Adaptive Learning for Segmentation and Detection

Jingjing Deng

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

June 23, 2017

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

For The Future.

For

R D

I would like to dedicate this thesis to my grandfather, who passed away during my Ph.D studies. Thank you Grandpa.

Abstract

Segmentation and detection are two fundamental problems in computer vision and medical image analysis, they are intrinsically interlinked by the nature of machine learning based classification, especially supervised learning methods. Many automatic segmentation methods have been proposed which heavily rely on hand-crafted discriminative features for specific geometry and powerful classifier for delinearating the foreground object and background region. The aim of this thesis is to investigate the adaptive schemes that can be used to derive efficient interactive segmentation methods for medical imaging applications, and adaptive detection methods for addressing generic computer vision problems. In this thesis, we consider adaptive learning as a progressive learning process that gradually builds the model given sequential supervision from user interactions. The learning process could be either adaptive re-training for small scale models and datasets or adaptive fine-tuning for medium-large scale. In addition, adaptive learning is considered as a progressive learning process that gradually subdivides a big and difficult problem into a set of smaller but easier problems, where a final solution can be found via combining individual solvers consecutively. We first show that when discriminative features are readily available, the adaptive learning scheme can lead to an efficient interactive method for segmenting the coronary artery, where promising segmentation results can be achieved with limited user intervention. We then present a more general interactive segmentation method that integrates a CNN based cascade classifier and a parametric implicit shape representation. The features are self-learnt during the supervised training process, no hand-crafting is required. Then, the segmentation can be obtained via imposing a piecewise constant constraint to the detection result through the proposed shape representation using region based deformation. Finally, we show the adaptive learning scheme can also be used to address the face detection problem in an unconstrained environment, where two CNN based cascade detectors are proposed. Qualitative and quantitative evaluations of proposed methods are reported, and show the efficiency of adaptive schemes for addressing segmentation and detection problems in general.

Acknowledgements

Throughout the time of Doctor of Philosophy studying, I had the luck to experience the guidance, mentoring, friendship, assistance, criticism and love of many great people. It is their commitment that brought me in the position of being able to carry out scientific research and finish this thesis.

I wish to express my sincere gratitude towards my advisor, Dr. Xianghua Xie, who guided and supported me throughout the years. His office door was always open to me whenever i needed his advice, and has frequently provided me with inspiration and guidance to carry out my research. I also want to thank him for giving me the opportunity to explore and deal with challenging research topics. Without his tireless help and encouragement, this thesis would not have been possible, and he has all my thanks for that. I am also grateful to my other supervisors, Dr. Gary Tam, Dr. Rita Borgo, Dr. Ben Daubney, Prof. Jianfeng Ma and Prof. Yikun Zhang for their help and support. A special thankyou to my parents, my family for their help and patience throughout my studies. Without their support this thesis would not be possible, and I am very grateful.

I am thankful to my colleagues in the SwanseaVision group for their tips and advice over the past few years. They are Dr. Huaizhong Zhang, Dr. Hui Fang, Dr. Feng Zhao, Dr. Dongbin Chen, Dr. Ehab Essa, Dr. Jonathan Jones, Dr. Robert Palmer, Mike Edwards, David George, Dafydd Ravenscroft, Yaxi Ye and Ren Liu. I would also like to especially thank Dr. Robert Palmer, Mike Edwards, and David George who have helped me with numerous tasks throughout my time in the group, and has constantly been there to bounce ideas off. Thanks also to my friends, especially to Kai An, Dr. Jianhai Bao, Dr. Yunqiu Li, Dr. Manduhu, Panpan Ren, Chao Tong and Jinjin Xiong who have distracted me when things have got stressful, and always been there when I needed to vent.

Table of Contents

List of Publications x

List of Acronyms xii

List of Tables xiv

List of Figures xvi

1 Introduction 1

1.1 Motivation . . . 1

1.2 Overview and Contribution . . . 4

1.2.1 Adaptive Learning for Coronary Artery Segmentation . . . 4

1.2.2 Adaptive Learning for Aorta Segmentation . . . 5

1.2.3 Adaptive Learning for Face Detection . . . 7

1.3 Outline . . . 8

2 Background 10

2.1 Introduction . . . 10

2.2 Supervised Machine Learning . . . 11

2.2.1 Random Forests . . . 11

2.2.2 Neural Networks . . . 13

2.3 Medical Image Analysis . . . 14

2.3.1 Medical Image Segmentation . . . 17

2.3.2 Cardiovascular Anatomy . . . 22

2.3.3 Medical Imaging Techniques . . . 29

2.4 Object Detection . . . 33

2.4.1 Binary Detection . . . 34
2.4.2 Object Localisation . . . 36
2.5 Summary . . . 38

3 Coronary Artery Segmentation **40**
3.1 Introduction . . . 41
3.2 Proposed Method . . . 42
3.2.1 Vessel Enhancing Diffusion . . . 42
3.2.2 Multi-Scale Coronary Feature Extraction . . . 44
3.2.3 Voxel Classification using Random Forests . . . 46
3.2.4 MRF Regularization with Primal Dual Algorithm . . . 49
3.3 Experimental Result . . . 52
3.3.1 Segmentation Software . . . 52
3.3.2 Segmentation Result . . . 55
3.4 Summary . . . 57

4 Aorta Segmentation **65**
4.1 Introduction . . . 66
4.2 Proposed Method . . . 72
4.2.1 Overview . . . 72
4.2.2 Intensity-based Naive-Bayesian Detector . . . 73
4.2.3 Pseudo-3D CNN Detector . . . 74
4.2.4 Localised Interactive Refining . . . 78
4.2.5 Non-Uniform Implicit B-spline Surface . . . 79
4.2.6 Segmentation as Region-based Deformation . . . 84
4.3 Evaluation . . . 85
4.3.1 3D CTA Dataset . . . 85
4.3.2 Experimental Result . . . 86
4.3.3 Speed Discussion . . . 94
4.4 Conclusion . . . 94

5 Face Detection **96**
5.1 Introduction . . . 97
5.2 Related Work . . . 99

5.3 Soft Cascade . . . 101
5.3.1 Proposed Method . . . 102
5.3.2 Experiment and Discussion . . . 106
5.4 Detection-Regression Cascade . . . 114
5.4.1 Proposed Method . . . 114
5.4.2 Experiment and Discussion . . . 120
5.5 Summary . . . 127

6 Conclusion and Future Work **128**
6.1 Conclusion . . . 128
6.2 Future Work . . . 130

Bibliography **134**

List of Publications

The following is a list of published papers as a result of the work in this thesis, a conference paper is under review, and a journal paper is in preparation.

1. J.Deng, and X.Xie. Segmenting 3D Medical Image via Adaptive Learning and Deforming a Non-Uniform B-Spline Implicit Representation. To be submitted to IEEE Transaction on Image Processing (TIP), In Preparation.

2. E.Boileau, S.Pant, C.Roobottom, I.Sazonov, J.Deng, X.Xie, and P.Nithiarasu, Estimating the Accuracy of a Reduced-Order Model for the Calculation of Fractional Flow Reserve (FFR). International Journal for Numerical Methods in Biomedical Engineering, 2017

3. J.Deng, and X.Xie. Detect Face in the Wild using CNN-Cascade with Feature Aggregation at Multi-Resolution. IEEE International Conference on Image Processing (ICIP), 2017.

4. J.Deng, and X.Xie. Nested Shallow CNN-Cascade for Face Detection in the Wild. IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2017.

5. J.Deng, X.Xie, R.Alcock, and C.Roobottom. 3D Interactive Coronary Artery Segmentation using Random Forests and Markov Random Field Optimization. IEEE International Conference on Image Processing (ICIP), 2014.

In addition to the publications listed above, there are several relevant works on applying machine learning to both biomedical and natural image analysis problems, which were published during the author's Ph.D studying period.

1. B.Daubney, X.Xie, J.Deng, N.M.Parthalin, and Reyer Zwiggelaar. Fixing the Root Node: Efficient tracking and detection of 3D human pose through local solutions. Image and Vision Computing (IVC), 2016.

2. M.Edwards, J.Deng, and X.Xie. From Pose to Activity: Surveying Datasets and Introducing CONVERSE. Computer Vision and Image Understanding (CVIU), 2016.

3. J.Deng, X.Xie, and M.Edwards. Combining Stacked Denoising AutoEncoders and Random Forests for Face Detection. Advanced Concepts for Intelligent Vision Systems (ACIVS), 2016.

4. J.Deng, X.Xie, L.Terry, A.Wood, N.White, T.H.Margrain, and R.V.North. Age-related Macular Degeneration Detection and Stage Classification using Choroidal OCT images. International Conference on Image Analysis and Recognition (ICIAR), 2016.

5. J.Deng, X.Xie, and B.Daubney. A Bag of Words Approach to Subject Specific 3D Human Pose Interaction Classification with Random Decision Forests. Graphical Models (GMOD), 2014.

6. J.Deng, X.Xie, and S.Zhou. Conversational Interaction Recognition based on Bodily and Facial Movement. International Conference on Image Analysis and Recognition (ICIAR), 2014.

7. A.Lacey, J.Deng, and X.Xie. Protein Classification using Hidden Markov Models and Randomised Decision Trees. International Conference on BioMedical Engineering and Informatics (BMEI), 2014.

8. D.Chen, J.Deng, X.Xie, P. Nithiarasu, and D.Smith. Efficient Rconstruction of Coronary Vessels from 2D Angiography. International Conference on Computational and Mathematical Biomedical Engineering (CMBE), 2013.

9. J.Deng, X.Xie, B.Daubney, H.Fang, and P.Grant. Recognizing Conversational Interaction based on 3D Human Pose. Advanced Concepts for Intelligent Vision Systems (ACIVS), 2013.

10. H.Fang, J.Deng, X.Xie, and P.Grant. From Clamped Local Shape Models to Global Shape Model. IEEE International Conference on Image Processing (ICIP), 2013.

List of Acronyms

AdaBoost Adaptive Boosting

AFLW Annotated Facial Landmarks in the Wild

AI Artificial Intelligence

Back-Prop Backwards Propagation of Errors

CMU-MIT CMU-MIT Face Dataset

CNN Convolutional Neural Network

CT Computed Tomography

CTA Computed Tomography Angiography

DDFD Deep Dense Face Detector

DICOM Digital Imaging and Communications in Medicine

DNN Deep Neural Network

DPM Deformable Part Model

DT Decision Tree

EM Expectation Maximization

FCN Fully Convolutional Neural Network

FDDB Face Detection Dataset and Benchmark

FFR Fractional Flow Reserve

GAC Geodesic Active Contour

GENKI GENKI Database

GMM Gaussian Mixture Model

GPGPU General-Purpose Graphics Processing Unit

GVF Gradient Vector Flow

Haar Haar wavelets

HOG Histogram of Oriented Gradients

IBS Implicit B-spline Surface

IoM Intersection over Minimum

IoU Intersection over Union

LBP Local Binary Patterns

MAP Maximizing A Posteriori

MRF Markov Random Field

MPR Multi-Planar Reconstruction

MRFA Multi-Resolution Feature Aggregation

NMS Non-Maximal Suppression

NN Neural Network

NU-IBS Non-Uniform Implicit B-spline Surface

ODE Ordinary Differential Equation

PAVR Percutaneous Aortic Valve Replacement

PDE Partial Differential Equation

PIR Parametric Implicit Representation

R-CNN Regions with Convolutional Neural Network Features

RBF Radial Basis Function

ReLU Rectified Linear Unit

RF Random Forests

ROI Region of Interest

SIFT Scale-Invariant Feature Transform

SIMD Single Instruction Multiple Data

SGD Stochastic Gradient Descent

SPPnet Spatial Pyramid Pooling CNNs

SVM Support Vector Machine

TAVI Transcatheter Aortic Valve Implantation

TAVR Transcatheter Aortic Valve Replacement

TMT Tubularity Markov Tree

TP True Positive

WHO World Health Organization

List of Tables

3.1 The technical implementation detail of interactive segmentation software *SVMIST*. 53
3.2 The main functionalities of interactive segmentation software *SVMIST*. 58

4.1 caption 68
4.2 Parametric Implicit Representation for Surface Reconstruction and Image Segmentation 69
4.3 The parameter settings of the key components of proposed Pseudo-3D CNN network. 75
4.4 Quantitative classification results (TP: True Positive, FP: False Positive, in %) of each cascade stage and localised interactive refining. 89
4.5 Quantitative comparison of Uniform IBS and proposed Non-Uniform IBS. (BS: #B-splines; Sim: Similarity; Mut: Mutual Information; Hau: Hausdorff; Mah: Mahanabolis;) 90
4.6 Speed and Approximation Accuracy of proposed NU-IBS on a $256 \times 256 \times 200$ volume. (BS: #B-splines; Maxtrix: Basis Matrix Size; Dist: Signed Distance Transformation; Basis: Compute Basis Matrix; Chol: Cholesky Decomposition; Deform: 1-Iteration; Interp: Interpolation.) 94

5.1 The network architecture of *ElmNet* for fast elimination. 103
5.2 The network architecture of *LocNet* for precise localisation. 105
5.3 The network architecture of *DetNet* for face detector. 106
5.4 Recall rate and number of false positives of individual detection stage of the proposed method on FDDB dataset. 110
5.5 Speed of individual stage (hypotheses/second) 113
5.6 The network architecture of *ElmNet* with a batch normalisation and a drop-out regularisations for fast elimination. 116

5.7 The network architecture of multi-task *VefNet* for face detection and bounding box regression. . . . 118
5.8 Recall rate and number of false positives of individual detection stage of the proposed method on FDDB dataset. . . . 122
5.9 Comparison of detection rates (%) with both discrete and continuous metrics on Face Detection Dataset and Benchmark (FDDB). . . . 122
5.10 Speed (samples/second) and GPU load (usage percentage and memory consumption) of individual stages. . . . 125

List of Figures

2.1 An illustrative decision tree used to predict whether a photo represents an indoor or outdoor scene, adapted from [1]. 12
2.2 Left: The mathematical model of a single neuron. Right: Two-layer neural networks model. 14
2.3 Examples of medical image analysis techniques. (a) Image segmentation [2], (b) Image registration [3], (c) Image de-noising [4], (d) 3D reconstruction and rendering [5]. 17
2.4 Examples of active contour models. (a) Parametric active contour model [6], (b) Geodesic active contour model [7]. 21
2.5 The simplified human circulation system including both the cardiovascular system and the lymphatic system, adapted from [8]. Red indicates oxygenated blood carried in arteries, blue indicates de-oxygenated blood carried in veins. 23
2.6 The human heart viewed from the front (a), and behind (b), adapted from [9, 10]. . 24
2.7 (a) The heart, showing valves, arteries and veins, and the white arrows showing the normal direction of blood flow, adapted from [11]. (b) With the atria and major vessels removed, all four valves are clearly visible, adapted from [12]. 25
2.8 Illustration of the aorta root, adapted from [13]. 26
2.9 Illustration of the coronary arteries, adapted from [14]. 27
2.10 (a) The progression of atherosclerosis, adapted from [15]. (b) Illustration of coronary artery disease caused by narrowing of artery, adapted from [16]. 28
2.11 (a) Illustration of transfemoral and transapical approaches for transcatheter aortic valve replacement, adapted from [17]. (b) Illustration of angioplasty and coronary artery stent placement, adapted from [18]. 29

2.12 Examples of medical imaging techniques. First row left: intravascular ultrasound. First row middle: optical coherence tomography. First row right: phase contrast microscopy. Second row left: brain MRI. Second row right: computed tomography angiography. Third row left: brain PET [19]. Third row right: fetal ultrasound [20]. 31
2.13 Modern CT scanner, adapted from [21, 22]. 32
2.14 (a) Illustration of CTA images and 3D reconstruction of coronary artery, adapted from [23]. (b) Illustration of a CTA image of human torso, adapted from [24]. . . . 33
2.15 The Viola-Jones object detection framework, adapted from [25, 26]. 35
2.16 The HOG feature for human detection, adapted from [27]. 36
2.17 Deep learning methods in object localization. The network architecture of AlexNet (Top) [28], and Fast R-CNN (Bottom) [29]. 38

3.1 The pipeline of proposed method for interactive 3D coronary artery segmentation. . 43
3.2 The examples of coronary arteries in 3D CTA images. 45
3.3 The examples of other anatomical structures in 3D CTA images. 46
3.4 Random Forests is an ensemble classifier consisting of a set of Decision Tree (DT). 47
3.5 The plots of proposed non-linear transformation function $T_{\eta=1,2,3}$ compared to the linear function $Y = X$. 51
3.6 The graphical representation of MRF with 6 neighbourhood system, $< p, q_{1,\cdots,6} >$. 52
3.7 To create or open a DICOM database at a specific storage location. 54
3.8 An example of database configuration XML file of *SVMIST*. 55
3.9 The database management GUI of *SVMIST*. 56
3.10 The snapshots and meta information of DICOM images are loaded when a sequence is selected. 57
3.11 Investigate the volumetric image using orthogonal MPR. 58
3.12 Create curved MPR via clicking surface control points on a 2D viewer pane. 59
3.13 Visualize the curved MPR surface on the 3D viewer pane on the right, and the projection image is shown in the 2D viewer pane on the left. 60
3.14 The examples of user strokes. 60
3.15 An example of exported labelling XML file of *SVMIST*. 61
3.16 The out-of-bag error of different number of grown trees. 61
3.17 The iso-surface rendering of the Computed Tomography Angiography (CTA) image. 62

3.18 The examples of user provided strokes. (blue: background strokes; yellow: foreground strokes; dot: control points of user strokes.) 62
3.19 The examples of interactive segmentation process: (a) RF-based voxel classification result; (b) final segmentation result of the proposed method. 63
3.20 The examples of interactive segmentation results: (a), (c) and (e) iso-surfaces rendering of the original CTA images; (b), (d) and (f) final segmentation results of the proposed method. 64

4.1 The network architecture of pseudo-3D CNN detector which consists of four components as follows: (a) multi-scale pseudo-3D sampling, (b) primitive feature extraction, (c) feature aggregation and generalisation, and (d) foreground-background prediction. 76
4.2 (a.1) A 1D example of cubic B-splines basis functions that are made out of scaling and translating the uniform blending functions. (a.2) An example of unweighted uniform kernel functions. (b.1) A 1D example of cubic B-Spline basis functions that are made out of scaling and translating the non-uniform blending functions. (b.2) An example of unweighted non-uniform kernel functions. 80
4.3 (a) A 1D signal is constructed using a set of sine functions that have different frequencies. (b) The heat map of continuous wavelet coefficients of the signal given in (a). 83
4.4 An example of 3D CTA TAVI image from 3 orthorgonal views and surface rendering created using *3DimViewer* [30]. The images from the top to the bottom in the left column are axial view, coronal view and sagittal view respectively. The right column shows the mesh model of aorta root (top) and 3D surface rending of the volume (bottom), where the aorta root is highlighted with the organ circle. 87
4.5 The visualization of three Naive-Bayesian classifiers trained for different dataset folds. The blue and cyan curves are conditional probabilities of foreground and background that are modelled using GMMs. The red and black curves are posterior probabilities obtained through Bayesian rule given a pre-defined prior. 88
4.6 Qualitative comparisons of Uniform IBS (top row) and proposed NU-IBS (bottom row). The uniform method turns to smooth out the geometrical details of aorta valves that are well preserved by our method. 90

4.7 The deformation process at the 1st, 4th, 8th, 10th, 12th and 20th iterations using different step sizes τ. 91
4.8 Qualitative results of detection and segmentation from three fold testing at different stages. (a) Naive-Bayesian classification results. (b) Pseudo-3D CNN classification results. (c) The first round of localised interactive refining results. (d) The last round of localised interactive refining results. (e) The final segmentation results. Green, yellow, and blue correspond to true positive, false positive, and false negative respectively. 92
4.9 Additional qualitative results of detection and segmentation from three fold testing at different stages. (a) Naive-Bayesian classification results. (b) Pseudo-3D CNN classification results. (c) The first round of localised interactive refining results. (d) The last round of localised interactive refining results. (e) The final segmentation results. Green, yellow, and blue correspond to true positive, false positive, and false negative respectively. 93

5.1 The pipeline of the proposed nested cascade face detector. 102
5.2 Network architecture of *ElmNet*. 104
5.3 Network architecture of *LocNet*. 104
5.4 Network architecture of *DetNet*. 106
5.5 Examples training images. (a) Positive images are cropped face from AFLW dataset; (b) negative images are generated by replacing the face region with non-face patches sampled from PASCAL VOC datasets. 107
5.6 ROC curves of the proposed detector and recent methods on FDDB database with the discrete score metric.) 109
5.7 Typical detection results on FDDB dataset (red: ground truth, blue: true positive). 110
5.8 Examples of false positives and false negatives on FDDB dataset (red: ground truth, blue: true positive, yellow: false positive, green: false negatives). 111
5.9 Examples of correct detections but counted as false positives (red: ground truth, blue: true positive, yellow: false positive). 111
5.10 Examples of qualitative results on AFW dataset. (green: ground truth, blue: detection results of the proposed method). 112
5.11 Examples of qualitative results on CMU-MIT dataset. 113
5.12 Examples of qualitative results on GENKI database. 113

5.13 The pipeline of the proposed MRFA detector. 115
5.14 Network architecture of *ElmNet* with batch normalisation and drop-out regularisations. 116
5.15 Network architecture of *VefNet*. 119
5.16 ROC curves of the proposed detector and recent methods on FDDB with the discrete score metric.) 123
5.17 Typical detection results on FDDB dataset (red: ground truth, blue: true positive). 124
5.18 Examples of false positives and false negatives on FDDB dataset (red: ground truth, blue: true positive, yellow: false positive, green: false negatives). 124
5.19 Examples of correct detections but counted as false positives (red: ground truth, blue: true positive, yellow: false positive). 124
5.20 Examples of qualitative results on CMU-MIT Face Dataset (CMU-MIT) and GENKI Database (GENKI) datasets, showing in (a) and (b) respectively. 126

6.1 The new interface of *SVMIST* that uses a cross-platform GUI library, QT. 133

Chapter 1

Introduction

Contents

1.1 Motivation 1
1.2 Overview and Contribution 4
1.2.1 Adaptive Learning for Coronary Artery Segmentation 4
1.2.2 Adaptive Learning for Aorta Segmentation 5
1.2.3 Adaptive Learning for Face Detection 7
1.3 Outline 8

1.1 Motivation

Over the past two decades, Artificial Intelligence (AI) has achieved remarkable successes, especially the advances in statistical machine learning, and deep neural networks. The emergences of the pioneering works in statistical machine learning such as Support Vector Machine (SVM) [31], Random Forests (RF) [32], Adaptive Boosting (AdaBoost) [33], and the robust feature descriptors such as Haar wavelets (Haar) [25], Scale-Invariant Feature Transform (SIFT) [34], Histogram of Oriented Gradients (HOG) [27], mark the third renaissance of AI. They have shown significant performance boosts in almost all traditional visual recognition problems, such as semantic segmentation, face and pedestrian detection, and object recognition. Some of the success was due to increasing computer power and some was achieved by focusing on specific isolated problems and pursuing them with the highest standards of scientific accountability. It motivates the development of more challenging and realistic datasets

which also reveal the limitations of these statistical methods in generalization ability, and the difficulties of feature crafting. To fulfil the dream of human level intelligence that has captured the imagination of the world in the 1960s, there are many research scientists from both academia and industry who are contributing their efforts to this fascinating topic. It is worth noting that Geoffrey E. Hinton's works [35, 36] make the Deep Neural Network (DNN) regain the public sight. Especially, AlexNet [28] in visual object recognition achieved significantly higher accuracy compared to the traditional methods that rely on stronger statistical classifiers and discriminative hand-crafted features. Since then, DNNs are becoming more and more mainstream [37]. There are four of the key reasons that have contributed to the success of DNNs. First, DNNs are able to learn the visual features hierarchically via training in supervised fashion, which avoids hand-crafting features. Second, layer-wise unsupervised pre-training methods [36] were developed and have proved to be more efficient compared with random initialisation. Third, a large amount of labelled datasets [38, 39, 40] are vital important to the advance in supervised training. Moreover, advances in hardware makes both forward pass and backward propagation computationally efficient. Especially, General-Purpose Graphics Processing Unit (GPGPU) are well placed for learning deep neural network structures [41]. Many industrial products and services have been developing using the deep learning, e.g. Google AlphaGo [42], and Self-Driving Car [43].

Medical image analysis has a strong connection with machine learning techniques with respect to decision making, e.g. in computer-aided diagnosis, image segmentation, and lesion detection. Both traditional statistical methods and state-of-the-art deep models have been successfully applied to analyse medical images [44, 45], where most of the work is focused on developing a fully automatic method. However, there are still major challenges in machine learning based medical image analysis which also form the major motivations of this thesis.

- **The knowledge gap between computer scientist and radiologist.** The statistical methods heavily rely on the discriminative visual features which requires years of working experience on radiology and relevant fields in hospital. However, radiologists and medical experts in clinics who are professional in identifying these visual features have little experience on machine learning, which makes rather difficult to transform the expert knowledge into a practical and automatic system.

- **The deep learning methods partially solves the feature crafting and decision making issues, where a large amount of labelled data is required but is not readily avail-**

able in general. In order to train an efficient and accurate deep model, a large amount of supervision data is required to initialise and optimise millions or billions of parameters that the model usually contains. Due to the complexity and large variability of medical image, to obtain reliable training data, it normally requires to collect multiple annotations from different experts, and then perform cross validation on the collected labellings. However, the time of experienced radiologist and clinician who are able to spend on preparing the dataset is very much valuable and generally rare resource. Hence, in general, huge image datasets are not publicly available to the academia, although the situation is improving with the joint effort of computer scientists and clinicians.

- **Adapting an existed method to another anatomy structure could be difficult, and it is certainly a non-trivial task.** Medical image segmentation is divided into multiple sub-fields, a number of methods were proposed for a specific anatomy structure, where some of them are also incorporated with targeted prior knowledges. It is rather difficult to apply an anatomy specific method to another subject due to large variations of imaging and geometrical structure, and strong hypothesis constrains that are used. A generalizable method with uniform segmentation scheme is still the ultimate goal that researchers are going to purse. In addition, image segmentation and anatomy reconstruction are usually used for disease diagnosis and surgery planning. An accurate and robust model can greatly reduce the chance of false prediction and unexpected surgical accident.

We believe that adaptive learning in cooperation with the user interaction can be an effective approach for addressing these challenges. **Here, adaptive learning refers to a progressive learning process that gradually builds the model given sequential supervision from user interactions. The learning process could be either adaptive re-training for small scale models and datasets, or adaptive fine-tuning for medium-large scale.** We show that such a scheme can lead to methods of interactive image segmentation and that are efficient for medical image analysis. Studies on segmenting two anatomies in human circulation system, i.e. coronary artery and aorta, are presented in Chapter 3 and Chapter 4, respectively. **In this thesis, adaptive learning is also considered as a progressive learning process that gradually divides a big and difficult problem into a set of smaller but easier problems, where a final solution can be found via combining individual solvers consecutively.** In Chapter 5, we show that adaptive learning is feasible in solving generic computer vision problems as well, i.e. face detection in the wild.

1.2 Overview and Contribution

The aim of this work is to utilise adaptive learning schemes to investigate and analyse medical images and natural images, particularly for image segmentation, and object detection problems. In Chapter 3 and Chapter 4, we will present two image segmentation methods for coronary artery and aorta using random forests and Convolutional Neural Networks (CNNs), respectively. In Chapter 5, we will show the feasibility of combining cascade based adaptive learning with shallow CNNs for detecting human face in unconstrained environment. In the rest of this section, we will briefly introduce the works presented in individual chapters, where the connections to the motivations that we discussed in Section 1.1, and contributions of each proposed methods are also discussed.

1.2.1 Adaptive Learning for Coronary Artery Segmentation

Coronary artery segmentation plays a vital important role in coronary disease diagnosis and treatment. The coronary artery is a small tubular-like structure that rings the heart, which can be well presented and distinguished using vessel feature descriptor extracted from the second order derivative information, so-called Hessian matrix. However, there are two main difficulties for coronary artery segmentation using 3D CTA images. First, it is very difficult to label the coronary artery due to the small size and poor connectivity showing in the images. Second, there are many similar small vessels in the whole volume scans, such as blood vessels in the lung. In chapter 3, we present a machine learning based interactive coronary artery segmentation method, where a random forest is gradually built given the knowledge interactively obtained from user. We first apply vessel diffusion to reduce noise interference and enhance the tubular structures in the images. A few user strokes are required to specify region of interest and background. Various image features for detecting the coronary arteries are then extracted in a multi-scale fashion, and are fed into a random forest classifier, which assigns each voxel with probability values of being coronary artery and background. The final segmentation is carried out using a Markov Random Field (MRF) [46] based optimisation with Primal Dual algorithm [47], and followed by a connective component analysis as post processing to remove isolated, small regions to produce the segmented coronary arterial vessels. The contributions of this work are threefold.

- We present an interactive segmentation method for the coronary artery that requires lim-

ited user interference and achieves robust segmentation results. It transfers the knowledge from radiologists to a practical segmentation system through the foreground and background guiding strokes.

- The proposed feature descriptor constructed on the eigen system of Hessian matrix at multi-scales is efficient for distinguishing tubular-like, plate-like, and sphere-like structures in shape analysis. Meanwhile, we showcase how to apply spatial piecewise constant on the binary prediction using MRF.

- We develop an interactive medical image segmentation platform, *SVMIST*. It has following functionalities that can be reused and further developed into a practical system: image dataset management, 3D viewing, volumetric rendering, interactive labelling, and segmentation.

Research outcomes from this chapter, including methods and experimental results, are presented within the following publication.

- J.Deng, X.Xie, R.Alcock, and C.Roobottom. 3D Interactive Coronary Artery Segmentation using Random Forests and Markov Random Field Optimization. IEEE International Conference on Image Processing (ICIP), 2014.

- E.Boileau, S.Pant, C.Roobottom, I.Sazonov, J.Deng, X.Xie, and P.Nithiarasu, Estimating the Accuracy of a Reduced-Order Model for the Calculation of Fractional Flow Reserve (FFR). International Journal for Numerical Methods in Biomedical Engineering, 2017

1.2.2 Adaptive Learning for Aorta Segmentation

It is difficult to hand-craft discriminative features for complex anatomies. A typical example could be the aorta that contains an arch tube and a root with three leaflets. Often, large datasets are not readily available to train a deeper model for mining the discriminative features, which makes it even a harder problem to solve. In Chapter 4, we present a semi-automatic method for segmenting aorta that is achieved by a *Classification-Refining-Regularizing* procedure in an interactive manner as follows: (1) detect the object via voxel-wise region classification; (2) interactively refine the predicted region using adaptive learning scheme; (3) regularise the results with a piecewise constant constraint that uses an implicit non-uniform B-spline model

for shape representation. A 2-stage cascade detector is used to leverage the overall classification accuracy and speed efficiency, where a simple Naive-Bayesian model was trained based on the intensity information for fast background voxel elimination, and a stronger Pseudo-3D CNN multi-scales detector was built to precisely identify the foreground objects. In addition to fully automatic voxel classification, an interactive refining scheme is introduced to boost the detection accuracy further by utilising the information gained from user's interventions, in our case, the foreground and background guiding strokes. However, it is worth noting that voxel-wise object detection is not equivalent to binary segmentation, as it does not take any prior knowledge into consideration. For example, the piecewise constant that is commonly used in the deformable segmentation. The proposed method addresses this problem by introducing a Non-Uniform Implicit B-spline Surface (NU-IBS) model to represent shape geometry, where the regularisation constrain can then be imposed via region based deformation given the classification confidence of each voxel. Our contributions are fourfold.

- A cascade detector is proposed to efficiently delinerate the foreground objects and background regions which contains a intensity-based Naive-Bayesian classifier for fast elimination, and a pseudo-3D CNN classifier for precise classification. The representative features for region-based detection is automatically learnt in a supervised fashion, hence, no hand-crafting is needed.

- An adaptive learning and localised refining strategies are introduced which further improve the detection result and boosts the accuracy with help of user interventions that are taken on-the-fly. The foreground and background guiding strokes provided by user are used as supervised labelling to adaptively update the classifier, where the spatial information of the strokes is used to localise the regions that need to be refined.

- We propose a novel shape representation method which embeds the shape into the zero manifold of a level set function that is approximated using locally supported B-spline patches in parametric form. The geometrical complexity is estimated using the proposed wavelet-based filtering method, according to which the control knots are placed, therefore, it is able to adapt according to the local topology. This contribution is non-trivial because it bridges implicit representation and explicit representation, and allows topological changes and localised user interaction.

- We derive the formulation of region based deformation for proposed NU-IBS, which can be used to impose the piece-wise constant constrain to the detection result. A smooth interface is iteratively propagated according to the classification confidence, where both geometrical and characteristic homogeneity are co-optimised, and an optimal solution to the joint object function can be achieved simultaneously.

Research outcomes from this chapter, including methods and experimental results, are presented within the following publication.

- J.Deng and X.Xie. Segmenting 3D Medical Image via Adaptive Learning and Deforming a Non-Uniform B-Spline Implicit Representation. To be submitted to IEEE Transaction on Image Processing (TIP), In Preparation.

1.2.3 Adaptive Learning for Face Detection

Face detection in the wild is a challenging computer vision problem due to large variations and unpredictable ambiguities commonly existed in real world images. Whilst using hand-crafted features is generally problematic, introducing powerful but complex models is often computationally inefficient. Especially, some recent works on adapting pre-trained large scale recognition models to face detection problem often requires excessive resource expenditure. In Chapter 5, we first propose a nested CNN-cascade learning algorithm that uses shallow neural network architectures and allows efficient and progressive elimination of negative hypothesis from easy to hard via self-learning discriminative representations from coarse to fine scales. The face detection problem is considered as solving three sub-problems: eliminating easy background with a simple but fast model, then localising the face region with a soft-cascade, followed by precise detection and localisation by verifying retained regions with a deeper and stronger model. Furthermore, we investigate multi-resolution feature aggregation strategies for detecting face in cooperation with cascade CNNs. We show that such strategies can be integrated into the architecture design of CNN via average pooling and channel-wise feature concatenation. Two proposed methods are tested on several public benchmarks with across dataset evaluation. Both quantitative and qualitative results show promising performance improvements on detecting faces in unconstrained environment. The contributions of this work can be summarised as follows:

- We show that face detection in the wild can be solved by dividing difficult problem into several sub problems using cascade-based adaptive learning. Individual sub problem is solved using a shallow CNN, where the features are learnt automatically and tuned to be optimal for different stages.
- We propose an adaptive cascade schemes such that the depth of CNN model is progressively increased with the number of stages. As the binary classification problem in the later stage is significantly more difficult than earlier stages, hence stronger models are used adaptively.
- We investigate nested soft decision method, feature aggregation via average pooling and channel-wise feature concatenation, and multi-task training schemes for CNN-based cascade, which proved to be effective for boosting the detection accuracy.

Research outcomes from this chapter, including methods and experimental results, are presented within the following publications.

- J.Deng, and X.Xie. Nested Shallow CNN-Cascade for Face Detection in the Wild. IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2017.
- J.Deng, and X.Xie. Detect Face in the Wild using CNN-Cascade with Feature Aggregation at Multi-Resolution. IEEE International Conference on Image Processing (ICIP), 2017.

1.3 Outline

The rest of the thesis is organised as follows.

Chapter 2 - Background: provides the necessary background information that includes introduction to supervised machine learning and medical image analysis, anatomical structure of coronary artery and aorta, relevant diseases, and an overview of face detection methods.

Chapter 3 - Coronary Artery Segmentation: presents an interactive classification method for segmenting coronary artery, where the random forests classifier is built gradually

given a sequential foreground and background labels from user. MRF is used to regularise the binary decision. The proposed method requires limited user interference and achieves robust segmentation results.

