resulting isotropically resolved volume (c), extract the mean curvature flow skeleton [Tagliasacchi et al., 2012] (black lines in (d)), use the resulting points to support a curved surface plane (e) and visualise this plane with curved planar reformation [Kanitsar et al., 2002] (f). The plane in (f) covers only relevant areas, hence grey value mapping can be adjusted automatically to emphasis placental structures.

4.2.3 Experiments and Results

Data

The segmentation pipeline is tested on two dissimilar datasets that are different in health status, gestational ages and acquired using different scanning parameters. All scans have been ethically approved. *Dataset I* contains 44 MR scans of healthy foetuses at young gestational age between 20–25 weeks. The data has been acquired on a Philips Achieva 1.5T, the mother lying 20° tilt on the left side to avoid pressure on the inferior vena cava. SSFSE T2-weighted sequences are used to acquire stacks of images that are aligned to the main axes of the fetus.

Usually three to six stacks are acquired for the whole womb and the placenta with a voxel size of $1.25 \times 1.25 \times 2.50$mm. *Dataset II* contains 22 MR scans of healthy foetuses and foetuses with intrauterine foetal growth restriction (IUGR) at wide gestational age between 20–38 weeks. The data was acquired with a 1.5T Philips MRI system using SSFSE sequences and a voxel size of $0.8398 \times 0.8398 \times 4$mm. Ground truth labels for both datasets have been obtained manually slice-by-slice in 2D views from the original motion-corrupted stacks by a clinical expert.

Evaluation Metrics

The performance is evaluated using three main metrics: (*i*) Dice similarity coefficient (Section 4.1.3) to measure the accuracy of the segmentation; (*ii*) absolute volume similarity to measure the volumetric error between the segmented and the ground truth volumes; and (*iii*) average Hausdorff distance as a distance error metric between the segmented and the ground truth surfaces. The Hausdorff distance (HD) Huttenlocher et al. [1993] is measured between surfaces of two segmentations S_1 and S_2 as the maximum distance of the surface of S_1 to the nearest point on the surface S_2 (as shown in Figure **??**). HD can be formulated as:

$$H(S_1, S_1) = \max_{c \in S_1}\{\min_{e \in S_2}\{d(c, e)\}\}, \tag{4.7}$$

where c and e denote points on S_1 and S_2 respectively, and $d(c, e)$ is the Euclidean distance between c and e. The *bidirectional* HD between the segmented placenta S and its ground truth G is then defined as:

$$H_{Bi}(G, S) = max\{H(G, S), H(S, G)\}. \tag{4.8}$$

Experiments

A first experiment [exp-1] evaluates the automatic segmentation of the placenta on *Dataset I* using a 4-fold cross validation (11 test patients and 33 training patients per fold). The main aim of this experiment is to evaluate the performance of the segmentation framework on a healthy

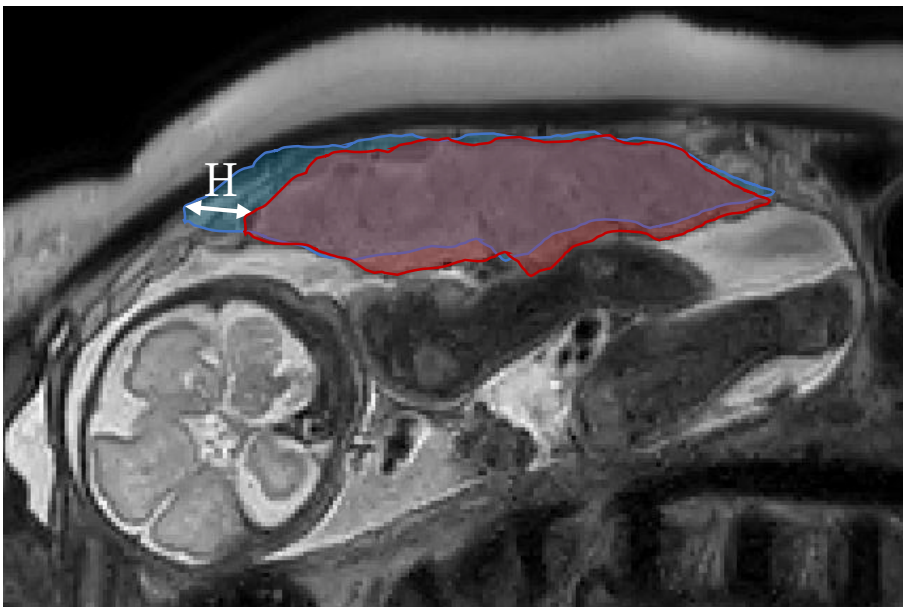
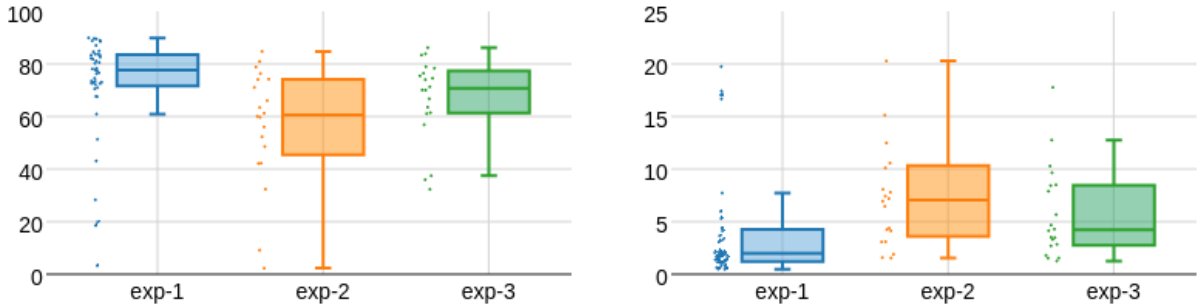


Figure 4.13: A schematic illustration for the Haussdorf distance (HD) calculation between ground truth (*red*) and segmented (*blue*) placenta using Equation 4.8.

homogeneous dataset. The results for this experiment are $71.95 \pm 19.79\%$ Dice, $30.92 \pm 33.68\%$ absolute volume difference, and $4.94 \pm 6.93mm$ average Hausdorff distance.

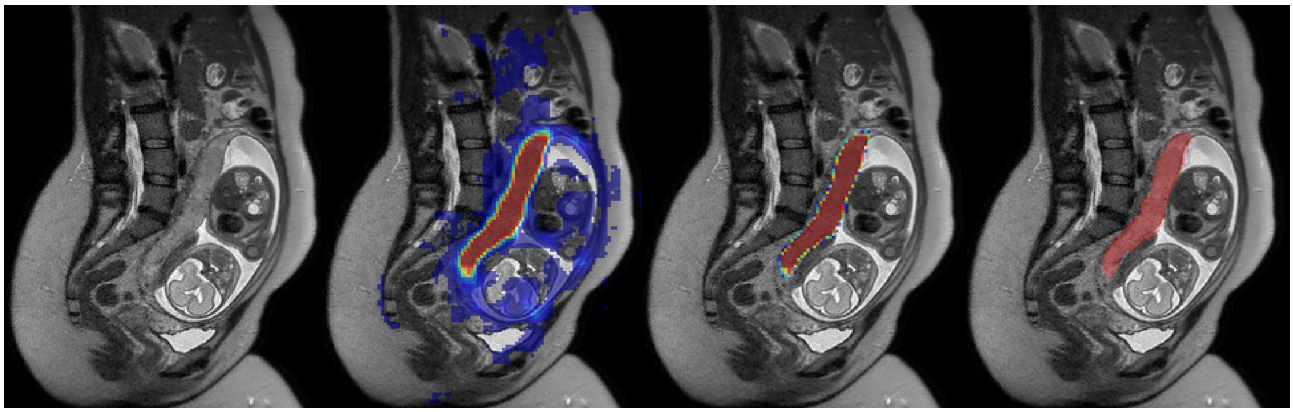
In a second experiment [exp-2], the CNN is trained using the whole 44 subject from *Dataset I* and test it on the 22 subjects from *Dataset II*. Where datasets *I* and *II* are significantly different using different scanners and scanning parameters. In addition, the gestational age range of the foetuses in *Dataset II* is wider, which has a big influence on the foetal body and placenta sizes. Hence, the performance of the segmentation framework is evaluated on a test data from a different environment. The results of this experiment are $56.78 \pm 21.86\%$ Dice, $48.19 \pm 46.96\%$ absolute volume difference, and $8.41 \pm 7.1mm$ average Hausdorff distance.

To resemble a realistic transfer learning application, a third experiment [exp-3] is designed using both datasets. The network is evaluated with 2-fold cross validation, 10 test subjects from *Dataset II*, and 44+10 training subjects from *Dataset I* and *Dataset II*. This experiment yielded a Dice accuracy of $66.89 \pm 15.35\%$, an absolute volume difference of $33.05 \pm 30.71\%$, and an average Hausdorff distance of $5.8 \pm 4.24mm$. Detailed results are shown in Figure 4.14. Training one fold takes approximately 40 hours and inference can be done within 2 minutes on an Nvidia Tesla K40.



(a) Dice (b) AHD

Figure 4.14: Evaluation of the proposed method using (a) average Dice coefficient and (b) average Hausdorff distance. [exp-1] refers to 4-fold evaluation on *Dataset I*, [exp-2] uses *Dataset I* for training and *Dataset II* for testing, and [exp-3] mixes both datasets for training and uses unseen examples from *Dataset II* for testing.

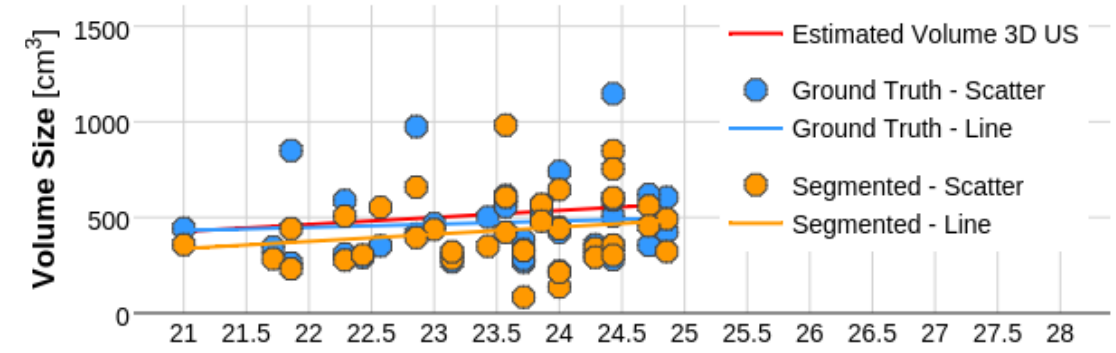


(a) Input (b) CNN (c) CRF (d) Output

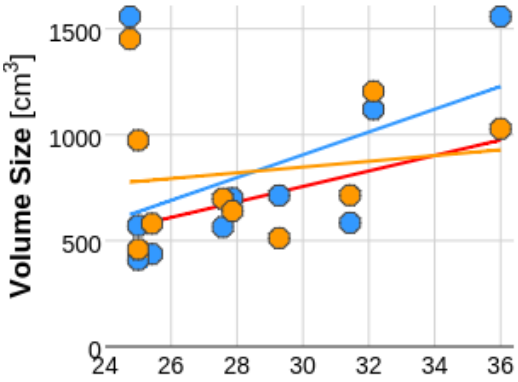
Figure 4.15: A sample 2D plane of the whole uterus from foetal MRI. (a) is the original greyscale input image. (b) is the output probability map overlaid with the input image. (c) is the regularised probability map after applied CRF. (d) is the final segmented placental mask.

Evaluation of Clinical Parameters

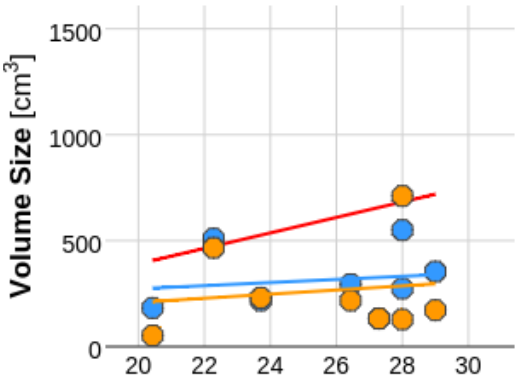
The segmentation results are compared to known values from the clinical practice. Routledge et al. [2015] shows that the average placental volume increases from $252.4cm^3$ at 20 weeks to $1421.5cm^3$ at 37 weeks. Figure 4.16 compares these values from the clinical literature [Thame et al., 2004] to the automatically measured volumes from testing datasets. It shows that the presented approach achieves very similar volumetric results compared to both expert estimations and clinical literature. In addition, the slope parameters of automatic segmentation and ground truth are not significantly different with p-value 0.94. Pathological cases from *Dataset II* show differences to scans of healthy placentas in Figure 4.16(c).



