Imperial College London
Department of Computing

Machine learning for efficient recognition of anatomical structures and abnormalities in biomedical images

Zhongliu Xie

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London
August 2016

Abstract

Three studies have been carried out to investigate new approaches to efficient image segmentation and anomaly detection. The first study investigates the use of deep learning in patch based segmentation. Current approaches to patch based segmentation use low level features such as the sum of squared differences between patches. We argue that better segmentation can be achieved by harnessing the power of deep neural networks. Currently these networks make extensive use of convolutional layers. However, we argue that in the context of patch based segmentation, convolutional layers have little advantage over the canonical artificial neural network architecture. This is because a patch is small, and does not need decomposition and thus will not benefit from convolution. Instead, we make use of the canonical architecture in which neurons only compute dot products, but also incorporate modern techniques of deep learning. The resulting classifier is much faster and less memory-hungry than convolution based networks. In a test application to the segmentation of hippocampus in human brain MR images, we significantly outperformed prior art with a median Dice score up to 90.98% at a near real-time speed (<1s).

The second study is an investigation into mouse phenotyping, and develops a high-throughput framework to detect morphological abnormality in mouse embryo μ-CT images. Existing work in this line is centred on, either the detection of phenotype-specific features or comparative analytics. The former approach lacks generality and the latter can often fail, for example, when the abnormality is not associated with severe volume variation. Both these approaches often require image segmentation as a pre-requisite, which is very challenging when applied to embryo phenotyping. A new approach to this problem in which non-rigid registration is

combined with robust principal component analysis (RPCA), is proposed. The new framework is able to efficiently perform abnormality detection in a batch of images. It is sensitive to both volumetric and non-volumetric variations, and does not require image segmentation. In a validation study, it successfully distinguished the abnormal VSD and polydactyly phenotypes from the normal, respectively, at 85.19% and 88.89% specificities, with 100% sensitivity in both cases.

The third study investigates the RPCA technique in more depth. RPCA is an extension of PCA that tolerates certain levels of data distortion during feature extraction, and is able to decompose images into regular and singular components. It has previously been applied to many computer vision problems (e.g. video surveillance), attaining excellent performance. However these applications commonly rest on a critical condition: in the majority of images being processed, there is a background with very little variation. By contrast in biomedical imaging there is significant natural variation across different images, resulting from inter-subject variability and physiological movements. Non-rigid registration can go some way towards reducing this variance, but cannot eliminate it entirely. To address this problem we propose a modified framework (RPCA-P) that is able to incorporate natural variation priors and adjust outlier tolerance locally, so that voxels associated with structures of higher variability are compensated with a higher tolerance in regularity estimation. An experimental study was applied to the same mouse embryo μ-CT data, and notably improved the detection specificity to 94.12% for the VSD and 90.97 % for the polydactyly, while maintaining the sensitivity at 100%.

Acknowledgements

I owe my deepest gratitude to my supervisor, Professor Duncan Gillies, for his sage advice, patient guidance and constant encouragement throughout the course of my PhD. It has been a great fortune and honour to pursue my study under his supervision.

I am also enormously grateful to my second supervisor, Professor Daniel Rueckert, for his enlightening suggestions and important support from time to time.

I extend my sincerest thanks to Professor Asanobu Kitamoto from Japan National Institute of Informatics, Professor Toshihiko Shiroishi from Japan National Institute of Genetics, and Dr. Masaru Tamura from RIKEN BioResource Centre. This thesis would not have been possible without their invaluable input.

Special thanks go to Clare Turner, Erika Helms, Jonathan Picken, Lynda Chandler, Amani El-Kholy and Hassan Patel, for their important help to secure my funding extension in the final year.

I would also like to thank my friends and colleagues for their help and support over the past years, which have enriched my life in many ways, especially Dr. Loizos Markides, Dr. Kate Reed, Dr. Zena Hira, Dr. Xi Liang, Dr. Tong Tong, Dr. Christian Ledig, Dr. Wenjia Bai, Dr. Qinquan Gao, Dr. Wenzhe Shi, Andreas Schuh, Lisa Koch, Konstantinos Kamnitsas, Dr. Jie Shen, Shiyang Cheng, Xiaoping Fan, Joao Amaral, Ting Zhao, Yu Xia, Yining Shen, Candy Shen, Yong Zhou, Yangyang Peng and Panwen Chen.

Finally, I am much obliged to my family for their constant support and love, which shaped me into who I am today.

"In theory, there is no difference between theory and practice; in practice, there is."

— *attributed to several computer scientists, most notably Jan L. A. van de Snepscheut and Yogi Berra*

Declaration of Originality

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

Zhongliu Xie

Copyright Declaration

Acronyms

AD Alzheimer's disease

ADNI Alzheimer's Disease Neuroimaging Initiative

ALM augmented Lagrangian multiplier

ANN artificial neural network

APG accelerated proximal gradient

CanonNet canonical neural network

CN cognitively normal

CT computed tomography

ConvNet convolutional neural network

dpc days post-coitum

EALM exact augmented Lagrangian multiplier

EM expectation maximisation

FN false negative

FP false positive

HU Hounsfield unit

IALM inexact augmented Lagrangian multiplier

IKMC International Knockout Mouse Consortium

IMPC International Mouse Phenotyping Consortium

MALP multi-atlas label propagation

MCI mild cognitive impairment

MI mutual information

MR magnetic resonance

MRI magnetic resonance imaging

N3 non-parametric non-uniformity normalisation

NCC normalised cross correlation

NMI normalised mutual information

PatchDNN patch-based deep neural network

PBS patch-based segmentation

PCA principal component analysis

PCP principal component pursuit

ReLU rectified linear unit

RF radio frequency

ROI region of interest

RPCA robust principal component analysis

RPCA-P robust principal component analysis with variation priors

SSD sum of squared differences

STAPLE simultaneous truth and performance level estimation

TN true negative

TP true positive

VSD ventricular septal defect

Contents

List of Tables

List of Figures

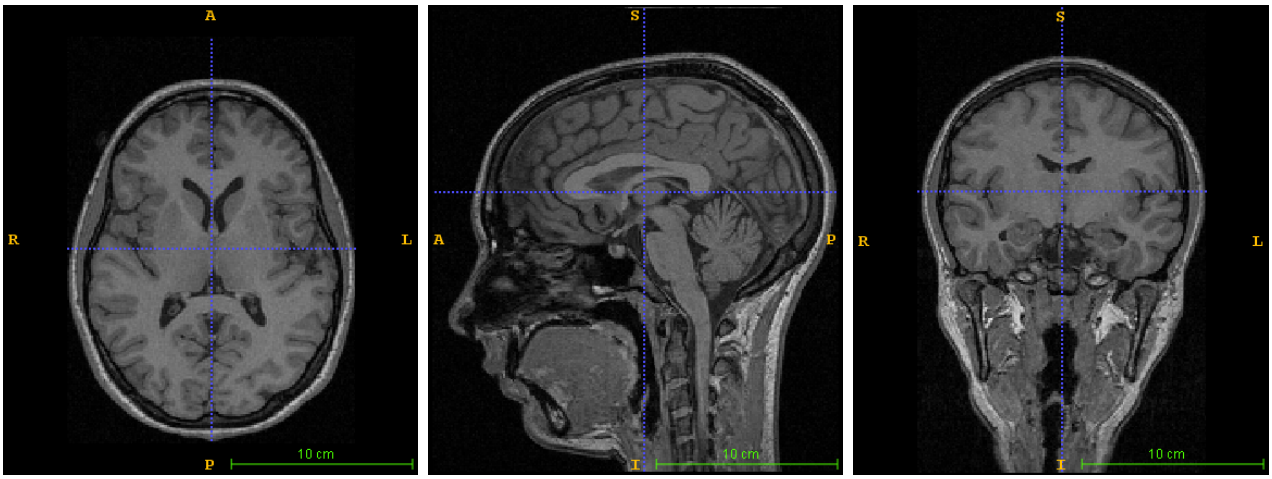
Chapter 1

Introduction

Modern biomedical studies often involve the use of imaging technology to examine biological matters in vivo in a non-invasive and efficient manner. The availability of biomedical images significantly enhances our capability to observe and understand the anatomy and biological process in humans and animals. Based on relevant imaging physics, popular biomedical imaging modalities range from ultrasonography, X-ray computed tomography (CT), magnetic resonance imaging (MRI), to positron emission tomography (PET) and single photon emission computed tomography (SPECT), etc. The imaging modalities related to this thesis, CT and MRI, will be described in more details in Section 1.2.

In the digital era, biomedical images are often stored in computer hardware, with popular file formats including Dicom, Analyze, Nifti and Minc [92], which are usually visualised using tools tailored for biomedical images, such as ITK-SNAP[1], ImageJ[2], and MIPAV[3], etc. Depending on the acquisition procedure, a biomedical image can be either a 2D slice or a 3D volume, where the latter is often visualised using three orthogonal transverse planes that, respectively, display the imaging subject in axial, saggital and coronal views. Figure 1.1 illustrates a sample 3D image of a human brain under MRI scanning. The size of a biomedical image may vary significantly based on the size of imaging subject and the spatial resolution of imaging device.

[1]ITK-SNAP official website: `http://www.itksnap.org/`
[2]ImageJ official website: `https://imagej.nih.gov/`
[3]MIPAV official website: `http://mipav.cit.nih.gov/`



(a) axial view (b) saggital view (c) coronal view

Figure 1.1: A sample 3D image of a human brain under MRI scanning in (a) axial, (b) saggital, and (c) coronal views

Biomedical image analytics is an interdisciplinary field of study involving computer science, mathematics, medicine and so on, with a primary purpose centred on extracting clinically relevant information or knowledge from biomedical images. Example tasks include organ recognition and disease diagnostics, etc. Traditionally, image analytics was carried out by domain experts manually, which was often a tedious and extremely time-consuming task. More importantly, past decades have witnessed a period of "data explosion", where the scale of image production at present day is unprecedented in history and has quickly overwhelmed the possibility of solely manual analytical work.

Computer-assisted analytics, or sometimes (bio-)medical image computing, on the other hand has become a major trend for image-based biomedical research and practices. Among a large variety of studies and projects, there are two major branches of research in this field, which aim:

1. To help recognise and/or process specific tissues, anatomical structures, organs and so forth in biomedical images.

2. To facilitate the procedures regarding the diagnosis, classification and/or identification of certain diseases or abnormalities of the imaging subjects.

One of the primary focuses in the first branch of work is the development of a series of fast

and accurate image segmentation methods, whereas the second branch is often associated with pattern recognition using disease-specific or pathology-dependent information.

To help achieve these goals, a range of image analytical techniques has been developed over the past decades. A large proportion of these techniques fall into a more general category of computing technology called "signal processing", in which an image is treated as a (multi-dimensional) matrix of intensity signals. To this day, signal processing theories and methods have turned out to be one of the major pillars underpinning the modern research of biomedical image computing, helping to shape it into its present form. Widely-used image computing techniques within this scope include image reconstruction, denoising, inhomogeneity correction, spatial encoding, intensity normalisation and image registration: a background study is easily accessible with a range of introductory materials [149, 177, 44]. However, computational approaches solely leveraging image processing techniques often have significant limitations when it comes to discovering underlying patterns within imaging data, or to making predictions on new data, for example the classification of imaging subjects, or detection of data anomaly, etc.

Recent years have witnessed a growing trend toward the employment of machine learning technology. Machine learning is a subfield of computer science centred on developing algorithms to train the computer so that it is able to learn from and make predictions on data. Based on the learning mechanism, it can be divided into supervised learning and unsupervised learning. In the supervised category, a ground truth label class will be provided for each training data entry, and the learning process is centred on fitting a classification model to the data, so that it can make effective generalisations over existing data and perform accurate classification on previously unseen data entries. Popular supervised learning frameworks include support vector machines, random forests and artificial neural networks, etc.

In the unsupervised category, such label information is not available, and the learning process is instead centred on discovering some hidden patterns or structures inside the data, such as data clusters. Popular unsupervised learning frameworks include nearest-neighbour clustering, principal component analysis (PCA) and manifold learning, etc. Introduction to machine learning is also easily accessible with a number of excellent textbooks [113, 20].

This thesis will present three studies that explore both research branches of biomedical image analytics, with a focus on the development of novel and efficient machine learning approaches to segmenting anatomical structures in human brain MRI data, as well as to identifying morphological anomalies in mouse embryo μ-CT imaging data.

1.1 Research Overview

1.1.1 Study One: efficient image segmentation using a patch-based canonical neural network

In terms of image analytics and image-powered medical interventions, identifying relevant anatomical structures in the biomedical image is a fundamental step. It is often a prerequisite to carrying out advanced procedures such as diagnostics. In biomedical image computing, this process is usually formalised as image segmentation or image annotation. In a typical procedure, the pixels/voxels pertaining to a structure of interest (such as heart or lung) or tissue type (such as grey matter or white matter) in the image will be classified using designated labels.

However, despite decades of active research in this field, the development of an accurate, robust and fast segmentation algorithm remains an open challenge. The human brain for example, consists of very complex anatomical structures that nevertheless look similar under prevailing imaging modalities such as MRI, making it difficult to achieve this goal. To address this challenge, modern approaches tend to utilise a set of labelled training images (called "atlases") to help with image segmentation. Normally, an atlas would be an image similar to the one being segmented, with target structures being manually annotated by a domain expert in advance. Due to the structural congruity of anatomy, the developed computer system is expected to mimic human behaviour and conduct similar annotation on the target image, either fully automatically (automatic segmentation), or with partial human involvement (semi-automatic segmentation).

Over the past years, a large variety of computerised segmentation frameworks have been pro-

posed. To this day, a large proportion of high-profile frameworks can be categorised into one of the two major paradigms:

1. **Label propagation via image registration [71, 91, 64, 85, 36, 65, 56, 129, 6, 40]:** the central philosophy behind this paradigm is to establish a voxel-level one-to-one correspondence between the target image and one or more atlases, so that the labels in the atlases can be propagated over the target image accordingly to perform segmentation. The spatial correspondence is often established using the "image registration" technique, and in most cases that of the "non-rigid registration" category. The developed segmentation methodology, based on the actual studies and application domains, may vary significantly in terms of image pre-processing, atlas selection, image registration scheme, label fusion, use of intermediate template, and post-processing, etc.

   However, although decades of research efforts have been devoted to the image registration framework, state-of-the-art methods still frequently encounter substantial difficulties to accurately and efficiently establish the desired voxel-wise correspondence, due to a range of reasons, including but not limited to, physiological movements, inter-subject variability, and different imaging protocols. Furthermore, computational efficiency is often limited, where the processing time to register a target image with multiple atlases can easily increase to the order of hours.

2. **Patch-based pattern matching [37, 133, 10, 49, 143, 151, 166, 156]:** As an alternative approach, instead of securing a spatial one-to-one correspondence, a patch of voxels, for example of size $5 \times 5 \times 5$, centred at each voxel of interest is retrieved in this paradigm, to perform contextual pattern matching. Typically, each target patch is pair-wise compared with a set of training patches retrieved from the atlases, and its centre voxel is then classified via label fusion using these training patches based on specific similarity measures. Depending on the actual applications, the methodology may also vary significantly, in terms of image pre-processing, atlas selection, patch search method, similarity measure and label fusion mechanism (including feature extraction), as well as image post-processing, etc.

Many methods under this paradigm are also limited by low computational efficiency as well as concerns regarding atlas selection. Moreover, patch comparison is often realised using some low-level hand-engineered features extracted from the patches, such as the sum of square differences, which generally have limited capability of feature representation.

More recently, the employment of machine learning for biomedical image segmentation has received increasing attention, in particular the deep learning framework. These approaches are centred on training a classifier based on hierarchical feature composition using deep neural networks. An artificial neural network consists of a number of inter-connected layers, each composed of a set of independent computational neurons. In the canonical architecture (CanonNet), each neuron stores a feature extraction function that takes an input in vector form and outputs a scalar-valued feature response. Each feature function contains a number of learning weights (or parameters), which are self-optimised by the computer during the learning process. In a deep network structure with many layers, the aggregate feature function can become very complex and enable high-level feature representation and classification capabilities. In the past few years, deep learning has brought about a revolution in computer vision, speech recognition, natural language understanding, sentiment analysis and so forth [94].

In the case of image classification in particular, since an image may easily scale to millions of pixels/voxels, leading to many more millions of weights to be trained using a relatively smaller number of images, the CanonNet architecture often quickly ends up over-fitting. As a result, the convolutional neural network (ConvNet) has become extremely popular, especially since the seminal work AlexNet [90] that outperformed the runner-up method by a substantial margin in the ImageNet challenge in 2012. Instead of taking an entire image directly as the input, a ConvNet neuron typically filters it with a sliding convolutional kernel which only addresses a local image patch at a time, and outputs a feature image for further processing. As a result, deep ConvNets are able to achieve advanced feature representation while using far fewer learning weights, making them more robust and popular than CanonNets for image-level classification.

Following its major success in computer vision, researchers in biomedical imaging quickly followed, and a number of image segmentation studies based on ConvNets have since been pro-

posed [131, 78, 132, 45, 126]. Typically, segmentation of an image is broken down into a collection of voxel-level classification sub-problems in a patch-based setting, in which each patch is treated as a mini-image for the labelling of its centre voxel. However, we argue that for patch-based segmentation, the ConvNet's advantage is largely diminished due to the nature of the segmentation approach, and propose a novel segmentation method based on the CanonNet architecture with substantial re-engineering. An evaluation was applied to the segmentation of hippocampus in the human brain, with a set of MRI data retrieved from a benchmark database.

1.1.2 Study Two: high-throughput mouse phenotyping using non-rigid registration and robust principal component analysis

This study explores the second major branch of research in biomedical image analytics and carries out an investigation into the development of a high-throughput image computing approach to mouse phenotyping. In this work, we will propose a morphological anomaly detection framework, based on the combined employment of non-rigid registration and robust principal component analysis (RPCA).

International efforts have been underway toward phenotyping the mouse genome, by systematically modifying mouse genes one-by-one for comparative analysis, in order to study the impact of gene-mutation with respect to morphology, metabolism or other biological traits (collectively known as the "phenotype"). In terms of phenotype examination, current practices still rest on the traditional method using microscopic histological examination, which is not only labour intensive and highly time consuming, but is also restricted to limited anatomical coverage and prone to errors during histological sectioning.

Fortunately, recent years have witnessed a growing trend regarding the employment of image analytics to facilitate the phenotyping process, especially that concerning the recognition of morphological abnormality. Similar to other biomedical imaging work, as the modern scale of data production grows at an unprecedented pace, the research community has been calling for some high-throughput image computing approaches, with the use of automatic or semi-

automatic analytical tools to address this increasing challenge.

Broadly speaking, there are two major branches of study toward computer-aided phenotyping.

1. **Phenotyping via detection of phenotype-specific features [167, 137, 161]:** The first branch is focused on developing some niche algorithms that capture phenotype-specific features to help recognise target phenotypes, for example, by detecting the connectivity between cardiac ventricles to identify ventricular septal defects (VSD) in the heart. This line of research ultimately leads to automatic classification of specific known phenotypes. However such niche approaches generally fail to serve a general phenotyping purpose, in particular the discovery of unknown phenotypes, since the corresponding phenotypical information is not available.

2. **Phenotyping via comparative analytics [35, 120, 172, 34, 89, 168, 134]:** The second branch of research is focused on anomaly detection, via data-driven comparative analytics. Existing work in this line is primarily centred on volumetric comparison of target anatomical structures between the normal and gene-modified subjects to identify anomaly. However, volume contrast merely achieves a superficial level of screening for identifying the morphological phenotype, and will generally fail when there is no severe volume variation involved, as there is, for example, in the VSD. Alternative approaches include leveraging distinguishing deformation features derived from group-wise image registration onto a purpose-built template to identify the anomaly. However, such detection approaches are often not very robust, as deformation features may vary significantly across different registration settings and the use of different templates.

Furthermore, both branches of work often require image segmentation on certain structures of interest before proceeding to further analysis. As discussed in study one above, many biomedical image segmentation methods require the availability of atlases to carry out the task. In contrast, the rapid organogenesis throughout the brief embryo development period (approximately 18.5 days) poses a significant obstacle to securing a suitable atlas, making image segmentation very challenging in the case of mouse embryo phenotyping. All these reasons lead to a critical

demand for a robust and efficient general-purpose anomaly detection framework without prior knowledge of the phenotype and the need of image segmentation, which is the purpose of this study.

We propose a systematic framework that is able to efficiently detect morphological anomalies in a batch of images, sensitive to both volumetric and non-volumetric variations, and does not require image segmentation, nor resort to unreliable deformation features, and is therefore robust to various registration settings and template use. The core of the proposed approach lies on jointly employing the non-rigid registration and RPCA methods. RPCA [32] is an unsupervised machine learning technique that is able to decompose the dataset under study into a regular and a singular component. The former is an estimated low-dimensional subspace that models the regular/standard structure of mouse anatomy, whilst the latter is mainly composed by sparse data that captures the anomalous information of each image. The purpose of non-rigid registration in this case is to group-wise align all images to ensure the effectiveness of the subspace estimation in the RPCA process. An evaluation was carried out on two separate mouse embryo μ-CT imaging datasets regarding the detection of two abnormal phenotypes: VSD and polydactyly.

1.1.3 Study Three: robust principal component analysis with variation priors (RPCA-P)

The third study investigates the RPCA framework in more depth and proposes a novel RPCA-P framework that is better able to address the prevailing natural variations in biomedical data. RPCA is an extension of the classic PCA technique, which is arguably the most widely used statistical data analytical method concerning feature extraction and dimensionality reduction. PCA uses an orthogonal transformation to convert a given dataset with possibly correlated variables into a low-dimensional subspace composed by linearly uncorrelated variables called principal components [75]. The classic PCA algorithm may be derived by assuming that the data is drawn from a multivariate Gaussian distribution. It is, however, well-known and well-documented [32, 169, 41, 42] that the classic PCA algorithm is fragile to some distributions

that depart from this assumption, such as those containing large outliers. Consequently, data distributed in a non-Gaussian manner will be represented less accurately, and a single data point with gross corruption in this case could significantly lower the quality of the resulting subspace representation. Such limitation alongside the wide prevalence of data corruption sparked the development of the RPCA framework [32, 169, 41, 42, 21, 81].

Due to the strong performance guarantees, RPCA via principal component pursuit (PCP) [32] has gradually become the standard RPCA method (the word RPCA in this thesis generally refers to PCP-based RPCA unless otherwise stated). It has previously been applied to many computer vision problems, including including video surveillance [163, 32, 8, 175, 174], face recovery [163, 32, 175, 174] and batch linear alignment of face images with partial occlusion/corruption [124], attaining excellent performances. However these applications commonly rest on a critical condition, which is that in the majority of images being processed, there is a background with very little variation.

By contrast, in biomedical imaging, there is a significant natural variation across different image samples, such as inter-subject variability, physiological movements (such as heartbeat), different postures and orientations at different imaging procedures. Also, different anatomical structures often manifest varying individual natural variations. For that reason, the method needs to be upgraded to address this condition. Although non-rigid image registration can go some way toward reducing this variance, it cannot eliminate the problem entirely. Moreover, in the baseline RPCA framework, the level of outlier tolerance for marginal data to be included into the regular or singular component is solely controlled by a single parameter, applied globally regardless of local variability.

To improve this purely unsupervised machine learning approach without leveraging any prior knowledge, we propose a modified RPCA framework (RPCA-P) that is able to incorporate natural variation priors in the model and adjusts outlier tolerance locally so that voxels associated to structures of higher natural variability are compensated by allowing a higher tolerance during feature decomposition. The variation priors are learned from the data itself. In this case, the proposed method is able to significantly improve the performance of feature decomposition

in the biomedical domain. An evaluation was carried out on a revisit to the mouse embryo phenotyping application, in order to investigate whether the RPCA-P improves the detection performance.

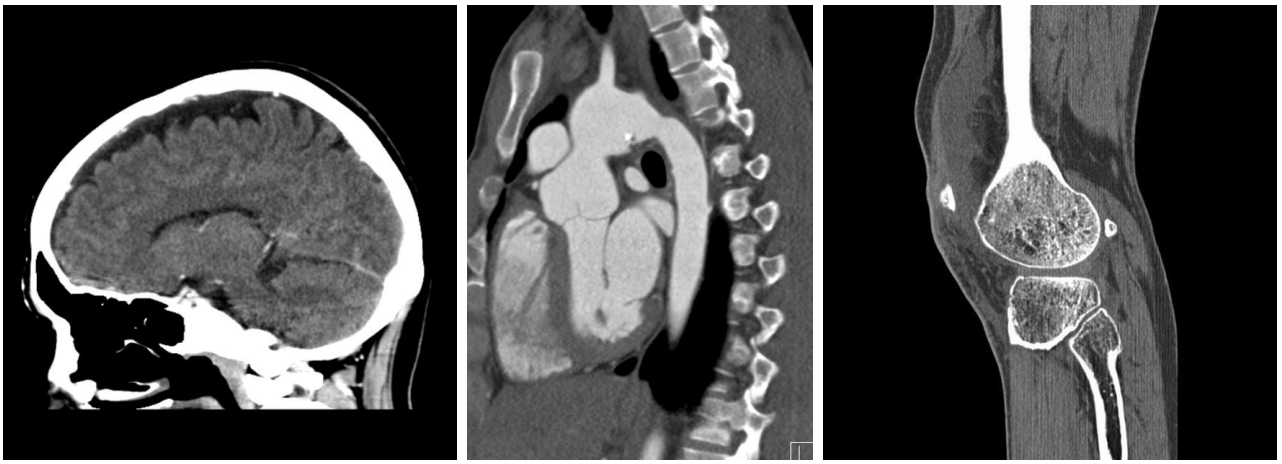
1.2 Imaging modalities in this thesis

As mentioned earlier, there are a variety of imaging modalities used in biomedical practices. Each modality has a unique visualisation property based on its imaging physics. In general, imaging technologies can be broadly divided into two categories: ionising radiation and non-ionising radiation. The ionising radiation category primarily consists of modalities that utilise X-rays (such as CT, Radiography) or Gamma rays (such as SPECT) to produce images. The non-ionising radiation category mainly utilises either acoustic pulses (such as ultrasonography) or radiowaves in combination with magnetic fields (such as MRI). Depending on the actual application, there could be one or multiple imaging modalities used at the same time. Here, we briefly describe some basics of CT and MRI, which are the imaging modalities used in the work presented in this thesis.

1.2.1 Computed tomography (CT)

CT is a prevailing imaging technology that produces tomographic images based on the combined use of X-ray and computer reconstruction. When performing a CT scan, the imaging subject is often positioned on a table, which gradually moves across the centre of an X-ray machine. The machine consists of a rotating radiation source that emits X-rays passing through the subject, and a signal receiver that measures the attenuation level of X-ray beam that arrives. As the source rotates, radiation decay is measured from multiple orientations. A computer tomographic reconstruction process such as Feldkamp's filtered back-projection [50] can then be applied and generate a grey-level intensity image based on the distribution of X-ray attenuation.

The intensity value of a pixel/voxel x in the generated CT image reflects the level of signal



(a) brain (b) heart (c) knee

Figure 1.2: Some sample 2D slices of CT scans on (all in sagittal view): (a) brain, (b) heart and (c) knee (image source: `http://radiopaedia.org/`)

attenuation as X-ray passes through the location, and is often computed based on the Hounsfield Unit (HU) [142]:

$$HU_x \;=\; 1000 \times \frac{\mu_x - \mu_{water}}{\mu_{water}}$$

where μ_x is the average linear attenuation coefficient at voxel x and μ_{water} represents the attenuation level of water. Since different types of tissue absorb X-rays at different levels, the HU value varies from voxel to voxel, resulting in a signal intensity image that visualises various parts and structures of the imaging subject, including bones, muscles, fat, tumours and internal organs.

The HU values in a CT image usually range from -1000 to +1000, with the air corresponding to around -1000 HU, water to around 0 HU, and bone to around +1000 HU. Some sample CT scans of the brain, the heart and the knee are shown in Figure 1.2, in which the brain skull and bones appear bright, the soft tissues grey, and the air dark.

1.2.2 Magnetic resonance imaging (MRI)

MRI is another widely-used tomographic imaging technology introduced some 30 years ago, and is able to generate high quality images based on the phenomenon of nuclear magnetic resonance. The fundamentals of MRI physics will be outlined here, yet for a more detailed

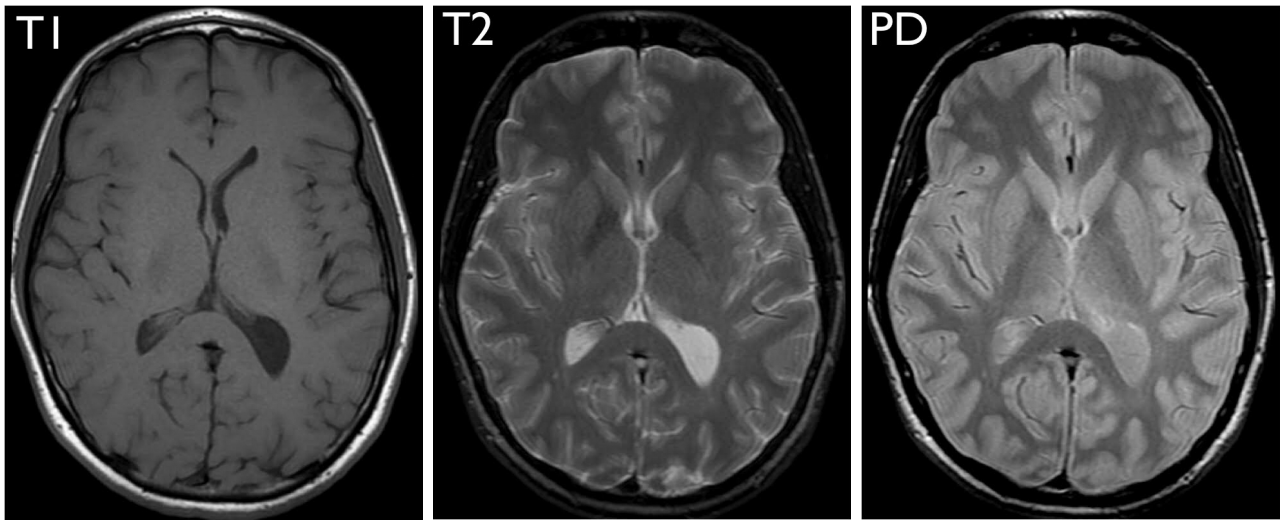
description one should refer to further readings [60].

When performing an MRI scan, the imaging subject is often placed in the bore of a superconducting magnet that generates a strong magnetic field. The magnetic field then aligns the spins of hydrogen nuclei (or protons) within the subject, most notably in water molecules. Such alignment can be disrupted by applying an external radio frequency (RF) electromagnetic pulse. In response to the force returning them to the equilibrium state, the nuclei precess causing a changing magnetic flux, leading to a differential voltage in receiver coils as a measurable response signal. Moreover, by applying additional magnetic field gradients that vary linearly over space, the source of response signals can be located using the frequency of resonance signals. Then an image of the subject anatomy can be produced, by applying a Fourier transform that decodes such frequencies into a spatial map.

The image quality and signal-to-noise-ratio of MRI scanning are determined by the strength of magnetic field, where the common setting in existing clinical systems is either 1.5T or 3T. The signal-to-noise-ratio of a 3T scan in theory is twice as good as a 1.5T scan, with an improved spatial resolution and reduced acquisition time. Nevertheless, MRI suffers from wide prevalence of artefacts such as spatial distortion, which tends to be stronger in higher magnetic field strength.

Furthermore, the signal intensity of MRI is collectively determined by a number of factors, most notably the longitudinal relaxation time (T1), transverse relaxation time (T2) and proton density. T1 and T2 affect how protons return to the equilibrium state after an RF pulse is applied, making a direct impact on image contrast. More specifically, there are two main parameters in the pulse sequence: the repetition time (TR) and echo time (TE). The former indicates the interval between two RF pulses whereas the latter indicates the duration from the start of an RF pulse to the detection of its response signal.

Based on the settings of TR and TE, there are three basic forms of MRI scanning, respectively, T1-weighted, T2-weighted and PD-weighted MR images [66]. The T1-weighted imaging sequence applies a short TR and short TE, the T2-weighted sequence applies a long TR and long TE, and the PD-weighted sequence uses a long TR and short TE. An example brain scan



(a) (b) (c)

Figure 1.3: Sample scans of human brain under (a) T1-weighted, (b) T2-weighted, and (3) PD-weighted MRI sequences (image source: `https://radiology.ucsf.edu/blog/neuroradiology/exploring-the-brain-how-are-brain-images-made-with-mri`)

of each imaging sequence is illustrated in Figure 1.3.

Relevant studies have suggested that T2-weighted scan is most suitable for the analysis of brain pathologies such as vascular damage or other intra-cranial diseases [46]. In contrast, T1-weighted and PD-weighted scans are preferable when characterising lesions. Moreover, a T1-weighted scan is often considered the best option in the case of analysing pathologies such as brain atrophy [63]. In addition, sometimes multiple sequences are used at the same time to enhance image analytics, such as to improve the segmentation of brain tumours using a combination of T1-weighted and T2-weighted scans [112].

1.2.3 CT vs MRI

Both CT and MRI are popular imaging technologies, each has a unique set of advantages and disadvantages, suitable for different imaging requirements and making them complementary to each other. Compared to MRI, CT offers a higher spatial resolution and requires a relatively short acquisition time (in the order of tens of seconds). Furthermore, there is a good imaging contrast between bones and soft tissues in CT scans, which generally outperform MRI when it comes to examining trauma, or other similar situations. On the other hand, CT suffers from

a poor soft tissue contrast, which makes it difficult to distinguish various soft tissues and to conduct relevant pathological analysis, though such limitation can be significantly relieved with the use of contrast agents. In addition, X-rays are an ionising form of radiation and could be harmful to the human body.

In contrast, MRI is non-ionising and considered a safe imaging technology without measurable harm. Another major advantage of MRI is its ability to adjust image contrast, and thus can be employed to highlight different types of soft tissue by applying customised pulse sequences. Moreover, MRI can be tailored for functional [106] imaging (fMRI) or diffusion [19] imaging (DWI or DW-MRI), satisfying a wider range of observational and analytical demands. However, since there is little presence of hydrogen nuclei for magnetic resonance in the bones, MRI is generally limited to a poor contrast between bones and soft tissues.

1.2.4 Micro-level imaging: μ-CT and μ-MRI

Regular CT and MRI often produce images at spatial resolutions in the order of millimetres. In some circumstances, imaging needs to be applied to small animals, microfossils or certain biomedical samples, etc., in which case the imaging subject is often too small to meet the regular CT and MRI standards. X-ray micro-computed tomography (micro-CT or μ-CT) on the other hand, is a special type of CT that produces images with a spatial resolution in the micrometre range. Its counterpart using MRI is known as micro-MRI or μ-MRI. Such high-resolution imaging allows for much refined visualisation to cater for the demand of micro-level observation. In study two and study three, image analytics will be carried out on a set of normal and gene-modified embryos of laboratory mice, in order to observe the subject morphology at prenatal stage. In this condition, μ-CT is used for data acquisition instead of the regular CT. In other similar studies, μ-MRI is also frequently used.

1.3 Summary of Key Contributions

The key contributions of the work presented in this thesis are summarised below:

- A novel biomedical image segmentation framework using a patch-based deep neural network, which in contrast to following the popular ConvNet architecture, employs the CanonNet architecture instead: in our evaluation study applied to the segmentation of hippocampus in human brain MRI data, the proposed framework outperformed the prior state-of-the-art with an improved accuracy and a near real-time speed.
- A systematic morphological anomaly detection framework based on the combined employment of non-rigid registration and RPCA techniques. The proposed framework offers a high-throughput approach to the widely-interesting mouse embryo phenotyping work.
- An RPCA-P technique that incorporates variation priors into the state-of-the-art PCP-RPCA framework, allowing it to work much more robustly in biomedical imaging.

1.4 Structure of the Thesis

Following this introduction chapter, the thesis is organised as follows:

- To start with, study one is presented in Chapter 2. It includes a comprehensive review of well-known existing image segmentation techniques based on registration-based label propagation and patch-based pattern matching, as well as recent work on ConvNet segmentation, before proceeding to describe the proposed framework.
- Study two is then detailed in Chapter 3. It starts by motivating the mouse phenotyping work using biomedical image analytics, then reviews existing phenotyping approaches and limitations, and further points out the challenges regarding mouse embryo phenotyping, followed by the methodological development of our high-throughput anomaly detection framework using non-rigid registration and RPCA.

- Based on the insights from study two, study three is described in Chapter 4. It first introduces the background of RPCA, with an emphasis on the work related to the principal component pursuit method, including the formulations of different optimisation problems and popular existing algorithms to solve them.

- Chapter 5 provides a summary of the work presented in this thesis, our major contributions and the key novelties. In addition, some suggestions for future studies carrying on these lines of work will be also be included.

1.5 List of publications

This thesis is based on the following publications:

- Z. Xie and D. Gillies. Near real-time hippocampus segmentation using a patch-based canonical neural network. *IEEE Transactions on Medical Imaging* (in submission).

- Z. Xie, X. Liang, L. Guo, A. Kitamoto, M. Tamura, T. Shiroishi, and D. Gillies. Automatic classification framework for ventricular septal defects: a pilot study on high-throughput mouse embryo cardiac phenotyping. *Journal of Medical Imaging* 2(4):041003, 2015.

- Z. Xie, A. Kitamoto, M. Tamura, T. Shiroishi, and D. Gillies. Non-rigid registration and robust principal component analysis with variation priors: a high-throughput mouse phenotyping approach. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1118–1122. Prague, Czech Republic, 2016.

- Z. Xie, A. Kitamoto, M. Tamura, T. Shiroishi, and D. Gillies. High-throughput mouse phenotyping using non-rigid registration and robust principal component analysis," In: *SPIE Medical Imaging*, vol. 9784, no. 978415. San Diego, USA, 2016.

- Z. Xie and D. Gillies. Patch forest: a hybrid framework of random forest and patch-based segmentation. In: *SPIE Medical Imaging*, vol. 9784, no. 978428. San Diego, USA, 2016.

- Z. Xie*, X. Liang*, M. Tamura, T. Shiroishi, and A. Kitamoto. Towards high-throughput mouse embryonic phenotyping: a novel approach to classifying ventricular septal defects. In: *SPIE Medical Imaging*, vol. 9413, no. 94131V. Orlando, USA, 2015.

- Z. Xie*, X. Liang*, A. Kitamoto, M. Tamura, T. Shiroishi, and R. Kotagiri. Novel atlas-based approach to the detection of mouse embryo ventricular septal defects. In: *MICCAI Workshop on Imaging Genetics*. Boston, USA, 2014.

- A. Kitamoto, S. Roy, W. Grimes, S. Kerjose, X. Liang, Z. Xie, M. Tamura, T. Shiroishi. Mouse phenotyping using registration-based deformation features (in Japanese). In: *Bioimage Informatics Workshop*. Okazaki, Japan, 2014.

— * indicates an equal contribution by these authors

Chapter 2

Efficient Image Segmentation Using A Patch-based Canonical Neural Network

2.1 Introduction

In biomedical image analytics, segmentation of relevant anatomical structures in the medical image is a fundamental step, and often a prerequisite to advanced procedures such as diagnostics. Traditionally this is done by an expert segmenting the image by hand, which is extremely time-consuming, prone to errors, hard to reproduce, unscalable, and also subject to inter-/intra-annotator variability [138]. Furthermore, the vast amount of data produced everyday also makes manual segmentation rather prohibitive. This naturally leads to the need of (semi-)automatic segmentation technology. However, despite decades of active research in this field, the development of an accurate, robust and fast segmentation algorithm remains an open challenge. The brain for example, composed of complex anatomical structures with similar appearances under the widely used MRI scanning, places a strict barrier to achieving this goal.

To address this difficulty, state-of-the-art approaches tend to leverage a set of labelled training data (called "atlases") to help with segmentation. The term "atlas" originates from cartography regarding the study and practice of drawing maps; in the biomedical domain, its definition extends to the annotation of structures of interest in the biomedical images. The interpretation

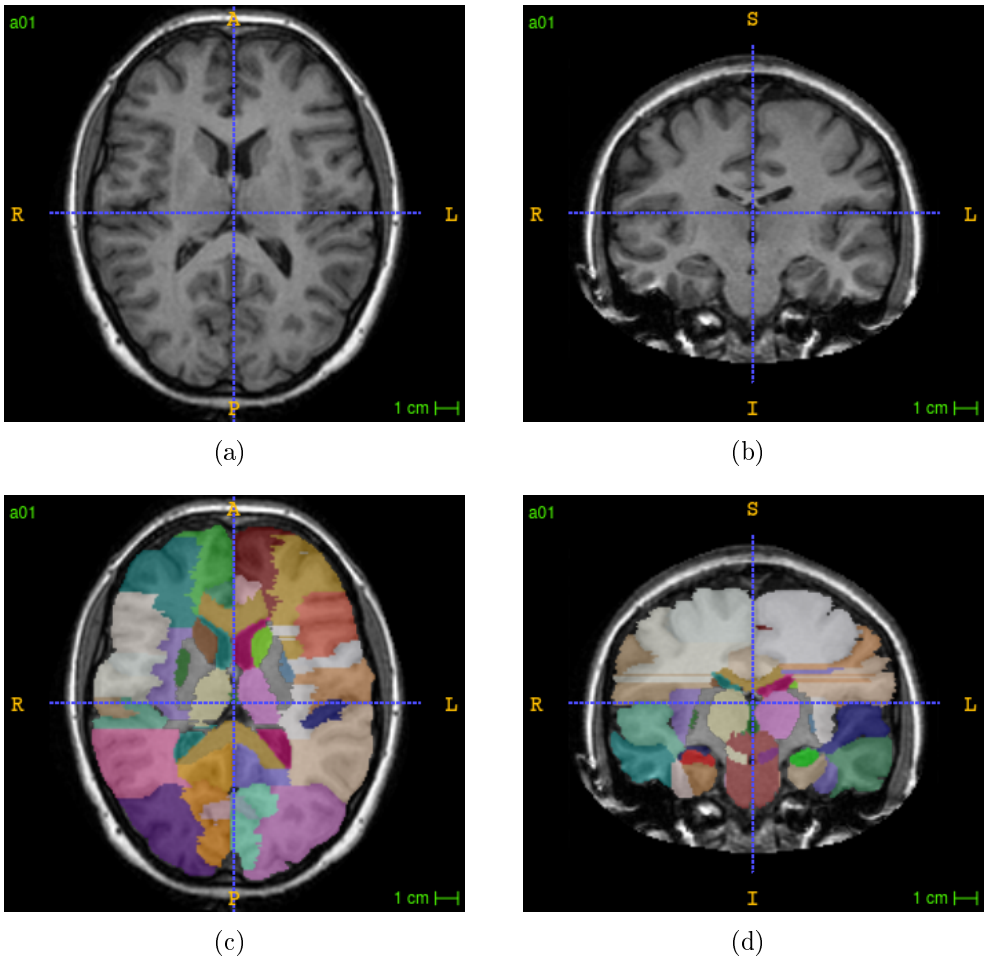
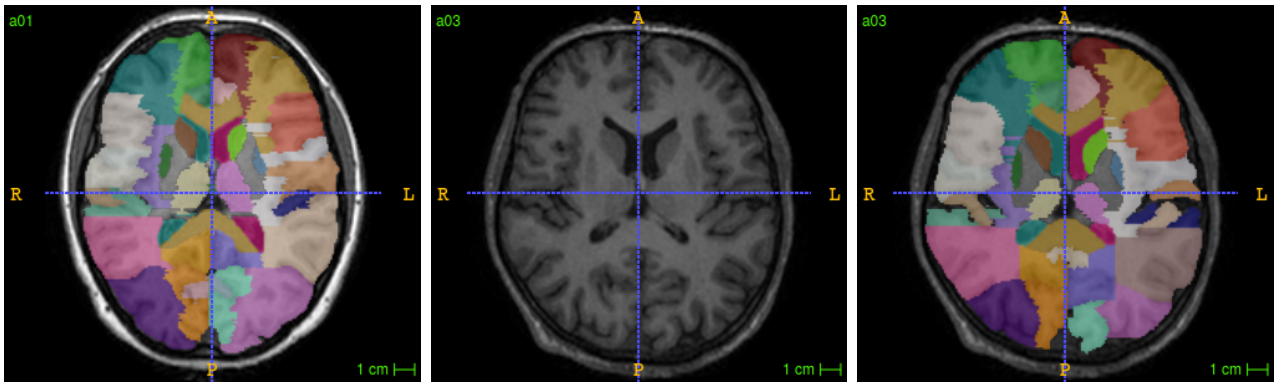


Figure 2.1: An example MRI atlas of a human brain: (a-b) the axial and coronal views of the grey-level image, (c-d) superimposed with its label map

of an atlas varies across different research communities: some interpret it as a pair of image and label map, others extend it to a collection of such pairs. Another less popular interpretation associates the term with a model template image without label annotations. To clarify its use in this thesis, we adopt the first interpretation and define an atlas as a tuple (I, L), where I represents the image and L represents its label map of target structures. An example atlas is shown in Figure 2.1. The production of a label map may be carried out by an expert manually, or with the use of automatic or semi-automatic segmentation tools. In early studies, researchers tended to create a model atlas based on a set of images for label map production. However an atlas created in this way cannot characterise the anatomical variability within the population. For that reason, in modern work each training image is often used to create a separate atlas, and the whole atlas collection is then used for segmentation in order to embrace the full spectrum of population diversity.



(a) atlas (b) target (c) segmented target

Figure 2.2: Atlas-based segmentation: (a) an atlas (b) the target image (c) the segmented target image

In segmentation problems, an atlas is normally similar to the image being segmented. By leveraging the structural congruity of anatomy, the developed algorithm is expected to mimic human behaviour and conduct similar annotation on the target image, as shown in Figure 2.2. This is often known as "atlas-based segmentation". Over the past years, a large collection of atlas-based segmentation methods has been proposed. A significant proportion of high-profile segmentation frameworks to this day, can be categorised into two major signal processing-based paradigms: label propagation via image registration and patch-based pattern matching.

In the label propagation paradigm, typically a voxel-level one-to-one correspondence will be established between the atlases and target image using image registration techniques, and the labels are then propagated over the target space to perform segmentation. To improve segmentation accuracy, usually multiple atlases are used at the same time, followed by some label fusion techniques, such as majority voting [64, 85, 36, 65, 56] or weighted label fusion [129, 6, 40]. Many algorithms of this type have been proposed in the literature, with high levels of accuracy reported in many studies [71, 91, 64, 85, 36, 65, 56, 129, 6, 40]. However these approaches typically require a non-rigid registration between every target-atlas pair, which can be very time-consuming with a large number of atlases. Some variant methods have been proposed to improve the efficiency, such as label propagation via composite transformation [108, 109], or using a probabilistic atlas in combination with statistical models to conduct label inference [123, 52, 12, 125, 98, 107]. However, the efficiency improvement is often achieved at a certain level of compromise on segmentation accuracy. Meanwhile, these methods often raise

an additional issue regarding the selection/creation of an intermediate template, which is not only application-specific but also data-dependent [76].