Chapter 4 - Aorta Segmentation: presents an interactive refining method for segmenting aorta root and arch, where given the user interaction the CNN classifier is adaptively fine-tuning to refine the classification result. A novel parametric implicit shape representation method is proposed, and the data driven deformation is derived from region based level set Partial Differential Equation (PDE). Hence, the segmentation can be achieved by regularising the refined classification result using region based shape deformation. The method is evaluated on a CTA dataset that has 36 volumes.

Chapter 5 - Face Detection: presents two cascade CNN methods for detecting face in unconstrained environment. The key concept is to divide the a difficult problem into several sub problems that can be solved adaptively. The proposed methods are evaluated on public datasets, and comparative studies are also presented.

Chapter 6 - Conclusions and Future Work: concludes the thesis with discussions of the proposed methods and possible extensions.

Chapter 2

Background

Contents

2.1 Introduction **10**
2.2 Supervised Machine Learning **11**
2.2.1 Random Forests 11
2.2.2 Neural Networks 13
2.3 Medical Image Analysis **14**
2.3.1 Medical Image Segmentation 17
2.3.2 Cardiovascular Anatomy 22
2.3.3 Medical Imaging Techniques 29
2.4 Object Detection **33**
2.4.1 Binary Detection 34
2.4.2 Object Localisation 36
2.5 Summary **38**

2.1 Introduction

Over the course of this chapter, the background knowledges that are required for this thesis are provided. In Section 2.2, we first introduce the basic concept of two supervised machine learning techniques that are used in the thesis, random forests and neural networks. In Section 2.3, an overview of medical image analysis is given, which includes an introduction of

image segmentation method, the anatomical structures and relevant diseases of the cardiovascular system, and medical imaging techniques. Especially, we focus on aorta and coronary artery in Section 2.3.2.1 and Section 2.3.2.2, which form the major subjects of this thesis. We also investigate the application of adaptive learning scheme, more specifically cascade method, in face detection, where an overview of object detection methods is provided in Section 2.4. This chapter is closed with a summary in Section 2.5.

2.2 Supervised Machine Learning

Supervised learning is one of the most important type of machine learning methods that infers a function or model from labelled training data. The training data consist of a set of pair examples which contains an input feature, typically a vector and a target output value, also called the ground-truth. The process of inferring the function or model is known as the training or learning procedure. The inferred function and model can then be used for mapping new examples to the targets, which is called the testing or prediction procedure. An optimal scenario will allow for the algorithm to correctly determine the class labels or regression value for unseen instances. Supervised learning is closely related to computational statistics and has strong ties to mathematical optimisation, where a function or model can be found via minimising the prediction error given input features on training dataset. Numerous supervised machine learning algorithms have been proposed and widely used in variety of applications. Examples include linear discriminant analysis (LDA), support vector machines (SVM), decision trees, classification and regression trees (CART), random forests, adaptive boosting (AdaBoost), stacked auto-encoders, neural networks (NN) [48, 49, 50]. A brief introduction to random forests (RF) and convolutional neural networks (CNN) that are used in the thesis is provided in the rest of this section.

2.2.1 Random Forests

A decision tree [1] is a tree-like predictive model which maps observations to targets. Each leaf node contains a target value from training dataset, each non-leaf node contains a test unit of certain features, and branches represent conjunctions of features that lead from the root to leaf nodes. Decision tree, where the target variable is continuous is called regression tree, when the target variable takes a finite set of values is called classification trees. A tree can be grown by

splitting the training set into subsets based on a feature value test. This recursive partitioning process is repeated on each derived subset in a recursive and top-down manner. The so-called best split can be quantitatively measured using homogeneity difference between a parent node and its children nodes. Popular homogeneity metric are Gini impurity, information gain, and variance reduction. Fig. 2.1 demonstrates a visual example of random decision tree.

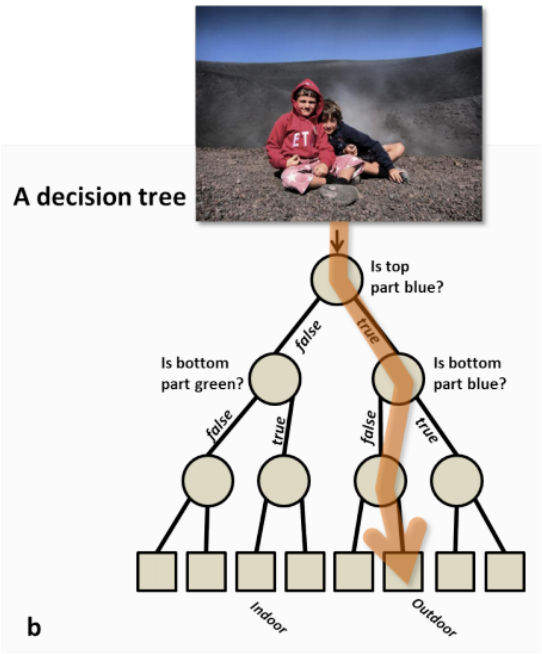


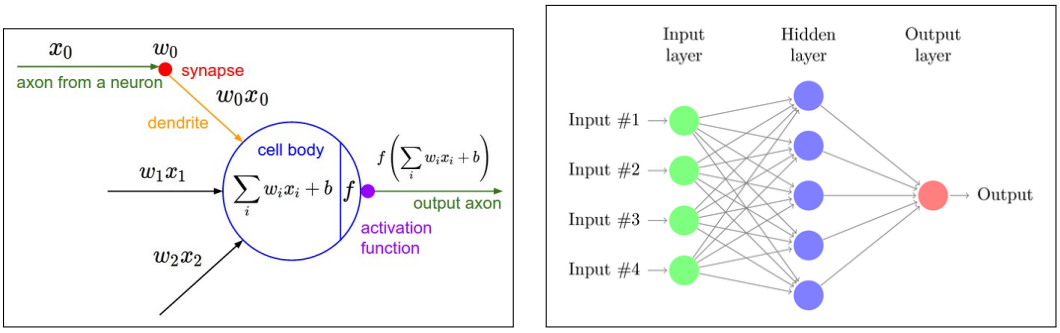
Figure 2.1: An illustrative decision tree used to predict whether a photo represents an indoor or outdoor scene, adapted from [1].

A RF [32, 1] is an ensemble classifier consisting of a set of decision trees, which significantly improves the generalisation ability of the classifier compared to a single decision tree. At the bootstrap aggregating stage (bagging), assuming that the data sample is independent and identically distributed, new training sets are generated by randomly sampling with replacement from the complete training set. For each new training set, one decision tree is constructed which consists of a set of split nodes and linking edges. Each non-leaf node stores a random test function which is applied to the input data, and leads to the leaf node. In the leaf nodes, the final predictor is stored. At the prediction stage, all the trees classify the incoming data independently, the most voted class given by the trees is considered as the final classification of the forest. Two parameters must therefore not be chosen determined for random forest training.

The first is the number of trees, and the other is the amount of weak classifiers allowed in the randomised subset to identify each non-leaf node. There are also some important parameters that control the depth of tree, and the minimum number of targets in the leaf node. Random forests are known to be accurate, and have reasonably good generalisation ability. However, due to it relying on weak classifiers, it is susceptible to over-fitting on noisy dataset, and heavily relies on hand-crafted discriminative features.

2.2.2 Neural Networks

Neural networks (NNs) are composed of layers where nodes in adjacent layers are linked together with weighted connections [51, 52]. Fig. 2.2 shows the mathematical model of neural networks with a single neuron on the left, and a two-layer neural network model. Mathematically, individual hidden node can be considered as an activation of linear combination of connected nodes from previous layer. A common choice of activation function is the sigmoid function, since it takes a real-valued input and squashes it to the range between 0 and 1. There are many alternatives proposed such as: Tanh, and ReLU family [53]. Neural networks are modelled as collections of neurons that are connected in an acyclic graph. The outputs of neurons in current layer can become inputs to other neurons in next layer. Cycles are strictly not allowed since that would imply an infinite loop in the forward pass of a network. Instead of an amorphous blobs of connected neurons, neural network models are often organised into distinct layers of neurons. For regular neural networks, the most common layer type is the fully-connected layer in which neurons between two adjacent layers are fully pairwise connected, but neurons within a single layer share no connections. A deeper model can be created via increasing the number of stacked layers in the network. Too few layers, and nodes in layers can lead to a network that is too generalised, whereas too many can easily lead to over-fitting the data. A balance between the two must be found depending on the data used in its application. Gradient descent optimisation is commonly used to train the neural networks which involves two consecutive stages. The parameters of the model are first initialised with small randomized real value. During the forward pass stage, the outputs are computed given the inputs of training samples layer by layer. Then, the outputs computed by the current model, and targets given by training dataset are used to compute the error loss that back-propagates to the network in an opposite direction. Once the analytic gradient is computed with back-propagation, the gradients are used to perform a parameter update.



(a) (b)

Figure 2.2: Left: The mathematical model of a single neuron. Right: Two-layer neural networks model.

CNNs [54] can be considered as extensions to Neural Networks (NNs). It makes the explicit assumption that the inputs are images, which allows us to force spatial connectivity on the node connection into the architecture using fixed size of local filtering kernels. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. Four main types of layers are used to build CNN architectures: convolutional layer, pooling layer, activation layer, and fully-connected layer. Convolutional layers create multiple feature maps, yielding a large number of parameters to be optimised, whereas pooling layers reduce the size of the feature maps with algorithms such as max, min or mean pooling. As a result, CNNs hierarchically create simpler, more generalised versions feature descriptor of the original image via stacking network building blocks. There are also many regularisation layers proposed to speed up training and avoid over-fitting issues to some extent, such as two popular options by using drop-out layer [55] and batch normalization layer [56]. The advantage of deep learning is its ability to learn suitable features. However, due to the large amount of parameters that need to be optimised, training deep learning algorithms is slow and require huge amount of labelled data. Furthermore, designing the architecture of network often requires trial and error which can be very time consuming.

2.3 Medical Image Analysis

Medical image analysis is an interdisciplinary field which focuses on applying the image processing, computer vision, machine learning, pattern recognition etc. techniques to signals and images that acquired for medical diagnosis and treatment purposes. It includes diverse research

topics which can be broadly grouped into five key directions from three different levels.

- **Image De-noising and Enhancement:** Image de-noising is a special case of noise reduction that is the process of removing noise from a signal [57, 58, 59]. An image is considered as a multi-dimensional signal where the spatial information is encoded into a regular grid. One goal in image de-noising is to remove the noise that recorded inevitably from the image in such a way that the so-called original image is restored. It is an important image processing topic, both as a inverse problem itself, and as a pre-processing component for other tasks. Many methods have been proposed such as chroma and luminance noise separation, linear smoothing filters, anisotropic diffusion, non-local means, non-linear filters, wavelet based approaches, and statistical methods. Sometimes the image de-noising is considered as a sub-topic of image enhancement which is defined as the process of adjusting digital images such that the results are more suitable for further image analysis. In addition to de-noising, it also includes image sharpening, feature highlighting and so on. For example, in Chapter 3, an anisotropic diffusion method is used to enhance the vessel-like structures.

- **Segmentation:** Image segmentation can be largely considered as the process of partitioning an image into multiple regions that have similar semantic and/or geometrical properties in terms of image appearance [60, 61]. The partitioning process generally relies on shape edges, homogeneous texture, and similar feature patterns that are able to differentiate the target object from the background. Hence, segmentation methods can also be categorised into three categories in terms of the image information used, as follows: edge-based segmentation, region-based segmentation, and feature-based segmentation. Considering the object functions that are constructed, there are continuous optimisation based segmentation that use PDE and variational methods, and discrete optimisation based methods that use graph partitioning methods. Since the majority of this thesis will investigate the application of adaptive learning in image segmentation problem, we provide a detailed overview of the segmentation methods in the next subsection.

- **Registration:** In medical diagnosis scenarios, imaging scans are generally collected over time to analyse pathological changes over a certain period, where different imaging techniques may be used, and the images with diverse modalities may be obtained. Image registration is the process of transforming different sets of image data into one

consistent coordinate system, such that comparison and integration studies can be carried out [62, 63]. An image is selected as the reference or source, and the other images are spatially transformed to align with the reference image using either rigid estimation or non-rigid estimation. In terms of the information that is used to estimate the transformation, there are two broadly defined categories, intensity based registration and feature based registration. Intensity based methods match intensity patterns between reference and target images via correlation metrics, while feature based methods establish a correspondence between a number of especially distinct points based on the feature similarity.

- **3D Reconstruction:** 3D reconstruction is the process of recovering the three-dimensional structure from a set of images. Understanding the geometrical structure of the target organ and tissue is very important for disease diagnosis, pre-surgery planning, and computer assisted surgery. The spatial information is one of the keys for 3D reconstruction which can be estimated using the geometrical relationship of multiple calibrated cameras, or the imaging parameters of a volumetric scanner [64]. 3D model of target object is created and visualised which can then be used for realistically investigating its geometrical structure and shape property. It may also involve some other image processing steps, such as image segmentation to obtain the target object, and image registration to correct the miss-alignment during volumetric scanning procedure.

- **Recognition:** Visual recognition is a high level image understanding task that normally involves feature extraction and decision making. The features can be extracted from different sources, such as low-level image based textural information [65], middle-level geometrical shape information, and so on. Hence, image de-noising and enhancement, segmentation, and registration can be considered as the pre-processes for visual recognition task. Machine learning methods, particularly supervised approached are widely used for modelling the common patterns from extracted features, and making decision on disease prediction [44, 66]. It is noteworthy that recently deep learning is becoming more and more main stream, as it shows superior performance boost in visual recognition task on both natural images and also on medical applications [45].

Fig. 2.3 shows some examples of medical image analysis techniques. In the rest of this section, we will first overview the image segmentation techniques especially in deformable model and graph cuts method. Then, we will discuss the anatomical structures and related

diseases of the subjects that we are going to segment in this thesis. In addition, the background of corresponding medical imaging techniques will also be provided.

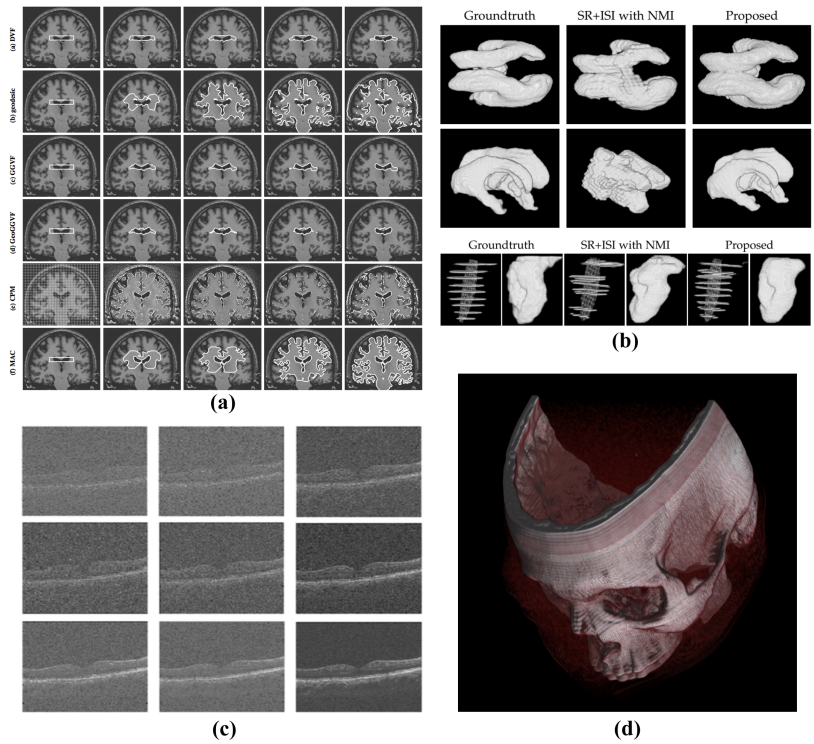


Figure 2.3: Examples of medical image analysis techniques. (a) Image segmentation [2], (b) Image registration [3], (c) Image de-noising [4], (d) 3D reconstruction and rendering [5].

2.3.1 Medical Image Segmentation

Image segmentation is a fundamental problem in computer vision, image processing, and medical image analysis, which has been widely used in motion tracking, 3D object reconstruction, especially in medical diagnosis and treatment. It separates the image into several disjointed parts that have semantic meaning of its own. Segmentation can generally be divided into three categories as follows: manual, automatic, and semi-automatic methods. Manual segmentation is generally carried out by experienced medical experts such as doctors or radiologists, which results in higher accuracy of delineating the Region of Interest (ROI) provided tools are sufficiently sophisticated and user-friendly. However, manual segmentation methods have a number of disadvantages. The most obvious one is the amount of time and effort it takes to

manually label a series of images. Fully-automatic segmentation methods can partially avoid such disadvantages to free up the time of doctors or radiologists. In addition, the segmentation will no longer be subjective, as an automatic segmentation method is a generally deterministic process. However, developing an efficient automatic method is a non-trivial task, and has many challenges. First, due to the large variation in terms of different anatomic structure, image modality, and quality of data, it is difficult to design a general automatic segmentation method to solve this problem in one go. Second, to transfer the knowledge of experienced medical experts that are accumulated during their whole career is extremely challenging, not only on developing the efficient models, but also on preparing huge amount of well formatted medical records. The last but not the least, the segmentation quality of automatic method is tightly bounded by the generalisation ability of the model that is used. Case-by-case variations are very common for medical images which may not be learnt by the pre-trained model, and result in incorrect prediction.

Semi-automatic methods, or so-called interactive segmentation, have been proposed to avoid these issues via constructing a controllable automatic scheme, where the user interaction is used to guide the segmentation towards the optimal target while minimising the user effort. An overview of interactive segmentation methods in medical image can be found in [67]. The user interaction can be applied to the process of segmentation via three different ways. First, the user provided information is used to build or correct a delineation model for separating foreground objects and backgrounds, where the statistical model is normally used as a classifier. Second, the user interaction is used as an initialisation segmentation or seed for some segmentation methods. The former is very common for deformable models, while the later is for region growing based methods. In the third interaction scheme, the user is asked to manually correct the segmentation result that is produced by an automatic method. All these interaction segmentation methods can also be used jointly in order to achieve the best performance. However, it is worth noting that the interaction schemes are auxiliary which initialise, guide, and correct the automatic method.

2.3.1.1 Active Contour

Active contour model [68] is considered as a pioneering works in deformable model which formulates image segmentation as an energy minimising problem, where deformable spline is influenced by smoothness constraint and image forces that pull it towards object contours and

internal forces that resist deformation. Active contours are parametric curves where one tries to fit to an image, usually to the edges within an image. Active contours may be understood as a special case of the general technique of matching a deformable model to an image by means of energy minimisation [69, 70]. The energy functional is given by:

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{snake}(\mathbf{V}(s))ds \\ &= \int_0^1 (E_{internal}(\mathbf{V}(s)) + E_{image}(\mathbf{V}(s)) + E_{con}(\mathbf{V}(s)))ds \end{aligned} \tag{2.1}$$

A parametric snake is defined by a set of n points $\mathbf{v}_i$, the internal elastic energy term $E_{internal}$, and the external edge-based energy term $E_{external}$. The purpose of the internal energy term is to control the deformations made to the snake, and the purpose of the external energy term is to control the fitting of the contour onto the image. The external energy is usually a combination of the forces due to the image itself E_{image} and the constraint forces introduced by the user E_{con}. The internal energy of the snake is composed of the continuity of the contour E_{cont} and the smoothness of the contour E_{curv}, which is meant to enforce smoothness of the parametric curve. Energy in the image is some function of the features of the image. This is one of the most common points of modification in derivative methods. Features in images and images themselves can be processed in many and various ways. For example, lines, edges, and terminations present in the image can be used to construct the image forces.

$$E_{image} = w_{line}E_{line} + w_{edge}E_{edge} + w_{term}E_{term} \tag{2.2}$$

where $w_{line}, w_{edge}, w_{term}$ are weights of these salient features. Higher weights indicate that the salient feature will have a larger contribution to the image force. Some systems, including the original snakes implementation, allowed for user interaction to guide the snakes, not only in initial placement but also in their energy terms. Such constraint energy E_{con} can be used to interactively guide the snakes towards or away from particular features. Given an initial guess for a snake, the energy function of the snake is iteratively minimized. Gradient descent minimization is one of the simplest optimizations which can be used to minimize snake energy. Each iteration takes one step in the negative gradient of the point with controlled step size to find local minima.

The original model is due to Kaas et al. [68], but many modifications have been proposed in the literature, though with their own trade-offs. First, one much discussed point on active contour is their inability to move into concavities of an objects boundary and their inability

to find the borders when it is initialised too far distant from the actual border location. The Gradient Vector Flow (GVF) snake model [71] addresses these two issues via introducing the so-called GVF field that is constructed on the diffused edge map to replace the external image force. The primary issue with using GVF is the smoothing term causes rounding of the edges of the contour. Reducing the value of smoothing term reduces the rounding but weakens the amount of smoothing. Second, the parametric curve has less topological flexibility, it is not able to break or emerge with regards to the image data naturally. Geodesic Active Contour (GAC) [72] employs ideas from Euclidean curve shortening evolution, where the contours split and merge depending on the detection of objects in the image. These models are largely inspired by level sets methods, and have been extensively employed in medical image segmentation. Level sets implicitly define lower dimensional structures such as surfaces via a function embedding. The 2 dimensional case is given by

$$\begin{aligned} \Gamma &= \{(x,y)|\phi(x,y)=0\} \\ \frac{\partial \phi}{\partial t} &= v|\nabla \phi| \end{aligned} \tag{2.3}$$

The curve evaluation is equivalent to solve a time dependant partial differential equation, in particular a Hamilton-Jacobi equation, which can be solved numerically, for example by using finite differences on a Cartesian grid. Hence, geodesic active contours have the advantage that they can change topology during evolution, and the result has proven to be very useful because it combines the intuitive active contours concepts with efficient implementations of level set methods. More recent developments in active contours address modeling of regional properties, incorporation of flexible shape prior and fully automatic segmentation. Fig. 2.4 (a) and (b) show two examples of active contour models with a parametric representation and an implicit representation respectively.

2.3.1.2 Graph Cut

Graph cut is a generic method for minimising a particular form of energy the so-called MRF energy [73]. As applied in the field of computer vision, graph cuts can be employed to efficiently solve a wide variety of low-level computer vision problems, such as background removal, image smoothing, the stereo correspondence problem, and many other computer vision problems that can be formulated in terms of energy minimisation. Such energy minimization problems can be reduced to instances of the maximum flow problem in a graph. Under most formulations

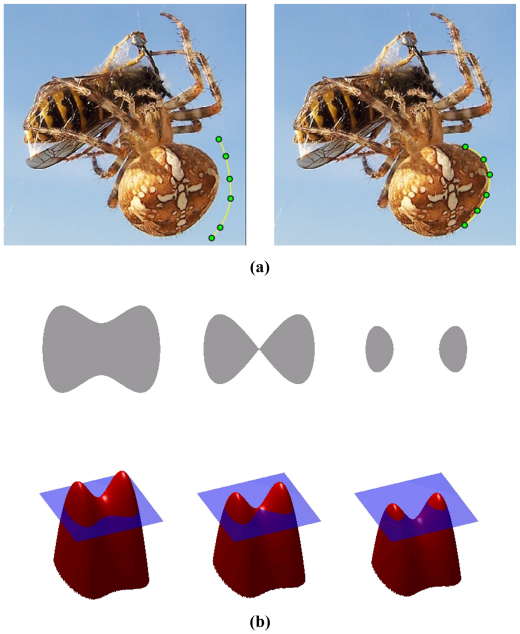


Figure 2.4: Examples of active contour models. (a) Parametric active contour model [6], (b) Geodesic active contour model [7].

of such problems in computer vision, the minimum energy solution corresponds to the maximum a posteriori estimate of a solution. Although many computer vision algorithms involve cutting a graph, the term "graph cuts" is applied specifically to those models which employ a max-flow/min-cut optimisation, while other graph cutting algorithms may be considered as graph partitioning algorithms. The solution of the graph cut is globally optimal with respect to a cost function which has a general form given by

$$\mathscr{E}(f) = C_{data}(f) + C_{coherence}(f) \tag{2.4}$$

where the C_{data} is the cost of data fitness, and $C_{coherent}$ is the cost of piecewise constant of pre-defined local neighbourhood. It has similar energy minimisation principle to active contour models, and has been widely applied to image segmentation. It sometimes outperforms the level set method when the model is MRF or can be approximated by MRF, as a global optimiser is often able to be found [47]. However, it also suffers from some significant limitations. First, when an image is represented by a 4-connected lattice, graph cuts methods can exhibit unwanted "blockiness" artifacts. Various methods have been proposed for addressing this issue, such as using additional edges [74] or by formulating the max-flow problem in

continuous space [75]. Second, the algorithm can be biased toward producing a small contour. Third, graph cuts is only able to find a global optimum for binary labeling problems, such as foreground-background image segmentation. Extensions have been proposed that can find approximate solutions for multi-label graph cuts problems [76]. Furthermore, graph cut segmentation is known to require high storage memory, and is computationally inefficient.

2.3.2 Cardiovascular Anatomy

In the human body, every organ, tissue and cell consumes the nutrients and oxygen, and then produces waste and carbon dioxide. The exchange processes cannot work without the circulatory system which consists of two separate subsystems, the cardiovascular system and the lymphatic system shown in Fig. 2.5. The cardiovascular system enables the blood to circulate and transport the essentials to every corner of human body, and expel metabolic waste. The lymphatic system circulates lymph which is also a vital part of immune system, where the fluids and immunological cells are transported from and to the blood and interstitial spaces. For most vertebrates including human, the cardiovascular system is a closed circulation, while the lymphatic system is an open system.

The human cardiovascular system consists of three essential parts as follows: the heart, blood and blood vessels, which form the pulmonary circulation and the systemic circulation. In a complete pulmonary cycle, the de-oxygenated blood is pumped away from the heart to the lungs, where carbon dioxide is released and oxygen is picked up during respiration, and then the oxygenated blood returns back to the heart. The systemic circulation, also known as the bronchial circulation, supplies nutrients and oxygen to the cells except the lungs, while carrying metabolic waste products away. It can further be divided into a macro-circulation and a micro-circulation, where the former circulate the blood from and to organ, the latter is the circulation of the blood in the smallest blood vessels to the tissues. The whole cardiovascular circulation system can be simply thought of as the heart works as a blood pump to which the blood vessels are connected, while the useful materials and produced waste are exchanged, although the reality is far more complex. Fig. 2.6 shows the superficial heart anatomy from both anterior (a) and posterior (b) views. In Fig. 2.6, the red and the blue blood vessels indicate the oxygenated and de-oxygenated blood that they are carrying respectively. The heart is a special muscle organ with four cavities inside. During the whole of human life, it continuously performs contraction and relaxation operations which results in blood flowing within the blood

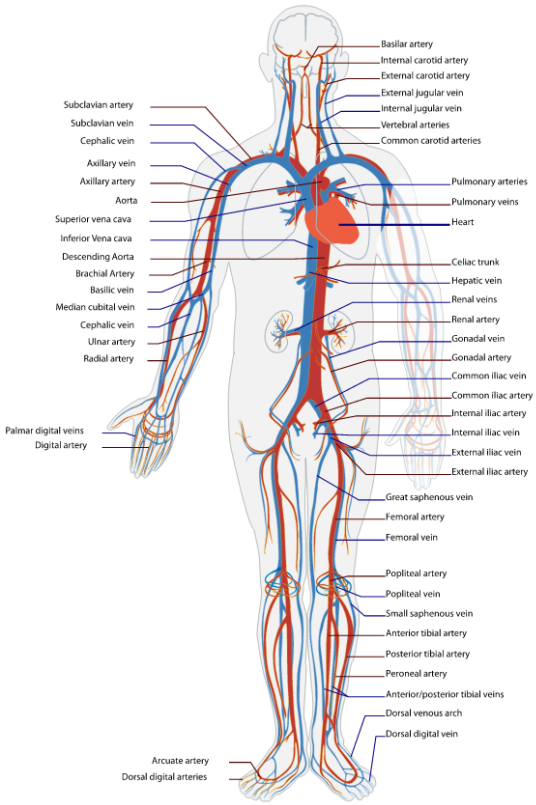
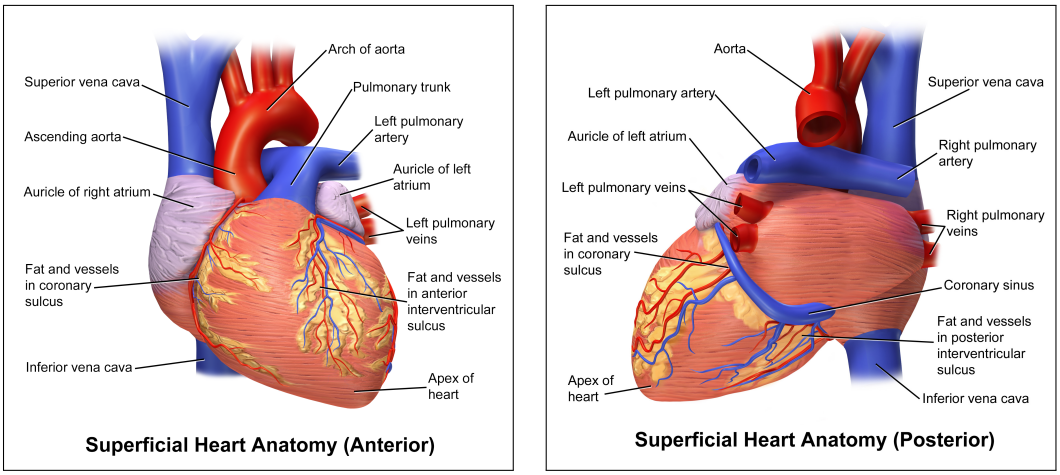


Figure 2.5: The simplified human circulation system including both the cardiovascular system and the lymphatic system, adapted from [8]. Red indicates oxygenated blood carried in arteries, blue indicates de-oxygenated blood carried in veins.

vessels. In a normal cycle, the contraction lasts 0.1 second, and the relation lasts 0.7 second. For an adult, the weight of the heart is generally 0.5% of his weight, whereas during every cycle of contraction and relaxation the heart can pump out or drain in 70ml blood on average resulting in 5L per minute.

As shown in Fig. 2.7 (a), the heart has four chambers, the left and the right atria on the top, and the left and the right ventricles at the bottom. The atria are receiving chambers where the blood flow is drained into the heart, while the ventricles are discharging chambers where the blood flow is pumped out of the heart. When the heart contracts, the oxygenated blood in the left ventricle passes through the aorta, and is transported to all parts of body except the lungs. At the same time, the de-oxygenated blood in the right ventricle passes through the pulmonary artery, and is transported to the lungs. When the heart relaxes, the blood carrying the oxygen is drained from the lungs through the pulmonary vein back to the left atrium, while the blood

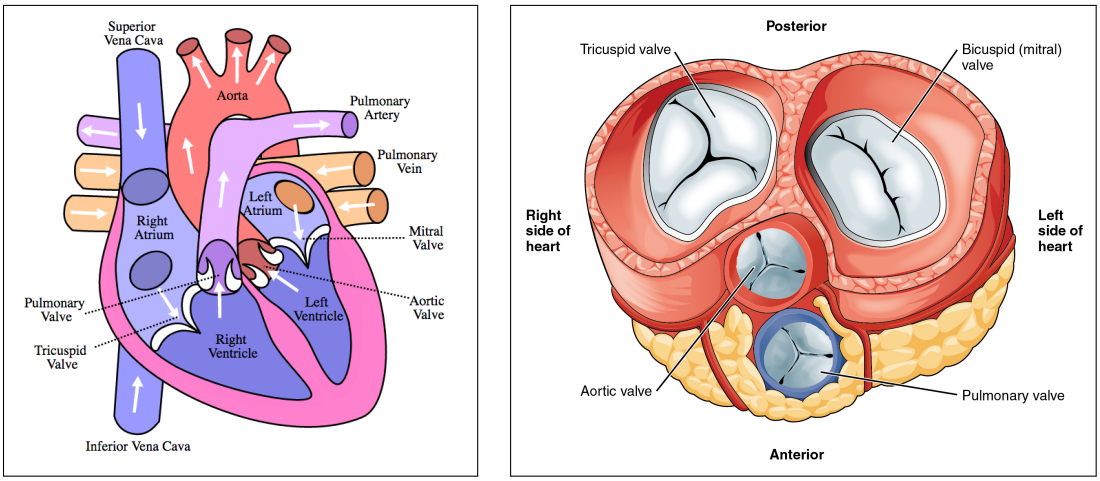


(a) (b)

Figure 2.6: The human heart viewed from the front (a), and behind (b), adapted from [9, 10].

carrying carbon dioxide is drained from the rest of the body through the vena cava back to the right atrium. The blood flows are one-way that can only go out from the ventricle and back to the atrium. In order to prevent refluxing, there are four valves separating the chambers, where one valve lies between each atrium and ventricle, and one valve rests at the exit of each ventricle. Fig. 2.7 (b) shows all four interior valves by removing the atria and major vessels. The tricuspid valve and the bicuspid valve lie between the right and left atria and ventricles respectively, where the valves open to enable the blood flows from atria to ventricles when the heart contracts, and close to prevent the back flow when the heart relaxes. The aortic valve and the pulmonary valve sit at the exit of each of the ventricles, where the valves open to enable the blood flows from ventricles to the main vessels when the heart contracts, and close when the heart relaxes. The motions of four valves are well synchronised to ensure the right blood circulation flow.

In this thesis, we will focus on the aorta root which is the start point of the systemic circulation, and the coronary artery which provides the blood circulation to the heart muscles. In the following two sub-sections, we will discuss their anatomical structures, functionalities, and the relevant diseases in cardiovascular circulation in more detail.



(a) (b)

Figure 2.7: (a) The heart, showing valves, arteries and veins, and the white arrows showing the normal direction of blood flow, adapted from [11]. (b) With the atria and major vessels removed, all four valves are clearly visible, adapted from [12].

2.3.2.1 Aorta Root and Arch

The aorta is the large blood vessel that carries oxygen-rich blood from the left ventricle of the heart to other parts of the body, where its root attaches to the heart. The aortic root consists of three aortic valve leaflets and the coronary ostia which are the openings for the coronary arteries. Fig. 2.8 shows the anatomical structure of the aorta root. The valve leaflets open to allow the blood flow into the ascending aorta. The coronary ostia consists of two main vessels that are just above the valve leaflets, and attached to the ascending aorta. The ascending and descending aortas form an arch shape, where there are three main arteries, the innominate artery, the left common carotid artery, and the left subclavian artery. The innominate artery is generally larger than the other two, and divided into the right common carotid artery and the right subclavian artery. The carotid arteries provide blood to the brain, and the subclavian arteries provide blood to the upper limbs. The blood is transported to the upper limbs and the other organs through descending aorta.

2.3.2.2 Coronary Artery

The cardiac muscle, also known as myocardium, produces powerful pressure to force the blood flowing within the blood vessels via periodical contraction and relaxation. The coronary ar-

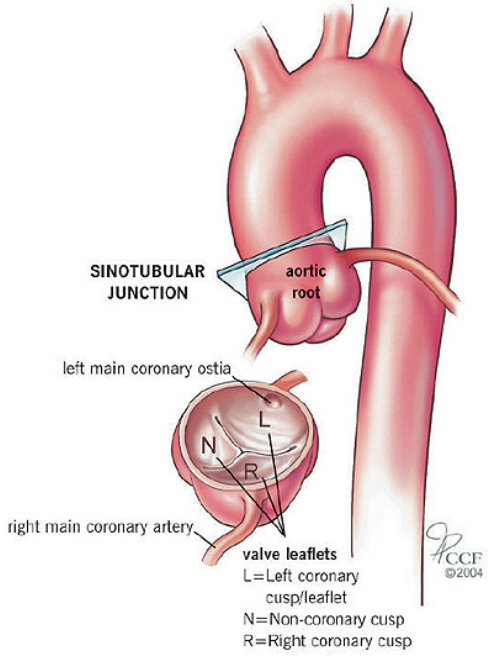


Figure 2.8: Illustration of the aorta root, adapted from [13].

teries are the blood vessels that are in charge of continuously transporting the blood with rich oxygen and rich nutrient to the myocardium. Fig. 2.9 shows that the left and the right main coronary arteries branch off the aorta root, and further subdivide into smaller branches that ring the heart. The right branches of the coronary artery transports the blood to the right atrium and ventricles and the atrioventricular septum. The left branches of the coronary artery provide blood supply to the rest of the left atrium and ventricle as well as the ventricular septum. Although the coronary arteries are generally narrow compared to other major vessels, the density of the blood capillary on myocardium is very high, 2,500 vessels per mm^2 on average. For a healthy adult, the total amount of blood flow in the coronary arteries reaches 225ml per minute that is about 5% of total amount of blood flow of the whole heart. The heart never stop beating during the whole life, it is the most efficient and busy organ, where the myocardium consumes 65% to 70% of oxygen from the blood that flows through to enable uninterruptible blood supply.