(a) Dataset I



(b) Dataset II (healthy)



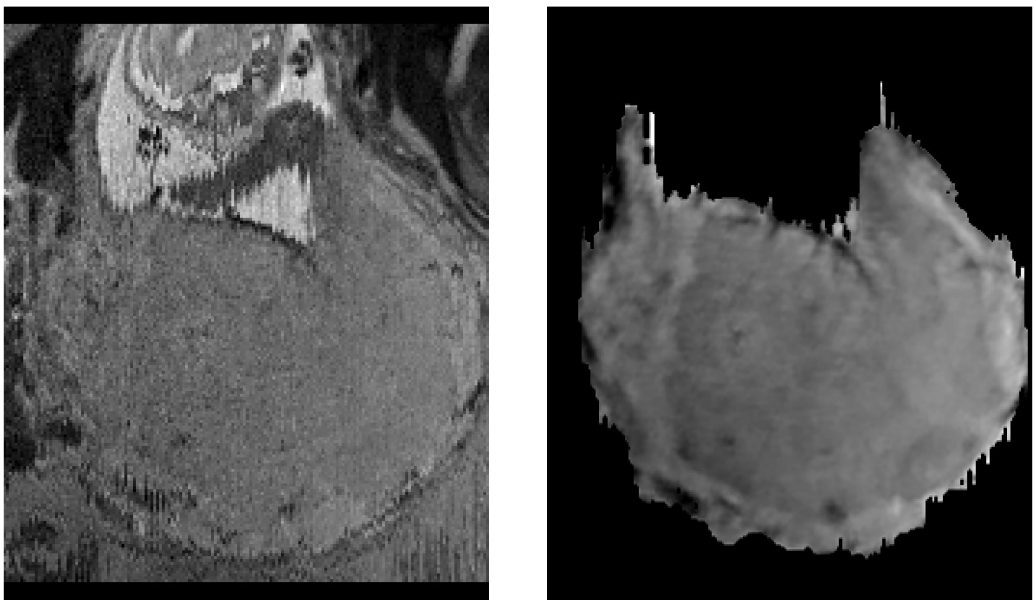
(c) Dataset II (IUGR)

Figure 4.16: A graph comparing automatic segmentations from the proposed automatic approach, the ground truth (expert) and the linear estimations [Thame et al., 2004] of the placental volumes versus their gestational ages. (a) shows that volumes automatically segmented placentas from the first experiment [exp-1] using *Dataset I* are very close to both the expert and theoretical estimations. However, the third experiment [exp-3] using healthy subjects from *Dataset II* shows less consistency of the segmented volumes (b) due to the large dissimilarity test data. Moreover, (c) shows more inconsistency by testing on foetuses with IUGR from [exp-3]. (foetuses with unknown gestational age were excluded)

Motion Compensation

Evaluating the quality of a motion compensated reconstruction is challenging due to the absence of the motion-free ground truth data. Assuming that the 2D in-plane patches from the original 3D stacks have no motion artefacts, reconstructed patches are evaluated using these motion-free 2D patches as ground truth. The average peak signal-to-noise ratio (PSNR) is calculated for all the patches of each subject. The baseline represents the quality of patches from non-native slice orientations, which is comparable to using a single motion corrupted stack for diagnostics

with arbitrary cutting planes. The baseline has low PSNR values due to the motion artefacts between the input stacks. These values increase during the reconstruction iterations as a result of reducing the motion artefacts of the segmented placenta, see Figure 4.18.



(a) Input (b) After segmentation and PVR

Figure 4.17: A sample 2D plane of the whole uterus cropped around the placenta from foetal MRI. (a) is the original greyscale input image. (b) is the reconstructed high-resolution segmented placenta using PVR.

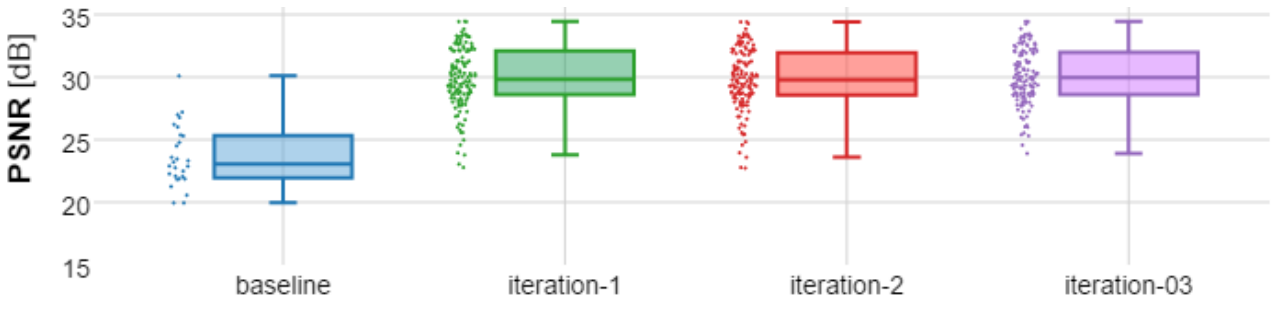


Figure 4.18: Comparisons between the 3D reconstructed placenta using superpixel-based PVR reconstructions [Alansary et al., 2017] with initial superpixel size $20x20$ pixels and a baseline using directly 2D slices from the input stacks for multi-planar examination. (stacks with high motion artefacts were excluded)

4.3 Conclusion

The first part of this Chapter have presented an automatic framework for localising the brain in foetal MRI scans using superpixel graphical models. Superpixels have enabled the proposed detection algorithm to be faster and more efficient than using pixels for classification. Also, extending the extracted features from the individual superpixels to include features from the neighbours using superpixel graphical models, have provided more information about the image context instead of using only the local information. The evaluation results achieved 94.55% accuracy for the brain detection and 98.18% for the centre of the brain, without the need of any prior information such as the gestational age or brain size. Moreover, the extracted brain can be be used for developing automatic motion correction and registration techniques for foetal MRI.

The second part have presented a fully automatic segmentation framework for the human placenta from motion corrupted foetal MRI scans. Rigorous experiments have been performed on two different testing datasets in order to evaluate thoroughly the presented segmentation approach, which is based on a 3D deep multi-scale convolutional neural network combined with conditional random field segmentation refinement. The experiments show that this framework can tackle motion artefacts by achieving segmentation accuracy of 71.95% for healthy foetuses. It is also capable of segmenting the placenta from dissimilar data by achieving segmentation accuracy of 66.89% for a cohort mixed with cases of intrauterine foetal growth restriction from different scanners. Moreover, the proposed framework scope was extended to real clinical applications by compensating motion artefacts using slice to volume registration techniques, as well as providing a novel standardized view into the placental structures using skeleton extraction and curved planar reformation. In future work, the potential use of the standardized placenta views will be investigated for image-based classification and automatic detection of abnormalities. Another future direction will be exploring how the CNN handles motion artefacts, and the possibility of applying the learned feature maps for motion correction.

Chapter 5

Anatomical Landmark Detection and View Planning

This chapter is based on:

Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. **Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection.** *MIDL Special Issue in Medical Image Analysis*, 2018. under review. [**?**]

Amir Alansary, Loic Le Folgoc, Ghislain Vaillant, Ozan Oktay, Yuanwei Li, Wenjia Bai, Jonathan Passerat-Palmbach, Ricardo Guerrero, Konstantinos Kamnitsas, Benjamin Hou, Steven McDonagh, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. **Automatic View Planning with Multi-scale Deep Reinforcement Learning Agents.** *In Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018. [Alansary et al., 2018]

Abstract

This chapter introduces artificial intelligent agents that are capable of learning to perform med-

ical imaging tasks such as landmark detection and view planning. The agents learns to mimic experienced operators using a reinforcement learning (RL) scenario, which enables a natural learning paradigm by interaction with the environment. The first part of this chapter evaluates novel deep RL strategies to train agents that can precisely and robustly localize target landmarks in medical scans. An artificial RL agent learns to identify the optimal path to the landmark by interacting with an environment, 3D images. It also proposes fixed- and multi-scale search strategies with novel hierarchical action steps in a coarse-to-fine manner. Several deep Q-network (DQN) architectures are evaluated for detecting multiple landmarks using three different medical imaging datasets: foetal head ultrasound, adult brain and cardiac MRI. The performance of the proposed agents surpasses state-of-the-art supervised and RL methods. The experiments also show that multi-scale search strategies perform significantly better than fixed-scale agents in images with large field of view and noisy background such as in cardiac MRI. The second part presents a fully automatic method to find standardized view planes in 3D image acquisitions. The proposed method employs a multi-scale RL agent framework and extensively evaluates several DQN-based strategies. The results are evaluated using the distance between the anatomical landmarks and detected planes, and the angles between their normal vector and target. The proposed algorithm is assessed on the mid-sagittal and anterior-posterior commissure planes of brain MRI, and the 4-chamber long-axis plane commonly used in cardiac MRI, achieving accuracy of $1.53mm$, $1.98mm$ *and* $4.84mm$, *respectively.*

5.1 Introduction

The methods presented in previous chapters may fail to solve medical imaging tasks that: contain few training samples (few landmarks and thousands of background points), or require navigation to find a target solution. This chapter introduces artificial intelligent agents that are capable of learning to perform such tasks. The agents learn to mimic experienced operators using an RL scenario, which enables a natural learning paradigm by interaction with the environment. The following sections propose two different RL agents that learns to find the optimum trajectory to target anatomical landmarks and standard views. The learning part is

done using reward signals that guide the agent towards the correct solution. However, during inference, the agent moves towards the correct solution without any guidance in unseen data (environments). RL has the advantage to solve non-differential problems using the power of reward signals. It is capable of solving different tasks using actions. Another major advantage is generated unlimited labelled samples for training with every step or action. The learning paradigm potentially can involve experts interactions (learning by imitations), which can open future directions to learn anatomical trajectories.

5.2 Anatomical Landmark Detection

Automatic detection of anatomical landmarks is an important step for a wide range of applications in medical image analysis. Whereas manual annotation of such landmarks is a tedious task and prone to observer errors. In the literature, most of the published works have adopted machine learning algorithms for landmark detection by learning a combined appearance and image based model [Criminisi et al., 2013, Gauriau et al., 2015, Li et al., 2018a, Oktay et al., 2017, Payer et al., 2016, Xu et al., 2017]. However, the majority of these methods rely on learning the search strategy and localisation separately, which may increase the possibility that the second stage misclassifies multiple candidates from the output of the first stage as positive. This section porposes a framework for learning the search strategy and localisation as one task. The proposed framework formulates the landmark detection problem as a sequential decision making process of a goal-oriented agent, navigating in an environment (the acquired image) towards a target landmark. At each time step, the agent decides which direction it has to proceed towards an optimal path to the target landmark. Reinforcement learning (RL) is used to learn an approximation of the optimal solution of this sequential decision making process. As explained in Chapter 2, RL is a computational approach for learning by interacting with an environment so as to maximize cumulative reward signals [Sutton and Barto, 1998b]. Different variants of deep Q networks (DQN) [Mnih et al., 2015, Van Hasselt et al., 2016, Wang et al., 2015b] architectures are levaraged for the landmark detection task, where a deep CNN is applied to learn complex representations of the input images (states) in order to estimate the

quality of taking an action a_t during a state s_t.

This section proposes: (*i*) use cases of several different DQN based models for anatomical landmark localisation; (*ii*) fixed- and multi-scale search strategies for the optimal path with novel hierarchical action steps for agent based landmark localisation frameworks; and (*iii*) extensive evaluations of the performance of the proposed agents by running multiple experiments on different MRI and US images, where the results outperform state-of-the-art.

5.2.1 Method

The problem of landmark detection is formulated as a Markov decision process (MDP), where an artificial agent learns to make a sequence of decisions towards the target landmark. In this setup, the input image defines the environment E, in which the agent navigates using a set of actions. The main goal of the agent is to find an anatomical landmark. This section explains the main elements of the MDP that includes a set of actions A, a set of states S, and a reward function R. Figure 5.1 shows an illustrative figure of the proposed 3D medical image environment, where the agents learn to optimise the best path to the target landmark (optimal policy). During testing, the agent does not receive any rewards and does not update the model neither, it just follows the learned policy.

Navigation Actions

The agent interacts with E by taking movement action steps $a \in A$ that imply a change in the current point of interest location. The set of actions A is composed of six actions, $\{\pm a_x, \pm a_y, \pm a_z\}$, in the positive or negative direction of x, y or z. For instance, taking a $+a_x$ action means that the agent will move a fixed step size in the positive x-direction. Figure 5.2 shows a schematic visualization of these navigation actions in a 3D scan.

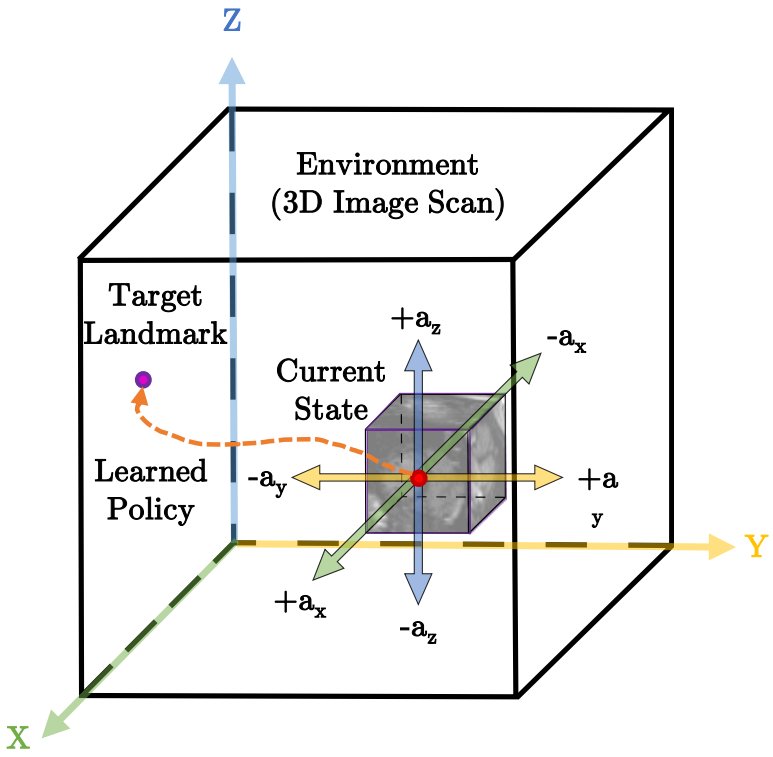


Figure 5.1: Schematic diagram of the proposed RL agent interacting with the 3D image environment E. At each step the agent takes an action towards the target landmark. These sequential actions forms a learned policy forming a path between the starting point and the target landmark.