As an alternative, another school of researchers have developed a series of patch-based segmentation methods [37, 133, 10, 49, 143, 151, 166, 156]. In these approaches, the strict voxel-level one-to-one correspondence for label propagation is relaxed to patch-based pattern matching, where patches are typically retrieved from the atlases using a localised search window. Patch relevance will then be measured using some hand-engineered similarity metrics (such as the sum of squared differences) in the form of pair-wise comparison. Subsequently, a set of best matching patches drawn from all atlases are used collectively (rather than atlas-by-atlas) to perform label fusion. A particular advantage of this paradigm is that it no longer requires any form of non-rigid registration, and thus avoids a series of troubles including the risk of registration failure, the issues regarding template creation and composite transformation, etc. State-of-the-art segmentation accuracy has also been reported in many relevant applications, however segmentation efficiency generally remains low. The key bottleneck lies on the time-consuming patch search and label fusion computation based on pair-wise patch comparison, although a series of more efficient patch search techniques have been proposed to mitigate this problem.

Recently, machine learning using deep neural networks (often referred to simply as "deep learning") is quickly gaining wide attention. These approaches are centred on training a classifier based on high-level hierarchical feature representations. The classifier delivers an abstract generalisation on the training data, and when applied to the test data, classification can then be carried out efficiently without explicitly re-using the training data. More specifically, a neural network consists of a number of inter-connected layers, each is composed by a set of independent computational neurons. In a canonical architecture (we will call it "CanonNet" hereafter), each neuron stores a feature extraction function that takes input in vector form and outputs a scalar valued feature response. With a deep structure (many layers), the aggregate feature function learned from the data can become very intricate and enable an excellent capacity of feature representation and classification, as in contrast to the simple hand-engineered features used in conventional approaches.

In terms of image classification however, an image may easily scale to millions of pixels/voxels, leading to millions of learning parameters for a single neuron and potentially millions of neurons in a network, quickly ending up over-fitting. As a result, the convolutional neural network ("ConvNet" hereafter) has become particularly popular. In the ConvNets, instead of fully connecting to the image for global processing, the neurons filter it by convolution, in which each neuron is associated with a convolutional kernel that only locally connects to a small image patch at a time, and produces a feature image (rather than a scalar) for further processing. Deep ConvNets can achieve intricate feature representation while using far fewer learning parameters, making them more robust and popular than CanonNets for image classification. Over the past few years, deep ConvNets have brought about a revolution in computer vision, most notably since the ImageNet challenge in 2012, when the AlexNet [90] significantly outperformed the runner-up with almost a half error rate. Following these striking successes, deep ConvNets have increasingly been applied to biomedical image segmentation lately. Notable works include the U-Net [131][1] and DeepMedic [78][2], amongst many others [132, 45, 126]. Image segmentation is often broken down to voxel-wise labelling on a patch-based setting, where a patch is treated as a mini-image for classification of its centre voxel.

However, ConvNet classification generally involves millions of neuron-wise convolution on each single run, requiring long training periods and demanding memory consumption (for example, the AlexNet required 5-6 days of training on two modern GPUs [90]). Moreover, many existing segmentation studies are based on a nesting structure that integrates with other models such as superpixels [132] or conditional random fields [78, 45], further complicating the computation. By contrast, we argue that for patch-based segmentation, the ConvNet's advantage is largely diminished, as a patch is only sized around 9×9 to 15×15, and the nature of the classification approach is that patches do not need decomposition and thus will not benefit from further convolution. Instead, we revisit the CanonNet architecture in which neurons only compute dot products, and are therefore much faster and less memory-hungry than convolution. Furthermore, we have also substantially re-engineered the CanonNet architecture with state-of-the-art

[1]The U-Net has won multiple segmentation challenges, including the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks in 2012, ISBI challenge for computer-automated detection of caries in bitewing radiography in 2015, and ISBI challenge for cell tracking challenge in 2015

[2]The DeepMedic has won the MICCAI challenge for ischemic stroke lesion segmentation in 2015

deep learning techniques, including the use of ReLU activation [90], 2.5D tri-planar patch setting [126] and dropout layers [68], in addition to GPU programming that significantly boosts computational efficiency. Although our approach is general, we conducted experimentation on the segmentation of hippocampus in the human brain, and achieved a median Dice score up to 90.98% at a near real-time speed (<1s). To the best of our knowledge, this is by far the fastest algorithm with highest segmentation accuracy ever reported in hippocampus segmentation.

2.2 Segmentation via registration-based label propagation

The label propagation approach heavily rests on the use of the image registration technique, and is therefore sometimes known as "registration-based segmentation". Image registration is the process to estimate a spatial transformation that maps one image (often called the "source image" or "moving image") to another (often called the "reference image", "target image" or "fixed image"), so that a voxel-level correspondence can be established between the two images. When an atlas is successfully registered with a target image (normally through a non-rigid registration scheme), its labels are propagated to the target space using the derived transformation to perform voxel-wise label classification, as illustrated in Figure 2.3. The image registration technique will be introduced in Section 2.2.1, followed by a detailed description of the label propagation framework in Section 2.2.2. Image registration will also be used in Chapter 3 and Chapter 4.

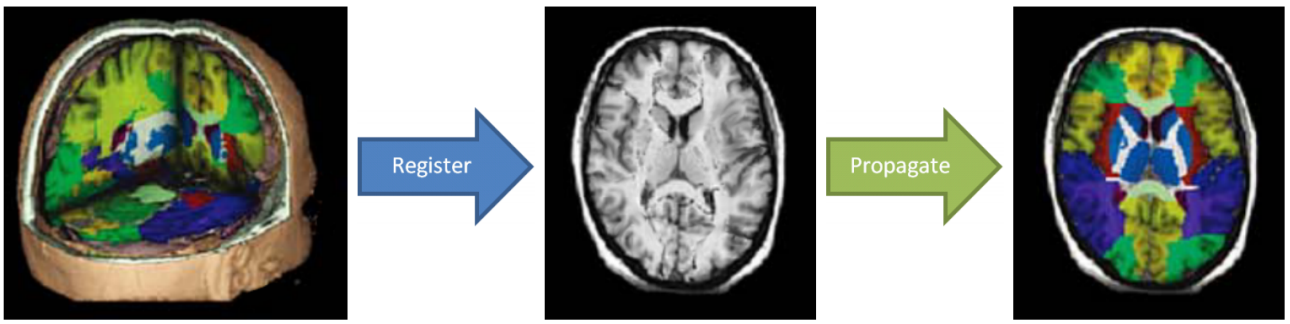


Figure 2.3: Image segmentation via registration-based label propagation (image source: [159])

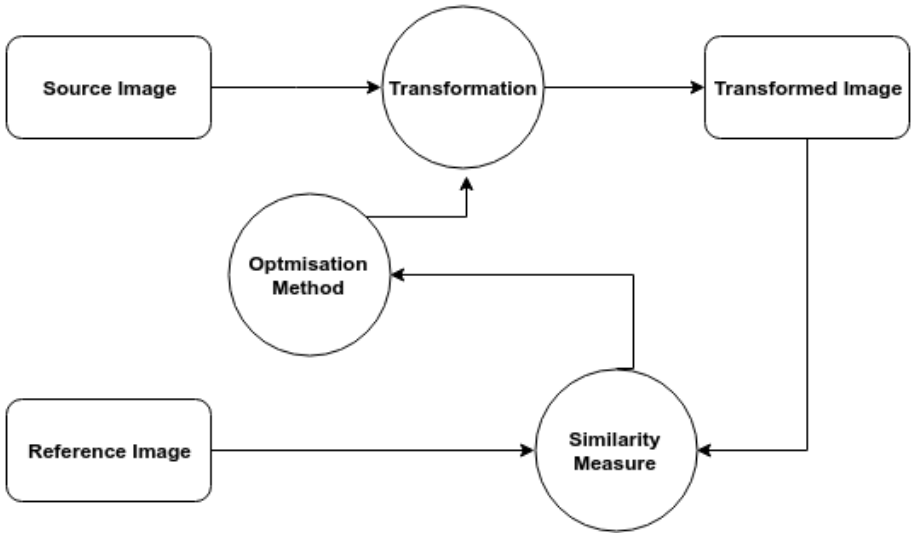


Figure 2.4: Overview of the image registration process

2.2.1 Image registration

Image registration is one of the major cornerstones underpinning modern biomedical image analytics. Despite a range of variant schemes having been proposed in the literature (which have been extensively reviewed in a number of survey studies [61, 67, 181, 115]), image registration is generally based on the process illustrated in Figure 2.4, with a number of major components to build up its computational pipeline, including similarity metrics, transformation models and optimisation methods.

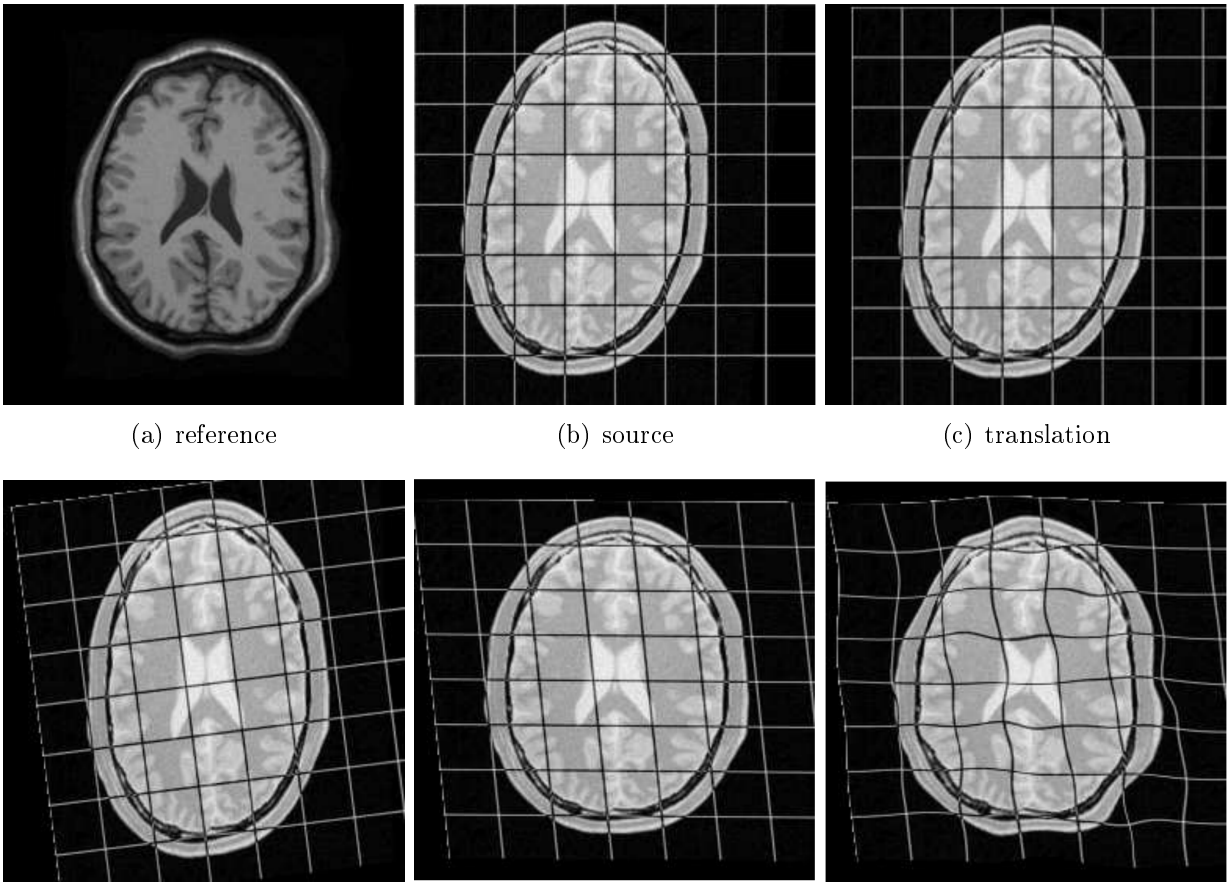
Denote the source image as I_S, the reference image as I_R, a physical point x in an image I as $\Pi(I, x)$. The registration process is to find a transformation (also called "deformation field") T such that $\Pi(I_S, T(x))$ spatially aligns with $\Pi(I_R, x)$. Technically, the mapping is from the reference space to the source, in order to ensure the completeness of a deformed image. The quality of alignment is defined by a cost function $f(T, I_S, I_R)$ based on a certain similarity measure (also called "distance measure") ζ, and in the case of non-rigid registration, since the problem is ill-posed, usually an additional penalty term (also called "regularisation term") ψ will be included to constrain T. Mathematically, the registration process can be formulated as:

$$\arg\min_{T} . \qquad f(T, I_S, I_R) = -\zeta(T, I_S, I_R) + \gamma \cdot \psi(T) \tag{2.1}$$

To improve the efficiency of registration, a multi-resolution scheme is often applied, where the estimation of T starts with the images at a relatively low resolution, which gradually increases, until reaching full resolution. There are a number of multi-resolution strategies, for example, based on down-sampling, Gaussian smoothing or both [101].

Transformation model

The transformation model determines how the source image can deform in order to align with the reference image. Based on the degree of freedom, in increasing order, major transformation models include translation, rigid, affine and non-rigid transformations. A visual demonstration of brain transformation with different models is illustrated in Figure 2.5.



(a) reference (b) source (c) translation

(d) rigid (e) affine (f) B-spline

Figure 2.5: Visual demonstration of major transformation models: (a) the reference image, (b) the source image overlaid with a grid, (c) deformed by a translation, (d) a rigid transformation, (e) an affine transformation, and (f) a B-spline non-rigid transformation. (Image source: [83])

To begin with, the first three models fall into the category of global transformation, in which the same computation is applied to all data in the image or a region of interest (ROI). Supposing $x = \langle x_1, x_2, x_3 \rangle$ denotes any physical point in 3D space (the 2D case can be easily deduced accordingly) and θ is the set of parameters associated with the transformation model, these models are formulated as follows:

- **Translation:**

$$T(x) = x + t \tag{2.2}$$

where $t = \langle t_1, t_2, t_3 \rangle$ is a translation vector that shifts x along each of the axes. The model in this case has three degrees of freedom and is parametrised by $\theta = (t_1, t_2, t_3)$.

- **Rigid transformation:**

$$T(x) = R(x - c) + t + c \tag{2.3}$$

where $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$ is a rotation matrix, $c = \langle c_1, c_2, c_3 \rangle$ is the centre of rotation, and t is a translation vector like in the previous model. In this model, the image is treated as a rigid body that can translate and rotate, but cannot be scaled or sheared. The rotation matrix is defined by three Euler angles α, β, γ, respectively, for the 2D planes over every two axes (note in the case of 2D there would be only one Euler angle): $R =$

$$\begin{bmatrix} r_{11} = cos\beta \cdot cos\gamma & r_{12} = cos\alpha \cdot sin\gamma + sin\alpha \cdot sin\beta \cdot cos\gamma & r_{13} = sin\alpha \cdot sin\gamma + cos\alpha \cdot cos\beta \cdot cos\gamma \\ r_{21} = -cos\beta \cdot cos\gamma & r_{22} = cos\alpha \cdot cos\beta - sin\alpha \cdot sin\beta \cdot sin\gamma & r_{23} = sin\alpha \cdot cos\gamma + cos\alpha \cdot sin\beta \cdot sin\gamma \\ r_{31} = sin\beta & r_{32} = -sin\alpha \cdot cos\beta & r_{33} = cos\alpha \cdot cos\beta \end{bmatrix}$$

Therefore this model is parametrised by $\theta = (\alpha, \beta, \gamma, t_1, t_2, t_3)$, increasing the degree of freedom to six.

- **Affine transformation:**

$$T(x) = A(x - c) + t + c \tag{2.4}$$

where the transform matrix $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ has no restrictions as previously, and the image can translate, rotate, and also scale and shear. For that reason, the model has 12 degrees of freedom and is parametrised by $\theta = (a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, a_{31}, a_{32}, a_{33}, t_1, t_2, t_3)$. Alternatively, some researchers prefer to parametrise A in terms of rotation, scale and shear, and define the transform matrix as

$$A = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{bmatrix} \begin{bmatrix} 1 & s_3 & s_1 \\ 0 & 1 & s_2 \\ 0 & 0 & 1 \end{bmatrix}$$

where α, β, γ are the Euler angles for rotation, k_1, k_2, k_3 and s_1, s_2, s_3 are the scaling and shearing factors along each axis. In this case, the model still has 12 degrees of freedom but is instead parametrised by $\theta = (\alpha, \beta, \gamma, k_1, k_2, k_3, s_1, s_2, s_3, t_1, t_2, t_3)$. Also, when $k_1 = k_2 = k_3 = 1$ and $s_1 = s_2 = s_3 = 0$, it is equivalent to a rigid transformation. Another notable point is that an affine transformation preserves parallels and only certain shears will do this.

Although such global transformations can achieve an overall alignment between I_S and I_R, local correspondence is not well addressed, in particular soft tissues often exhibit non-rigid deformations. To address this issue, there has been a range of non-rigid transformation models proposed in the literature, including the spline models [23, 135], elastic-solid models [153], physical models [154], viscous-fluid models [39], linear combination of some basic functions such as wavelet basis functions [164], and smoothed displacement fields [5].

Among them, the thin-plate spline and B-spline models are probably the most commonly used, and both embody transformation using control points. The thin-plate splines were initially introduced to model the deformation of shapes in medical image analysis by Bookstein [23]. An important merit of thin-plate splines is that they provide a close-form solution of T estimation using landmarks, which on the other hand however, also makes it very sensitive to landmark

placement. In particular, the movement of any landmark will affect the entire deformation field, and as a consequence makes the model computationally inefficient.

By contrast, Rueckert et al. [135] proposed free-form deformation based on B-splines, which has arguably become a standard non-rigid transformation model. This method features deformation with local control point support, which enables efficient implementation, including the use of GPU programming [114]. In this model, the image is overlaid with an $n_1 \times n_2 \times n_3$ grid Φ, composed by a set of evenly spaced control points $\phi_{i,j,k}$. Then the local transformation is modelled in the form of cubic B-splines:

$$T(x) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l(u)B_m(v)B_n(w)\phi_{i+l,j+m,k+n} \tag{2.5}$$

where $i = \lfloor \frac{x_1}{n_1} \rfloor - 1$, $j = \lfloor \frac{x_2}{n_2} \rfloor - 1$, $k = \lfloor \frac{x_3}{n_3} \rfloor - 1$, $u = \frac{x_1}{n_1} - \lfloor \frac{x_1}{n_1} \rfloor$, $v = \frac{x_2}{n_2} - \lfloor \frac{x_2}{n_2} \rfloor$ $w = \frac{x_3}{n_3} - \lfloor \frac{x_3}{n_3} \rfloor$ and $B_l(\cdot)$ represents the lth basis function of the B-spline [95, 96]

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

In this case, the degree of freedom is $n_1 \times n_2 \times n_3 \times 3$, which is determined by the resolution of Φ (note each control point can deform in a 3D space). In practice, the B-spline deformation is normally applied after (or combined with) a global transformation (usually through affine registration) that secures an overall correspondence.

Furthermore, as mentioned earlier, in non-rigid registration the deformation often needs to be regularised to secure a good performance, for example, in terms of smoothness [135], curvature [51] or rigidity [140], etc. The bending energy penalty proposed in the original free-form deformation framework [135], for example, is a widely used in combination with B-spline registration to regularise deformation smoothness. It is defined by:

$$\begin{aligned}\psi(T) = \frac{1}{|\Omega(I_R)|}\int_0^{x_1}\int_0^{x_2}\int_0^{x_3} \Big[\big(\frac{\partial^2 T}{\partial x_1^2}\big)^2 + \big(\frac{\partial^2 T}{\partial x_2^2}\big)^2 + \big(\frac{\partial^2 T}{\partial x_3^2}\big)^2 \\ +2\big(\frac{\partial^2 T}{\partial x_1 x_2}\big)^2 + 2\big(\frac{\partial^2 T}{\partial x_2 x_3}\big)^2 + 2\big(\frac{\partial^2 T}{\partial x_1 x_3}\big)^2\Big] dx_1 dx_2 dx_3\end{aligned} \tag{2.6}$$

Another notable point is that transformation is applied to physical points in the image space, which may correspond to individual voxel centres, as well as inter-voxel locations.

Similarity metrics

Similarity metrics are used to evaluate the quality of alignment between I_R and the deformed I_S. Unlike the transformation model that can be applied to any physical points, the similarity measure is based on the actual voxels in the images or ROIs. Since the voxel centres in I_R often do not exactly align with their counterparts in the deformed I_S, an interpolation process is usually involved. Popular interpolation methods include nearest neighbour interpolation, linear interpolation (sometimes called "tri-linear interpolation" in the case of 3D), B-spline interpolation, and C-spline interpolation [99]. Higher-order interpolations often lead to higher computational complexity, thus in practice there is often a trade-off between quality and speed.

Based on the nature of comparison, similarity metrics can be divided into intensity-based and feature-based metrics. Widely-used similarity metrics include the sum of squared differences [53], cross correlation [127], normalised cross correlation [127], mutual information [152, 147], normalised mutual information [141], joint entropy [128] and label consistency [18]. Supposing $I(v)$ denotes the intensity value of voxel v in image I, and $\Omega(I)$ denotes the set of all voxels in I or the ROI within it. The four most widely-used metrics are formulated as follows.

- **Sum of squared differences (SSD)**

$$\zeta_{SSD}(T, I_S, I_R) = \sum_{v \in \Omega(I_R)} \Big(I_S(T(v)) - I_R(v) \Big)^2 \tag{2.7}$$

  Sometimes the average value will be used instead of the sum, which turns it into another common metric: the mean of squared differences.

- **Normalised cross correlation (NCC)**

$$\zeta_{NCC}(T, I_S, I_R) = \frac{\sum_{v \in \Omega(I_R)} \big(I_S(T(v)) - \mu[I_S] \big) \big(I_R(v) - \mu[I_R] \big)}{\sqrt{\sum_{v \in \Omega(I_R)} \big(I_S(T(v)) - \mu[I_S] \big)^2 \sum_{v \in \Omega(I_R)} \big(I_R(v) - \mu[I_R] \big)^2}} \tag{2.8}$$

where $\mu[I] = \frac{1}{|\Omega(I)|}\sum_{v\in\Omega(I)} I(v)$ is the average intensity value of image I. This is a normalised version of the cross correlation metric [127].

- **Mutual information (MI)**

$$\zeta_{MI}(T, I_S, I_R) = \sum_{s\in B_S}\sum_{r\in B_R} p(T,s,r) log_2\Big(\frac{p(T,s,r)}{p_R(r)p_S(T,s)}\Big) \tag{2.9}$$

where B_S, B_R are, respectively, the sets of bin centres of the intensity histograms of I_S, I_R, and $p_S(T,s)$, $p_R(r)$, respectively, denote the marginal probability distribution functions of the two images. $p(T,s,r)$ denotes the joint probability distribution function:

$$p(T,s,r) = \frac{1}{|\Omega(I_R)|}\sum_{v\in\Omega(I_R)} w_S\Big(\frac{s - I_S(T(v))}{\delta_S}\Big) w_R\Big(\frac{r - I_R(v)}{\delta_R}\Big) \tag{2.10}$$

where $w_S(\cdot)$ and $w_R(\cdot)$ indicate the source and reference B-spline Parzen Windows, and the scaling constants δ_S and δ_R equal to the bin widths of B_S and B_R. The number of bins to create the histograms is a tunable parameter adjusted by the user.

- **Normalised mutual information (NMI)**

$$\zeta_{NMI}(T, I_S, I_R) = \frac{\sum_{s\in B_S}\sum_{r\in B_R} p(T,s,r) log_2\Big(p_R(r)p_S(T,s)\Big)}{\sum_{s\in B_S}\sum_{r\in B_R} p(T,s,r) log_2 p(T,s,r)} \tag{2.11}$$

This is essentially a normalised version of the MI metric.

The SSD and NCC measures are usually used for images of the same modality. The SSD measure is only suited in the case that the two images have a compatible intensity distribution, such as the Hounsfield scale in the CT modality. The NCC measure is less strict and only assumes a linear relationship between the intensity values of I_S and I_R, and is thus applicable to, for example, two MR images of different intensity scales. The MI and NMI are more general and rest only on the assumption of a relation between the intensity probability distributions of the two images in the information-theoretic sense. More intuitively, they are used to measure the ability of one image to explain the other, and are thereby suitable for both mono- and multi-modality image registration.

Optimisation method

To solve the registration problem Eq. (2.1), an iterative optimisation strategy is often employed:

$$T_{i+1} = T_i + \eta_i \cdot d_i \tag{2.12}$$

where d_i and η_i, respectively, indicate the search direction and step size for optimisation at iteration i. Common optimisation methods include gradient descent, quasi-Newton, non-linear conjugate gradient and Robbins-Monro algorithms [84]. Among them, the gradient descent method is probably most widely used. It takes the opposite of the gradient of the cost function as the search direction at each iteration, and is formulated as:

$$T_{i+1} = T_i - \eta_i \cdot \frac{\partial f(T_i, I_S, I_R)}{\partial T_i} \tag{2.13}$$

Furthermore, in modern applications, a stochastic gradient descent technique is often utilised to approximate this procedure in an efficient manner. Instead of taking each optimisation step using all voxels in the image/ROI as in the standard gradient descent, the stochastic scheme performs optimisation on a small subset (for example 2000 voxels) randomly sampled from the image/ROI. It has widely proven to be able to improve the registration process significantly without compromising registration accuracy [84].

2.2.2 The label propagation framework

The label propagation framework is centred on the employment of non-rigid registration to establish a voxel-level one-to-one correspondence between the atlases and the target image. Although an accurate registration may be achieved more easily between images of similar subjects, it is far more challenging on imaging data that manifest substantial inter-subject variability: unfortunately this condition applies to most real-world scenarios. An inaccurate registration could cause significant misalignment of anatomical structures, leading to incorrect label propagation. To improve segmentation performance, state-of-the-art methods often use

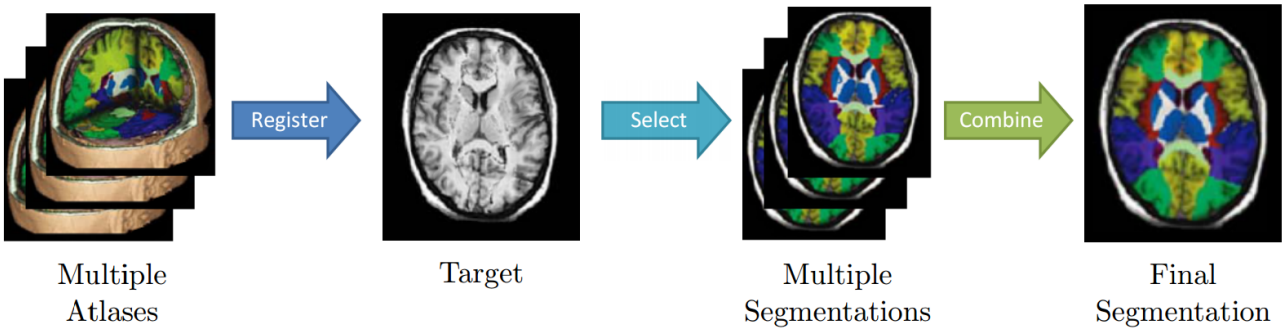


Figure 2.6: Multi-atlas label propagation (image source: [159])

a number of atlases at the same time to minimise the impact of registration errors, which is widely known as "multi-atlas label propagation" (MALP). Many MALP algorithms have been proposed in the literature, generally based on a computational pipeline illustrated in Figure 2.6. The variations are primarily based on the atlas selection and label fusion methods, in addition to the registration methods discussed earlier. A comprehensive survey has been conducted by Iglesias and Sabuncu [71].

Atlas selection

Suppose there are N atlases in total collectively represented by U, and denote an atlas as $(I_a, L_a) \in U$ with an ID number a and the target image as I_t. In terms of segmentation accuracy, relevant studies [4] have demonstrated that MALP using a properly selected subset of atlases (denoted as U_K) often yields a better performance than using the whole atlas set, or a single best atlas, or a randomly selected subset. This is because the aggregate registration error can be reduced in this case, making label classification more robust, especially when label fusion is based on majority voting. The issue of atlas selection on the other hand, often boils down to two factors: the number of atlases to use (denoted as K) and the selection criteria.

In terms of selection criteria, some similarity metrics are generally used to to rank the relevance of atlases. Popular metrics in image registration such as the SSD and NMI are often borrowed and calculated on a voxel-to-voxel basis over the whole image or an ROI. The exact choice is normally application-dependent, for example, SSD is not suitable for images without proper intensity normalisation, whereas NMI is more robust to differential levels of intensity contrast

and appearance, but is less sensitive to small local differences.

Moreover, based on the stage to perform selection, it can be further divided into pre-registration and post-registration selections. In general, post-registration tends to be more effective and widely-used, but requires a separate registration between every target-atlas pair, which significantly increases the computational complexity. Several survey studies have been conducted to comparatively evaluate common post-registration methods in the literature [4, 130]. In contrast, pre-registration selection is generally considered much more challenging, less effective and also less well-studied by far, with the existing work mostly based on content-based image retrieval. A review has been carried out by Akgul et al. [2]

Label fusion

Denote $(I'_a, L'_a) \in U'$ as the atlases warped to the target space, and U'_K as the collection of selected atlases. Label fusion (also known as "decision fusion") is the process to derive the final segmentation L_t using U'_K. In principle, it is similar to the traditional classification problem with multiple independent classifiers, since each atlas generates a separate label proposal. The simplest fusion strategy is probably *majority voting*, in which each selected atlas casts a single vote and the label class that wins the most votes is chosen as the final label of the target voxel. Majority voting is one of the first and most widely-used fusion schemes, easy to implement, with its effectiveness demonstrated in many applications [64, 85, 36, 65, 56]. However, since transformation is continuous and voxels generally do not align perfectly, it often leads to several neighbouring label proposals in the same atlas bidding for the same target voxel.

To address this issue, an interpolation scheme is often employed. Common schemes, as described in Section 2.2.1, include nearest neighbour interpolation, linear interpolation and B-spline interpolation. Except for the nearest neighbour scheme, partial volume interpolation is realised in most other schemes, in which a portfolio of label proposals is included for *weighted voting*. A representative weighted voting approach is to directly sum up the weights per label class derived from U_K, and pick the class with the highest weight as the final label, which is often known as the *sum rule* [129]. Alternatively, label proposals may be weighted on intensity-based

similarity, such as the NMI, between the atlases and target image [6], or label-based similarity, for example, by pair-wise comparison between the atlases only [40], or other weighting schemes. Moreover, weighting may be carried out at the global, regional or local level [71, 7, 102].

Another popular label fusion technique is the simultaneous truth and performance level estimation (STAPLE) method [160], which alternates between estimating the consensus segmentation and the reliability of contribution by each atlas, based on the well-known expectation-maximisation (EM) framework [43]. In each iteration, the provisional consensus segmentation is used to estimate contribution reliability, and measure atlas weights based on individual sensitivity and specificity. Subsequently, the derived weights are then used to update the estimate of consensus segmentation, which triggers another iteration. In that regard, the STAPLE method enables atlas weighting with minimal impact of dissimilar atlases. It is particularly effective when there is a large variation in the atlases.

2.2.3 Notable variants

Label propagation using composite transformation

A high accuracy has been reported in many studies using segmentation approaches based on the MALP framework. However, a major drawback which limits its wider applicability is its high computational burden. In a typical MALP application, non-rigid registration needs to be performed between every atlas $(I_a, L_a) \in U$ and the target image I_t in order to estimate a separate transformation. This will require N times of independent registration processes to segment each single image, which is extremely time-consuming, especially with a large set of atlases. For example, segmenting a single brain image using state-of-the-art MALP methods with 15 atlases in practice could take up to several hours to complete [91].

As a remedy, some researchers proposed a more efficient approach using composite transformation [108, 109]. Instead of a direct registration with the target image, in this approach all atlases are registered with an intermediate template in advance, each generates an atlas-template transformation $T_a^{'}$. At test-time, only a single run of non-rigid registration needs to

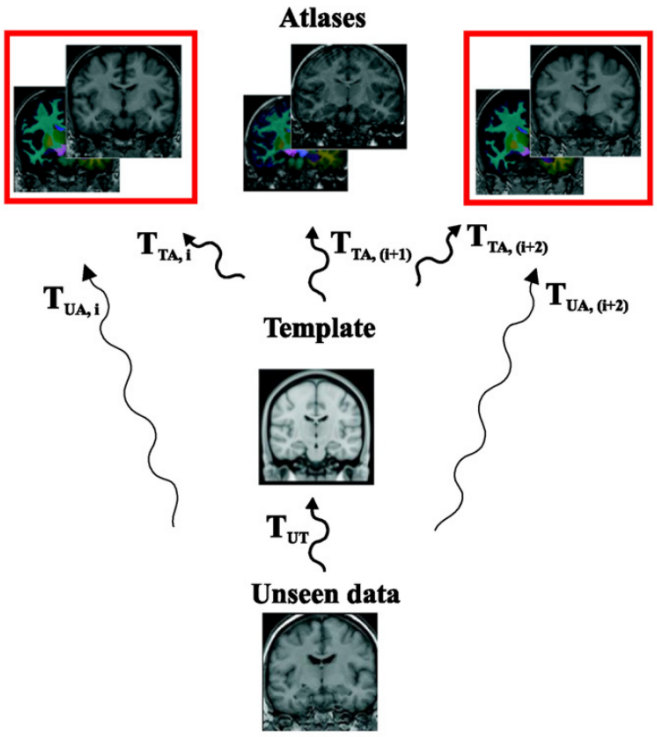


Figure 2.7: Label propagation using composite transformation (image source: [109])

be carried out to estimate the transformation T_0 between the template and the target image, followed by the concatenation $T_a = T_a' \circ T_0$ to perform label propagation on each atlas, as shown in Figure 2.7. This approach could approximately reduce computation time to $1/N$ of that using a standard MALP method, dramatically boosting segmentation efficiency. Although theoretically sound, composite transformation however is prone to errors in practice. Moreover, it also raises the issue concerning the selection/creation of an ideal template to secure a good composite transformation, which is often not only application-specific but also data-dependent [76].

Segmentation with a probabilistic atlas

Another approach that avoids pair-wise registration between every atlas and the target image is to use a probabilistic atlas. A probabilistic atlas is usually created using the segmentations of multiple subjects based on a group-wise non-rigid alignment in a common template space [123], for example the MNI152 template [3] widely used in brain MRI applications. More specifically, a

[3]MNI152 brain MRI template: `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases`

probabilistic atlas provides voxel-wise varying prior information, which, when propagated to the target image via non-rigid registration, can be used in combination with a range of parametric statistical models to perform label inference [123, 52, 12, 125] .

A popular approach, for instance, is to combine prior information such as label probability statistics, with intensity models such as the Gaussian mixture model, which captures each tissue class with a separate Gaussian distribution. Parameter optimisation may be achieved using the EM algorithm [43], which tries to maximally explain the intensities in I_S using the estimated model. A typical work is the segmentation of white matter, grey matter and cerebrospinal fluid in brain MR images [98], where each tissue class is captured separately in the Gaussian mixture model, and a Markov random field is used to correct intensity inhomogeneity occurring during image acquisition. An accurate and smooth segmentation for each tissue class was obtained in that study, and a number of related methods have since been proposed in other studies [52, 107].

In these approaches, normally only a single registration process is required, in order to warp the target image to the template space, or the other way round. However, similar to label propagation using a composite transformation, the improved efficiency is often achieved at a certain level of compromise on segmentation accuracy. In addition, registration performance is critical to both the creation of the probabilistic atlas and the quality of prior information for segmentation, and securing a suitable template remains a challenge, pending on the actual level of inter-sample variability.

2.3 Segmentation via patch-based pattern matching

In this paradigm, the strict voxel-level one-to-one correspondence for label propagation between the target image and each atlas is relaxed to patch-based pattern matching, and therefore it is often known as "patch-based segmentation" (PBS). A patch (sometimes also called "template") is usually a 2D or 3D box centred at the voxel under study, containing its contextual information. For each patch in the target image, a patch search process will be performed to retrieve similar

patches from the atlases, followed by label fusion using the retrieved patches to classify the centre voxel of the target patch. Often, a group-wise affine registration will be applied in advance, to establish an overall alignment between the target image and atlases, in order to enhance the patch search process.

Patch matching and label fusion are often carried out using some hand-engineered features and similarity metrics, for example based on the SSD. Sometimes, some image pre-processing may be included to enhance pattern matching. In particular, for pattern matching based on the SSD metric, a tissue-standardising normalisation [121] is often applied to MRI data to regulate and ensure the compatibility of intensity scales across all images. Other standard pre-processing techniques including image denoising [38] and inhomogeneity correction [139] are also frequently employed as a preliminary step to improve MR image quality.

Comparing to the label propagation approach, the key differences of the PBS framework include:

1. *Localised similarity measure*: in the label propagation framework, local disparities are aggregated to an overall cost function in the registration process. By contrast, in the PBS framework, a patch is essentially a local context descriptor, and the similarity measure is localised to the patch level.

2. *Voxel-wise optimisation*: In image registration, the optimisation of cost function is realised using model parameters and is subject to the corresponding degree of freedom, which may be fine-grained to the control point resolution level (using the B-spline model) at the best. In the PBS framework, segmentation is optimised voxel-wise and carried out independently from one voxel to another.

3. *Pattern matching beyond immediate neighbourhood*: in the label propagation framework, pattern matching is based on information within the vicinity of a target voxel. In the PBS framework, it is well extended beyond the immediate neighbourhood, where patch search was carried out over a large search window, or even an entire ROI.

4. *Increased proposals for label fusion*: in the label propagation framework, each (selected) atlas typically generates a single label proposal per voxel for label fusion. Whereas in

the PBS framework, a large and customisable number of patches are retrieved from the atlases, generating many more proposals for label fusion. A notable point is that the patches in atlases are collectively addressed at the same time, rather than atlas-by-atlas.

2.3.1 The patch-based segmentation framework

The PBS framework was originally proposed by Coupe et al. [37] and Rousseau et al. [133] independently around the same time, and was, respectively, applied to the segmentation of hippocampus and lateral ventricle, and to the segmentation of tissues and a range of anatomical structures. Both studies were validated on a set of human brain MRI data. This framework primarily consists of three components to build up its computational pipeline: atlas selection, patch search and label fusion, as illustrated in Figure 2.8.

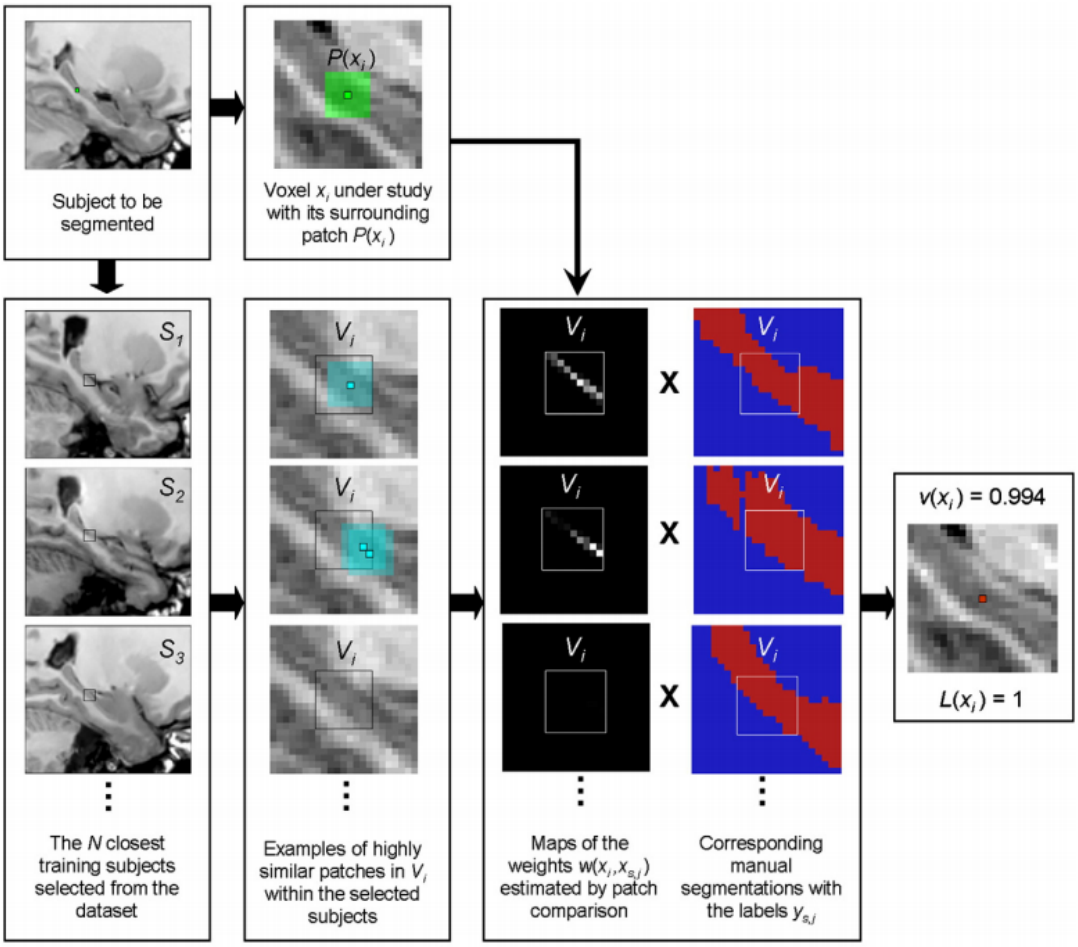


Figure 2.8: Overview of the standard patch-based segmentation framework (image source: [37])

Atlas selection

Similar to the label propagation framework, segmentation performance is usually improved with the use of a properly selected subset of atlases U_K, rather than all atlases U collectively. Yet unlike the label propagation framework where atlas selection is generally performed after a time-consuming pair-wise non-rigid registration, the selection process is carried out at a much earlier stage in the PBS framework. In the work of Coupe et al. [37], an ROI that covers only the hippocampus, lateral ventricle and their immediate neighbourhoods was applied as an initialisation mask to the linearly aligned images. Atlas selection was then performed by identifying the most similar K atlases based on the SSD metric computed over the ROI mask. Other popular selection strategies discussed in relevant studies [4, 130] may be employed instead, including the use of NCC or MI metrics.

Patch search

Denote $P(v) \in \mathbb{R}^{p \times p \times p}$ as a patch (typically sized from $3 \times 3 \times 3$ to $9 \times 9 \times 9$) centred at voxel v in the target image, $P_a(v)$ as its counterpart at the corresponding location in atlas (I_a, L_a) (note all images are linearly aligned), and $S_a(v) \in \mathbb{R}^{s \times s \times s}$ as the search window (typically sized from $3 \times 3 \times 3$ to $15 \times 15 \times 15$) deployed around v in the atlas space. The purpose of the patch search algorithm is to efficiently retrieve a set of matching patches Υ_*, from a patch library $\Upsilon = \{P(v'), l' \mid v' \in S_a(v), l' = L_a(v'), (I_a, L_a) \in U_K\}$ created using all selected atlases U_K.

In the work of Coupe et al. [37], the patch search algorithm was simply a brute-force search over Υ, screening out dissimilar patches using the structural similarity measure [157]:

$$ss(P(v), P(v')) = \frac{2\mu[P(v)] \cdot \mu[P(v')]}{\mu[P(v)]^2 + \mu[P(v')]^2} \times \frac{2\sigma[P(v)] \cdot \sigma[P(v')]}{\sigma[P(v)]^2 + \sigma[P(v')]^2} > \tau \tag{2.14}$$

where $\mu[P(v)]$ and $\sigma[P(v)]$, respectively, indicate the mean intensity and standard deviation of patch $P(v)$, and all training patches $(P(v'), l') \in \Upsilon$ with $ss(P(v), P(v'))$ not above the given threshold τ are to be discarded. In their experiment, τ was manually set to 0.95 based on empirical experience. Other patch search algorithms will be discussed in Section 2.3.2.

Label fusion

Once patch search is complete, a weighted label fusion process will be carried out using all the retrieved patches $(P(v'), l') \in \Upsilon_*$ in the form of pair-wise comparison. Label weighting is typically carried out based on the SSD metric. For instance, in the work of Coupe et al. [37], it is defined by:

$$w_l(P(v), P(v')) = \begin{cases} \exp\left(\frac{-||P(v)-P(v')||_2^2}{h(v)}\right), & \text{if} \quad l = l' \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

where $||\cdot||_2$ represents the element-wise L2 norm and $||P(v) - P(v')||_2^2$ is equivalent to the SSD between $P(v)$ and $P(v')$ in matrix form. The $h(v)$ is a local adjustment factor borrowed from relevant work [111] and is defined by:

$$h(v) = \min_{v'} . \, ||P(v) - P(v')||_2 + \epsilon \qquad \forall (P(v'), l') \in \Upsilon_* \tag{2.16}$$

where ϵ is a small constant to ensure numerical stability. As the last step, the target voxel v is labelled with the label class carrying the highest aggregate weight:

$$L(v) = \arg\max_{l} . \sum_{(P(v'),l') \in \Upsilon_*} w_l(P(v), P(v')) \tag{2.17}$$

2.3.2 Notable variants

Despite its excellent segmentation accuracy, the standard PBS method has been widely criticised for its low computational efficiency. Segmentation of an image may take hours to complete when applied, for instance, to the whole brain. The key bottleneck lies on the lack of efficient patch search and label fusion techniques. A number of approaches have since been proposed to improve its efficiency, such as a multi-resolution approach that gradually refines segmentation from a coarser to a finer level [49, 158], efficient patch search using search trees [156], random forests [179, 180, 166, 87, 88], or the PatchMatch algorithm [143].