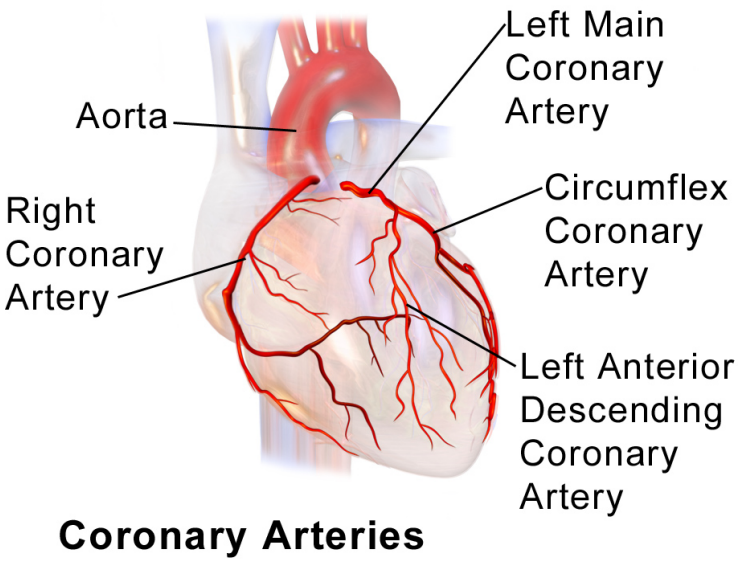
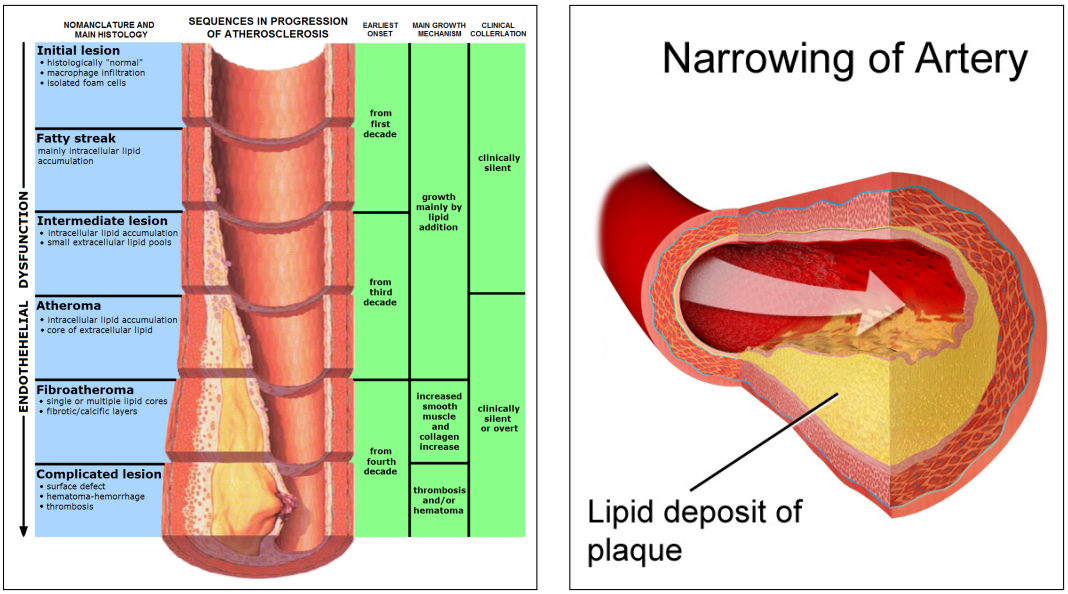


Figure 2.9: Illustration of the coronary arteries, adapted from [14].

2.3.2.3 Cardiovascular Disease

Cardiovascular diseases are the diseases of the heart and blood vessels, such as stroke, heart failure, cardiomyopathy, valvular heart disease, peripheral artery disease, and venous thrombosis. According to the report published by World Health Organization (WHO), cardiovascular diseases are the leading cause of death in the world [77]. There were 12.3 million death due to cardiovascular disease in 1990, with the number increasing to 17.3 million in 2013 [78]. An important and most hazardous category of cardiovascular diseases is so called stenosis, which is an abnormal narrowing in a blood vessel. Aortic stenosis is usually a result of calcium or plaque deposited in the artery which narrows the valve, meaning the aortic valve cannot fully open. This leads to decreasing blood flow from the heart to the body. In turn, it can lead to severe hypertension and angina. Coronary stenosis reduces or blocks the oxygen-rich blood flowing to myocardium, which is the leading cause of ischaemia and myocardium damage. Artery wall thickening diseases are specific forms of arteriosclerosis, also known as atherosclerosis. They are the results of an accumulation of atheromatous and fibrofatty plaques inside of blood vesseles that are generally produced by white blood cells and the proliferation of intimal-smooth-muscle cells. Fig. 2.10 shows the progression of atherosclerosis.

Arteriosclerotic vascular diseases including both aorta stenosis and coronary stenosis result

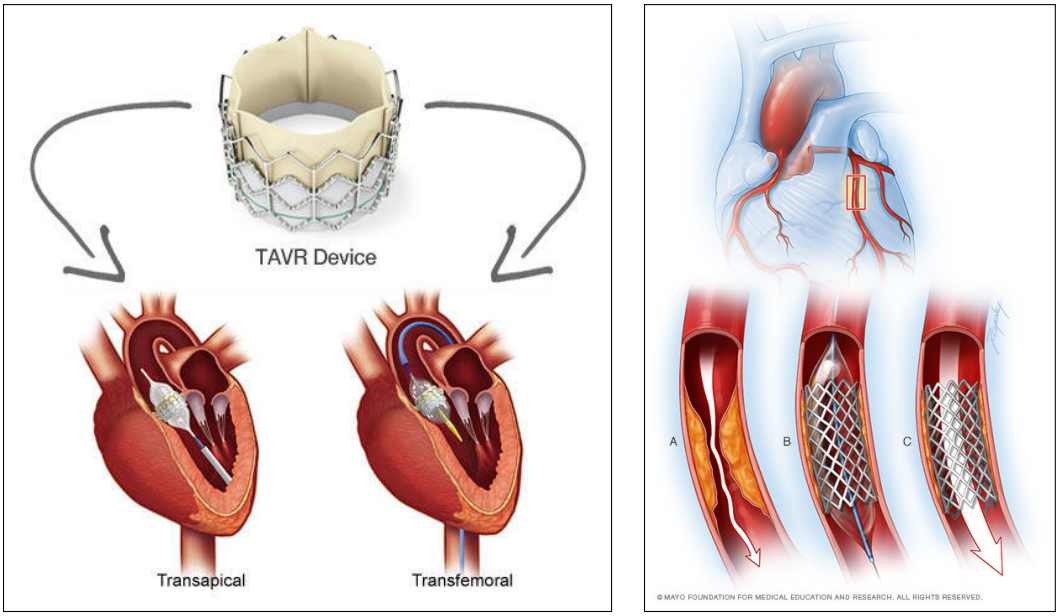


(a) (b)

Figure 2.10: (a) The progression of atherosclerosis, adapted from [15]. (b) Illustration of coronary artery disease caused by narrowing of artery, adapted from [16].

in occlusion of any of these vessels that can interrupt the blood supply to the body and the heart. Left without treatment, severe stenosis will lead to functional deterioration, heart failure, and often death. Open heart surgery requires cutting the chest open, and it is performed on the muscles, valves, or arteries of the heart, which are high risk operations to the life of patients especially for elder patient. Nowadays, for the patients who are not well enough to have open heart surgeries, transcatheter based surgeries are normally given instead. For aorta stenosis, Percutaneous Aortic Valve Replacement (PAVR), also known as Transcatheter Aortic Valve Implantation (TAVI) or Transcatheter Aortic Valve Replacement (TAVR), is the surgery of replacing of the aortic valve of the heart using transcatheter through the blood vessels. For coronary stenosis, a coronary angioplasty is given to widen blocked or narrowed coronary arteries. Fig. 2.11 (a) shows transfemoral and transapical approaches for transcatheter aortic valve replacement, and Fig. 2.11 (b) illustrates angioplasty and coronary artery stent placement approaches. The valve replacement device and the artery stents are transported to the heart via the catheter from one of the large vessels in the limbs, and then are positioned in the right locations. In both cases, the path and the geometrical structure that the catheter takes must be assessed. In chapter 3 and 4 of this thesis, we will present two interactive segmentation

methods that are used to analyse the geometrical structure of the coronary artery and aorta root respectively.



(a) (b)

Figure 2.11: (a) Illustration of transfemoral and transapical approaches for transcatheter aortic valve replacement, adapted from [17]. (b) Illustration of angioplasty and coronary artery stent placement, adapted from [18].

2.3.3 Medical Imaging Techniques

Medical imaging is the technique and process of revealing internal structures of human body for clinical analysis and medical intervention. Since 1895 when the German physicist, Wilhelm Conrad Rontgen discovered that X-rays can identify bone structures, the terminology "imaging" has developed into much wider sense instead of producing signal in an image form only. In order to create visual representations of the interior of a body, many physical techniques are used, such as medical radiography, ultrasound, magnetic resonance, radionuclide, optical laser, thermography, fluoroscopy, and so on. Those medical imaging techniques can be further divided into two categories, invasive and non-invasive based on whether instruments are introduced into a patient's body or not. In an unrestricted sense, the process of producing visual representations of the interior of a body is the solution of a mathematical inverse problem given the observed signals. For example, X-ray radiation imaging technique is based on the fact of

that it is absorbed at different rates by different tissue types such as bone, muscle and fat, where the interior structures can be inferred via measuring the signal difference between incidence and emergence. Taking image scans for patient is an essential procedure for diagnosis, treatment and surgery planning. Fig. 2.12 shows some examples of acquired medical images. In the next two sub-sections, we will introduce two special medical imaging techniques that are used commonly for analysing and treating cardiovascular diseases.

2.3.3.1 Computed Tomography and Angiography

Computed Tomography (CT) is nowadays a common medical imaging technique. It produces cross-sectional (tomographic) images of a scanned object from many X-ray images taken from different angles. The process of computing tomographic images is so-called digital geometry processing, where a series of two-dimensional radiographic images are taken via rotating a single axis, and then combined into a single cross-sectional image. Fig. 2.13 shows a modern 3D CT scanner, where a series of 2D X-ray images are generated via rotating the X-ray source. The patient lays on the motorised table, and is moved through the target locations in the scanner progressively. By stacking all cross-sectional images, the CT scanner produce a volumetric data that can be reformatted in various reconstruction planes for better representations of target structures.

Generally an invasive catheter coronary angiography is given to patients in order to detect narrowing of blood vessels, especially for those who have coronary artery diseases. However, such invasive imaging techniques could be risky and painful to the patients. CTA is a computed tomography technique used to visualize arterial and venous vessels using non-invasive CT scanning technique. As blood has very low capacity of absorbing X-ray radiation, it is difficult for the normal CT imaging to differentiate tissue and blood vessels. In order to reveal the internal geometrical structure of vessels, radiocontrast agents are injected into the human body or consumed by patients to increase the visibility of blood under X-ray radiation. The typical radiocontrast agents are iodine or barium compounds that lack therapeutic effects for patients. Before the radiocontrast agents are metabolised, the patient is given a normal CT scan procedure, hence the produced tomography image is able to reflect blood flow inside of body as well as the geometrical structure of blood vessel. CTA are used to examine blood vessels in many vital regions of the body, such as the brain, kidneys, the heart, and lungs. Fig. 2.14 (a) shows typical examples of CTA images of coronary artery, and the corresponding recon-

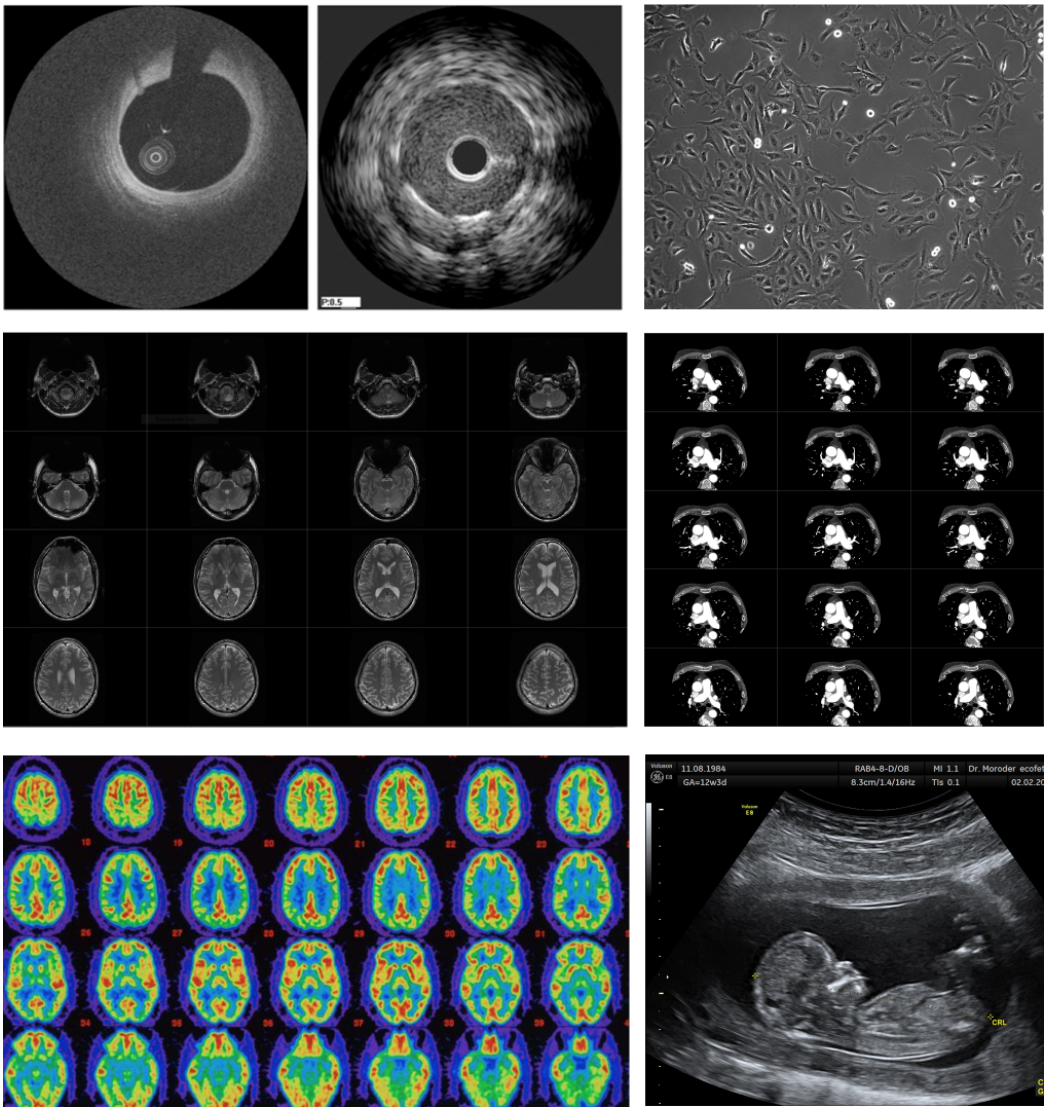


Figure 2.12: Examples of medical imaging techniques. First row left: intravascular ultrasound. First row middle: optical coherence tomography. First row right: phase contrast microscopy. Second row left: brain MRI. Second row right: computed tomography angiography. Third row left: brain PET [19]. Third row right: fetal ultrasound [20].

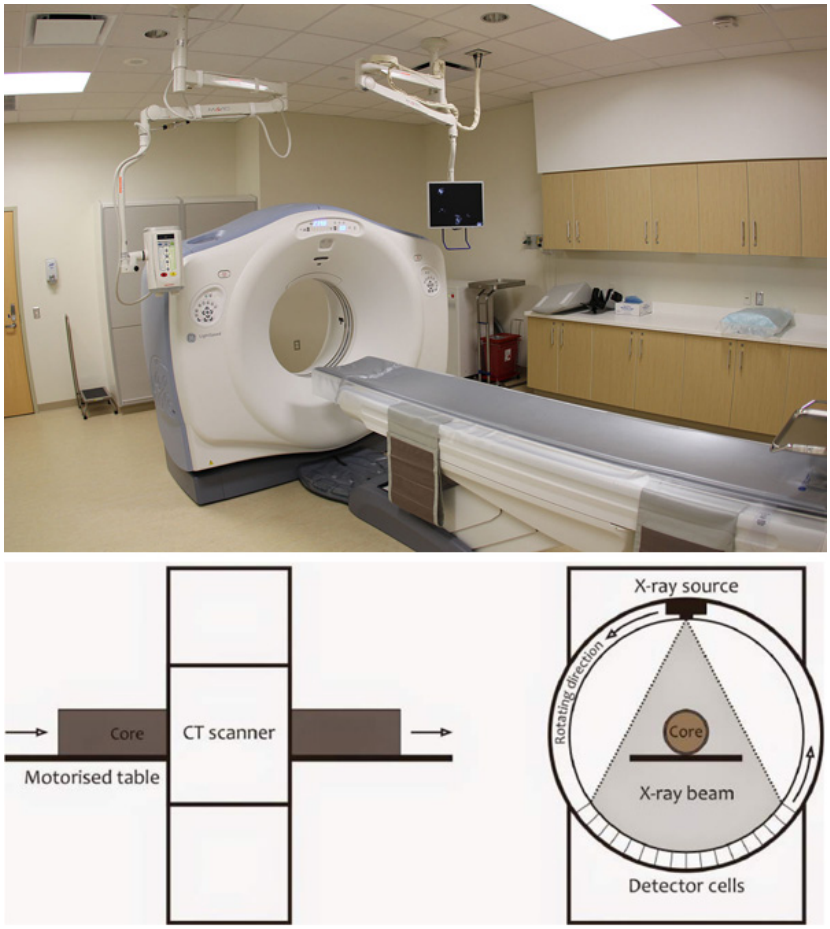
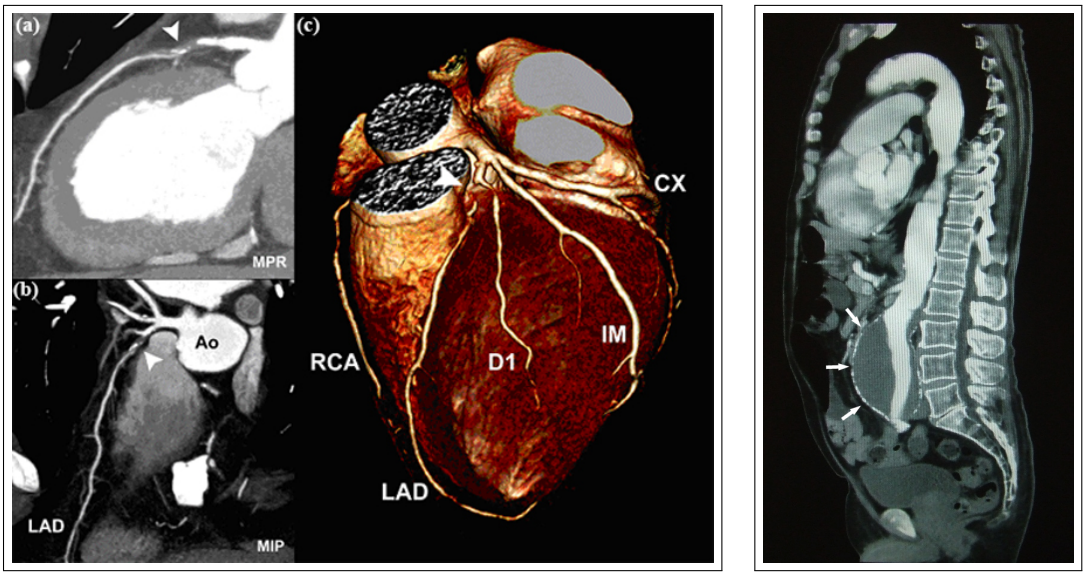


Figure 2.13: Modern CT scanner, adapted from [21, 22].

struction, where the blood vessels (bright regions) have relatively high intensity compared to the myocardium. Fig. 2.14 (b) shows a CTA image of a human torso, where the blood vessel are highlighted in CT images that have similar appearance with high density regions, such as bone.

2.3.3.2 Fractional Flow Reserve

CTA can provide accurate measurements of the geometrical structure of blood vessels. However, sometimes it is not sufficient to precisely locate the regions of stenoses and pathological changes, especially for small vessels, such as coronary artery. In order to reduce the risk of TAVI procedure, and achieve the best treatment, Fractional Flow Reserve (FFR) is normally



(a) (b)

Figure 2.14: (a) Illustration of CTA images and 3D reconstruction of coronary artery, adapted from [23]. (b) Illustration of a CTA image of human torso, adapted from [24].

given before the surgery. It is a guide wire-based procedure that can measure blood pressure, temperature, and flow through a specific part of the vessels using a small sensor on the tip of the wire. FFR is performed through a standard diagnostic catheter at the time of a coronary angiogram. The measurement of FFR can then be used to determine the exact severity of the stenosis, which has been shown useful in assessing whether or not to perform angioplasty or stenting at pre-surgery stage. Although, FFR is an invasive procedure, it has certain advantage over non-invasive CTA. The most significant benefit is that FFR can quantitatively estimate the narrowing whereas CTA only visualises contrast inside a vessel.

2.4 Object Detection

Object detection is the process of finding instances of real-world objects, such as faces, pedestrian, bicycles, and buildings in images or videos. Object detection algorithms typically use extracted features and learning algorithms to recognise instances of an object category. It is commonly used in applications such as image retrieval, security, surveillance, and automated vehicle parking systems. It can be broadly categorised into binary detection and object localisation in terms of the number of object category that it tries to find. Binary detection involves

detecting instances of objects from a particular class in an image, where a binary decision is made. In order to differentiate binary detection, the term "object localization" is used to refer to multi-class detection problem. We will briefly review some popular methods in binary and multi-class detection.

2.4.1 Binary Detection

Face detection and pedestrian detection are two most typical binary detection problems where a foreground and background classifier is built to find the objects in an image from a particular category. It generally has two main stages and one optional stage as follows: hypotheses generation, foreground object prediction, and location refinement.

The Viola-Jones object detection framework [25] proposed in 2001 by Paul Viola and Michael Jones is the first object detection framework to provide competitive object detection rates in real-time. Although it can be trained to detect a variety of object classes, it was motivated primarily by the problem of face detection [26]. The hypotheses are generated using sliding window technique at multiple scales, such that different sizes and locations of face are all covered in the whole hypotheses set. Haar features are extracted efficiently using an integral image algorithm [79]. A set of binary AdaBoost classifier are trained and connected consecutively to form a cascade model, where the background hypotheses are progressively eliminated. Those ones are retained by the cascade model are considered as face regions that detected in the image. The key contributions of Viola-Jones object detection framework are threefold. First, an integral image was used that allows for very fast feature evaluation. It can be computed from an image using a few operations per pixel. Once computed, any one of these Haar-like features can be computed at any scale or location in constant time. The second contribution is a method for constructing a classifier by selecting a small number of important features using AdaBoost. In order to ensure fast classification, the learning process must exclude a large majority of the available features, and focus on a small set of critical features. AdaBoost evaluates the discriminative power of those weaker classifier, and combines them through critical weighting, which can be viewed as a feature selection process. It provides an effective learning algorithm and strong bounds on generalisation performance. The third major contribution is a method for combining successively more complex classifiers in a cascade structure which dramatically increases the speed of the detector by progressively eliminating negatives and focussing on promising regions of the image. The pipeline of Viola-Jones object

detection framework is shown in Fig. 2.15.

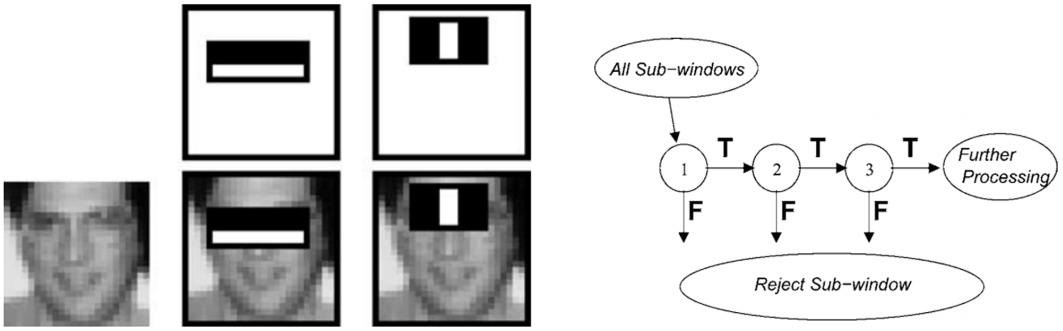


Figure 2.15: The Viola-Jones object detection framework, adapted from [25, 26].

More powerful discriminative feature descriptors and stronger classifiers have also been proposed to solve binary detection problem. The most successful work is [27] where the HOG feature and an SVM classifier are used to detect the whole human body. The essential idea behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is computed. Fig. 2.16 shows an example of HOG feature for human detection. The descriptor is the concatenation of these histograms. For improved accuracy, the local histograms can be normalised with respect to the contrast by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalise all cells within the block. This normalisation results in better invariance to changes in illumination and shadowing effects. The HOG descriptor has a few key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation. Such changes would only appear in larger spatial regions. Moreover, coarse spatial sampling, fine orientation sampling, and strong local photometric normalisation permit the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HOG descriptor is thus particularly suited for human detection in images. Zhu *et al.* [80] presented an algorithm to significantly speed up human detection using HOG descriptor methods, where HOG descriptors were used in combination with the cascading classifiers algorithm normally applied with great success to face detection. Their proposed algorithm achieved comparable

performance to the original algorithm, but operated at speeds up to 70 times faster.

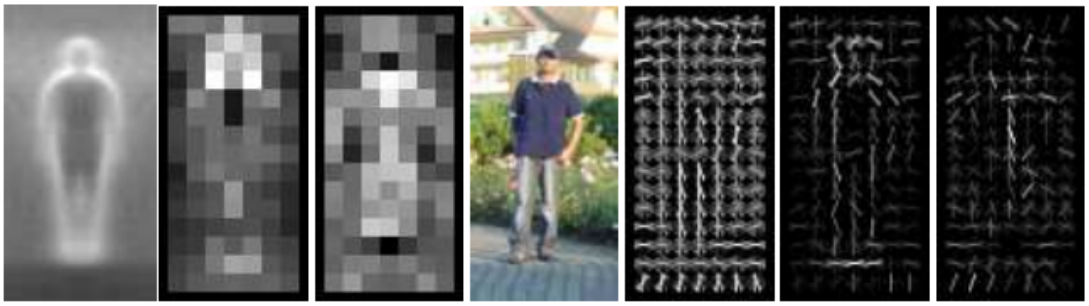


Figure 2.16: The HOG feature for human detection, adapted from [27].

2.4.2 Object Localisation

Traditional object recognition can be considered as a multi-class extension of object detection, whose algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. Common techniques include edges, gradients, HOG, Haar, and Local Binary Patterns (LBP). However, traditional object recognition algorithm is heavily limited by the discriminative power of features that are used. DNN is becoming more and more mainstream [37], as it has been shown superior over many other methods, especially for visual recognition tasks. It is able to learn the visual features hierarchically via training in supervised fashion, which avoids hand-crafting features. The following can be considered as three of the key reasons that contributed to the success of DNNs. First, training a multi-layer neural network involves finding a local minimum of a highly non-linear function. In order to obtain a reasonable local minimum, gradient descent based methods require a good initialisation. Layer-wise unsupervised pre-training methods [36] were developed and have been proved to be more efficient compared with random initialisation. Second, a large amount of labelled datasets [38, 39, 40] are vitally important to the advance in supervised training. For example, Microsoft COCO dataset [40] contains more than 300,000 images, over 2,000,000 instances from 80 object categories, where each image has 5 caption labels. Moreover, advances in hardware makes both forward pass and backward propagation computationally efficient. Especially, with dedicated high speed memory module and Single Instruction Multiple Data (SIMD) architecture, GPGPU is particularly well placed for learning deep neural network structures [41].

One of the well known work is AlexNet [28] which ranked the first in the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). The network was made of 5 convolutional layers, max-pooling layers, dropout layers, and 3 fully connected layers. The network was used for classification with 1000 possible categories. The most notable improvement is that AlexNet achieved a top 5 test error rate of 15.4%, while the next best entry achieved an error of 26.2%. Very recently, several works have shown that Regions with Convolutional Neural Network Features (R-CNN) [81, 29, 82] and Spatial Pyramid Pooling CNNs (SPPnet) [83] are effective in simultaneous object localisation and recognition. These methods contain four main components: convolutional feature extraction, region proposal generation, ROI classification, and bounding box refinement. In [81], the authors showed that the representation feature learnt with CNN using deep structure can be effectively used for visual classification and ROI regression. By introducing spatial pyramidal pooling layer to generate a fixed length output feature regardless the size of input image, [83] overcame the limitation of [81] without cropping or wrapping the images that are problematic as they result in information loss and distortion. The work in [29, 82] improved the computational efficiency further by sharing the deep convolutional layers with region proposal, classification and regression networks. However, for small objects, R-CNNs have difficulty detecting them in small scales due to low resolution and lack of visual context.

As for face detection, Farfade *et al.* [84] proposed a multi-view face detection method, Deep Dense Face Detector (DDFD), which uses a fine-tuned 8-layer AlexNet [28] that was initially designed for object recognition. It has 5 convolutional layers and 3 fully connected layers. A pre-trained AlexNet was fine-tuned for face detection on 200,000 face patches and 20,000,000 background patches, which were all resized to 227×227 pixels in order to match the input size of AlexNet. During the testing stage, the sliding window approach was used to generate hypotheses. DDFD classifies each candidate into face or background, and decision confidence scores is obtained. Non-Maximal Suppression (NMS) is followed to remove redundant bounding boxes. Fully Convolutional Neural Network (FCN) [85] was firstly introduced for semantic segmentation, and then adapted to solve object detection problems. In contrast to classifying each object hypothesis into face or background, FCN based approaches take the whole image as input, and the convolutional outputs of forward pass are considered as a set of feature maps. Detection then can be achieved by investigating the region pattern of the target object on the feature maps. UnitBox [86] is derived from VGG-16 [87] model, and

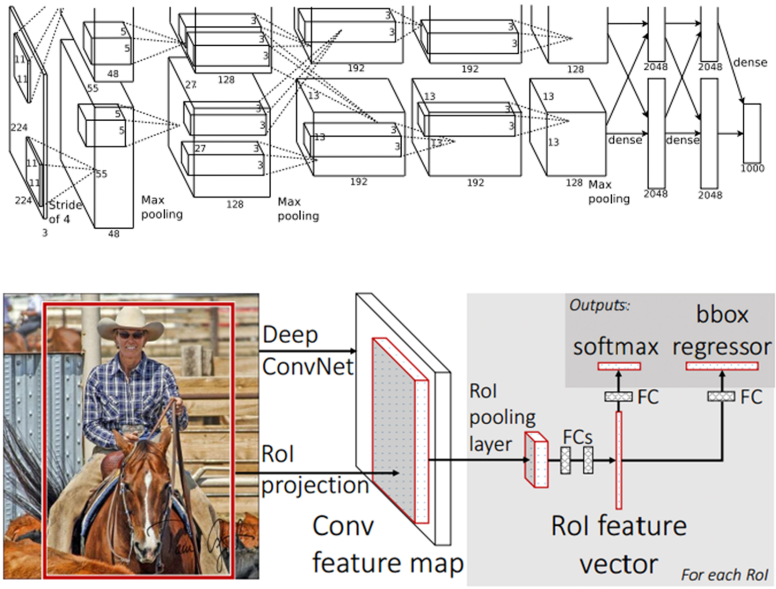


Figure 2.17: Deep learning methods in object localization. The network architecture of AlexNet (Top) [28], and Fast R-CNN (Bottom) [29].

replaces the original fully connected layer with two pixel-wise bounding box prediction layers. The network can then be trained via minimising the Intersection over Union (IoU) loss, which quantitatively measures how well the predicted bounding boxes are aligned with ground truths. Yang *et al.* [88] and Bai *et al.* [89] also showed that incorporating FCN with multi-scale strategy helps to boost detection accuracy. In addition to solving detection problems, it is common to use those deep models for multi-tasks jointly, such as fiducial landmark localisation, face pose estimation, gender recognition, and 3D face modeling [90, 91, 92, 93].

2.5 Summary

In this chapter necessary background information has been discussed in preparation for presenting the proposed methods in the following chapters. The chapter started by providing an overview on supervised machine learning algorithms, which includes RF, CNN, and cascade classifier. This was followed by an overview of conventional approaches for medical image segmentation, especially in deformable models, and graph-cut segmentation. A discussion on geometric and parametric deformable modelling is included as well. The anatomical structures

of the cardiovascular system, particularly the aorta root and arch, and the coronary artery, and relevant disease is provided. An overview of object detection and recently advance in deep learning based methods are provided as well.

In the rest of thesis, we will first investigate two general segmentation methods for 3D medical images given user interventions, such as foreground and background guiding strokes, in an interactive manner using adaptive learning methods. Particularly, we focus on segmenting two cardiovascular anatomies, the coronary artery and the aorta root in Chapter 3 and Chapter 4 respectively. In Chapter 5, we also show the feasibility of combining cascade scheme with CNNs to solve face detection problem in unconstrained environment.

Chapter 3

Coronary Artery Segmentation

Contents

3.1 Introduction **41**

3.2 Proposed Method **42**

3.2.1 Vessel Enhancing Diffusion 42

3.2.2 Multi-Scale Coronary Feature Extraction 44

3.2.3 Voxel Classification using Random Forests 46

3.2.4 MRF Regularization with Primal Dual Algorithm 49

3.3 Experimental Result **52**

3.3.1 Segmentation Software 52

3.3.2 Segmentation Result 55

3.4 Summary **57**

In Chapter 1, we defined the adaptive learning as a progressive learning process that gradually builds the model given a sequential supervision data from user interactions. In the case of using small scale model and dataset, the learning process that involves adaptive re-training given the accumulated interaction is often affordable. Especially, when the discriminative features for the well-defined classification problem are available, adaptive re-training strategy usually lead to a better model compared to on-line strategies in terms of prediction accuracy and generalisation. In this chapter, we show such an adaptive learning scheme can lead to an efficient interactive method for segmenting the coronary artery from 3D CTA images.

3.1 Introduction

An accurate segmentation algorithm for extracting the vessel structure from the heart is often considered essential for patient-specific modelling of cardiovascular diseases. A number of vessel extraction methods for different modalities have been developed in recent years, e.g. [94, 95, 96]. Li and Yezzi [95] proposed a 4D representation for 3D vessel by combining both the spatial coordinates and the thickness of the vessel. With two user specified endpoints, the surface as well as the center line of the vessel are extracted using the generalized 4-D global minimal paths algorithm. Esneault *et al.* [97] proposed a 3-D geometrical moment-based detector to extract the centre line of the vessel, as well as its diameter and orientation. Finally, a graph cut algorithm was applied to regularise the final segmentation with a local continuity constraint. In [98], shape prior of 3D tubular tree structure is used to formulate the regularisation to refine the initial vessel segmentation or detection. Zhu and Chung [99] proposed the Tubularity Markov Tree (TMT) method to model and detect vessel structure, with a graph cut algorithm applied to solve the energy minimisation problem in order to obtain the final segmentation. Deformable models, particularly those that are capable of capturing complex geometries such as [100], may be applied to vessel segmentation. Efficient model representations and numerical methods are desirable and semi-implicit schemes have been shown effective in segmenting complex objects, e.g. [101, 102].