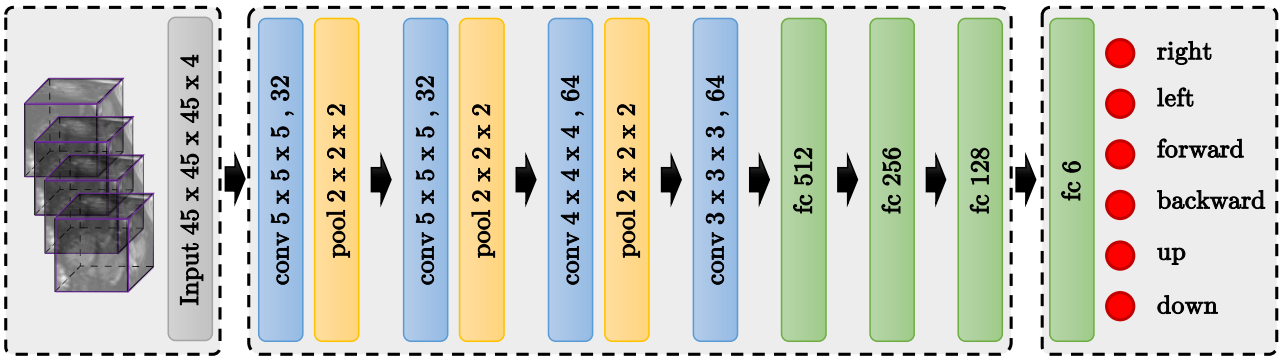


Figure 5.2: Schematic illustration of the proposed DQN-based network architecture for anatomical landmark detection. The input are volumetric samples along the 3D trajectory of the region centred around the current location of the agent. The output is the approximated Q-value for the six possible actions. The agent will pick the action with the highest Q-value. This is done sequentially until the agent finds the target landmark.

States

An environment E is represented by a 3D image, where each state s defines a 3D Region of Interest (ROI) centred around the target landmark. A frame history buffer is used to capture the last 4 action steps (ROIs) taken by the agent in its search for the landmark. This stabilizes the search trajectories and prevents the agent from getting stuck in repeated cycles.

Reward Function

Designing good empirical reward functions R is often difficult as RL agents can easily overfit the specified reward, and thereby produce undesirable or unexpected results. For the view planning task, the difficulty arises from designing a reward that encourages the agent to move towards the target plane while still being learnable. Thus, R should be proportional to the improvement that the agent makes to detect a landmark after selecting a particular action. Here, the reward function is defined as: $R = D(P_{i-1}, P_t) - D(P_i, P_t)$, where D represents the Euclidean distance between two points. P_i is the current predicted landmark's position at step i, with P_t the target ground truth landmark's location. The difference between the two Euclidean distances, the previous step and current step, signifies whether the agent is moving closer to or further away from the desired target location.

Terminal State

The final state is reached when there are no further transition states for the agent to take. This means that the agent has found the target landmark P_t. The terminal state during training is triggered when the distance between the current point of interest and the target landmark are less than or equal to 1mm. Finding a terminal state during testing is more challenging, due to the absence of the landmark's true location. One solution is to define a new trigger action that terminates the sequence of the current search when the target state is reached [Caicedo and Lazebnik, 2015, Maicas et al., 2017]. Although this modifies the environment by marking the region that is centred around the correct location of the target landmark, it increases the

complexity of the task to be learned by increasing the action space size. It also introduces a new parameter, maximum number of interactions, which needs to be set manually. It may also slow down the testing time in cases where the terminal action is not triggered. Riedmiller [1998] found that the agent shows strong oscillating behavior around the terminal state. The oscillation property is adopted to terminate the search process during testing without defining an explicit terminal state. In contrast to [Ghesu et al., 2016], the terminating state is chosen based on the corresponding lower Q-value. Q-values are lower when the agent is closer to the target point, and higher when it is far. Intuitively, by awarding higher Q-values, DQN encourages the agent to take any action from states that are far away from the target landmark, and conversely for closer states.

Multi-scale Agent

In images with large field of view, noisy background can deteriorate the performance of the agent for finding the target landmark. In order to capture spatial relations within a global neighborhood, a multi-scale search strategy [Ghesu et al., 2017a,b] is adopted in a coarse-to-fine fashion with novel hierarchical action steps. The environment E samples a fixed size image-grid with initial spacing (S_x, S_y, S_z) mm around the current location P_o, and the agent searches for the target landmark with initial large action steps. Once the target point is found, E samples the new image-grid with smaller spacing, as well as the agent uses smaller action steps. Coarser levels in the hierarchy provide additional guidance to the optimization process by enabling the agent to see more structural information. Finer scales, on the other hand, provide more precise adjustments for the final estimation of the plane. Similarly, larger action steps speed convergence towards the target plane, while smaller steps fine tune the final estimation of plane parameters. The same DQN is shared between all levels in the hierarchy.

5.2.2 Experiments and Results

The performance of different RL agents for anatomical landmark detection is evaluated on three different US and MR datasets. Fixed- and multi-scale search strategies are experimentd

by sampling with different spacing values. During testing, the initial selected points arefixed for all models for a fair comparison between different variants of the proposed method. A set of 19 different starting points are selected from the whole image for every testing subject in order to report more robust results. The accuracy is measured based on the Euclidean distance error between detected and target landmarks. Finally, extensive comparison are run between different DQN-based architecture, namely DQN, DDQN, Duel DQN, and Duel DDQN.

Experiments

During training, a random point is sampled from a region with size 80% of the whole image dimensions around the centre. An ROI of size 45x45x45 voxels is sampled around the selected point. The agent follows an ϵ-greedy exploration strategy, where at every step it selects an action uniformly at random with probability $(1 - \epsilon)$. Every trial to find the target landmark is called an *episode*. Here, heuristically, 1500 frames are used to limit the maximum number of frames per episode. During testing, the agent follows the learned policy by selecting the action with highest Q-value at each step.

Comparison with State-of-the-art

The performance of the presented agents agents are assessed compared to recent published works based on fixed-scale [Ghesu et al., 2016] and multi-scale [Ghesu et al., 2017b] RL, and fully-supervised deep CNNs [Li et al., 2018a] for the detection of the cavum septum pellucidum point in 3D foetal US head scans. In [Ghesu et al., 2016, 2017b], fixed- and multi-scale DQN-based agents detect the target landmark in a similar RL formulation. The method from Li et al. [2018a] is based on repeatedly passing patches to a CNN until the estimated point position converges to the true landmark location with full supervision, called patch-based iterative network (PIN). For this experiment, a set of 72 foetal head ultrasound scans are used, divided into 21 training and 51 testing images, as detailed in Experiment 5.2.2. The performance of the multi-scale agents improve upon the state-of-the-art methods as shown in Table 5.1.

RL Fixed-scale [Ghesu et al., 2016]	RL Multi-scale [Ghesu et al., 2017b]	Supervised PIN Single-Landmark [Li et al., 2018a]	Supervised PIN Multiple-Landmarks [Li et al., 2018a]
7.37 ± 5.86	6.51 ± 5.41	5.47 ± 4.23	5.50 ± 2.79

(a) State-of-the-art results

	DQN	DDQN	Duel DQN	Duel DDQN
Fixed-Scale	4.95 ± 3.09	5.01 ± 2.48	6.29 ± 3.95	5.12 ± 3.15
Multi-Scale	$\mathbf{3.66 \pm 2.11}$	4.02 ± 2.20	4.17 ± 2.62	4.02 ± 1.55

(b) Our results

Table 5.1: Comparison with stat-of-the-art RL [Ghesu et al., 2016, 2017b] and supervised [Li et al., 2018a] for detecting the cavum septum pellucidum point in foetal ultrasound images. Distance errors are in *mm*.

Experiment-I: Foetal Head Ultrasound

Finding the target landmarks in such images is a challenging task because of ultrasound artefacts such as shadowing, mirror images, refraction, and foetal motion. Three levels are used for the multi-scale agent with spacing values from 3mm to 1mm, decreasing by one at each level. Hierarchical action steps are chosen from 9 to 1 steps per iteration, dividing by 3 at each level of the hierarchy.

Dataset: 72 foetal head US scans [1] are randomly divided into 21 and 51 images for training and testing. Three landmarks are selected, the right and left cerebellum and cavum septum pellucidum, that define the transcerebellar (TC) plane, commonly used for foetal sonographic screening examination, see Figure 5.3. The selected landmarks were manually annotated by clinical experts using three orthogonal views. All images were roughly aligned to the same orientation and re-sampled to isotropic 0.5mm spacing.

Results: Table 5.2 shows the comparative results of the performance of different agents. In general, all methods share similar performance including speed and accuracy. However, Duel DQN achieves the best accuracy detecting the right and left cerebellum points, while DQN performs the best for finding the cavum septum pellucidum point. Additionally, the multiscale

[1] `http://www.ifindproject.com/`

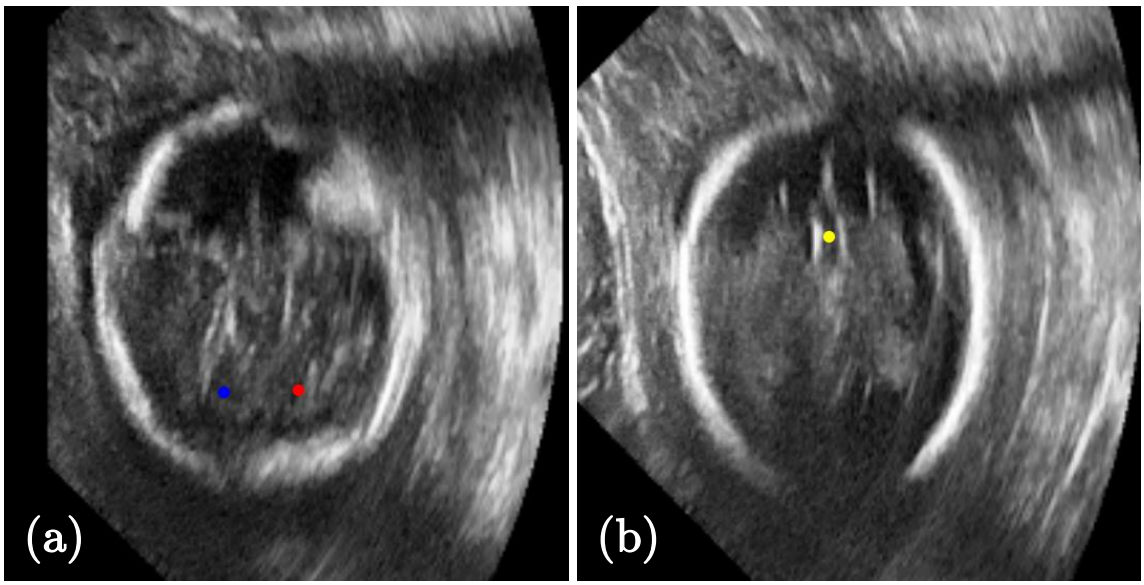


Figure 5.3: Sample 2D images from foetal head ultrasound showing the target landmarks: (a) right (red) and left (blue) cerebellum, and (b) cavum septum pellucidum (yellow) points.

strategy improves the performance of the agents and increase pace to the target point thanks to the hierarchical action steps.

Method	Right Cerebellum		Left Cerebellum		Cavum Septum Pellucidum	
	FS	MS	FS	MS	FS	MS
DQN	4.17 ± 2.32	3.37 ± 1.54	2.78 ± 2.01	3.25 ± 1.59	$\mathbf{4.95 \pm 3.09}$	$\mathbf{3.66 \pm 2.11}$
DDQN	3.44 ± 2.31	3.41 ± 1.54	2.85 ± 1.52	2.95 ± 1.00	5.01 ± 2.84	4.02 ± 2.20
Duel DQN	$\mathbf{2.37 \pm 0.86}$	3.57 ± 2.23	$\mathbf{2.73 \pm 1.38}$	$\mathbf{2.79 \pm 1.24}$	6.29 ± 3.95	4.17 ± 2.62
Duel DDQN	3.85 ± 2.78	$\mathbf{3.05 \pm 1.51}$	3.27 ± 1.89	3.50 ± 1.70	5.12 ± 3.15	4.02 ± 1.55

Table 5.2: Comparison of different DQN-based agents using fixed-scale (FS) and multi-scale (MS) search strategies for the detection of right and left cerebellum, and cavum septum pellucidum landmarks in foetal US images. Distance errors are in *mm*.

Experiment-II: Brain MRI

Two landmarks are selected for this experiment: anterior and posterior commissure points, commonly used by the neuroimaging community to define the axial plane during image acquisition, see Figure 5.4.

Dataset: 832 isotropic 1mm MR scans were obtained from the ADNI database [Mueller et al., 2005], randomly divided into 728 and 104 images for training and testing, respectively. All brain images were skull stripped and affinely registered to the same space.

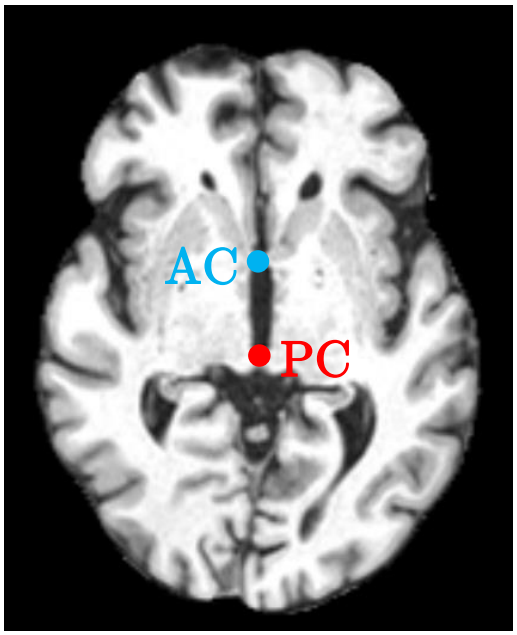


Figure 5.4: The anterior commissure (AC) and posterior commissure (PC) points in brain MRI.

Results: Similar to the foetal US experiment, Table 5.3 shows that the best performing agent varies for each landmark. However, the multi-scale strategy improves the performance of fixed-scale agents.