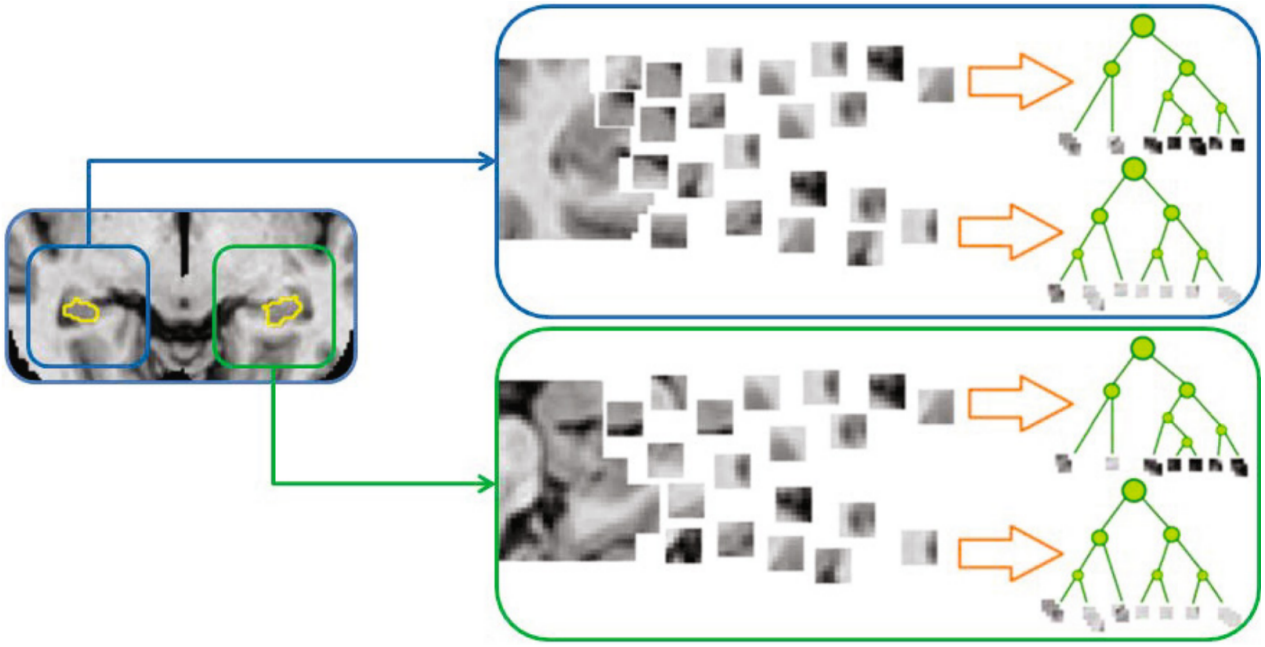


Figure 2.9: Example of patch search using tree models: a search tree is created per label class with training patches drawn from each ROI of each atlas (image source: [156])

PBS with search trees or random forests

Probably the most straightforward idea is to use search trees to facilitate patch retrieval. More specifically, by employing a tree-based search mechanism, patch search (at test-time) no longer requires time-consuming exhaustive search across the whole patch library for pair-wise similarity measure, but through a number of quick binary tests instead. Figure 2.9 illustrates an example idea of tree construction for patch search. Notable studies in this line include the use of ball trees [156], atlas forests [179, 180], patch forests [166] and neighbourhood approximation forests [87, 88]. The variations are primarily centred on the methods regarding tree creation, patch search and label fusion.

For instance, in one of our earlier studies, we proposed an augmented PBS method with an efficient patch search technique using random forests (called "patch forests") [166]. Each atlas (I_a, L_a) is used to train a separate patch forest that contains n number of trees: $PF_a = \{t_a^{(1)}, .., t_a^{(n)}\}$. Once trained, each tree $t_a \in PF_a$ is able to efficiently classify/navigate a target patch $P(v)$ to a leaf node, which stores a collection of training patches drawn from (I_a, L_a), denoted by $t_a[P(v)]$, where each $(P(v'), l') \in t_a[P(v)]$ is similar to $P(v)$ in appearance with label information. Starting from the root, each splitting node is associated with a binary test

composed of a feature projection function $f \in F :: \mathbb{R}^{p\times p\times p} \rightarrow \mathbb{R}$ and a threshold $\tau \in \mathbb{R}$. The tests are applied to $P(v)$ and redirect it to the left/right child for further processing, until a leaf node is reached. Since all training patches in the same leaf node have passed the same set of tests, they are considered similar to $P(v)$. With n trees in PF_a, it then outputs $\Upsilon_a = t_a^{(1)}[P(v)] \cup .. t_a^{(n)}[P(v)]$ as the search outcome, and the patches are ranked by the frequency included in different tree search results:

$$\Gamma_{(P(v'),l')} = \frac{|\{t_a|(P(v'),l') \in t_a[P(v)],\, t_a \in PF_a\}|}{n} \qquad (P(v'),l') \in \Upsilon_a \tag{2.18}$$

Suppose there are K atlases used for segmentation $PF = \{PF_1, .., PF_K\}$, which generate $\Upsilon_1 \cup .. \Upsilon_K$. Then the k patches with highest $\Gamma_{(P(v'),l')}$ scores across all atlases are then selected (denoted by Υ_*) for label fusion, using the weighting scheme:

$$w_l(P(v),P(v')) = \begin{cases} \exp\Big(-||P(v)-P(v')||_2^2\Big), & \text{if} \quad l = l' \\ 0, & \text{otherwise} \end{cases} \tag{2.19}$$

which is similar to Eq. (2.15), but without the need of $h(v)$. As the last step, the target voxel v is labelled with the class carrying the highest weight, as in Eq. (2.17).

In terms of forest training, each tree is trained independently with the same objective function at each splitting node:

$$\min_{f,\tau} . \quad \frac{|\Psi_{sL}|}{|\Psi_s|} C(\Psi_{sL}) + \frac{|\Psi_{sR}|}{|\Psi_s|} C(\Psi_{sR}) \qquad \text{s.t.} \quad \Psi_s = \Psi_{sL} + \Psi_{sR} \tag{2.20}$$

based on the compactness measure borrowed from other random forest studies [88]:

$$C(\Psi_s) = \frac{1}{|\Psi_s|^2} \sum_{P(v_1),P(v_2)\in\Psi_s} \rho(P(v_1),P(v_2)) \tag{2.21}$$

where Ψ_s, Ψ_{sL}, Ψ_{sR}, respectively, represent the set of training patches arriving at the node, and the subsets to the left and right children. $\rho(\cdot,\cdot)$ is a customisable distance measure, which in our work was modelled as a combination of squared SSD and spatial regularisation with

borrowed insights from related work [156]:

$$\rho(P(v_1), P(v_2)) = ||P(v_1) - P(v_2)||_2^2 + \alpha \cdot ||v_1 - v_2||_2^2 \tag{2.22}$$

where α is a weighting parameter to adjust $||v_1 - v_2||_2^2$, the squared Euclidean distance between voxel coordinates v_1 and v_2. Intuitively, tree training is centred on finding a parcellation that minimises data compactness, in terms of both intensity information and spatial distance. When an optimal parcellation ends up with $|\Psi_{oL}| = 0$ or $|\Psi_{oR}| = 0$, the node becomes a leaf node. The detailed training algorithm is based on a heuristic technique with iterative randomisation of f and τ, which will not be elaborated here.

PBS with PatchMatch

Another notable variant is to combine PBS with the PatchMatch algorithm. PatchMatch was originally proposed by Barnes et al. [11] for structural editing of real-world images, and was later adapted as a fast patch search technique to improve the PBS efficiency [143]. The algorithm consists of three steps: initialisation, propagation and random search. In the PBS adaptation [143], it is initialised by randomly associating each patch in the target image to another in an atlas image, within a constrained window. Subsequently, a series of propagation and random search steps are iteratively carried out in alternation to improve the match.

Denote $\Theta[P(v)]_i = P_a(v')$ as the PatchMatch function, which associates $P(v)$ with $P_a(v')$ in atlas (I_a, L_a) at iteration i. In the propagation step, supposing $\Theta[P(v+d)]_i = P_b(v''+d)$ is the current match of an offset patch $P(v+d)$, a provisional match between $P(v)$ and $P_b(v'')$ will be examined to check whether a better match can be established in comparison to the original match. In their work, the SSD was simply used as the distance metric for patch comparison, and the offsets were restricted to the six adjacent patches with $d \in \{\langle \pm 1, 0, 0\rangle, \langle 0, \pm 1, 0\rangle, \langle 0, 0, \pm 1\rangle\}$. Then, out of the seven proposals derived by all adjacent patches and the original match, $\Theta_{ps}[P(v)]_{i+1} = P_e(v''')$ is updated by the one with the lowest SSD. A notable point is that an updated match could be a patch in a different atlas: it is not necessary for $a = e$.

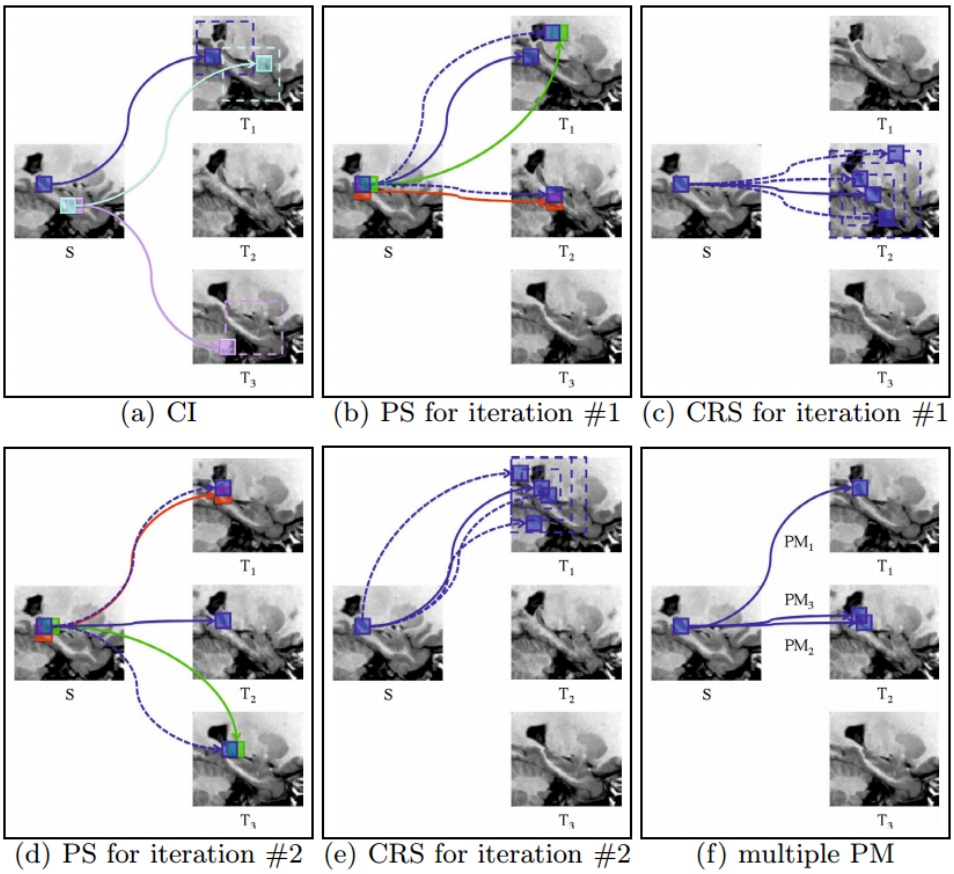


Figure 2.10: PBS with PatchMatch (image source: [143]): the CI here stands for constrained initialisation, PS stands for a propagation step, CRS stands for a constrained random search, and PM stands for a separate PatchMatch instance

In the random search step, supposing $\Theta_{ps}[P(v)]_{i+1} = P_e(v''')$ is the updated match in the propagation step, a set of offset patches beyond adjacency will be randomly sampled from the same atlas (I_e, L_e), constrained within a gradually narrowing search window around v''', in order to explore a better match $\Theta_{crs}[P(v)]_{i+1} = P_e(v'''')$. This then triggers another iteration, until the termination condition is met. The algorithm is often set to terminate after around five iterations [11, 143]. Figure 2.10 (a-e) illustrates an example procedure from an initialisation step, to two iterations of propagation and random search.

In addition, a parallel setting is employed to improve both computational efficiency and segmentation accuracy, with k independent instances of PatchMatch running at the same time, as shown in Figure 2.10 (f). As the last step, segmentation is completed with the same label fusion process as in the standard PBS method, using the final match $\{\Theta_*[P(v)]^{(1)}, .., \Theta_*[P(v)]^{(k)}\}$ to classify each target voxel v.

2.4 Deep learning and the Proposed Approach

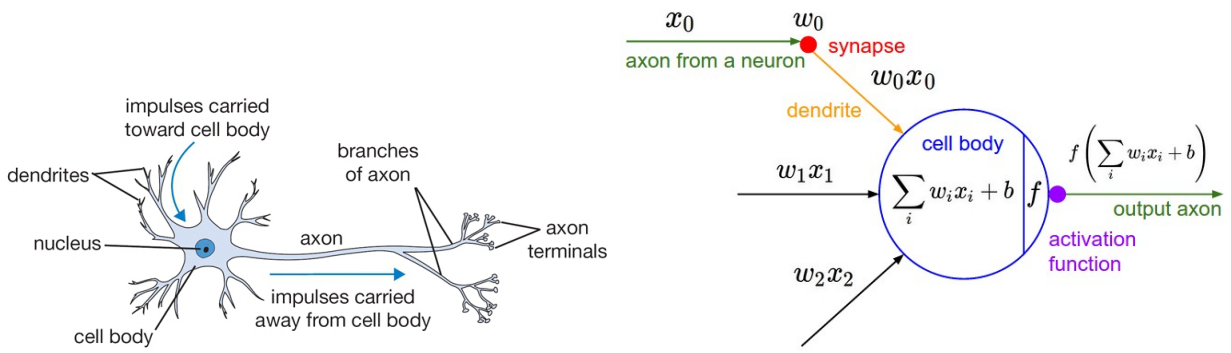
Our approach carries on the line of research on patch-based segmentation by incorporating deep learning. In contrast to the use of low-level hand-engineered features (such as the SSD) for patch search and label fusion in the conventional PBS approaches, we train a patch-based deep neural network (PatchDNN) that serves as both an advanced feature extraction mechanism and a classifier. The PatchDNN takes patch data as input and projects it to a high-level feature space through a set of deeply learned non-linear functions, followed by a simple softmax classification. In this case, segmentation at run-time becomes extremely fast and highly accurate, without the need to explicitly re-use the atlas data for further processing.

Moreover, sometimes there is a variety of manual annotation protocols used in different studies, leading to certain levels of discrepancy on the reference segmentation, such as the case of the hippocampus segmentation work. Although these are mitigated by recent protocol standardisation practice [22], they can still cause considerable complexity, when adapting a signal processing-based segmentation algorithm to multiple datasets obtained from earlier studies. By contrast, with our approach, it simply requires a network re-training process, without significant human involvement.

2.4.1 The neural network framework

Mainstream studies into artificial neural networks (ANN) date back to the 1980s, when researchers started to train multi-layer computational architectures with the back-propagation technique [136]. Since then, there has been only limited success, due to the complexity of network training and limited computing power, until the advent of modern GPUs and simplified GPU programming techniques. Since its revival in the late 2000s, deep learning has brought about a revolution in computer vision, speech recognition, natural language understanding, sentiment analysis and so forth: LeCun et al. [94] have carried out an insightful review.

The development of the ANN architecture was originally inspired by the simulation of biological neural systems. In a typical neural system, there are billions of neurons intricately connected



(a) biological neuron (b) computational neuron

Figure 2.11: Computational simulation of a biological neuron (image source: `http://cs231n.github.io/neural-networks-1`)

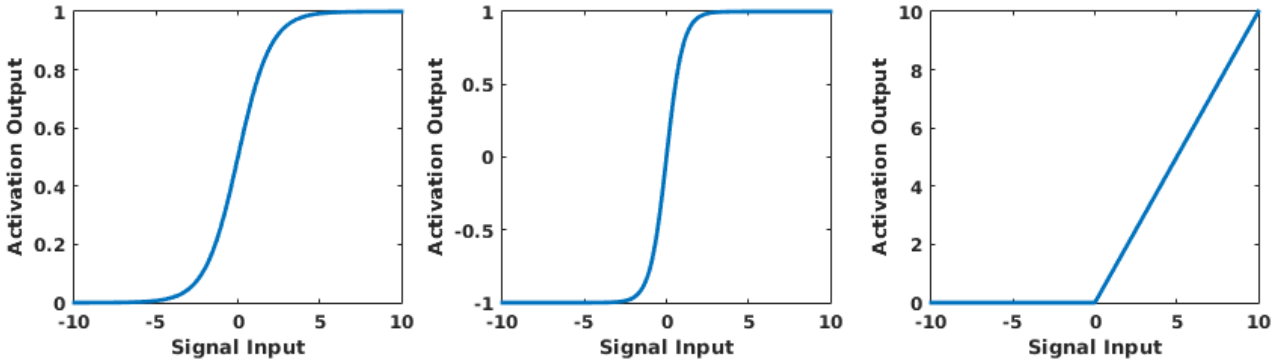
through an extremely large collection of synapses. Each neuron receives input signals from its dendrites and produces output signals in its axon, which then channels out to other neurons via the synaptic connections. In a similar course, an ANN consists of a number of inter-connected layers, each is composed of a set of independent computational units named after the "neurons" (sometimes also called "nodes" or "modules"). Figure 2.11 illustrates the basic idea regarding the computational simulation of a biological neuron. More specifically, each neuron receives a vector of signals $\boldsymbol{x}$ and outputs a scalar-valued response:

$$y = f(\boldsymbol{w} \cdot \boldsymbol{x} + b) = f\Big(\sum_i w_i \cdot x_i + b\Big) \tag{2.23}$$

where $\boldsymbol{w}$ is a vector of adjustable parameters (often called "weights") that "interacts" with the data, and is to be learned in the network training process; b is a bias term that differs from one neuron to another, and f is an activation function, which triggers signal emission in a particular pattern and is usually used to introduce non-linearity. Common activation functions include

- Sigmoid activation: $f_{sgm}(x) = \frac{1}{1+e^{-x}}$
- Tanh activation: $f_{tanh}(x) = \frac{2}{1+e^{-2x}} - 1$
- Rectified linear unit (ReLU) activation: $f_{relu}(x) = max(x, 0)$

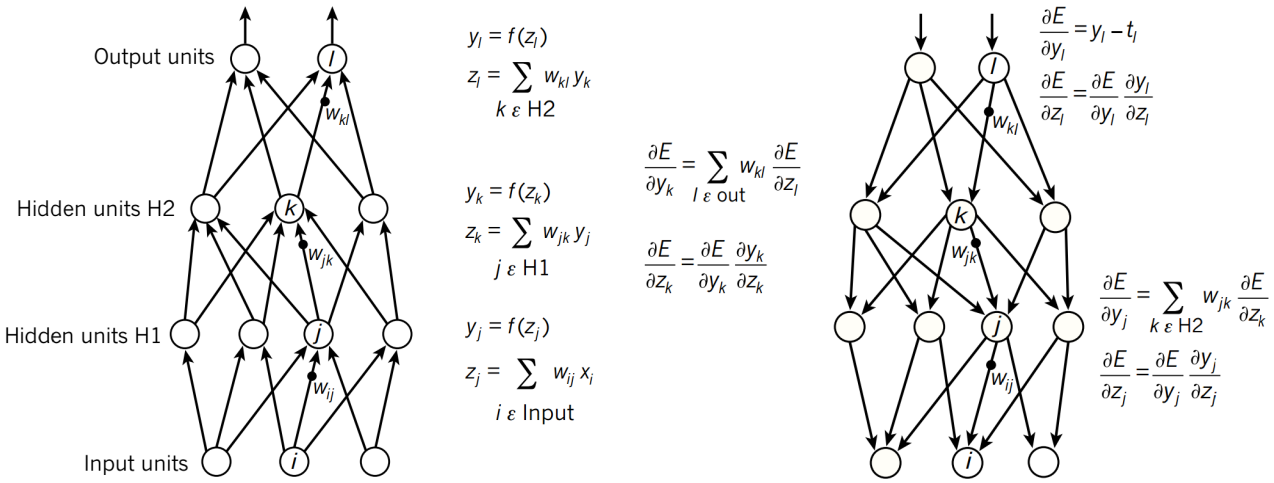
The graphs of these activation functions are shown in Figure 2.12 in a comparative fashion.



(a) Sigmoid (b) Tanh (c) ReLU

Figure 2.12: Graphs of the (a) Sigmoid, (b) Tanh, and (c) ReLU activation functions

In a more abstract sense, each neuron is a feature extraction function that projects input signals over a feature dimension and generates a feature response. In fact itself can be used for classification directly, but the feature representation capability of a single neuron is limited. The ANN model extends this property by constructing a multi-layer architecture, where the input signals can be fed forward layer-after-layer, as the example shown in Figure 2.13 (a). By harnessing the property of feature composition, the aggregate function can become very intricate and achieve an excellent capacity of feature representation. Finally at the output layer, the network returns a vector of class scores, and then the input data is normally labelled by the class with the highest score.



(a) Feature composition via forward propagation (b) Network training via backward propagation

Figure 2.13: Artificial neural network architecture (image source: [94])

Network training, on the other hand, is through backward propagation, by employing the chain rule of derivatives. In a supervised learning scenario, the provisional classification output will be compared with the ground truth, and an objective function will be formulated to measure the level of discrepancy. Then the classic gradient descent technique is utilised to reduce the discrepancy by making a minor adjustment on the weights in the opposite direction to their gradients. The adjustment at the output layer is then propagated backwards, layer-after-layer, until reaching the input layer, as shown in Figure 2.13 (b). This procedure is typically repeated for millions of times before the network converges to an optimal solution that yields an excellent classification accuracy.

2.4.2 Convolutional neural network

We call the ANN model described above the canonical network or CanonNet architecture. In image classification problems however, an image may easily scale to millions of pixels/voxels, leading to millions of learning weights for a single neuron, and there could be millions of neurons in a network. A CanonNet in this case, often quickly ends up over-fitting. As a result, a variant model, the convolutional network or ConvNet has become particularly popular.

Instead of "interacting" with an entire image directly, ConvNet neurons (in a convolutional layer) filters it by convolution, in which each neuron only "interacts" with a small local patch of the input image at a time, and produces a feature image rather than a scalar. Each convolutional layer contains a collection of neurons, each uses a separate convolutional kernel with different weights but of the same size, and all neurons collectively output a multi-channel feature image for further processing. A notable point is that such local connectivity dramatically reduces the number of weights to be learned in each neuron and hence the whole network. Activation functions such as ReLU are often incorporated as well. An example ConvNet application to the classification of a 2D RGB image of a Samoyed is illustrated in Figure 2.14.

Deep ConvNets are able to achieve advanced feature extraction while using far fewer learning weights, making them more robust and popular than CanonNets for image classification. In particular, since the ImageNet challenge 2012 when the AlexNet [90] significantly outperformed

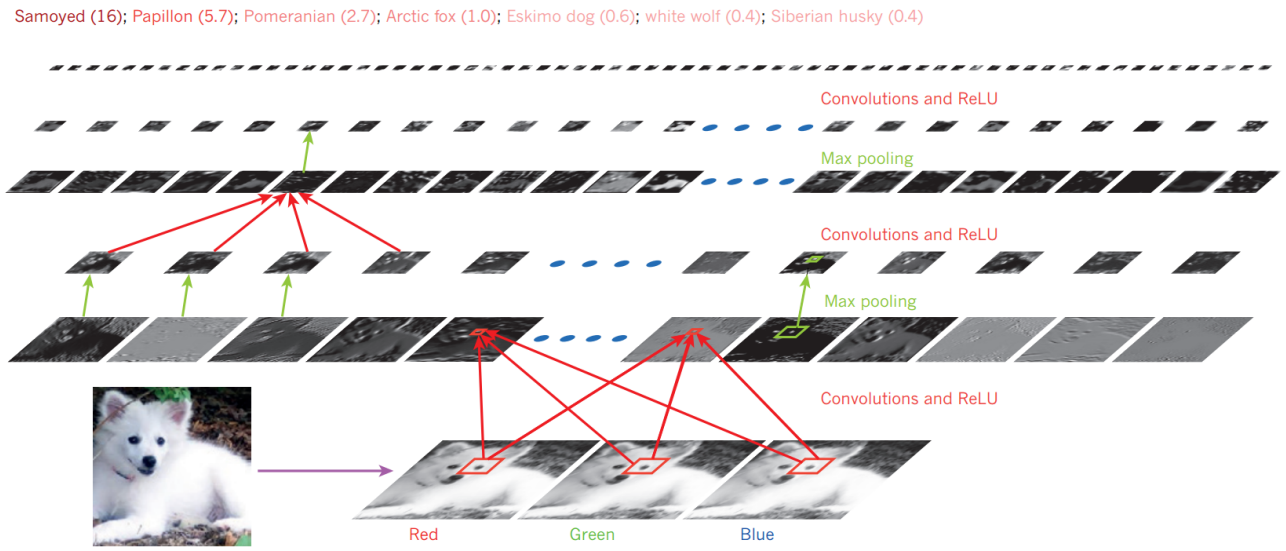


Figure 2.14: Convolutional neural network architecture (image source: [94])

the runner-up with almost half the error rate, the use of ConvNet has now become a dominant approach in computer vision. That accomplishment was primarily attributed to the efficient use of modern GPUs that significantly boosted computational efficiency, the ReLU activation functions that reportedly sped up network training by 5-6 times, and a new regularisation technique (called "dropout") [68] to reduce over-fitting.

Following the remarkable success in computer vision, although still at a limited scale, deep learning using ConvNets (in fact deep learning in general) are quickly gaining popularity in biomedical imaging, especially in the applications to image segmentation lately [131, 78, 132, 45, 126]. Typically, segmentation is broken down to voxel-wise labelling on a patch-based setting, where a patch is treated as a mini-image for the classification of its centre voxel. However, ConvNet classification generally involves millions of neuron-wise convolutions in each single run, requiring long training periods (for example, the AlexNet required 5-6 days of training on two modern GPUs [90]), as well as demanding memory consumption, since each neuron generates a separate feature image. In addition, existing ConvNet segmentation approaches are often based on a nesting structure that integrates with other models such as superpixels [132] or conditional random fields [78, 45], further complicating the computation.

2.4.3 Efficient biomedical image segmentation using a patch-based canonical neural network: the proposed approach

Unlike other deep learning work, in this study we abandon the popular ConvNet architecture. We argue that for patch (mini-image) based segmentation, the ConvNet's advantage is largely diminished, as a patch is only sized around 9×9 to 15×15, and the nature of the classification approach is that patches do not need decomposition and thus will not benefit from further convolution. Instead, we revisit the CanonNet architecture in which neurons only compute dot products, and are therefore much faster and less memory-hungry than convolution.

Furthermore, we substantially re-engineer the CanonNet architecture with state-of-the-art deep learning techniques, including use of ReLU activation and dropout layers. Moreover, we also adopt a 2.5D patch setting, which takes three 2D patches, respectively, from the axial, coronal and sagittal planes as the input. Such tri-planar patch setting is able to achieve comparable (sometimes even better) segmentation accuracies as the conventional 3D patch setting at a much smaller computational cost [126]. In an abstract sense, our PatchDNN model can be formulated as:

$$F = h\left(P_1(v), P_2(v), P_3(v)\right), \qquad V = softmax(F), \qquad l = \underset{c}{argmax}\left(V_c\right) \tag{2.24}$$

where $P_1(v)$, $P_2(v)$, $P_3(v)$ are the tri-planar patches for target voxel v, F is a feature vector generated by the feature extraction function $h(\cdot,\cdot,\cdot)$, V is a vector of label values obtained by the softmax classifier, and l is the output label, which is assigned to the label class c carrying the highest label value V_c. In addition, modern GPU programming techniques are also employed to improve the computational efficiency.

2.4.4 Network architecture

The network contains 6 feature extraction layers, 2 dropout layers, and a softmax layer at the end to take the aggregate features for classification, as shown in Figure 2.15. All layers

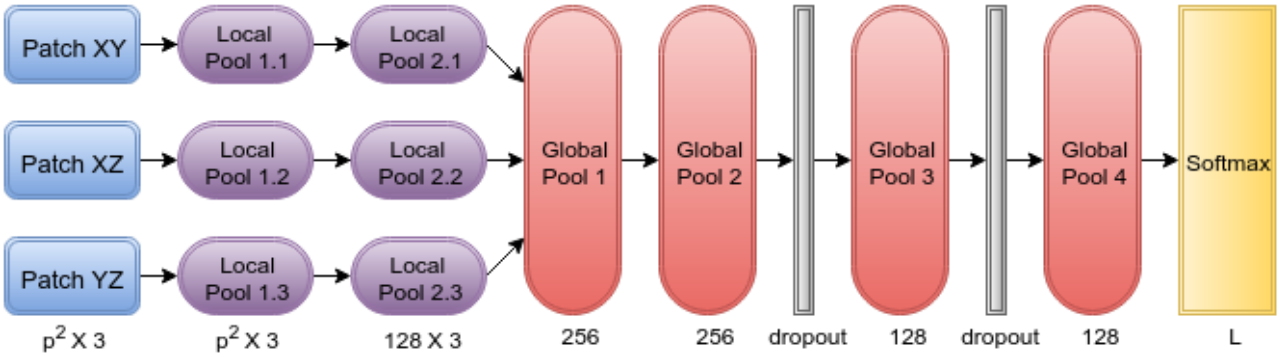


Figure 2.15: Architecture of the proposed PatchDNN model

preceding the softmax layer collectively model the $h(\cdot,\cdot,\cdot)$ in Eq. (2.24). Each neuron in a feature extraction layer, pairing with a subsequent ReLU activation, independently generates a feature response by the following function:

$$f^{(j)} = max\left(\sum w_i^{(j)} \cdot x_i + b^{(j)}, 0\right) \tag{2.25}$$

where $w_i^{(j)}$ and $b^{(j)}$ are the learnable weights for the ith entry of the (interim) input x at the jth neuron.

The ReLU introduces non-saturating non-linearity, which is able to speed up gradient descent training of large networks multiple times compared to the traditional Sigmoid and Tanh activation functions [90, 94]. Moreover, the first two feature extraction layers contain three pathways, respectively, for each of the tri-planar patches. Starting the third layer, all three pathways merge into one, which becomes a standard fully connected layer. Such design significantly reduces the number of weights to train compared to a scheme with full connectivity from the beginning, while retaining excellent feature representation capabilities. In addition, Layer 5 and 7 are dropout layers. A dropout layer randomly disconnects each neuron from the network at probability ϵ, which is typically set to 0.5 during training and 0 (no dropout) at test-time [68]. The purpose of dropout is to simulate the network as a combination of several networks trained separately, which can empirically reduce over-fitting.

2.4.5 Network training

Our framework performs a supervised learning scheme, where the tuple $\langle P_1(v),\ P_2(v),\ P_3(v)\rangle$ collectively represents a single training data entry d and the centre voxel label in the reference segmentation indicates its supervisory output l^*. In terms of cost function, we resort to the widely-used cross entropy:

$$\min_{\xi}. \quad E = H(\xi^{(k)}, X, Y) = -\sum_{o=1..n} \pi(l_o^*) \cdot log\ \xi^{(k)}(d_o) \tag{2.26}$$

where $X = [d_1, .., d_n]$ and $Y = [l_1^*, .., l_n^*]$, respectively, indicate the training data and ground truth information, $\pi(\cdot)$ denotes the probability distribution function of label classes, and $\xi^{(k)}(\cdot)$ denotes the network function that outputs classification scores, at training step k. Moreover, a number of leading-edge training techniques are also employed in this work:

Mini-batch stochastic gradient descent: The modern mini-batch gradient descent optimisation scheme is utilised to train the network. It takes a small batch of training data (denoted as X' and Y'), randomly sampled from X and Y with a customisable batch size δ, to perform optimisation in each step, instead of using the entirety of the training pool.

Reducing training pool with an ROI mask: Due to the structural congruity of anatomy, the target structures being segmented are spatially close in different images (after affine alignment). We therefore apply a mask over the corresponding ROI, created by taking the union of all foreground voxels in all atlases, followed with a minor dilation. The mask should be manually checked to ensure it fully covers the target structures and their immediate neighbourhoods in each image. This can reduce training pool to a small fraction, dramatically lowering the computational volume and memory consumption.

Foreground/background separate sampling: In segmentation studies, background labels often substantially outnumber foreground labels, even with the use of an ROI mask. For example, the ratio exceeded 10:1 in our application to hippocampus segmentation described in Section 2.4.6. This is often known as the "class imbalance problem" in machine learning

research, which could lead to a trained classifier over-fitting the background label class. To address this issue, for each mini-batch we draw half the samples from foreground and half from background. This enables each foreground data entry to be trained multiple times with different background counterparts, and was proven able to effectively improve segmentation accuracy.

2.4.6 Application to hippocampus segmentation

Although our approach is general, in this work a validation study was applied to the segmentation of the hippocampus in the human brain, which is a problem domain frequently visited in existing work on patch-based approaches [37, 143, 166, 151, 156], allowing for easy comparison. Hippocampus segmentation is a particularly challenging problem, due to its small size, high variability, low contrast, and discontinuous boundaries in conventional MR images [37].

Furthermore, another reason to carry out the validation study on hippocampus segmentation was due in part to the easy access to data, in particular the open source ADNI database[4] that our methods were tested on. The ADNI database contains a series of brain MR images acquired at regular temporal intervals from a population of around 800 people, including some 200 cognitively normal (CN) elderly individuals, 400 with mild cognitive impairment (MCI) and another 200 with confirmed Alzheimer's disease (AD). A more detailed description is available in relevant ADNI studies [117].

As a pilot study, 100 samples were randomly picked from the ADNI database. Among the test dataset, 34 subjects were CN controls, 33 subjects were confirmed AD patients and the other 33 subjects were diagnosed with MCI. The demographic profile is shown in Table 2.1, and a simple Student's t-test was conducted, concluding no statistically significant difference on age and MMSE score ($p\text{-}value > 0.1$) from the entire ADNI database, indicating that the sample pack was representative.

In addition, all the imaging data were acquired via the standard ADNI pipeline [73], and a reference segmentation of the hippocampus for each image was created semi-automatically using a

[4]The Alzheimer's Disease Neuroimaging Initiative (ADNI), official website: `http://adni.loni.usc.edu`

Table 2.1: Demographic profile of the test dataset

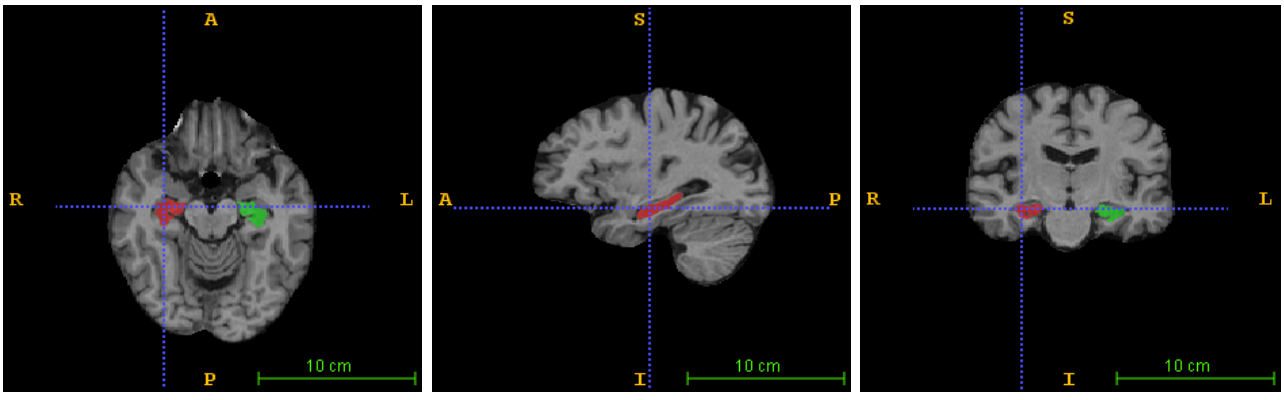
	Number	Age	MMSE score
CN	34	77.13 ± 6.10	29.77 ± 1.08 [26-30]
MCI	33	75.01 ± 7.35	27.40 ± 1.52 [24-30]
AD	33	76.28 ± 6.97	23.54 ± 1.88 [20-26]

high dimensional brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO), and later inspected and manually corrected by qualified reviewers [70]. Intensity inhomogeneity was also corrected using the well-known non-parametric non-uniformity normalisation (N3) technique [139]. An example brain image superimposed with its reference segmentation is shown in Figure 2.16.

2.5 Results

2.5.1 Experimental setting and evaluation method

Experimentation was carried out using cross-validation. The 100 images in the test dataset were divided into ten equally-sized folds via random distribution. A leave-one-out strategy was applied in each experiment, with nine folds to train a PatchDNN and the remaining fold for testing. In total, there were ten instances of network training and 100 instances of segmentation



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

Figure 2.16: An example brain MR image superimposed with its hippocampus reference segmentation in (a) axial view, (b) sagittal view, and (c) coronal view. The green and pink coloured regions, respectively, indicate the left and right hippocampi

for each experimental setting, which is considered adequate for our proof-of-concept purpose.

Segmentation accuracy was measured using the Dice score (also known as the "kappa index" or "similarity index"). It is a prevailing metric standard in biomedical image segmentation, computed by $Dice(A, B) = \frac{2|A \cap B|}{|A|+|B|}$, which is the number of matching labels between segmentation A and ground truth B, divided by the total number of labels in both label maps.

2.5.2 Training parameters

In our work, the hippocampal ROI mask covered around 70,000 voxels. With the use of 90 atlases in each test, there were around 6.3 million data entries collectively used to train a PatchDNN with $O(10^5)$ learning weights. At the training stage, the only major parameters to tune were: 1) patch size $p \times p$, 2) learning rate η, 3) mini-batch size δ.

The batch size is relatively simple, because it does not have significant impact, as long as adequate training is secured. For simplicity, δ was fixed at 200 entries for each training step. The learning rate on the other hand, is more difficult, because it can dramatically affect training outcome. Figure 2.17 illustrates the impact of three rates: $\eta = 10^{-4}$, $\eta = 10^{-5}$ and $\eta = 10^{-6}$, on the segmentation performance during the training of a PatchDNN with 13×13 patches. In this case, $\eta = 10^{-4}$ and $\eta = 10^{-6}$ were, respectively, set too high to train a good network and too low for the network to converge to an optimal solution in an efficient manner, whereas $\eta = 10^{-5}$ was considered the best rate, which was in fact the setting used in our final configuration.

In terms of patch size, we tested four settings, respectively, 9×9, 11×11, 13×13 and 15×15. The performance metrics of segmentation are shown in Figure 2.18. The highest median dice score achieved (using 13×13 patch setting) was 90.98%, which, to the best of our knowledge, is by far the highest accuracy level ever reported on hippocampus segmentation, compared to previous work using a comparable size of validation dataset (80-202 images [37, 143, 166, 155, 156]). The 9×9 setting scored slightly lower than the others at 87.75% median level, yet still outperformed the prior state-of-the-art (to be detailed in Section 2.5.4). Sample segmentations of the best, median and worst cases (in terms of the 13×13 setting) are compared in Figure 2.19.

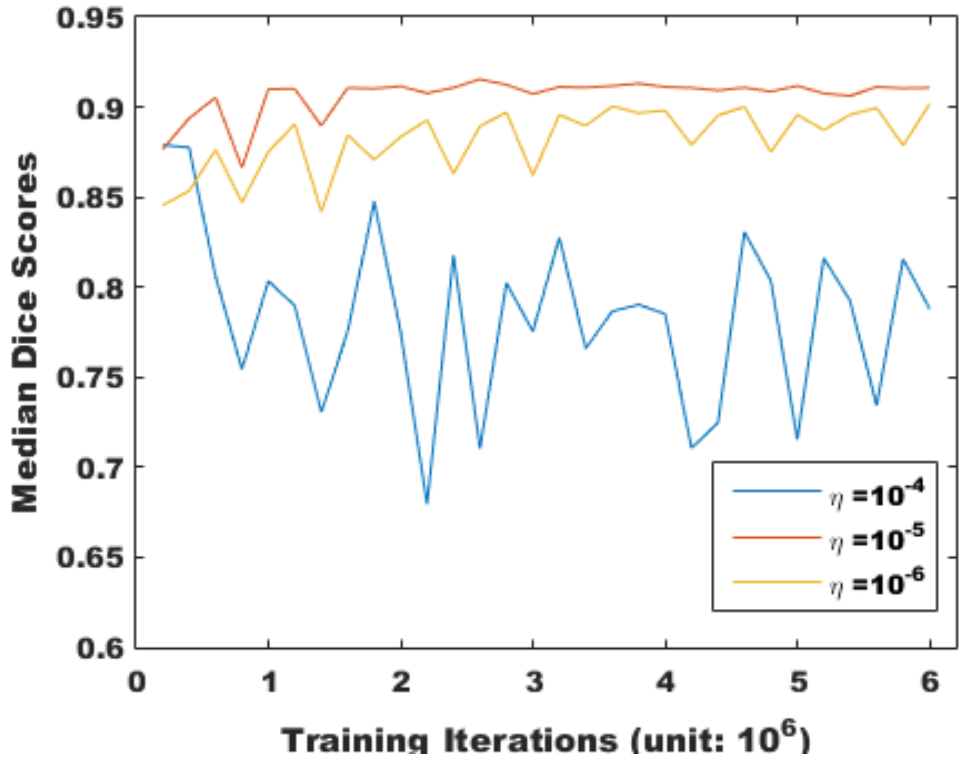


Figure 2.17: Impact of training parameters: learning rate η

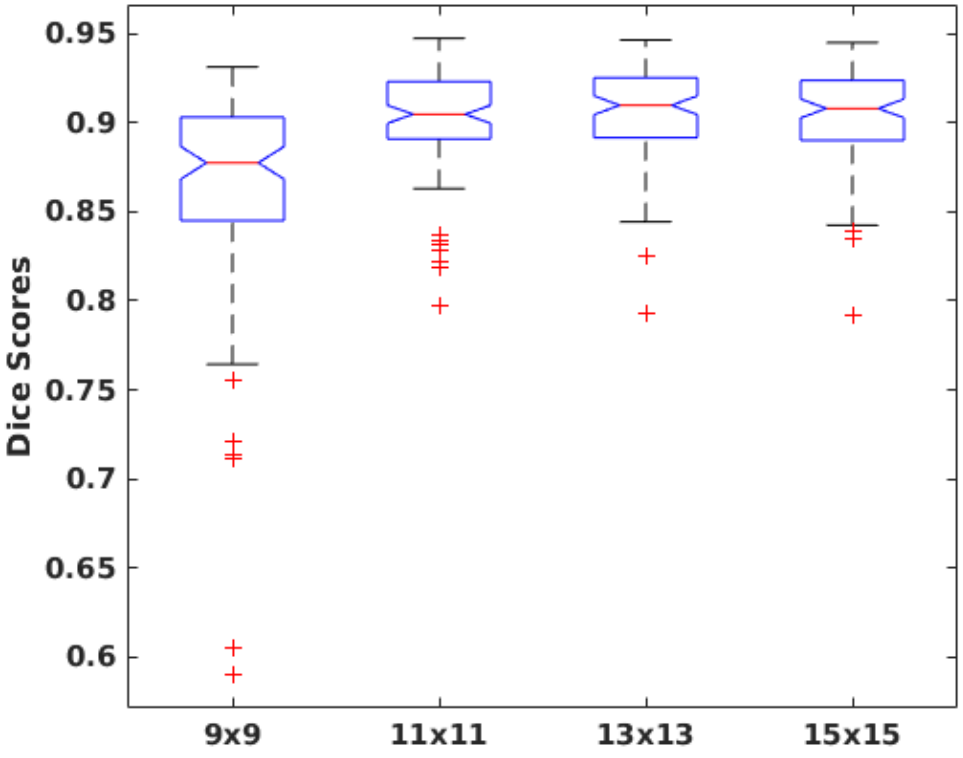


Figure 2.18: Impact of training parameters: patch size

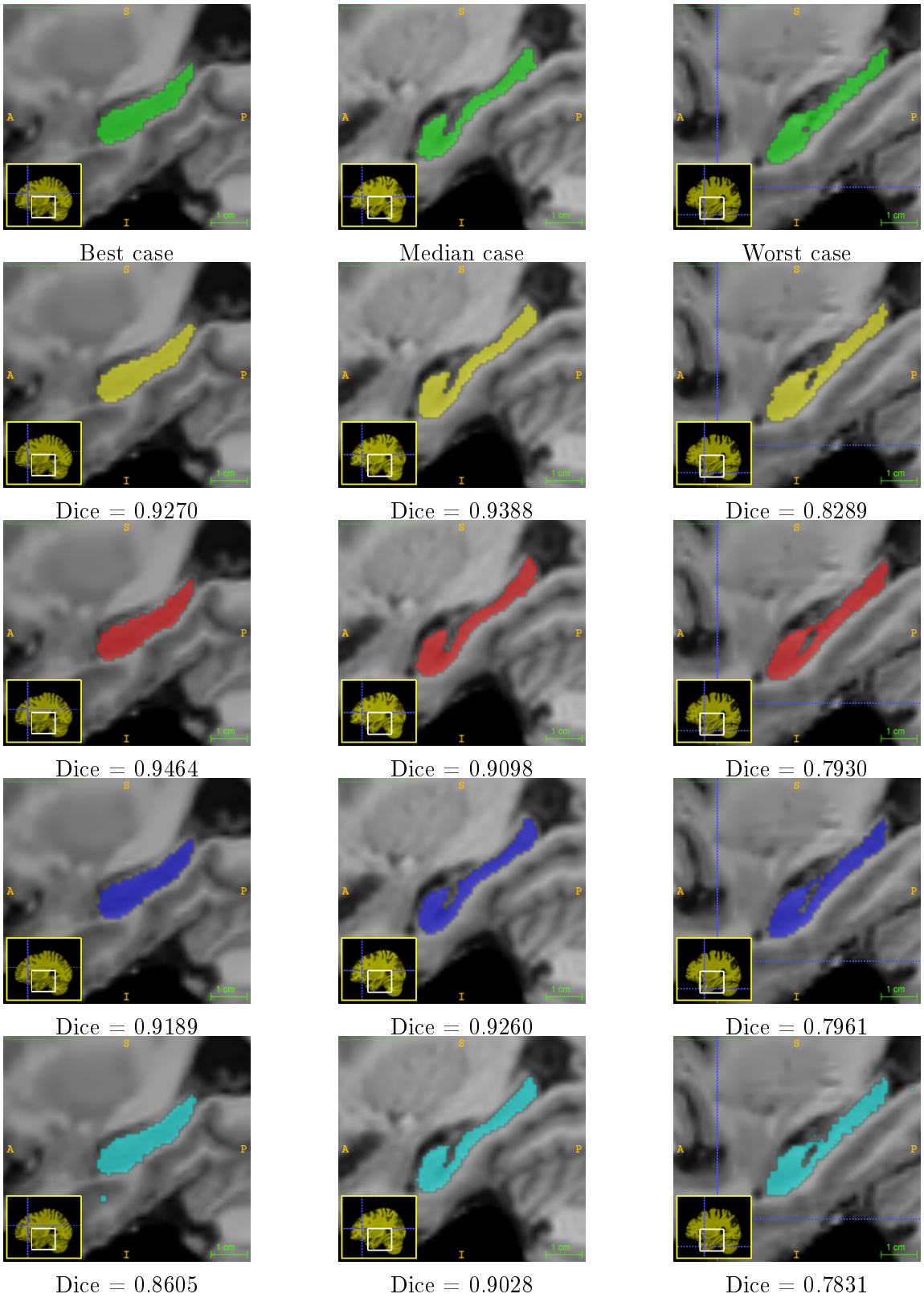


Figure 2.19: Sample segmentation outcome (the images are zoomed in around the hippocampus region in saggital view): (Row 1) reference segmentation, (Row 2) PatchDNN with 15×15 patches, (Row 3) 13×13 patches, (Row 4) 11×11 patches, and (Row 5) 9×9 patches. The best, median and worst cases are defined in terms of the 13×13 setting.