The segmentation of coronary artery is not a trivial problem. Coronary arteries are relatively small blood vessels in the heart which branch off from the root of aorta, and is divided into two main sub-branches. These two main sub-branches further split and grow a tree-like structure. First, the coronary artery is attached to the myocardium and surrounded by other tissues. Second, compared to the aorta, the size of the coronary artery is much smaller, which makes it difficult to segment and maintain its vessel connectivity. The size of blood vessel is getting even smaller when it splits into multiple sub-branches, i.e. 2-4 pixels wide. Third, labelling such anatomy is extremely difficult as there is no appropriated viewing plan to visualise, and the geometry variations are large across different patients. Last but not least, there are several other blood vessels nearby, such as pulmonary blood vessels in the lung, which has very similar appearance and geometry. This makes an automated, global detection or classification a difficult task. For example, region-growing based methods [103] are commonly used for vessel segmentation, however, they are strongly limited by the quality of the images, where the connectivity of such subtle structure is often broken. Interactive region growing [104] was

proposed to address this issue by selecting the new growing seeds at the broken regions manually, where a large amount of interactions and labour efforts are required. In this chapter, we present an interactive coronary artery 3D segmentation method for CTA volumetric image. An initial vessel classification is given by a random forest classifier which is trained on a few user strokes: the foreground stroke labels the coronary artery and the background stroke indicates the other tissues. Based on the label population in the leaf nodes of the randomised decision trees, we formulate the final segmentation as an MRF based optimisation with local consistency constraints. The primal dual algorithm with graph cut is used to solve the energy minimisation problem.

The rest of the chapter is organized as follows. Section 3.2 presents our proposed approach including vessel enhancement, feature extraction, RF classification and MRF optimisation. The interactive segmentation software, typical segmentation process, and experimental results are presented in Section 3.3. Section 3.4 concludes the proposed method.

3.2 Proposed Method

Given a 3D CTA image of the human heart region, the coronary artery is segmented by taking the following four steps. First, the original image is smoothed, but the tubular-like structure is preserved and enhanced using multi-scale vessel enhancing diffusion. Next, the features designed for detecting the target vessel is computed in a multi-scale fashion. Third, according to the user provided strokes the random forests classifier is trained, and classifies each vertex in the volume into a positive point (coronary artery) or a negative point (other tissues). Finally, the MRF model is used to formulate the energy function for regularising the initial classification result with local consistency constraint. The finally segmentation is achieved by solving the energy minimisation problem using integer programming with primal dual strategy. The pipeline of proposed method for interactive 3D coronary artery segmentation is shown in Fig. 3.1.

3.2.1 Vessel Enhancing Diffusion

It is desirable to enhance the brightness of coronary vessel and the sharpness of vessel edges in 3D before carrying out segmentation, particularly for this type of thin tubular structures. Frangi et al. [105] proposed a vesselness function by analyzing the eigenvalues of the second order information (Hessian) in a local neighborhood at multiple scales. The eigenvalues decomposed

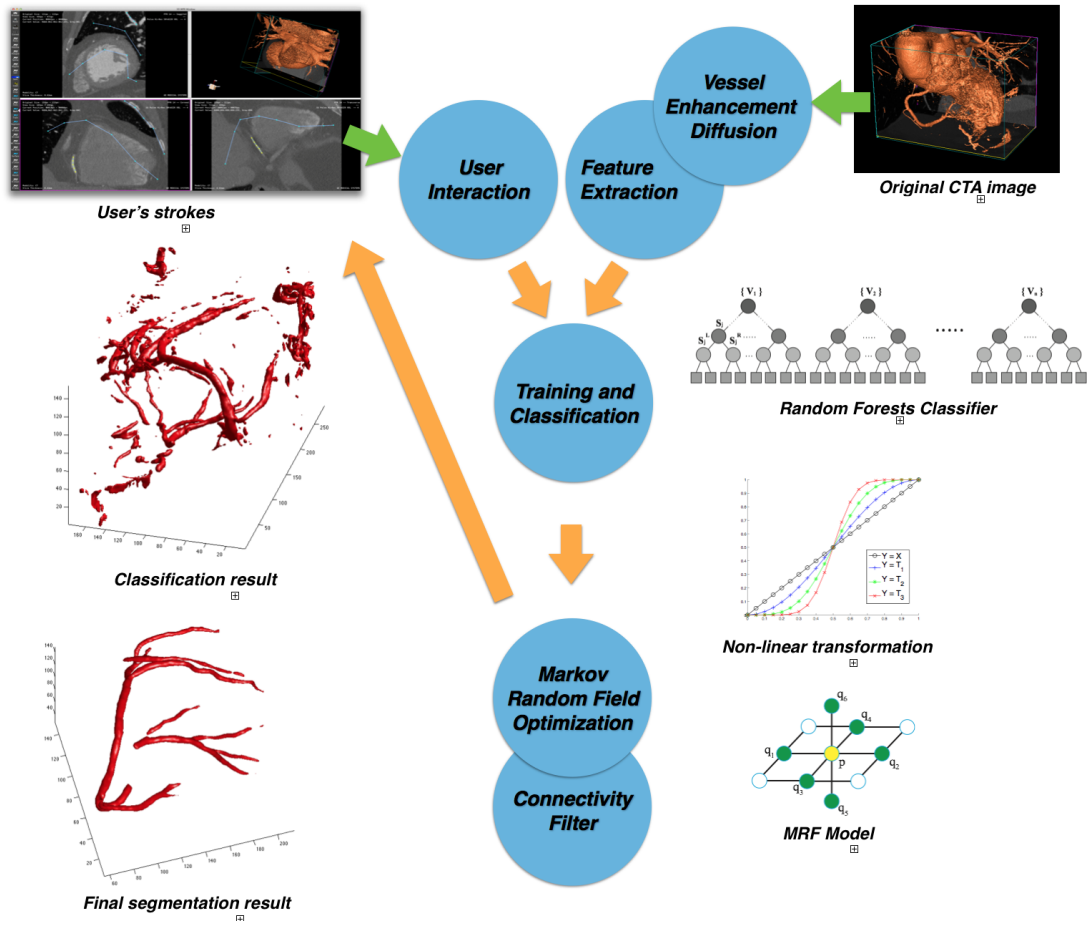


Figure 3.1: The pipeline of proposed method for interactive 3D coronary artery segmentation.

from the Hessian matrix were used to locally differentiate the tubular-like structure from other structures, including blob-like structure, plate-like structure and background. Manniesing *et al.* [106] extended it to a continuous, n-th order differentiable function for measuring the vesselness. Based on the proposed vesselness function, the diffusion tensor was constructed to enhance the image at vessel region along minimal local curvature direction, while smoothing the image isotropically at non-vessel region. In this work, we follow this approach to enhance the coronary vessel structures in the CTA images. Later, the measurements derived from this vesselness analysis are also used as part of the coronary features.

Given the vesselness function (Eq. 3.1) and the orthogonal eigenvectors of Hessian matrix, at the vessel region, we construct the anisotropic diffusion tensor to preserve and enhance the tubular-like structure by maximizing the strength of diffusion in the minimal curvature

direction, and minimizing the diffusion in the rests of two directions. At the same time, at the non-vessel region, the isotropic diffusion tensor is required as well, in order to reduce the background noise. The diffusion tensor is defined as follows:

$$\mathscr{D} \triangleq \mathscr{Q}\Lambda'\mathscr{Q}^{\mathscr{T}} \tag{3.1}$$

where the $\mathscr{Q}$ is the orthogonal eigenvectors of Hessian matrix $\mathscr{H}$, and the Λ' is a 3 by 3 diagonal matrix, with the following values on its diagonal

$$\lambda_1' \triangleq 1 + (\omega - 1) \cdot \mathscr{V}^{\frac{1}{R}} \tag{3.2}$$

$$\lambda_2' = \lambda_3' \triangleq 1 + (\varepsilon - 1) \cdot \mathscr{V}^{\frac{1}{R}} \tag{3.3}$$

where $\omega \gg \varepsilon > 0$, and $0 \leq \mathscr{V} \leq 1$. When the vesselness $\mathscr{V}$ is approximating to the maximum ($\mathscr{V}_{max} = 1$), the maximum diffusion factor $\lambda_1' = \omega$ in the minimal curvature direction, and the minimum $\lambda_2' = \lambda_3' = \varepsilon$ in the rests are achieved, which ensures the anisotropic diffusion along the vessel direction. On the contrary, the isotropic diffusion tensor with $\lambda_1' = \lambda_2' = \lambda_3' = 1$ is obtained when the vesselness $\mathscr{V} = \mathscr{V}_{min} = 0$. The parameter R controls the sensitivity to the vesselness response.

3.2.2 Multi-Scale Coronary Feature Extraction

Coronary arteries are the blood vessels that circulate the blood with oxygen into heart muscle myocardium. At the root of the aorta, it branches off into two main coronary arteries, and then these coronary arteries branch off into smaller arteries to form a tree structure. The left and right coronary arteries run on the surface of the heart. The examples of coronary arteries in 3D CTA images shown in Fig. 3.2. Segmenting coronary arteries is a challenge problem as there are many anatomical structures forming similar geometries and appearances, such as pulmonary vessels, bones of rib cage, myocardium, and adventitia. Fig. 3.3 shows the examples of other anatomical structures in 3D CTA images.

The features we use to highlight coronary vessels can be categorised as texture or appearance based and shape based. The texture features are derived as intensity and image gradient magnitude distribution in a local neighborhood across multiple scales. These appearance features are useful in differentiating myocardium, bone, and adventitia. Since they are extracted from multiple scales, the difference between other vascular structures and coronary vessels can be highlighted. For example, although aorta exhibits similar brightness to coronary, it has

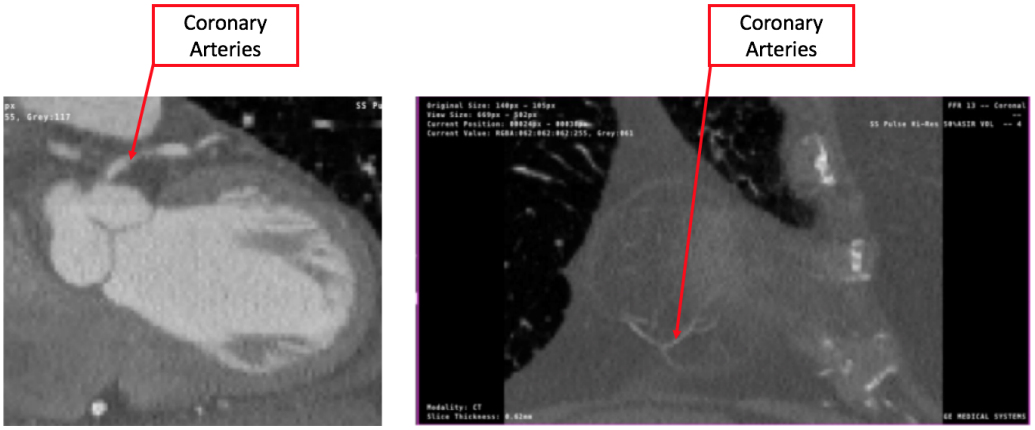


Figure 3.2: The examples of coronary arteries in 3D CTA images.

different intensity distribution across scales because aorta is a much larger vessel. Pulmonary vessels has similar geometry to coronary arteries but their neighborhood appearances are different. The second set of features are designed to highlight the narrow, tubular-like structure of coronary vessels. We derive multiscale local geometrical features, following those that have been used in vessel enhancement.

At scale $\mathscr{S}$, the Hessian matrix $\mathscr{H}$ at each voxel $\mathscr{P}$ is computed by convolving the volumetric image with derivatives of Gaussian. The eigenvalues λ of Hessian matrix are then computed. In the case of 3D, we define the ordering eigenvalues as $\mathscr{H}$ as $\lambda_1, \lambda_2, \lambda_3$, where $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$. At scale s, the eigenvalues of Hessian indicate the strengths of intensity variation between the inside and outside of the region $(-s, s)$ along the direction of the corresponding eigenvectors. The coronary vessels are assumed at the region in which $|\lambda_1| \approx 0$, $|\lambda_1| \ll |\lambda_2|$, $\lambda_2 \approx \lambda_3$, while the eigenvector corresponding to the eigenvalue λ_1 indicates the vessel direction. We adopt the Manniesing's vesselness function $\mathscr{V}$, which is defined as follows:

$$\mathscr{V} \triangleq \begin{cases} 0 & \lambda_2 \geq 0 \; or \; \lambda_3 \geq 0 \\ (1 - e^{\frac{-A^2}{2\alpha^2}}) \cdot e^{\frac{-B^2}{2\beta^2}} \cdot (1 - e^{\frac{-S^2}{2\gamma^2}}) \cdot e^{\frac{2c^2}{|\lambda_2|\lambda_3^2}} & otherwise \end{cases}$$

where

$$A = \frac{|\lambda_2|}{|\lambda_3|}, \;\; B = \frac{|\lambda_1|}{\sqrt{|\lambda_2 \lambda_3|}}, \;\; S = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \tag{3.4}$$

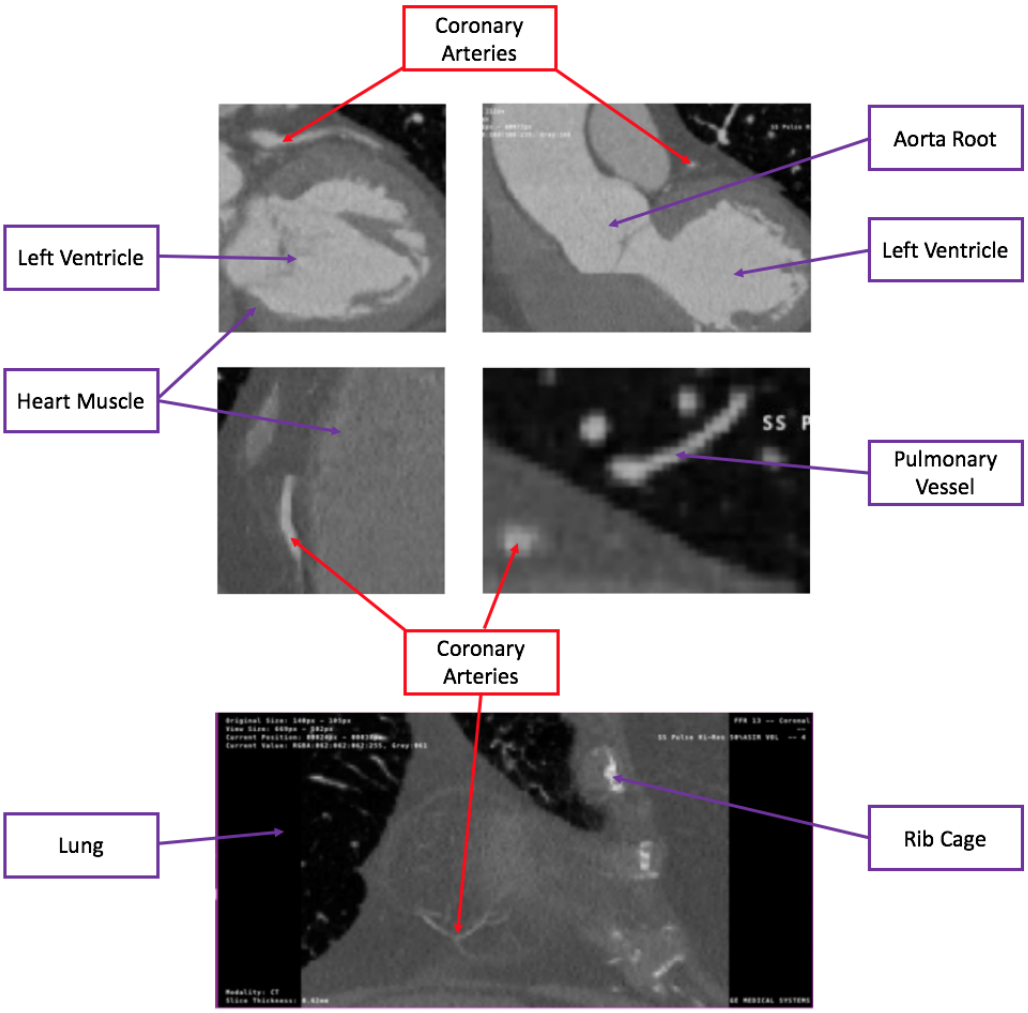


Figure 3.3: The examples of other anatomical structures in 3D CTA images.

The parameters α, β, γ are the weighting variables controlling the contributions of the measurements to the response of vesselness function. The vesselness measurement are computed at multiple scales and the maximum response $\mathscr{V}$ over the scale-spaces is selected.

3.2.3 Voxel Classification using Random Forests

DTs are a popular method for various supervised learning problems, as it is invariant to scaling and various other feature transformations. However, one of major limitation of decision trees is that it tends to learn highly irregular patterns, and over-fits to training sets, especially when

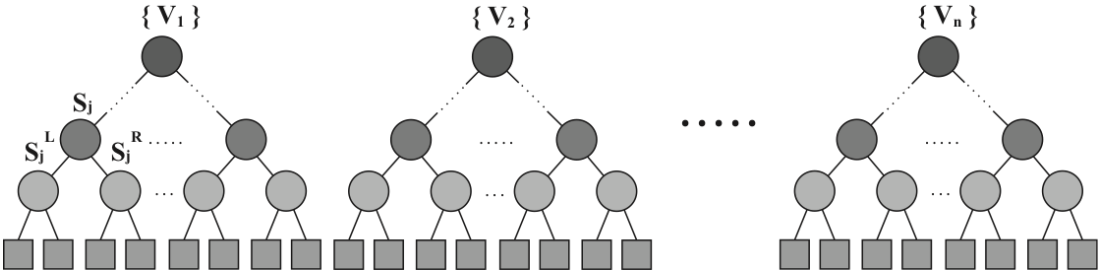


Figure 3.4: Random Forests is an ensemble classifier consisting of a set of DT.

very deep structures are used. RF is a supervised machine learning method which aims to overcome the problems caused by poor generalization ability of single decision tree by a way of averaging multiple deep decision trees given different subsets of the training set. The general methodology of random forests was introduced by Tin Kam Ho [107, 108], and then Leo Breiman formed the basis of the modern practice of RF [32], where the out-of-bag error was proposed to evaluated the generalization error, and the measurement of feature importance was introduced through variable permutation. The modern RF is an ensemble classifier consisting of a set of decision trees shown in Fig. 3.4, which significantly improves the generalisation ability of the classifier compared to a single decision tree.

At the bootstrap aggregating stage (bagging), assuming that the data sample is independent and identically distributed, new training sets are generated by randomly sampling with replacement from the complete training set. Given a training set $\mathscr{T}$ of size N, bagging creates M sub training sets $\mathscr{T}'$ with size of N' via uniformly sampling $\mathscr{T}$ with replacement. Hence, some observations are repeated. For each new training set of $\mathscr{T}'$, one decision tree is constructed which consists of a set of split nodes and linking edges. Each non-leaf node stores a random test function which is applied to the input data, and leads to the leaf node. The information gain and Gini impurity are two popular metrics used to evaluate the performance of random test function. The information gain measures the entropy difference between the parent node, and weighted sum of its direct children nodes, which can be computed as:

$$I = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i) \tag{3.5}$$

In discrete case the Shannon entropy is defined as:

$$H(S) = -\sum_{c \in C} p(c) log(p(c)) \tag{3.6}$$

In the continuous case, for example, the entropy of a Gaussian distribution with d multi-variable can be computed using the equation:

$$H(S) = \frac{1}{2}\log\Big((2\pi e)^d|\beta(S)|\Big) \tag{3.7}$$

where $\beta(S)$ is the covariance matrix. Hence, finding the best split for each non-leaf node is equivalent to maximizing the information gain. Gini impurity is alternative choice for growing a decision tree, which measures how often a randomly chosen sample from the training set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset. Given a set of training samples from $\mathbf{J}$ categories, $i \in \{1,2,...,\mathbf{J}\}$ is the index of categories, and g_i be the fraction of items labelled with category i in the set, the Gini impurity can be computed as:

$$G = \sum_{i=1}^{\mathbf{J}} g_i(1-g_i) = 1 - \sum_{i=1}^{\mathbf{J}} {g_i}^2 = \sum_{i!=k} g_i g_k \tag{3.8}$$

It is noteworthy that modern random forests introduce the idea of searching over a random subset of the feature when splitting a node, which decorrelates individual decision trees further to improve the generalization ability.

In the leaf nodes, the final predictor is stored. At the testing stage, all the trees predict the incoming data independently, and the final prediction is combined by averaging the output for regression, or voting for classification.

$$f = \frac{1}{M}\sum_{i=1}^{M} f_i(x) \tag{3.9}$$

where f, and f_i are overall combined predictor, and individual predictors for each bagging sub training set respectively. In our case, the classification RF is used, such that the most voted class given by the trees is considered as the final classification of the RF.

In this work, classifying each voxel into coronary and non-coronary is required so that the confidence value of the classification at each voxel is used to construct the cost function for the MRF based optimisation. The RF grows a number of DTs independently using subsets of training data by randomly sampling with replacement from the complete training set. For each single decision tree, it grows recursively by finding the best splitting function for each non-leaf node using the entropy or Gini index to evaluate the information loss, until the stopping criteria are satisfied. The non-leaf nodes consist of the splitting functions, each testing sample could follow the tests and reach the leaf node in the end. The leaf nodes, at the bottom layer

of the tree, store the training samples which fell in in the training stage, and it votes the class with largest proposition for the prediction. The random forests combines the prediction of each single decision tree, the most voted class given by the forests is considered as the final classification for the test sample. From the implementation point of view, the random forests is supremely adequate for paralleling. By taking advantage of GPU computing technique, classifying each pixel of a 500×300 2D image can be achieved in 140ms [109].

3D Multi-Planar Reconstruction (MPR) and curved MPR are used to produce the longitudinal view of the coronary artery, in which a few strokes are placed to indicate the region of interest at the foreground. Also, the non-coronary artery tissues, such as: aorta, ventricle, heart muscle, pulmonary blood vessels and so on, are obtained through additional user strokes as background, negative samples. We sample the voxels following the strokes with equal spacing, and the features of those voxels as described in Section. 3.2.2 are collected as training set. Then the whole volume is tested, the classification result may be considered as an initial segmentation. However, the proposition of voting by these randomized decision trees for each voxel can be considered as segmentation cue. In the next section, we show how to use these proposition values to carry out segmentation that can be more coherent than RF classification.

3.2.4 MRF Regularization with Primal Dual Algorithm

The MRF has been widely applied in different computer vision applications to address the regularisation problems. Especially, the grid-like, pairwise MRF model in image segmentation area has shown to be an effective approach, e.g. [110]. In general, the MRF energy is formulated over the graph $G(\mathscr{P},\mathscr{E})$ as follows:

$$E(p) = \sum_{i \in \mathscr{P}} U(p_i) + \sum_{<i,j> \in \mathscr{E}} O(p_i, p_j) \tag{3.10}$$

where $\mathscr{P}$ and $\mathscr{E}$ represent the node set and the two-tuples set of undirected edge of G respectively. $U(.)$ is the unary potentials defined on the node $\mathscr{P}$, and $O(.)$ is the pairwise potentials defined on the edge $\mathscr{E}$. The first term of Eq. 3.10 is considered as point-wise data term which provides the segmentation cue, the second term is considered as pair-wise smoothness term which constrains the consistency between neighbour nodes. For example, the Potts pairwise potentials is defined on the distance of two linked nodes x_i, x_j, as follows:

$$O(p_i, p_j) = w_{ij} \cdot (1 - \delta(p_i - p_j)) \tag{3.11}$$

where $w_{ij} \geq 0$ is the weighting coefficient of smoothing penalty for the edge $< i, j >$, and the Kronecker delta δ is defined as:

$$\delta(p) = \begin{cases} 1 & x = 0 \\ 0 & x = 1 \end{cases} \tag{3.12}$$

Here, the segmentation is to assign each voxel/vertex with a label l_p ($l_p \in \mathscr{L}$; $l_p = 1$ when p is coronary artery; $l_p = 0$ when p is not). In Section. 3.2.3, the binary classification result of each vertex is given by the classifier as well as the voting proposition k_{l_p} which could be considered as the likelihood or confidence of being categorised to the class. So, the regularisation can be formulated as solving the discrete MRF optimisation problem by minimising the following MRF energy function:

$$min\left(\sum_{p\in\mathscr{P}} U(p) + \sum_{<p,q>\in\mathscr{E}} O(p,q)\right) \tag{3.13}$$

In binary classification case, we have $k_{l_p=0} = 1 - k_{l_p=1}$, so the point-wise potentials is defined as:

$$U(p) = \begin{cases} T(1-k_{l_p=1}) & if \quad l_p = 1 \\ T(1-k_{l_p=0}) & if \quad l_p = 0 \end{cases} \tag{3.14}$$

which implies, for example, the cost of assigning the class label 0 to the vertex p is equal to $T(1-k_{l_p=1}) = T(k_{l_p=0})$, the non-linear transformation of the confidence of assigning it with label 1. One disadvantage of graph-cut based method is the shrink bias which results in smaller contour, and becomes even worse in the corner region [111]. The goal of the non-linear transformation function T is to enlarge the difference between $k_{l_p=0}$ and $k_{l_p=1}$ when their values are getting similar, which is very common for the vertexes around the vessel surfaces. By applying the non-linear transformation, to a large extent, the shrink bias caused by the pair-wise term will be reduced. We propose the following non-linear transformation function T:

$$T_\eta(k) = \overbrace{t \circ t \circ \cdots \circ t}^{\eta} \tag{3.15}$$

$$t(k) = \frac{1}{2} + \frac{1}{2}\sin(\pi k - \frac{\pi}{2}) \tag{3.16}$$

where $\circ$ is the function composition operator (see Fig. 3.5). It can be proved that $T_\eta(1-k) = 1 - T_\eta(k)$ when $k \in [0,1]$. In addition, we use grid-like MRF with 6-neighbourhoods system in our experiment, and choose the Potts model as pair-wise potentials (see Eq. 3.11, 3.12).

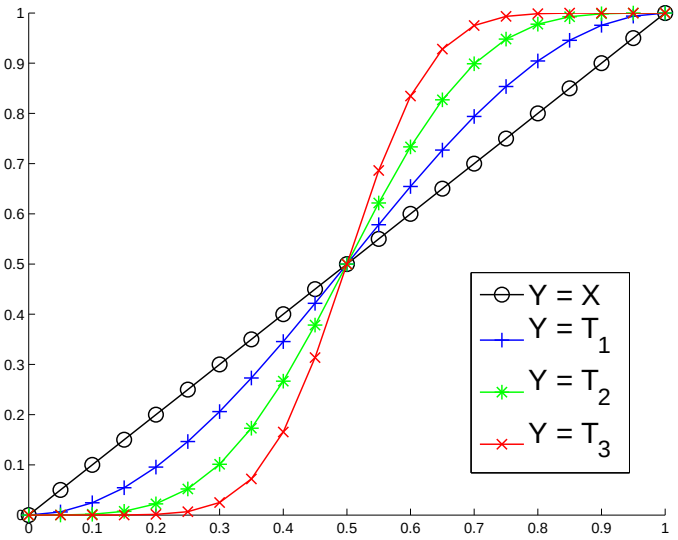


Figure 3.5: The plots of proposed non-linear transformation function $T_{\eta=1,2,3}$ compared to the linear function $Y = X$.

A number of approaches have been proposed in the literature to solve the energy minimisation problem of discrete pair-wise MRF, such as graph cuts based methods [112, 113], and belief propagation algorithm [114]. Especially, the dual-decomposition approach with linear programming method attracts a great attention in the last decade [115]. Chekuri et al. [116] have proved that the solution of metric labeling problem given by the form of minimizing the MRF energy (Eq. 3.13) can be approximated using the following integer programming formulation:

$$min\left(\sum_{p\in\mathscr{P},a\in\mathscr{L}} c_p(a)x_p(a) + \sum_{<p,q>\in\mathscr{E}} w_{pq}\sum_{a,b\in\mathscr{L}} d(a,b)x_{pg}(a,b)\right) \tag{3.17}$$

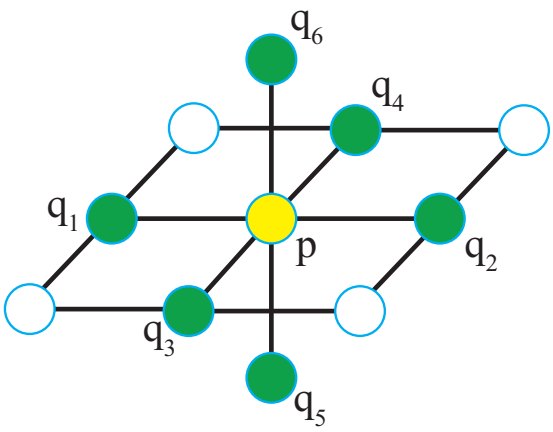
which subjects to the following constraints:

$$\sum_{a\in\mathscr{L}} x_p(a) = 1 \qquad \forall p \in \mathscr{P} \tag{3.18}$$

$$\sum_{a\in\mathscr{L}} x_{pg}(a,b) = x_q(b) \quad \forall b \in \mathscr{L}, < p,q > \in \mathscr{E} \tag{3.19}$$

$$\sum_{b\in\mathscr{L}} x_{pg}(a,b) = x_p(a) \quad \forall a \in \mathscr{L}, < p,q > \in \mathscr{E} \tag{3.20}$$

where $x_p(.) = 0,1$, and $x_{pq}(.,.) = 0,1$. The constraint Eq. 3.18 ensures for each vertex p, a label is assigned to, and the constraints Eq. 3.19 and 3.20 ensure the consistency of the label between



(b)

Figure 3.6: The graphical representation of MRF with 6 neighbourhood system, $< p, q_{1,\cdots,6} >$.

the neighbours. Komodakis *et al.* gave the solution of above optimisation problem Eq. 3.17 via dual decomposition, and proposed a family of PD (Primal-Dual) algorithms. In this work, the PD1 algorithm [117, 118] is adopted, which solves the decomposed sub-problems via graph cut in each iteration.

3.3 Experimental Result

3.3.1 Segmentation Software

We developed an interactive segmentation software namely *"SwanseaVision Medical Image Segmentation Toolkit (SVMIST)"* to evaluate the proposed method. The summary of technical implementation details of *SVMIST* is listed in Table 3.1. It implements a 3D image viewer supporting standard Digital Imaging and Communications in Medicine (DICOM) format, such as DICOM disk and sequential DICOM files. It provides standard Multi-Planar Reconstruction (MPR), and curved MPR which enables user to inspect Region Of Interests (ROI), and place labelling contours, foreground, and background guiding strokes. The main functionalities and corresponding screen shots are listed in Table 3.2.

SVMIST stores the DICOM images in a fold-based file system database, which is organised according to patient ID, study ID, and series ID in a hierarchical manner (see Fig. 3.7 and Fig. 3.9). The database configuration contains the file locations of all items in the hierarchical

Table 3.1: The technical implementation detail of interactive segmentation software *SVMIST*.

Platform	Mac OS X
Programming Language	Object-C & C++
Graphical User Interface	Cocoa Framework
Database Format	XML-based Markup
Third-Part Library	The Visualization Toolkit (VTK) http://www.vtk.org/
	Segmentation & Registration Toolkit (ITK) http://www.itk.org/
	Grassroots DICOM Library (GDCM) http://gdcm.sourceforge.net/
	DCMTK Library http://dicom.offis.de/
	OpenCV Library http://www.opencv.org/
	Intel Threading Building Blocks (TBB) http://www.threadingbuildingblocks.org/

tree, and the contents are serialized onto local storage media in XML format. Fig. 3.8 shows an example of database configuration file. Given a selected item on the database GUI, the meta information of DICOM sequence is loaded on the fly, and displayed in sliding pane on the right. The snapshots of all DICOM images in the selected sequence are shown slice by slice in the bottom pane with a sliding navigation control, where the auto-play option is available for users to view in a movie mode (see Fig. 3.10). In Fig. 3.11 and Fig. 3.12, the 3D MPR viewer is created when the corresponding function is toggled, where the 3D MPR viewer GUI contains four 2D viewer pane, and one 3D viewer pane. The 2D viewers not only offer image projections from sagital, axial, and coronal views given a rotatable orthogonal coordinate system (see Fig. 3.11), but also provide 2D projection from a curved surface in 3D (see Fig. 3.12 and Fig. 3.13). Rotation and translation of orthogonal coordinate system is operated in 3D scene viewer on the top right pane, which offers an intuitive viewing experience compared to most of popular DICOM viewers, such as *OsiriX*. For curved MPR, the projection surface is constructed via selecting surface control points from 2D viewer, where the surface with the sampled image is overlaid onto 3D viewer (see Fig. 3.13). *SVMIST* provides 2D annotation functions, which allows creating open or close contour and regions using both polynomial and BSpline interpolations, where the annotations can also be visualised on the 3D viewer (see Fig. 3.14). The annotations created by the user together with the information of coordinate

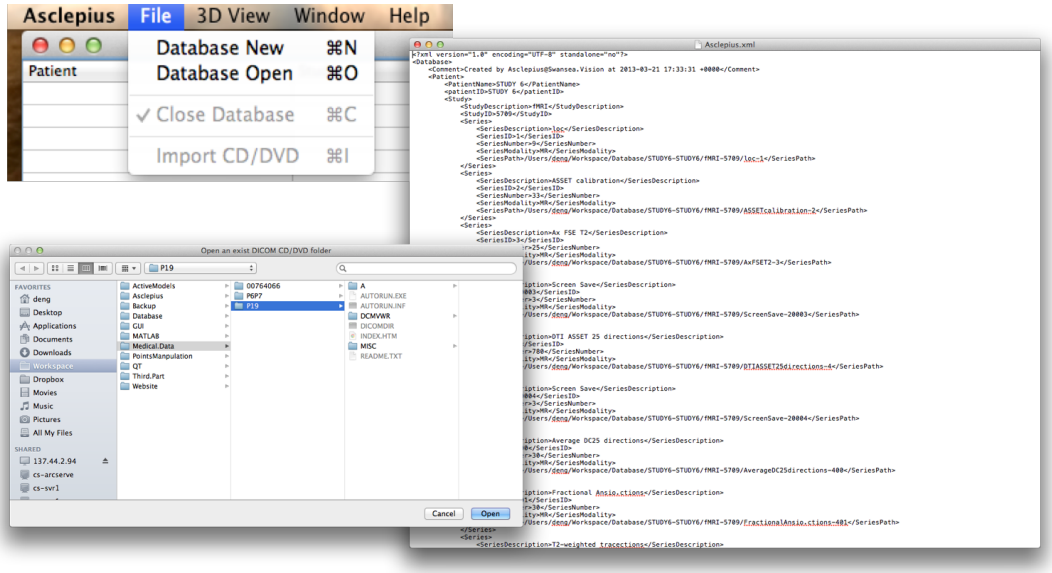


Figure 3.7: To create or open a DICOM database at a specific storage location.

system can be exported in XML format, and loaded back whenever needed. Fig. 3.15 shows an example of XML file for storing the annotations. Meanwhile, the user input strokes for guiding the interactive segmentation is implemented as a special case of open contour, which offers adequate flexibility for the users to communicate with the software.