Method	Anterior Commissure		Posterior Commissure	
	FS	**MS**	**FS**	**MS**
DQN	3.04 ± 1.70	2.46 ± 1.44	$\mathbf{2.03 \pm 0.97}$	2.05 ± 1.14
DDQN	$\mathbf{2.62 \pm 1.24}$	2.61 ± 1.64	3.31 ± 1.2	$\mathbf{1.86 \pm 1.07}$
Duel DQN	3.04 ± 1.28	2.4 ± 1.42	3.6 ± 1.46	2.15 ± 1.24
Duel DDQN	2.97 ± 1.23	$\mathbf{2.01 \pm 1.29}$	2.04 ± 1.04	2.27 ± 1.22

Table 5.3: Performance of different agents using fixed-scale (FS) and multi-scale (MS) search strategies for the detection of anterior and posterior commissure landmarks in brain MR images. Distance errors are in *mm*.

Experiment-III: Cardiac MRI

Two landmarks are selected for this experiment: apex and centre of mitral valve, commonly used for defining the short axis view during image acquisitions, see Figure 5.5. They are also used to assist automatic segmentation methods by defining starting and ending slices in the acquired cardiac stack of 2D image sequence.

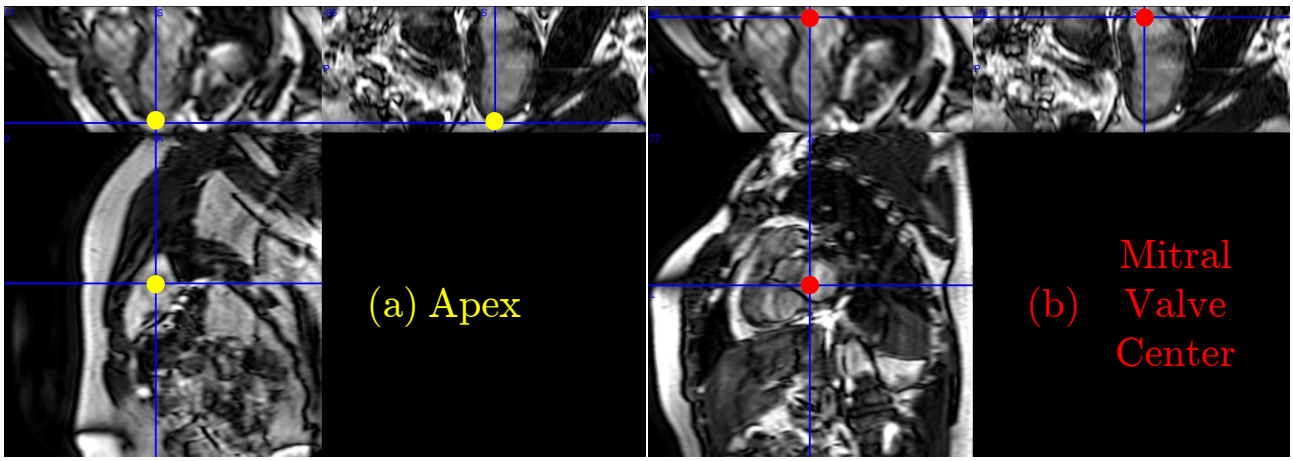


Figure 5.5: A 2D cardiac MR-image showing the apex and centre of the mitral valve points.

Dataset: 455 short-axis cardiac MR images of resolution $1.25 \times 1.25 \times 2$ mm obtained from the UK Digital Heart Project [de Marvao et al., 2014], randomly divided into 364 and 91 images for training and testing, respectively. All cardiac images are re-sampled to isotropic 1mm spacing.

Results: Table 5.4 shows that Duel DQN agents perform the best for detecting the apex, while all agents performs similarly for detecting the mitral valve center. Multi-scale agents achieve a slight decrease in the detection error compared to inter-observer errors. They also significantly improve upon stratified decision forests [Oktay et al., 2017] and fixed-scale agents, which is reasonable in cardiac imaging because of the bigger field of view and noisy background.

Method	**Apex**		**Mitral valve**	
Inter-observer Errors	5.79 ± 3.28		5.30 ± 2.98	
Decision Forests [Oktay et al., 2017]	6.74 ± 4.12		6.32 ± 3.95	
Our agents	**FS**	**MS**	**FS**	**MS**
DQN	7.49 ± 4.05	4.47 ± 2.63	$\mathbf{8.33 \pm 4.70}$	5.73 ± 4.16
DDQN	8.13 ± 5.60	4.53 ± 2.78	8.82 ± 4.80	$\mathbf{5.20 \pm 2.82}$
Duel DQN	$\mathbf{7.17 \pm 4.21}$	$\mathbf{4.42 \pm 2.67}$	8.82 ± 4.80	5.76 ± 3.89
Duel DDQN	7.59 ± 4.17	5.43 ± 3.37	8.63 ± 4.58	5.28 ± 2.61

Table 5.4: A comparison between inter-observer errors, stratified decision forests [Oktay et al., 2017], and different agents using fixed-scale (FS) and multi-scale (MS) search strategies for the detection of apex and centre of mitral valve landmarks in cardiac MR images. Distance errors are in mm.

Implementation

Training times are around 24 and 48 hours for individual landmarks from brain and cardiac MRI, respectively, using an NVIDIA GTX 1080Ti GPU. The experiments show that the agent is capable of finding the target landmark in less than 1 second for any random initialisation. During inference, the agent finds the target location using sequential steps, where each step takes around 0.5-1 milliseconds. The proposed agents are implemented with a batch size of 48, experience replay memory of size $1e5$, activation function PReLU for convolutional layers and leakyReLU for fully connected layers, ADAM optimizer, $\gamma = 0.9$, and $\epsilon = 0.9 - 0.1$. Figure 5.2 shows the architecture of the proposed DQN. The source code of this implementation is publicly available[2]. More visualizations are on the github repository showing different animated visual examples of trained agents searching for the target landmarks.

[2]https://git.io/fNUoS

5.3 Automatic View Planning

Standard view planning through a defined anatomy is commonly used in clinical practice to establish comparable metrics. Obtaining accurate biometric measurements that are comparable across populations is essential for diagnosis and support clinical decisions. In addition to evaluating and assessing anomalies, the standardized views can be used as an initialisation step for several registration or image alignment methods. They are often constrained to the native orientation of a 3D image acquisition, while the appearance of relevant structures can exhibit large variance according to the positioning of the imaging plane. Moreover, navigating through target anatomy to find the required view plane is tedious and operator-dependent. Figure **??** shows different view plans from cardiac and brain MRI images.

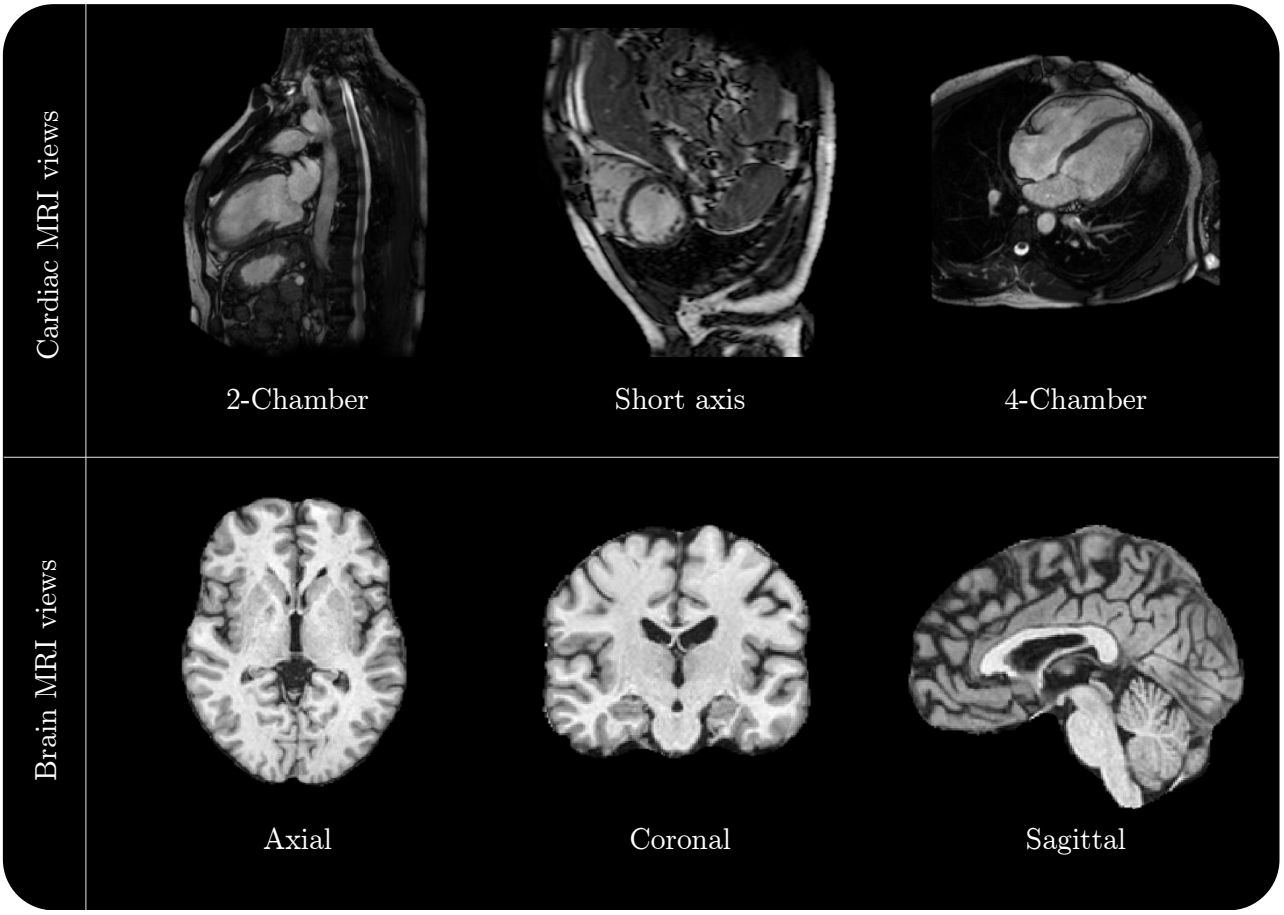


Figure 5.6: Standardised views in brain and cardiac MRI images.

For instance, the acquisition of the 4-chamber (4CH) imaging plane utilized in cardiac MRI requires a multi-step approach: (*i*) localising three planes (axial, coronal, and sagittal); (*ii*)

acquiring an axial stack scanned from above the aortic arch to below the level of the heart; (*iii*) defining the 2-chamber (2CH) view plane perpendicular to the axial plane and parallel to the inter-ventricular septum (IVS), and in the middle of left ventricle (LV); (*iv*) defining pseudo short axis (SA) view plane perpendicular to the 2CH view and aligned with the mitral valve; (*v*) defining 4CH view plane perpendicular to SA, and through the centre of the LV and the intersection of interior and inferior of the free wall, and dividing the heart along the long axis in the 2CH view. Figure **??** shows the corresponding 4CH and 2CH view planes from a short-axis cardiac MRI scan.

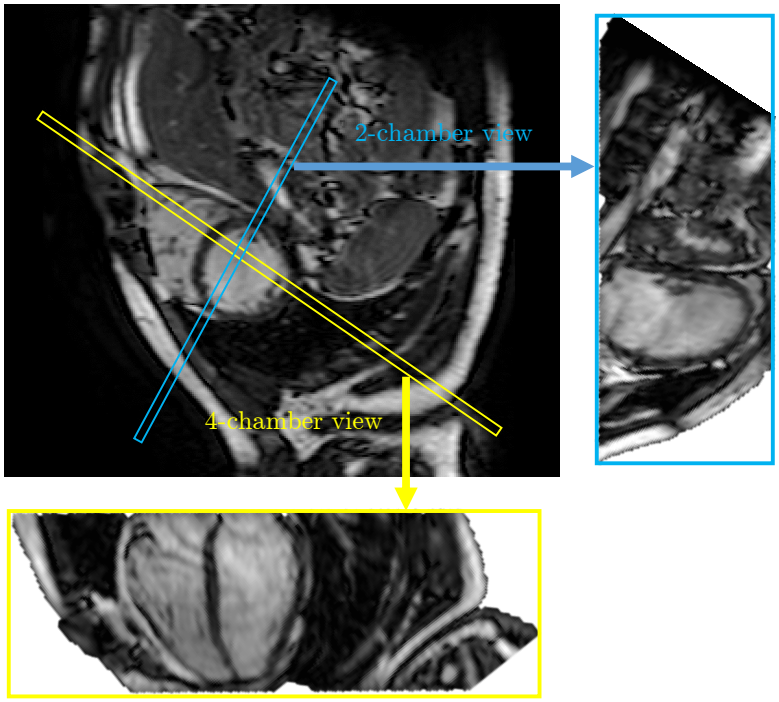


Figure 5.7: The acquisition of 4-chamber and 2-chamber view planes from a short-axis cardiac MRI scan.

In previous works [Le et al., 2017, Lu et al., 2011], automatic view planning in cardiac MRI was performed based on learning the anatomy segmentation and detecting anchor landmarks. These methods require prior knowledge of the whole 3D image for the purpose of plane detection. This also involves manual annotation of anatomical landmarks, which is a tedious and time-consuming task. This section proposes a novel artificial agent that is trained using RL to navigate through a 3D image scan toward the standard view (target) plane without any manual labelling. The navigation process is performed by sampling new planes at every step during the agent's lifetime. The proposed view planning framework is

shown in Figure 5.8.

5.3.1 Method

The plane P is defined in the Cartesian coordinates of the 3D image as: $ax + by + cz + d = 0$. Where (a, b, c) represent the normal direction (cosine) to this plane and d is the distance of the plane from the origin. To automate standard view planning, the agent learns to find the appropriate parametrisation of the target plane. Following are the main elements of the proposed RL framework for view planning.

States

An environment E is represented by a 3D scan and s is a 3D region of interest that contains P. A frame history buffer of size 4 is used for storing the last planes from previous steps to stabilize search trajectories and prevent getting stuck in repeated cycles.