2.5.3 Training and testing time

The experimentation environment was deployed on a standard PC with an NVIDIA GTX Titan X graphics card. During training, each gradient descent step took approximately only $0.01s$. Since it generally took millions of steps to secure good performance, the total training time could take up to 10+ hours. In stark contrast, image segmentation at test-time was achieved at a near real-time speed, taking $< 1s$ for each target image. To the best of our knowledge, this is by far the fastest hippocampus segmentation system, with only the PBS-PatchMatch method [143] described in Section 2.3.2 having reported a comparable speed.

2.5.4 Comparison with prior state-of-the-art

In further evaluation we also compared our work with the best previous results in hippocampus segmentation, in which the conventional PBS methods have been very successful. Although there is a variety of such algorithms proposed in the literature, the fundamental principle is always pair-wise patch similarity measure with some hand-engineered low-level features (usually SSD variants). We therefore used the standard PBS method [37] and PBS with Patchforest [166] for comparison, both were described in Section 2.3.

The implementation of the standard PBS method was based on the built-in PBS framework in the open source IRTK repository[5]. To ensure fairness, we rigorously applied the same pre-processing described in the original work [37], including non-local means denoising [38], N3 bias correction and tissue-standardising normalisation [121]. For each target image, we then selected 10 atlases with lowest overall SSD within the masked ROI to perform segmentation. Two patch settings were tested: $5 \times 5 \times 5$ (PBS-1) and $7 \times 7 \times 7$ (PBS-2), with a $11 \times 11 \times 11$ search window, which were the two best performing configurations in the original work [37].

Figure 2.20 shows the performance metrics in comparison to the PatchDNN with the 13×13 patch setting. The best median Dice score obtained was 86.65% (PBS-2): a level somewhat below the 88.4% in the authors' own work, which was probably due to the use of different

[5]Image Registration Toolkit (IRTK), official website: `https://github.com/BioMedIA/IRTK`

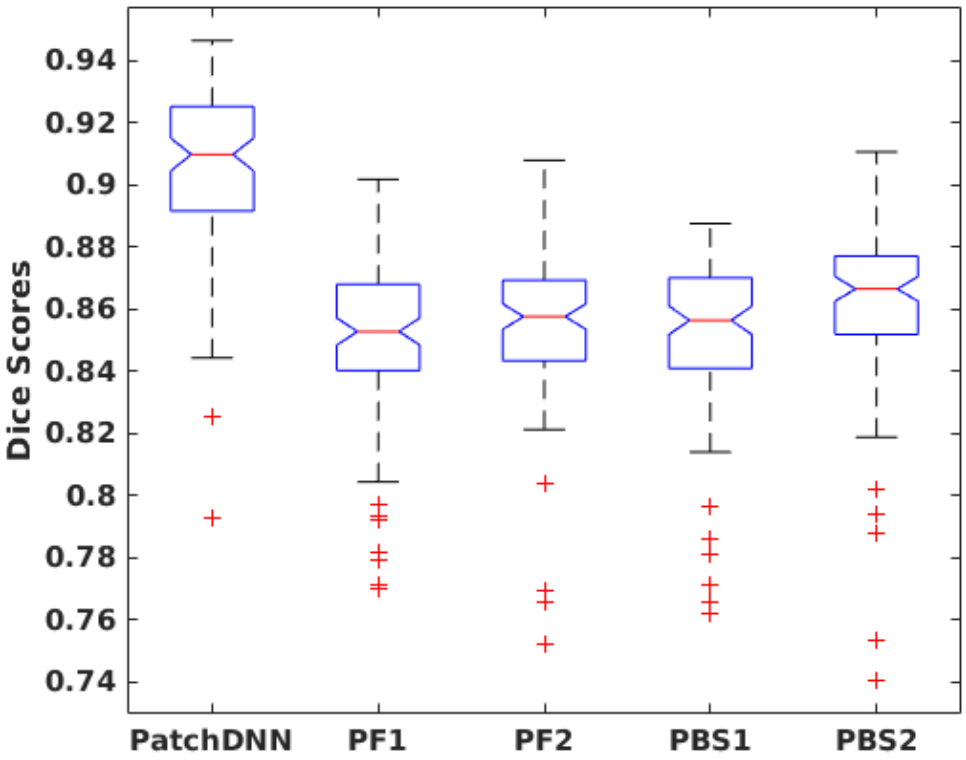


Figure 2.20: Comparison of segmentation accuracies: (from left to right) PatchDNN using 13×13 tri-planar patches, PF using $5 \times 5 \times 5$ and $7 \times 7 \times 7$ patches, and PBS using $5 \times 5 \times 5$ and $7 \times 7 \times 7$ patches

experimental datasets, in particular there were only healthy controls in their dataset whereas ours also includes MCI and AD subjects. Yet by any means the accuracy was notably lower than our PatchDNN approach. The segmentation outcome of the same three cases are illustrated in Figure 2.21 for a comparative view. Furthermore, efficiency boost was even more evident, with PBS-1 and PBS-2, respectively, taking an average $197s$ and $645s$ (excluding pre-processing) to segment a target image, compared to our near real-time level.

The PBS with Patchforest method was implemented based on the Microsoft Sherwood Library[6]. This method only modifies the patch search and label fusion processes, and thereby the remaining computation was identical to the standard PBS, including the pre-processing and atlas selection. The same two patch settings $5 \times 5 \times 5$ (PF-1) and $7 \times 7 \times 7$ (PF-2) were used, but the search window was expanded to $15 \times 15 \times 15$ with its improved patch search capability: which were also the best performing configurations in the original work [166].

[6]Microsoft Sherwood Library, official website: `https://www.microsoft.com/en-us/research/project/decision-forests/`

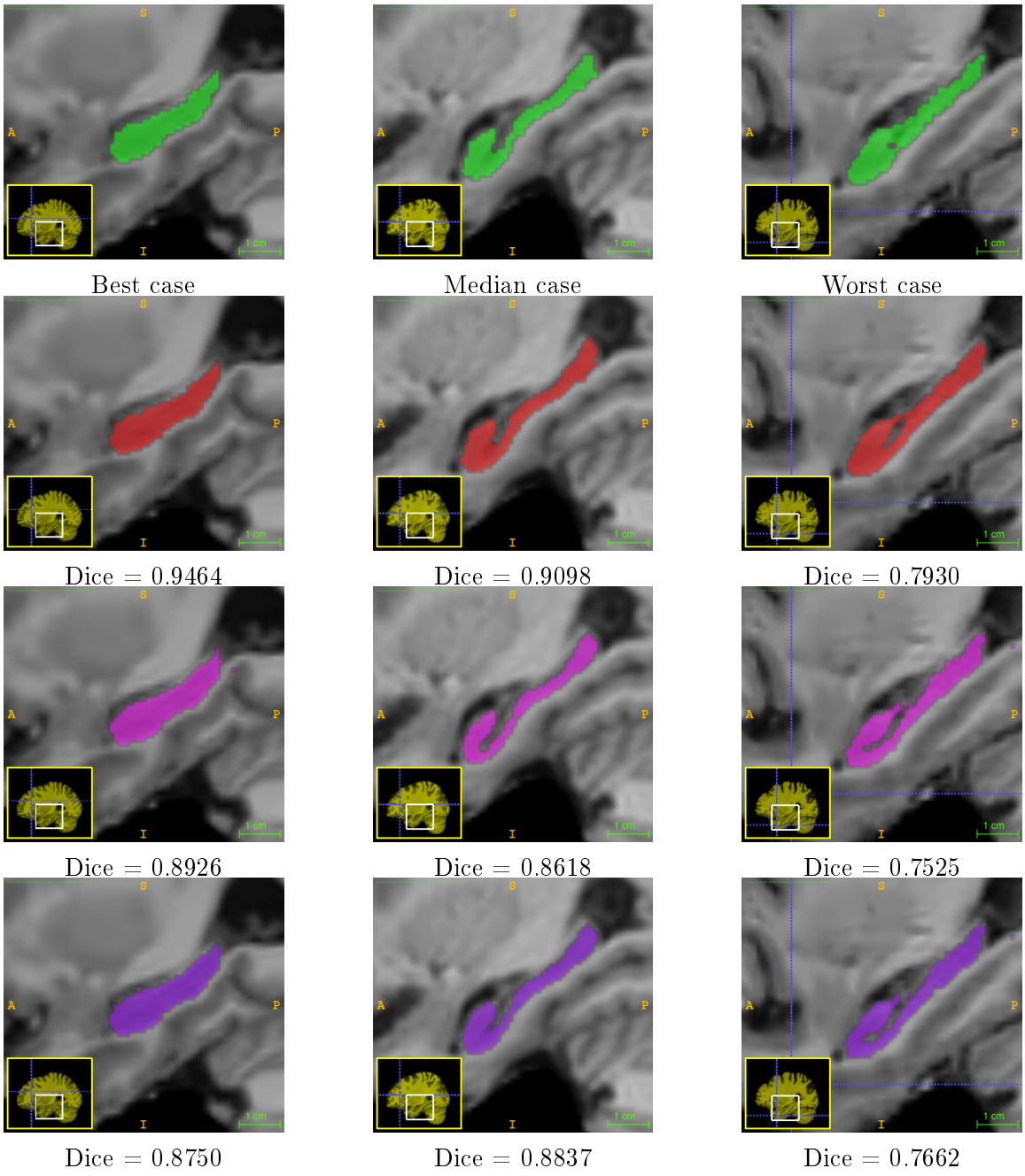


Figure 2.21: Comparison of segmentation outcome: (Row 1) reference segmentation, (Row 2) PatchDNN using 13×13 tri-planar patches, (Row 3) standard PBS using $7 \times 7 \times 7$ patches, (Row 4) PF using $7 \times 7 \times 7$ patches, applied to the same three subjects as in Figure 2.19.

The accuracy metrics are also shown in Figure 2.20 for direct comparison. The best median Dice score obtained was 85.12% (PF-2), which was slightly below the standard PBS method, but the difference was not significant. Segmentation outcome of the same sample subjects is also illustrated in Figure 2.21 for comparison. The computational time on the other hand was reduced to 115 and 130, which were considerably lower than the PBS method, especially in

the case of $7 \times 7 \times 7$ patch setting. Nonetheless, the PatchDNN approach again substantially outperformed this method in terms of both the segmentation accuracy and speed.

2.6 Discussion and Conclusion

Over the past decades, image segmentation in the biomedical domain in general has heavily rested on signal processing paradigms, with registration-based label propagation and pair-wise patch-based pattern matching being the cornerstones of prior state-of-the-art. The computational efficiency however, has been a major obstacle to a wider application in clinical environments. Although at a much limited scale, machine learning has been increasingly practised in recent years. In general, machine learning approaches tend to be much faster than signal processing counterparts, but the level of sophistication for feature representation has been a limiting factor in traditional (shallow learning) frameworks, such as random forests or support vector machines, to secure a comparable segmentation accuracy. Therefore machine learning is often used in combination with the conventional art to improve the overall performance, for example, to help with atlas selection [88], patch search [143, 166] and so on, rather than used for image segmentation alone.

Deep learning in contrast, by projecting contextual information over a highly intricate network composed of a large collection of neurons each associated with non-linear computation, is able to achieve an excellent capability of feature representation and use it to train a high performance classifier in terms of both accuracy and efficiency. Deep ConvNets are the model behind many ground-breaking image classification studies that overtook prior state-of-the-art in computer vision over recent years. Its major advantage for image classification is a comparatively small number of learning weights (although many systems still scale to millions) by employing convolution, whereas a fully connected CanonNet model applied to a large image would end up with an overwhelming number of learning weights that easily lead to over-fitting.

In the case of biomedical image segmentation with a patch-based approach however, we argue such concern is no longer as significant as at full image scale, and can be further relieved by the

employment of modern techniques such as dropout layers and tri-planar patch multi-pathway setting. For that reason, we abandon the popular ConvNet model and propose the PatchDNN using, which can be considered a substantially modernised CanonNet architecture. Without large-scale neuron-wise convolution on each single run, the network trains much faster than a deep ConvNet while consuming far less memory.

In addition, state-of-the-art network training techniques have also been utilised, including ReLU activation to introduce non-saturating non-linearity, mini-batch stochastic gradient descent, foreground/background separate sampling, which collectively can reduce the training time dramatically, making such an approach very practical. The effectiveness was clearly reflected by the quick segmentation and training speed. An important insight resulting from our work is that a ConvNet is not necessarily superior to an CanonNet despite its popularity, and we predict a shift away from ConvNet to other deep learning models for patch-based approaches in coming years.

In an evaluation study, we tested the proposed approach on the application of hippocampus segmentation using 100 brain MR images drawn from the widely-used ADNI database. Our framework was able to significantly outperform the prior state-of-the-art approaches, in terms of both segmentation accuracy and speed: scoring a median Dice score up to 90.98% with a near real-time performance ($< 1s$) on a modern GPU.

Chapter 3

High-throughput Mouse Phenotyping Using Non-rigid Registration and Robust Principal Component Analysis

3.1 Introduction

For many decades, scientists have been studying genetic impacts on mammalian development and disease pathology. The Human Genome Project[1] for example, was a major milestone marking the completion of human DNA sequencing and gene mapping, since which researchers have carried on to study gene-associated functional information. This process, known as phenotyping as compared to genotyping, however, is very difficult to follow due to numerous issues regarding modifying human genes, especially the moral concerns. The mouse genome, therefore, has widely been used as a surrogate model, due, in part, to the findings that 99% of mouse genes have a homologue in the human genome [116], as well as other advantages including the low cost and short time required for breeding mice.

Intensive global efforts have been underway toward understanding all the approximately 25,000 genes in the mouse genome, most notably the International Knockout Mouse Consortium

[1]The Human Genome Project, `https://www.genome.gov/10001772`.

(IKMC)[2] [72] and the International Mouse Phenotyping Consortium (IMPC)[3] [24], which are centred on one-by-one systematic gene modifications for comparative analysis. More specifically, an entire series of mouse embryonic stem cells is under development, with individual genes to be mutated using gene targeting [33] or other relevant techniques, to generate tailored mouse strains, which are then raised in order to investigate the impact of gene-mutation with respect to morphology, metabolism, or other biological traits (also known as the "phenotype"). Furthermore, it was estimated that around 30% of mutant mouse lines would end up embryo lethal [1], which in turn has stimulated considerable research effort into prenatal phenotyping.

In terms of morphological phenotyping, current practices still heavily rely on the traditional method using microscopic histological examination, which is not only labour intensive, highly time consuming and prone to errors during histological sectioning, but is also restricted to limited anatomical coverage. With the data volume and analytical workload scaling up exponentially in modern studies, the research community has been calling for some high-throughput phenotyping approaches [25]. Biomedical imaging technology, such as CT and MRI, has started being used to facilitate phenotyping practices concerning morphological anomaly, especially with the employment of image informatics to automate defect recognition.

Broadly speaking, there are two branches of study toward computer-aided phenotyping. The first is focused on developing some tailored algorithms that capture phenotype-specific features to help recognise target phenotypes. For example, certain heart phenotypes can be characterised by the detection of connectivity between cardiac ventricles [167], or diameter measurement of great arteries and semilunar valves [161]. This line of research ultimately leads to automatic classification of specific known phenotypes. However these tailored approaches often fail to serve a general phenotyping purpose, in particular the discovery of new phenotypes, since the corresponding phenotypical information is not known.

This naturally leads to the second branch of research centred on anomaly detection, which

[2]The International Knockout Mouse Consortium (IKMC, `http://www.mousephenotype.org/about-ikmc`) includes a number of projects, most notably the Knockout Mouse Project (KOMP), European Conditional Mouse Mutagenesis Program (EUCOMM) and the North American Conditional Mouse Mutagenesis Project (NorCOMM).

[3]International Mouse Phenotyping Consortium (IMPC), `http://www.mousephenotype.org`.

is often achieved by conducting data-driven comparative analytics between the normal and gene-modified mice. Existing work in this line primarily leverages volumetric contrast of target anatomical structures to identify anomaly [35, 120, 172, 89]. Nevertheless, volume contrast merely achieves a superficial level of screening for morphological phenotype, and would generally fail when it is not associated with severe volume variation, such as the ventricular septal defects (VSD) in the heart. Another major approach is to employ morphometrics based on a range of distinguishing deformation features, derived from non-rigid registration to a purpose-built template image [34, 134, 168]. However, such deformation-based morphometry is often not very robust because deformation features may vary significantly with the use of different registration settings and different reference templates.

Furthermore, phenotyping on mouse embryo data is particularly challenging for a number of reasons. First of all, mouse embryo development typically takes only around 18.5 days, and one day difference often leads to dramatic changes due to rapid organogenesis. In addition, the variety of nurturing conditions can also end up in differential growth rates. In this case, it is sometimes difficult to ensure subjects under study are acquired from the same developmental stage, especially when conducting comparative analysis on data retrieved from multiple sources. More importantly, this poses a substantial challenge to secure a suitable atlas for image segmentation, which happens to be a cornerstone of existing phenotyping work in both branches of research. Meanwhile the situation is further undermined by the limited availability of public atlases. On the other hand, segmentation via label propagation is also much more challenging on mouse embryos undergoing anomalous deformation, even if an atlas is secured.

All these reasons lead to a critical demand for a robust and efficient general-purpose anomaly detection framework without prior knowledge of the phenotype and the need of image segmentation, which is the purpose of this study. We propose a systematic framework that is able to efficiently detect morphological anomaly in a batch of images simultaneously, sensitive to both volumetric variations like polydactyly and non-volumetric variations like VSD, and does not require image segmentation, nor resort to unreliable deformation features, and is thus robust to various registration settings and template use. The key to the proposed approach lies on the combined employment of non-rigid registration and robust principal component analysis

(RPCA) to achieve feature decomposition into a regular and a singular component, the latter of which is then used for anomaly detection.

This chapter will be focused on methodological development, with a comprehensive pipeline that starts from image denoising, mouse embryo extraction, creation of a model template using local data, and goes on to group-wise non-rigid image alignment and RPCA feature processing for anomaly detection. The RPCA technique will be studied in more depth in the next chapter, and a novel RPCA-P method will also be developed to better address the wide prevalence and varying levels of natural variation in biomedical data, which is able to significantly improve RPCA's practical performance in the biomedical domain.

3.2 Existing phenotyping work and limitations

3.2.1 Phenotyping via comparative analytics

Volumetric contrast

The first major branch of existing image informatics work is primarily centred on data-driven comparative analytics, most notably by leveraging volumetric contrast of anatomical structures between the normal and gene-modified mice for abnormality detection [35, 120, 172, 168, 34].

These approaches generally rest on image segmentation over the structures of interest as a pre-requisite. Typically, an atlas is created in advance with target structures manually labelled by experts. At test time, target images are non-rigidly registered with the atlas, and labels are propagated over the target space to perform segmentation, which is widely recognised as atlas-based label propagation (also known as segmentation propagation) and has previously been detailed in Chapter 2 Section 2.2. Once structures of interest have been segmented, volumetric evaluations can then be conducted on separate abnormal subjects from the normal.

However, volumetric contrast merely achieves a superficial level of screening for morphological abnormality, and generally fails when the anomalous phenotype is not associated with severe

volume variations, such as for the case of VSD. Furthermore, performing atlas-based segmentation on images of mouse embryos undergoing abnormal deformations can be very challenging, which will be detailed in Section 3.3.

Morphometrics on deformation features

In response to these limitations, another school of researchers has suggested employing morphometrics based on a range of distinguishing deformation features, derived from non-rigid registration to some purpose-built template image [34, 89, 168, 134]. Such methods are sometimes also recognised as "deformation-based morphometry" or "tensor-based morphometry". Typically, a group-wise non-rigid image registration process will be carried out first, to collectively align target images in the designated template space. Then a set of quantitative features, often based on the deformation fields resulting from the non-rigid image registration, are leveraged to capture distinct deformation properties across different subjects, groups or genotypes. It is hypothesized that in such settings, phenotypes associated with morphological anomalies would present abnormal deformation features that are highly distinguishable from the normal ones.

Suppose the deformation field (also known as "transformation") of an image I is denoted as F_I, and $F_I(v)$ represents the displacement vector that maps a voxel v to its counterpart in the template space. The final distinguishing model applied to "normal-vs-abnormal" classification in practice is often a portfolio of primitive features that are direct derivatives of the deformation fields, such as:

- voxel-wise displacement magnitude (also called "positional shift") [172, 34, 89]: obtained by computing the Euclidean norm $||F_I(v)||$ for each displacement vector $F_I(v)$
- voxel-wise rate of volumetric expansion/contraction [172, 168, 34, 134]: obtained by computing the determinant $J_I(v)$ of the local Jacobian matrix. $J_I(v) > 1$ indicates voxel expansion and $J_I(v) < 1$ indicates contraction
- block-wise deformation stress [134]: obtained by computing the entropy of varying deformation directions within a local block of neighbouring voxels $B_I(v)$

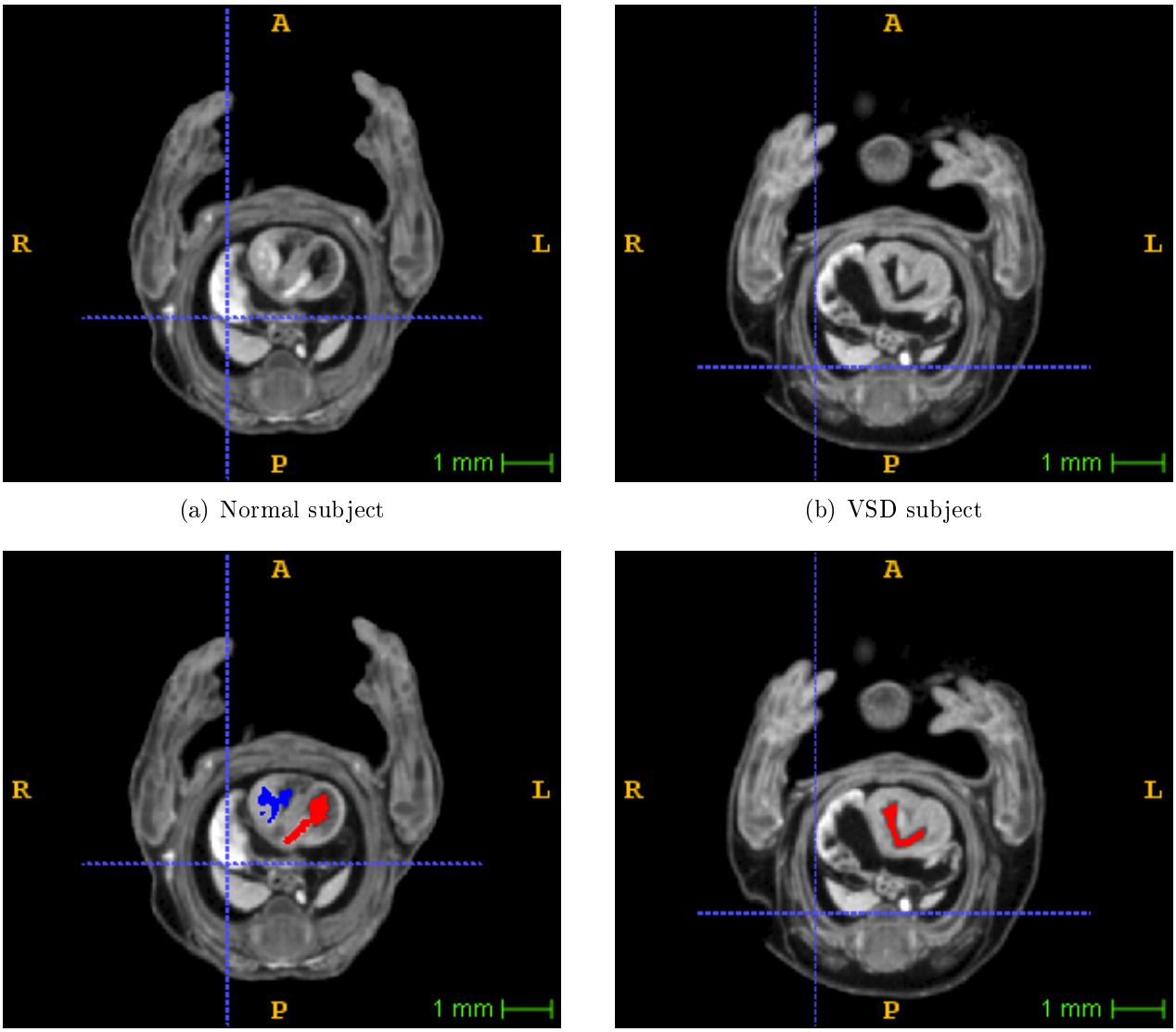
Meanwhile, functions like mean, variance, and standard deviation are frequently applied to aggregate individual metrics to capture systematic differences between the normal and gene-mutated populations. Moreover, image segmentation, or simply a mask roughly covering the region of interest, is often applied as well in order to obtain aggregate feature metrics at the structure or region level.

Nevertheless, this type of phenotyping approach is often not very robust, since deformation features could vary significantly with the use of different registration settings and different reference templates. In addition, the build-up of feature portfolios is often carried out empirically on a trial and error basis, without strong theoretical support for the choice of feature combination.

3.2.2 Phenotyping via detection of phenotype-specific features

The second major branch of existing work is generally centred on leveraging phenotype-specific information to help recognise target phenotypes. A number of phenotyping methods have been proposed to identify some (well-known) defective phenotypes via the detection of their well-studied distinctive morphological features, such as the connectivity between cardiac ventricles to identify the VSD [167] and diameter measurement of great arteries and semilunar valves to identify vascular and valvular stenosis [161], as well as cavity analysis regarding various heart malformations [137], etc. This line of research ultimately leads to automatic classification of specific known phenotypes.

Probably one of the best examples is the class of phenotypes concerning congenital heart diseases, which has attracted wide attention in the domain of mouse embryo phenotyping [69]. Surprisingly however, there is very limited research progress harnessing image computing technology to the recognition of heart phenotypes. This is not only due to the challenges of embryo heart segmentation in which only a few semi-automatic segmentation studies were found [182], but also because heart defects tend to be more subtle and often difficult to identify using, for example, volumetric contrast. The VSD for instance, is one of the most common congenital heart diseases (amongst others such as cardiomyopathy and atrial septal defects) [69]. Figure



(a) Normal subject

(b) VSD subject

(c) Ventricle connectivity detection: negative

(d) Ventricle connectivity detection: positive

Figure 3.1: VSD classification via the detection of ventricular connectivity: (a) a normal subject, (b) a VSD subject, and (c-d) ventricular connectivity detections of (a-b) using joint ventricle segmentation and snake evolution [167]

3.1 (a) and (b), respectively, illustrate the μ-CT images of a normal subject and a VSD subject in a comparative fashion. This particular phenotype features the presence of a hole or similar defect in the cardiac ventricular septum that divides the left and right ventricles, for which no severe heart volume difference may be observed between the normal and abnormal populations.

In order to address this challenge, in our earlier work [167] we developed, to the best of our knowledge, the first fully automatic VSD classification system in mouse embryo phenotyping. The core of the approach is to perform a (coarse) atlas-based segmentation on individual ven-

tricles, followed by a snake evolution (also known as active contour) algorithm that gradually grows the contours of the two ventricles. VSD classification is achieved by checking whether the two ventricle segmentations eventually border or overlap with each other. A sample test applied to the two aforementioned cases are shown in Figure 3.1 (c) and (d), where, respectively, a negative and a positive result of ventricular connectivity detection were generated accordingly.

However, all these approaches are tailored for specific known phenotypes, and are normally unable to deliver a general phenotyping purpose, in particular the discovery of new phenotypes, since the corresponding phenotypical information is not known.

3.3 Further challenges of mouse embryo phenotyping

3.3.1 Rapidity of mouse embryo development

Prenatal mouse phenotyping is particularly challenging due to the short embryo development period. Typically, it only lasts for around 18.5 days, and one day difference often leads to dramatic changes due to rapid organogenesis. The approximate timeline of mouse embryo development is illustrated in Figure 3.2. Moreover, the variety of nurturing conditions across different facilities can also contribute to mildly differential growth rates. For that reason, it is sometimes difficult to ensure subjects under study are acquired from the same developmental stage, especially when data is retrieved from different cohorts at multiple institutions.

To help cross-comparative analysis, a standardising system is often employed to categorise imaging subjects into different stages. Probably the most common and straightforward approach is to categorise via timed pregnancy, or more specifically, based on the measure of days post-coitum (dpc). Normally, mouse embryos are collected from natural mating, and noon on the day, when the presence of vaginal plug is detected, is designated as 0.5 dpc or the E0.5 stage. This is also the standard used in Figure 3.2. However, the timeline could be significantly affected by different nurturing conditions as mentioned earlier, and meanwhile is also subject to inter-sample variability.

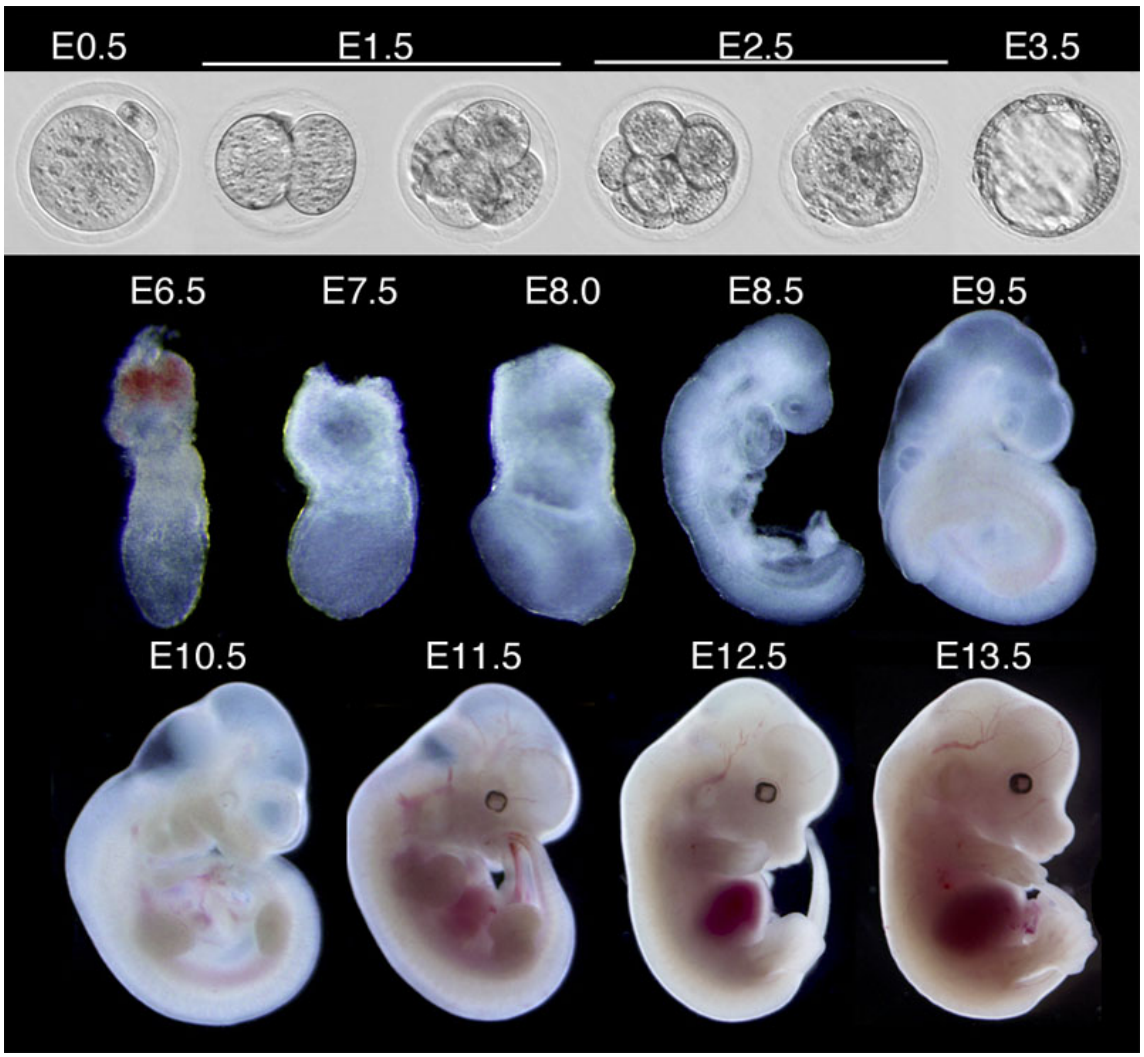


Figure 3.2: The approximate timeline of mouse embryo development: the E0.5-13.5 notations refer to 0.5-13.5 days post-coitum. (image source: the e-Mouse Atlas Project, `http://www.emouseatlas.org`)

Alternatively, some other researchers tend to use a set of morphological criteria to stage fetal development. The Theiler stage is arguably the most widely used such system, which divides fetal development into 26 prenatal and 2 postnatal stages. The staging criteria is detailed in the seminal work by Karl Theiler [146], and is to some extent similar to the Carnegie stage system in human embryology [122]. The Downs and Davies stage [48] is another well-known and more recent system, but is much less widely-used in practice. However, although such a staging system in theory is able to establish a better standard to measure developmental status of normal embryos, subjects of anomalous phenotypes sometimes do not follow the same rule, especially the ones leading to prenatal mortality. Therefore, to some extent it is actually more common to use the timed pregnancy system in phenotyping practices.

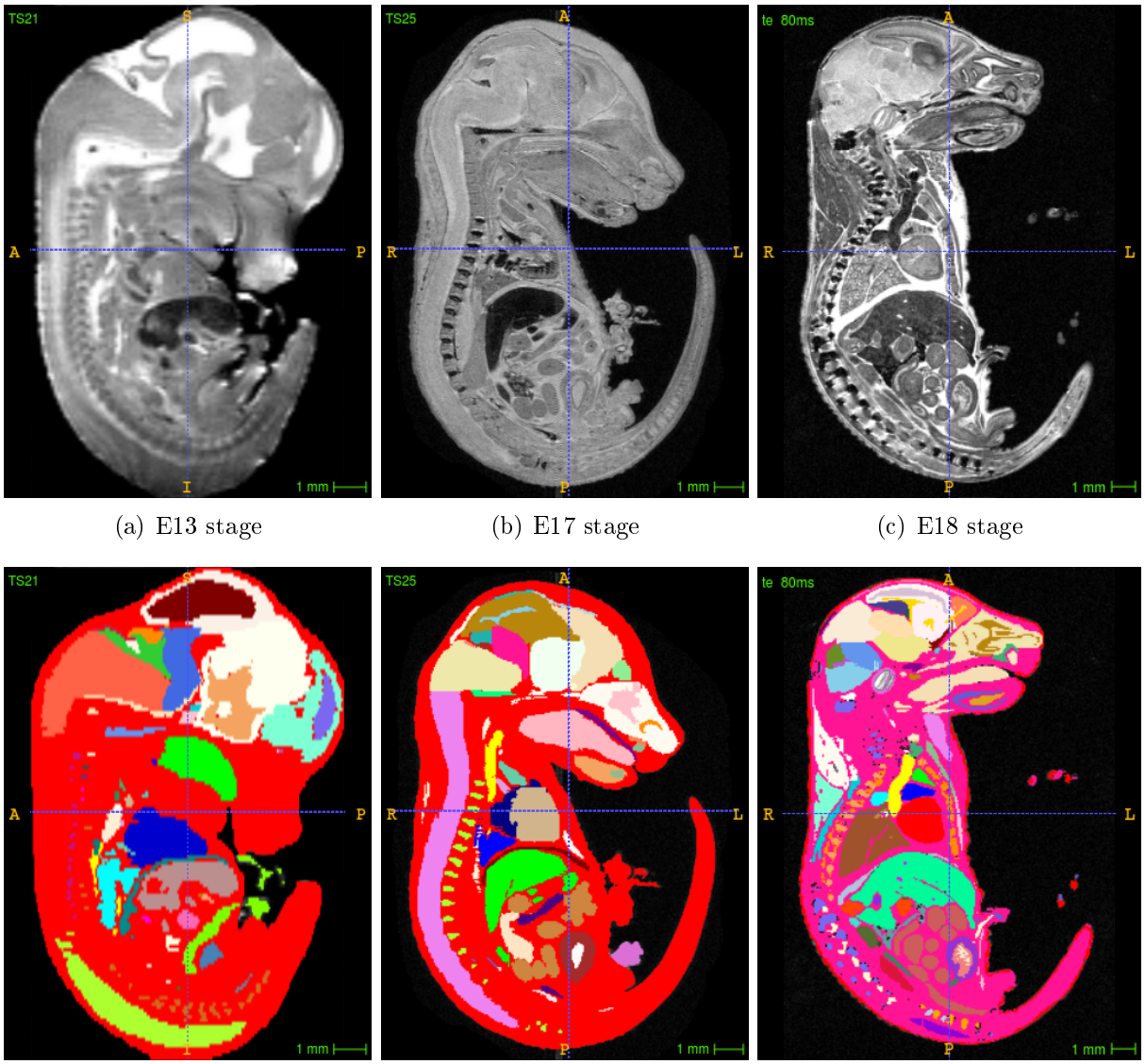
Another significant challenge stems from the incompleteness of knowledge over phenotypes. Not only does this pose substantial obstacles to phenotyping using tailored anomalous features, but it also raises the question from what stage the relevant phenotypical features start to be manifest, and consequently when to collect data. An optimal strategy to obtain desirable results in the shortest possible time is a common matter of consideration in embryo phenotyping work. Typically, to ensure productive image analysis, the embryos are usually collected at later stages, often after 12.5 or 14.5 dpc, when organogenesis generally finalises and salient phenotypical features have emerged to allow effective image processing.

3.3.2 Challenges of image segmentation

A large proportion of existing image analytics-based phenotyping studies rely heavily on image segmentation, which is particularly challenging on mouse embryo data, especially in the presence of anatomical defects. Considerable progress has been made in recent years on the model of the adult mouse, including the creation of mouse atlases [89, 47, 74, 110] and brain segmentation work [3, 165, 93, 100, 9, 97, 119]). On the other hand, embryo segmentation has encountered more obstacles. An obvious challenge directly resulting from the rapid and differential developmental status of mouse embryos, is the difficulty in practice to secure a suitable public atlas, since the studies are often conducted in different environments with different cohorts of mice. To make situation worse, the availability of public atlases is actually very limited, with the only ones we found being:

(1) **a set of multi-stage μ-MRI atlases** [80]: six atlases in total, respectively, at E8.5, E13, E15, E16, E17 and E18 stages, created by directly conducting manual segmentation on selected μ-MRI mouse embryo images, with labels covering all parts of the embryos, as shown in Figure 3.3. However, the image quality is relatively poor.

(2) **an E15.5+ μ-MRI atlas** [35]: during its creation, a template image was generated first, using 19 μ-MRI embryo images ranging from E15.5 to E18.5 stages, followed by manual segmentation on the template image, with labels mainly covering the brain and the heart,



(a) E13 stage (b) E17 stage (c) E18 stage

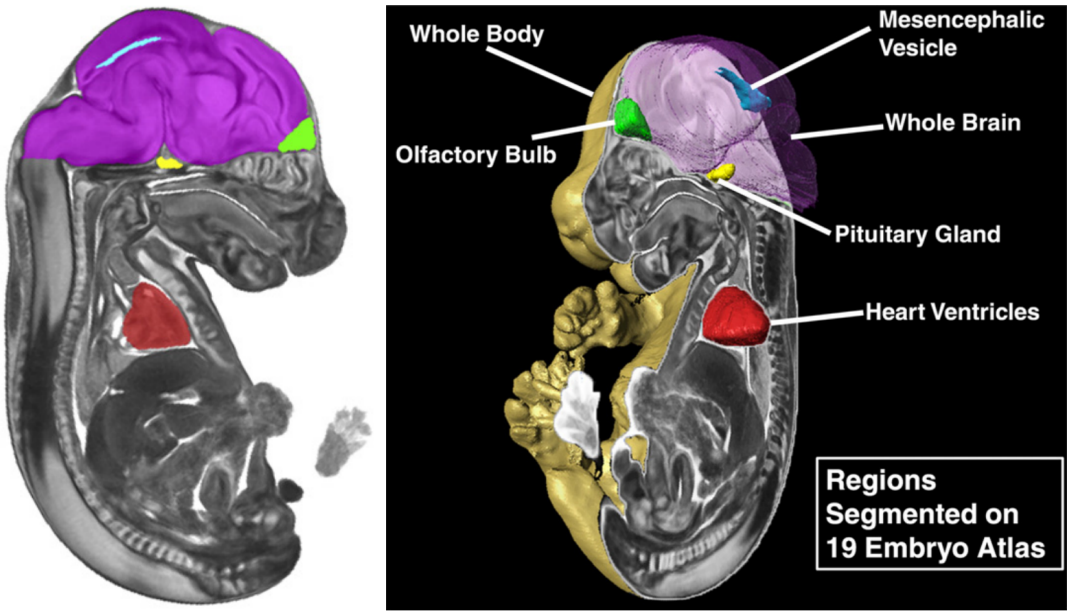
(d) E13 atlas with labels (e) E17 atlas with labels (f) E18 atlas with labels

Figure 3.3: The multi-stage μ-MRI atlas set [80]: (a-c) sample images at different stages, (d-f) superimposed with label maps

as shown in Figure 3.4.

(3) **an E15.5 μ-CT atlas** [162]: during its creation, a template image was created using 35 E15.5 μ-CT embryo images first, followed by manual segmentation on 48 anatomical structures, including 25 structures in the brain as well as 23 in other organs such as the heart, lung, liver, stomach, as illustrated in Figure 3.5.

Among them, the modality of atlases (1) and (2) are incompatible to the μ-CT images used in our work. Meanwhile, atlas (3) is inapplicable to phenotypes such as polydactyly (extra fin-



(a) 2D view (b) 3D view

Figure 3.4: The E15.5+ μ-MRI atlas [35]

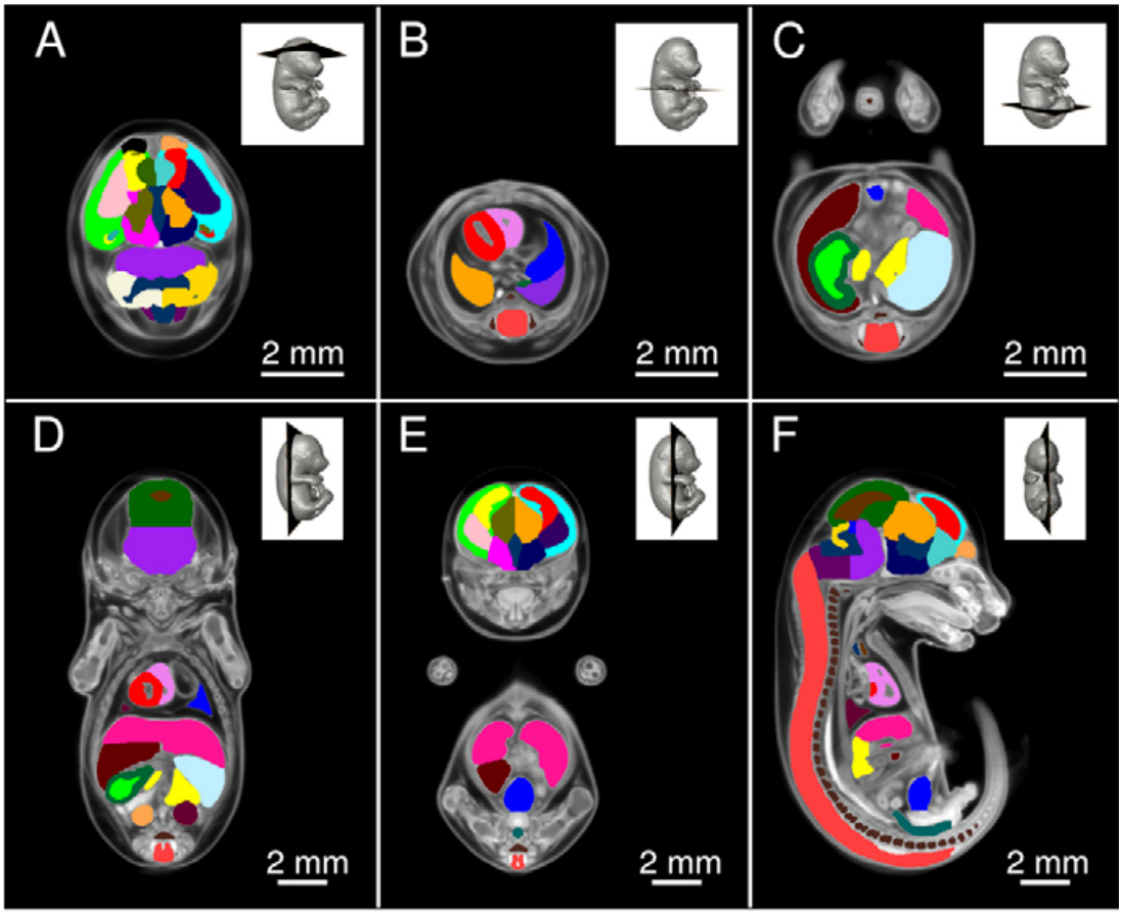


Figure 3.5: The E15.5 μ-CT atlas [162]

gers/toes), which is to be tested in our study, due to the absence of limb labels for segmentation propagation. Creating a local atlas on the other hand, is an expensive and time-consuming job requiring domain expertise, and thus may well be impractical in many scenarios. Furthermore, even allowing for the presence of a suitable atlas, it could still be difficult to establish a voxel-level one-to-one correspondence on a region undergoing anomalous deformation with a normal counterpart, for example the heart of the VSD subject with that of the control subject along the ventricle regions in Figure 3.1 .

3.4 Overview: Methods and Materials

3.4.1 Anomaly detection by non-rigid registration and RPCA

In essence, our approach is centred on group-wise feature extraction and decomposition into regular and singular components, where the singular features are then used for detection of morphological abnormalities. The core of this approach lies on the combined use of group-wise non-rigid registration and RPCA techniques.

As a preliminary step, image denoising is applied to each raw image of mouse embryo, which is then extracted from its unwanted surroundings. Subsequently, all extracted mouse embryos are group-wise non-rigidly aligned with a template image, using a standard three-step registration scheme including rigid, affine and B-spline registrations. The template is created locally using only the normal control data. Once all target images are group-wise aligned in the template space, we then proceed to decompose each image with the model $I_i = I_{i,r} + I_{i,s}$. In a superficial sense, this could be understood as that each embryo's observed appearance is a distortion of its regular form by some singular factors, in particular anomalous deformations.