Given *SVMIST*, the typical interactive segmentation processes are as follows:

1. Create or open an *SVMIST* database from a specified local folder;
2. Import the 3D DICOM image sequence to the created or opened database;
3. View the imported sequence using 3D or curved MPR;
4. Provide foreground and background strokes by annotating the 2D images;
5. Train a binary classifier (Random Forests) using the supervised information given by the user;
6. Classify the whole volume using the trained classifier;
7. Refine the classifier by providing more strokes until the user is satisfied with the classification result;
8. Obtain the segmentation result by applying the regularization scheme (MRF) to the classification result;
9. Apply some post processes to refine the segmentation.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Database>
    <Comment>Created by Asclepius@Swansea.Vision at 2014-05-07 09:17:15 +0000</Comment>
    <Patient>
        <PatientName>TAVI 3^CT </PatientName>
        <patientID>AW1287029794.339.1372764774 </patientID>
        <Study>
            <StudyDescription>CT Cardiac angiogram coronary </StudyDescription>
            <StudyID></StudyID>
            <Series>
                <SeriesDescription>HR 65BPM OR LESS</SeriesDescription>
                <SeriesID>2 </SeriesID>
                <SeriesNumber>501</SeriesNumber>
                <SeriesModality>CT</SeriesModality>
                <SeriesPath>/Users/deng/Desktop/Demo/TAVI3^CT-AW1287029794.339.1372764774/CTCardiacangiogramcoronary-/HR65BPMORLESS-2</SeriesPath>
            </Series>
        </Study>
    </Patient>
    <Patient>
        <PatientName>STUDY 1 </PatientName>
        <patientID>STUDY 1 </patientID>
        <Study>
            <StudyDescription>fMRI BRAIN</StudyDescription>
            <StudyID>5638</StudyID>
            <Series>
                <SeriesDescription>FSPGR BRAVO </SeriesDescription>
                <SeriesID>5 </SeriesID>
                <SeriesNumber>120</SeriesNumber>
                <SeriesModality>MR</SeriesModality>
                <SeriesPath>/Users/deng/Desktop/Demo/STUDY1-STUDY1/fMRIBRAIN-5638/FSPGRBRAVO-5</SeriesPath>
            </Series>
        </Study>
    </Patient>
    <Patient>
        <PatientName>FFR 14</PatientName>
        <patientID>AW1109887332.768.1372427767 </patientID>
        <Study>
            <StudyDescription></StudyDescription>
            <StudyID></StudyID>
            <Series>
                <SeriesDescription>SS Pulse Hi-Res 50%ASIR VOL </SeriesDescription>
                <SeriesID>4 </SeriesID>
                <SeriesNumber>224</SeriesNumber>
                <SeriesModality>CT</SeriesModality>
                <SeriesPath>/Users/deng/Desktop/Demo/FFR14-AW1109887332.768.1372427767/-/SSPulseHi-Res50%ASIRVOL-4</SeriesPath>
            </Series>
        </Study>
    </Patient>
</Database>
```

Figure 3.8: An example of database configuration XML file of *SVMIST*.

3.3.2 Segmentation Result

The method is evaluated on the clinical CTA volumes. The volumes have different number of slices with 0.65mm inter-slice spacing, each slice has 512×512 pixels with 0.38mm intra-slice spacing. At first, we cropped out the region of interest which contains the whole heart and a part of region in the lung, then followed by 10-iteration vessel enhancing diffusion filtering. The features for every vertex in the sub-volume are computed in multi-scale spaces with $\sigma \in \{0,1,2,3,4\}$, which results in a 85-components feature vector. An RF classifier is interactively trained on the training set which provided by sampling the vertexs from the user's foreground and background strokes. Fig. 3.16 shows the out-of-bag error against the number of decision trees that are used to build the RF. It can be observed that the out-of-bag error steadily converged at 200. Therefore, we empirically set the number of decision trees to 200

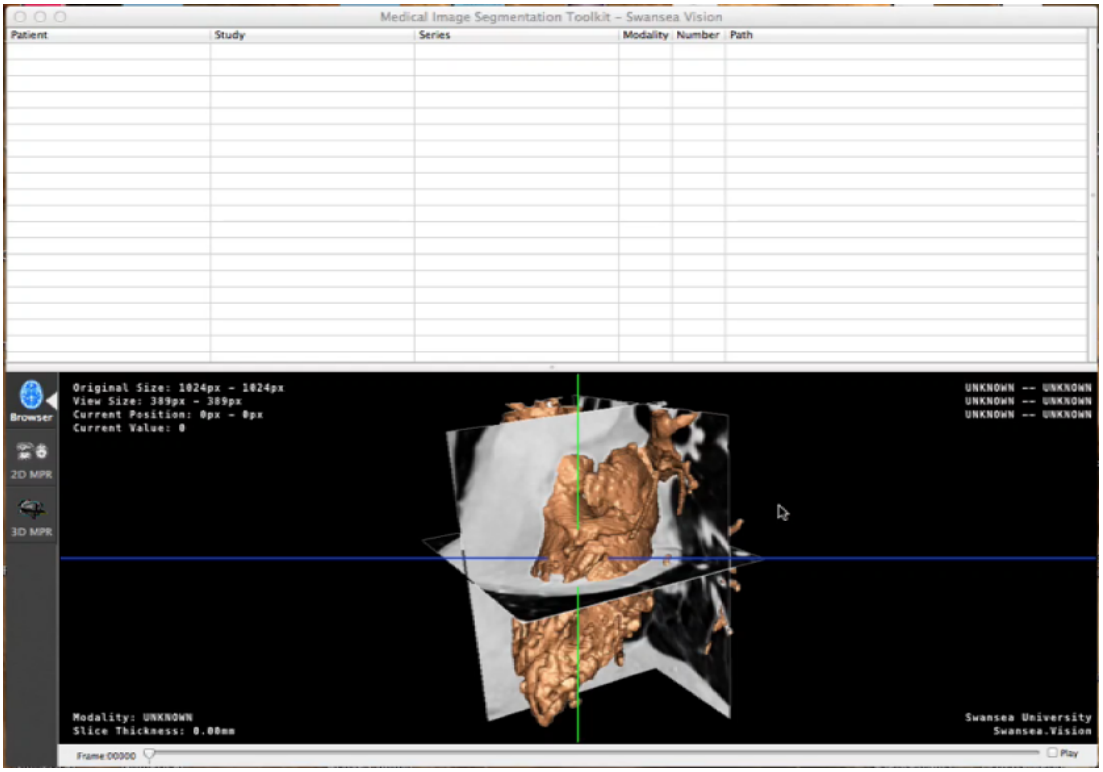


Figure 3.9: The database management GUI of *SVMIST*.

for all iterations. The sub-volume is segmented by optimising the classification result given by the RF using grid-like Markov random field model with 6 neighbours system and PD1 algorithm. Given the binary volume, the segmented result is visualized as iso-surface using marching cube algorithm. Once the user adds more strokes, we repeat the classifier training, sub-volume classification, label optimisation and result rendering processes, until no more user stroke is detected, and the final segmentation is achieved. A connected component analysis was also carried out to remove isolated, small regions.

Fig. 3.17, Fig. 3.18, and 3.19 provide examples of the segmentation process and result. In Fig. 3.17, we show the iso-surface rendering of the vascular structures, including the ventricles. It is clear from this that coronary vessels are only a small proportion of those structures. To isolate them and to obtain a coherent structure with good connectivity is not a trivial task. Fig. 3.19 (a) shows the classification result from RF classifier. Most of the non-coronary structures are removed, but there are still plenty of isolated thin, tubular structures. The final result of the proposed method is shown in Fig. 3.19 (b), where the coronary structures are well seg-

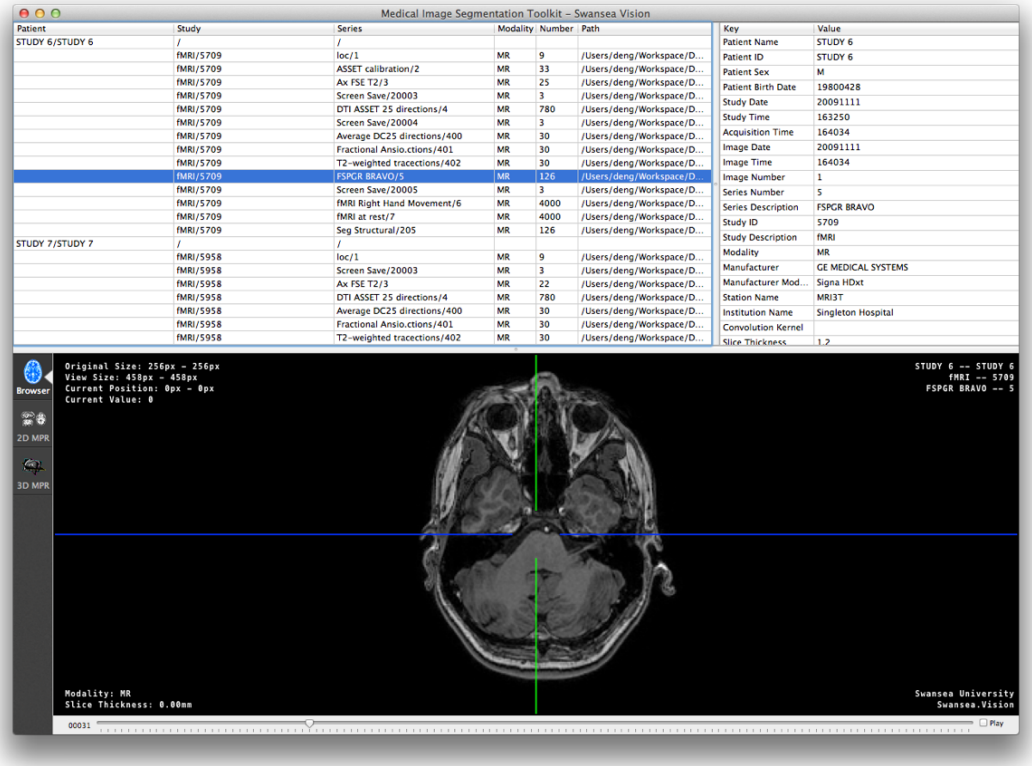


Figure 3.10: The snapshots and meta information of DICOM images are loaded when a sequence is selected.

mented. Two further examples are provided in Fig. 3.20. The examples provided here are typical results we achieve using the proposed method. The user interactions are minimal, i.e. only a few strokes on the foreground and background.

3.4 Summary

Coronary artery segmentation plays a vital important role in coronary disease diagnosis and treatment. In this chapter, we present a machine learning based interactive coronary artery segmentation method for 3D CTA images. We first apply vessel diffusion to reduce noise interference and enhance the tubular structures in the images. A few user strokes are required to specify region of interest and background. Various image features for detecting the coronary arteries are then extracted in a multi-scale fashion, and are fed into a random forests classifier,

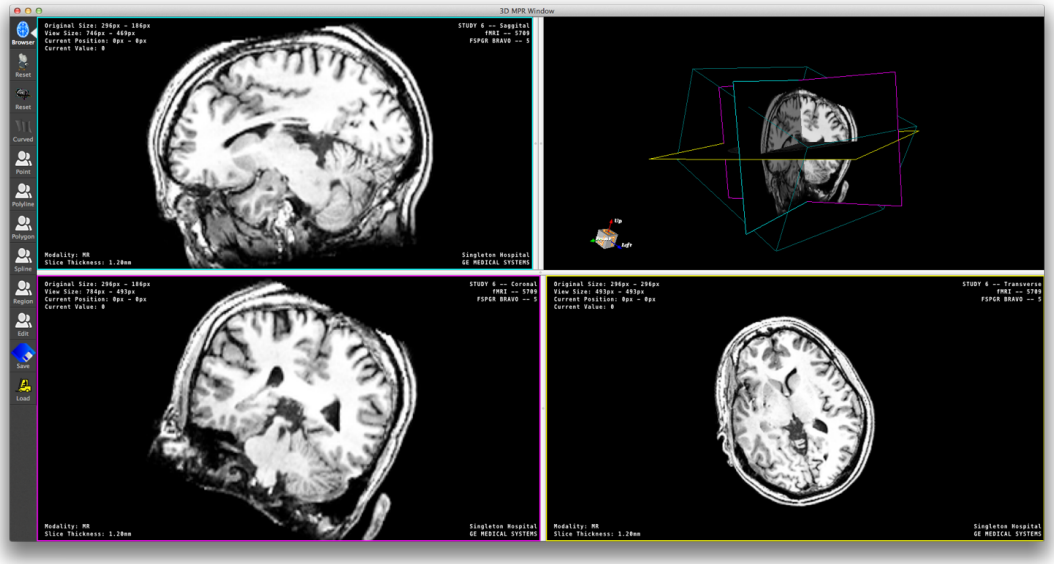


Figure 3.11: Investigate the volumetric image using orthogonal MPR.

Table 3.2: The main functionalities of interactive segmentation software *SVMIST*.

Figure	**Detail**
Fig. 3.7	Create and open a DICOM database.
Fig. 3.9 & 3.8	The GUI of database management, and an example of database configuration file in XML format.
Fig. 3.10	Load meta information and slice snapshots of the selected DICOM sequence.
Fig. 3.11	Visualize the volumetric image using orthogonal MPR.
Fig. 3.12 & 3.13	Visualize the volumetric image using curved MPR by interactively constructing the projection surface.
Fig. 3.14 & 3.15	Annotations can be created using open or close contours, and an example of exported annotation file in XML format

which assigns each voxel with probability values of being coronary artery and background. The final segmentation is carried through an MRF based optimisation using primal dual algorithm. A connectivity component analysis is carried out as post processing to remove isolated, small regions to produce the segmented coronary arterial vessels. The proposed method requires limited user intervention and achieves robust segmentation results. A segmentation software namely *SVMIST* was developed, and promising segmentation results are achieved with just a few user strokes. The clinicians evaluated segmentation results which are considered to be

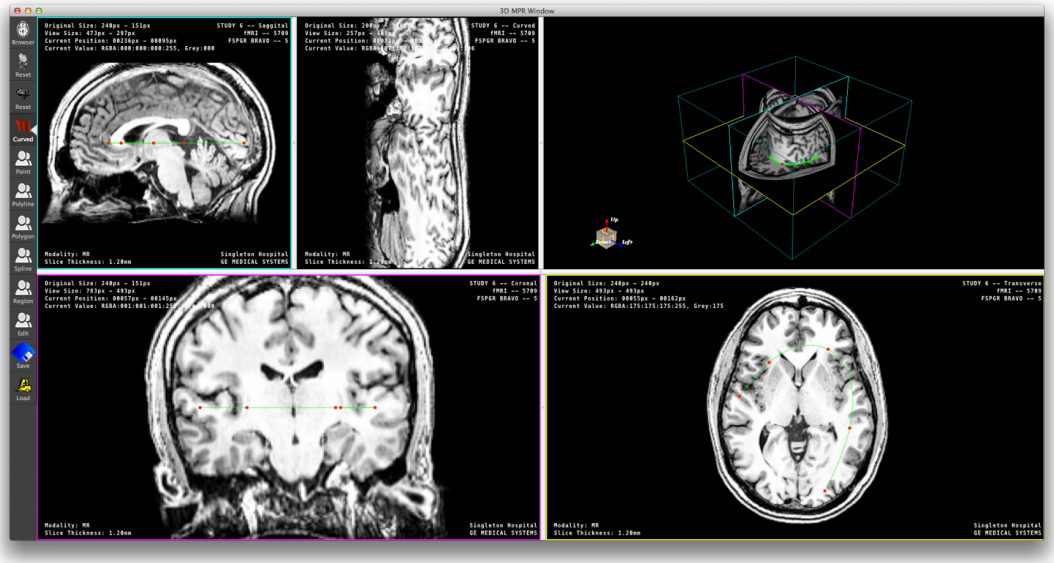


Figure 3.12: Create curved MPR via clicking surface control points on a 2D viewer pane.

consistent with the real anatomical structures. In addition, the segmented geometries were used to calculate fractional flow reserve in the blood vessel with a reduced-order model, which suggests that our approach can be used as a part of a broader risk assessment tool that aims at increasing the diagnostic yield of cardiac catheterisation for in-hospital evaluation of significant stenoses. This part of work has been published in the following journal paper.

- E. Boileau, S. Pant, C. Roobottom, I. Sazonov, J. Deng, X. Xie, and P. Nithiarasu, Estimating the Accuracy of a Reduced-Order Model for the Calculation of Fractional Flow Reserve (FFR). International Journal for Numerical Methods in Biomedical Engineering, 2017

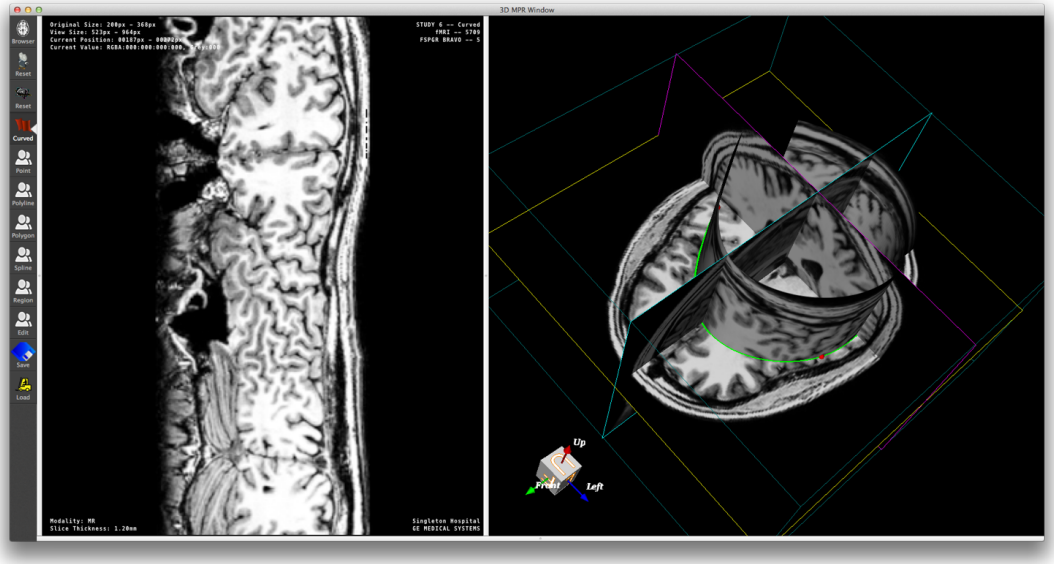


Figure 3.13: Visualize the curved MPR surface on the 3D viewer pane on the right, and the projection image is shown in the 2D viewer pane on the left.

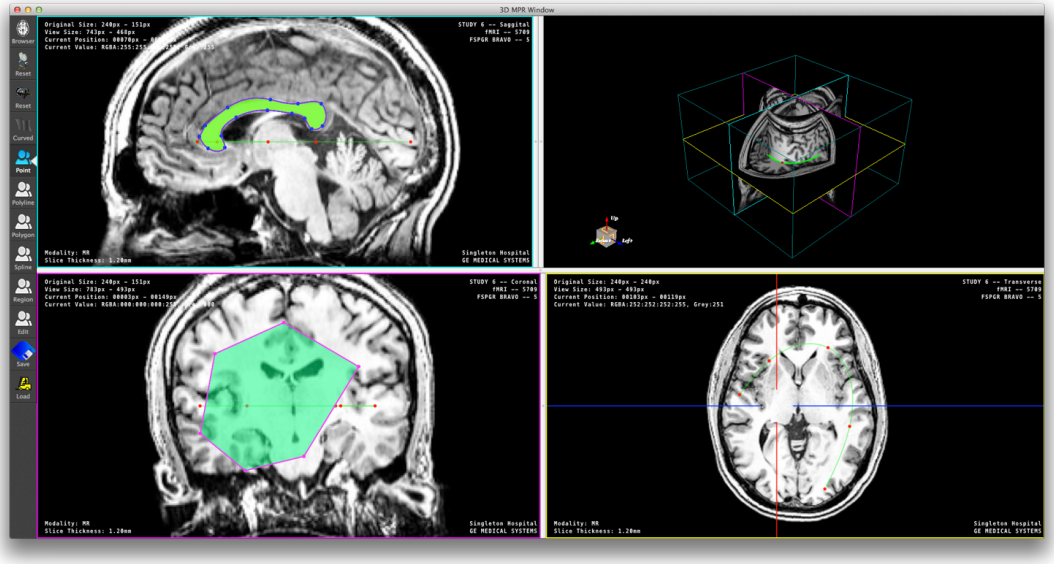


Figure 3.14: The examples of user strokes.