Actions

The agent interacts with E by taking action steps $a \in A$ to modify the position parameters of the plane. The action space consists of eight actions, $\{\pm a_{\theta_x}, \pm a_{\theta_y}, \pm a_{\theta_z}, \pm a_d\}$, which update the plane parameters $a = cos(\theta_x + a_{\theta_x})$, $b = cos(\theta_y + a_{\theta_y})$, $c = cos(\theta_z + a_{\theta_z})$ and $d = d + a_d$.

Rewards

The reward is defined as $R = \text{sgn}(D(P_{i-1}, P_t) - D(P_i, P_t))$, where D is a function to take the Euclidean distance between plane parameters. P_i is the current predicted plane at step i, and P_t is the target ground truth plane. The difference of the parameter distances, between the previous and current steps, signifies whether the agent is moving closer to or further away from the desired plane parameters. $R \in \{+1, 0, -1\}$ provide the agent with a per step (non-sparse) reward signal, with zero-valued R presents plane oscillations around the correct solution.

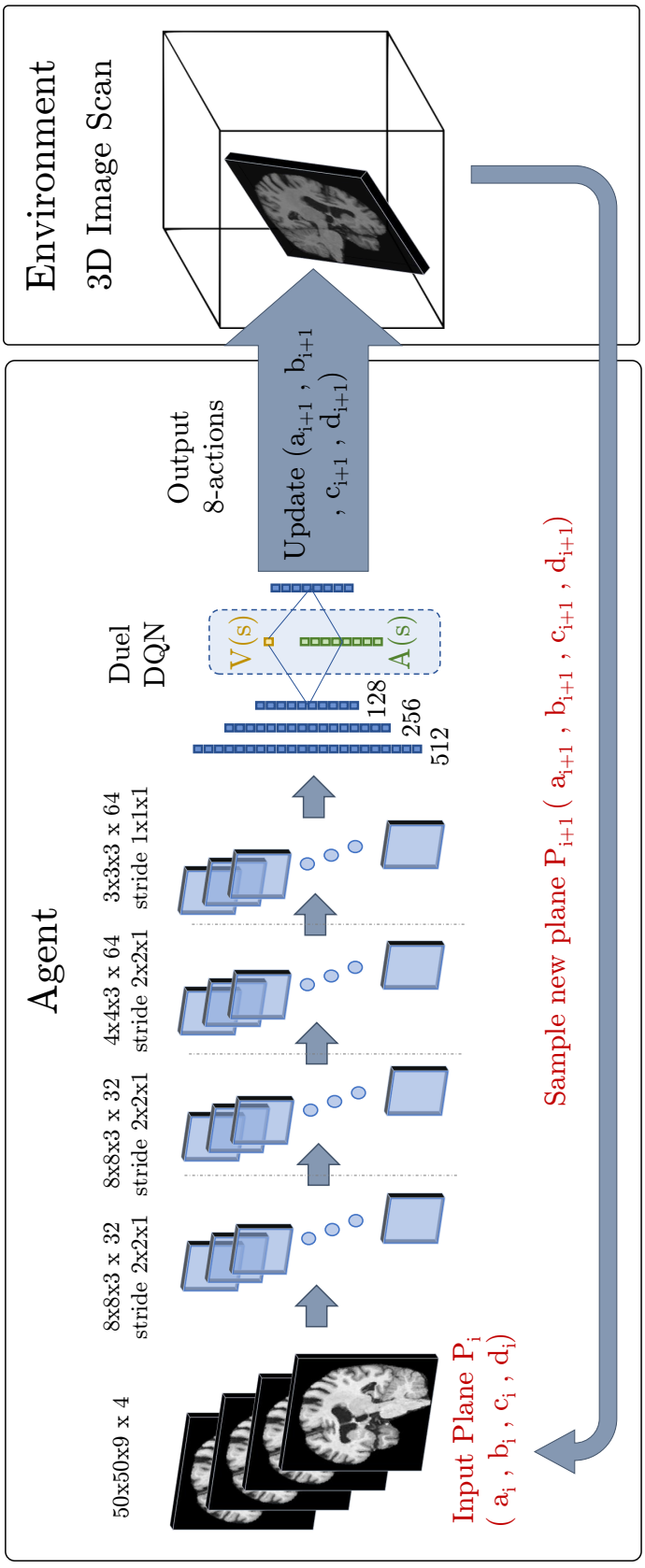


Figure 5.8: The pipeline of the proposed multi-scale RL agent. Initially, the environment samples a plane: $ax + by + cz + d = 0$, from the 3D image scan. The agent selects an action to update a single parameter for sampling the next plane. This process is repeated till the agent reaches a terminal state (oscillation).

Terminal State

The final state is defined as the state in which the agent finds the target plane P_t. The terminal condition is triggered when oscillation is detected, without the need to expand the action space by adding a new trigger terminal action. The terminating state is chosen according to the action with the lowest Q-value.

Multi-scale Agent

In order to provide more structural information, the agent use a multi-resolution search strategy in a coarse-to-fine fashion with hierarchical action steps. In this scenario, E samples a grid of a fixed plane size (P_x, P_y, P_z) of voxels around the plane origin P_o and initial spacing (S_x, S_y, S_z) mm. Initially, the agent searches for the plane with higher action steps. Once the target plane is found, E samples the new planes with smaller spacing and the agent uses smaller action steps. Coarser levels enables the agent to see more contextual information, while finer scales provide more precise adjustments for the final estimation of the plane. Larger step actions speed up the solution towards the target plane, while smaller steps make small adjustments to the final estimation of plane parameters.

5.3.2 Experiments and Results

The proposed algorithm is assessed using 12 different experiments; a combination of four different DQN-based methods with three target planes. The results are evaluated using two main metrics: (*i*) distance error by measuring the distance between anatomical landmarks and the detected planes; and (*ii*) orientation error by calculating the angle between normal vectors of the detected and target planes.

Training

A random point is sampled from the 3D input image. The initial random plane is then defined using the normal vector between the centre of the image and the random point. The origin of this plane is the projected point of the centre of the input image. Finally, a plane of size $(50, 50, 9)$ voxels is sampled around the plane origin with initial spacing $3 \times 3 \times 3$ mm. Initial $a_{\theta_x}, a_{\theta_y}, a_{\theta_z}$ equal 8 and a_d equals 4. With every new scale $a_{\theta_x}, a_{\theta_y}, a_{\theta_z}$ decrease by a factor of 2 and a_d decrease 1 unit. 3-levels of scale with spacing from 3 to 1 mm are used for the brain experiments, and 4-levels of scale from 5 to 2 mm for the cardiac experiment. For experiments on cardiac images, initial planes are sampled randomly from the 3D input image within 20% around the centre of the image, to avoid sampling outside the field of view.

Testing

During inference, the environment samples a plane and the agent updates sequentially new plane's parameters until reaching the terminal state. In order to have a fair comparison between different variants of the proposed method, the initial plane is chosen randomly and fixed for all models during evaluation.

[Experiment-I] 4CH View Planning in Cardiac MRI

Datasets

A set of 455 short-axis cardiac MRI images of resolution $(1.25 \times 1.25 \times 2)$ mm obtained from the UK Digital Heart Project [de Marvao et al., 2014]. A subset of 364 and 91 images are used for training and testing. Six landmarks are manually annotated by expert, and projected on the 4-chamber plane: (*i-ii*) the two right ventricle (RV) insertion points; (*iii-iv*) lateral wall turning points for the right and left ventricles; (*v*) apex, and (*vi*) the centre of the mitral valve, see Figure 5.9.

Results

View planning in cardiac images is challenging because of the complexity of cardiac anatomy,

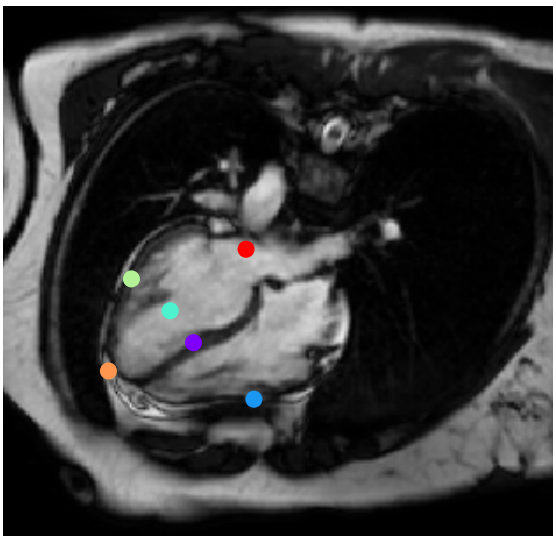


Figure 5.9: Ground truth 4CH view from cardiac MRI with the projected two RV insertion points(violet, green), RV and LV lateral wall turning points (blue, lime), apex (orange), and the centre of the mitral valve (red).

and the low quality and high variability between different scans. The agent has to navigate in a bigger field of view compared to brain images in Experiment 5.3.2. Thus Duel DQN-based architectures achieve the best results for detecting the 4-chamber plane with 4.84 ± 3.03 mm and $8.86 \pm 12.42°$ accuracy, as a result from learning a better state value function by decoupling it from action-value function. These results surpass the state-of-the-art [Lu et al., 2011], which achieves an accuracy of 5.7 ± 8.5mm and $17.6 \pm 19.2°$. Unlike [Lu et al., 2011], the proposed method does not require manual annotation of landmarks.

Error	**Landmark-based** [Lu et al., 2011]	**RL Agents**			
		DQN	**DDQN**	**Duel DQN**	**Duel DDQN**
Distance (mm)	5.70 ± 8.50	5.61 ± 4.09	5.79 ± 4.58	$\mathbf{4.84 \pm 3.03}$	5.07 ± 3.33
Angle (°)	17.16 ± 19.20	10.16 ± 10.62	11.20 ± 14.86	8.86 ± 12.42	$\mathbf{8.72 \pm 7.44}$

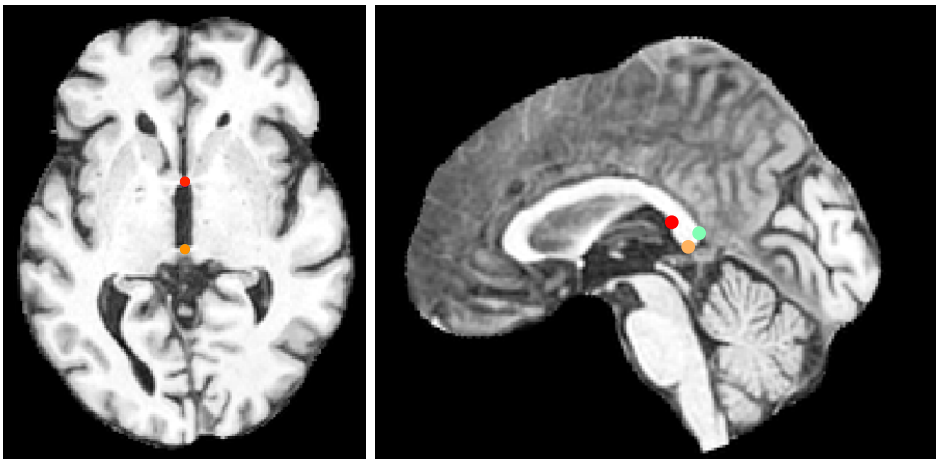
Table 5.5: Results of multi-scale RL agent detecting three different MRI planes.

[Experiment-II] ACPC axial and mid sagittal view planning in brain MRI

Datasets:

A set of 832 isotropic 1mm MR scans were obtained from the ADNI database [Mueller et al., 2005] to evaluate the proposed method. While, a subset of 728 and 104 images are used for

training and testing. All brain images were skull stripped and affinely aligned to the MNI space, thus allowing ground truth planes to be extracted in the standard directions. ACPC plane is defined in the axial direction containing AC and PC points. While the mid-sagittal plane is defined perpendicular to the ACPC plane in the sagittal direction and passing through Splenium of corpus callosum (outer aspect, inferior tip and inner aspect) points, see Figure 5.10. ACPC planes are evaluated using the AC and PC landmarks for the distance error calculation. Similarly, the outer aspect, inferior tip and inner aspect points of splenium of corpus callosum are used for measuring distance errors of the detected mid-sagittal planes.



(a) Axial ACPC plane (b) Mid-sagittal plane

Figure 5.10: Ground truth planes from brain and cardiac MRI scans. (a) ACPC axial plane marking AC (red) and PC (yellow) points. (b) Mid-sagittal plane with outer aspect (green), inferior tip (yellow) and inner aspect (red) points of splenium of corpus callosum.

Results

Table 5.6 shows the results from these comparative experiments. All methods share similar performance including speed and accuracy, and there is no unique winner for the best overall method. Best performing agents for detecting the mid-sagittal and ACPC planes achieve accuracy of 1.53 ± 2.2 mm and $2.44 \pm 5.04°$, and 1.98 ± 2.23 mm and $4.48 \pm 14.0°$, respectively. More visualization results are published on github.

Model	Mid-sagittal brain		ACPC brain	
	Distance (mm)	**Angle** ($^\circ$)	**Distance** (mm)	**Angle** ($^\circ$)
DQN	1.65 ± 1.99	2.42 ± 5.27	2.61 ± 5.44	**3.23±6.03**
DDQN	2.08 ± 2.58	3.44 ± 7.46	**1.98±2.23**	4.48 ± 14.00
Duel DQN	1.69 ± 1.98	3.82 ± 7.15	2.13 ± 1.99	5.24 ± 13.75
Duel DDQN	**1.53±2.20**	**2.44±5.04**	5.30 ± 11.19	5.25 ± 12.64

Table 5.6: Results of multi-scale RL agent detecting three different MRI planes.

Implementation

Training times are around 12-24 hours for the brain experiments and 2-4 days for the cardiac experiments using an NVIDIA GTX 1080Ti GPU. During inference, the agent finds the target plane using iterative steps, where each step takes ~0.02s. The details of the proposed network for DQN are in Figure 5.8. The source code of this implementation is publicly available on github https://git.io/vhuMZ.