Suppose there are n images $I_1, .., I_n \in \mathbb{R}^{w \times h \times d}$. Concatenating each image into a vector of size $m = w \times h \times d$ and stacking them together generates a matrix $D = R + S$, where $D, R, S \in \mathbb{R}^{m \times n}$. Since D is the observed data, the goal then is to estimate an $\hat{R}$ such that it best represents its regular form. Due to the structural congruity of mouse anatomy, R should be low-rank,

which essentially turns into a low-dimensional subspace learning problem with the presence of anomalous data. It can then be formulated as:

$$\min_{R,S} . \quad ||S||_0 \qquad \text{s.t.} \quad rank(R) < min(m,n), D = R + S \tag{3.1}$$

where $||.||_0$ denotes the element-wise 0-norm, the total number of non-zero entries in the matrix.

Principal component analysis (PCA) is a well-known feature extraction technique that projects high-dimensional data into a low-dimensional feature space [75]. However, the classic PCA algorithm only works well when the data points are independent and identically distributed (i.i.d.) in a Gaussian manner. Phenotypical deformation readily breaks this condition with arbitrarily large levels of variation.

In recent years, a number of approaches have been proposed to make PCA more robust, most notably the RPCA framework based on principal component pursuit [32]. It is an extension of the classic PCA that tolerates a certain level of gross data corruption, and tackles Eq. (3.1) via an approximate conversion to:

$$\min_{R,S} . \quad ||R||_* + \lambda||S||_1 \qquad \text{s.t.} \quad O = R + S \tag{3.2}$$

where $||M||_* = \sum_i \sigma_i(M)$ denotes the nuclear norm of a matrix M, which is the sum of its singular values, and $||M||_1 = \sum_{ij} |M_{ij}|$ denotes the element-wise 1-norm, the sum of its absolute values. The theoretical foundations of RPCA and popular algorithms to solve Eq. (3.2) will be detailed later in Chapter 4. The derived R and S can then be used to reconstruct $I_{i,r}$ and $I_{i,s}$ for each subject i, and $I_{i,s}$ is subsequently used for anomaly detection.

RPCA (including its variants) has previously been applied to a number of computer vision problems and achieved remarkable success. Notable examples include video surveillance [163, 32, 8, 175, 174], face recovery from shadow and specularity [163, 32, 175, 174], and batch linear alignment of face images with partial occlusion/corruption [124]. Sample work of video surveillance and face recovery are described below, which are primarily focused on deriving the S and R, respectively.

Figure 3.6: RPCA application to video surveillance [32]: (left) sample original video frames in D, (middle) the invariant background recovered by R, and (right) the moving objects (such as people and luggage) captured by S

Example RPCA work: video surveillance

Figure 3.6 illustrates some sample decomposition results retrieved from an RPCA study applied to video surveillance [32]. In this application, a recorded video clip was acquired from a stationery camera set up to monitor a scene in an airport. The video clip contained 200 frames of the scene without significant change in illumination condition. During the period of recording, there was a lot of human activity occurring in the scene, mostly consisting of different people walking along the corridor. All 200 frames were vectorised and stacked together to create an observation matrix D for RPCA processing. As a result, the background was captured in the derived R, whereas the moving objects (such as people and luggage) were successfully separated into S.

Figure 3.7: RPCA application to face recovery [32]: (left) sample original face images in D, (middle) the faces recovered by R, and (right) the distortions resulting from the shadow and specularity captured by S

Example RPCA work: face recovery from shadow and specularity

Figure 3.7 illustrates sample decomposition results retrieved from another RPCA application, which was aimed at recovering faces from shadow and specularity [32]. In this application, a set of face images were downloaded from a popular benchmark database[4]. For each subject, a

[4]The Extended Yale Face Database B: `http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html`

total of 31 photos taken under different illumination conditions were collected, all of which were cropped and well-aligned. Similarly, all these images were vectorised and stacked together to create an observation matrix D for RPCA processing. As a result, the faces in the images were successfully recovered in R, and the distortions associated with the shadow and specularity were captured by S.

Limitations of existing RPCA work and the proposed approach

Nevertheless, almost all these previous applications commonly rest on a critical condition, which is that in the majority of images being processed, there is a background with very little variation. By contrast in biomedical imaging there is a significant natural variation across different image samples, hence the method needs to be upgraded to address this condition.

A comprehensive study of the RPCA technique will be carried out in Chapter 4, including the development of a novel RPCA-P method that is able to incorporate natural variation priors to better address this challenge in biomedical imaging. In this study, non-rigid image registration is employed to deal with this issue. Non-rigid registration estimates a combined global and local transformation from one image to another, and establishes a voxel-level one-to-one correspondence between two images undergoing certain levels of physiological motions (such as heartbeat) and inter-subject variation. A detailed description of non-rigid registration has been covered in Chapter 2 Section 2.2.1.

The proposed approach rests on the assumption that the phenotypes of interest will cause salient topological distortion to a subject's regular form, which cannot be well aligned with the normal counterparts. A validation study has been carried out using two different defective phenotypes: VSD and polydactyly. The former is associated with the presence of a hole in the cardiac ventricular septum, leading to the (lethal) connectivity between left and right ventricles, whereas the latter indicates the presence of extra fingers/toes. Such abnormal features do not match with normal features and will not be "corrected" by non-rigid registration. Another assumption is that anatomical abnormalities across different subjects do not appear in consistent patterns even for the same phenotype, making anomalous features always sparse.

Both assumptions hold quite generally for many phenotypes, and this approach should work robustly in practice.

3.4.2 Data acquisition

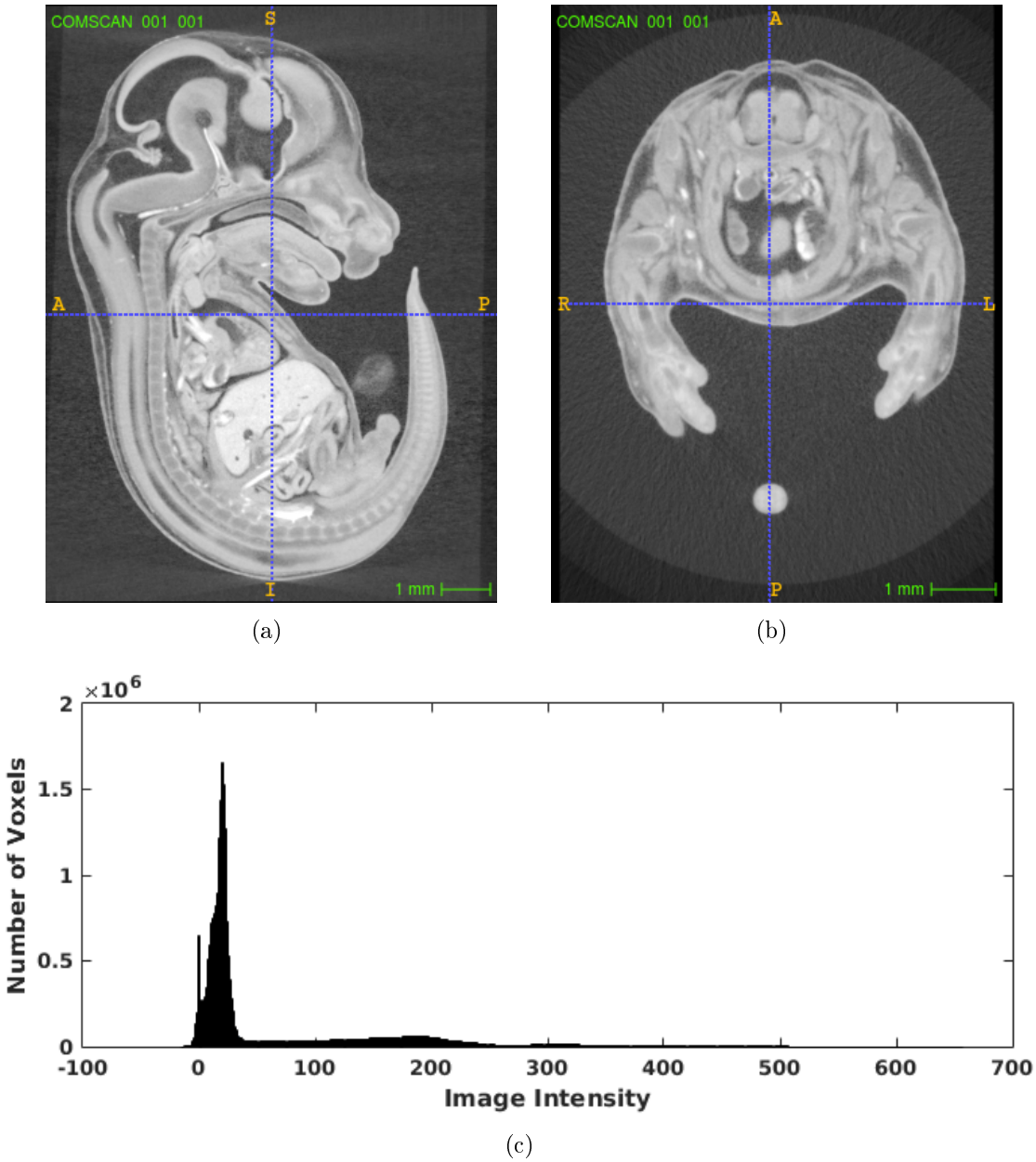
All the test data was produced by an expert team at the National Institute of Genetics[5] and the RIKEN BioResource Centre[6] in Japan. All our animal experiments were approved by the Animal Care and Use Committee, and all mouse subjects used in this study were maintained at the Genetic Strains Research Centre, both within the National Institute of Genetics. The mouse embryos were stored in a specific pathogen free facility, with 12-hour light and dark cycles. The popular C57BL/10 mouse strain was used as both the source of normal subjects and the basis for gene-modified lines. The embryos were generated by natural mating and timed pregnancy was used as the staging system.

All mouse embryos in this study were collected at 14.5 dpc. A notable point is that sample collection at earlier stages might lead to the absence of certain anatomical structures and substantial inter-subject variability due to ongoing organogenesis, making it difficult to perform non-rigid image registration and RPCA. All the collected embryos were then washed in Phosphate Buffered Saline solution and maintained in 4% Paraformaldehyde solution. Just before imaging, the embryos were soaked in a contrast agent, created by using a 1:3 mixture of Lugol solution and double distilled water. The scanning was then performed on a SCANXMATE-E090S 3D μ-CT machine (Comscan Techno, Japan), during which each embryo was fitted in a separate 1.5 ml Eppendorf tube and fixed by wet paper. The X-ray radiation was applied at a tube voltage peak of 60 kVp and a cube current of 130 μA. The subject was rotated by 360^o at 0.36^o per step, generating 1000 projections and reconstructed in 3D at an isotropic resolution.

There were two datasets used for experimentation in this work:

[5]Principal researcher: Professor Toshihiko Shiroishi, Mammalian Genetics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. Official website: `https://www.nig.ac.jp/nig/research/organization-top/organization/shiroishi`

[6]Principal researcher: Dr. Masaru Tamura, RIKEN BioResource Centre, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan. Official website: `http://en.brc.riken.jp/`

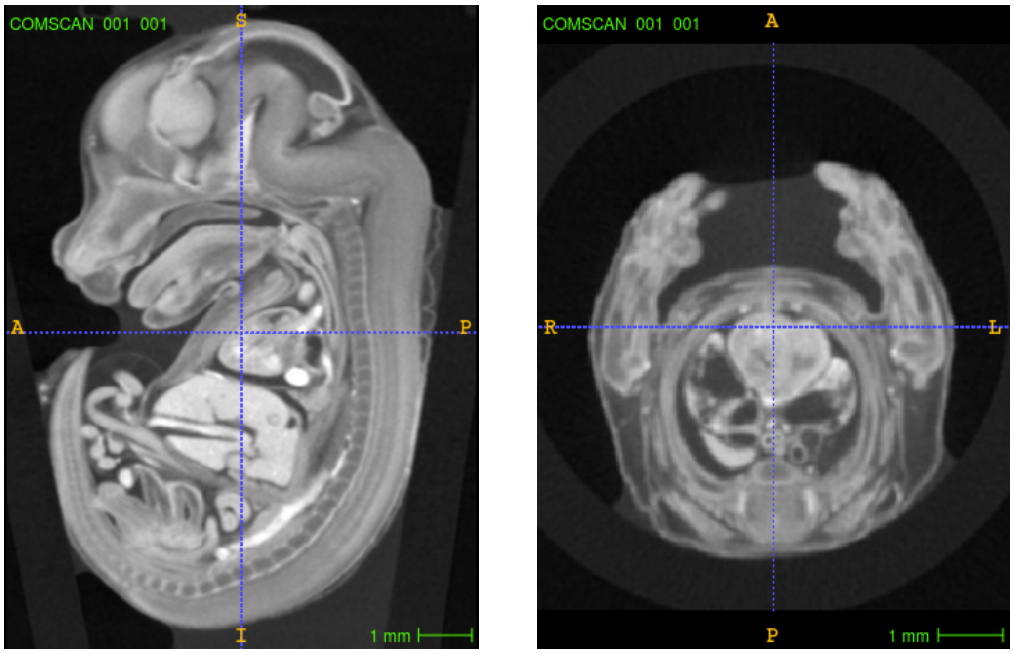


(a) (b)

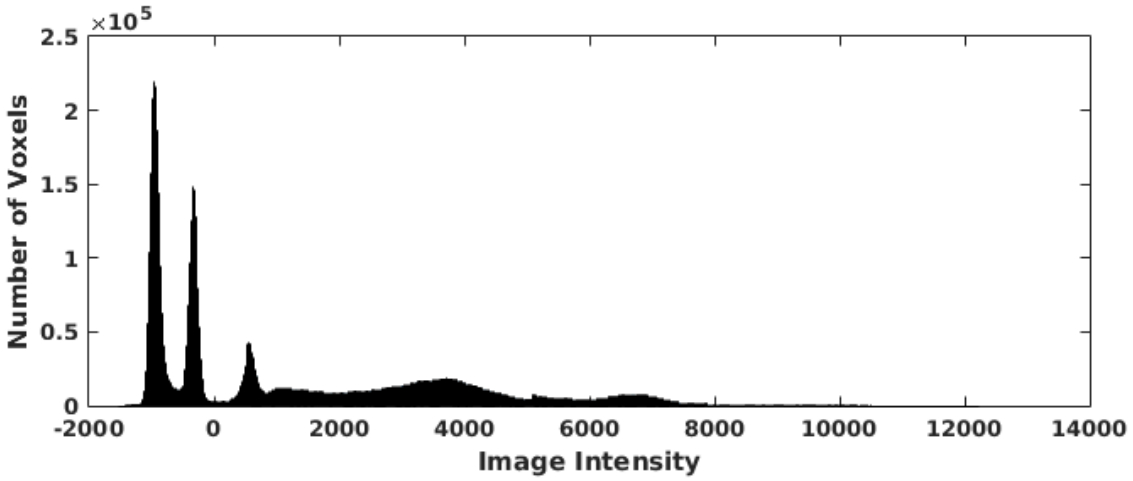
(c)

Figure 3.8: An example raw image of a mouse embryo from Dataset A in (a) sagittal view, (b) axial view, and (c) its intensity histogram

- **Dataset A** consisted of 31 μ-CT scans, each contained a separate mouse embryo from the gene knockout project, acquired via an imaging protocol conforming to IMPC standards. Each image was approximately sized $560 \times 640 \times 950$ at a spatial resolution around $12 \times 12 \times 12\ \mu m^3$. Among them, 26 subjects were standard C57BL/10 embryos showing the normal phenotype, while the remaining 5 were knockout subjects with 4 samples manifesting the VSD phenotype. Figure 3.8 illustrates an example raw image of a mouse embryo from Dataset A in sagittal and axial views, as well as its intensity histogram.



(a) (b)



(c)

Figure 3.9: An example raw image of a mouse embryo from Dataset B in (a) sagittal view, (b) axial view, and (c) its intensity histogram

- **Dataset B** consisted of 15 μ-CT scans, each contained a separate mouse embryo from the recombination-induced mutation 4 (Rim4) project [144]. These embryos were imaged using a slightly different protocol with intensities rescaled to a different standard. The images were sized around $300 \times 360 \times 480$ at a spatial resolution around $27 \times 27 \times 27$ μm^3. In this dataset, polydactyly was identified in 10 out of the total 60 limbs. Figure 3.9 illustrates an example raw image of a mouse embryo from Dataset B in sagittal and axial views, and its intensity histogram.

3.5 Detailed Methodology

3.5.1 Image denoising

To start with, all raw images undergo a standard image denoising process using Gaussian smoothing. This is to reduce artefacts generated during imaging and to smooth intensity variance across a small neighbourhood for each voxel.

3.5.2 Mouse embryo extraction

Since each mouse embryo is contained in a glass test tube and soaked in the maintenance solution (which may also contain air such as in our Dataset B), the first major computational step is to extract the embryo from its unwanted surroundings for subsequent processing. A simple histogram study reveals that signals tend to be distributed into several clusters, with the first two being *tube* and *solution* (denoted k_1, k_2) respectively (if there is a significant amount of *air* present, there would be an additional cluster k_0 preceding these two). The residuals are generally part of the *mouse* (k_3), with a rather bizarre pattern due to different reflection properties of tissue types.

Given this, we perform a simple multi-scale histogram smoothing, with the voxel count $\theta_b^{(i)}$ of bin b at each stage i updated by averaging values over $[b^{(i)} - \alpha^{(i)} \cdot N_b,\ b^{(i)} + \alpha^{(i)} \cdot N_b]$, where N_b indicates the total number of bins and $\alpha^{(i)}$ is a radius factor. The histogram shall then become much less steep, with most local maxima/minima smoothed out.

Next, we proceed to localise the clusters. For simplicity we use the first three (four) peaks in the histogram, respectively, p_1, p_2, p_3 (and p_0) drawn from the set $S_p = \left\{ b \;\middle|\; \frac{d\theta_b}{db} = 0,\ \frac{d^2\theta_b}{db^2} < 0 \right\}$ in ascending order, to represent the positions of k_1, k_2, k_3 (and k_0). A binary thresholding is applied thereafter to separate k_3 from k_1, k_2 (and k_0), by setting the threshold to the valley between p_2 and p_3:

$$\Phi_1 = \min \left\{ b \;\middle|\; \frac{d\theta_b}{db} = 0,\ \frac{d^2\theta_b}{db^2} > 0,\ b > p_2 \right\} \tag{3.3}$$

After that, the largest foreground region is extracted, while the smaller regions, mostly being k_2 voxels misidentified as k_3, are re-labeled as background. The label map is then dilated with a standard ball-shaped kernel to re-include k_3 voxels being accidentally ruled out during thresholding.

3.5.3 Creation of the local mouse template

As described earlier, in order to effectively estimate a representative feature space during RPCA processing, all mouse embryo images must be group-wise aligned in advance. To reduce bias, we create an unbiased control template using local normal subject images via the following procedure (as sketched in Figure 3.10):

First, one of the candidate images is randomly selected as an initial reference, with the rest of them rigidly registered to it. The rigidly-aligned images are then averaged to generate a new reference, which is blurry but unbiased toward the geometry of the initial reference. The aligned images are then affinely registered to the rigid-mean reference, followed by further averaging to generate an affine-mean reference. After that, a B-spline non-rigid registration is iteratively performed for local alignment, with both the reference and images updated after each iteration, until boundaries become sharp and anatomical structures become clear.

The B-spline registration uses a metric function composed of mutual information and a bending energy penalty term, and runs a multi-resolution scheme with the spacing of control points

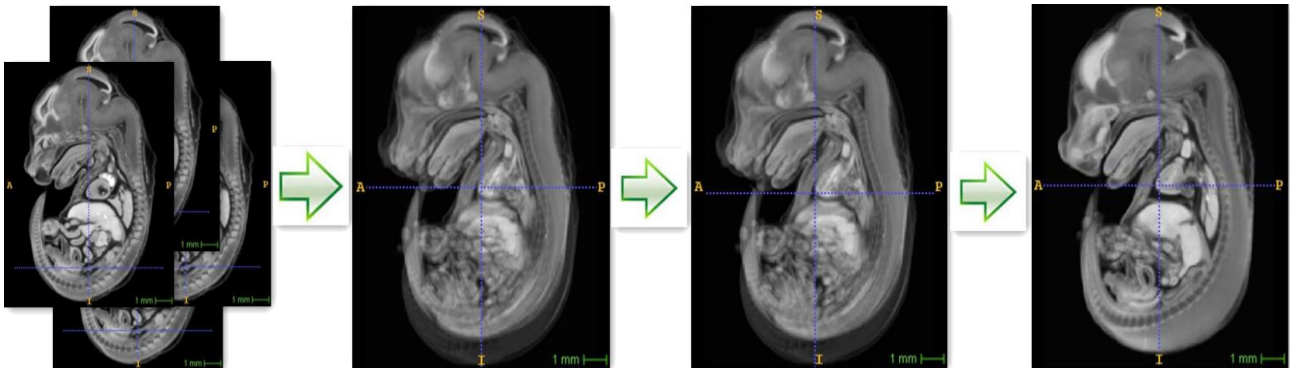


Figure 3.10: Creation of the mouse template through rigid, affine and (iterative) non-rigid image alignment and averaging, using local normal control subject images

gradually reducing over time. Furthermore, a stochastic gradient descent technique is employed using small sub-samples retrieved from a randomised region for each optimisation step. Such setting can significantly reduce the computation time for registration while retraining an equal level of performance. The actual implementation of the registration algorithm in our work was based on the open source Elastix toolbox[7] [86], and the parameter settings in different registration steps are listed in the Appendix.

3.5.4 Group-wise non-rigid image alignment

Once the template is generated, we proceed to perform group-wise non-rigid image alignment. The purpose of this process is centred on reducing physiological motions and inter-subject variations in the test dataset, so that the feature space can be best estimated. The alignment scheme follows a standard pipeline with three steps: a rigid and then an affine registration to secure a global correspondence, followed by a B-spline non-rigid registration to establish a localised one-to-one correspondence. All target images after embryo extraction in each experiment are collectively registered to the template, in order to achieve group-wise alignment with minimal bias. The registration configuration is similar to that used in template creation, and the influence of key parameter settings will be examined in Section 3.6.

3.5.5 Feature decomposition using RPCA

Due to the individualism of natural variation from one anatomical structure to another, which will be discussed in more details in Chapter 4, in this study we apply a mask around the heart region for the detection of VSD, and another mask covering the limb regions for polydactyly, in order to secure a good feature decomposition performance regarding the target phenotypes. These masks are manually checked prior to RPCA computation to ensure they cover the desirable regions in each of the target images. The use of mask is a constraint which will be released in the next Chapter by the introduction of a modified technique (RPCA-P) that incorporates prior knowledge of natural variation. On the other hand, it also dramatically lowers compu-

[7] Elastix official website: `http://elastix.isi.uu.nl`

tational volume and memory consumption. In our case, it reduces $m = w \times h \times d$ to a small fraction m', by which we formulate:

$$\{R', S'\} \quad = \quad \underset{R',S'}{\arg\min}. \quad ||R'||_* + \lambda||S'||_1 \qquad \text{s.t.} \quad D' = R' + S' \tag{3.4}$$

$$\text{where} \quad D', R', S' \in \mathbb{R}^{m' \times n}$$

and solve it using the IALM algorithm [103], which will be detailed in Chapter 4 Section 4.2.2. After that, a target image I_i and its decompositions $I_{i,r}$, $I_{i,s}$ can be easily reconstructed using the corresponding D'_i, R'_i, $S'_i \in \mathbb{R}^{m' \times 1}$.

3.5.6 'Normal-vs-Abnormal' classification

In theory, a perfect non-rigid registration should be able to critically deform all normal subject images so that they perfectly align with the reference template, while leaving morphological abnormalities in the defective subjects poorly aligned. These are then solely captured by S' in the RPCA process. However, registration often does not work to the demanding quality in practice, and some minor misalignment almost always occurs near the boundaries between different anatomical structures, due to significant inter-subject variations.

In order to deal with this problem, we develop a special metric, the anomaly rate (denoted as Ω), to calculate the level of morphological abnormality identified in each image/region, and compare it with a baseline level to determine whether the corresponding image/region would be annotated as normal or abnormal:

$$\Omega_i = \frac{||S'_i||_0}{m'}, \qquad L_i = \begin{cases} 1, & \Omega_i > \mu[\Omega] + \eta \cdot \sigma[\Omega] \\ 0, & \Omega_i \leq \mu[\Omega] + \eta \cdot \sigma[\Omega] \end{cases} \tag{3.5}$$

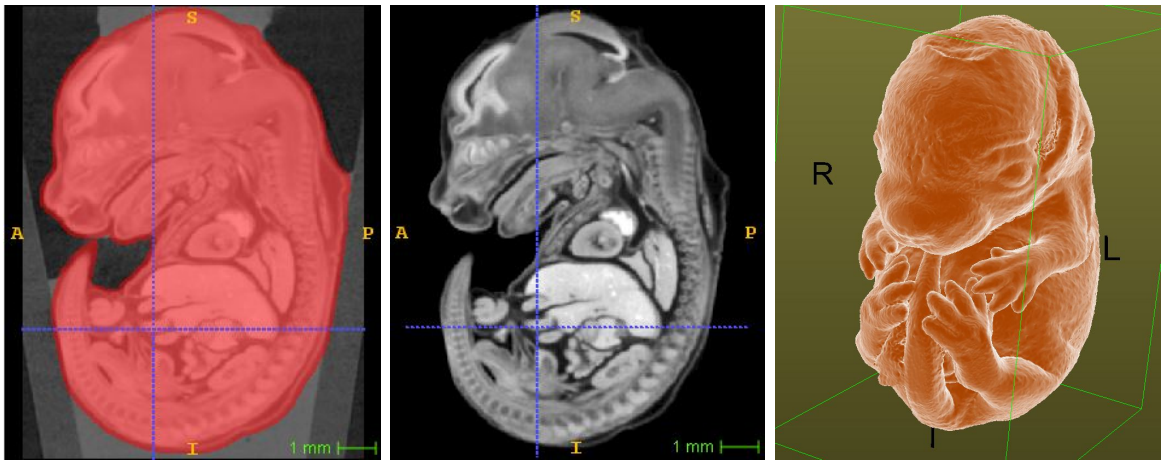
where $\mu[\Omega]$ and $\sigma[\Omega]$ represent the mean and standard deviation of anomaly rates across the whole dataset, and η is a tunable parameter used to adjust the baseline anomaly rate. $L_i = 1$ indicates image I_i (or I'_i at the region level) being considered abnormal and vice versa.

3.6 Evaluation

A series of experiments have been carried out on the aforementioned two test datasets to evaluate the proposed framework. The key to a successful feature decomposition and abnormality detection has been found to be associated with the setting of control point spacing in non-rigid registration and the tuning of the tolerance parameter λ in Eq. (3.4). The impact of major parameter settings will be detailed in the following sections, prior to which, the results of mouse embryo extraction and template creation will be briefly discussed to ensure the completeness of evaluation.

3.6.1 Results of mouse embryo extraction

The mouse embryo extraction results were assessed qualitatively by domain experts and found to be satisfactory, in particular a 3D model was reconstructed for each extracted mouse image to facilitate evaluation. For instance, a sample raw image of mouse embryo superimposed with its label map generated by the algorithm is shown in Figure 3.11 (a), and was used to obtain the extracted embryo in Figure 3.11 (b), which was then successfully portrayed in 3D with a high level of detail, including head, body, tail as well as the number of fingers and toes, as



(a) (b) (c)

Figure 3.11: Example mouse embryo extraction outcome: (a) a raw image superimposed with the label map generated by our algorithm (b) the embryo image after extraction (c) 3D reconstruction of the extracted embryo

shown in Figure 3.11 (c). Quantitative measures (such as Dice score) however were difficult to compute since there was no ground truth information available. On the other hand, such measures were not part of the goal here for morphological abnormality detection.

In our best experimental performance, we set $N_b = 1000$ and chose a three-stage processing with $a^{(i)} = 1/25, 1/35, 1/45$, respectively, for the multi-scale histogram smoothing. This however is a simple guideline since there could be numerous settings to achieve this task, as long as most local maxima/minima are smoothed out and intensity distribution of the clusters follows a Gaussian-like pattern. Given such a condition, a more straightforward idea for cluster localisation may be to fit a Gaussian mixture model as used in human brain tissue classification [173]. The model is expected to capture the mean μ_k and standard deviation σ_k of each cluster, and once obtained the threshold can then be simply set to $\Phi = \mu_2 + 2\sigma_2$ or similar values. However this does not work well in practice, because the mouse cluster does not follow a simple Gaussian distribution. Modelling the mouse itself as a Gaussian mixture is also difficult, as there is no method to determine how many sub-clusters would fit the data.

In addition, another approach that might be considered with CT images is to define an absolute threshold using Hounsfield Unit (HU). As explained in Chapter 1 Section 1.2.1, HU is a normalised CT intensity scale with respect to water, computed by:

$$HU_x \;=\; 1000 \times \frac{\mu_x - \mu_{water}}{\mu_{water}} \tag{3.6}$$

where μ_x is the average linear attenuation coefficient at voxel x and μ_{water} represents the attenuation level of water. The HU values in a CT image usually range from -1000 to +1000, with the air corresponding to around -1000 HU, water to 0 HU, and bone to around +1000 HU. In this case, if the HU distribution of the maintenance solution is known in advance and consistent over different datasets, the threshold can be derived using its upper bound. This however was unknown in our study. In addition, the intensity scale in Dataset A conforms to the HU standard while Dataset B does not, therefore a universal threshold was not applicable. On the other hand, our approach is more general and identifies the desirable threshold automatically, regardless of the use of different imaging protocols and maintenance solutions.

3.6.2 Results of template creation

All the normal subject images in each dataset were collectively used to create a local template. The interim result right after group-wise linear processing (alignment and averaging) was very blurry, yet it was quickly refined as non-rigid processing took place. After five iterations of non-rigid processing, boundaries had already become sharp. After eight/nine iterations improvement became marginal. The reference template right after linear processing, and after five, eight, ten iterations of non-rigid processing are shown in Figure 3.12. A notable point is

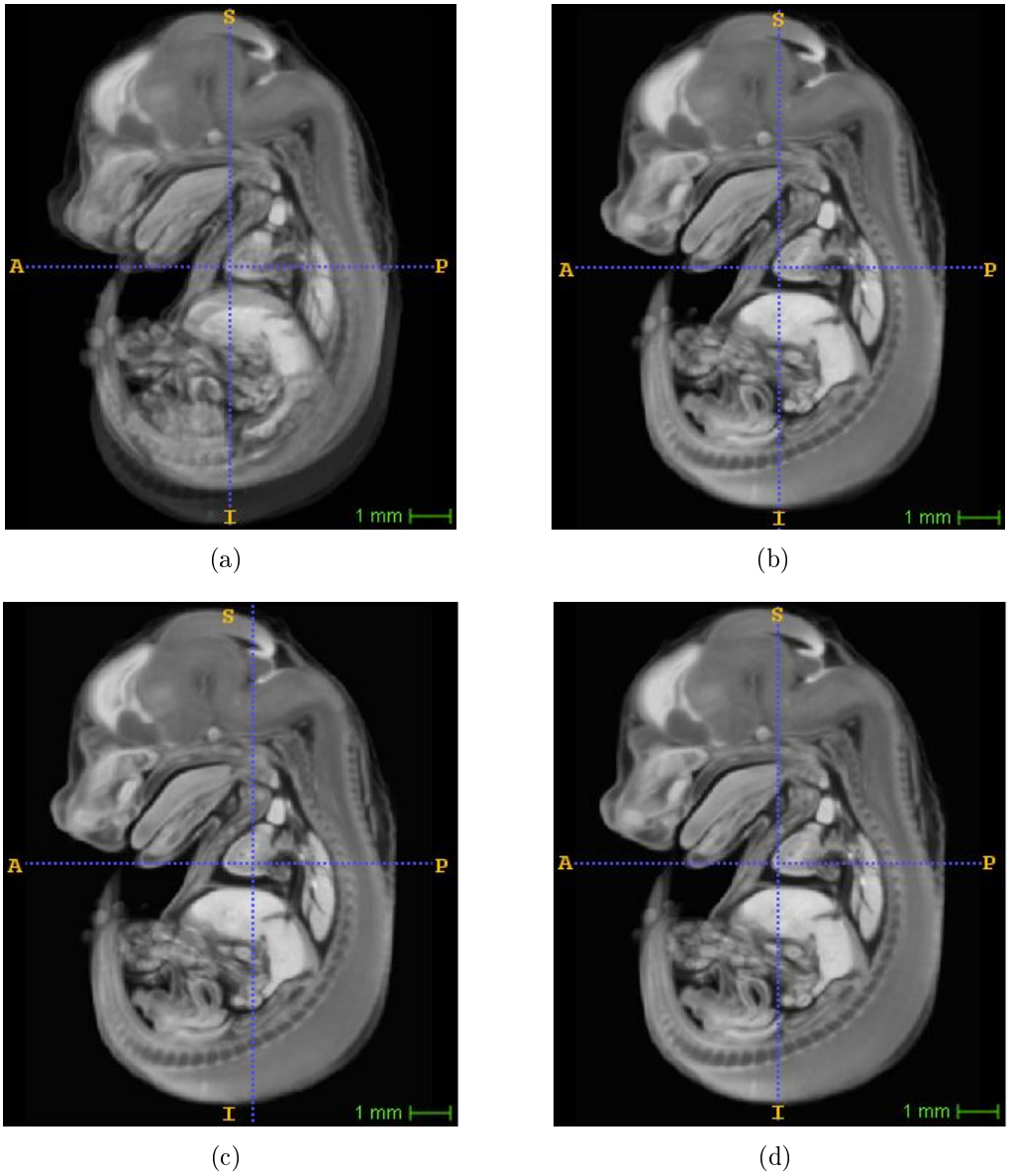


Figure 3.12: Template updates (a) right after group-wise linear processing (b) after 5 iterations (c) 8 iterations (d) 10 iterations of non-rigid processing

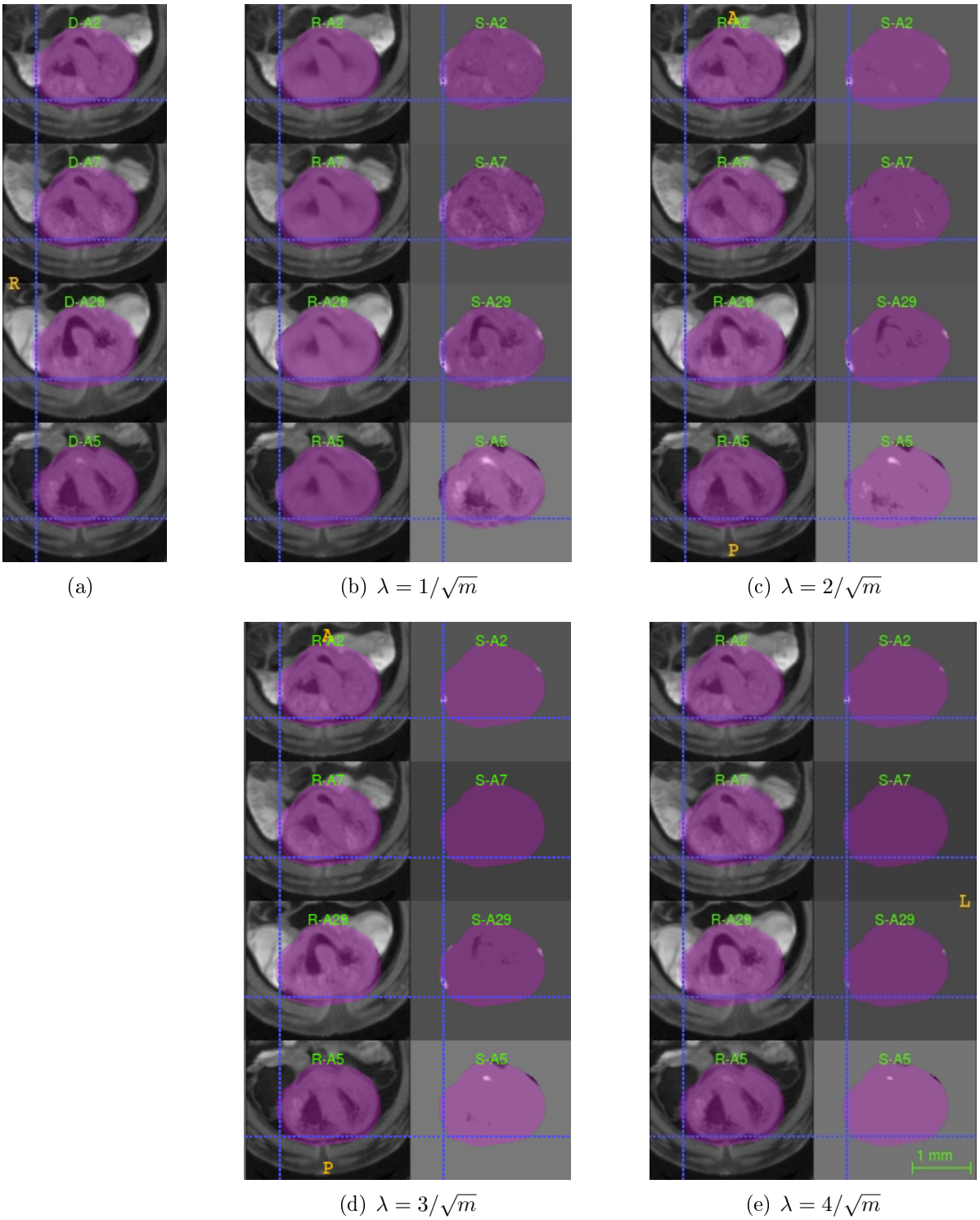
that abdominal structures will always appear more blurry due to substantial natural variations. In our work, we carried out ten iterations before finalising the template, which was then used as the reference for group-wise image alignment.

In fact, warping images to a representative template space for comparative analysis is by any means not new. Rather it is a common practice in many biomedical image computing studies. There have been a number of paradigms proposed for unbiased template creation [162, 134, 77], more or less akin to ours. The key insight here is that by creating a "standard mouse" using local normal control data for group-wise non-rigid alignment, we are able to minimise the overall inter-subject variations across the local imaging dataset, so that the regular component can include as much data as possible in the RPCA decomposition process, ideally with only abnormalities remaining in the singular component.

3.6.3 Results of feature decomposition: the influence of tolerance parameter λ in RPCA processing

The parameter λ governs the level of tolerance/strictness for potential outliers to be considered normal. In other words, it affects whether marginal features should be included in R' or S'. The tolerance factor takes effect at the whole batch level rather than at single image level. A working setting is generally within the order of $1/\sqrt{m}$, based on the insights from previous studies [32]. Figure 3.13 compares the decomposition results of four settings, respectively, $\lambda = 1/\sqrt{m}$, $\lambda = 2/\sqrt{m}$, $\lambda = 3/\sqrt{m}$ and $\lambda = 4/\sqrt{m}$, applied to four different subject images from Dataset A regarding the VSD phenotype. Similar comparison across three cases from Dataset B regarding the polydactyly phenotype are illustrated in Figure 3.14. In both examples, the final control point spacing for non-rigid registration was set to $200\mu m$ (to be detailed in Section 3.6.4). A close observation will discover that the reconstructed I_r images are more similar to each other when λ is lower. Then as λ increases, more information in S' is transferred to R', where only salient features that are more distant from "normality" are retained.

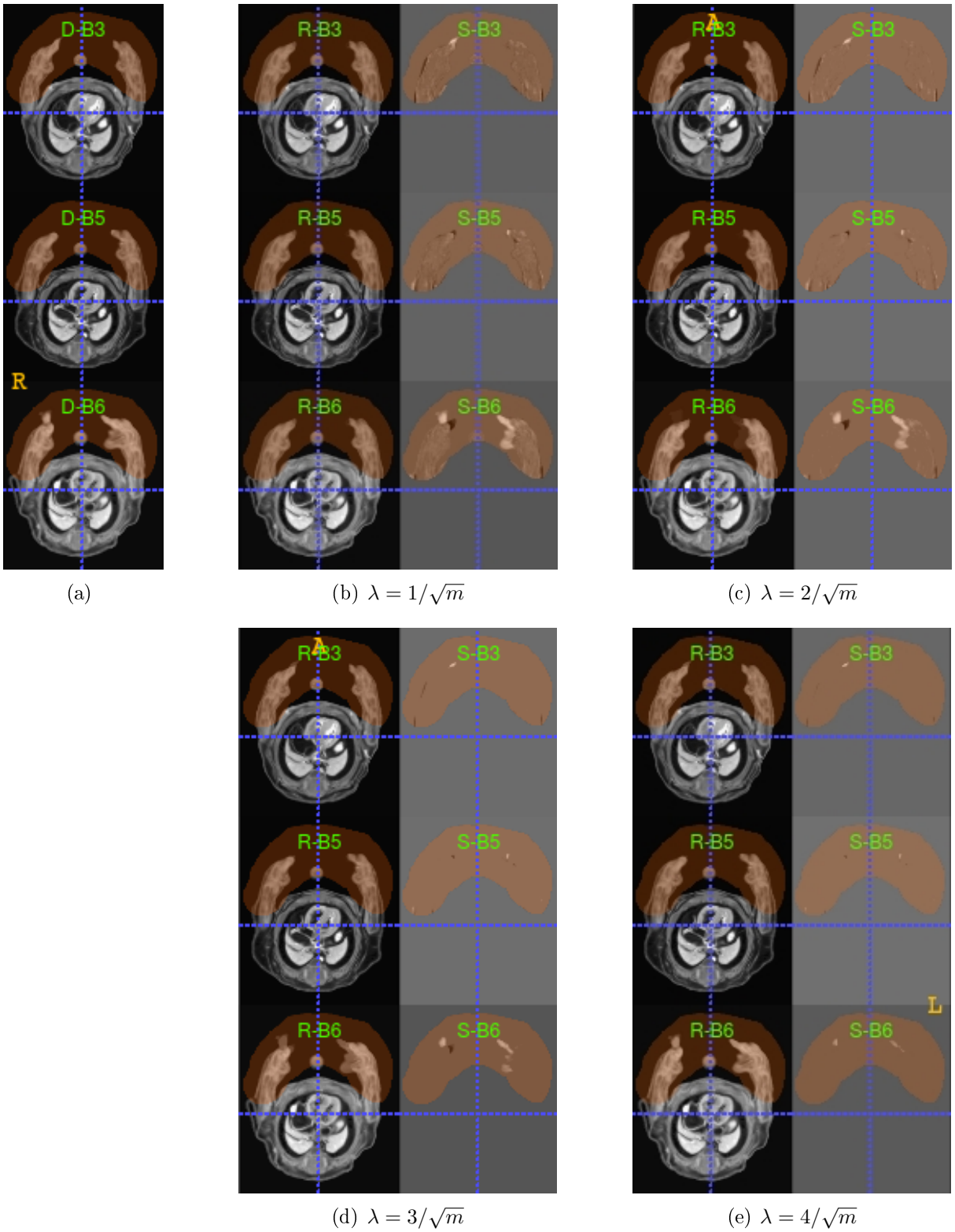
In the test to Dataset A, when $\lambda = 3/\sqrt{m}$, the majority of normal subjects ended up with an



(a) (b) $\lambda = 1/\sqrt{m}$ (c) $\lambda = 2/\sqrt{m}$

(d) $\lambda = 3/\sqrt{m}$ (e) $\lambda = 4/\sqrt{m}$

Figure 3.13: Example RPCA decomposition results on Dataset A (the images are cropped to better show the heart region, which is coloured in purple): the original images (a) are decomposed into a regular (left) and a singular component (right), with (b) $\lambda = 1/\sqrt{m}$, (c) $\lambda = 2/\sqrt{m}$, (d) $\lambda = 3/\sqrt{m}$, and (e) $\lambda = 4/\sqrt{m}$. Each row in the subfigures shows a different subject. Rows 1-2 are typical of normal controls and Row 3 is a subject of the VSD phenotype. Row 4 on the other hand shows a normal subject with significant natural variation, in which case its individual features are more likely to be retained in the singular component.



(a) (b) $\lambda = 1/\sqrt{m}$ (c) $\lambda = 2/\sqrt{m}$

(d) $\lambda = 3/\sqrt{m}$ (e) $\lambda = 4/\sqrt{m}$

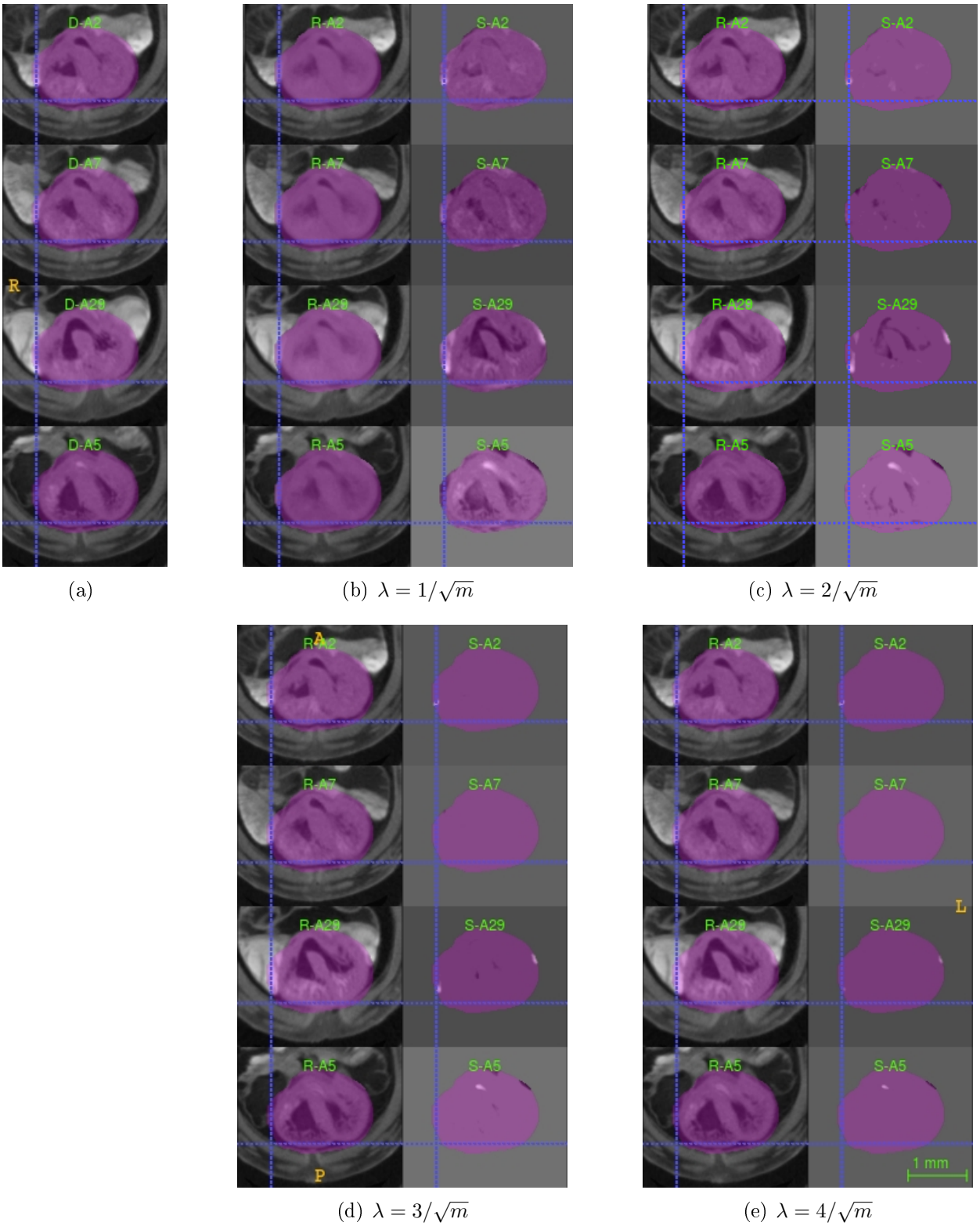
Figure 3.14: Example RPCA decomposition results on Dataset B (the limb region is coloured in orange): similarly, the original images (a) are decomposed to a regular (left) and a singular component (right), with (b) $\lambda = 1/\sqrt{m}$, (c) $\lambda = 2/\sqrt{m}$, (d) $\lambda = 3/\sqrt{m}$, and (e) $\lambda = 4/\sqrt{m}$. Each row shows a different subject. Rows 1-2 are typical of normal controls, and Rows 3 is a subject with polydactyly phenotype.

almost empty I_s. Whereas in subjects with VSD, the features signifying ventricular connectivity were retained in I_s. However, raising λ to $4/\sqrt{m}$ led to most anomalous features being included in R' and remaining undetected. The test to Dataset D showed a similar pattern, yet features signifying polydactyly remained more salient than the normal controls at $\lambda = 4/\sqrt{m}$. In addition, a small proportion of normal subjects manifested significant individual traits, and the associated salient features were more likely to be retained in I_s, causing false positive detection described in Section 3.6.5. This situation occurred more commonly in the detection of cardiac abnormality than limb abnormality, probably due to more significant inter-subject variability in the heart.