```
<UserInput>
    <PointSet Type="2" ID="0" isOpenSet="YES" StrokeType="1">
        <Points ID="0">108.06:138.81:55.80</Points>
        <Points ID="1">106.83:129.89:55.80</Points>
        <Points ID="2">103.93:123.17:55.80</Points>
        <Points ID="3">99.86:113.87:55.80</Points>
        <Points ID="4">97.34:106.07:55.80</Points>
        <Points ID="5">95.92:101.73:55.80</Points>
    </PointSet>
    <PlaneNormal>
        <SaggitalPlane X="1.000000" Y="0.000000" Z="0.000000"></SaggitalPlane>
        <CoronalPlane X="0.000000" Y="1.000000" Z="0.000000"></CoronalPlane>
        <TransversePlane X="0.000000" Y="0.000000" Z="1.000000"></TransversePlane>
        <SaggitalPlane_P1 X="119.531250" Y="239.531250" Z="-0.599998"></SaggitalPlane_P1>
        <SaggitalPlane_P2 X="119.531250" Y="-0.468750" Z="143.399635"></SaggitalPlane_P2>
        <SaggitalPlane_Or X="119.531250" Y="-0.468750" Z="-0.599998"></SaggitalPlane_Or>
        <CoronalPlane_P1 X="-0.468750" Y="119.531250" Z="143.399635"></CoronalPlane_P1>
        <CoronalPlane_P2 X="239.531250" Y="119.531250" Z="-0.599998"></CoronalPlane_P2>
        <CoronalPlane_Or X="-0.468750" Y="119.531250" Z="-0.599998"></CoronalPlane_Or>
        <TransversePlane_P1 X="239.531250" Y="-0.468750" Z="113.399712"></TransversePlane_P1>
        <TransversePlane_P2 X="-0.468750" Y="239.531250" Z="113.399712"></TransversePlane_P2>
        <TransversePlane_Or X="-0.468750" Y="-0.468750" Z="113.399712"></TransversePlane_Or>
    </PlaneNormal>
</UserInput>
```

Figure 3.15: An example of exported labelling XML file of *SVMIST*.

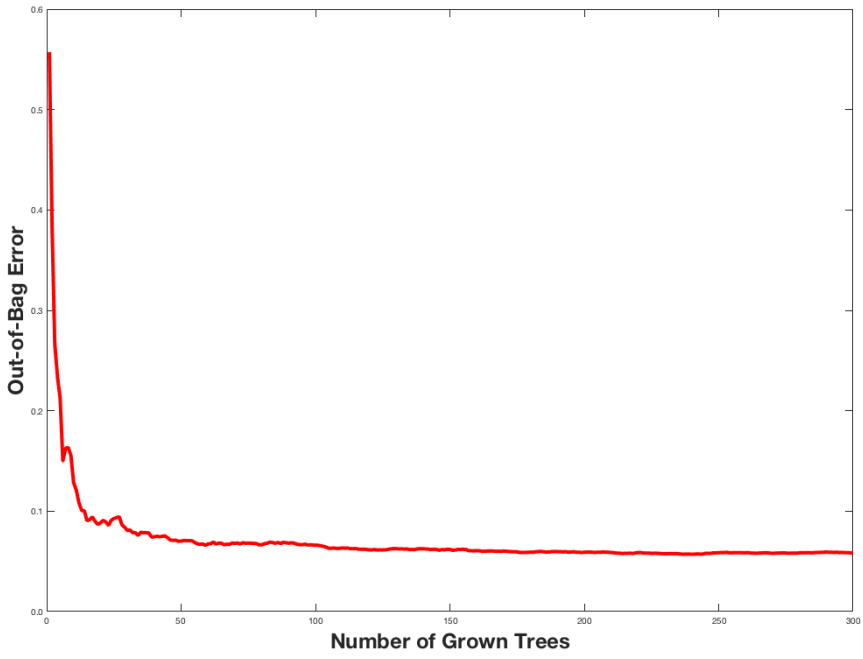


Figure 3.16: The out-of-bag error of different number of grown trees.

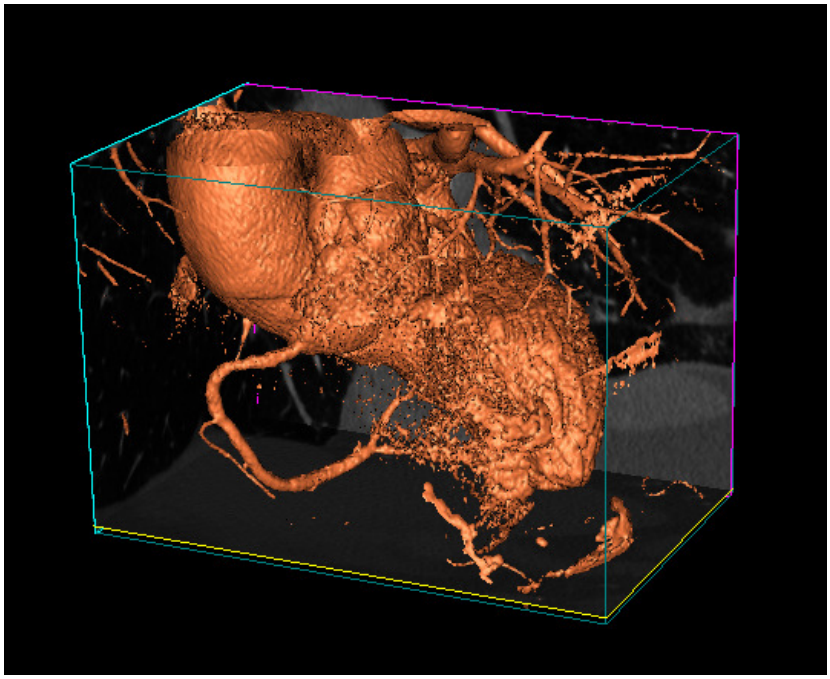


Figure 3.17: The iso-surface rendering of the CTA image.

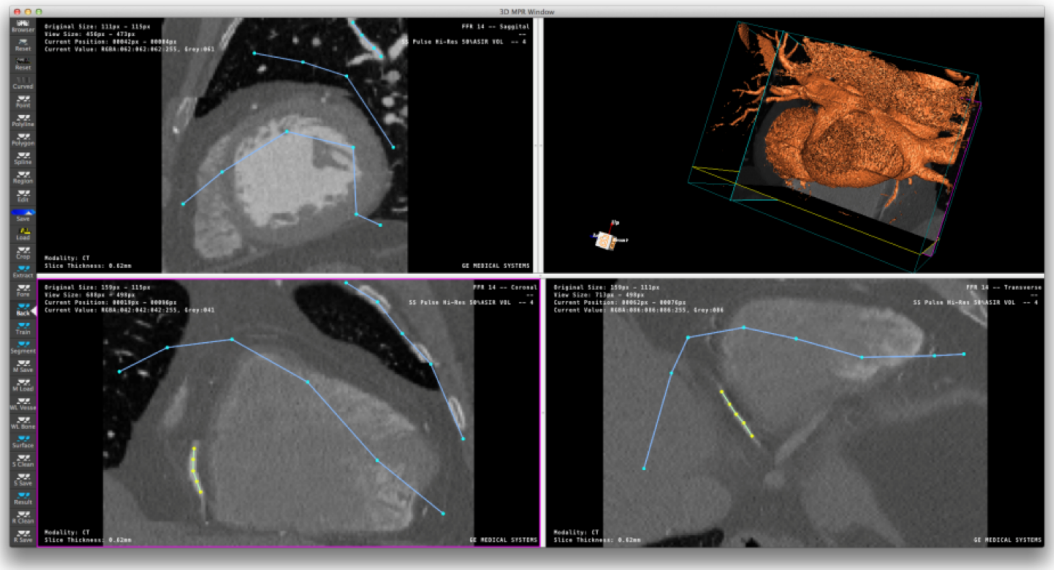
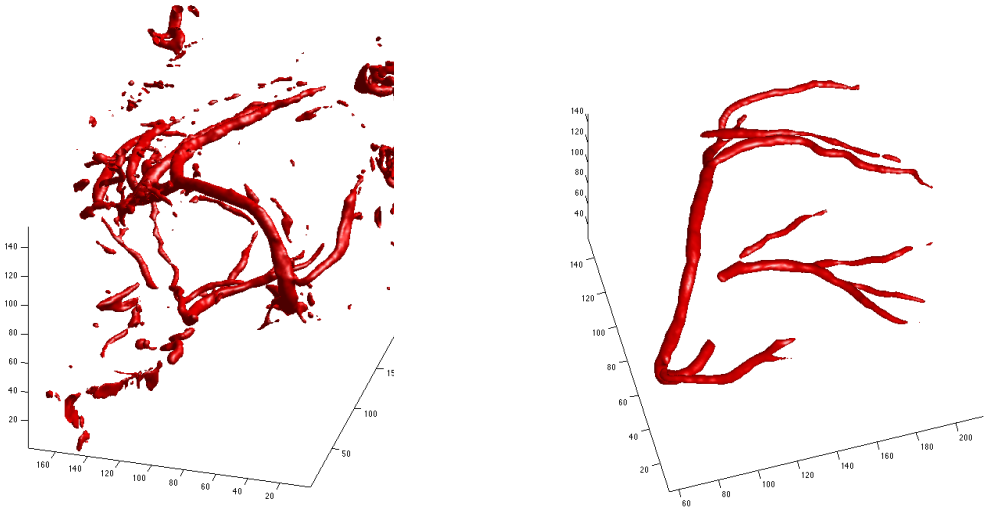
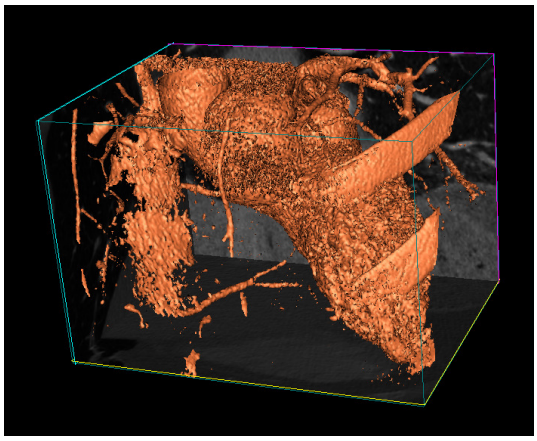


Figure 3.18: The examples of user provided strokes. (blue: background strokes; yellow: foreground strokes; dot: control points of user strokes.)

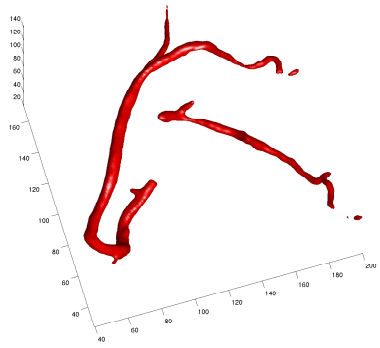


(a) (b)

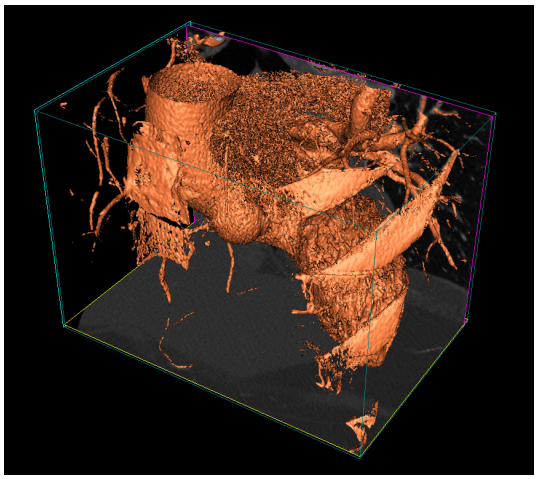
Figure 3.19: The examples of interactive segmentation process: (a) RF-based voxel classification result; (b) final segmentation result of the proposed method.



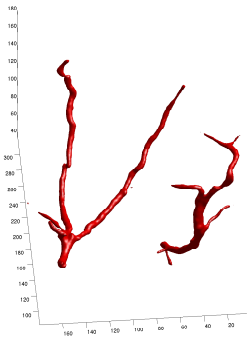
(a)



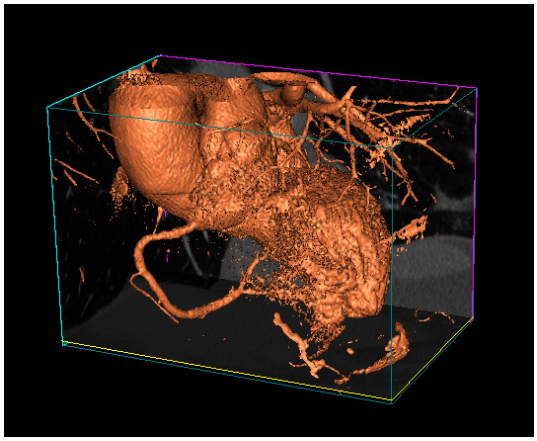
(b)



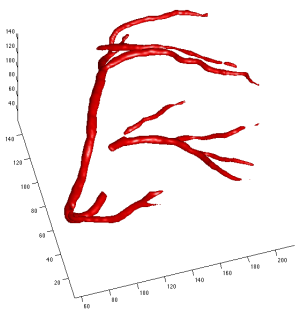
(c)



(d)



(e)



(f)

Figure 3.20: The examples of interactive segmentation results: (a), (c) and (e) iso-surfaces rendering of the original CTA images; (b), (d) and (f) final segmentation results of the proposed method.

Chapter 4

Aorta Segmentation

Contents

4.1 Introduction **66**
4.2 Proposed Method **72**
4.2.1 Overview 72
4.2.2 Intensity-based Naive-Bayesian Detector 73
4.2.3 Pseudo-3D CNN Detector 74
4.2.4 Localised Interactive Refining 78
4.2.5 Non-Uniform Implicit B-spline Surface 79
4.2.6 Segmentation as Region-based Deformation 84
4.3 Evaluation **85**
4.3.1 3D CTA Dataset 85
4.3.2 Experimental Result 86
4.3.3 Speed Discussion 94
4.4 Conclusion **94**

In the previous chapter we proposed an interactive segmentation method, where an off-line RF model was repeatedly trained using the accumulated foreground and background strokes acquired from user in an interactive fashion. There are several limitations that prevent it being an efficiently segmentation method for other complex anatomies. First, it requires hand-crafting discriminative features for the anatomy to be segmented, whereas representative features are often not available, and it is very time-consuming to design them by hand. Second, during each

round of interaction, the RF model is re-trained completely using all collected training samples. Generally the off-line model converges to better locally optimal solutions compared to the on-line model, however, it is computationally inefficient for large training sets and complex learning models [119]. Third, it has no geometrical representation for the segmented object, which makes it rather difficult to incorporate with topology analysis, shape prior or structure manipulation. In order to address the issues that we discussed above, in this chapter, we present a novel interactive segmentation method that is built upon an adaptive learning scheme which utilises CNN based cascade detector and an implicit parametric shape representation.

4.1 Introduction

Interactive image segmentation plays a vital important role in computer vision, graphics, and medical image analysis especially, where user intervention as additional source of information for guiding the process is incorporated with procedural segmentation to form a semi-automatic method. An overview of interactive medical image segmentation can be found in [67]. Interactive scheme bridges expert knowledge of user required on the fly with fully-automatic method, and generally produces more accurate segmentation result than using non-interactive method alone. Many traditional segmentation methods, such as active contour models, region growing based models, and statistical models, are capable of combining with interactive schemes, where user interactions are used in diverse ways, such as geometrical initialization, foreground-background clue, parameter tuning, or correction for miss-segmentation. Especially, interactive segmentation with statistical model has proved to be an efficient strategy for image segmentation over full-automatic approaches in differentiating foreground and background, where the supervised interaction from user is directly applied to the image to be segmented. The user interaction is considered as supervision labels to form either region based texture model, or edge based boundary model to partition the image into a number of subregions [120, 121, 122, 123]. The segmentation can then be obtained via solving a either combinational or continuous optimization problem, where the user interaction also can be used as regularization constraints [124, 125]. Semi-automatic methods are also able to be extended to segment higher dimensional data, i.e. 3D volumetric image, where a computational efficiency, and a flexible and natural user interaction scheme are particularly in need. It is worth noting that these semi-automatic schemes still heavily rely on the traditional fully-automatic methods. Hence, the challenges of designing interactive segmentation method can be summarized into

four fundamental problems as follows: delinearating foreground object from background, user interaction scheme, shape representation and segmentation regularization.

An ideal interactive segmentation method provides accurate segmentation results with minimum user interaction effort given a natural and friendly interface. A list of related interactive image segmentation methods is shown in Table. 4.1. Boykov *et al.* [126] proposed a general interactive segmentation method, *Interactive Graph Cuts* for N-dimensional images, where the user marks certain strokes as object and background to guide the semi-automatic partition process. The delinearating model is built as two grey-level distributions for object and background respectively using intensity histograms obtained from user interactions. The segmentation is equivalent to a binary labelling problem given the data support and assuming local smoothness of pixel intensity. It is achieved by solving a discrete energy minimization using *Graph Cut* method using standard minimum cut algorithm, where a globally optimal solution can be found for binary segmentation. Based on the original *Interactive Graph Cuts* method, Rother *et al.* proposed a so-called *GrabCut* [120] for foreground and background segmentation in still images using RGB color space. An initialization ROI bounding box is provided by user, where the texture appearances of foreground and background are modelled and delinearated using two separate Gaussian Mixture Models (GMMs). The piecewise constant is then imposed by solving the same Gibbs energy minimization problem as proposed in [126] using minimum cut. The segmentation is achieved via a periodical procedure, where the method interactively updates the delinearating model based on the border matting provided by user, and assigns all pixels with optimal labels that minimize the energy objective function. Such interactive segmentation procedure terminates when satisfying results are achieved. Han *et al.* [127] extended the *GrabCut* by introducing multi-scale non-linear structure tensor texture feature to overcome the difficulty of delinearating the scale difference of textured image. Given a few user marked foreground and background lines, *Lazy Snapping* [128] builds two K-Means models to partition the image into many small pre-segmented regions based on the color similarity. Therefore, the segmentation can be obtained by formulating as a binary labelling problem for pre-segmented blocks using graph cut. Unger *et al.* [122] showed that such interactive framework can be incorporated with total variation regularization, which solves a minimization of the geodesic active contour energy. It is clearly that the interactive segmentation methods fall into an uniform framework with a series of periodical processes as follows: acquiring object and background clues from user, image delinearating and segments regularization. Therefore, in order to boost

the efficiency of method, stronger delinearating models with more discriminative features were proposed, such as naive Bayesian classifier with geodesic distance features [129], and RF with arbitrary features [123]. Most recently, Feng *et al.* [130] showed the feasibility of applying such interactive segmentation method to RGBD images.

Table 4.1: Related Interactive Image Segmentation Methods (F-B: Foreground-Background, ROI: Region Of Interest, MRF: Markov Random Fields.)

Method	Features	Delinearating	Regularization	Interactive Scheme
[126]	Grey Texture	Histogram	Graph Cut	F-B Stroke
[120]	Color Texture	GMMs	Graph Cut	ROI Box, Border Matting
[128]	Color Texture, Image Gradient	K-Means	Graph Cut	F-B Stroke, Boundary Editing
[122]	Color Texture	Histogram	Total Variation	ROI Box, Border Matting
[127]	Color Texture, MSNST	GMMs	Graph Cut	ROI Box
[123]	Color Texture	Random Forests	Total Variation	F-B Stroke
[129]	Color Texture, Geodesic Distance	Guass Naive Baysian	Graph Cut	F-B Stroke
[130]	Geodesic Distance	Naive Baysian	MRF	F-B Stroke
Our	**Texture, Self-Learned Feature**	**Naive Baysian, CNN**	**Total Variation**	**F-B Stroke, Locally Refining**

Adaptive learning is an efficient strategy for constructing interactive image segmentation methods which can be used to build classifiers to differentiate foreground and background for region based segmentation. In contrast to the traditional machine learning based fully-automatic segmentation method, interactive segmentation requires the supervision interactions from user on the fly. These interactions are directly applied to an image that is to be segmented via adaptively tuning the classifiers. The user interactions are considered as supervision labels to form either a region based texture model or edge based boundary model, partitioning the image into a number of sub-regions [120, 121, 122, 123]. The segmentation can then be obtained via solving either a combinational or a continuous optimization problem, where the user interaction also can be used as regularisation constraints [124, 125]. Statistical models are often built from the user interaction, and then used to delinearate the foreground objects and background regions [123]. However, the discriminative features are hand-crafted for a specific structure, and user inputs are relatively simple and often biased, which generally leads to failure in learning the right decision boundary for predicting the foreground and background locations. For example, in Chapter 3, the features for the coronary artery segmentation are extracted from the Hessian matrix and the intensities of local neighbourhood at multi-scale. These features are designed for representing the small tubular-like vessels that attach to the myocardium, hence, they are informative for differentiating the coronary artery and other tissues, however they are not suitable for segmenting the aorta root and arch. Recently, deep learning based methods are becoming more mainstream [35, 37], as it has been found superior over many traditional

methods for visual recognition tasks. An overview on deep learning method in medical image segmentation can be found in [66]. By using deep CNN models, the features are automatically learnt through a supervised classification training process, where the low-level features can be further generalised by stacking multiple convolutional layers, and the decision boundary is also learnt jointly. Backwards Propagation of Errors (Back-Prop) with mini-batch based gradient descent is often used to train the CNN model. Therefore, on-line learning schemes can be developed naturally via fine-tuning the pre-learnt model with the training data batch acquired from user interaction. However, training a deep model requires a large amount of supervision data that are generally not available in the scene of interactive segmentation, which more likely results in an under-fitted model.

Table 4.2: Parametric Implicit Representation for Surface Reconstruction and Image Segmentation

Param.	Support	Method
Polynomial	Global	Polynomial Kernel [131]
	Local	B-splines Kernel [132, 133, 134]
RBF	Global	Thin-Plate [135], Gaus [136], Multi-Quadric [137]
	Local	Wendland's RBF [138]

Implicit functions that are widely used in segmentation provide smooth and topologically flexible shape representations, where the surface of a shape is embedded into a zero level set that is able to deform and visualize naturally [72, 139]. Compared to the parametric models [68, 71], the geometric models have more topological flexibility as the shape is embedded in higher dimensional space which can break, merge and vanish naturally during the level set function evolution that is driven by a time-dependant PDE. The high dimensional implicit function can be approximated using parametric form, which is so-called Parametric Implicit Representation (PIR). During the level set evolution, for segmenting volumetric data in particular, shape or flat gradients may be developed, which requires re-initialisation to avoid inaccuracy numerical approximation. PIR approximates the level set function via interpolating its parametric formulation, which avoids developing the numerical errors, and a more concise Ordinary Differential Equation (ODE) solution is often available. PIR can be broadly divided into a polynomial based approach, or a Radial Basis Function (RBF) based approach, both of which can be further categorised into globally support method and locally support method considering the kernel function that is used. Table 4.2 lists some representative PIR approaches that are used for surface reconstruction and image segmentation. The idea of ap-

plying PIR to image segmentation was first proposed by Morse *et al.* [135] in 2005, where a continuous representation of the level set function is parametrised using globally supported Thin-Plate RBF. The deformation is driven by an external image and balloon forces to move the locations of zero RBF constraints towards the boundary of the object. However, this is an incomplete solution, as the locations of RBF centres are updated during each iteration while their coefficients are fixed, where periodical interpolation is required to re-initialise the implicit function in order to ensure the functional continuity. Xie *et al.* [137] overcame this numerical intractability by introducing fixed location RBF centres. The formulation of coefficient based deformation can then be derived, where the level set PDE problem is converted to an ODE problem, and re-initialisation is no longer needed. Paiement *et al.* [136] showed such a strategy is able to solve the segmentation and interpolation problems jointly when the image data is partially missing. However, computational complexity is the major limitation of the globally supported RBF approaches [135, 136, 137], which is the same case for polynomial fitting [131], as those methods generally involve decomposing a large and dense kernel matrix that is computationally expensive operation. Gelas *et al.* [138] introduced compactly support RBF (Wendland's RBF [140]), where a sparse linear system is obtained. According to [141], the computational complexity of sparse matrix factorization can be considered to be $O(N_{nzf})$ where N_{nzf} is the number of non-zero factors. This is much simpler compared to the dense formulation of $O(N \log N)$. Bernard *et al.* [132] proposed variational B-spline level set model that approximates the level set function using a number of B-spline basis. It shows that the parametric representation can be deformed as a sequential 1-D convolution. Rouhani *et al.* proposed an Implicit B-spline Surface (IBS) based reconstruction method that recovers shape from point cloud [133], and showed its feasibility of solving a shape registration problem [134]. However, the over-smoothing issue is one of the major drawbacks of the interpolation process where unnecessary loss of geometrical detail is inevitable.

In this chapter, we present a novel volumetric image segmentation method that bridges state-of-the-art deep learning based object detection approach, and deformable parametric implicit shape representation. First, to combine the deep model with an interactive scheme, we propose a two-stage cascade detector that contains a Naive-Bayesian classifier for fast elimination, and a pseudo-3D CNN classifier for precise detection. Instead of learning the model from scratch using the interaction, the pseudo-3D CNN is trained on pre-built dataset, such that the discriminative features are first learnt using sufficient data, which completely avoids

hand-crafting features. Then, user interaction is only used to fine-tune the pre-learnt model. The refined model is used to correct the miss-classified regions that are selected, again based on user interactions, such that the effect of any biased model is imprisoned only in the local regions marked for refining. Second, it is worth noting that the geometrical complexity of an individual object has not been considered by the previous PIR approaches, whereas our proposed NU-IBS method measures the complexity density of the local topology using scale weighted wavelet coefficients. It is able to adapt the density and supporting radius of individual B-splines, where dense and compact bases are placed at the regions that have more subtle structures. The level set PDE is transformed to an ODE problem, where the formulation of coefficient deformation is derived using the region based velocity function. In addition, the proposed method is equipped with an efficient foreground detector, where the segmentation no longer relies on shape edge, homogeneous region, or good initialisation. It is able to distinguish the object from complex background, which is more practical for real-world applications. Our contributions are fourfold.

- To efficiently delinearate the foreground objects and background regions, a cascade detector is proposed which contains an intensity-based Naive-Bayesian classifier for fast elimination, and a pseudo-3D CNN classifier for precise classification. The representative features for region-based detection are automatically learnt in a supervised fashion together with the decision boundary for binary classification, no hand-feature-crafting is needed.The use of pseudo-3D CNN avoids 3D convolution over the volumetric data via aggregating 2D convolutional features that extracted from 3 orthogonal planes at multi-scales, which is a more computationally economical scheme, and makes no sacrifice in accuracy.

- An adaptive learning and localised refining strategy is proposed which further improves the detection result and boosts accuracy with help of user interactions that are taken on-the-fly. It requires minimum effort from user to provide foreground and background guiding strokes interactively, where the supervision information is used to adaptively update the classifier, and the geometrical information is used to localise the regions that need to be refined. The method is able to compensate the case by case variations that are not pre-learnt by the detector, meanwhile avoid the outliers contaminating the well-modelled common patterns.

- We proposed a novel shape representation method, NU-IBS. It embeds the shape into the zero manifold of a level set function that is approximated using locally supported B-spline patches in parametric form. In contrast to the uniform knot distribution, the geometrical complexity is estimated using the proposed wavelet-based filtering method, and the control knots are placed according to the complexity density. It is able to adapt according to the local topology, where highly curved regions are blended using more compact patches to avoid over-smoothing or adding unnecessary knots.

- Piecewise constant is the most common and successful regularisation scheme widely used for image de-noising, restoration and segmentation problems, which assumes that the appearance of an image or the geometrical structure of an object are locally homogeneous. The regularisations are generally imposed through adding homogeneity measurements to the objective function that penalise the discrepancies. In contrast to the traditional approaches, we impose the piecewise constant on the classification results given by the proposed cascade detector through NU-IBS that ensures geometrical smoothness naturally. The region based deformation scheme are derived from level set PDE which iteratively propagates a smooth interface according to data support, where both geometrical and characteristic homogeneity are co-optimised, and a optimal solution to the joint object is achieved.

The rest of the chapter is organized as follows. The proposed methods including cascade detection, NU-IBS and region based deformation are introduced in Section 4.2. The evaluation is performed on segmenting aorta root and arch given a 3D CTA image dataset, with details of dataset and experimental results presented in Section 4.3. The concluding remarks are provided in Section 4.4.

4.2 Proposed Method

4.2.1 Overview

The goal of the proposed method is to segment the target object from a volumetric data provided limited user interaction, i.e. the strokes that indicate the foreground and background regions. The delinearated model is initially learnt given a set of labelled training images, where the foreground and background ground-truth is provided in the form of binary volume. The

segmentation task can be completed by a *Classification-Refining-Regularising* procedure in an interactive manner as follows: (1) detect the object via voxel-wise region classification; (2) interactively refine the predicted region using adaptive learning scheme; (3) regularise the results with a piecewise constant constraint that uses an NU-IBS model for shape representation.

From a machine learning perspective, detecting an object is equivalent to a binary classification problem, which groups individual voxel into foreground objects or background regions. Voxel-wise classification on volumetric images is a computational expensive task due to the large number of hypotheses, and feature variations compared to lower dimension data. When more discriminative but complex models were used, the speed of segmentation procedure is reduced dramatically. In order to boost the runtime performance, a 2-stage cascade detector is used to leverage the overall classification accuracy and speed efficiency, where a simple Naive-Bayesian model was trained based on the intensity information for fast background voxel elimination, and a stronger pseudo-3D CNN multi-scale detector was built to precisely identify the foreground objects. In addition to fully automatic voxel classification, an interactive refining scheme is introduced to boost the detection accuracy further by utilizing the information gained from user interventions, in our case, the foreground and background guiding strokes. However, it is noteworthy that voxel-wise object detection is not equivalent to binary segmentation, as it does not take any prior knowledge into consideration, such as the piecewise constant that is commonly used in the deformable segmentation. The proposed method solves this problem by introducing an NU-IBS model to represent shape geometry, where the regularisation constrain can then be imposed via region based deformation given the classification confidence of each voxel.

4.2.2 Intensity-based Naive-Bayesian Detector

Given $\mathscr{V}$, a set of $\mathbf{N}$ training voxels, each sample in $\mathscr{V}$ is a pair tuple defined as $\mathbf{v}_i \in \mathscr{V}$, and $\mathbf{v}_i := \langle \mathbf{t}_i, \mathbf{c}_i \rangle$, where i is the index of the training sample, $\mathbf{t}_i$ is a scaled integer intensity within the range of $[0,255]$, and $\mathbf{c}_i \in \{0,1\}$ is its corresponding binary category label that indicates either background ($\mathbf{c}_i = 0$), or foreground ($\mathbf{c}_i = 1$). Hence, the likelihood of background and foreground can be empirically estimated using two GMMs with K components as follows:

$$\mathscr{P}\,(t|C) = \sum_{k=1}^{K} a_k\, \mathscr{N}(t, \mu_k, \sigma_k^2), \quad C \in \{0,1\} \tag{4.1}$$

$$\mathscr{N}(t, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{4.2}$$

where the mixture weights a, means μ, and stand deviations σ of K Gaussian components can be obtained via Expectation Maximization (EM) given the observations from training dataset. In our case, the foreground and background likelihoods are equivalent to two probability density functions of GMMs, where the parameters of their Gaussian components are estimated independently on two sets of training samples from the distinct categories. Hence, the naive Bayesian classifier can be constructed via choosing an appropriate prior probability $\mathscr{P}(C)$ for each category, and then applying the Bayesian rule as follows:

$$\mathscr{P}\ (C=\mathbf{c}|t) = \frac{\mathscr{P}(C=\mathbf{c})\ \mathscr{P}(t|C=\mathbf{c})}{\mathscr{P}(t)} \tag{4.3}$$

$$\mathscr{P}\ (t) = \sum_{\mathbf{c}}^{\{0,1\}} \mathscr{P}(C=\mathbf{c})\ \mathscr{P}(t|C=\mathbf{c}) \tag{4.4}$$

There are a number of approaches to select the prior probability, such as empirical estimation of the category frequency, or inference by Maximizing A Posteriori (MAP) from given training dataset. Whereas for the stage classifier in a cascade framework, it is sensible to sacrifice the fallout rate to some extent in order to retain a high recall rate, which can be achieved by manually setting a biased prior probability that is weighted towards the foreground category. Such a strategy compensates the limitation of lacking positive evidence for general detection problem, especially for those extremely unbalanced datasets, it levers the biased data distribution to some extent. Hence, the classifier can be constructed via computing the posterior probabilities of t for all $C \in \{0,1\}$ given the prior probabilities, and then mapping t to the category label which maximises the posterior, where t is an 1D intensity value of the target voxel.

4.2.3 Pseudo-3D CNN Detector

Intensity only is not sufficient to distinguish the object from background, as 1D features lack structural information which forms variations in terms of appearance and geometry. Inspired by the commonly used MPR in radiography, a pseudo-3D CNN detector is proposed using the local image patch from three perpendicular panels at multi-scale to precisely identify the foreground object from the hypotheses retained by the first stage intensity-based naive-Bayesian detector. 2D image patches are a set of appearance projections of the original 3D geometry data from different view angles at a certain location that is within the volume. For most cases, the information from a single projection is ambiguous and biased, whereas the uncertainty can be reduced significantly when more projections are available and integrated, especially from

uncorrelated views. For example, coronal, sagittal, and axial views are perpendicular to each others, and the mutual information is minimum that can only be found on the intersecting lines. Hence, the pseudo-3D CNN learns the primitive features from those three views independently, which are then aggregated to further generalise abstract descriptors to represent foreground and background elements where a compact classification boundary can also be found. Rotations to the coordinates system can also be applied at the same time in order to obtain the best views and projections for the given anotomical structures. In addition, different to 3D CNN [142] which normally computes 3D convolutional features from a volumetric data, the proposed pseudo-3D CNN applies 2D convolution operators to the images sampled from coronal, sagittal, and axial views that centred at the target voxels. The proposed method has many fewer computational operations and much less complexity, such that it is more efficient in speed. The details of the network architecture are illustrated in Fig. 4.1, and the parameter settings of the key components are listed in Table 4.3.

Table 4.3: The parameter settings of the key components of proposed Pseudo-3D CNN network.

BLK	**Type**	**Parameter**
(a)	C1, C2	13×13 patches from coronal view at two scales
	S1, S2	13×13 patches from sagittal view at two scales
	A1, A2	13×13 patches from axial view at two scales
(b)	Conv. 3	16 (3×3) Conv. filters with stride of 2 pixels
	BNorm	Batch Normalisation
	ReLU	Rectified Linear Unit activation function
(c).B5	Conv. 5	192 (5×5) Conv. filters with stride of 2 pixels
	BNorm	Batch Normalisation
	ReLU	Rectified Linear Unit activation function
(c).B3	Conv. 3	192 (3×3) Conv. filters with stride of 1 pixel
	BNorm	Batch Normalisation
	ReLU	Rectified Linear Unit activation function
	Conv. 3	192 (3×3) Conv. filters with stride of 2 pixels
	BNorm	Batch Normalisation
	ReLU	Rectified Linear Unit activation function
(d)	Ave. Pool	4×4 average pooling filter
	FC. 2	Fully Connected layer with 2 outputs
	Softmax	Softmax layer for binary classification

The pseduo 3D CNN network consists of four components as follows: (a) multi-scale pseudo-3D sampling, (b) primitive feature extraction, (c) feature aggregation and generali-

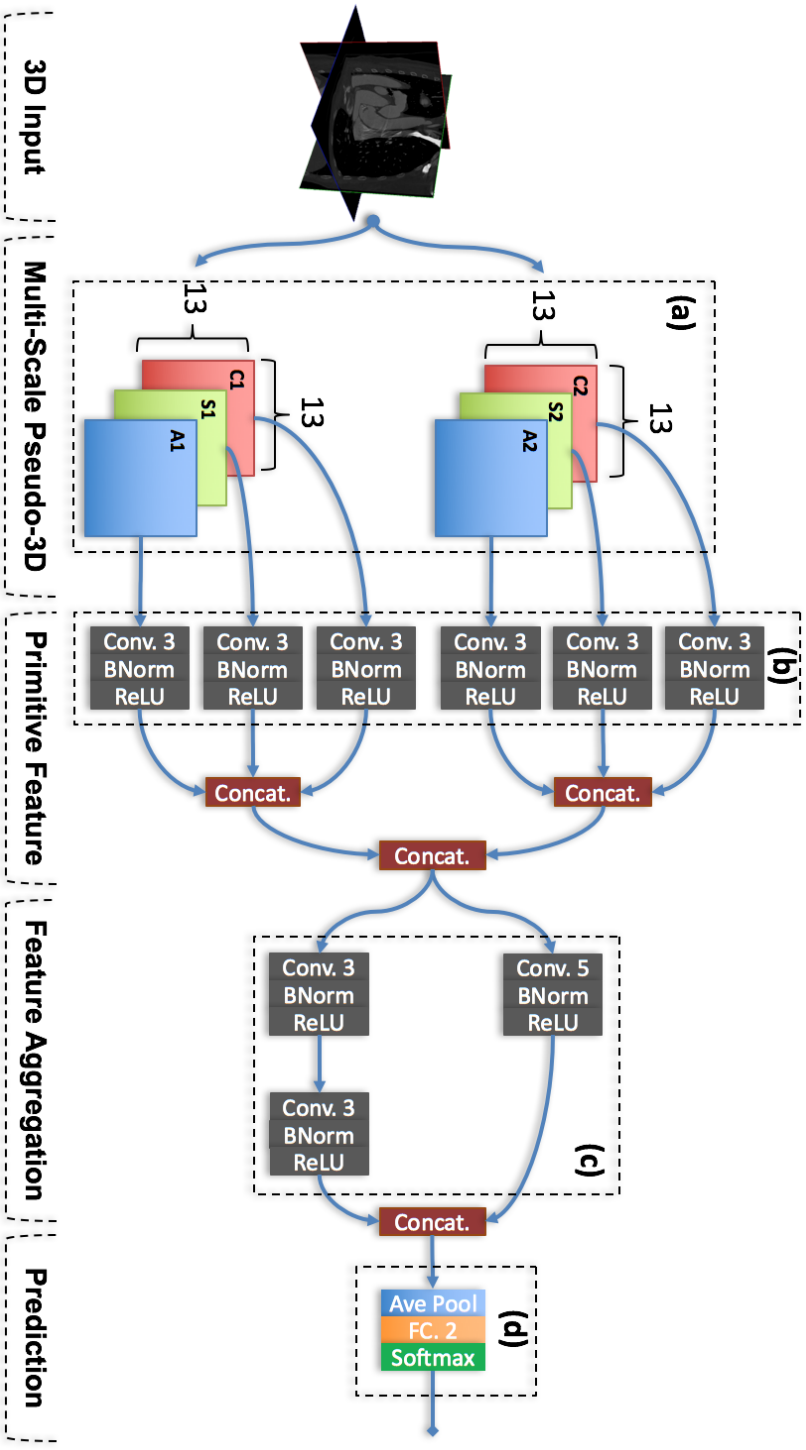


Figure 4.1: The network architecture of pseudo-3D CNN detector which consists of four components as follows: (a) multi-scale pseudo-3D sampling, (b) primitive feature extraction, (c) feature aggregation and generalisation, and (d) foreground-background prediction.