5.4 Discussion and Conclusion

This chapter presented a novel framework based on training intelligent artificial agents for medical imaging tasks such as anatomical landmark detection and view planning. Different variants of DQN-based architectures were evaluated in Section 5.2.2 and 5.3.2. The first part of this chapter proposed different RL agents for automatic landmark detection in medical images. These RL-agents are capable to automatically find landmarks by moving towards the target sequentially step-by-step. Fixed- and multi-scale optimal path search strategies were also exploited. The extensive evaluations demonstrated high detection accuracy on three different datasets: foetal head ultrasound, adult brain and cardiac MRI. The results showed that multi-scale search strategies significantly improve the performance in images with large field of view and noisy background such as cardiac MRI. Hierarchical action steps speed up the searching process using larger steps, as well as smaller steps to fine-tune the final target location.

The second part of this chapter proposed a novel approach based on multi-scale reinforcement learning agents for automatic standard view extraction. The proposed approach is capable of finding standardized planes in fast time without the need of any manual annotated landmarks, which in turn enables accelerated image acquisition. Consequently, it can alleviate the comparison between different imaging examinations using anatomically standardized biometric measurements. The RL-agents were extensively evaluated using several DQN based strategies for the detection of three different planes. Results achieved state-of-the-art for the automatic detection of the ACPC and mid-sagittal planes from brain MRI with distance error less than 2 mm, and for the detection of the 4-chamber plane from cardiac MRI with distance error around 5 mm. Despite the fact that RL is a difficult problem that needs a careful formulation of its elements such as states, rewards and actions, this chapter has shown that the proposed RL-based framework was capable of detecting landmarks and planes automatically with high accuracy in a very short time.

Challenges

The results show that the optimal algorithm for achieving the best performance is environment-dependant. RL is a difficult problem that needs a careful formulation of its elements, where a bad design of the reward function can result in overfitting or underfitting to the reward signals causing unexpected behaviours. Therefore the design of the reward function has to capture exactly the desired task, and still be learnable.

Future work

There are open several open directions that can be explored in the future such as: (*i*) using a continuous action space to improve the performance through reduction of quantization errors introduced by fixed action steps; (*ii*) using of either competitive or collaborative multi-agents to detect the same or different anatomical landmarks or planes; (*iii*) using intrinsic geometry instead of intensity patterns for the RL environment to improve performance.

Chapter 6

Conclusions

The main goal of this thesis is to develop automatic methods for fast and accurate decision-making processes in the clinical practice. Medical imaging plays a crucial role for identifying a medical condition, therapeutic response, disease level or the chance of developing a future complication. This thesis proposed several scientific contributions to the main components of the medical imaging pipeline from acquisition, and through reconstruction, to analysis. The summary of these contributions and potential future directions are provided as follows.

6.1 Thesis Summary

6.1.1 Motion Correction and Reconstruction in Foetal MRI

The foetal is not sedated during an examination and the mother breaths normally, which leads to motion artefacts especially between slices in the out-of-plane direction. Traditional automatic foetal reconstruction methods rely on the assumption of the rigid motion. This limits these methods to organs that move rigidly such as the foetal brain. It also requires a mask around the organ of interest, which is usually a tedious task to be performed manually.Chapter 3 presented and evaluated a new paradigm for motion correction from foetal MRI. The proposed method is based on slice-to-volume registration (SVR) and flexible subdivision of the input space into over-

lapping, highly redundant and partly rigid image patches, thus solving the motion compensation problem for large field of view reconstructions. The patch-to-volume reconstruction (PVR) also compensates non-rigid motion artefacts from foetal MRI scans without requiring a pre-defined region of interest. Input stacks were first split into small overlapping areas, and followed by finding areas that contain rigid components. This allowed to iteratively learn their consistency compared to a global reconstruction optimization volume. Corrupted and inconsistent data was automatically identified and excluded using robust statistics. PVRs reconstruction quality and computational performance were evaluated quantitatively and qualitatively on an adult phantom T2-weighted brain with synthetic non-rigid motion artefacts, as well as on the whole uterus from motion corrupted foetal MRI data including foetal brain, placenta and cases with multiple births. These extensive experiments showed good reconstruction quality of the whole uterus, selected foetal organs, and secondary, non-rigidly moving pregnancy structures (*e.g.* placenta).

6.1.2 Foetal Brain Localisation

Foetal brain localisation is important for assessing the foetal brain development and maturation. It is also a primary step for the majority of automatic motion correction and reconstruction methods. Manual localisation of the foetal brain is usually a laborious and time consuming task. Chapter 4 introduced a novel fully-automated framework for localising the foetal brain in MRI scans. Superpixels are capable of reducing the effect of motion artefacts between neighbouring pixels, as they typically represent rigid regions in the image. In addition, superpixels enable faster and more efficient performance compared to using individual pixels by significantly reducing number of data samples. The proposed localisation framework is based on training a random forest classifier using superpixel graphs. Including features from the neighbouring superpixels using weighted graphs provides larger contextual information compared to local features of individual superpixels. The graphs were modelled using both spatial and intensity distances in 3D. The proposed localisation framework achieved 94.55% accuracy for the brain detection and 98.18% prediction accuracy for the centre of the brain, without the need of any

landmarks or prior information such as the gestational age of the foetus or the volumetric size of the brain.

6.1.3 Automatic Placenta Segmentation

Quantitative measurements such as placental volume and surface attachment to the uterine wall, are required for identifying abnormalities during pregnancy. In addition, recording the structural appearance (*e.g.* placental cotyledons and shape) is essential for clinical qualitative analysis. The automatic segmentation of the placenta is very challenging due to the high variability of its position, orientation, thickness, shape and appearance. Chapter 4 proposed the first fully automatic segmentation framework for the placenta from motion corrupted foetal MRI. The proposed framework deploys a 3D multi-scale convolutional neural network (CNN) as a strong classifier for image segmentation followed by a conditional random field (CRF) for refinement. The multi-scale architecture enables the network to capture larger 3D contextual information, which is essential for detecting an organ that appears in different locations and sizes. The CRF is then applied to regularise inconsistencies of the segmentation using priors with the relational consistency of their neighbours. Rigorous experiments have been performed on two different testing datasets in order to evaluate thoroughly the presented segmentation approach. The results showed that this framework could achieve segmentation accuracy of 71.95% for healthy foetuses despite the motion artefacts. It was also capable of segmenting the placenta from dissimilar data by achieving segmentation accuracy of 66.89% for a cohort mixed with cases of intrauterine foetal growth restriction from different scanners. The results showed also the ability of this approach to scale to real clinical applications by comparing the volumetric size of the segmented placenta in clinical practice. Moreover, the proposed framework scope was extended to real clinical applications by compensating motion artefacts using slice to volume registration techniques, as well as providing a novel standardized view into the placental structures using skeleton extraction. This was done by applying the placental mask for the initialization of PVR techniques to compensate for motion artefacts. Whereas the resulting reconstructed placental volume can be used to provide a novel standardized view into

the placental structures by applying shape skeleton extraction and curved planar reformation for shape abstraction.

6.1.4 Anatomical Landmark Detection

Accurate detection of anatomical landmarks from medical images is an essential step for many image analysis and interpretation methods. Since manual landmark annotation is time consuming and error prone, automatic methods were developed to tackle this problem. The design of such methods is challenging due to varying organ morphology, orientation, pathology, and image quality. Chapter 5 presented and evaluated fixed-scale and multi-scale RL based agents for the landmark detection task. The proposed RL framework formulated the landmark detection problem as a sequential decision making process of a goal-oriented agent, navigating in an environment (the acquired image) towards a target landmark. At each time step, the agent decides which direction it has to proceed towards an optimal path to the target landmark. RL is used to learn an approximation of the optimal solution of this sequential decision making process. Extensive experiments were performed to evaluate the performance of different DQN based models for anatomical landmark localisation using both fixed-scale and multi-scale search strategies with novel hierarchical action steps by running multiple experiments on different MRI and US images. Hierarchical action steps speed up the searching process using large steps, as well as small steps to fine-tune the final target location. The extensive evaluations demonstrated state-of-the-art detection accuracy on three different datasets: foetal head ultrasound, adult brain and cardiac MRI. The agent was capable of finding the target landmark in less than 1 second starting from any random initialization. The results also showed that multi-scale search strategies significantly improve the performance in images with large field of view and noisy background such as cardiac MRI.

6.1.5 Automatic View Planning

Standardised views through a defined anatomy are commonly used in clinical practice to establish comparable metrics. Manual view planing requires expertise in the scanned organ anatomy and geometry. Besides, it is subject to operator bias, and the process is time-consuming. Chapter 5 presented a new method for automatic view planning in volumetric MRI images using novel multi-scale reinforcement learning (RL) agents. The artificial agent learns to make comprehensive and sensible decisions by mimicking navigation processes, in a manner that allows medical experts to gain confidence in fully automatic methods. RL constitutes a sub-field of machine learning concerned with how agents take actions in an environment. In contrast to supervised learning, RL involves learning by interacting with an environment instead of using a set of labelled examples that is typically provided by a knowledgeable supervisor. This learning paradigm allowed these agents to learn the view planning task iteratively by sampling a new plane from the 3D volume at every step towards the target plane. This was done by giving positive rewards when the agent moves towards the target plane, and negative rewards otherwise. Eventually, the agent learns the optimal trajectory to the target plane starting from any location in the image. The proposed framework follows a multi-scale search strategy with hierarchical action steps in a coarse-to-fine fashion. During testing, the artificial agents were capable of finding standardized planes in fast time without the need of any manual annotated landmarks, which in turn could enable *accelerated image acquisition*. Extensive experiments were applied for evaluating different deep Q-networks (DQN) baselines on detecting three different planes from brain and cardiac MRI. The proposed algorithm is assessed on the mid-sagittal and anterior-posterior commissure planes of brain MRI, and the 4-chamber long-axis plane commonly used in cardiac MRI, achieving state-of-the-art accuracy of 1.53mm, 1.98mm and 4.84mm, respectively.

6.2 Challenges and Future Directions

6.2.1 Deep Learning for Image Reconstruction

Recent advances in deep learning methods have proved to outperform state-of-the-art conventional existing techniques especially for tasks such as image localisation and segmentation, as shown in chapters 4 and 5. However, it is still challenging to apply these methods for tasks with small or no ground truth labels. For instance, the problem of correcting motion artefacts in foetal MRI, as introduced in Chapter 3, has no ground truth high-resolution motion-free image. McDonagh et al. [2017] proposed to tackle this by learning the up-sampling step in the reconstruction pipeline using a 3D CNN. Training samples were generated by naïvely down-sampling the input images, and using the high-resolution images as ground truth targets. Although, the results showed improvement in the up-sampled images, it was limited only to input 2D slices with no motion artefacts. Another major challenge is the evaluation of the reconstructed output. Unlike in computer vision, every pixel or point in medical images represent a tissue or organ. Deep CNNs are still used as a black box approach, which can produce unexpected results with no explanation or interpretation. This may result in hallucinating output pixels in the reconstructed images that are not clinically or anatomically correct. Without a ground truth and proper metrics, it is even more difficult to evaluate the output. Given the potential power of deep learning for different imaging tasks, a future direction will be to explore the possibility to apply new advances to the reconstruction problem.

6.2.2 Learning from Small Data

The current deep learning methods rely on massive training data for boosting the performance accuracy. Chapter 4 showed a drop in the segmentation accuracy, when the same trained network applied to different test data. In computer vision, researchers make use of tools like Amazon Mechanical Turk [1] to harvest big training data. Performing repetitive tasks such as labelling objects in images does not require experienced skills, thus it is cheap and efficient.

[1] https://www.mturk.com/

While, in medical imaging, annotating one image can take several hours and requires significantly greater amounts of expertise. Hence, preparing the training data for such methods can be more crucial for achieving high accuracy compared to building the network architecture. This lead to a growing interest in the machine learning research in exploring novel training strategies with small or no labelled images (*e.g.* unsupervised, semi-supervised, one-shot, and transfer learning). The current research for solving this task are based on three main approaches: (*i*) applying Bayesian models that involve prior knowledge; (*ii*) decreasing the complexity of the model to avoid overfitting; and (*iii*) using topological or geometrical information. Moreover, if the training data does not contain rare or edge cases such as pathologies, the trained models will most likely fail to capture what they have not seen during training. The RL method presented in Chapter 5 implicitly handles the sparse sampling problem using interactive instead of supervised learning. This was done by defining different actions to navigate towards the target landmark. It will be interesting to investigate the use of different learning paradigms to tackle the learning task from small medical datasets in a future study.

6.2.3 Interpretability and Uncertainty

One of the main challenges of the current state-of-the-art methods, despite their good performance, is their high complexity in terms of model architecture and number of parameters. This makes interpreting the results extremely difficult and computationally impracticable. In medical applications, it is essential to provide an explanation associated with any decision making process. This is another hot topic of research that has already attracted the research community for the last couple of years. A very recent work [Arslan et al., 2018] has explored the use of graph convolutional networks (GCN) to identify regions of the brain that corresponds to sex classification and has provided visual attribution method based on class activation mapping. They demonstrated the robustness of the performance of their method in highlighting reproducible regions across individuals. However, their method is restricted to finite input dimensions from brain connectivity networks. Another direction for understanding the behaviour of different model structures, is the use of Bayesian models to provide a measure of uncertainty

of the data in an interpretable way [Gal, 2016]. This can open an avenue of future research in medical images.

6.2.4 Active Learning

In active learning, the algorithm interacts with the user to obtain the desired output from new data. For instance, RL agents presented in Chapter 5 learns by interaction with an environment. This work can be extended to include user interactions to boost the performance of the artificial agents, especially on new environments. This offers a novel future direction, similar to AlphaGo [Silver et al., 2016], where an RL agent could mimic and improve upon the moves of a human expert by accumulating the learning experiences, to learn from experienced operators during real time observation (learning by demonstrations).