3.6.4 Results of feature decomposition: the influence of registration parameters in group-wise image alignment

The rigid and affine registrations are more standardised and do not involve complex parameter settings, whilst the non-rigid registration part is a more difficult. We tested three settings, all using a four-scale multi-resolution scheme with maximally 4000 iterations at each resolution level. The spacing of control points was halved at every new resolution level, with the final spacing, respectively, set to $400\,\mu m$, $200\,\mu m$, $100\,\mu m$ in these settings.

We found feature decomposition did not work desirably with the $400\,\mu m$ setting with respect to the VSD phenotype, which, as shown in Figure 3.15, almost led to empty I_s images at $\lambda = 3/\sqrt{m}$ for the same four subjects in Figure 3.13. Decomposition performance improved dramatically as the spacing narrowed down to $200\,\mu m$ and $100\,\mu m$. This implies feature similarity and outlier tolerance are determined in relative terms across the image batch rather than in absolute terms. RPCA performance tends to improve when normal samples are better aligned against subjects with salient anomalous features. This can be achieved through a more localised non-rigid registration, in which the registration is properly regularised so that anomalous features are not eliminated. However such improvement may diminish quickly while computational cost rises dramatically: a trade-off needs to be balanced in practice. In general, the $200\mu m$ and $100\mu m$ settings were found to be both effective and relatively efficient in our experiments.



(a) (b) $\lambda = 1/\sqrt{m}$ (c) $\lambda = 2/\sqrt{m}$

(d) $\lambda = 3/\sqrt{m}$ (e) $\lambda = 4/\sqrt{m}$

Figure 3.15: Example RPCA decomposition results on the same four samples from Dataset A, with the $400\mu m$ non-rigid registration setting: the original images (a) are decomposed into a regular (left) and a singular component (right), with (b) $\lambda = 1/\sqrt{m}$, (c) $\lambda = 2/\sqrt{m}$, (d) $\lambda = 3/\sqrt{m}$, and (e) $\lambda = 4/\sqrt{m}$. In this setting, the decomposition almost end up with an empty singular component at $\lambda = 3/\sqrt{m}$ for all four subjects.

3.6.5 Abnormality detection performance

In our study, each setting was tested 31 and 15 times, respectively, on Datasets A and B via a leave-one-out strategy. In each test, non-rigid registration and RPCA were applied on all the remaining 30 and 14 images, with $\eta = 0.5$ for heart anomaly and $\eta = 0.3$ for limb anomaly, based on empirical experience. Anomaly detection was measured at the region-level, where the heart and each of the four limbs in a mouse were treated separately. A true positive (TP) or false negative (FN) detection indicates an abnormal region being correctly labelled as abnormal or otherwise, and a true negative (TN) or false positive (FP) indicates a normal case being correctly labelled normal or otherwise. In terms of metrics, the sensitivity, specificity as well as the overall accuracy were calculated:

$$SEN = \frac{TP}{TP + FN} \qquad SPE = \frac{TN}{TN + FP} \qquad ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.7}$$

The performance metrics on both datasets across all three registration settings with $\lambda = 3/\sqrt{m}$ are illustrated in Table 3.1 and Table 3.2 in a comparative fashion. The low standard deviation

Table 3.1: Performance metrics of abnormality detection on Dataset A across all three settings of final control point spacing for non-rigid registration

Dataset A			
	SEN	SPE	ACC
$400\mu m$	$50.00 \pm 6.09\%$	$88.89 \pm 1.15\%$	$83.87 \pm 1.25\%$
$200\mu m$	$100.00 \pm 0.00\%$	$85.19 \pm 1.29\%$	$87.10 \pm 1.13\%$
$100\mu m$	$100.00 \pm 0.00\%$	$81.48 \pm 1.42\%$	$83.87 \pm 1.25\%$

Table 3.2: Performance metrics of abnormality detection on Dataset B across all three settings of final control point spacing for non-rigid registration

Dataset B			
	SEN	SPE	ACC
$400\mu m$	$100.00 \pm 0.00\%$	$66.67 \pm 4.72\%$	$80.00 \pm 2.96\%$
$200\mu m$	$100.00 \pm 0.00\%$	$88.89 \pm 3.15\%$	$93.34 \pm 1.84\%$
$100\mu m$	$83.33 \pm 4.88\%$	$88.89 \pm 3.15\%$	$86.66 \pm 2.52\%$

in the metrics indicates a stable performance across different experimental trials. A notable point is that in this particular problem domain, sensitivity is considered the first priority, and the general guideline is normally to maximise sensitivity before specificity (although this may not necessarily always be the case). For that reason, with a mean sensitivity of 50%, the performance of the 400 μm registration setting was considered almost a failure for VSD phenotype detection, although the mean specificity was the highest, at 88.89%.

Its performance on polydactyly was more acceptable with 100% sensitivity and 66.67% mean specificity. Significant improvements were seen on both datasets as the spacing narrowed down to 200 μm: VSD sensitivity quickly rose to the desirable level of 100%, although mean specificity lowered to 85.19%; and in the case of polydactyly detection, the mean specificity rose to 88.89% while maintaining sensitivity at 100%. Further narrowing down the control point spacing to the 100 μm setting however, did not produce further improvement. Rather, a somewhat surprising minor decline was witnessed, with the mean VSD specificity lowered to 81.48% and the mean polydactyly sensitivity lowered to 83.33%.

3.6.6 Experimental Environment and Computation Time

The runtime environment of our experiments was deployed on a standard PC with an Intel i7 3.4GHz quad-core CPU and 32GB RAM memory. Taking all runs of cross validation into account, in the case of Dataset A, on average mouse embryo extraction took 57s per image, non-rigid registration took around 14 min per image and RPCA took only 68s for each image batch, which summed to around 451 min for the whole batch (30 images) or 15 min per image.

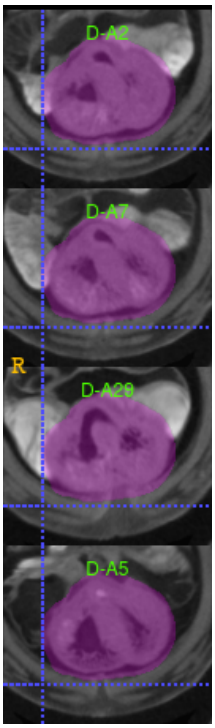
For Dataset B, on average mouse embryo extraction took 42s per image, non-rigid registration to the template took 10 min per image, RPCA processing took 301s for the whole batch, which summed to around 155 min at the batch level (14 images) or 11 min per image. This is considered substantially high-throughput in phenotyping practices. The reason for RPCA on Dataset B to be longer than Dataset A was mostly due to the use of a larger mask to cover limbs than the mask for heart. In addition, template creation may take up to several hours depending on data size and the number of iterations but is done in advance and once for all.

Furthermore, since all steps other than RPCA are independent for each image sample, computational efficiency can be further boosted with parallel setting for all non-RPCA processing. In a separate experiment under a cluster computing environment, the computation time reduced dramatically to a small fraction, with only 22 *min* and 25 *min* for Datasets A and B at the batch level respectively.

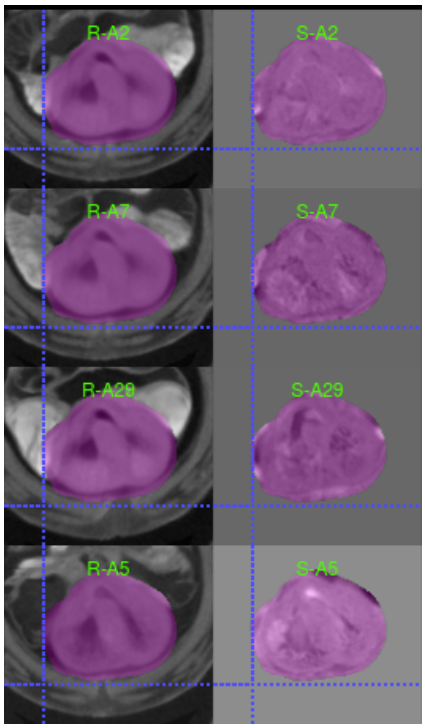
3.7 Comparison with the baseline PCA approach

In further evaluation, additional experimentation was conducted using the baseline PCA approach for direct comparison, in order to examine our earlier suggestion regarding its infeasibility in terms of feature decomposition and abnormality detection in this scenario. To ensure a fair comparison, the computation followed an identical procedure to the original methodology detailed in Section 3.5, except that the RPCA decomposition was replaced with the classic PCA method. More specifically, a set of principal components were learned from the observation data D, where the most representative r components were then used to reconstruct R, and the residual $S = D - R$ was used for abnormality detection. In this case, r is the only parameter to tune and resembles the function of λ in Eq. (3.2).

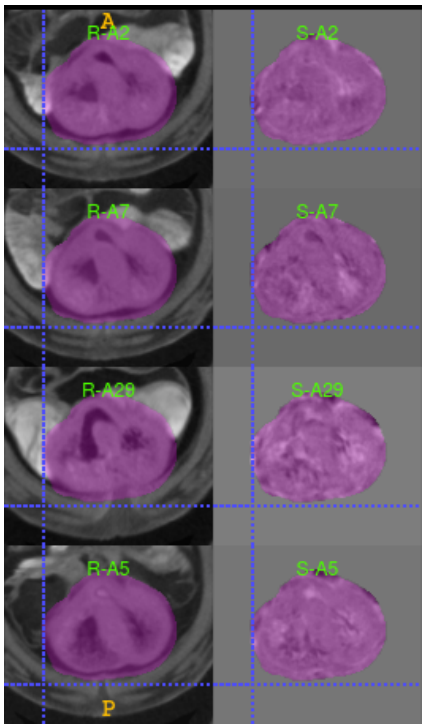
In the test applied to Dataset A, the outcome of the PCA-based feature decomposition on the same four subjects as in the original experiments are illustrated in Figure 3.16, with an increasing number of principal components used for R reconstruction. By comparison with Figure 3.13, it can be easily identified that replacing RPCA with baseline PCA led to a wide prevalence of noise that almost corrupted every single entry in S, except for the case using all principal components ($r = 30$, note there are only 31 images in Dataset A) which perfectly reconstructed R, leaving S completely empty. This shows that PCA lacks the ability to separate data anomaly from regularity, and its detection performance with the use of 400μ, 200μ and 100μ registration settings is shown in Table 3.3 in comparison to the best result obtained in the RPCA-based approach. A similar finding was confirmed in the counterpart test applied to Dataset B, where Figure 3.17 illustrates the feature decomposition outcome for comparison



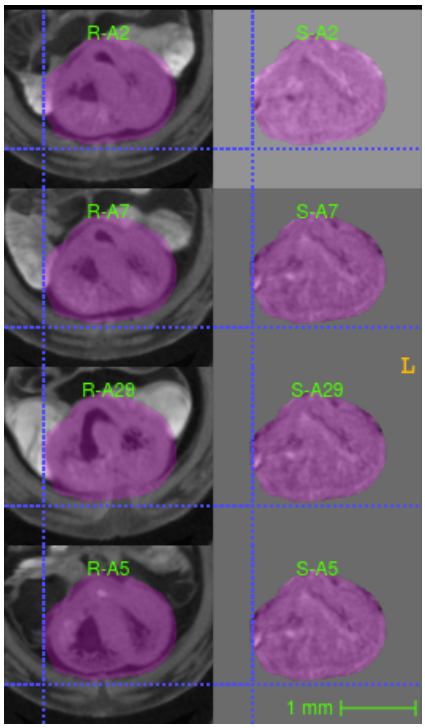
(a)



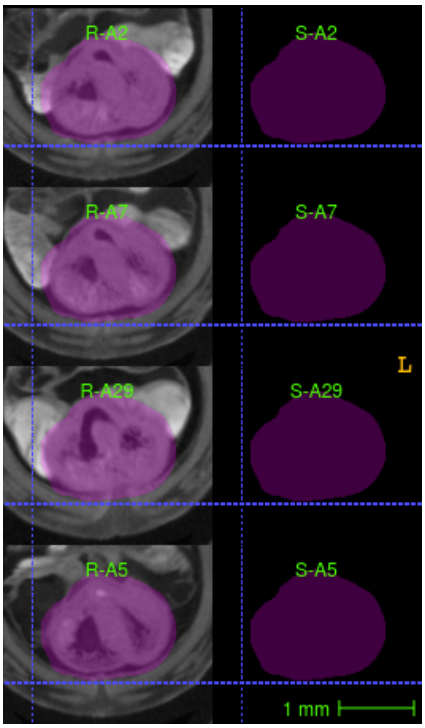
(b) $r = 3$



(c) $r = 17$

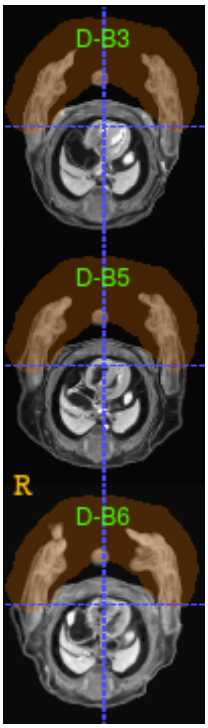


(d) $r = 29$

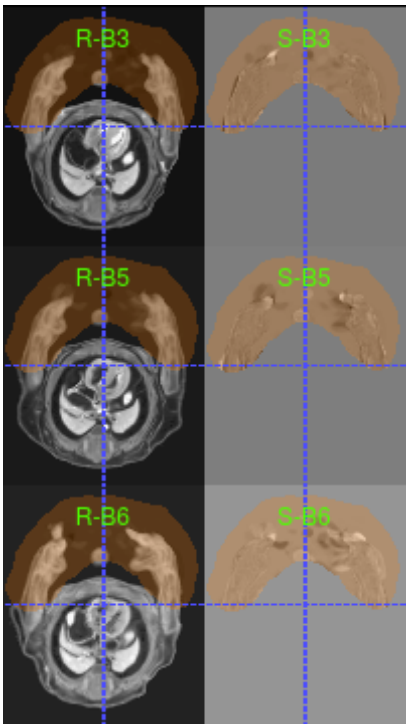


(e) $r = 30$

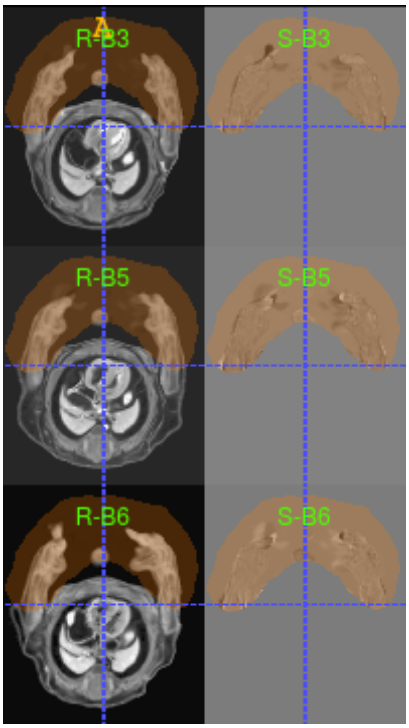
Figure 3.16: PCA-based feature decomposition on Dataset A, applied to (a) the same four samples as in Figure 3.13 based on the 200μm non-rigid registration setting: with the use of (b) $r = 3$, (c) $r = 17$, (d) $r = 29$, and (e) $r = 30$ principal components for regular component reconstruction.



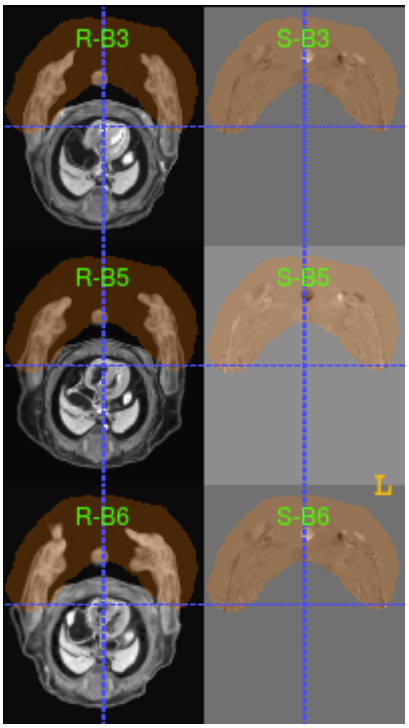
(a)



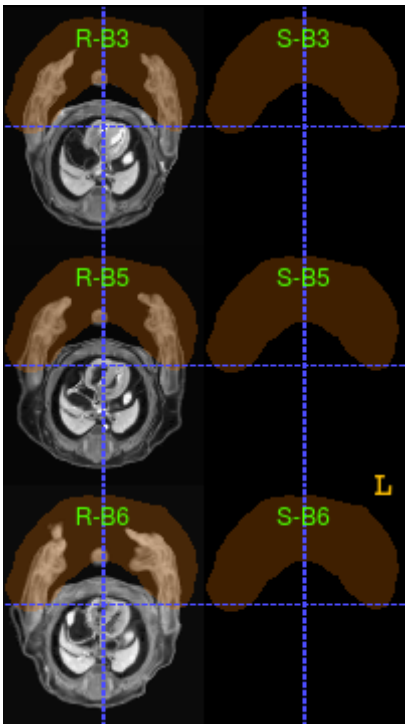
(b) $r = 3$



(c) $r = 11$



(d) $r = 13$



(e) $r = 14$

Figure 3.17: PCA-based feature decomposition on Dataset B, applied to (a) the three four samples as in Figure 3.14 based on the 200μm non-rigid registration setting: with the use of (b) $r = 3$, (c) $r = 11$, (d) $r = 13$, and $r = 14$ principal components for regular component reconstruction

Table 3.3: Performance metrics of abnormality detection on Dataset A using the baseline PCA approach as compared to the RPCA approach

Dataset A			
	SEN	SPE	ACC
RPCA	$100.00 \pm 0.00\%$	$85.19 \pm 1.29\%$	$87.10 \pm 1.13\%$
PCA $400\mu m$	$100.00 \pm 0.00\%$	$5.88 \pm 1.36\%$	$23.81 \pm 2.18\%$
PCA $200\mu m$	$100.00 \pm 0.00\%$	$29.41 \pm 3.15\%$	$42.86 \pm 2.54\%$
PCA $100\mu m$	$100.00 \pm 0.00\%$	$17.65 \pm 2.20\%$	$33.33 \pm 2.42\%$

Table 3.4: Performance metrics of abnormality detection on Dataset B using the baseline PCA approach as compared to the RPCA approach

Dataset B			
	SEN	SPE	ACC
RPCA	$100.00 \pm 0.00\%$	$88.89 \pm 3.15\%$	$93.34 \pm 1.84\%$
PCA $400\mu m$	$100.00 \pm 0.00\%$	$7.41 \pm 5.86\%$	$18.67 \pm 3.05\%$
PCA $200\mu m$	$100.00 \pm 0.00\%$	$14.81 \pm 6.32\%$	$28.00 \pm 1.69\%$
PCA $100\mu m$	$100.00 \pm 0.00\%$	$22.22 \pm 4.17\%$	$22.81 \pm 2.24\%$

with that in Figure 3.14. Table 3.4 shows the corresponding abnormality detection performance. Both these tests corroborated our suggestion and demonstrated the superiority of RPCA to the classic PCA in this particular problem.

3.8 Discussion

Phenotype assessment is an advanced process requiring high levels of domain knowledge. Some studies have been proposed to automatically classify certain phenotypes, however these approaches generally tend to be over-specific to phenotypical features and unable to achieve general-purpose detection of abnormal phenotypes. Moreover, even after the use of such automatic classifiers, the confirmation of exact phenotypes in practice usually still requires expert manual assessment (and often involve advanced histological examination), due in part to the complex nature of pathology diagnostics, as well as the problem that the C57BL/10 strain is still under study and hence the complete phenotypical characteristics regarding morphological abnormality are yet unknown.

The proposed approach on the other hand aims to equip domain experts with a mechanism that is able to automatically identify salient morphological anomaly in imaging data, so that the search space can be narrowed considerably for the discovery of potential defective phenotypes. Yet a notable point is that an "anomalous region" signified by the algorithm does not necessarily always indicate the presence of an abnormal phenotype, as there might be other factors causing the anomaly, although the detection of a morphological anomaly is usually correlated to certain biological significance. For instance, an accidental removal of the umbilical cord (locating close to the toes) during data acquisition had occurred in one particular subject image, and constantly led to false positive detection of polydactyly. Also, stronger individual variation in a subject is more likely to trigger false positive detection.

In the wake of substantial obstacles toward direct phenotype classification, identifying major morphological anomalies for manual phenotype examination is a widely used approach as in other phenotyping studies based on data-driven comparative analytics. With its capability to perform batch-wise anomaly detection in a short time, this high-throughput framework is able to significantly boost phenotyping efficiency, especially when applied to large data volumes.

Another notable point is that the 'normal-vs-abnormal' classification technique based on the anomaly rate described in Section 3.5.6, to some extent, can be cast as a special type of (one-tailed) hypothesis testing, where an Ω statistic is computed over the target data and then compared to a baseline level, which is derived from the estimation of overall data distribution (parametrised by $\mu[\Omega]$ and $\sigma[\Omega]$) and a manually chosen significance level (parametrised by η). By simple extension, a p-value can be calculated once the distribution is estimated, and the use of parameter η to adjust the baseline level can be substituted by a p-value threshold instead.

The merits of such re-formulation include: (1) hypothesis testing is a widely-used concept familiar to a broad readership, (2) it links abnormality detection with probability theory, offering an easier way to make sense of the classification, (3) it also provides strong theoretical support to make the actual choice of threshold (especially when it comes to p-value), whereas the choice of η in Eq. (3.5) tends to be more *ad hoc*. This is particularly favourable when there are many repetitive tests involved, in which case state-of-the-art studies often leverage the techniques

regarding false discovery rate control [15, 16, 55]. The false discovery rate is defined as:

$$q \;=\; E\Big(\frac{FP}{FP+TP}\Big) \tag{3.8}$$

which is the expectation of the probability for a test to be false positive given it yields a positive output, in the context of multiple inferences. This concept was originally proposed by Benjamini and Hochberg [15], who also suggested a well-known technique to control q:

> Suppose there are m tests $H_1, .., H_m$ which yeild p-values $P_1 \leq .. \leq P_m$ sorted in ascending order, the false discovery rate can be controlled at q_* level if the threshold θ is set to
>
> $$\theta \;=\; P_k \qquad \text{where} \qquad k \;=\; \arg\max_i . \quad P_i \leq \frac{i}{m} \cdot q_* \tag{3.9}$$
>
> whereby $H_1, .., H_k$ are rejected (classified abnormal) and $H_{k+1}, .., H_m$ are accepted (classified normal). The value of q_* is adjustable and often set to 0.01 or 0.05, etc. A number of studies have since been carried out to enhance false discovery rate control in various scenarios [16, 55].

The line of research on hypothesis testing, especially multi-testing with false discovery rate control, can be highly relevant when our work is extended to consider the detection of multiple abnormalities in very large datasets. Its applicability however is limited in this pilot study, due to the difficulty to accurately estimate a data distribution with a small sample size, in particular an accurate estimation often requires the exclusion of outliers in the training stage, which does not align with the nature of this unsupervised learning problem.

3.9 Conclusion

In conclusion, we have proposed a high-throughput general-purpose mouse phenotyping framework. It features a systematic methodology that starts from image denoising, extraction of mouse embryo in the image, then goes on to group-wise non-rigid registration with reference to a template image created using local normal control data, followed by RPCA feature decomposition into a regular and a singular component, where the latter is then used to detect morphological abnormality.

The proposed framework has been tested on two μ-CT datasets, which, respectively, contain 31 mouse embryo subjects with four manifesting the VSD phenotype, and 15 subjects with 10 out of the 60 limbs manifesting the polydactyly phenotype. A set of empirical cross-validating experiments have been conducted, in which the best setting (200 μm for non-rigid registration and $\lambda = 3/\sqrt{m}$ for RPCA processing) achieved 100% detection sensitivity for both phenotypes, as well as a $85.19 \pm 1.29\%$ specificity and $87.10 \pm 1.13\%$ overall accuracy for the heart defects, and a $88.89 \pm 3.15\%$ specificity and $93.34 \pm 1.84\%$ overall accuracy for the limb defects.

This pilot study has verified our framework to be 1) both theoretically sound and empirically viable; 2) high-throughput, where anomaly detection can be performed on a batch of images simultaneously in a single run; 3) effective to both volumetric abnormalities such as polydactyly and non-volumetric such as VSD; 4) no segmentation is involved and no external atlas of any kind is required, making it low-cost and widely applicable. 5) feature extraction is performed post image alignment and does not rest on unreliable deformation features, thereby it is robust to various registration settings and template uses.

Chapter 4

Robust Principal Component Analysis with Variation Priors

4.1 Introduction

The mouse phenotyping framework using non-rigid image registration and robust principal component analysis (RPCA) was introduced in the last chapter, with a comprehensive methodology starting from image denoising, mouse embryo extraction, template creation, to group-wise non-rigid image alignment, RPCA processing and abnormality detection. As outlined in Chapter 3 Section 3.4, this framework is centred on group-wise feature extraction and decomposition into a regular and a singular component, where the latter is then used to detect morphological abnormalities. However, the methodology in Chapter 3 simply employed the baseline RPCA to implement a purely unsupervised data-driven approach to abnormality detection, without incorporating any prior information of natural variation. Furthermore, in terms of feature decomposition there was only a single criterion of outlier tolerance, controlled solely by the parameter λ and applied globally regardless of local variability. For that reason feature decomposition was applied to a dedicated region of interest (ROI) to overcome the individual natural variation from one anatomical structure to another, rather than to multiple structures/regions at the same time. In addition, the technical details of the RPCA method were not explained.

As mentioned earlier, the RPCA technique stemmed from the work of principal component analysis (PCA) [75], which is arguably the most widely used statistical data analytical technique concerning feature extraction and dimensionality reduction. PCA uses an orthogonal transformation to convert a given dataset with possibly correlated variables into a low-dimensional subspace (sometimes known as feature space or a new coordinate system) composed by linearly uncorrelated variables called principal components. However, it is well-known [32, 169, 41, 42] that the classic PCA algorithm is fragile to the presence of data anomalies or outliers, for its subspace learning assumes a multivariate Gaussian data distribution, and a single data point with gross corruption could significantly lower the quality of estimation. Such limitation alongside the wide prevalence of grossly corrupted data in real-world applications has stimulated considerable research effort toward the RPCA or robust subspace learning problem [32, 169, 41, 42, 21, 81], in particular in the computer vision domain.

In contrast to the classic PCA, RPCA aims to separate outliers from the regular data during subspace estimation. Early RPCA works include learning linear multivariate representations of the imaging data via robust M-estimations [41, 42], or via self-organising rules that incorporate additional binary decision fields and outlier prior distributions [169]. Other representative approaches include using iteratively re-weighted least squares [21], or alternating convex programming on robust L1 norm factorisation [81]. Nevertheless, none of these approaches yielded a polynomial-time algorithm with strong performance guarantees under broad conditions, until the seminal work of principal component pursuit (PCP) [163, 32], which has gradually become the standard approach of RPCA.

Prior to this study, the RPCA technique (PCP and its variants) has been applied to a number of computer vision problems, including video surveillance [163, 32, 8, 175, 174], face recovery [163, 32, 175, 174] and batch linear alignment of face images with partial occlusion/corruption [124], attaining excellent performances. Nevertheless, an important pre-condition of the baseline RPCA is the existence of a stationery background, which is mostly invariant throughout the entire dataset. In biomedical imaging, there is usually a structural congruity of biological anatomy across different subjects, which coarsely corresponds to this condition and yields the relevance of low-dimensional subspace learning. Yet meanwhile, biomedical data often

undergoes significant natural variations from one subject to another, leading to a number of challenges to the practical application of baseline RPCA.

This chapter will carry on this line of work and explore the RPCA framework in more depth. In particular, a modified framework (which we call RPCA-P) that extends the baseline technique to incorporate variation priors for feature decomposition will be proposed, which can significantly improve its practical performance in biomedical imaging and achieve better abnormality detection results.

4.2 The RPCA Framework

4.2.1 PCP problem formulation

The RPCA technique aims to decompose data (also known as observations) into a regular component and singular component that features a collection of gross but generally sparse corruptions, based on the following model:

$$D = R_0 + S_0 \quad \text{where} \quad D, R_0, S_0 \in \mathbb{R}^{m \times n} \tag{4.1}$$

The D here represents the observed data with corruptions, R_0 represents the uncorrupted form of D which lies in a low-dimensional subspace, and S_0 models the erroneous factors that cause the corruptions. In this sense, RPCA is closely related to two lines of research:

- The first line of work is often formalised as "matrix completion" [28, 29, 30, 59, 58, 82], which, in brief, is about recovering a low-rank matrix from the condition where only a small fraction of its entries are available. In some circumstances, it is extended to completing a matrix with only a small number of linear components/variables. The estimation of R_0 in the RPCA problem is somewhat similar to matrix completion. A prevailing and mathematically validated approach to it is to conduct convex optimisation on nuclear norm minimisation [28, 29, 30, 58, 82].

- The estimation of S_0 on the other hand, is inspired from another line of work that concerns robustly recovering under-determined linear systems of equations from arbitrary but sparse errors in polynomial time. This can be achieved by performing convex optimisation on $L1$ norm minimisation [31].

Combining the insights from both lines of work, the original PCP approach [32] was initially introduced to tackle to an idealised RPCA problem, in which S_0 is strictly sparse and contains errors that can be arbitrarily large in magnitude but affect only a small fraction of D. It was demonstrated that under minimal assumptions, R_0 and S_0 can be exactly recovered from D by formulating the following tractable convex optimisation problem:

$$\min_{R,S} . \quad ||R||_* + \lambda||S||_1 \qquad \text{s.t.} \quad D = R + S \tag{4.2}$$

where R and S are, respectively, the estimations of R_0 and S_0, $||M||_* = \sum_i \sigma_i(M)$ denotes the nuclear norm of a matrix M, which is the sum of its singular values, and $||M||_1 = \sum_{ij} |M_{ij}|$ denotes the element-wise 1-norm, the sum of its absolute values. Moreover, λ is a global weighting parameter that balances the optimisation process between the two sub-problems that, respectively, estimate R_0 and S_0. Based on the insights from their study [32], λ is usually set to $\lambda = 1/\sqrt{max(m,n)}$.

The idealised RPCA condition however, significantly limits PCP's ability to apply to many real-world problems. This is because besides the gross but sparse errors, data acquisition in real-world applications is often undermined by other types of noise, which may cause small perturbations but affect many or even all entries of D, in either a deterministic or stochastic pattern. For instance, there might be a mild change of background luminance from one acquisition procedure to another. In face recognition on the other hand, the human face is not a strictly convex, and does not exhibit the Lambertian reflectance property with perfect isotropic diffusion. Therefore a collection of face images acquired under lights from different sources often do not exactly satisfy the low-rank condition.

To address this issue, some researchers proposed a relaxed version of PCP formulation (called

"Stable PCP" [176]), which introduces another noise term N and converts Eq. (4.2) to:

$$D_0 = R_0 + S_0 + N_0 \quad \text{where} \quad D_0, R_0, S_0, N_0 \in \mathbb{R}^{m \times n} \tag{4.3}$$

The optimisation problem in this case is re-formulated as:

$$\min_{R,S} . \quad ||R||_* + \lambda ||S||_1 \qquad \text{s.t.} \quad ||N||_F \leq \delta, D = R + S + N \tag{4.4}$$

where δ is a tunable parameter that models the aggregate magnitude of the estimated noise N. Such a relaxed condition secures a better stability to achieve the desired decomposition from noisy data.

In the special case where quantisation errors (denoted E) during image acquisition are primarily concerned, the following modified formulation (called "Quantised PCP" [14]) is often used instead:

$$\min_{R,S} . \quad ||R||_* + \lambda ||S||_1 \qquad \text{s.t.} \quad ||E||_\infty \leq \epsilon, D = R + S + E \tag{4.5}$$

where $||M||_\infty = max\{|M_{1,1}|, |M_{1,2}|, .., |M_{m,n}|\}$ is the infinity norm or maximum norm, which is the largest absolute value of all matrix entries. In the case of RGB images with three 8-bit channels, quantisation can induce an error of at most $\epsilon = 0.5$ at pixel/voxel level. In fact, in terms of algorithmic solutions (to be discussed in Section 4.2.2), the Stable PCP and Quantised PCP are almost identical to the original PCP, with major differences being the corresponding termination conditions.

In addition, when the errors are sparse while corrupting the data in a column-wise (or row-wise) pattern, a tailored version of PCP (called "block-sparse PCP" [145] by the authors, but should not be confused with some other block-sparse RPCA approaches [54]) can be used to improve decomposition performance by formulating:

$$\min_{R,S} . \quad ||R||_* + \kappa(1-\lambda)||R||_{2,1} + \kappa\lambda||S||_{2,1} \qquad \text{s.t.} \quad D = R + S \tag{4.6}$$

where $||M||_{2,1} = \sum_i ||M_i||_2$ is the L2,1 norm, which calculates the L1 norm of the vector

resulting from column-wise taking L2 norms of the matrix M.

4.2.2 Algorithms to solve PCP optimisation problems

A number of algorithms have been developed to solve the PCP optimisation problems. An early and straightforward solution is to employ the standard interior point methods (such as CVX [57]). However, off-the-shelf interior point solvers are generally limited to $O(m^6)$ complexity, and thus are difficult to apply when the problem scales up to, for example $m > 10^2$. To address this issue, a number of more advanced algorithms have been proposed, most notably the accelerated proximal gradient (APG) and augmented Lagrangian multiplier (ALM) algorithms, each will be discussed in the following.

In an abstract sense, all these algorithms are based on two sub-algorithms: the Bregman iterative algorithm for L1-minimisation [62, 170] and the singular value thresholding algorithm for nuclear-norm minimisation [27], which are used to optimise S and R estimations, respectively. In particular, an efficient approach is employed based on the iterative shrinkage-thresholding scheme [13], with the use of a special shrinkage (soft-thresholding) operator [148]:

$$\Phi_\epsilon[x] = \begin{cases} x - \epsilon, & \text{if } x > \epsilon \\ x + \epsilon, & \text{if } x < -\epsilon \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

where x and ϵ are real-valued scalars and ϵ must be positive. At the matrix/vector level, the operator is simply extended to:

$$\Phi_\epsilon[X] = sgn(X) \circ max(|X| - \epsilon, \mathbf{0}) \tag{4.8}$$

which performs element-wise absolute value shrinkage on all entries of the target matrix/vector X towards the zero matrix/vector $\mathbf{0}$.

Accelerated proximal gradient algorithm

The APG algorithm was originally proposed for matrix completion [150], and a similar algorithm was developed later to tackle the PCP problem [163, 104]. In this algorithm, Eq. (4.2) is first of all relaxed to the following form:

$$\min_{R,S}. \quad ||R||_* + \lambda||S||_1 + \frac{1}{2\mu}||D - R - S||_F^2 \tag{4.9}$$

where μ is a positive scalar, and as $\mu \to 0$, Eq. (4.9) approximates the original objective function. Supposing $X = (R, S)$, $g(X) = ||R||_* + \lambda||S||_1$, and $f(X) = \frac{1}{2\mu}||D - R - S||_F^2$, Eq. (4.9) is then converted to:

$$\min_{X}. \quad F(X) = f(X) + g(X) \tag{4.10}$$

Instead of directly minimising $F(X)$, the APG algorithm proceeds to iteratively minimise a sequence of separable quadratic approximations to $F(X)$, denoted $Q(X, Z)$, formulated using a set of purposefully chosen milestones Z:

$$\min_{X}. \quad Q(X, Z) = f(Z) + \langle \Delta f(X), X - Z \rangle + \frac{L_f}{2}||X - Z||_F^2 + g(X) \tag{4.11}$$

where L_f represents the Lipschitz constant and in this case $L_f = 2$. Algorithm 1 is a piece of pseudocode that outlines the computational procedure.

The convergence behaviour of this algorithm depends heavily on the selection of Z. A straightforward choice would be setting $Z_i = X_i$, which would lead to a convergence rate of $O(i^{-1})$ [104, 13]. Alternatively, it was discovered that by setting $Z_i = X_i + \frac{t_{i-1}-1}{t_i}(X_i - X_{i-1})$ at each iteration i, with a sequence $\{t_i\}$ satisfying the condition $t_i^2 - t_i \leq t_{i-1}^2$, the optimisation is able to achieve a convergence rate of $O(i^{-2})$ [118]. Furthermore, to improve practical performance, a gradual decay scheme (or continuation scheme) is often applied to μ, which starts with a large initial value μ_0 followed by geometric decrease, until reaching a pre-set minimum μ_{min}. Nevertheless, it is generally difficult to derive a generic continuation setting that guarantees both good accuracy and convergence rate across a wide range of problem settings. For that

Input: $D \in \mathbb{R}^{m \times n}$, λ, μ

$R_0 = R_1 = \mathbf{0}, S_0 = S_1 = \mathbf{0}, t_0 = t_1 = 1, \mu_{min} > 0, \eta < 1, L_f = 2;$

while *not converged* **do**

  $Z_{i,R} = R_i + \frac{t_{i-1}-1}{t_i}(R_i - R_{i-1})$, $Z_{i,S} = S_i + \frac{t_{i-1}-1}{t_i}(S_i - S_{i-1})$;

  $G_{i,R} = Z_{i,R} - \frac{1}{L_f}(Z_{i,R} + Z_{i,S} - D)$, $G_{i,S} = Z_{i,S} - \frac{1}{L_f}(Z_{i,R} + Z_{i,S} - D)$;

  $(U, \Sigma, V) = svd(G_{i,R})$;

  $R_{i+1} = U\Phi_{\frac{\mu_i}{L_f}}[\Sigma]V^*$;

  $S_{i+1} = \Phi_{\frac{\lambda\mu_i}{L_f}}[G_{i,S}]$;

  $t_{i+1} = \frac{1+\sqrt{4t_i^2+1}}{2}$;

  $\mu_{k+1} = max(\eta\mu_i, \mu_{min})$;

  $k = k + 1$;

end

Output: $R^*, S^* \in \mathbb{R}^{m \times n}$

Algorithm 1: RPCA by APG algorithm

reason, it often requires a series of empirical tests before reaching a good solution. In addition, a pre-requisite of the APG approach, is that $f(X)$ must be convex and smooth, and follows the Lipschitz continuity $||\Delta f(X_1) - \Delta f(X_2)||_F \leq L_f||X_1 - X_2||_F$ [104]. Fortunately, this condition is generally satisfied for RPCA problems on image analytics.

Augmented Lagrangian multiplier algorithms

The ALM is one of the most widely used methods to solve optimisation problems. A general description of the method with practical examples can be easily found in the literature [17]. Two tailored solutions to the PCP problem have been proposed, namely the exact ALM (EALM) algorithm and inexact ALM (IALM) algorithm [103], the latter of which is similar to the alternating direction method developed separately at the same time [171]. Both EALM and IALM algorithms start with a standard conversion of Eq. (4.2) to the canonical augmented Lagrangian function form:

$$\min. \quad L \doteq ||R||_* + \lambda||S||_1 + \langle Y, D - R - S\rangle + \frac{\mu}{2}||D - R - S||_F^2 \tag{4.12}$$

In this case, R, S and Y are then updated iteratively: in each iteration, R and S are updated first by minimising L with respect to R and S while keeping Y fixed, and then the discrepancy

Input: $D \in \mathbb{R}^{m \times n}$, λ
$R_{0,0} = \mathbf{0}$, $S_{0,0} = \mathbf{0}$;
$Y_0 = D \,/\, max\,(||D||, \lambda^{-1}||D||_\infty)$;
$i = 0$, $\mu_0 > 0$, $\rho > 1$;
while *not converged* **do**
  $j = 0$;
  while *not converged* **do**
    $(U, \Sigma, V) = svd(D - S_{i,j} + \mu_i^{-1} Y_i)$;
    $R_{i,j+1} = U \Phi_{\mu_i^{-1}}[\Sigma] V^*$;
    $S_{i,j+1} = \Phi_{\lambda \mu_i^{-1}}[D - R_{i,j+1} + \mu_i^{-1} Y_i]$;
    $j = j + 1$;
  end
  $Y_{i+1} = Y_i + \mu_i (D - R_{i,j} - S_{i,j})$;
  $R_{i+1,0} = R_{i,j}$, $S_{i+1,0} = S_{i,j}$;
  $\mu_{i+1} = \rho \mu_i$;
  $i = i + 1$;
end
Output: $R_*, S_* \in \mathbb{R}^{m \times n}$

Algorithm 2: RPCA by EALM algorithm

$(D - R - S)$ is used in turn to update Y.

EALM algorithm: Once the objective function is formulated, the EALM algorithm then carries out the computation described in Algorithm 2, where $||M|| = max\{\sigma_i\}$ denotes the spectral norm of the matrix M, which calculates its largest singular value. To secure a good convergence speed and improve practical performance, Y is often initialised to $Y_0 = D/max\,(||D||, \lambda^{-1}||D||_\infty)$, the merit of which is described in the dual problem [104]. Furthermore, based on relevant analysis [103], the EALM algorithm converges Q-linearly. In particular, the outer-loop converges faster as the geometric series μ_i grows faster, however the inner-loop converges slower when μ_i is larger, which makes the setting of μ and ρ critical to the final computational efficiency.

IALM algorithm: Further analysis [103] on the other hand, concluded that the inner-loop that tackles the sub-problem:

$$(R_{i*}, S_{i*}) = \underset{R_i, S_i}{\arg\min}\,. \quad L(R_i, S_i, Y_i, \mu_i) \qquad \text{while keeping } Y_i \text{ fixed} \tag{4.13}$$

does not need to be solved exactly in each outer-loop iteration, in order for the final R_*, S_* to

Input: $D \in \mathbb{R}^{m \times n}$, λ
$R_0 = \mathbf{0}$, $S_0 = \mathbf{0}$;
$Y_0 = D \,/\, max\,(||D||, \lambda^{-1}||D||_\infty)$;
$i = 0$, $\mu_0 > 0$, $\rho > 1$;
while *not converged* **do**
  $S_{i+1} = \Phi_{\lambda\mu_i^{-1}}[D - R_i + \mu_i^{-1}Y_i]$;
  $(U, \Sigma, V) = svd(D - S_{i+1} + \mu_i^{-1}Y_i)$;
  $R_{i+1} = U\Phi_{\mu_i^{-1}}[\Sigma]V^*$;
  $Y_{i+1} = Y_i + \mu_i(D - R_{i+1} - S_{i+1})$;
  $\mu_{i+1} = \rho\mu_i$;
  $i = i + 1$;
end
Output: $R_*, S_* \in \mathbb{R}^{m \times n}$

Algorithm 3: RPCA by IALM algorithm

converge to the optimal solution. For that reason, the IALM algorithm was proposed based on a much faster inexact solution, as outlined in Algorithm 3. In this algorithm, R_i and S_i are only updated once when tackling the sub-problem Eq. (4.13), and all (R_i, S_i, Y_i) are updated collectively in each single (overall) loop. Furthermore, it was shown that the algorithm still converges Q-linearly with a geometrically growing μ_i series, yet if the growth rate ρ was set too high, the algorithm may no longer converge to the optimum [103]. The recommended setting was somewhere near $\rho = 1.5$ based on the authors' empirical experience. In a set of simulation experiments [103], the IALM algorithm converged at least five times faster than the APG algorithm. For that reason, the IALM method has become one of the most widely used solution to the PCP optimisation problem nowadays.

4.3 RPCA with Variation Priors

4.3.1 Challenges of RPCA in biomedical imaging

Based on the description in Section 4.2.1, it is not difficult to identify that the condition of stationery background in previous RPCA applications simply corresponds to the idealised RPCA assumption, in which R_0 lies in a low-dimensional subspace while S_0 being sparse and affecting only a small number of entries in D. However, such idealised assumption is often

unrealistic when applied to the biomedical domain.

In biomedical imaging, there is usually a structural congruity of biological anatomy across different subjects, which in many circumstances implies the existence of a regular structure in imaging data, or in RPCA terms, a low-dimensional subspace for high-level feature representation. Nevertheless, in contrast to those computer vision problems, there is often a significant natural variation within biomedical data, leading to a number of fundamental challenges regarding the applicability of the RPCA technique, most notably:

(a) When acquiring multiple images of the same subject, the imaging subject may be positioned differently with slightly different postures or orientations in different acquisition procedures.

(b) A subject may undertake physiological movements when imaging is performed at different time points. For instance, this problem is particularly significant when applied to in-vivo cardiac imaging, due to frequent heartbeats.

(c) When imaging is performed on multiple subjects, there is usually a significant inter-subject variability due to the idiosyncratic nature of biological structures.

(d) Different anatomical structures are generally associated with individualistic natural variability rather than a common variation property. For example, the intestine generally has higher natural variation than the bones.