sation, and (d) foreground-background prediction. Given a 3D volumetric data and a voxel location, block (a) constructs the image patches with size of $W \times H$ pixels from coronal, sagittal, and axial views using MPR sampling at multiple scales $\{S_1 ... S_n\}$. Therefore, there are in total $3 \times |S|$ images fed into the networks as inputs, in our case, $(W,H) = 13$, and $|S| = 2$. In block (b), each primitive feature extractor includes a 3×3 convolutional filtering layer, a batch normalisation layer, and a Rectified Linear Unit (ReLU) layer connected consecutively. The primitive feature extraction processing is applied to individual images per view per scale, where the learnt responses of kernel filters then join together via channel-wise concatenation. A batch normalization layer as a regularisation scheme is inserted between the convolutional layer and the activation layer to compensate the internal covariate shift that is introduced by mini-batch gradient descent via Back-Prop learning [56]. A ReLU is a favorable and efficient choice for non-linear activation function, where the vanishing gradient issue no longer exists at the positive axis, and highly sparse models are often obtained to ensure a reliable local minima to some extent [143]. Block (c) consists of two branches, a single feature extractor with a 5×5 convolutional layer, and two consecutive feature extractors with two 3×3 convolutional layers. Two branches have the same size of receptive field (5×5 pixels), and join the features via channel-wise concatenation at the end, whereas the depths of feature abstraction are different (1 level for the top branch, 2 levels for the bottom branch, see Fig. 4.1 and Table 4.3). This strategy of enforcing feature aggregation from different abstraction levels was proved to be efficient to boost the accuracy [144]. No pooling layer is used for feature extraction blocks (b & c), spatial down-sampling is applied by setting the stride of convolutional filter to 2 with boundary padding. To note that for the bottom branch of block (c), only the stride of the last feature extractor is set to 2 pixels in order to achieve the consistent spatial resolution with the top branch. Therefore, the feature outputs of block (b & c) have the spatial resolutions of 7×7 pixels and 4×4 pixels respectively. In block (d), an average pooling layer with the filter size of 4×4 pixels is used to reduce the spatial resolution to 1 pixel, whereas the features are encoded within the channels. Then, a fully connected layer with 2 output nodes and a *Softmax* prediction layer are followed to perform binary classification. During the training procedure, the *Softmax* is replaced by its *log* loss version for Back-Prop. For a voxel, the pseduo 3D CNN network encodes the aggregated features using the perpendicular perspective projections from multi-scales, where a set of rich descriptors of local appearances and geometrical structures can be extracted hierarchically, and then be identified using discriminative analysis for

segmentation.

4.2.4 Localised Interactive Refining

Due to different clinical conditions of patients, the medical scans are normally acquired in a case by case basis. The patient-specific variations are often observed when different radiation doses, the scanning angles are used. A fully automatic off-line model that can achieve the optimal classification accuracy is extremely hard to train, and is generally not available. User intervention from the experienced clinician is informative and very helpful to overcome the difficulties that are introduced by the personalised image analysis. In order to minimise the intervention effort of the user, the stroke based foreground and background guiding curves are introduced in our method. There are two key bits of information implied inside the guiding strokes: the supervision knowledge, and the spatial location that need to be corrected. The voxels bypassed by the foreground strokes are considered as positive regions, whereas the ones along the background strokes indicate the negative regions. Although, at the testing stage, such supervision knowledge is limited in terms of the number of strokes that are acquired from the user, the labelled regions are more representative and informative for the testing volume which can be used to revise the pre-learnt model, and add the missing variances. Since the pseudo-3D CNN detector is trained using Back-Prop with Stochastic Gradient Descent (SGD), it can then be fine-tuned on-the-fly whenever the training data sampled from the new guiding strokes is available. However, fine-tuning is very sensitive to the training data in terms of adjusting feature patterns, and shifting the decision boundary, whereas the strokes normally indicate specific case by case patterns or the outliers. Training a global classifier using this supervision information with SGD overtime will lead to a biased model, and the overall accuracy can drop dramatically. In order to overcome such difficulty, there are three training strategies that are introduced to suppress the learning oscillation. First, in addition to the voxels interactively acquired from user, there are a number of pseudo-3D patches sampled from original training data to form the fine-tuning set. Hence, the model revision will not be dominated by the training data given by the supervision strokes, and the gradient of each mini-batch is corrected to some extent towards the direction of common pattern over all variations. Second, a relatively lower learning rate, and a smaller number of training epochs are used to ensure a stable gradient descent optimisation, and prevent largely adapting the model towards the special variations. The last but the most important strategy is localised refining, which takes the location information

embedded inside the supervision strokes into account. It considers that the trajectories of the strokes imply the bypassed local regions require refining, where the re-classification is only applied to the sub-volume that are k-neighbours of the supervision strokes. This procedure is performed iteratively until the satisfying result is achieved, through which the case by case patterns are learnt, and classification are refined. Algorithm 1 shows the detail of proposed localised interactive refining scheme.

Algorithm 1: Localised Interactive Refining

Input : $\mathscr{C}$ is a trained pseudo-3D CNN detector.
Input : $\mathscr{V}$ is a 3D volumetric image.
Input : $\mathscr{D}_s$ is the binary classification of $\mathscr{V}$ given $\mathscr{C}$.
Output: $\mathscr{D}_i$, $\mathscr{S}_i$ are the refined binary classification and the confidence score of $\mathscr{V}$ respectively.

$\mathscr{D}_i \leftarrow \mathscr{D}_s$, Initiate the classification result;
$n \leftarrow 0$, Initiate the interactive classification counter;
while *the user provides foreground strokes* $\mathscr{F}$*, and background strokes* $\mathscr{B}$ **do**
  $n \leftarrow n+1$, increase the counter;
  $\mathscr{P}_o \leftarrow$ randomly sample foreground and background pseudo-3D patches from the original dataset for training $\mathscr{C}$;
  $\mathscr{P}_i \leftarrow$ sample foreground and background pseudo-3D patches and corresponding labels along the guiding strokes $\mathscr{F}$ and $\mathscr{B}$;
  $\mathscr{C}_i^n \leftarrow$ fine-tune the pre-trained CNN detector $\mathscr{C}$ using $\mathscr{P}_o$ and $\mathscr{P}_i$ with a relatively lower learning rate $\mathscr{L}_i$;
  $\mathscr{V}_i^n \leftarrow$ find the irregular sub-volume which contains the voxels that are the k-neighbour of the guiding strokes $\mathscr{F}$ and $\mathscr{B}$;
  $\mathscr{D}_i^n, \mathscr{S}_i^n \leftarrow$ classify and score the sub-volume $\mathscr{V}_i^n$ using the fine-tuned model $\mathscr{C}_i^n$;
  $\mathscr{D}_i, \mathscr{S}_i \leftarrow$ merge the refined classification result $\mathscr{D}_i^n$, and confidence $\mathscr{S}_i^n$;
end
return $\mathscr{D}_i$ *and* $\mathscr{S}_i$;

4.2.5 Non-Uniform Implicit B-spline Surface

A binary volume is obtained using the proposed 2-stage cascade detector, and interactive refining process, where each voxel is assigned with either a foreground or background label independently. It is considered as a loose presentation of object, while a compact model that has rich geometrical interpolation is missing and in need. NU-IBS model is proposed in this section which can be used to represent shape using a set of parametric basis functions that have non-uniform local supports. It offers topological flexibility, sparse and local control, and is able

to adapt in terms of the shape complexity. The key idea is that the shape is first embedded in an implicit representation using a signed distance function, it then can be approximated using non-uniform B-spline patches in a parametric form:

$$\mathscr{L}(\mathbf{X}) = \mathscr{C}^T \, D(\mathbf{X}) \tag{4.5}$$

where $\mathbf{X} \in \mathbf{R}^3$ is the control knots in 3-dimensional space represented using *xyz*-coordinates, D is the B-spline basis vector given $\mathbf{X}$, $\mathscr{C}$ is the coefficient vector for all B-spline bases, and $\mathscr{L}$ is the approximated level-set function. We will introduce the formulation of uniform implicit B-splines surface first, and then show its non-uniform expansion using density mapping of control knots that is based on the estimation of shape complexity. Hence, constructing the NU-IBS representation of a shape is equivalent to solving a non-linear least square problem, while the surface can then be reconstructed via interpolation given its implicit parametric formulation.

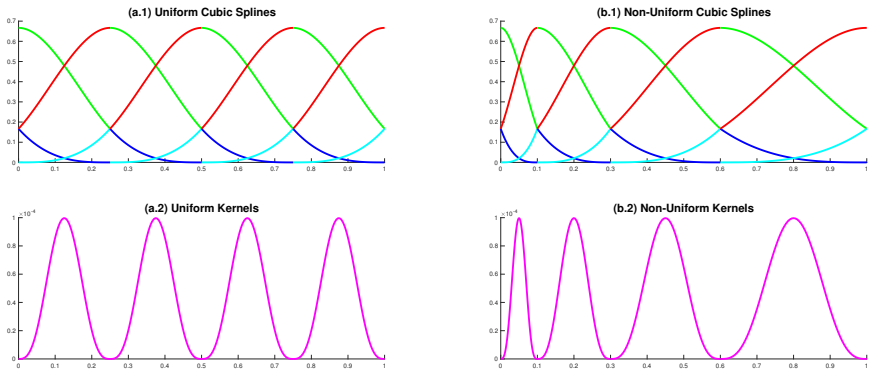


Figure 4.2: (a.1) A 1D example of cubic B-splines basis functions that are made out of scaling and translating the uniform blending functions. (a.2) An example of unweighted uniform kernel functions. (b.1) A 1D example of cubic B-Spline basis functions that are made out of scaling and translating the non-uniform blending functions. (b.2) An example of unweighted non-uniform kernel functions.

4.2.5.1 Uniform Implicit B-spline Surface

In the case of cubic splines that are used in proposed method, the basis vector is constructed using four 3^{rd} degree polynomial blending functions as follows:

$$\begin{aligned} b_0(u) &= (1-u)^3/6 \\ b_1(u) &= (3u^3 - 6u^2)/6 \\ b_2(u) &= (-3u^3 + 3u^2 + 3u + 1)/6 \\ b_3(u) &= u^3/6 \end{aligned} \tag{4.6}$$

Let r,s,t be the indexes of blending functions, N be the number of basis functions which are uniformly placed over the definition interval $[0,1]$, and $c_{i,j,k}$ be the coefficient of knot $\{i,j,k\}$, such that given a local control point in xyz-coordinates, the indexes of knot and corresponding u,v,w for each axis can be mapped as follows:

$$\begin{gathered} \delta = 1/(N-3) \\ i = \lceil x/\delta \rceil,\ j = \lceil y/\delta \rceil,\ k = \lceil z/\delta \rceil \\ u = x/\delta - \lfloor x/\delta \rfloor,\ v = y/\delta - \lfloor y/\delta \rfloor,\ w = z/\delta - \lfloor z/\delta \rfloor \end{gathered} \tag{4.7}$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are ceil and floor rounding operators respectively. Then the approximated level-set function $\mathscr{L}$ can be computed based on Eq. 4.6 and 4.7 as follows:

$$\mathscr{L}(\mathbf{X}) = \sum_{r,s,t=0}^{3} c_{i+r,j+s,k+t}\ b_r(u) b_s(v) b_t(w) \tag{4.8}$$

Fig. 4.2 (a.1) shows a 1D example of cubic B-Spline basis functions that are made out of scaling and translating uniform blending functions, and (a.2) are the unweighed uniform kernel functions that are computed using Eq. 4.8 with constant coefficients ($\forall c = 1$).

4.2.5.2 Non-Uniform Expansion

The uniform model places the knots evenly on the definition intervals, however, for the shapes that have both simple and complex structures, densely distributed knots are required to cover the highly curved parts, while the computational resource is wasted at the smooth regions. Hence, it is worth considering a non-uniform model that can adaptively distribute the knots based on the local complexity of the shape. Fig. 4.2 (b.1) shows an 1D example of cubic B-Spline basis functions that are made out of scaling and translating non-uniform blending

functions, and (b.2) are the unweighed non-uniform kernel functions that are computed using Eq. 4.8 with constant coefficients ($\forall c = 1$). In Fig. 4.2 (b.2), the densely distributed knots (towards left) that has small support radius provide more compact representation compared to the loose ones (towards right).

The surface lays on the zero level set of $\mathscr{L}$, where those highly curved structures are presented as large oscillations and high frequency signals over the signed distance field. A scale weighted complexity estimation method is proposed to determine the density of knots for each axes. Fig. 4.3 (a) show a 1D signal that is constructed using a set of sine functions that have different frequencies. Fig. 4.3 (b) show the heat map of continuous wavelet coefficients of the signal in multiple scales. Given an implicit shape representation $\mathscr{L}_0$, the density of shape complexity along one axis can be estimated via marginalising the scale weighted amplitudes of Guassian wavelet responses over other two axes, as follows:

$$\begin{aligned}
\mathscr{W}_x &= \sum_{y=1}^{Y}\sum_{z=1}^{Z}\sum_{m=1}^{M}\frac{1}{m}\,||\mathscr{L}_0(:,y,z)\ \otimes\ \mathscr{K}_{gaus}||_1 \\
\mathscr{W}_y &= \sum_{x=1}^{X}\sum_{z=1}^{Z}\sum_{m=1}^{M}\frac{1}{m}\,||\mathscr{L}_0(x,:,z)\ \otimes\ \mathscr{K}_{gaus}||_1 \\
\mathscr{W}_z &= \sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{m=1}^{M}\frac{1}{m}\,||\mathscr{L}_0(x,y,:)\ \otimes\ \mathscr{K}_{gaus}||_1
\end{aligned} \tag{4.9}$$

where M is the number of scales, $\mathscr{K}_{gaus}$ is the kernel filter of Gaussian wavelet, and $\otimes$ is the convolution operator. The amplitudes ($L1$ norm) of wavelet coefficients are weighted by the reciprocal of the scales, which lowers the contributions of detected signal oscillation in large scales, while concentrates it in small scales. Therefore, the density of shape complexity can be interpreted as the density histogram of subtle changes of geometrical structures along a certain axis over the whole volume. N B-splines divide the definition domain into $N-3$ intervals where the knots are placed at the intersections. An NU-IBS divides the intervals according to density histogram $\mathscr{W}_x, \mathscr{W}_y$, and $\mathscr{W}_z$, ensures that each intervals has even accumulated density, where such mapping can be easily obtained by using histogram equalisation algorithm. For NU-IBS, the uniform distributed knots (Eq. 4.7) are replaced by this adaptive mapping method, and the basis vector D can be constructed accordingly. By doing so, the regions that have complex geometrical structures are approximated using more B-spline patches with compact supports, when the total number of B-splines is fixed.

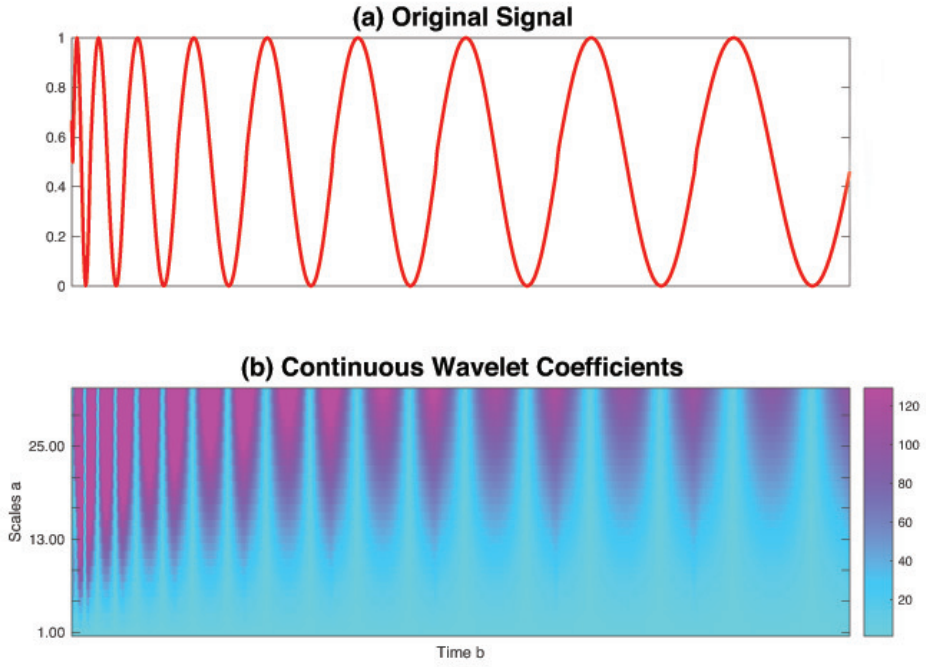


Figure 4.3: (a) A 1D signal is constructed using a set of sine functions that have different frequencies. (b) The heat map of continuous wavelet coefficients of the signal given in (a).

4.2.5.3 Surface Parametrisation and Reconstruction

Given a binary volume, a signed distance function $\mathscr{L}'$ is computed, where the voxels on the object surface are assigned to 0, and the others are the Euclidean distance to the surface. In order to ensure the numerical stability, we follow the convention that the distance values inside the object are positive, and outsides are negative. The surface parametrisation is equivalent to solving the following non-linear least square problem with a ridge regularisation:

$$\begin{aligned} \mathscr{C}^* &= \arg\min_{\mathscr{C}} \{||\mathscr{L}' - \mathscr{L}(\mathbf{X})||_2 + \mu(\mathscr{C}^T \mathbf{I}\, \mathscr{C})\} \\ &= \arg\min_{\mathscr{C}} \{||\mathscr{L}' - \mathscr{C}^T D(\mathbf{X})||_2 + \mu(\mathscr{C}^T \mathbf{I}\, \mathscr{C})\} \end{aligned} \tag{4.10}$$

where the μ is the regularisation parameter, and $\mathbf{I}$ is the identity matrix. Given a uniformly sampled sub-volume $\mathbf{S}$ from $\mathscr{L}'$, the vectorised distance values $\mathbf{B}_s$, and basis matrix $D_s(\mathbf{X})$ can be constructed by concatenating the basis vector of corresponding points in $\mathbf{S}$ row-by-row,

such that the approximated solution can be found as follows:

$$\begin{aligned} \mathscr{C} &= D_s^{\dagger}(\mathbf{X})\ \mathbf{B}_s = (D_s(\mathbf{X})^T D_s(\mathbf{X}))^{-1} D_s(\mathbf{X})^T \mathbf{B}_s \\ \mathscr{C}^* &= (D_s(\mathbf{X})^T D_s(\mathbf{X}) + \mu\ \mathbf{I})^{-1}\ D_s(\mathbf{X})^T\ \mathbf{B}_s \end{aligned} \tag{4.11}$$

where $D_s^{\dagger}(\mathbf{X})$ denotes the pseudo inverse of $D_s(\mathbf{X})$. The basis matrix $(D_s(\mathbf{X})^T D_s(\mathbf{X}))^{-1}$ is a highly sparse matrix, where much faster factorisation algorithms are available compared to dense matrix [145]. Given the parametric representation of a shape, the level-set function can be computed via interpolating the distance field of the shape within the definition domain using Eq. 4.5. Hence, the surface is reconstructed on the zero level-set manifold.

4.2.6 Segmentation as Region-based Deformation

The NU-IBS is the approximated parametric form of the level set function given a shape which can be constructed directly from the binary decision volume. It imposes the geometrical smoothness constraint to the loosely detected object, however, the classification score has not been taken into account. In order to incorporate the data support, level set based segmentation is introduced which captures the shape via propagating the zero interface Γ according to a PDE derived from an energy functional. In this chapter, we consider the classical Chan-Vese energy functional [146] that leads to a solution partitioning the definition domain into two regions with piecewise constant data support, and delimiting the boundaries of the objects:

$$\begin{aligned} J(\mathscr{L}) = \lambda_1 &\int_{\Omega} \delta(\mathscr{L}) ||\nabla \mathscr{L}||\ d\mathscr{C} \\ &+ \lambda_2 \int_{\Omega} (\mathscr{S}(\mathscr{C}) - C_1(\mathscr{L}))^2 \cdot u(\mathscr{L})\ d\mathscr{C} \\ &+ \lambda_3 \int_{\Omega} (\mathscr{S}(\mathscr{C}) - C_2(\mathscr{L}))^2 (1 - u(\mathscr{L}))\ d\mathscr{C} \end{aligned} \tag{4.12}$$

where u and δ are the Heaviside and Dirac univariate functions respectively, $\lambda_1, \lambda_2, \lambda_3$ are positive hyper-parameters that control the contributions from the surface smoothness, inside and outside of the object. $C_1(\mathscr{L}), C_2(\mathscr{L})$ are computed during the interface propagation at each iteration using the following expression:

$$\begin{aligned} C_1(\mathscr{L}) &= \frac{\int_{\Omega} \mathscr{S}(\mathscr{C}) \cdot u(\mathscr{L}(\mathscr{C}, t))\ d\mathscr{C}}{\int_{\Omega} u(\mathscr{L}(\mathscr{C}, t))\ d\mathscr{C}} \\ C_2(\mathscr{L}) &= \frac{\int_{\Omega} \mathscr{S}(\mathscr{C}) \cdot (1 - u(\mathscr{L}(\mathscr{C}, t)))\ d\mathscr{C}}{\int_{\Omega} (1 - u(\mathscr{L}(\mathscr{C}, t)))\ d\mathscr{C}} \end{aligned} \tag{4.13}$$

The general minimisation solution of $J(\mathscr{L})$ can be found using variational calculus and gradient descent method [146, 147, 148], as follows:

$$\begin{aligned} \frac{\partial \mathscr{L}(\mathscr{C},t)}{\partial t} + \mathbf{V}(\mathscr{C},t) \cdot \delta_\varepsilon(\mathscr{L}(\mathscr{C},t)) = 0 \\ \delta_\varepsilon(x) = \frac{1}{\pi\varepsilon \cdot (1+(\frac{x}{\varepsilon})^2)} \end{aligned} \tag{4.14}$$

where δ_ε is a regularised Dirac function. It is noteworthy that in Eq. 4.12, λ_1 controls the contribution weight of surface smoothness which is already assured by the intrinsic property NU-IBS, we can simply set $\lambda_1 = 0$, and $\lambda_2 = \lambda_3 = 1$. Then, the velocity term is given as:

$$\mathbf{V}(\mathscr{C},t) = -(\mathscr{S}(\mathscr{C}) - C_1(\mathscr{L}))^2 + (\mathscr{S}(\mathscr{C}) - C_2(\mathscr{L}))^2 \tag{4.15}$$

By combining Eq. 4.5, 4.10, 4.14, the PDE equation can then be transformed to an ODE equation with respect to the B-spline coefficients $\mathscr{C}$ of NU-IBS, where the optimal segmentation can be found by iteratively updating $\mathscr{C}$ according to the detection confidence score $\mathscr{S}$ until the steady state is reached. The gradient descent solution is given as:

$$\begin{aligned} \frac{d\mathscr{C}}{dt} = -\{(D_s(\mathbf{X})^T D_s(\mathbf{X}) + \mu\mathbf{I})^{-1} D_s(\mathbf{X})^T \\ \times (\mathbf{V}(\mathscr{C},t) \cdot \delta_\varepsilon(\mathscr{L}(\mathscr{C},t)))\} \\ \mathscr{C}_{(n+1)} = \mathscr{C}_{(n)} + \tau \frac{d\,\mathscr{C}_{(n)}}{dt} \end{aligned} \tag{4.16}$$

where τ is the step size. A small τ value enables a steady numerical solution while more iterations are required to converge.

4.3 Evaluation

4.3.1 3D CTA Dataset

The proposed method was evaluated on a 3D CTA dataset which contains 36 volumetric TAVI scans. The number of slice of each scan varies, while the image size of each slice is fixed at 256×256 pixels across all scans. The anatomical structure to be segmented is the aorta which is the large blood vessel that carries oxygen-rich blood from the left ventricle of the heart to other parts of the body, and its root attaches into the heart. The aortic root consists of three valve leaflets which open to allow the blood in the left ventricle to flow into the ascending aorta when the heart contracts. The ascending and descending aortas form an arch-like shape.

An example of 3D TAVI image and its 3D surface rendering is shown in Fig. 4.4, where the aorta root is highlighted with the organ circle. This is an ideal case to study our method. First, the aorta has heterogeneous local geometrical complexity, where its arch is a generally smooth structure while its root formed by three valve leaflets has far more complex topology. Second, differentiating the aorta from the volume is not a trivial problem, as there are many similar object in the whole volume in terms of image appearance and geometry structure, for example, the pulmonary artery and the superior vena cava. To label the ground-truth, for each scan the ROI was cropped out, and a reconstruction plane was found manually. The representative plane is perpendicular to the ascending direction of the root. Then, the root including three valve leaflets, and the arch were labelled slice by slice up to the top of the arch using closed contours. Hence, a binary volume can be constructed where insides of the contours were assigned to 1, and outsides were 0 indicating foreground objects and background respectively.

4.3.2 Experimental Result

3-Fold cross validation was used, where the 36 volumes were randomly divided into 3 subsets each of which contains 12 volumes. In each evaluation round, one subset was retained for testing, and the rests were used for training. The intensity value was scaled into the range of $[0, 255]$ given the optimal window size and window level that were provided in the Digital Imaging and Communications in Medicine (DICOM) image meta information. To train the Naive-Bayesian detector, 300K foreground and 300K background voxel intensities from each training volume were collected, which was about 6% of the total number of voxels in the whole volume. A GMM with 5 Gaussian components was used, and the conditional posterior probabilities were computed given an even foreground and background prior. Fig. 4.5 shows three Naive-Bayesian classifiers that were constructed for different fold tests. The blue and cyan curves are conditional probabilities of foreground and background that are modelled using the GMM, and the red and black curves are posterior probabilities obtained through Bayesian rule given a pre-defined prior. To train a pseudo-3D CNN detector, the false positive and the false negative voxels were collected from which the multi-scale Pseudo-3D patches were sampled from each volume. The size of mini batch was 512, the number of epochs was set to 6, which leaded to 63,420 iterations on average. The initial learning rate was set to 0.1, and then divided by a factor of 10 every two epochs. To simulate the localised interactive refining procedure, we randomly selected 1,280 voxels from both false positives and false negatives that

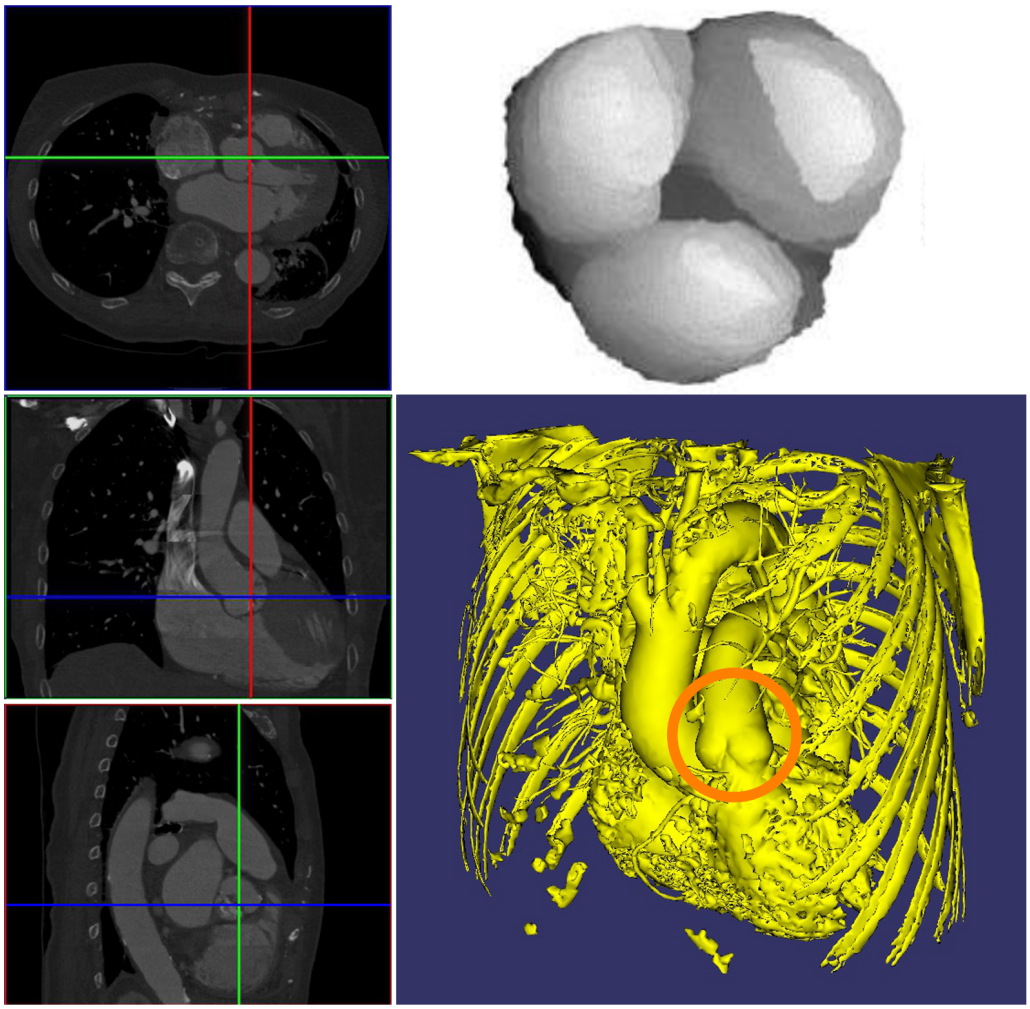


Figure 4.4: An example of 3D CTA TAVI image from 3 orthorgonal views and surface rendering created using *3DimViewer* [30]. The images from the top to the bottom in the left column are axial view, coronal view and sagittal view respectively. The right column shows the mesh model of aorta root (top) and 3D surface rending of the volume (bottom), where the aorta root is highlighted with the organ circle.

were given by the Pseudo-3D detector as user guiding strokes. In addition, 3,840 voxels were randomly sampled from the original dataset, which was together with the simulated guiding strokes making a fine-tuning dataset with 5,120 samples in total. The learning rate and training epoch for fine-tuning were set to 10^{-5} and 10 respectively to avoid large decision boundary shifting. The localised refinement was applied to $9 \times 9 \times 9$ sub-volumes that were centred at the voxels from simulated guiding strokes.

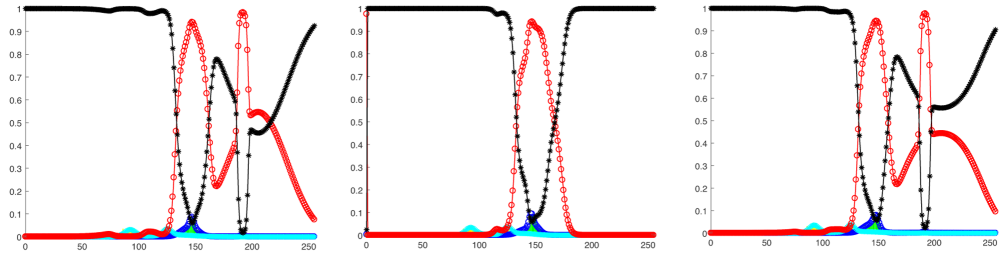


Figure 4.5: The visualization of three Naive-Bayesian classifiers trained for different dataset folds. The blue and cyan curves are conditional probabilities of foreground and background that are modelled using GMMs. The red and black curves are posterior probabilities obtained through Bayesian rule given a pre-defined prior.

The classification results of individual stages from different fold tests are listed in Table 4.4. The Naive-Bayesian classifier achieves 93.20% true positive rate on average, which is required to tolerate 9.63% false positive. However, as the majority of the volume is background the false positive rate is considerably high, where the structures that have similar intensities are preserved inevitably, such as rib cage, pulmonary artery, blood vessels in the lungs, ventricles and atria (See Fig. 4.8 row (a)). In the next stage, the pseudo-3D CNN detector dramatically eliminates those objects by learning the spatial feature hierarchically. It achieves a 0.82% false positive rate, while it makes a sacrifice in true positive rate which was reduced by 8.06% on average. Generally, the miss-classification happens around the outer boundaries of the arch and the tips of leaflets shown in blue in Fig. 4.8 row (b), where those regions either have no sharp edge or are too close to other large blood vessels. The localised interactive refining scheme greatly improves the detection accuracy. In particular, the first round of refining boosts the true positive rate from 85.14% to 92.16% while reduces the false positive rate further. The major reason could be that the fine-tuning procedure was trained using the samples directly from the testing volume that are far more representative. Moreover, the localised refining strategy prevents contaminating the well classified regions. Fig. 4.8 row (c) and (d) show the examples

of the first round and the last round of refinement respectively. The false positive (yellow regions) and false negative (blue regions) are eliminated progressively, especially the example in the second column. It is worth noting that the false positive rate goes higher slowly after 2 rounds of refinement, which is caused by the miss-classifications within the localized regions, and shows the upper limit of discrimination power of pseudo-3D CNN to some extent. A connectivity component analysis was applied as a post-processing step to remove the isolated small regions.

Table 4.4: Quantitative classification results (TP: True Positive, FP: False Positive, in %) of each cascade stage and localised interactive refining.

	Fold-1		Fold-2		Fold-3		Avg.	
	TP	FP	TP	FP	TP	FP	TP	FP
N-B	94.79	8.65	87.99	9.41	96.83	10.83	93.20	9.63
P-3D	84.96	0.66	81.56	0.86	88.91	0.95	85.14	0.82
Ref-1	93.81	0.44	90.35	0.45	92.31	0.61	92.16	0.50
Ref-2	94.89	0.46	91.14	0.39	93.02	0.64	93.02	0.50
Ref-3	95.30	0.48	92.11	0.40	93.55	0.66	93.65	0.51
Ref-4	95.59	0.49	93.07	0.43	94.02	0.69	94.23	0.54
Ref-5	95.89	0.50	94.06	0.47	94.53	0.71	94.83	0.56

The shape representation was initially constructed using the binary classification volume, and then deformed with regard to the normalised prediction scores until it converges ($\Delta C_1 + \Delta C_2$) $<5e$-4). The maximum number of iterations was set to 50. The regularization parameters of Heaviside and Dirac functions were set to $1e$-5 and $1e$-1 respectively, and the step size τ was set to $1e$-1 for all iterations. We compared the proposed NU-IBS with uniform IBS using 23 and 28 B-splines with different sampling rates. The quantitative mesurements are listed in Table 4.5 which were calculated using *EvaluateSegmentation Tool* [149]. Table 4.5 shows that given the same number of B-splines and sampling rate, NU-IBS outperforms IBS in all aspects. Fig. 4.8 row (e) shows three qualitative segmentation results, where the false negatives in blue can be largely observed at the tips of valves. The best performance is observed with 28 B-splines and a sampling rate of every 3 pixels, where the highest recall rate (91.66%) and lowest Hausdorff distance (5.5733) is achieved. Although compared to the recall rate achieved by the interactive refinement (94.83%), the region based deformation decreases by 3.17% on average, which is mainly caused by the intrinsic smoothness property of NU-IBS, where it is an inevitable issue for all PIR approaches. However, compared to the uniform IBS, the proposed NU-IBS has much richer details of subtle structures. Fig. 4.6 shows the qualitative comparison

of uniform IBS (top row) and proposed NU-IBS (bottom row), where the uniform method turns to smooth out the geometrical details of aorta valves that are well preserved by our method. The main reason is that the IBS has far less B-spline patches at the valve regions compared to the NU-IBS, which prevents the IBS deforming further to match the data support. Fig. 4.7 shows three examples of proposed region based deformation using a cube as an initialisation, where the shape can break naturally during the deformation. The speed of deformation is controlled by the step size τ, where the large value generally leads to quick convergence while it also increase the possibility of producing intractable numerical error.

Table 4.5: Quantitative comparison of Uniform IBS and proposed Non-Uniform IBS. (BS: #B-splines; Sim: Similarity; Mut: Mutual Information; Hau: Hausdorff; Mah: Mahanabolis;)

BS	Rate	Method	Dice	Jaccard	Sim	Mut	Hau	Mah	Recall	Fallout
23	6	IBS	0.9095	0.8346	0.9173	0.1180	7.5909	0.0739	0.8406	0.0002
		NU-IBS	0.9296	0.8688	0.9314	0.1240	5.7885	0.0407	0.8702	0.0001
	3	IBS	0.9233	0.8580	0.9332	0.1221	6.9412	0.0632	0.8658	0.0002
		NU-IBS	0.9404	0.8881	0.9441	0.1279	5.8728	0.0381	0.8912	0.0001
28	6	IBS	0.9264	0.8633	0.9372	0.1230	6.7426	0.0602	0.8720	0.0002
		NU-IBS	0.9470	0.8997	0.9521	0.1300	5.8706	0.0359	0.9041	0.0001
	3	IBS	0.9304	0.8701	0.9425	0.1243	6.2202	0.0550	0.8800	0.0002
		NU-IBS	0.9536	0.9115	0.9594	0.1321	5.5733	0.0315	0.9166	0.0001

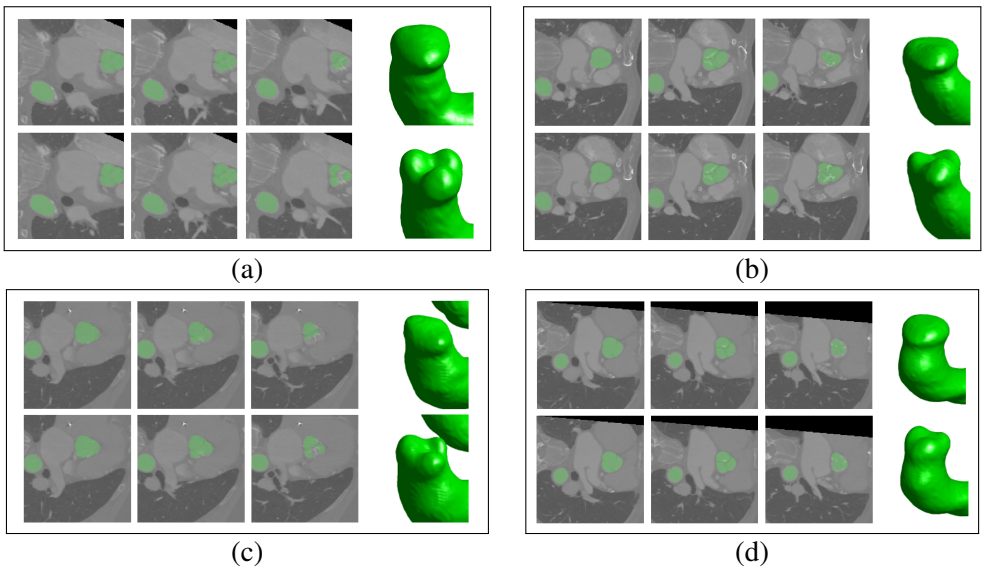
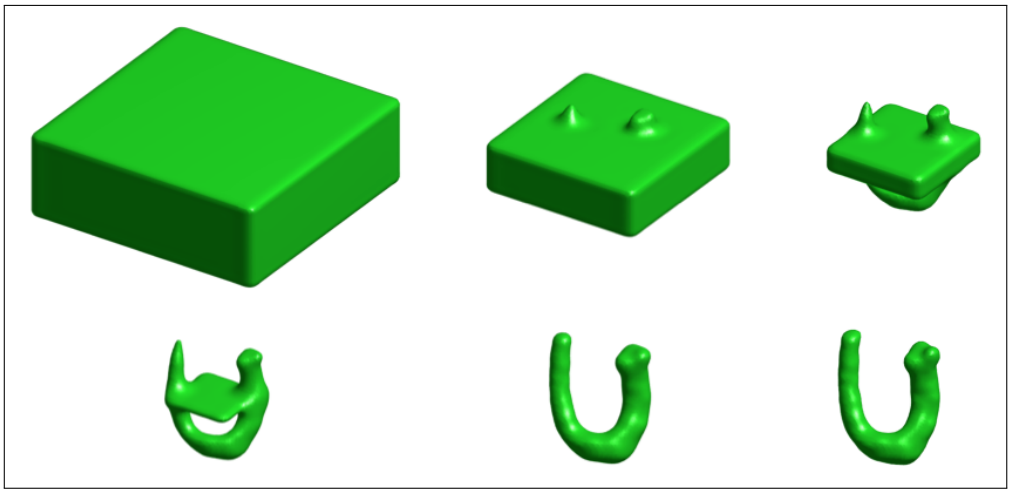
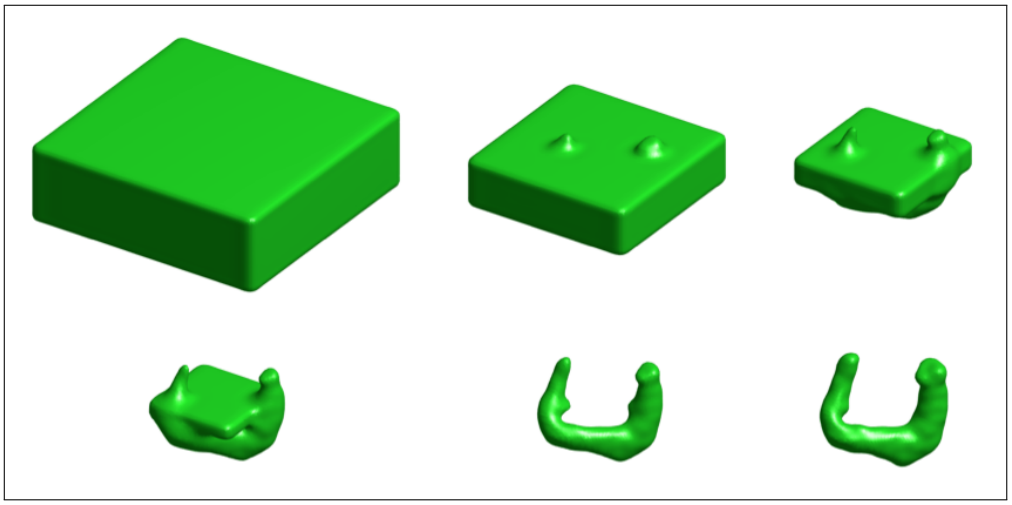


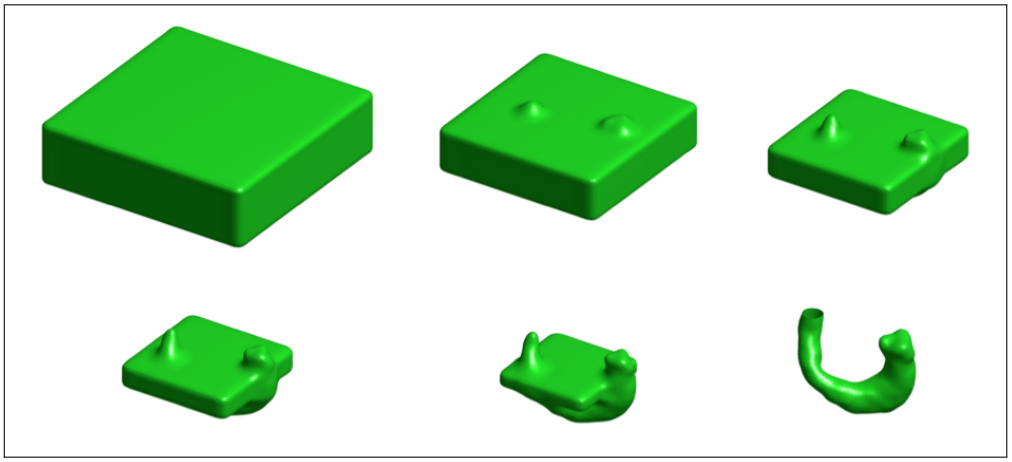
Figure 4.6: Qualitative comparisons of Uniform IBS (top row) and proposed NU-IBS (bottom row). The uniform method turns to smooth out the geometrical details of aorta valves that are well preserved by our method.



(a) $\tau = 2.50e\text{-}1$



(b) $\tau = 1.75e\text{-}1$



(c) $\tau = 1.00e\text{-}1$

Figure 4.7: The deformation process at the 1st, 4th, 8th, 10th, 12th and 20th iterations using different step sizes τ.

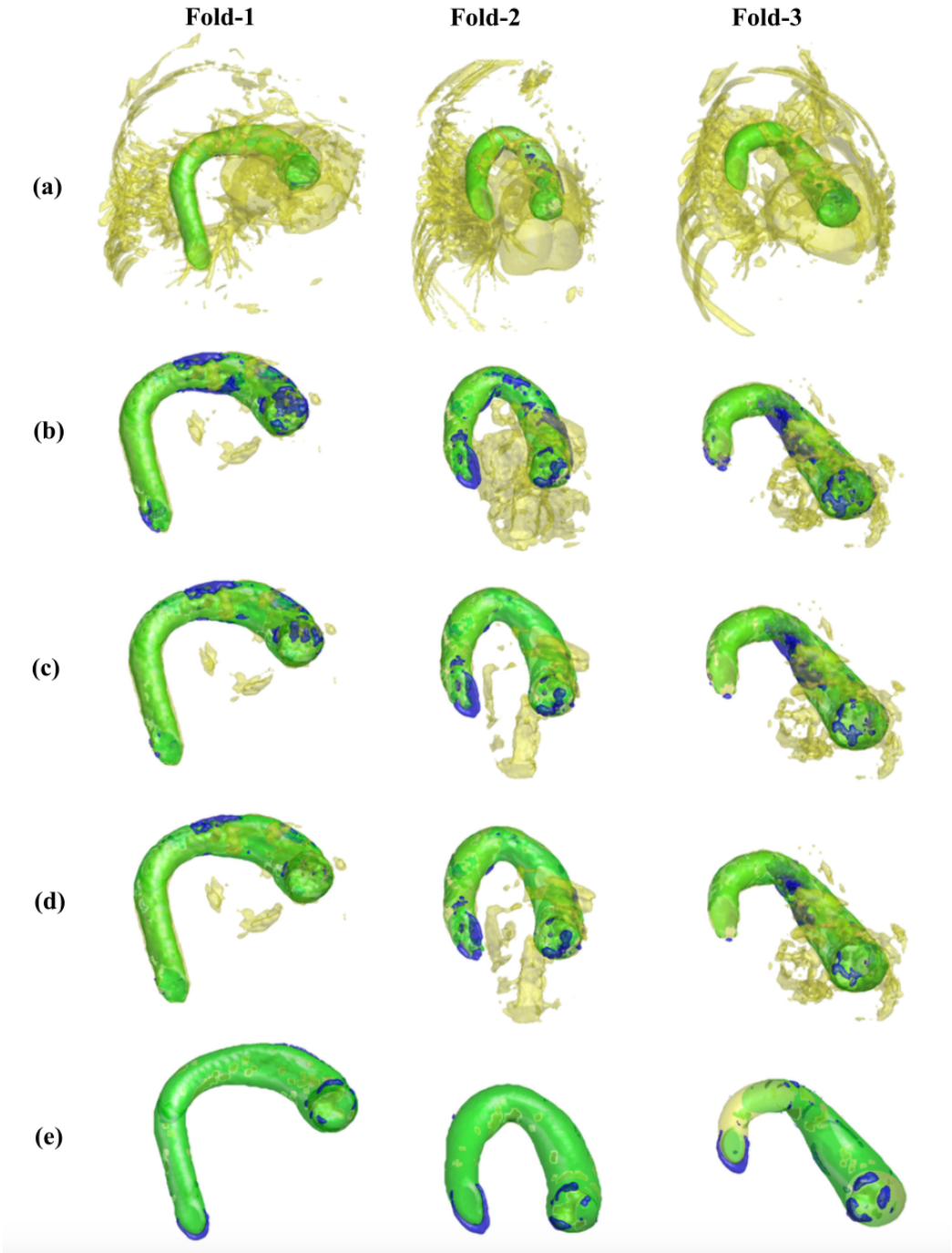


Figure 4.8: Qualitative results of detection and segmentation from three fold testing at different stages. (a) Naive-Bayesian classification results. (b) Pseudo-3D CNN classification results. (c) The first round of localised interactive refining results. (d) The last round of localised interactive refining results. (e) The final segmentation results. Green, yellow, and blue correspond to true positive, false positive, and false negative respectively.

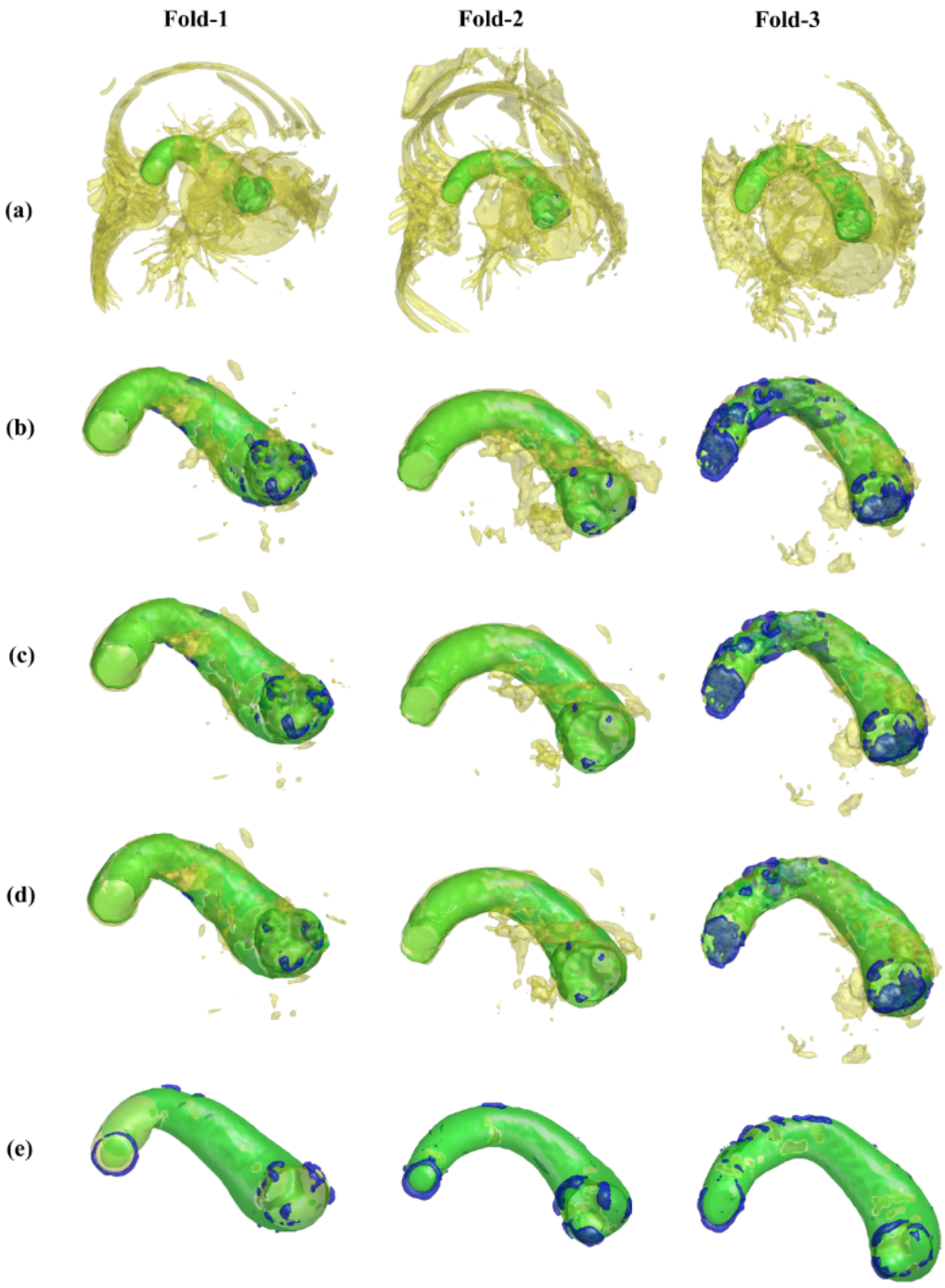


Figure 4.9: Additional qualitative results of detection and segmentation from three fold testing at different stages. (a) Naive-Bayesian classification results. (b) Pseudo-3D CNN classification results. (c) The first round of localised interactive refining results. (d) The last round of localised interactive refining results. (e) The final segmentation results. Green, yellow, and blue correspond to true positive, false positive, and false negative respectively.

4.3.3 Speed Discussion