Bibliography

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

Amir Alansary, Matthew Lee, Kevin Keraudren, Bernhard Kainz, Christina Malamateniou, Mary Rutherford, Joseph V Hajnal, Ben Glocker, and Daniel Rueckert. Automatic brain localization in fetal MRI using superpixel graphs. In *Medical Learning Meets Medical Imaging*, pages 13–22. Springer, 2015.

Amir Alansary, Konstantinos Kamnitsas, Alice Davidson, Rostislav Khlebnikov, Martin Rajchl, Christina Malamateniou, Mary Rutherford, Joseph V Hajnal, Ben Glocker, Daniel Rueckert, et al. Fast fully automatic segmentation of the human placenta from motion corrupted MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–597. Springer, 2016.

Amir Alansary, Martin Rajchl, Steven G McDonagh, Maria Murgasova, Mellisa Damodaram, David FA Lloyd, Alice Davidson, Mary Rutherford, Joseph V Hajnal, Daniel Rueckert, et al. PVR: patch-to-volume reconstruction for large area motion correction of fetal MRI. *IEEE transactions on medical imaging*, 36(10):2031–2044, 2017.

Amir Alansary, Loic Le Folgoc, Ghislain Vaillant, Ozan Oktay, Yuanwei Li, Wenjia Bai, Jonathan Passerat-Palmbach, Ricardo Guerrero, Konstantinos Kamnitsas, Benjamin Hou, Steven McDonagh, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Automatic View Planning with Multi-scale Deep Reinforcement Learning Agents. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018.

Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Kamnitsas Konstantinos, Vlontzos Athanasios, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection. 2019.

Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

Simon Andermatt, Simon Pezold, Michael Amann, and Philippe C Cattin. Multi-dimensional Gated Recurrent Units for Automated Anatomical Landmark Localization. *arXiv preprint arXiv:1708.02766*, 2017.

J. Anquez, E. D. Angelini, and I. Bloch. Automatic segmentation of head structures on fetal MRI. In *in Proc. IEEE International Symposium on Biomedical Imaging*, pages 109–112. IEEE Press, 2009. ISBN 978-1-4244-3931-7.

Babak A Ardekani, Jeff Kershaw, Michael Braun, and I Kanuo. Automatic detection of the mid-sagittal plane in 3-D brain images. *TMI*, 16(6):947–952, 1997.

Vincent Arsigny, Olivier Commowick, Nicholas Ayache, and Xavier Pennec. A fast and log-euclidean polyaffine framework for locally linear registration. *J. Math. Imaging Vision*, 33 (2):222–238, 2009.

Salim Arslan, Sofia Ira Ktena, Ben Glocker, and Daniel Rueckert. Graph Saliency Maps through Spectral Convolutional Networks: Application to Sex Classification with Brain Connectivity. *arXiv preprint arXiv:1806.01764*, 2018.

Wenjia Bai, Wenzhe Shi, Declan P O'Regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S Peters, and Daniel Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. *TMI*, 32(7):1302–1315, 2013.

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics, 2001.

C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

Richard Bellman. *Dynamic programming*. Courier Corporation, 2013.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Margrit Betke, Harrison Hong, Deborah Thomas, Chekema Prince, and Jane P Ko. Landmark detection in the chest and registration of lung surfaces with an application to nodule registration. *MedIA*, 7(3):265–281, 2003.

Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient nd image segmentation. *International journal of computer vision*, 70(2):109–131, 2006.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

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

Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2488–2496. IEEE, 2015.

Mieke M Cannie, Frederik De Keyzer, Sigrid Van Laere, Astrid Leus, Johan de Mey, Catherine Fourneau, Filip De Ridder, Toon Van Cauteren, Inneke Willekens, and Jacques C Jani. Potential heating effect in the gravid uterus by using 3-T MR imaging protocols: experimental study in miniature pigs. *Radiology*, 279(3):754–761, 2015.

P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.*, 6(2):298–311, Feb 1997. ISSN 1057-7149. doi: 10.1109/83.551699.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International*

Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 424–432. Springer, 2016.

Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

Chris A. Cocosco, Vasken Kollokian, Remi K.-S. Kwan, G. Bruce Pike, and Alan C. Evans. BrainWeb: Online Interface to a 3D MRI Simulated Brain Database. *NeuroImage*, 5:425, 1997.

Olivier Commowick, Vincent Arsigny, Aurélie Isambert, Jimena Costa, Frédéric Dhermain, François Bidault, P-Y Bondiau, Nicholas Ayache, and Grégoire Malandain. An efficient locally affine framework for the smooth registration of anatomical structures. *Medical Image Analysis*, 12(4):427–441, 2008.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer-Verlag New York Incorporated, 2013.

Antonio Criminisi, Jamie Shotton, and Stefano Bucciarelli. Decision forests with long-range spatial context for organ localization in CT volumes. *Medical Image Computing and Computer-Assisted Interventation (MICCAI)*, pages 69–80, 2009.

Antonio Criminisi, Jamie Shotton, Duncan Robertson, and Ender Konukoglu. Regression forests for efficient anatomy detection and localization in CT studies. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 106–117. Springer, 2011.

Antonio Criminisi, Duncan Robertson, Ender Konukoglu, Jamie Shotton, Sayan Pathak, Steve White, and Khan Siddiqpotesil2015personalizedui. Regression forests for efficient anatomy detection and localization in computed tomography scans. *MedIA*, 17(8):1293–1303, 2013.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8609–8613. IEEE, 2013.

Antonio de Marvao, Timothy JW Dawes, Wenzhe Shi, Christopher Minas, Niall G Keenan, Tamara Diamond, Giuliana Durighel, et al. Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power. *Journal of Cardiovascular Magnetic Resonance*, 16(1):16, 2014.

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

René Donner, Erich Birngruber, Helmut Steiner, Horst Bischof, and Georg Langs. Localization of 3d anatomical structures using random forests and discrete optimization. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 86–95. Springer, 2011.

Birgit Ertl-Wagner, Andreas Lienemann, Alexander Strauss, and Maximilian F Reiser. Fetal magnetic resonance imaging: indications, technique, anatomical considerations and a review of fetal abnormalities. *European Radiology*, 12(8):1931–1940, 2002.

Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

Baowei Fei, Jeffrey L Duerk, Daniel T Boll, Jonathan S Lewin, and David L Wilson. Slice-to-volume registration and its potential application to interventional MRI-guided radio-frequency thermal ablation of prostate cancer. *Medical Imaging, IEEE Transactions on*, 22(4):515–525, 2003.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *Int J Comput Vision*, 59(2):167–181, 2004.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Mary C Frates, Ada J Kumar, Carol B Benson, Valerie L Ward, and Clare M Tempany. Fetal Anomalies: Comparison of MR Imaging and US for Diagnosis 1. *Radiology*, 232(2):398–404, 2004.

William T. Freeman and Edward H Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9):891–906, 1991.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

Masatoshi Fukushima. *Dirichlet forms and Markov processes*, volume 23. Elsevier, 1980.

Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 670–677. IEEE, 2009.

Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

Catherine Garel. Imaging the fetus: when does MRI really help? *Pediatric radiology*, 38: 467–470, 2008.

Romane Gauriau, Rémi Cuingnet, David Lesage, and Isabelle Bloch. Multi-organ localization with cascaded global-to-local regression and shape prior. *MedIA*, 23(1):70–83, 2015.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Florin C Ghesu, Bogdan Georgescu, Tommaso Mansi, Dominik Neumann, Joachim Hornegger, and Dorin Comaniciu. An artificial agent for anatomical landmark detection in medical images. In *MICCAI*, pages 229–237. Springer, 2016.

Florin C Ghesu, Bogdan Georgescu, Sasa Grbic, Andreas K Maier, Joachim Hornegger, and Dorin Comaniciu. Robust Multi-scale Anatomical Landmark Detection in Incomplete 3D-CT Data. In *MICCAI*, pages 194–202. Springer, 2017a.

Florin Cristian Ghesu, Bogdan Georgescu, Yefeng Zheng, Sasa Grbic, Andreas Maier, Joachim Hornegger, and Dorin Comaniciu. Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans. *PAMI*, 2017b.

Ali Gholipour, Judy A Estroff, and Simon K Warfield. Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain MRI. *IEEE Trans. Med. Imag.*, 29(10):1739–1758, 2010.

N Girard, C Raybaud, C Dercole, L Boubli, C Chau, S Cahen, A Potier, and M Gamerre. In vivo MRI of the fetal brain. *Neuroradiology*, 35(6):431–436, 1993.

O.A. Glenn and A.J. Barkovich. Magnetic Resonance Imaging of the Fetal Brain and Spine: An Increasingly Important Tool in Prenatal Diagnosis, Part 1 & 2. *AJNR*, 27(8):1604–1611, 2006a.

OA Glenn and AJ Barkovich. Magnetic resonance imaging of the fetal brain and spine: an increasingly important tool in prenatal diagnosis, part 1. *AJNR*, 27(8):1604–1611, 2006b.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Comput. Surv.*, 23(1):5–48, March 1991. ISSN 0360-0300. doi: 10.1145/103162.103163.

Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and Ronald L Tatham. *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

Hado V Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

DG Hebbs. The organization of behavior. *Wiely and Sons, New York, NY, USA*, 1949.

Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

Tin Kam Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE, 1995.

Benjamin Hou, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Predicting slice-to-volume transformation in presence of arbitrary subject motion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 296–304. Springer, 2017.

Benjamin Hou, Bishesh Khanal, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. 3-D Reconstruction in Canonical Co-ordinate Space from Arbitrarily Oriented 2D Images. *IEEE Transactions on Medical Imaging*, 37:1737–1750, 2018a.

Benjamin Hou, Nina Miolane, Bishesh Khanal, Matthew CH Lee, Amir Alansary, Steven McDonagh, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Computing CNN Loss and Gradients for Pose Estimation with Riemannian Geometry. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018b.

Bonnie J. Huppert, Kathy R. Brandt, Kirk D. Ramin, and Bernard F. King. Single-shot fast spin-echo MR imaging of the fetus: A pictorial essay. *Radiographics*, 19(SPEC.ISS.), 10 1999. ISSN 0271-5333.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

Mark Ison, R Donner, E Dittrich, Gregor Kasprian, Daniela Prayer, and Georg Langs. Fully automated brain extraction and orientation in raw fetal MRI. In *Proc. Workshop on Paediatric and Perinatal Imaging, MICCAI*, volume 12, pages 17–24, 2012.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

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

Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

B. Kainz, M. Steinberger, W. Wein, M. Murgasova, C. Malamateniou, K. Keraudren, Th. Torsney-Weir, K., P. Aljabar, M. Rutherford, J. Hajnal, and D. Rueckert. Fast Volume Reconstruction from Motion Corrupted Stacks of 2D Slices. *IEEE Trans. Med. Imag.*, 34(9): 1901–13, 2015a. ISSN 0278-0062. doi: 10.1109/TMI.2015.2415453.

Bernhard Kainz, Kevin Keraudren, Vanessa Kyriakopoulou, Mary Rutherford, Joseph V Hajnal, and Daniel Rueckert. Fast fully automatic brain detection in fetal MRI using dense rotation invariant image descriptors. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 1230–1233. IEEE, 2014a.

Bernhard Kainz, Christina Malamateniou, Maria Murgasova, Kevin Keraudren, Mary Rutherford, Joseph V Hajnal, and Daniel Rueckert. Motion Corrected 3D Reconstruction of the Fetal Thorax from Prenatal MRI. In *MICCAI*, pages 284–291, 2014b.

Bernhard Kainz, Amir Alansary, Christina Malamateniou, Kevin Keraudren, Mary Rutherford,

Joseph V Hajnal, and Daniel Rueckert. Flexible reconstruction and correction of unpredictable motion from stacks of 2D images. In *MICCAI*, pages 555–562. Springer, 2015b.

Bernhard Kainz, D Lloyd, Amir Alansary, M Kuklisova Murgasova, Rostislav Khlebnikov, Daniel Rueckert, Mary Rutherford, Reza Razavi, and J Hajnal. High-performance motion correction of fetal MRI. 2016.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

A. Kanitsar, D. Fleischmann, R. Wegenkittl, P. Felkel, and E. Groller. CPR – curved planar reformation. In *IEEE VIS 2002*, pages 37–44, Nov 2002. doi: 10.1109/VISUAL.2002.1183754.

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

Kevin Keraudren, Vanessa Kyriakopoulou, Mary Rutherford, Joseph V Hajnal, and Daniel Rueckert. Localisation of the brain in fetal MRI using bundled SIFT features. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 582–589. Springer, 2013.

Kevin Keraudren, Bernhard Kainz, Ozan Oktay, V Kyriakopoulou, MA Rutherford, Joseph V Hajnal, and Daniel Rueckert. Automated Localization of Fetal Organs in MRI Using Random Forests with Steerable Features. In *Medical Image Computing and Computer-Assisted Interventation–MICCAI*, page In Press. 2015.

Kio Kim, P.A Habas, F. Rousseau, O.A Glenn, AJ. Barkovich, and C. Studholme. Intersection Based Motion Correction of Multislice MRI for 3-D in Utero Fetal Brain Image Formation. *IEEE Trans. Med. Imag.*, 29(1):146–158, Jan 2010. ISSN 0278-0062. doi: 10.1109/TMI.2009.2030679.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2012.

Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *MICCAI*, pages 344–352. Springer, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel. Reconstruction of Fetal Brain MRI with Intensity Matching and Complete Outlier Removal. *Medical Image Analysis*, 16(8):1550–60, 2012.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Matthieu Le, Jesse Lieman-Sifry, Felix Lau, Sean Sall, Albert Hsiao, and Daniel Golden. Computationally efficient cardiac views projection using 3D Convolutional Neural Networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 109–116. Springer, 2017.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1): 29–44, 2001.

D. Levine. Fetal Magnetic Resonance Imaging. *J Matern Fetal Neonatal Med*, 15(2):85–94, February 2004. ISSN 1476-7058.

D Levine, P D Barnes, S Sher, R C Semelka, W Li, C R McArdle, S Worawattanakul, and R R Edelman. Fetal fast MR imaging: reproducibility, technical quality, and conspicuity of anatomy. *Radiology*, 206(2):549–554, 1998. doi: 10.1148/radiology.206.2.9457211.