(e) There is a high prevalence of noise and artefacts in biomedical images acquired using contemporary imaging facilities.

(f) For the MRI modality, different images often have different scales of intensity values

To cope with the challenges above, a number of methods have been employed in the framework described in Chapter 3. To start with, Gaussian smoothing was applied to denoise images, in order to mitigate problem (e). Secondly, group-wise linear (rigid and affine) registration was employed to deal with problem (a), where images with different postures and orientations

can be effectively aligned. This is also a common practice in computer vision applications such as face recovery. Furthermore, challenge (b) can be addressed by the application of non-rigid registration on intra-subject images, whilst (c) can be tackled (although not eliminated entirely) by group-wise non-rigid alignment in a locally-created representative template space. In addition, in the case of MRI data, some tissue-standardising normalisation techniques [121] are often utilised to overcome problem (f).

However, challenge (d) was not directly addressed using the methodology described in Chapter 3. Instead, an ROI mask was applied to restrict RPCA decomposition only to the target structures, in order to avoid dealing with multiple structures at the same time. More importantly, this points out the limitation of the baseline RPCA technique for biomedical imaging, in the sense that there is only a single criterion of outlier tolerance in feature decomposition, solely controlled by the parameter λ, and applied globally regardless of local variation properties.

4.3.2 The RPCA-P framework

In contrast to the baseline RPCA technique that implements a purely unsupervised data-driven approach without leveraging any prior knowledge, we propose a modified framework, the RPCA-P, which is able to incorporate variation priors in feature decomposition. In the proposed framework, outlier tolerance can be adjusted locally so that voxels associated with structures/regions that experience higher variability in nature are compensated by allowing a higher tolerance during feature decomposition. The priors are learned from the variation patterns in the data itself. In this case, the RPCA-P framework rests on a more relaxed assumption, in which there could be differential levels of variation across different regions in imaging data. On the other hand, in terms of anomaly detection, the anomalous distortion should exceed the normal range of variation to distinguish themselves from normality.

Suppose there are n images in total, all of which are properly denoised and group-wise aligned non-rigidly in a common template space (denoted as $I_1, .., I_n \in \mathbb{R}^{w \times h \times d}$). Concatenating each image into a vector of size $m = w \times h \times d$ and stacking them together will create a matrix

$D \in \mathbb{R}^{m \times n}$, then we have:

$$D = R + S \quad \text{where} \quad D, R, S \in \mathbb{R}^{m \times n} \tag{4.14}$$

Similar to the previous setting, the R here estimates the intrinsic regular structure of D that lies in a low-dimensional subspace, and S captures the abnormal deformations that distort subjects from their regular form. However R may be subject to moderate normal variation across all regions in every data sample, whereas the abnormal variation in S is assumed to be sparse, only affecting a small percentage of entries in D. Then in contrast to the standard PCP formulation, we instead formulate:

$$\min_{R,S} . \quad ||R||_* + \lambda ||\xi^{(n)} \circ S||_1 \qquad \text{s.t.} \quad D = R + S \tag{4.15}$$

where $\xi^{(n)} \in \mathbb{R}^{m \times n}$ is the n-time column-wise replication of $\xi \in \mathbb{R}^{m \times 1}$, which is the vectorised form of the variation prior map $I_\xi \in \mathbb{R}^{w \times h \times d}$ that voxel-wise adjusts outlier tolerance. The prior I_ξ is learned from data, and the learning methods will be discussed in Section 4.3.3.

In order to solve the above optimisation problem, we developed an algorithm based on the state-of-the-art IALM algorithm described in Section 4.2.2. To start with, Eq. (4.15) is converted to the canonical augmented Lagrangian function form:

$$L \doteq ||R||_* + \lambda ||\xi^{(n)} \circ S||_1 + \langle Y, D - R - S \rangle + \frac{\mu}{2} ||D - R - S||_F \tag{4.16}$$

We then solve it using Algorithm 4. In particular, the shrinkage operator is extended to include a two-input version:

$$\Phi_{\tau, \mathbf{w}}[X] = sgn(X) \circ max(|X| - \tau \cdot \mathbf{w}^{(n)}, \mathbf{0}) \tag{4.17}$$

where X represents the target matrix to perform the shrinkage operation, $\tau \cdot \mathbf{w}^{(n)}$ models the locally adjusted shrinkage level and is based on the combination of a baseline shrinkage τ and a weight vector $\mathbf{w} \in \mathbb{R}^{m \times 1}$, in which shrinkage is performed on each image independently.

Input: $D \in \mathbb{R}^{m \times n}$, $\xi \in \mathbb{R}^{1 \times m}$, λ

$Y_0 = D \,/\, max\left(||D||, \lambda^{-1}||D||_\infty\right)$;

$R_0 = \mathbf{0}$, $S_0 = \mathbf{0}$, $i = 0$, $\mu_0 > 0$, $\rho > 1$;

while *not converged* **do**

  $S_{i+1} = \Phi_{\lambda\mu_i^{-1}, \xi^*}[D - R_i + \mu_i^{-1} Y_i]$;

  $(U, \Sigma, V) = svd(D - S_{i+1} + \mu_i^{-1} Y_i)$;

  $R_{i+1} = U \Phi_{\mu_i^{-1}}[\Sigma] V^*$;

  $Y_{i+1} = Y_i + \mu_i (D - R_{i+1} - S_{i+1})$;

  $\mu_{i+1} = \rho \mu_i$;

  $i = i + 1$;

end

Output: $R^*, S^* \in \mathbb{R}^{m \times n}$

Algorithm 4: RPCA-P by modified ALM Method

4.3.3 RPCA-P application to mouse embryo phenotyping

In order to test our hypothesis regarding the effectiveness of RPCA-P, we apply it to the mouse embryo phenotyping problem for direct comparison with the baseline PCA and RPCA methods. We employ a similar methodology to that detailed in Chapter 3 Section 3.5:

Data pre-processing: image denoising and mouse embryo extraction

The computation in this part is identical to the original methodology in Chapter 3, with image denoising described in Section 3.5.1 and mouse embryo extraction in Section 3.5.2. The effectiveness of such pre-processing has been evaluated in Section 3.6.

Group-wise image alignment

Non-rigid image registration is employed to collectively align all target images before RPCA-P processing, in order to ensure effective subspace estimation. For simplicity, we directly use the same template created earlier in Chapter 3 Section 3.5.3 as the reference for group-wise registration. The non-rigid registration scheme again contains three steps: a rigid, an affine and then a B-spline registration. Based on our previous study, we use the best parameter setting derived in Section 3.6.4. In particular, in terms of B-spline registration we apply a four-scale multi-resolution configuration with maximally 4000 iterations at each resolution level.

The objective function is based on the combination of mutual information and bending energy penalty, and the spacing of control points is halved at every new resolution level, with the final spacing set to 200 μm.

Estimation of natural variation priors

In order to perform RPCA-P processing, we need to estimate the variation prior ξ first. As described earlier, $\xi = vec(I_\xi)$ where $I_\xi \in \mathbb{R}^{w \times h \times d}$ is a weight map that captures the local variability at voxel-level. A notable point is that I_ξ can be learned from (1) the whole dataset, or (2) a smaller control group. The former case requires the abnormal deformation features to be sparse across the whole dataset, so that they will carry a low weight in the learned I_ξ. The latter case relaxes this condition and is able to completely exclude the influence of abnormal deformation in I_ξ estimation, however it requires some subjects known to be normal in advance. Since the purpose of the phenotyping study is to detect morphological abnormalities in the whole test dataset without knowing label information a priori, and defective phenotypes are assumed to be sparse with individualistic deformation features, we take the former approach.

Depending on the actual application, there could be many possible learning models. Three typical models are used in this study, respectively, based on:

- **Voxel-wise intensity variance:**

$$I'_\xi = \frac{1}{n} \cdot \sum_{i \in 1..n} \Big((I_i - I_\mu) \circ (I_i - I_\mu) \Big) \tag{4.18}$$

  where $I_\mu = \frac{1}{n} \sum_{i \in 1..n} I_i$ is the average image of the dataset, and $A \circ B$ indicates entry-wise multiplication of matrices A and B.

- **Voxel-wise standard deviation:**

$$I'_\xi = \sqrt{\frac{1}{n} \cdot \sum_{i \in 1..n} (I_i - I_\mu) \circ (I_i - I_\mu)} \tag{4.19}$$

where $\sqrt{M}$ is an entry-wise operator that computes the square root of every entry in the matrix M. By definition, this is simply the voxel-wise square root of the variance model.

- **Average pair-wise discrepancy:**

$$I'_\xi = \Big(\frac{n!}{2!(n-2)!}\Big)^{-1} \cdot \sum_{i,j \in 1..n} |I_i - I_j| \tag{4.20}$$

where $|M|$ is an entry-wise operator that computes the absolute value of every entry in the matrix M. This model essentially sums up the pair-wise absolute differences between every two images and then takes the average.

To normalise the weight scale across different prior models, the intensities in the generated I'_ξ are rescaled with reference to their mean intensity values, and a standard Gaussian smoothing is then applied before vectorisation to finalise ξ estimation:

$$I_\xi = G\Big(\frac{I'_\xi}{\mu(I'_\xi)}\Big) \qquad\qquad \xi = vec(I_\xi) \tag{4.21}$$

The effectiveness of each model will be empirically tested on the mouse embryo μ-CT data regarding anomaly detection in Section 4.4.

RPCA-P feature decomposition

With the variation prior model ξ estimated, the optimisation problem Eq. (4.15) is then formulated and solved using Algorithm 4. Since ξ is able to voxel-wise adjust outlier tolerance across the whole mouse embryo rather than restricted to a small anatomical structure or region, an ROI mask in this case is no longer needed. Instead we apply it to the entire embryo (excluding the background to reduce computational burden in the actual implementation).

In terms of the convergence condition, we use the same criterion recommended in the original work of the IALM algorithm [103] and set it to:

$$\pi = ||D - R^* - S^*||_F \,/\, ||D||_F < 10^{-7} \tag{4.22}$$

The only important parameter to tune then is the baseline tolerance λ, which will be detailed in Section 4.4. Once the algorithm terminates, the obtained R^* and S^* are then used to reconstruct $I^*_{i,r}$ and $I^*_{i,s}$, $\forall i \in 1..n$.

Abnormality detection

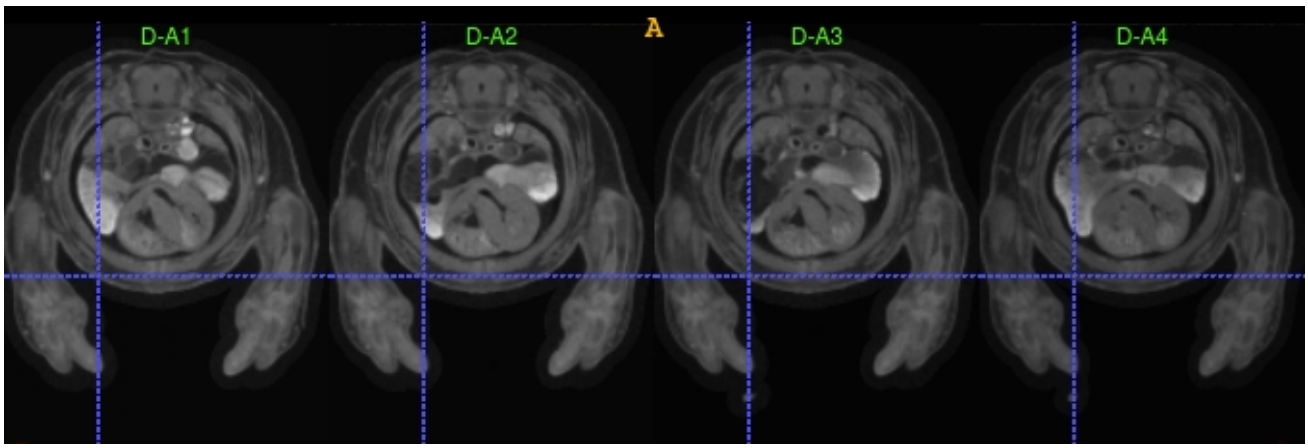
For easy comparison, abnormality detection is carried out using the same method as in the original methodology using the purpose-developed anomaly rate metric (denoted Ω), described in Chapter 3 Section 3.5.6. A notable point is that the 'Normal-vs-Abnormal" classification is applied separately from the RPCA-P process, and can be applied to the entire image or a specific region, regardless of whether feature decomposition is applied to the image or region level. In brief, this method calculates the level of morphological abnormality identified in each image or region, and compares it with a baseline level, in order to determine whether the corresponding image/region should be annotated normal or abnormal.

4.4 Evaluation

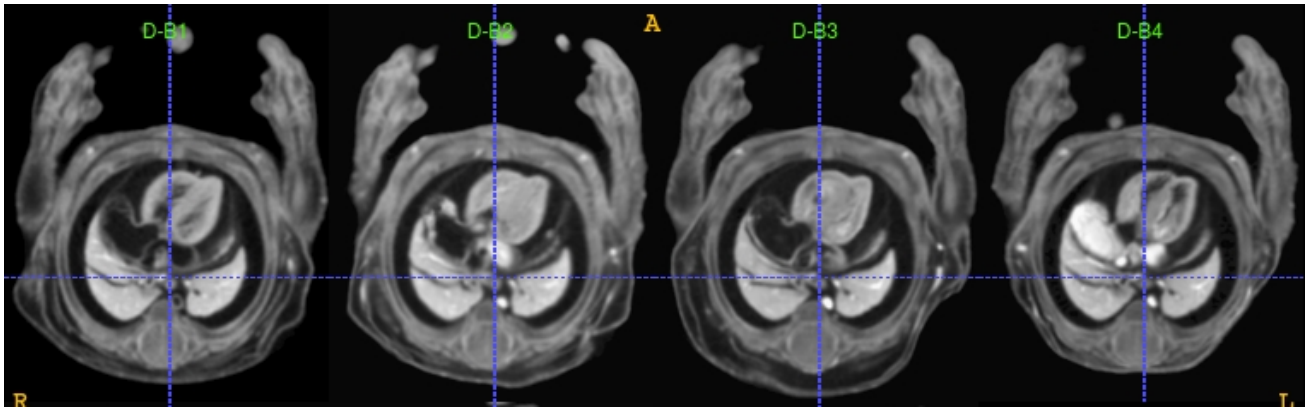
4.4.1 Test data

An evaluation study has been carried out on the same two datasets of mouse embryo μ-CT images, described in Chapter 3 Section 3.4.2. Figure 4.1 illustrates some sample images of Datasets A and B after pre-processing and group-wise alignment. In particular, since the images in Dataset A were too large to scale RPCA-P computation to the entire embryo region in our experimental environment (with 32GB memory only), all of them were down-sampled to the resolution $20 \times 20 \times 20\ \mu m^3$, and the reduced image size was $326 \times 392 \times 612$.

Furthermore, it is easy to identify that the images in each dataset are structurally similar, but there is also a notable natural variation across different subjects, especially in regions like the heart and lung. A series of experiments have been carried out on both test datasets to demonstrate the effectiveness of the RPCA-P in comparison with the baseline method. The



(a) Dataset A



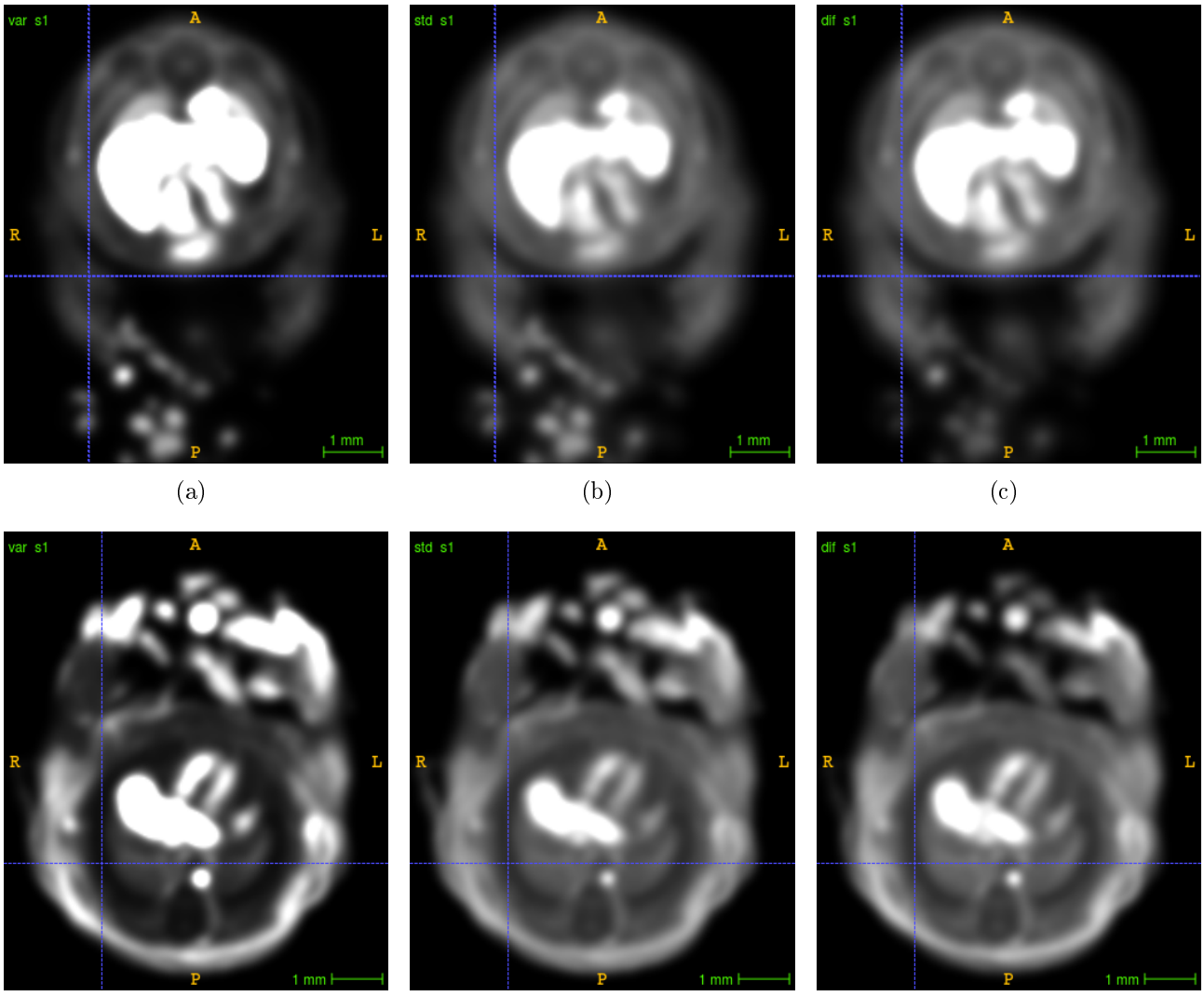
(b) Dataset B

Figure 4.1: Sample images in axial view from (a) Dataset A and (b) Dataset B. In contrast to Figures 3.8 and 3.9 in Chapter 3, all images displayed here are properly pre-processed and group-wise non-rigid aligned in the template space, ready for RPCA-P processing directly. In particular, images in Dataset A are down-sampled to the resolution $20 \times 20 \times 20\ \mu m^3$

results of variation prior estimation, and the combinatorial influences of baseline tolerance and local tolerance adjustment on feature decomposition and abnormality detection will be discussed in the following sections.

4.4.2 Results of variation prior ξ estimation

The variation prior ξ estimates the degree of natural variability at voxel-level, and is used to adjust the local outlier tolerance during RPCA-P feature decomposition. In anatomical structures or regions that naturally experience higher variability (such as the lung and the intestine), features more distant from "normality" will be tolerated and included in the regular



(a) (b) (c)

(d) (e) (f)

Figure 4.2: The weight maps generated by variation prior estimation using the (a) ξ_{var} model, (b) ξ_{std} model and (c) ξ_{apd} model for Dataset A, as well as the counterparts using the (d) ξ_{var} model, (e) ξ_{std} model and (f) ξ_{apd} model for Dataset B

component, rather than being considered anomalous.

As described earlier, in this study there were three models used for learning local natural variability, respectively, based on (1) voxel-wise intensity variance ξ_{var}, (2) voxel-wise standard deviation ξ_{std}, and (3) average pair-wise discrepancy ξ_{apd}. The local tolerance maps generated for Dataset A and Dataset B using these three learning models are illustrated in Figure 4.2, where the brighter part indicates a higher local variability and thus higher outlier tolerance, and vice versa.

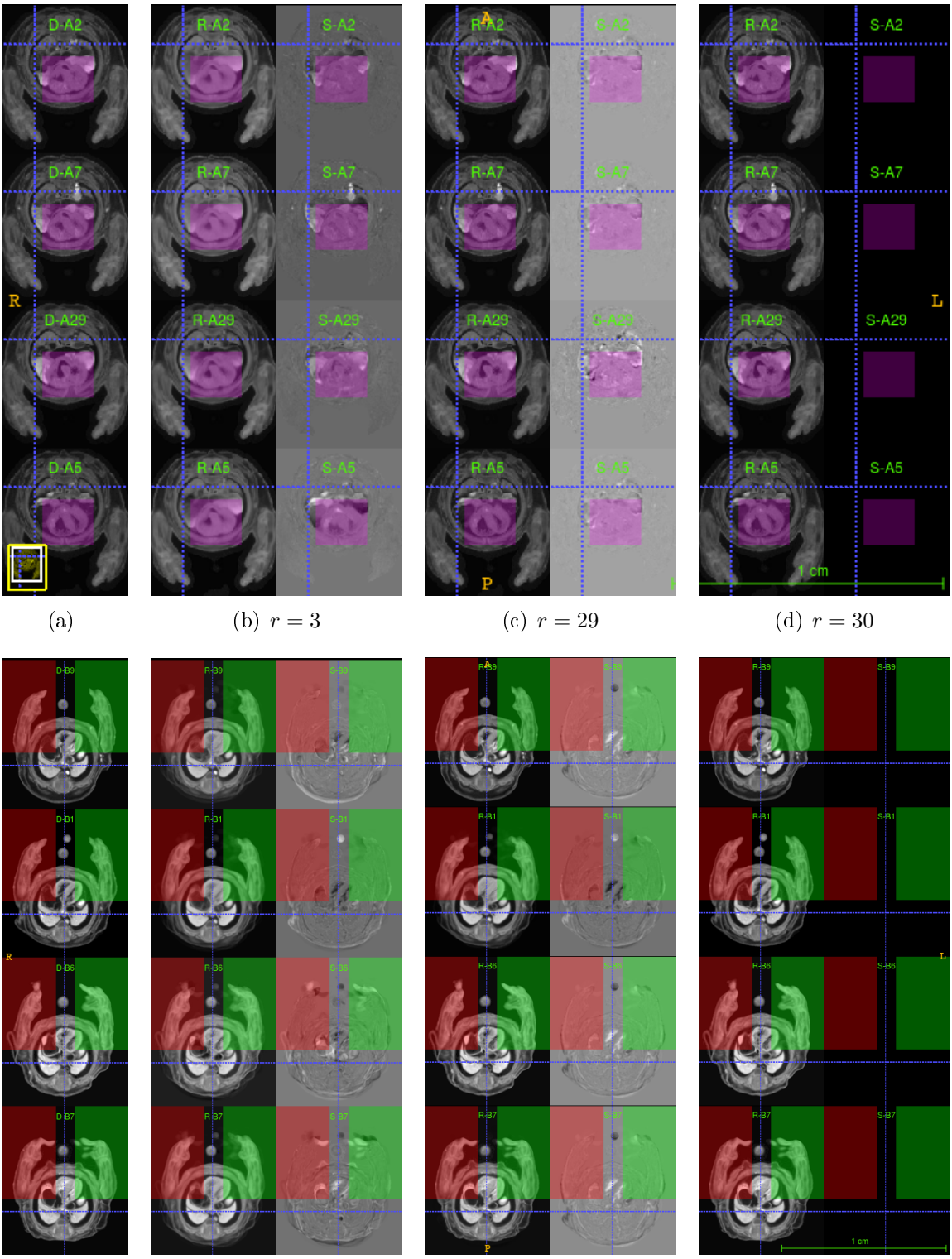
4.4.3 Results of feature decomposition: the influence of the baseline tolerance λ

The parameter λ governs the global baseline level of outlier tolerance for regular/singular decomposition. Similar to that in the original methodology, the reconstructed I_r images are more similar to each other when λ is smaller, and as λ increases, more individualistic features are included in I_r, leaving only features distant from "normality" retained in I_s.

In contrast to the simple guideline described in Chapter 3 Section 3.6.3, the setting of λ in the RPCA-P framework needs to be determined in combination with the use of variation prior ξ, which may seem more complicated at the first glance. However, we found that once the weight normalisation in Eq. (4.21) is applied, a working setting would still be somewhere near the order of $\lambda = 1/\sqrt{m}$; or $\lambda = 1/(\mu_\xi\sqrt{m})$ if not normalised. The decomposition results with the settings $\lambda = 0.5/\sqrt{m}$, $\lambda = 1/\sqrt{m}$ and $\lambda = 1.5/\sqrt{m}$ are compared across four different samples from Dataset A and four from Dataset B in Figure 4.5, Figure 4.6, and Figure 4.7.

4.4.4 Results of feature decomposition: the influence of the variation prior ξ

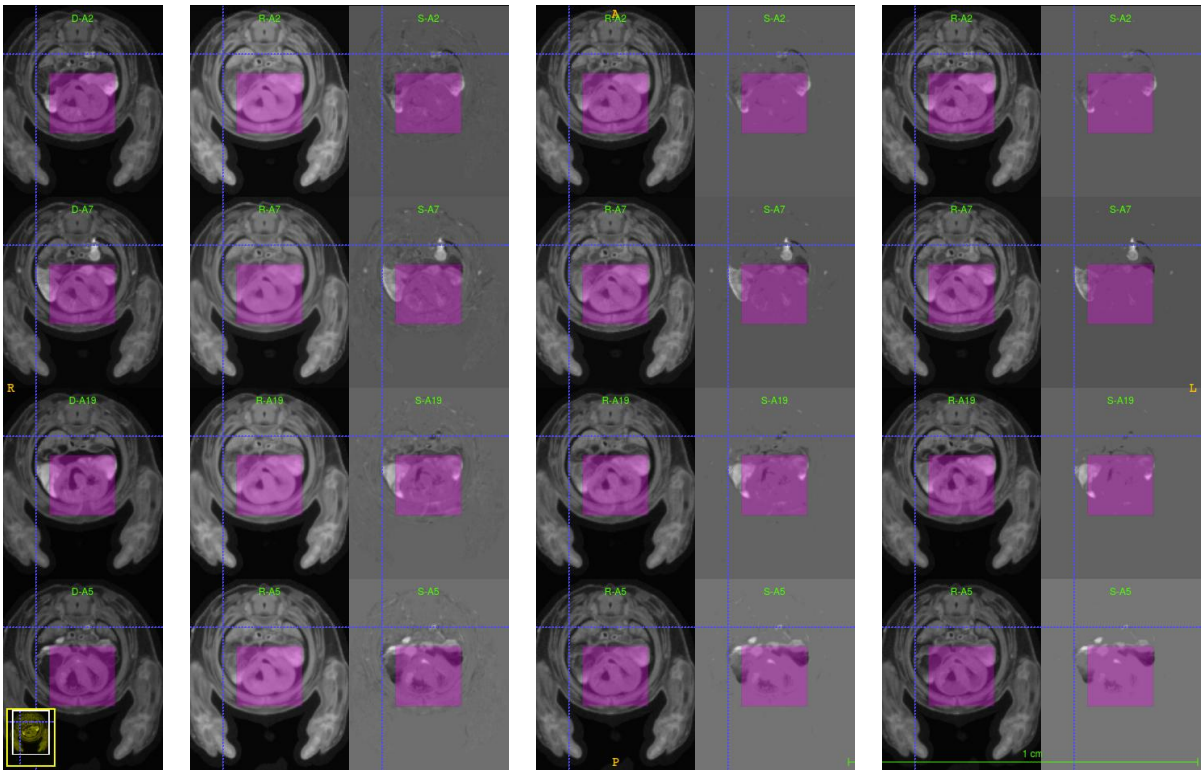
In general, the incorporation of variation priors significantly improves the performance of feature decomposition compared to the baseline method used in our original methodology. To better demonstrate the superiority of RPCA-P, an example outcome of feature decomposition applied to the same subjects in Dataset A and Dataset B, using the baseline PCA method and baseline RPCA method, as well as RPCA-P with ξ_{var}, ξ_{std}, and ξ_{apd} priors are, respectively, illustrated in Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, and Figure 4.7 in a comparative fashion. The region showing the heart in images from Dataset A, as well as the regions containing the left and right limbs in images from Dataset B are, respectively, coloured in purple, green and red, in order to help with visual inspection. In contrast to the original methodology however, as described earlier, the RPCA-P process did not use ROI masks for feature decomposition, and was instead applied to the entire embryo.



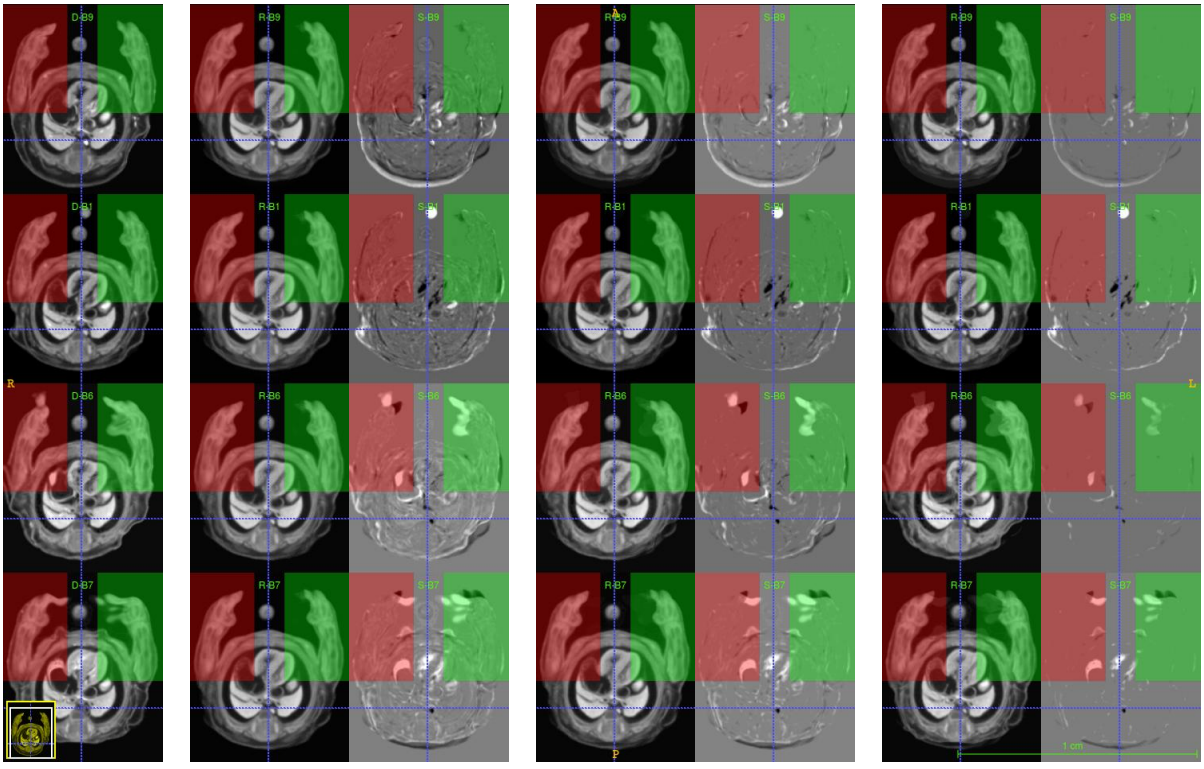
(a) (b) $r = 3$ (c) $r = 29$ (d) $r = 30$

(e) (f) $r = 3$ (g) $r = 13$ (h) $r = 14$

Figure 4.3: Sample results of feature decomposition using baseline PCA, applied to (a-d) Dataset A with $r = 3$, $r = 29$, and $r = 30$, and (e-h) Dataset B with $r = 3$, $r = 13$, and $r = 14$, respectively. Rows 1-2 in both cases are typical of normal controls, Row 3 is a subject of VSD/polydactyly phenotype, and Row 4 is a normal subject with particularly significant natural variation.

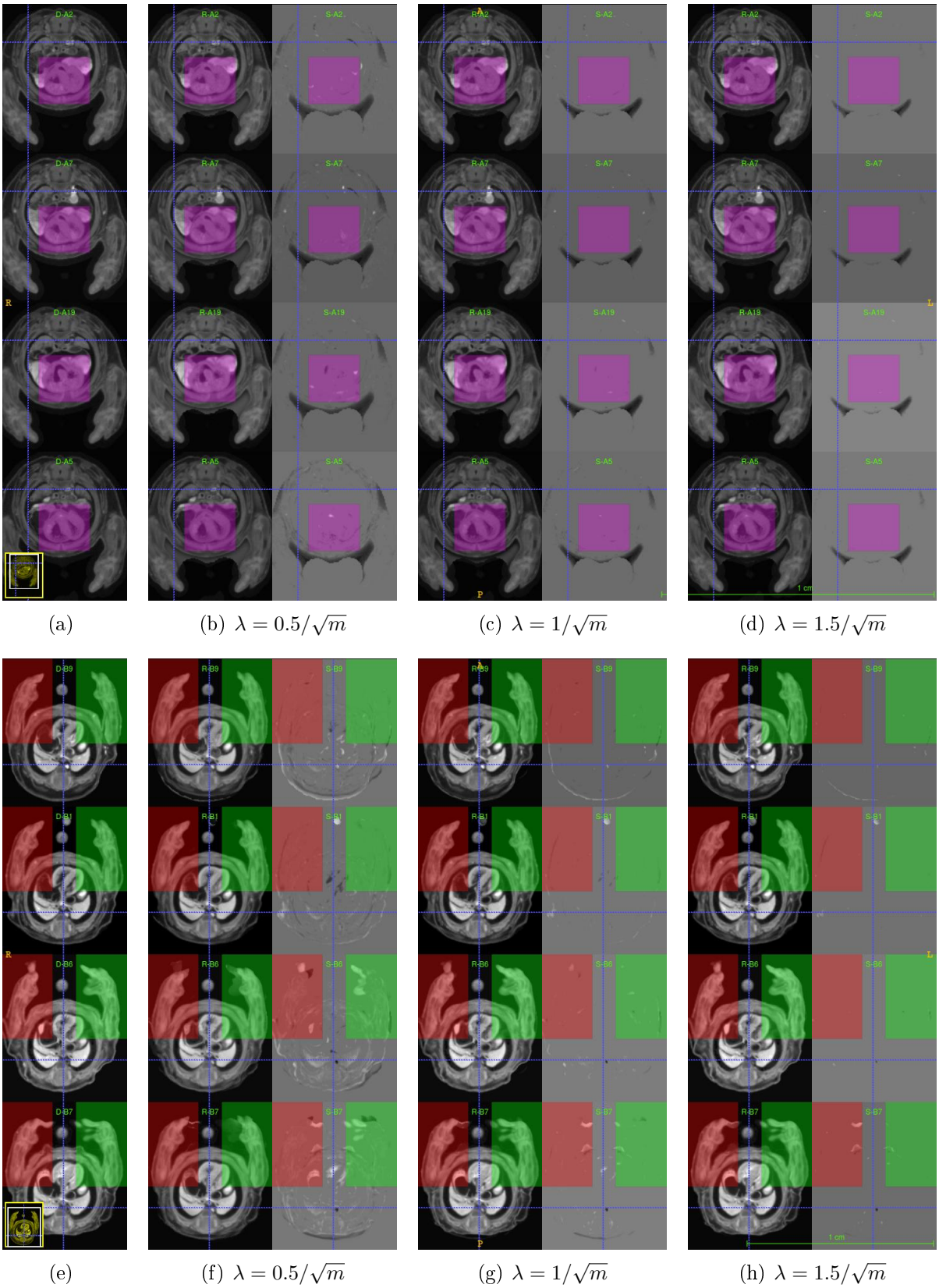


(a) (b) $\lambda = 0.5/\sqrt{m}$ (c) $\lambda = 1/\sqrt{m}$ (d) $\lambda = 1.5/\sqrt{m}$



(e) (f) $\lambda = 0.5/\sqrt{m}$ (g) $\lambda = 1/\sqrt{m}$ (h) $\lambda = 1.5/\sqrt{m}$

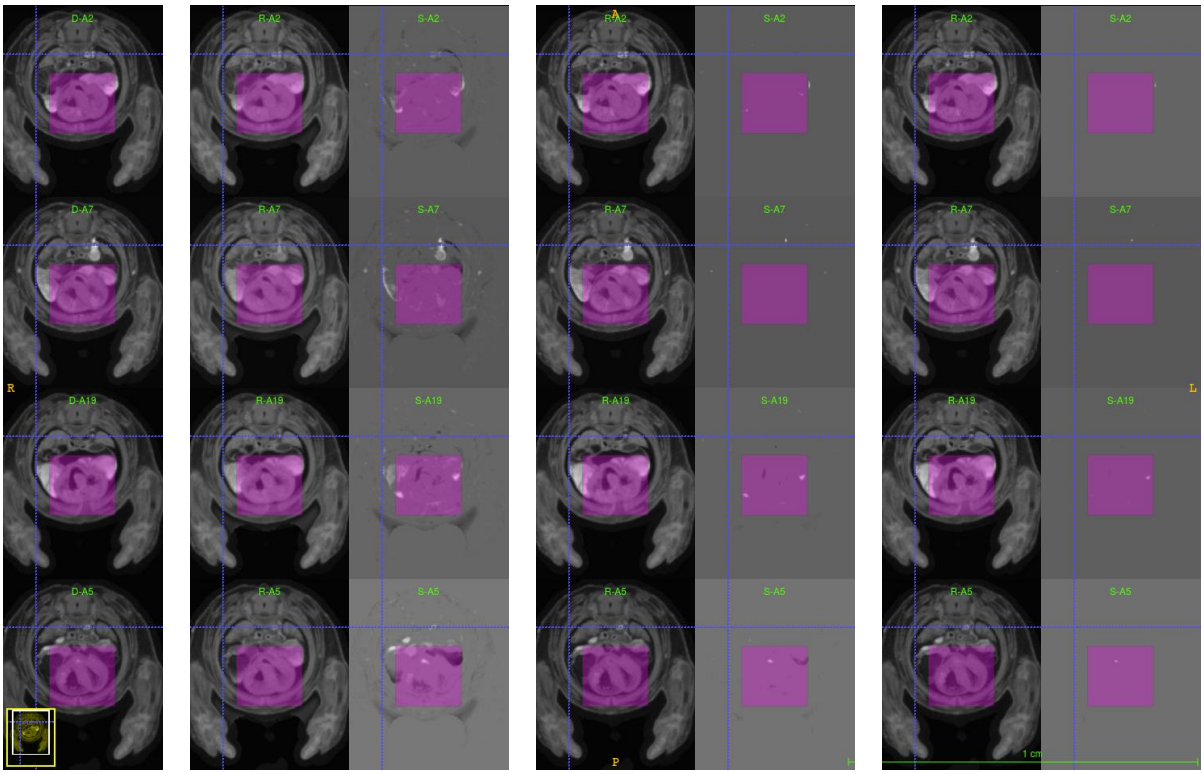
Figure 4.4: Sample results of feature decomposition using baseline RPCA, applied to the same subjects as above, with $\lambda = 0.5/\sqrt{m}$, $1/\sqrt{m}$, and $1.5/\sqrt{m}$, respectively.



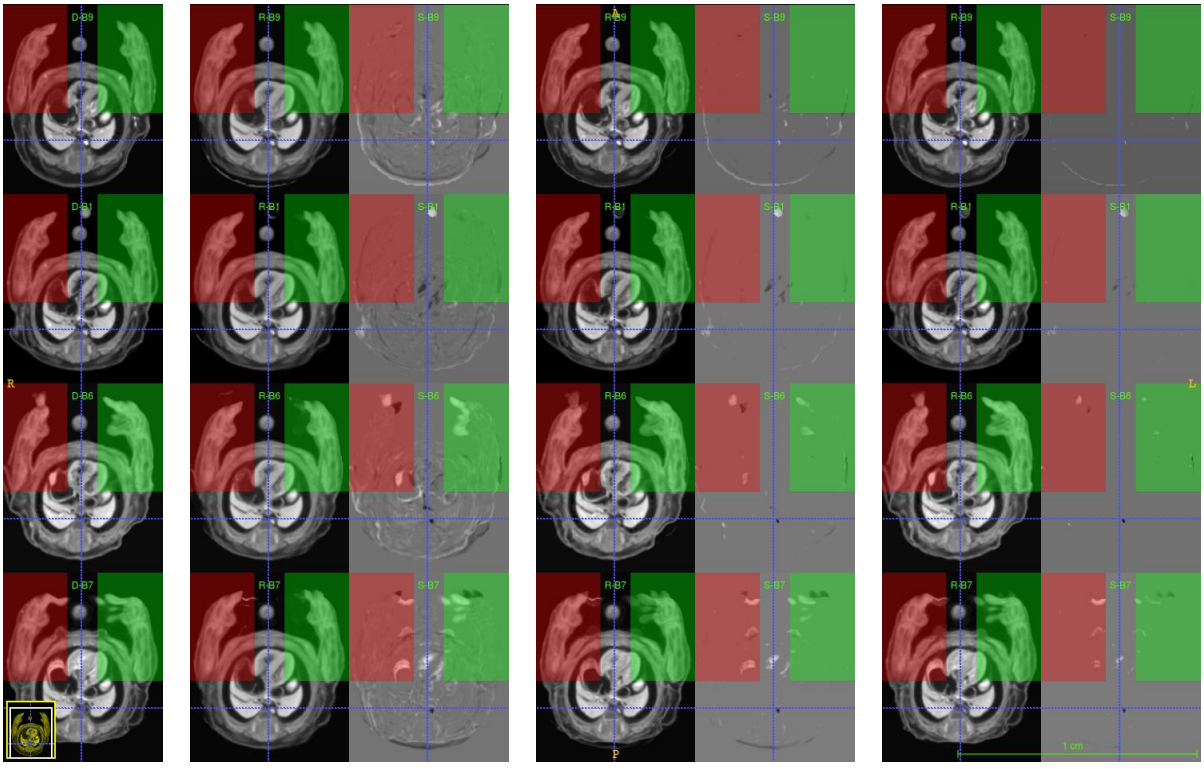
(a) (b) $\lambda = 0.5/\sqrt{m}$ (c) $\lambda = 1/\sqrt{m}$ (d) $\lambda = 1.5/\sqrt{m}$

(e) (f) $\lambda = 0.5/\sqrt{m}$ (g) $\lambda = 1/\sqrt{m}$ (h) $\lambda = 1.5/\sqrt{m}$

Figure 4.5: Sample results of feature decomposition using the RPCA-P method with ξ_{var}, applied to the same subjects as above



(a) (b) $\lambda = 0.5/\sqrt{m}$ (c) $\lambda = 1/\sqrt{m}$ (d) $\lambda = 1.5/\sqrt{m}$



(e) (f) $\lambda = 0.5/\sqrt{m}$ (g) $\lambda = 1/\sqrt{m}$ (h) $\lambda = 1.5/\sqrt{m}$

Figure 4.6: Sample results of feature decomposition using the RPCA-P method with ξ_{std}, applied to the same subjects as above

(a) (b) $\lambda = 0.5/\sqrt{m}$ (c) $\lambda = 1/\sqrt{m}$ (d) $\lambda = 1.5/\sqrt{m}$

(e) (f) $\lambda = 0.5/\sqrt{m}$ (g) $\lambda = 1/\sqrt{m}$ (h) $\lambda = 1.5/\sqrt{m}$

Figure 4.7: Sample results of feature decomposition using the RPCA-P method with ξ_{apd}, applied to the same subjects as above

By simple inspection, one can easily identify that feature decomposition by classic PCA was, again, ineffective and unable to separate outliers from the regularity. Decomposition by the baseline RPCA method, on the other hand, was still far from ideal, with a lot of normal features outside the heart region in the case of Dataset A, or the limb region in the case of Dataset B, being mistakenly included in the singular component. In stark contrast, the RPCA-P method substantially enhanced the decomposition performance, where the target abnormalities were generally captured with minimal noise present in the singular component. Moreover, feature decomposition using ξ_{std} and ξ_{apd} generally outperformed the counterpart using ξ_{var}, in terms of capturing the desired phenotypical features while tolerating considerable natural variations. The difference between ξ_{std} and ξ_{apd} on the other hand, was much less significant. Yet a closer observation will discover that there was a larger amount of phenotypical features remaining in the singular component with the use of ξ_{apd} prior model, especially when λ increased to $\lambda = 1.5/\sqrt{m}$. A notable point is that the RPCA-P application to Dataset B using ξ_{var} in particular, was not very successful, with the polydactyly features mostly being included in the regular component rather than separated out as singular features, leading to a lower anomaly detection accuracy, which will be discussed in the next section. The differential feature decomposition results indicate that the selection of variation prior is critical to RPCA-P performance.

4.4.5 Results of abnormality detection

Similar to our previous study, the performance of abnormality detection was evaluated using cross validation, in which each setting was applied 31 and 15 times, respectively, to Datasets A and B with a leave-one-out strategy. In each test, the baseline PCA, baseline RPCA and the proposed RPCA-P using ξ_{var}, ξ_{std}, ξ_{apd} priors were applied to the remaining 30 and 14 images. In terms of evaluating the detection performance, the same system based on the sensitivity, specificity and overall accuracy measures described in Chapter 3 Section 3.6.5 was applied.

To enforce a fair comparison, we apply the same set of ROI masks corresponding to, respectively, the heart for Dataset A and each of the four limbs in Dataset B, to compute the anomaly rates

Table 4.1: Performance metrics of abnormality detection on Dataset A

Dataset A			
	SEN	SPE	ACC
Baseline PCA	$100.00 \pm 0.00\%$	$17.65 \pm 2.20\%$	$33.33 \pm 2.42\%$
Baseline RPCA	$100.00 \pm 0.00\%$	$82.35 \pm 2.20\%$	$85.71 \pm 1.79\%$
RPCA-P with ξ_{var}	$100.00 \pm 0.00\%$	$88.24 \pm 1.86\%$	$90.48 \pm 1.50\%$
RPCA-P with ξ_{std}	$100.00 \pm 0.00\%$	$94.12 \pm 1.36\%$	$95.24 \pm 1.09\%$
RPCA-P with ξ_{apd}	$100.00 \pm 0.00\%$	$94.12 \pm 1.36\%$	$95.24 \pm 1.09\%$

for all cases. A notable point is that each ROI mask used here is a simple cuboid that contains a target organ as well as its surrounding tissues and structures (respectively depicted by the purple, green and red-coloured regions in the figures), in contrast to only covering the target organs as in our previous study in Chapter 3. This reduces the additional work to draw refined ROI masks as in the original methodology.