John P Lewis. Fast template matching. In *Vision interface*, volume 95, pages 15–19, 1995.

Yingying Li, Stanley Osher, et al. A new median formula with applications to PDE based denoising. *Commun. Math. Sci*, 7(3):741–753, 2009.

Yuanwei Li, Amir Alansary, Juan Cerrolaza, Bishesh Khanal, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Fast Multiple Landmark Localisation Using a Patch-based Iterative Network. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018a.

Yuanwei Li, Bishesh Khanal, Benjamin Hou, Amir Alansary, Juan J Cerrolaza, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, et al. Standard Plane Detection in 3D Fetal Ultrasound Using an Iterative Transformation Network. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018b.

Rui Liao, Shun Miao, Pierre de Tournemire, Sasa Grbic, Ali Kamen, Tommaso Mansi, and Dorin Comaniciu. An Artificial Agent for Robust Image Registration. In *AAAI*, pages 4168–4175, 2017.

Long-Ji Lin. Reinforcement learning for robots using neural networks. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.

N Linduska, S Dekan, A Messerschmidt, G Kasprian, PC Brugger, K Chalubinski, M Weber, and D Prayer. Placental pathologies in fetal MRI with pathohistological correlation. *Placenta*, 30(6):555–559, 2009.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

John A Little, Derek LG Hill, and David J Hawkes. Deformations incorporating rigid structures. In *Mathematical Methods in Biomedical Image Analysis, Proceedings of the Workshop on*, pages 104–113. IEEE, 1996.

DFA Lloyd, B KAINZ, JFP VAN AMEROM, K PUSHPARAJAH, JM SIMPSON, V ZIDERE, O MILLER, G SHAR-LAND, T ZHANG, M LOHEZIC, et al. Three-dimensional modelling of the fetal vasculature from prenatal MRI using motion corrected slice-to-volume registration. In *Proc. Int. Soc. Magn. Reson. Med.*, volume 24, page O413, 2016.

S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

Xiaoguang Lu, Marie-Pierre Jolly, Bogdan Georgescu, Carmel Hayes, Peter Speier, Michaela Schmidt, Xiaoming Bi, Randall Kroeker, Dorin Comaniciu, Peter Kellman, et al. Automatic view planning for cardiac MRI acquisition. In *MICCAI*, pages 479–486. Springer, 2011.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Gabriel Maicas, Gustavo Carneiro, Andrew P Bradley, Jacinto C Nascimento, and Ian Reid. Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI. In *MICCAI*, pages 665–673. Springer, 2017.

C. Malamateniou, S.J. Malik, S.J. Counsell, J.M. Allsop, A.K. McGuinness, T. Hayat, K. Broadhouse, R.G. Nunes, A.M. Ederies, J.V. Hajnal, and M.A. Rutherford. Motion-

Compensation Techniques in Neonatal and Fetal MR Imaging. *American Journal of Neuroradiology*, 34(6):1124–1136, 2013. doi: 10.3174/ajnr.A3128.

P Mansfield and A.A Maudsley. Planar spin imaging by NMR. *Journal of Magnetic Resonance (1969)*, 27(1):101 – 119, 1977. ISSN 0022-2364. doi: http://dx.doi.org/10.1016/0022-2364(77)90197-4.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Steven McDonagh, Benjamin Hou, Amir Alansary, Ozan Oktay, Konstantinos Kamnitsas, Mary Rutherford, Jo V Hajnal, and Bernhard Kainz. Context-sensitive super-resolution for fast fetal magnetic resonance imaging. In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pages 116–126. Springer, 2017.

Haichao Miao, Gabriel Mistelbauer, Alexey Karimov, Amir Alansary, Alice Davidson, David FA. Lloyd, Mellisa Damodaram, Lisa Story, Jana Hutter, Joseph V. Hajnal, Mary Rutherford, Bernhard Preim, Bernhard Kainz, and M. Eduard Gröller. Placenta Maps: In Utero Placental Health Assessment of the Human Fetus. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2017.

Stephen Milborrow and Fred Nicolls. Active shape models with SIFT descriptors and MARS. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, pages 380–387. IEEE, 2014.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable Parallel Programming with CUDA. *Queue*, 6(2):40–53, March 2008. ISSN 1542-7730.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

Ozan Oktay, Wenjia Bai, Ricardo Guerrero, Martin Rajchl, Antonio de Marvao, Declan P ORegan, Stuart A Cook, Mattias P Heinrich, Ben Glocker, and Daniel Rueckert. Stratified decision forests for accurate anatomical landmark localization in cardiac images. *TMI*, 36 (1):332–342, 2017.

Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE, 1998.

Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine, IEEE*, 20(3):21–36, 2003.

Y Patenaude, D Pugash, K Lim, L Morin, S Bly, K Butt, Y Cargill, G Davies, N Denis, G Hazlitt, et al. The use of magnetic resonance imaging in the obstetric patient. *Journal of obstetrics and gynaecology Canada: JOGC= Journal d'obstetrique et gynecologie du Canada: JOGC*, 36(4):349–363, 2014.

Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *MICCAI*, pages 230–238. Springer, 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Graeme P Penney, Jürgen Weese, John A Little, Paul Desmedt, Derek LG Hill, et al. A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE Trans. Med. Imag.*, 17(4):586–595, 1998.

Agostino Pierro, Edward M Kiely, and Lewis Spitz. Classification and clinical evaluation. In *Seminars in pediatric surgery*, volume 24, pages 207–211. Elsevier, 2015.

Vaclav Potesil, Timor Kadir, Günther Platsch, and Mike Brady. Improved Anatomical Landmark Localization in Medical Images Using Dense Matching of Graphical Models. In *BMVC*, volume 4, page 9, 2010.

Vaclav Potesil, Timor Kadir, Günther Platsch, and Michael Brady. Personalized graphical models for anatomical landmark localization in whole-body medical images. *International Journal of Computer Vision*, 111(1):29–49, 2015.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Bahbibi Rahmatullah, Aris T Papageorghiou, and J Alison Noble. Image analysis using machine learning: Anatomical landmarks detection in fetal ultrasound images. In *Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual*, pages 354–355. IEEE, 2012.

Martin Rajchl, Matthew CH Lee, Franklin Schrans, Alice Davidson, Jonathan Passerat-Palmbach, Giacomo Tarroni, Amir Alansary, Ozan Oktay, Bernhard Kainz, and Daniel Rueckert. Learning under distributed weak supervision. *arXiv preprint arXiv:1606.01100*, 2016.

Martin Riedmiller. Reinforcement learning without an explicit terminal state. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 3. IEEE, 1998.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Francois Rousseau, Orit A. Glenn, Bistra Iordanova, Claudia Rodriguez-Carranza, Daniel B. Vigneron, James A. Barkovich, and Colin Studholme. Registration-Based Approach for Reconstruction of High-Resolution In Utero Fetal MR Brain Images. *Academic Radiol.*, 13 (9):1072 – 1081, 2006. ISSN 1076-6332. doi: http://dx.doi.org/10.1016/j.acra.2006.05.003.

François Rousseau, Kio Kim, Colin Studholme, Meriam Koob, and J-L Dietemann. On super-resolution for fetal brain MRI. In *MICCAI*, pages 355–362. Springer, 2010.

François Rousseau, Estanislao Oubel, Julien Pontabry, Marc Schweitzer, Colin Studholme, Mériam Koob, and Jean-Louis Dietemann. BTK: An open-source toolkit for fetal brain MR image processing. *Computer methods and programs in biomedicine*, 109(1):65–73, 2013.

Emma Routledge, Christina Malamateniou, and Mary Rutherford. Can MR imaging of the placenta demonstrate a distinct placental phenotype in normal and abnormal pregnancies? Technical report, King's College London, University of London, April 2015.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imag.*, 18(8):712–721, 1999.

Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. *TMI*, 22(8):1014–1025, 2003.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Farhang Sahba, Hamid R Tizhoosh, and Magdy MA Salama. A reinforcement learning framework for medical image segmentation. In *IJCNN*, pages 511–517. IEEE, 2006.

C.M. Salafia, D.P. Misra, M. Yampolsky, A.K. Charles, and R.K. Miller. Allometric metabolic scaling and fetal and placental weight. *Placenta*, 30(4):355 – 360, 2009. ISSN 0143-4004.

Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

Jason Sanders and Edward Kandrot. *CUDA by Example: An Introduction to General-Purpose GPU Programming*. Addison-Wesley, 1st edition, 2010. ISBN 0131387685, 9780131387683.

Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3):2255–2265, 2012.

Maryam Shokri and Hamid R Tizhoosh. Using reinforcement learning for image thresholding. In *Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on*, volume 2, pages 1231–1234. IEEE, 2003.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304. Ieee, 2011.

Robin Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Henrik Skibbe, Marco Reisert, Thorsten Schmidt, Thomas Brox, Olaf Ronneberger, and Hans Burkhardt. Fast rotation invariant 3D feature computation utilizing efficient local neighbor-

hood operators. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8): 1563–1575, 2012.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, 1986.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2627435.2670313`.

Mikkel B Stegmann, Karl Skoglund, and Charlotte Ryberg. Mid-sagittal plane and mid-sagittal surface optimization in brain MRI using a local symmetry measure. In *Medical Imaging: Image Processing*, volume 5747, pages 568–580, 2005.

Markus Steinberger, Bernhard Kainz, Bernhard Kerbl, Stefan Hauswiesner, Michael Kenzel, and Dieter Schmalstieg. Softshell: Dynamic Scheduling on GPUs. *ACM Trans. Graph.*, 31 (6):161:1–161:11, November 2012. ISSN 0730-0301.

Gordon N Stevenson, Sally L Collins, Jane Ding, Lawrence Impey, and J Alison Noble. 3-D Ultrasound Segmentation of the Placenta Using the Random Walker Algorithm: Reliability and Agreement. *Ultrasound in medicine & biology*, 41(12):3182–3193, 2015.

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998a.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998b.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Andrea Tagliasacchi, Ibraheem Alhashim, Matt Olson, and Hao Zhang. Mean Curvature Skeletons. *Computer Graphics Forum*, 31(5):1735–1744, 2012. ISSN 1467-8659.

Vahid Taimouri, Ali Gholipour, Clemente Velasco-Annis, Judy A Estroff, and Simon K Warfield. A TEMPLATE-TO-SLICE BLOCK MATCHING APPROACH FOR AUTOMATIC LOCALIZATION OF BRAIN IN FETAL MRI. 2015.

Youssef Taleb, Marc Schweitzer, Colin Studholme, Meriam Koob, Jean-Louis Dietemann, and François Rousseau. Automatic Template-based Brain Extraction in Fetal MR Images. 2013.

M Thame, C Osmond, F Bennett, R Wilks, and T Forrester. Fetal growth is directly related to maternal anthropometry and placental volume. *European journal of clinical nutrition*, 58 (6):894–900, 2004.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

AN Tikhonov. On the solution of incorrectly posed problem and the method of regularization. *Soviet Math*, 4:1035–1038, 1963.

Sébastien Tourbier, Xavier Bresson, Patric Hagmann, Jean-Philippe Thiran, Reto Meuli, and Meritxell Bach Cuadra. An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. *NeuroImage*, 118:584–597, 2015.

Sébastien Tourbier, Clemente Velasco-Annis, Vahid Taimouri, Patric Hagmann, Reto Meuli, Simon K Warfield, Meritxell Bach Cuadra, and Ali Gholipour. Automated template-based brain localization and extraction for fetal brain MRI reconstruction. *NeuroImage*, 2017.

Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1744–1757, 2010.

Martin Urschler, Thomas Ebner, and Darko Štern. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. *MedIA*, 43:23–36, 2018.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep Reinforcement Learning with Double Q-Learning. In *AAAI*, volume 16, pages 2094–2100, 2016.

Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010.

Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *ECCV*, pages 705–718. Springer, 2008.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

Guotai Wang, Maria A Zuluaga, Rosalind Pratt, Michael Aertsen, Anna L David, Jan Deprest, Tom Vercauteren, and Sebastien Ourselin. Slic-Seg: slice-by-slice segmentation propagation of the placenta in fetal MRI using one-plane scribbles and online learning. In *MICCAI*, pages 29–37. Springer, 2015a.

Guotai Wang, Maria A Zuluaga, Rosalind Pratt, Michael Aertsen, Tom Doel, Maria Klusmann, Anna L David, Jan Deprest, Tom Vercauteren, and Sébastien Ourselin. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Medical image analysis*, 34:137–147, 2016.

Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Transactions on Medical Imaging*, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando

De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015b.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Zhoubing Xu, Qiangui Huang, JinHyeong Park, Mingqing Chen, Daguang Xu, Dong Yang, David Liu, and S Kevin Zhou. Supervised Action Classifier: Approaching Landmark Detection as Image Partitioning. In *MICCAI*, pages 338–346. Springer, 2017.

Irving J Zamora, Fariha Sheikh, Christopher I Cassady, Oluyinka O Olutoye, Amy R Mehollin-Ray, Rodrigo Ruano, Timothy C Lee, Stephen E Welty, Michael A Belfort, Cecilia G Ethun, et al. Fetal MRI lung volumes are predictive of perinatal outcomes in fetuses with congenital lung masses. *Journal of pediatric surgery*, 49(6):853–858, 2014.

Jun Zhang, Mingxia Liu, and Dinggang Shen. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *TIP*, 26(10):4753–4764, 2017.

Yefeng Zheng, Adrian Barbu, Bogdan Georgescu, Michael Scheuering, and Dorin Comaniciu. Fast automatic heart chamber segmentation from 3D CT data using marginal space learning and steerable features. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

Yefeng Zheng, David Liu, Bogdan Georgescu, Hien Nguyen, and Dorin Comaniciu. 3d deep learning for efficient and robust landmark detection in volumetric data. In *MICCAI*, pages 565–572. Springer, 2015.

Dianle Zhou, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi. Automatic landmark location with a combined active shape model. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pages 1–7. IEEE, 2009.