The metrics of phenotypical abnormality detection using the baseline PCA, baseline RPCA and RPCA-P with ξ_{var}, ξ_{std}, ξ_{apd} priors applied to Dataset A are shown in Table 4.1 in a comparative fashion. Unsurprisingly, the baseline PCA method obtained low detection scores, again, corroborating our suggestion of its infeasibility for abnormality detection. Then, in comparison to the baseline RPCA method, an improved detection performance regarding the VSD abnormality was observed in all three RPCA-P settings: where the statistics rose from $82.35 \pm 2.20\%$ specificity and $85.71 \pm 1.79\%$ overall accuracy obtained in the baseline case, to $88.24 \pm 1.86\%$ specificity and $90.48 \pm 1.50\%$ accuracy using RPCA-P with ξ_{var}, and further rose to $94.12 \pm 1.36\%$ specificity and $95.24 \pm 1.09\%$ accuracy with ξ_{std} or ξ_{apd}, while the sensitivity was maintained at 100% at all times.

In the case of Dataset B, the comparative performance metrics are shown in Table 4.2. Similarly, the baseline PCA was considered a failure and quickly ruled out of consideration, whereas the baseline RPCA approach obtained $77.42 \pm 2.93\%$ specificity and $84.44 \pm 2.06\%$ overall accuracy. On the other hand, by using RPCA-P with ξ_{std} the performance improved to $86.46 \pm 2.69\%$ specificity and $90.67 \pm 1.87\%$ overall accuracy, and by using RPCA-P with ξ_{apd} it further improved to $90.97 \pm 1.68\%$ specificity and $93.78 \pm 1.17\%$ overall accuracy. However, RPCA-P

Table 4.2: Performance metrics of abnormality detection on Dataset B

Dataset B			
	SEN	SPE	ACC
Baseline PCA	$100.00 \pm 0.00\%$	$11.11 \pm 11.11\%$	$20.00 \pm 21.78\%$
Baseline RPCA	$100.00 \pm 0.00\%$	$77.42 \pm 2.93\%$	$84.44 \pm 2.06\%$
RPCA-P with ξ_{var}	$100.00 \pm 0.00\%$	$63.86 \pm 3.47\%$	$75.11 \pm 2.48\%$
RPCA-P with ξ_{std}	$100.00 \pm 0.00\%$	$86.46 \pm 2.69\%$	$90.67 \pm 1.87\%$
RPCA-P with ξ_{apd}	$100.00 \pm 0.00\%$	$90.97 \pm 1.68\%$	$93.78 \pm 1.17\%$

application with ξ_{var} led to a lower performance, scoring only $63.86 \pm 3.47\%$ specificity and $75.11 \pm 2.48\%$ overall accuracy. The sensitivity was fixed at 100% in all cases.

A notable point is that the detection performance with the baseline PRCA method here is lower than that in Chapter 3 Section 3.6.5 using the original methodology. This is because feature decomposition was no longer applied to a dedicated ROI but to the whole embryo instead, in which the singular features in neighbouring structures could lead to false positive anomaly detections, and thus lowered the specificity and overall accuracy.

4.4.6 Computation time

The same runtime environment was deployed as in our previous study, on a standard PC with an Intel i7 3.4GHz quad-core CPU and 32GB RAM memory without parallel settings. However, this time feature decomposition was applied to the whole embryo rather than a small ROI. An embryo in Dataset A occupies $m_A \approx 3 \times 10^7$ voxels, and a counterpart in Dataset B occupies $m_B \approx 2 \times 10^7$ voxels.

Taking all runs of cross validation into consideration, the average RPCA-P processing time on Dataset A was around 22 min for all 30 images per run (equivalent to around 44s per image), where $size(D_A) \approx 9 \times 10^8$. No significant difference was identified between the use of ξ_{var}, ξ_{std} and ξ_{apd} prior models. The algorithm generally took about 20 iterations to converge. On the other hand, the baseline RPCA method generally took about 30 iterations to converge, which inreased the average processing time to 34 min (equivalent to around 68s per image).

In the case of Dataset B, the RPCA-P algorithm average processing was around 10 *min* (equivalent to around 42*s* per image), which was also faster than the baseline method which took an average 15.5 *min* (equivalent to around 66*s* per image) to process, with $size(D_B) \approx 3 \times 10^8$. There was no notable difference between the use of variation prior model either. In addition, a faster convergence rate was witnessed again with RPCA-P generally taking about 20 iterations to converge while the baseline RPCA taking about 30 iterations.

4.5 Discussion

Over the past decades, there have been numerous studies trying to make the PCA technique more robust to significant outlier presence or partial data distortion. With its substantial performance guarantees, the PCP approach has gradually become the standard RPCA framework. In addition to feature extraction and dimensionality reduction, this framework also provides a feature decomposition capability that is able to decompose a collection of (distorted) observation data, into a low-dimensional regular structure and a sparse component mainly composed of distortion factors. For that reason, it can be employed either to recover the undistorted form of the data (such as to recover a matrix with missing entries), or detect the presence of anomaly within the data (such as to detect intrusion in video surveillance). The decomposition process can be adjusted by a tunable parameter λ that balances the weights between including marginal data into the regular or singular components. In the mouse phenotyping study, our methodology is centred on leveraging its anomaly detection capability.

However, despite the considerable success in computer vision, the application of the baseline RPCA method to biomedical imaging has proven to be much more challenging. This is primarily due to the wide prevalence of various forms of natural variation in biomedical data, which substantially undermines the PCP assumption of a stationery background for subspace estimation. As a remedy, our proposed RPCA-P method can relax this condition by allowing a certain degree of background variation, and applying a tailored outlier tolerance for each single voxel/dimension. In other words, it allows voxel-wise adjustment of feature decomposition, so

that in the regions experiencing higher or lower natural variability, the level of strictness for outliers to be considered anomalous can be relaxed or tightened by some localised fine-tuning. In this case, RPCA-P can be applied to the entire embryo (or entire image if memory is not concerned) directly, instead of a dedicated ROI as in the case of the baseline RPCA in our original methodology.

The local adjustment is realised by the inclusion of a variation prior ξ in the modified PCP formulation. In the phenotyping application, it is essentially a weight image that estimates the degree of natural variation at voxel level, and is learned from the data itself, bearing the assumption that abnormal deformations are sparse and thus do not significantly affect the ξ estimation based on a majority of normal images in the test dataset. In this sense, the RPCA-P method could be regarded as a special semi-supervised learning approach, compared to the purely unsupervised approach using the baseline RPCA method and the purely supervised approach in a standard label classification problem. Three simple learning models were proposed in this study: voxel-wise intensity variance ξ_{var}, voxel-wise standard deviation ξ_{std} and average pair-wise discrepancy ξ_{apd}. Each was proven to be effective in terms of local tolerance adjustment, leading to different feature decomposition outcomes. In general, the use of PRCA-P led to improved performance of abnormal phenotype detection compared to the baseline RPCA, except for the case applied to Dataset B with ξ_{var}. In particular, both RPCA-P with ξ_{std} and ξ_{apd} contributed to around 12% increase of detection specificity in the case of VSD phenotype, and respectively, achieved 9% and 13.5% increase in the case of polydactyly phenotype.

4.6 Conclusion

In this study we proposed a novel RPCA-P framework that is able to incorporate prior information regarding the local degree of outlier tolerance at each data dimension in the feature decomposition process. In terms of anomaly detection in biomedical imaging data, the natural variation can be learned a priori at the voxel level, and a prior model is then incorporated in the subsequent RPCA-P process to locally adjust outlier tolerance. In this case, the proposed

method in theory should be able to address natural variation and improve the performance of feature decomposition in the biomedical domain.

In our test application to the mouse embryo phenotyping problem, three simple learning models were suggested for prior estimation, respectively, based on voxel-wise intensity variance, voxel-wise standard deviation and average pair-wise discrepancy, each was proven to be effective in terms of local decomposition adjustment. Moreover, the refined methodology using the RPCA-P technique (with ξ_{apd} in particular) significantly outperformed the original methodology using the baseline RPCA technique, without restricting feature decomposition to a dedicated ROI. The best anomaly detection performance achieved was 100% sensitivity, 94.12±1.36% specificity and 95.24±1.09% overall accuracy for the VSD phenotype, and 100% sensitivity, 90.97±1.68% specificity and 93.78 ± 1.17% overall accuracy for the polydactyly phenotype.

Chapter 5

Summary and Future Work

5.1 Summary

To sum up, the work presented in this thesis is centred on developing machine learning techniques for efficient recognition of anatomical structures and morphological abnormalities in biomedical images. Three investigations have been carried out during the course of this thesis:

Image segmentation using a patch-based canonical neural network

The first study explores the use of deep learning in patch-based image segmentation. Segmentation of anatomical structures in the biomedical image is a fundamental step of image analytics, and is often a prerequisite to advanced procedures such as diagnostics. A large proportion of existing image segmentation frameworks proposed over the past decades have heavily rested on two signal processing-based paradigms: label propagation via image registration and pair-wise patch-based pattern matching. Despite a high segmentation accuracy having been reported in many applications, the computational efficiency is generally limited, and an additional atlas selection process is often required to identify the most suitable atlases to help with segmentation.

More recently, the employment of machine learning has gained wider attention. In particular,

we argue that better patch-based classification can be achieved by harnessing the power of deep learning. In contrast to the low-level hand-engineered features (such as the sum of squared differences) used in the conventional patch-based segmentation approaches, we train a deep neural network that is able to achieve highly intricate feature representation and classification capabilities. Following the remarkable success in computer vision, some deep learning approaches, especially with the use of the convolutional neural network (ConvNet), have been frequently applied to biomedical imaging lately. The ConvNet model is a special type of neural network that features a structure with multiple layers, each is composed of a set of independent computational neurons that perform convolutional filterings on images. In each single convolution process, an image is filtered with a sliding kernel typically of a small size (such as $5 \times 5 \times 5$). By projecting images over many layers of convolution filters, a deep ConvNet architecture can yield advanced feature representation capacity while using far less learning weights compared to a canonical neural network (CanonNet) model.

In terms of image segmentation, the process is often broken down to voxel-wise label classification with a patch-based setting, where each patch is treated as a mini-image for the classification of its centre voxel. However, we argue that in the context of patch-based segmentation, the ConvNet has little advantage over the CanonNet architecture. This is because a patch is small (often around 9×9 to 15×15 only), and thus do not need further decomposition and will not benefit from convolution. Instead, we make use of the CanonNet in which neurons only compute dot products. Meanwhile we also incorporate modern techniques of deep learning, including GPU programming, Rectified Linear Unit (ReLU) activation, dropout layers and 2.5D tri-planar patch multi-pathway setting. The resulting classifier is much faster and less memory-hungry than convolution based networks. In our test application to the segmentation of hippocampus in human brain MR images, we significantly outperformed prior state-of-the-art with a median Dice score up to 90.98% at a near real-time speed (<1s). To the best of our knowledge, this is the fastest hippocampus segmentation algorithm with the highest segmentation accuracy ever reported on a comparable size of experiments at the time of the work.

Mouse phenotyping with the combined use of non-rigid registration and robust principal component analysis

The second study is an investigation into mouse phenotyping, and develops a high-throughput framework to detect morphological abnormality in imaging data with the combined use of non-rigid registration and robust principal component analysis (RPCA).

Significant research efforts have been underway toward leveraging gene modification and image analytics to help phenotype the mouse genome. Existing research can be broadly divided into two branches of work, respectively, based on the detection of phenotype-specific features and comparative analytics. The first branch is primarily centred on the classification of certain known phenotypes and is thus unable to deliver a general phenotyping purpose. The second branch, despite being more general, primarily rests on either volumetric contrast or deformation-based morphometrics to separate abnormal subjects from the normal. Detection of morphological abnormality via volumetric contrast often fails when there is no severe volume variation involved, whereas the detection via deformation-based morphometrics on the other hand is not very robust since the deformation features may vary significantly with the use of different registration settings and templates. Furthermore, existing approaches often require image segmentation on certain structures of interest before proceeding to further analysis, which is very challenging when applied to mouse embryo phenotyping.

In contrast, we propose a novel phenotyping approach centred on feature decomposition, which divides imaging data into a regular and a singular component, the latter of which is then used to detect morphological abnormality. In essence, we make use of non-rigid registration to group-wise align target images in a locally generated template space, followed by RPCA processing to realise the desired feature decomposition. The proposed framework is able to efficiently perform abnormality detection in a batch of images simultaneously, sensitive to both volumetric and non-volumetric variations, and does not require image segmentation. A validation study has been applied to two datasets of mouse embryo μ-CT images, and successfully distinguished the VSD and polydactyly from the normal phenotype with a 100% sensitivity. The detection specificities achieved for these two abnormal phenotypes were, respectively, 85.19$\pm$1.29% and 88.89$\pm$3.15%,

and the resulting overall accuracies were, respectively, $87.10 \pm 1.13\%$ and $93.34 \pm 1.84\%$.

Robust principal component analysis with variation priors (RPCA-P)

The third study investigates the RPCA framework in more depth and proposes a novel RPCA-P framework that is better able to address the prevailing natural variations in biomedical data. RPCA is an extension of the classic PCA, which is a widely used statistical data analytical technique concerning feature extraction and dimensionality reduction. Due to its strong performance guarantees, RPCA via principal component pursuit (PCP) has gradually become the standard RPCA method, and has previously been applied to many computer vision problems, attaining fruitful results. However, all these applications rest on a common condition, in which there is a stationery background that is invariant across the majority of the imaging data being processed.

In contrast, there is a significant natural variation in the biomedical domain, which distinguishes one subject from another. In this case, the baseline RPCA does not work to the demanding quality when directly applied to biomedical imaging. Although image registration can be used to group-wise align these images non-rigidly and to some extent reduces such variance. The problem however cannot be eliminated entirely. In particular, different anatomical structures often have individual natural variations, whereas the level of outlier tolerance for feature decomposition in the baseline RPCA framework is solely controlled by a single parameter that applies globally regardless of local variability, which further complicates the situation.

To improve this purely unsupervised machine learning approach without leveraging any prior knowledge, we propose a semi-supervised approach with a modified RPCA framework (RPCA-P), which is able to incorporate variation priors in the model and adjusts outlier tolerance locally so that voxels associated to structures of higher natural variability are compensated by allowing a higher tolerance during feature decomposition. Furthermore, the variation prior model can be learned in advance from the data itself using a range of learning models.

A revised methodology using RPCA-P was proposed in the application to the mouse embryo

phenotyping problem. In particular, feature decomposition was applied to the whole embryo rather than restricted to the structure of interest. Furthermore, we proposed three learning models, respectively, based on voxel-wise intensity variance, voxel-wise standard deviation and average pair-wise discrepancy, to estimate the variation priors using local data. In our evaluation using the same mouse embryo data, the revised methodology achieved better feature decomposition and anomaly detection results, with the performance metrics raised to 100% sensitivity, $94.12 \pm 1.36\%$ specificity and $95.24 \pm 1.09\%$ overall accuracy for the VSD phenotype, and to 100% sensitivity, $90.97 \pm 1.68\%$ specificity and $93.78 \pm 1.17\%$ overall accuracy for the polydactyly phenotype.

5.2 Future Work

Extensions on the patch-based deep learning segmentation framework

There are a number of future extensions that can be suggested based on the patch-based deep learning segmentation work presented in this thesis. To start with, as a pilot validation study, this framework has only been tested on the segmentation of hippocampus in the adult human brain with 100 MR images retrieved from one database. A broader experimentation on a range of anatomical structures, tissues or brain parcellations using larger, potentially multi-modality and/or cross-database image sets should be carried out, before a more comprehensive conclusion could be drawn regarding the full capacity of the proposed method in terms of both feature representation potential and segmentation performance (accuracy and speed).

Furthermore, while boasting the superiority of the proposed architecture based on the CanonNet, one of the limitations of the current work is the lack of a direct comparative study with ConvNet methods. A major reason is the lack of ConvNet work on hippocampus segmentation reporting a state-of-the-art accuracy, and adapting a more general framework from other studies (such as the U-Net [131] or DeepMedic [78], which were, respectively, validated on the segmentation of neuronal structures in microscopy images and brain lesion in multi-modality MR images) requires non-trivial engineering and experimentation work. Nevertheless, the pros

and cons of "CanonNet vs ConvNet" should be more comprehensively addressed in an extensional study.

Last but not least, recent years have witnessed an exponential growth of deep learning studies, applied to a large variety of problems, including image classification, face recognition, natural language understanding in addition to biomedical imaging. However, in the majority of these studies, usually only a fixed network architecture was described and tested. The reasons for choosing the hyper-parameter configuration, such as the number of layers in the network, the number of neurons in each layer, or the impact of a different configuration, were generally not elucidated. A detailed investigation into hyper-parameter configuration would be not only very demanding, but also highly constructive to the future development of deep learning in general.

Extensions on the mouse phenotyping framework

First of all, registering an image undergoing abnormal deformations with a normal subject in practice is challenging, since the abnormal features do not match with the normal counterparts. Although the nature of difference is utilised for feature decomposition, it may also cause minor displacement of the target structures, which in turn leads to more singular features generated around the boundaries and can potentially trigger false positive detection. Recently, a novel registration approach [105] has been proposed to address the presence of defects such as lesions and tumours, also by leveraging non-rigid registration and RPCA. It starts with a standard group-wise non-rigid image alignment. After that, RPCA is carried out to recover all the images into their regular forms (and separate the singular factors out), followed by a group-wise non-rigid registration performed on the recovered images, where the generated transformations are then used to align the original images. This process is iteratively executed until the images align to a satisfactory quality. A similar group-wise alignment scheme has also been used in computer vision to deal with images with partial occlusion or corruption [124]. This approach may be incorporated in our phenotyping framework to replace the current group-wise alignment method. We consider it might be able to improve the follow-up feature decomposition and anomaly detection performance, yet on the other hand would significantly increases

computational burden. Its practical merit requires further investigations.

Secondly, there were only two phenotypes tested in our evaluation studies: VSD and polydactyly. Also, the test datasets were relatively small, with a total of 46 mouse embryo images. To fully examine the performance of our phenotyping approach, a more comprehensive study should be carried out on many other phenotypes with a much larger test dataset. Moreover, all our test data were μ-CT images, and the performance on other modalities such as μ-MRI requires further experimentation.

Furthermore, anomaly detection in the current framework is a region-level classification based on measuring anomaly rate over a target region. A bad selection of such a region, for example a small one that covers the target structure only by a portion, or a large one that covers many other structures, could potentially undermine the detection performance. A more desirable anomaly detection outcome may be a label map that accurately annotates the anomalous part of each defective structure in the original image. This may be achieved by some post-processing on the singular feature images, yet it could easily lead to scattered false positive signals and may require additional fine-tuning. Last but not least, if a large pool of samples of different genotypes is available, further statistical analysis may be conducted on the regions consistently showing large inter-genotype discrepancy, to examine systematic correlation of the abnormality with genotype and gene modification.

Extensions on the RPCA-P technique

To start with, there were three learning models to estimate variation priors examined in this thesis, respectively, based on voxel-wise intensity variance, voxel-wise standard deviation and average pair-wise discrepancy. Although the effectiveness of local tolerance adjustment was demonstrated for each prior model, they all ended up with somewhat different feature decomposition outcomes, and led to different anomaly detection results. Also, there could be many other learning models besides the proposed three. Therefore a direct future study could be an extensional investigation regarding the optimal learning model for variation prior estimation, which in fact is probably application-dependent. Moreover, in our work we learned the

variation priors using all images in the test dataset, based on the assumption that abnormal deformations are sparse and thus will make limited impact on the estimation outcome. Learning approaches in other situations could also be explored, for example when there is a large proportion of subjects with abnormal phenotypes, and/or some of the subjects are known to be normal/abnormal in advance. In other words, an insightful generalisation over variation prior learning is highly desirable, and can significantly improve the applicability of the RPCA-P technique to different problem domains.

Furthermore, the nature of RPCA or RPCA-P lies on feature decomposition, and thereby we consider it to be widely applicable to many other problems regarding anomaly detection and regular structure recovery. The fundamental purpose of the RPCA-P technique is to enhance the performance of RPCA, in the conditions where the background undergoes certain levels of variation. Such variation is particularly common in biomedical data, but also applies to other computer vision problems, such as a partially moving background or the use of a non-stationery camera in video surveillance, and a direct RPCA application is likely to fail in these conditions. For that reason, a future study in this line of work could be an investigation regarding whether or how we would be able to improve the detection of anomalous objects in these scenarios using RPCA-P. For instance, if the background includes a vibrating or swinging object (like a pendulum) during the course of surveillance, one might be able to learn and incorporate variation priors into feature decomposition, so that it only triggers anomaly detection when there is a more substantial variation occurring in that region, such as the presence of another object.

Appendix: Elastix configuration in Chapter 3

Configuration of Rigid Registration

//ImageTypes

(FixedImagePixelType "float")

(FixedImageDimension 3)

(MovingImagePixelType "float")

(MovingImageDimension 3)

//Multi-Resolution

(Registration "MultiResolutionRegistration")

(NumberOfResolutions 3)

(MaximumNumberOfIterations 500)

//Pyramid

(FixedImagePyramid "FixedSmoothingImagePyramid")

(MovingImagePyramid "MovingSmoothingImagePyramid")

//ImageSampler

(ImageSampler "RandomCoordinate")

(NumberOfSpatialSamples 3000)

(NewSamplesEveryIteration "true")

(UseRandomSampleRegion "true")

(SampleRegionSize 2.5 2.25 3)

//Metric

(Metric "AdvancedMattesMutualInformation")

```
(NumberOfHistogramBins 64)
(CheckNumberOfSamples "true")

//Optimizer
(Optimizer "AdaptiveStochasticGradientDescent")
(AutomaticParameterEstimation "true")
(MaximumNumberOfSamplingAttempts 10)

//Transform
(Transform "EulerTransform")
(AutomaticScalesEstimation "true")
(AutomaticTransformInitialization "true")
(AutomaticTransformInitializationMethod "GeometricalCenter")
(UseDirectionCosines "true")
(HowToCombineTransforms "Compose")

//Interpolator and Resampler
(Interpolator "LinearInterpolator")
(Resampler "DefaultResampler")
(ResampleInterpolator "FinalBSplineInterpolator")
(FinalBSplineInterpolationOrder 3)
```

Configuration of Affine Registration

```
//ImageTypes
(FixedImagePixelType "float")
(FixedImageDimension 3)
(MovingImagePixelType "float")
(MovingImageDimension 3)

//Multi-Resolution
(Registration "MultiResolutionRegistration")
```

(NumberOfResolutions 3)

(MaximumNumberOfIterations 500)

//Pyramid

(FixedImagePyramid "FixedSmoothingImagePyramid")

(MovingImagePyramid "MovingSmoothingImagePyramid")

//ImageSampler

(ImageSampler "RandomCoordinate")

(NumberOfSpatialSamples 3000)

(NewSamplesEveryIteration "true")

(UseRandomSampleRegion "true")

(SampleRegionSize 2.5 2.25 3)

//Metric

(Metric "AdvancedMattesMutualInformation")

(NumberOfHistogramBins 64)

(CheckNumberOfSamples "true")

//Optimizer

(Optimizer "AdaptiveStochasticGradientDescent")

(AutomaticParameterEstimation "true")

(MaximumNumberOfSamplingAttempts 10)

//Transform

(Transform "AffineTransform")

(AutomaticScalesEstimation "true")

(UseDirectionCosines "true")

(HowToCombineTransforms "Compose")

//Interpolator and Resampler

(Interpolator "LinearInterpolator")

(Resampler "DefaultResampler")

(ResampleInterpolator "FinalBSplineInterpolator")

(FinalBSplineInterpolationOrder 3)

Configuration of B-spline Registration

```
//ImageTypes
(FixedImagePixelType "float")
(FixedImageDimension 3)
(MovingImagePixelType "float")
(MovingImageDimension 3)

//Multi-Resolution
(Registration "MultiMetricMultiResolutionRegistration")
(NumberOfResolutions 4)
(MaximumNumberOfIterations 4000)

//Pyramid
(FixedImagePyramid "FixedSmoothingImagePyramid")
(MovingImagePyramid "MovingSmoothingImagePyramid")

//ImageSampler
(ImageSampler "RandomCoordinate")
(NumberOfSpatialSamples 3000 )
(NewSamplesEveryIteration "true")
(UseRandomSampleRegion "true")
(SampleRegionSize 2.5 2.25 3)

//Metric
(Metric "AdvancedMattesMutualInformation" "TransformBendingEnergyPenalty")
(Metric0Weight 1.0)
(Metric1Weight 0.01)
(NumberOfHistogramBins 64)

//Optimizer
```

(Optimizer "AdaptiveStochasticGradientDescent")

(AutomaticParameterEstimation "true")

(MaximumNumberOfSamplingAttempts 5)

//Transform

(Transform "BSplineTransform")

(BSplineTransformSplineOrder 3)

(FinalGridSpacingInPhysicalUnits 0.2) // 0.4 for 400μm and 0.1 for 100μm settings

(UseDirectionCosines "true")

(HowToCombineTransforms "Compose")

//Interpolator and Resampler

(Interpolator "LinearInterpolator")

(Resampler "DefaultResampler")

(ResampleInterpolator "FinalBSplineInterpolator")

(FinalBSplineInterpolationOrder 3)

Bibliography

[1] D. Adams, R. Baldock, S. Bhattacharya, A. J. Copp, M. Dickinson, N. D. Greene, et al. Bloomsbury report on mouse embryo phenotyping: recommendations from the IMPC workshop on embryonic lethal screening. *Disease Models & Mechanisms*, 6(3):571–579, 2013.

[2] C. B. Akgul, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar. Content-based image retrieval in radiology: current status and future directions. *Journal of Digital Imaging*, 24(2), 208–222, 2011.

[3] A. A. Ali, A. M. Dale, A. Badea, and G. A. Johnson. Automated segmentation of neuroanatomical structures in multispectral MR microscopy of the mouse brain. *NeuroImage*, 27(2):425–435, 2005.

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

[5] P. R. Andresen and M. Nielsen. Non-rigid registration by geometry-constrained diffusion. *Medical Image Analysis*, 5(2):81–88, 2001.

[6] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano. Efficient classifier generation and weighted voting for atlas-based segmentation: two small steps faster and closer to the combination oracle. In: *SPIE Medical Imaging*, vol. 6914, no. 69141W, 2008.

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

[8] S. D. Babacan, M. Luessi, R. M. Molina, and A. Katsaggelos. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Transactions on Signal Processing*, 60(8):3964–3977, 2011.

[9] M. H. Bae, R. Pan, T. Wu, and A. Badea. Automated segmentation of mouse brain images using extended MRF. *NeuroImage*, 46(3):717–725, 2009.

[10] W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. *IEEE Transactions on Medical Imaging*, 32(7):1302–1315, 2013.

[11] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), 2009.

[12] P.-L. Bazin and D. L. Pham. Homeomorphic brain image segmentation with topological and statistical atlases. *Medical Image Analysis*, 12(5):616–625, 2008.

[13] A. Beck and M. Teboulle: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[14] S. Becker, E. Candes, and M. Grant. TFOCS: flexible first-order methods for rank minimization, In: *SIAM Conference on Optimization: Low-Rank Matrix Optimization Symposium*, 2011.

[15] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300, 1995.

[16] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4): 1165–1188, 2001.

[17] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Method*. Academic Press, 1982.

[18] K. K. Bhatia, J. Hajnal, A. Hammers, and D. Rueckert. Similarity metrics for group-wise non-rigid registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 4792, pp. 544–552. Springer, 2007.

[19] D. L. Bihan. Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience*, 4(6):469–480, 2003.

[20] C. Bishop . *Pattern Recognition and Machine Learning*. Springer, 2007.

[21] M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–92, 1996.

[22] M. Boccardi, R. Ganzola, M. Bocchetta, M. Pievani, A. Redolfi, G. Bartzokis, R. Camicioli, J. G. Csernansky, M. J. de Leon, L. deToledo-Morrell, et al. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. *Journal of Alzheimer's Disease*, 26:61–75, 2011.

[23] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6): 567–585, June 1989.

[24] S. D. M. Brown and M. W. Moore. Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Disease Models & Mechanisms*, 5(3):289–292, 2012.

[25] S. D. M. Brown, J. M. Hancock, and H. Gates. Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. *PLoS Genetics*, 2(8):1131–1137, 2006.

[26] M. Cabezas, A. Oliver, X. Llado, J. Freixenet, and M. B. Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177, 2011.

[27] J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[28] E. J. Candes and B. Recht. Exact matrix completion via convex optimzation. *Foundations of Computational Mathematics*, 9:717–772, 2009.

[29] E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[30] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[31] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[32] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis?. *Journal of the ACM*, 58(3):article 11, 2011.

[33] M. R. Capecchi. Generating mice with targeted mutations. *Nature Medicine*, 7(10):1086–1090, 2001.

[34] X. J. Chen, N. Kovacevic, N. J. Lobaugh, J. G. Sled, R. M. Henkelman, and J. T. Henderson. Neuroanatomical differences between mouse strains as shown by high-resolution 3D MRI. *NeuroImage*, 29(1):99–105, 2006.

[35] J. O. Cleary, M. Modat, F. C. Norris, A. N. Price, S. A. Jayakody, J. P. Martinez-Barbera, N. D. E. Greene, D. J. Hawkes, R. J. Ordidge, P. J. Scambler, S. Ourselin and M. F. Lythgoe. Magnetic resonance virtual histology for embryos: 3D atlases for automated high-throughput phenotyping. *NeuroImage*, 54(2):769–778, 2011.

[36] D. Collins and J. Pruessner. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52(4): 1355–1366, 2010.

[37] P. Coupe, J. V. Manjon, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.

[38] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. *IEEE Transactions on Medical Imaging*, 27(4):425–441, 2008.

[39] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. A viscous fluid model for multimodal non-rigid image registration using mutual information. *Medical Image Analysis*, 7(4):565–575, 2003.

[40] R. Datteri, A. Asman, B. Landman, and B. Dawan. Estimation of registration accuracy applied to multiatlas segmentation. In: *MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion*, 2011.

[41] F. De la Torre and M. J. Black. A robust principal component analysis for computer vision. In: *International Conference on Computer Vision*, Vancouver, Canada, 2001.

[42] F. De la Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1):117–142, 2003.

[43] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

[44] A. P. Dhawan. *Medical Image Analysis*. Wiley, 2011.

[45] N. Dhungel, G. Carneiro, and A.P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 9349, pp. 605–612. Springer, Heidelberg, 2015.

[46] L. Z. Diaz-de-Grenu, J. Acosta-Cabronero, J. Pereira, G. Pengas, G. B. Williams, and P. J. Nestor. MRI detection of tissue pathology beyond atrophy in Alzheimer's disease: introducing T2-VBM. *NeuroImage*, 56(4):1946–1953, 2011.

[47] A. E. Dorr, J. P. Lerch, S. Spring, N. Kabani, and R. M. Henkelman. High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult C57Bl/6J mice. *NeuroImage*, 42(1):60–69, 2008.

[48] K. M. Downs and T. Davies. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development*, 118(4):1255–1266, 1993.

[49] S.F. Eskildsen, P. Coupe, V. Fonov, J. V. Manjon, K. K. Leung, N. Guizard, S.N. Wassef, L. R. ÃŸstergaard, and D. L. Collins, Alzheimer's Disease Neuroimaging Initiative. BEaST: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373, 2012.

[50] L. Feldkamp, L. Davis, and J. Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984.

[51] B. Fischer and J. Modersitzki. A unified approach to fast image registration and a new curvature based registration technique. *Linear Algebra and its Applications*, 380(15):107–124, 2004.

[52] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[53] K. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, R. S. Frackowiak, and R. S. J. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995.

[54] Z. Gao, L. Cheong, and M. Shan. Block-sparse RPCA for consistent foreground detection. In: *European Conference on Computer Vision (ECCV)*, 2012.

[55] C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4): 870–878, 2002.

[56] I.S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40:672–684, 2008.

[57] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. URL: `http://cvxr.com/cvx/`, last visit: July 2016.

[58] D. Gross, Y.-K. Liu, S. T. Flammia, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010.

[59] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[60] E. M. Haacke, R. W. Brown, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, John Wiley & Sons, 1999.

[61] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes (eds). *Medical Image Registration.* CRC Press, 2001.

[62] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for L1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3)1107–1130, 2008.

[63] L. Harper, F. Barkhof, P. Scheltens, J. M. Schott, and N. C. Fox. An algorithmic approach to structural imaging in dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(6):692-698, 2014.

[64] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1), 115âĂŞ26, Oct. 2006.

[65] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Multiclassifier fusion in human brain MR segmentation: modelling convergence. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 4191, pp. 815âĂŞ822. Springer Berlin Heidelberg, 2006.

[66] J. R. Hesselink. Basic Principle of MR Imaging. URL: `http://spinwarp.ucsd.edu/neuroweb/Text/br-100.htm`, last visit: July 2016.

[67] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–R45, 2001.

[68] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.

[69] J. I. E. Hoffman and S. Kaplan. The incidence of congenital heart disease. *Journal of the American College of Cardiology*, 39(12):1890–1900, 2002.

[70] Y. Hsu, N. Schuff, A. Du, K. Mark, X. Zhu, D. Hardin, and M. Weiner. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. *Journal of Magnetic Resonance Imaging*, 16(3):305–310, 2002.

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

[72] International Mouse Knockout Consortium, F. S. Collins, J. Rossant, and W. Wurst. A mouse for all reasons. *Cell*, 128(1):9–13, 2007.

[73] C. Jack, M. Bernstein, N. C. Fox, P. Thompson, G. Alexander, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27 (4):685–691, 2008.

[74] G. A. Johnson, A. Badea, J. Brandenburg, G. Cofer, B. Fubara, S. Liu, and J. Nissanov. Waxholm space: an image-based reference for coordinating mouse brain research. *NeuroImage*, 53(2):365–372, 2010.

[75] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[76] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151-S160, 2004.

[77] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[78] K. Kamnitsas, C. Ledig, V. F.J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36: 61–78, 2016.

[79] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[80] M. H. Kaufman. *The Atlas of Mouse Development*. Academic Press, London, 1994.

[81] Q. Ke and T. Kanade. Robust L1-norm factorization in the presence of outliers and missing data. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[82] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[83] S. Klein and M. Staring. *Elastix: the manual*, 2015. URL: `http://elastix.isi.uu.nl/download/elastix_manual_v4.8.pdf`, last visit: July 2016.

[84] S. Klein, M. Staring, and J. P. W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, 16(12):2879–2890, 2007.

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

[86] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim. Elastix: a toolbox for intensity based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.

[87] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi. Neighbourhood approximation forests. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 7512, pp. 75–82, 2012.

[88] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi. Neighbourhood approximation using randomized forests. *Medical Image Analysis*, 17: 790–804, 2013.

[89] N. Kovacevic, J. T. Henderson, E. Chan, N. Lifshitz, J. Bishop, A. C. Evans, R. M. Henkelman, and X. J. Chen. A three-dimensional MRI atlas of the mouse brain with estimates of the average and variability. *Cerebral Cortex*, 15(5):639–645, 2005.

[90] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 1106-1114, 2012.

[91] B. Landman, and S. Warfield (eds). *MICCAI 2012 Workshop on Multi-Atlas Labeling*. CreateSpace, Nice, France, 2012.

[92] M. Larobina and L. Murino. Medical image file formats. *Journal of Digital Imaging*, 27(2):200–206, 2014.

[93] J. C. Lau, J. P. Lerch, J. G. Sled, R. M. Henkelman, A. C. Evans, and B. J. Bedell. Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of Alzheimer's disease. *NeuroImage*, 42(1):19–27, 2008.

[94] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[95] S. Lee, G. Wolberg, K.-Y. Chwa, and S. Y. Shin. Image metamorphosis with scattered feature constraints. *IEEE Transactions on Visualization and Computer Graphics*, 2(4):337–354, 1996.

[96] S. Lee, G. Wolberg, and S. Y. Shin. Scattered data interpolation with multilevel B-splines. *IEEE Transactions on Visualization and Computer Graphics*, 3(3):228–244, 1997.

[97] J. Lee, J. Jomier, S. Aylward, M. Tyszka, S. Moy, J. Lauder, and M. Styner. Evaluation of atlas based mouse brain segmentation. In: *SPIE Medical Imaging*, vol. 7259, no. 725943, 2009.

[98] K. van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999.

[99] T. M. Lehmann, C. Gonner, and K. Spitzer. Survey: Interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18(11):1049–1075, 1999.

[100] J. P. Lerch, J. B. Carroll, S. Spring, L. N. Bertram, C. Schwab, M. R. Hayden, and R. M. Henkelman. Automated deformation analysis in the YAC128 Huntington disease mouse model. *NeuroImage*, 39(1):32–39, 2008.

[101] H. Lester and S. R. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149, 1999.

[102] F. van der Lijn, T. den Heijer, M. M. B. Breteler, and W. J. Niessen. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43(4):708–720, 2008.

[103] Z. Lin, M. Chen, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055, 2009.

[104] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In: *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

[105] X. Liu, M. Niethammer, R. Kwitt, M. McCormick, and S. Aylward. Low-rank to the rescue – atlas-based analyses in the presence of pathologies. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 8675, pp. 97–104, 2014.

[106] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001.

[107] M. Lorenzo-Valdes, G. I. Sanchez-Ortiz, A. G. Elkington, R. H. Mohiaddin, and D. Rueckert: Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. *Medical Image Analysis*, 8(3):255–265, 2004.

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

[109] J. Lotjonen, R. Wolz, J. Koikkalainen, V. Julkunen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. *NeuroImage*, 56(1):185–196, 2011.

[110] Y. Ma, P. R. Hof, S. C. Grant, S. J. Blackband, R. Bennett, L. Slatest, M. D. McGuigan, and H. Benveniste. A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Neuroscience*, 135(4):1203–1215, 2005.

[111] J. V. Manjon, P. Coupe, L. Marti-Bonmati, D. L. Collins, and M. Robles. Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, 31:192–203, 2010.

[112] B. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[113] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[114] M. Modat, G. Ridgway, Z. Taylor, M. Lehmann, J. Barnes, D. Hawkes, N. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278–284, 2010.

[115] J. Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[116] Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, J Rogers, J. F. Abril, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[117] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005.

[118] Y. Nesterov. A method of solving a convex programming problem with convergence rate O(1/k2), *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[119] J. Nie and D. Shen. Automated segmentation of mouse brain images using multi-atlas multi-ROI deformation and label fusion. *Neuroinformatics*, 11(1):35–45, 2013.

[120] F. C. Norris, M. Modat, J. O. Cleary, A. N. Price, K. McCue, P. J. Scambler, S. Ourselin, and M. F. Lythgoe. Segmentation propagation using a 3D embryo atlas for high-throughput MRI phenotyping: comparison and validation with manual segmentation. *Magnetic Resonance in Medicine*, 69(3):877–883, 2012.

[121] L. Nyul and J. Udupa. Standardizing the MR image intensity scales: making MR intensities have tissuespecific meaning. In: *SPIE Medical Imaging*, vol. 3976, pp. 496–504, 2000.

[122] R. O'Rahilly and F. Muller (eds). *Developmental stages in human embryos*. Carnegie Institution of Washington, Washington, DC, 1987.

[123] H. Park, P. H. Bland and C. R. Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *IEEE Transactions on Medical Imaging*, 22(4):483–492, 2003.

[124] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012.

[125] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W.M. Wells. A Bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228–239, 2006.

[126] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 8150, pp. 246–253. Springer, Heidelberg, 2013.

[127] W. K. Pratt. *Digital Image Processing*. John Wiley & Sons, New York, 1978.

[128] C. E. Rodriguez-Carranza and M. H. Loew. Weighted and deterministic entropy measure for image registration using mutual information. In: *SPIE Medical Imaging*, vol. 3338, pp. 155–166, 1998.

[129] T. Rohlfing and C. R. Maurer. Multi-classifier framework for atlas-based image segmentation. *Pattern Recognition Letters*, 26(13):2070–2079, 2005.

[130] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004

[131] O. Ronneberger, P. Fischer, and T. Brox. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 9351, pp. 234–241. Springer, Heidelberg, 2015. 2015.

[132] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 9349, pp. 556–564. Springer, Heidelberg, 2015.

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

[134] S. Roy, X. Liang, A. Kitamoto, M. Tamura, T. Shiroishi, and M. S. Brown. Phenotype detection in morphological mutant mice using deformation features. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 8151, pp. 437–444, 2013.

[135] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[136] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[137] J. E. Schneider, S. D. Bamforth, C. R. Farthing, K. Clarke, S. Neubauer, and S. Bhattacharya. Rapid identification and 3D reconstruction of complex cardiac malformations in transgenic mouse embryos using fast gradient echo sequence magnetic resonance imaging. *Journal of Molecular and Cellular Cardiology*, 35(2):217–222, 2003.

[138] D. W. Shattuck, M. Mirza , V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3): 1064–1080, 2008.

[139] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998.

[140] M. Staring, S. Klein, and J. P. W. Pluim. A rigidity penalty term for nonrigid registration. *Medical Physics*, 34(11):4098–4108, 2007.

[141] C. Studholme, D. L. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[142] P. Suetens. *Fundamentals of Medical Imaging*. Cambridge University Press, 2009.

[143] V.-T. Ta, R.Giraud, D. L. Collins, and P. Coupe. Optimized patchMatch for near real time and accurate label fusion. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 8675, pp. 105–112. Springer, Heidelberg, 2014.

[144] M. Tamura, M. Hosoya, M. Fujita, T. Iida, T. Amano, A. Maeno, T. Kataoka, T. Otsuka, S. Tanaka, S. Tomizawa, and T. Shiroishi. Overdosage of hand2 causes limb and heart defects in the human chromosomal disorder partial trisomy distal 4q. *Human Molecular Genetics*, 22(12):2471–2481, 2013.

[145] G. Tang and A. Nehorai. Robust principal component analysis based on low-rank and block-sparse matrix decomposition, In: *Annual Conference on Information Sciences and Systems (CISS)*, 2011.

[146] K. Theiler. *The House Mouse: Atlas of Mouse Development*. Springer-Verlag, New York, 1989.

[147] P. Thevenaz and M. Unser. Optimization of mutual information for multi-resolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083–2099, 2000.

[148] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: series B*, 58:267–288, 1994.

[149] K. D. Toennies. *Guide to Medical Image Analysis: Methods and Algorithms*. Springer, 2012.

[150] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.

[151] T. Tong, R. Wolz, P. Coupe, J. V. Hajnal, and D. Rueckert. Segmentation of MR images via Discriminative Dictionary Learning and Sparse Coding: application to hippocampus labeling. *NeuroImage*, 76(1):11–23, 2013.

[152] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[153] Y. Wang and L. H. Staib. Elastic model based non-rigid registration incorporating statistical shape information. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 1496, pp. 1162–1173, 1998.

[154] Y. Wang and L. H. Staib. Physical model-based non-rigid registration incorporating statistical shape information. *Medical Image Analysis*, 4(1):7–21, 2000.

[155] H. Wang, J. W. Suh, S. Das, J. Pluta, C. Craige and P. Yushkevich. Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, 2013.

[156] Z. Wang, R.Wolz, T. Tong, and D. Rueckert. Spatially aware patch-based segmentation: an alternative patch-based segmentation framework. In: *Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging*, pp. 93–103, 2013.

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

[158] Z. Wang, C. Donoghue, and D. Rueckert. Patch-based segmentation without registration: application to knee MRI. In: *Machine Learning in Medical Imaging*, LNCS vol. 8184, pp. 98–105, 2013.

[159] Z. Wang. *Patch-based Segmentation with Spatial Context for Medical Image Analysis*. PhD Thesis, Imperial College London, 2014.

[160] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[161] W. J. Weninger, B. Maurer, B. Zendron, K. Dorfmeister, and S. H. Geyer. Measurements of the diameters of the great arteries and semi-lunar valves of chick and mouse embryos. *Journal of Microscopy*, 234(2):173–190, 2009.

[162] M. D. Wong, A. E. Dorr, J. R. Walls, J. P. Lerch, and R. M. Henkelman. A novel 3D mouse embryo atlas based on micro-CT. *Development*, 139(17):3248–3256, 2012.

[163] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: *Advances in Neural Information Processing Systems (NIPS)*, vol. 22, pp. 2080–2088, 2009.

[164] Y.-T. Wu, T. Kanade, C.-C. Li, and J. Cohn. Image registration using wavelet-based motion model. *International Journal of Computer Vision*, 38(2):129–152, 2000.

[165] G. Wu, F. Qi, and D. Shen. Learning-based deformable registration of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1145–1157, 2006.

[166] Z. Xie and D. Gillies. Patch forest: a hybrid framework of random forest and patch-based segmentation. In: *SPIE Medical Imaging*, vol. 9784, no. 978428. San Diego, USA, 2016.

[167] Z. Xie, X. Liang, L. Guo, A. Kitamoto, M. Tamura, T. Shiroishi, and D. Gillies. Automatic classification framework for ventricular septal defects: a pilot study on high-throughput mouse embryo cardiac phenotyping. *Journal of Medical Imaging*, 2(4):041003, 2015.

[168] Z. Xie, D. Yang, D. Stephenson, D. Morton, C. Hicks, T. Brown, and T. Bocan. Characterizing the regional structural difference of the brain between tau transgenic (rTg4510) and wild-type mice using MRI. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 6361, pp. 308–315, 2010.

[169] L. Xu and A. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.

[170] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for L1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1): 143–168, 2008.

[171] X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *Optimization Online*, 2009. URL: `http://www.optimization-online.org/DB_HTML/2009/11/2447.html`, last visit: July 2016.

[172] M. Zamyadi, L. Baghdadi, J. P. Lerch, S. Bhattacharya, J. E. Schneider, R. M. Henkelman, and J. G. Sled. Mouse embryonic phenotyping by morphometric analysis of MR images. *Physiological Genomics*, 42(2):89–95, 2010.

[173] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[174] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang. Robust principal component analysis with complex noise. In: *International Conference on Machine Learning*, Beijing, China, 2014.

[175] T. Zhou and D. Tao. GoDec: randomized low-rank & sparse matrix decomposition in noisy case. In: *International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[176] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, Stable principal component pursuit, In: *IEEE International Symposium on Information Theory (ISIT)*, pp. 1518–1522, 2010.

[177] S. Zhou. *Medical Image Recognition, Segmentation and Parsing*. Elsevier, 2015.

[178] S. C. Zhu, and A. Yuille. Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multi-band Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

[179] D. Zikic, B. Glocker, and A. Criminisi. Atlas encoding by randomized forests for efficient label propagation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS vol. 8151, pp. 66–73, 2013.

[180] D. Zikic, B. Glocker, and A. Criminisi. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Medical Image Analysis*, 18:1262–1273, 2014.

[181] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[182] T. Zouagui, E. Chereul, M. Janier, and C. Odet. 3D MRI heart segmentation of mouse embryos. *Computers in Biology and Medicine*, 40(1):64–74, 2010.