The proposed method were evaluated on a machine with a 3.4-GHz Intel i7 (Sandy Bridge) CPU, 32GiB of RAM, and a Nvidia GeForce Titan X (Maxwell) GPU. The classification speeds of Naive-Bayesian classifier and pseudo-3D CNN classifier are 2,725,033 voxels/second and 18,985 voxels/second on average respectively. The NU-IBS and uniform IBS have the same computational complexity, we evaluated the speed efficiency of our method on a volume with a fixed size of $256 \times 256 \times 200$. The single thread speeds and the approximation accuracies of NU-IBS are reported in Table 4.6 where a sampling rate of 6 pixels was used. The total computational time can be further reduced to 59s and 67s for 23 and 28 B-splines cases respectively by using multi-threading techniques and optimised factorisation libraries, where *Intel TBB* [150] and *SuiteSparse* [151] were used in our case.

Table 4.6: Speed and Approximation Accuracy of proposed NU-IBS on a $256 \times 256 \times 200$ volume. (BS: #B-splines; Maxtrix: Basis Matrix Size; Dist: Signed Distance Transformation; Basis: Compute Basis Matrix; Chol: Cholesky Decomposition; Deform: 1-Iteration; Interp: Interpolation.)

BS	Matrix	#Points	Dist	Basis	Chol	Ave Error	Deform	Interp
23	12167^2	62866	19.256s	12.041s	16.969s	0.020	30.253s	226.950s
28	21952^2	62866	19.704s	12.779s	48.390s	0.032	30.460s	226.973s

4.4 Conclusion

In this chapter, we first introduced a two-stage object detection cascade that contains a fast Naive-Bayesian classifier and a powerful pseudo-3D CNN classifier, which balances the speed efficiency and discrimination performance. Particularly, the pseudo-3D classifier learns the hierarchical features and decision boundary simultaneously through a supervised classification task, hence, no hand-crafting is required. In addition, it avoids using 3D convolution operator that is a computational expensive. The proposed localised interactive refining scheme enables user guided miss-classification correction on the fly. The segmentation is obtained via regularising the voxel-based classification results with NU-IBS deformation in terms of the prediction scores. The NU-IBS has non-uniform distribution of control knot that is adapted to the density of geometrical complexity, which can well preserve the subtle structures. The proposed method is evaluated on a 3D CTA dataset via segmenting the aorta root and arch. The qualitative and

quantitative comparisons are reported, and show the superiorities of the proposed method in both segmentation accuracy and subtle structure preserving.

Chapter 5

Face Detection

Contents

5.1 Introduction **97**
5.2 Related Work **99**
5.3 Soft Cascade **101**
5.3.1 Proposed Method 102
5.3.2 Experiment and Discussion 106
5.4 Detection-Regression Cascade **114**
5.4.1 Proposed Method 114
5.4.2 Experiment and Discussion 120
5.5 Summary **127**

In Chapter 4, there were two notable findings that can be further applied to address generic computer vision problems. First, CNN can be used to learn discriminative features given sufficient data and suitable architectures, hence no hand-crafting features is required. Second, the cascade detector can be constructed by combining simple but fast elimination classifiers and accurate but complex detection classifiers, which neatly balances the speed efficiency and accuracy performance. In this chapter, we show these strategies can be jointly used to address the challenging face detection problem.

5.1 Introduction

View-specific face detection under controlled environment is largely considered a solved problem due to recent advances in object detection, in particular the work by Viola-Jones (VJ) [26]. As a typical detection problem, the class distribution between face and background is extremely unbalanced and heavily biased towards the background. The traditional VJ framework uses a multi-stage cascade detector, where individual stage is a binary filter that classifies retained hypotheses from previous stage into face and non-face. For efficiency, the traditional methods use simple visual features or weak classifiers at multiple stages (typically over 15). However, they perform poorly on the so-called *Face in the Wild* problem, where faces are captured with large pose and facial expression variations, severe occlusions and clutters, and poor lighting scenarios. Built upon those classical detection frameworks, several works have been recently reported in developing discriminative visual features [152, 27, 153] and strong classifiers [84, 154, 155] to improve face detection performances in the wild. Deep learning methods [37], especially CNNs, have shown outstanding successes in representative feature learning and supervised classification for various computer vision problems. Our work leverages recent advances in deep learning for efficient face detection. More in-depth discussions to these related work are presented in the next section.

From image retrieval perspective, face detection can be considered as a visual matching problem, where a window candidate is determined as face by successfully finding reliable correspondences in a pre-built exemplar database. In [156, 157], an exemplar database is constructed using localised visual words, and detection is obtained by finding the high confidence regions on the voting map provided by matched exemplars. The performance of those non-parametric searching methods can be severely compromised by the quality of exemplar database, such as the discriminative power of visual features, and the variation in coverage of different poses, illuminations, occlusions and so on. In addition, using a large exemplar database also slows down the detection speed as exploring large search space is a time consuming task. The Deformable Part Model (DPM) was originally proposed for object recognition, and can be considered as an alternative searching based method for detecting faces [158, 159]. It considers that the target object is consisted of several deformable parts. The part candidates are proposed by individual part detectors, and then the entire object can be found by searching for a most plausible configuration of displaced parts. DPM helps to overcome the difficulties introduced by severe occlusion and clutter, provided reliable performance of part detectors.

However, assembling individual parts into objects is equivalent to solving a combinational optimisation problem which could also be computationally expensive even with approximation algorithms.

Computational efficiency is one of the main concerns for practical detection system, especially when dealing with large number of hypothesis, complex visual feature, and strong classifier. For example, to precisely locate faces in the image, exhaustive search methods, such as sliding window, are commonly used to generate candidates. However, examining all hypotheses is computationally expensive, thus relatively simple features and weak classifiers are typically used to reduce the complexity [26, 27]. It is worth noting by taking this approach the detection problem is divided into a set of sub-problems first and then solved by combining individual sub-problem solver into a multi-stage detector [26, 160]. For example, Koestinger [161] trained a 20-stage VJ face detector using LBP features. Object region-proposal methods are popular for image recognition and object localisation, such as objectness [162], selective search [163], category independent object proposals [164], combinatorial grouping [165], and segmentation based methods [166]. However, generating object candidates generally involves region segmentation, classification, and grouping, which slow down the detection speed drastically. Furthermore, the recall rate of region proposal is generally lower than exhaustive search, such as sliding window.

Whilst using hand-crafted features is generally problematic, introducing powerful but complex models is often computationally inefficient. Especially, some recent works on adapting pre-trained large scale recognition models to face detection problem often requires excessive resource expenditure. Cascading, feature aggregation and multi-resolution are three efficient strategies for traditional visual recognition methods. In this paper, we show that such strategies can be used and integrated into the architecture design of CNN via. Shallow networks with feature aggregation at multi-resolution enables the traditional cascade framework to tackle the challenging detection problems efficiently. In this chapter, we propose two different approaches for tackling the face detection problem in an unconstrained environment. The Soft-Cascade approach considering a loose decision boundary over multiple stages of the cascade detector, where it has maximum 2 convolutional layers, and the final detection is made by averaging the classification results in an ensemble fashion. The Detection-Regression approach uses more complex CNN models with maximum 4 convolutional layers to construct a hard cascade. Whereas it more focuses on refining the location of detected windows using a regression

net that shares the features computing with other classification nets. The rest of chapter is organised as follows: Section 5.2 reviews related works on CNN-based face detection methods. Section 5.3 provides detailed descriptions of Soft-Cascade method, and experimental results on public datasets. Section 5.4 provides detailed descriptions of Detection-Regression method together with experimental results on three public datasets. Conclusions and remarks are provided in Sections 5.5.

5.2 Related Work

Applying NNs to face detection dates back to at least early 1990s [167, 168, 169]. Back then, training a multi-layer neural networks was difficult as the number of parameters increases exponentially with the number of layers. However, Deep Neural Network (DNN) is becoming more and more mainstream [37, 170], as it has been shown superior over many other methods, especially for visual recognition tasks. The following can be considered as three of the key reasons that contributed to the success of DNNs. First, training a multi-layer neural network involves finding a local minimum of a highly non-linear function. In order to obtain a reasonable local minimum, gradient descent based methods require a good initialisation. Layer-wise unsupervised pre-training methods [36] were developed and have been proved to be more efficient compared with random initialisation. Second, a large amount of labelled datasets [38, 39, 40] are vitally important to the advance in supervised training. For example, Microsoft COCO dataset [40] contains more than 300,000 images, and over 2,000,000 instances from 80 object categories, where each image has 5 caption labels. Moreover, advances in hardware makes both forward pass and backward propagation computationally efficient. Especially, with dedicated high speed memory module and Single Instruction Multiple Data (SIMD) architecture, General-Purpose Graphics Processing Unit (GPGPU) are particularly well placed for learning deep neural network structures [41].

As for face detection, Farfade *et al.* [84] proposed a multi-view face detection method, so-called Deep Dense Face Detector (DDFD), which uses a fine-tuned 8-layer AlexNet [28] that was initially designed for object recognition. It has 5 convolutional layers and 3 fully connected layers. A pre-trained AlexNet was fine-tuned for face detection on 200,000 face patches and 20,000,000 background patches, which were all resized to 227×227 pixels in order to match the input size of AlexNet. During the testing stage, the sliding window approach was used to generate hypotheses. DDFD classifies each candidates into face or background, and decision

confidence scores is obtained. Non-Maximal Suppression (NMS) was followed to remove redundant bounding boxes.

In [154], the authors introduced a deep CNN based deformable part model for face detection. The whole face is decomposed into 5 facial regions: hair, eye, nose, mouth and beard. The part detectors are constructed using 5 binary CNN classifiers that shared the same deep layers for computational efficiency. The window candidates are generated using object proposal methods, such as selective search [163]. The confidence scores of each candidate can then be inferred via examining the spatial configurations of part detector responses. Finally, to further refine the detection results, a CNN with similar architecture to AlexNet is trained for face-background classification and bounding box regression.

FCN [85] was firstly introduced for semantic segmentation, and then adapted to solve object detection problems. In contrast to classifying each object hypothesis into face or background, FCN based approaches take the whole image as input, and the convolutional outputs of forward pass are considered as a set of feature maps. Detection then can be achieved by investigating the region pattern of the target object on the feature maps. UnitBox [86] is derived from VGG-16 [87] model, and replaces the original fully connected layer with two pixel-wise bounding box prediction layers. The network can then be trained via minimising the IoU loss, which quantitatively measures how well the predicted bounding boxes are aligned with ground truths. Yang *et al.* [88] and Bai *et al.* [89] also show that incorporating a FCN with multi-scale strategy helps to boost detection accuracy. In addition to solving detection problems, it is common to use those deep models for multi-tasks jointly, such as fiducial landmark localization, face pose estimation, gender recognition, and 3D face modeling [90, 91, 92, 93].

Very recently, several works have shown that Regions with CNN features (R-CNN) [81, 82] and Spatial Pyramid Pooling CNNs (SPPnet) [83] are effective in simultaneous object localisation and recognition. These methods contain four main components: convolutional feature extraction, obtaining region proposal, region of interest (ROI) classification, and bounding box refinement. In [81], the authors showed that the representation feature learnt with CNN using deep structure can be effectively used for visual classification and ROI regression. By introducing spatial pyramidal pooling layer to generate a fixed length output feature regardless the size of input image, [83] overcame the limitation of [81] without cropping or wrapping the images that are problematic as they result in information loss and distortion. The work in [82] improved the computational efficiency further by sharing the deep convolutional layers with

region proposal, classification and regression networks. However, for small objects, R-CNNs have difficulty to detect them in small scale due to low resolution and the lack of visual context.

Although deeper models generally outperforms shallow ones, training complex models is not a trivial task, especially for binary detection problems where the distribution of target object and background is extremely unbalanced. Given millions of parameters to optimise using back-propagation, deep nets have the tendency to overfit the data, even with strong regularisations such as dropout and batch normalisation. Due to the smaller amount of parameters, training shallow nets is significantly faster. Embedding shallow nets into traditional cascade framework can also significantly reduce the number of stages and drastically increase the discriminative power of the model [171]. One of limitation of shallow nets is that the recall rate drops quickly with the increase of the number of stages. In this chapter, we introduce a nested soft cascade to compensate the loss of recall while adding multiple stages to remove false positives.

Yang *et al.* [172] proposed a multi-scale cascade CNN, where 4 proposal nets and 4 FCN detectors with different resolutions were used in an ensemble fashion. The input image passes through 4 detection procedures, and the results are combined at the end. Ensembling multiple detectors boosts the accuracy, however, the computational efficiency is the bottleneck. The most relevant work to ours is [155], where 3 face-nonface classification CNNs are used for separating face regions from background and 3 calibration CNNs are used to refine the location of detected bounding box. Sliding window method is used to generate region candidates. These hypotheses pass through 3 classification-refinement components with different image resolutions, from coarse to fine, and the retained ones are considered as object regions. However, a cascade based method has to make a compromise between the number of stages, accuracy and efficiency. For example, in a hard cascade setting, adding more stages helps to reduce false positives, while it decreases the detection rate and speed, especially when a computationally intensive model is used such as CNN. In addition, refining the detected windows between stages introduces re-sampling the patches from the original image, which is non-trivial during the testing phase.

5.3 Soft Cascade

The hard-cascade method progressively eliminates the negative hypotheses by individual stage classifiers, while some of the hard positives are also dropped out during this filtering process. Inspired by [160], a nested soft-cascade is introduced as an ensemble classifier that averages

the predictions over multiple stages. In this section, we present a multi-resolution face detector, which embeds 5 shallow CNN classifiers into a nested cascade framework. Detecting faces in images is carried out in three phases: (1) A large amount of easy background patches from the whole hypothesis set generated by the sliding window are eliminated at the very early stage using a shallow but fast net at a coarse scale. (2) A nested soft cascade with 3 nets is used to further reject hard false positive hypotheses while keeping a high recall rate. (3) To precisely locate the face region, all retained hypotheses from previous stages are verified by a deeper net using higher resolution.

5.3.1 Proposed Method

Fig. 5.1 shows the basic flowchart of the proposed nested cascade face detector. It consists of three main phases as follows: fast elimination, nested soft cascade, and precise detection. Window patches are firstly generated by densely scanning the input image at multiple scales using sliding windows. Majority of those window patches are quickly eliminated as background by an *ElmNet* using a patch resolution of 12×12. A soft-cascade is built by combining 3 *LocNets* in a weighted fashion, which is used to further reject the hard false positives with a patch resolution of 24×24. Then, all retained candidates from the previous stages are verified by *DetNet* using a patch resolution of 48×48. The final detections are obtained via removing redundant detections with Non-Maximum Suppression (NMS).

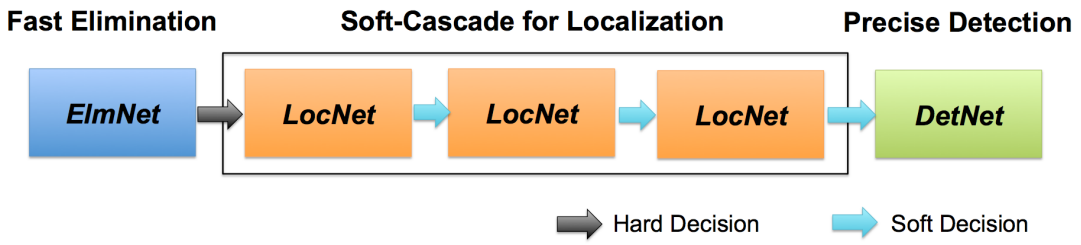


Figure 5.1: The pipeline of the proposed nested cascade face detector.

5.3.1.1 ElmNet: Fast Elimination

A large amount of patch candidates are generated by the sliding window method. The *ElmNet* is designed to quickly eliminate negative patches to reduce the computational cost for the

following phases. Table 5.1 and Fig. 5.2 provide the details of the architecture for *ElmNet*, where only one convolutional layer and one fully connected layer are used. Adopting such simply CNN structure is motivated by the following two reasons. Firstly, *ElmNet* has a small input size of 12×12, a small kernel size of 3×3, and a small number of filters of 16. Compared to other nets, *ElmNet* has significantly smaller number of parameters, which enables a lower memory consumption and a much lower computational cost. Secondly, at this fast elimination stage, low frequency image features extracted from coarse spatial resolution is more effective in rejecting easy negative hypothesis. Since there is no hierarchical feature extraction within *ElmNet*, the discriminative power is limited. In order to retain most positive windows for the following stage, a high recall rate can be achieved by shifting the decision boundary of Softmax layer towards zero. For example, using a minimal face size of 48×48, 87.16% recall can be achieved by shifting the decision boundary to 0.01, whereas 72.62% recall is achieved with 0.50.

Table 5.1: The network architecture of *ElmNet* for fast elimination.

No	Layer Type	Parameter Setting
1	Image Input	12x12x3 images scaled to the range [0,1]
2	Convolution	16 3x3 filters with stride 1
3	ReLU	Rectified linear unit
4	Max Pooling	3x3 filter with stride 2
5	Fully connected	Fully connected with 16 outputs
6	ReLU	Rectified linear unit
7	Fully connected	Fully connected with 2 outputs
8	Softmax	Softmax regression for binary classes
9	Classification	Classification output

5.3.1.2 LocNets: Nested Soft-Cascade

Each stage classifier in cascade is trained using the full set of true positives and the false positives passed through previous stages. Although over 90% of negative patches are eliminated by *ElmNet* at the first stage, the number of retained false positives for training following stage is still considerably large, especially when a large negative image set is used. In our case, 18,089 negative images are used. In order to retain high recall and remove hard non-face hypotheses further, multiple *LocNets* are trained on different subsets of negative images and then assembled in a soft-cascading fashion, where the final decision confidence is a weighted sum

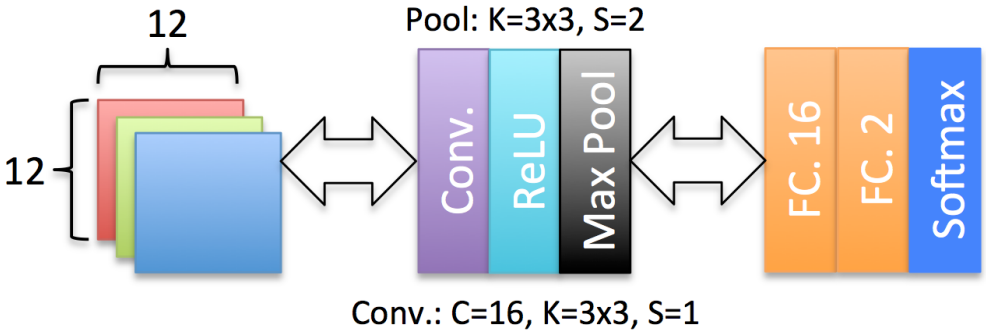


Figure 5.2: Network architecture of *ElmNet*.

of all Softmax outputs of *LocNets*. Within individual *LocNet*, see Table 5.2 and Fig. 5.3, there are two levels of feature abstraction using convolution layers, each of which is followed by a non-linear mapping and a spatial down-sampling. Such hierarchical network enables more discriminative descriptors being learnt through back-propagation, and lifting up from low-level features to high-level representations. The weights of each *LocNet* are estimated using linear regression by solving an over-conditioned least square problem without the interception term. This linear regression problem can be formally defined as

$$\arg\min_{W} \sum_{n=1}^{N} \|L_n - \sum_{s=1}^{S} W_s \times C_{sn}\|^2, \tag{5.1}$$

where W, C, L, S and N denote the weights, probability confidences of face category given by Softmax layers, ground truth labels, the number of LocNet stages, and the number of training samples, respectively. The decision boundary of nested soft-cascade is also shifted to 0.01 in order to achieve a high recall rate.

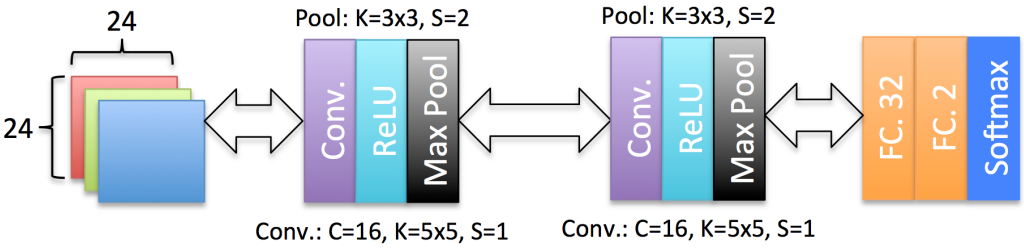


Figure 5.3: Network architecture of *LocNet*.

Table 5.2: The network architecture of *LocNet* for precise localisation.

No	Layer Type	Parameter Setting
1	Image Input	24x24x3 images scaled to the range [0,1]
2	Convolution	16 5x5 filters with stride 1
3	ReLU	Rectified linear unit
4	Max Pooling	3x3 filter with stride 2
5	Convolution	16 5x5 filters with stride 1
6	ReLU	Rectified linear unit
7	Max Pooling	3x3 filter with stride 2
8	Fully connected	Fully connected with 32 outputs
9	ReLU	Rectified linear unit
10	Fully connected	Fully connected with 2 outputs
11	Softmax	Softmax regression for binary classes
12	Classification	Classification output

5.3.1.3 DetNet: Precise Detection

DetNet is designed to precisely locate face regions by verifying retained face candidates at a higher image resolution. In order to capture features in detail, the resolution of input, and the number of filters are doubled compared to *LocNet*, while the size of convolutional kernel and the level of feature abstraction are kept as the same (See Table 5.3 and Fig. 5.4) for computational efficiency. Local response normalisation layers are added between the non-linear mapping layer and the maximum spatial pooling layers. Such inhibition scheme is only applied across channels to enforce regularisation to the networks. Since *DetNet* is the last phase of cascade, binary classification is carried out without shifting the decision boundary, and detected square bounding boxes are then refined using a 2-step NMS to remove redundancies. For the detections at the same scale, we iteratively select the detection with highest confidence score and remove the detections that has the intersection over union (IoU) ratio larger than 0.50 with selected window. Then, for the detections at different scales, the redundancies can be found by measuring the intersection over minimum (IoM) ratio, where the threshold is set to 0.90. The first step removes the redundant detections that are spatially offset to the correct location, and the second step enables removing redundancies in scale.

Table 5.3: The network architecture of *DetNet* for face detector.

No	Layer Type	Parameter Setting
1	Image Input	48x48x3 images scaled to the range [0,1]
2	Convolution	32 5x5 filters with stride 1
3	ReLU	Rectified linear unit
4	Normalisation	Cross channel (9) normalisation
5	Max Pooling	3x3 filter with stride 2
6	Convolution	32 5x5 filters with stride 1
7	ReLU	Rectified linear unit
8	Normalisation	Cross channel (9) normalisation
9	Max Pooling	3x3 filter with stride 2
10	Fully connected	Fully connected with 128 outputs
11	ReLU	Rectified linear unit
12	Fully connected	Fully connected with 2 outputs
13	Softmax	Softmax regression for binary-classes
14	Classification	Classification output

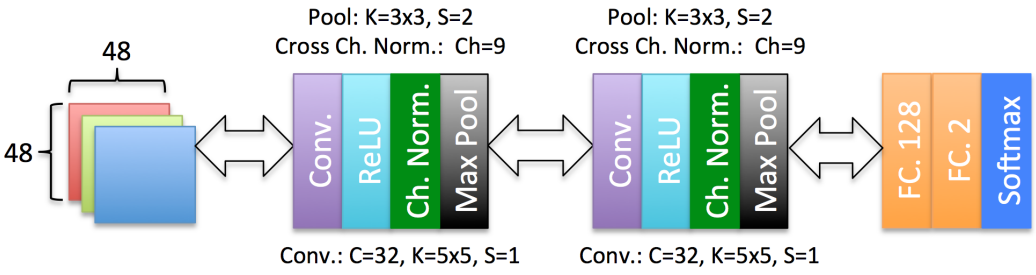


Figure 5.4: Network architecture of *DetNet*.

5.3.2 Experiment and Discussion

5.3.2.1 Detector Training

The AFLW (Annotated Facial Landmarks in the Wild [173]) dataset was used to train the face detector. The dataset contains 22,712 labelled faces out of 21,123 images. The positive face windows were further augmented by horizontal flipping. In total, 45,424 faces were used in the training procedure, and examples of face images are shown in Fig. 5.5 (a). The negative images contain no face. To bootstrap non-face images, labelled face windows were replaced with non-face patches which were randomly sampled from PASCAL VOC dataset [39] (the person subset was excluded). In total, 19,458 negative images were generated using this bootstrapping approach. However, there are considerable amount of unannotated faces in AFLW dataset, we

thus further applied Koestinger's VJ-LBP detector [161] on the negative images. After those ones which have positive response were removed, the negative set contains 18,089 images.

To train *ElmNet*, 904,450 non-face samples were cropped randomly from all negative images (50 patches per image), and then resized to 12×12. With cascading set-up, the negative samples for training the next stages were the residuals (false positives) generated by densely scanning the negative image set using all previous stages. It is useful to set a maximum negative-positive ratio (MNP) for LocNets and DetNet. For example, *ElmNet* would generate over 50 million false positives from 18,089 images, where MNP can thus avoid training with extremely imbalanced data. In our case, we used 48×48 scanning window with the stride of 16 pixels, scale factor of 1.18, and MNP of 10. All networks were trained using back-propagation with batch stochastic gradient descent.

(a) (b)

Figure 5.5: Examples training images. (a) Positive images are cropped face from AFLW dataset; (b) negative images are generated by replacing the face region with non-face patches sampled from PASCAL VOC datasets.

5.3.2.2 Evaluation on FDDB Dataset

The proposed face detector was quantitatively evaluated on the Face Detection Dataset and Benchmark (FDDB) [174] dataset that contains 5,171 annotated faces in 2,845 images. The quantitative results were generated following the standard evaluation procedure with the software provided by the authors. For discrete score evaluation, the detections that has over 0.50 IoU with annotations are counted as true positive. Since the groundtruth faces are labelled using ellipses, we also fitted ellipses to our bounding boxes for fair comparison. However, the faces were labelled using ellipses, whereas our method outputs square bounding boxes. In order to match the ground truth format, the square detection outputs were extended 20% vertically towards the top of image. Then for each upright rectangle, an ellipse was approximated

by setting the centre of ellipse, the length of major and minor axes, to the centre, height and width of rectangle respectively. The angle of major axis with the horizontal axis was set to a constant ($\pi/2$).

Table 5.4 shows the discrete metrics of individual stages of the proposed method using minimal face sizes of 36×36 and 48×48. Over 96% of hypotheses were eliminated, but reasonable recall rate was achieved by ElmNet at the first stage, which ensures deeper network can be computed effectively in the following cascade without overwhelming computational cost. The results of different stages shows that higher resolution and hierarchical feature abstraction are the key to build discriminative models. We also compared proposed method with state-of-the-art methods which are trained on the same dataset. The discrete ROC curves are shown in Fig. 5.6. DPM based methods, such as Yan et al. [175] and HeadHunter [176] are leading the performance, mainly because the variations of facial parts are relatively small, thus detecting facial parts are more robust than detecting face as a whole. Especially, HeadHunter [176] reports the optimal results that obtained through comprehensive studies on training strategies and parameter settings. However, DPM methods require training part detectors, and searching optimal configuration, which make building the detector a laborious, time-consuming task, and are known to be much slower than cascade based methods. ACF-Multiscale [153] method aggregates multiple features, such as colour, gradient, local histogram, into a rich representation, and then trains multiple soft cascade with depth-2 decision tree for different views. It shows that combining multiple models and features outperforms a single model. The computational cost of aggregating feature channels is considerably more. Significantly, Koestinger [161] shows that without rich features, the performance of multi-view based method drops by a significant margin. In addition, sophisticated post-processing is required to combine the multiple detection outputs given by detectors of different views. The proposed method requires no model aggregation. The features are self-learnt through training, and it outperforms the traditional methods which use the cascade framework such as NPDFace [177]. Also image retrieval based methods suffer from efficiency issue much more severely. For example, to process an image of size 1480×986 with minimal face size 80×80, Boosted Exemplar [157], and XZJY [156] take 900ms and 33000ms receptively, whereas our methods only takes 153ms using a non-optimised Matlab implementation.

Qualitative results on the FDDB dataset are shown in Figs. 5.7, 5.8, and 5.9. Red and blue ellipses represent groundtruth and true positives, whereas yellow and green ellipses represent

false positives and false negatives respectively. Fig. 5.7 illustrates some examples of typical detection results with large pose and facial expression variations, blurring, and severe occlusion and clutter. Fig. 5.8 shows some examples of false positives and false negatives. The false positives are usually observed at the region that contains partial face, and false negatives are mainly caused by severe blurring and faces in small scale. Fig. 5.9 shows some interesting detections in yellow, which are counted as false positives since there are no annotations to match. However, they are in fact correct detections.

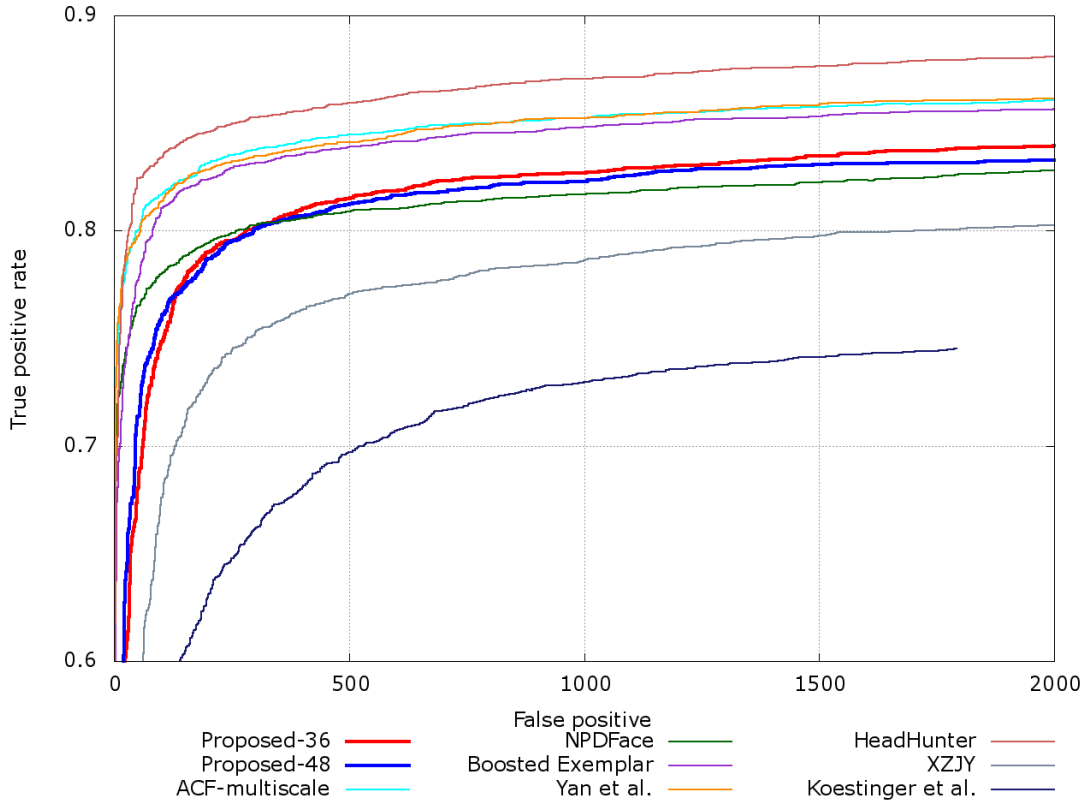


Figure 5.6: ROC curves of the proposed detector and recent methods on FDDB database with the discrete score metric.)

5.3.2.3 Evaluation on AFW Dataset

We quantitatively evaluated our face detector on another face detection benchmark, namely Annotated Face in the Wild (AFW) [178] that contains 205 images, and 468 annotated faces. 97.43% recall rate was achieved by our face detector, which is slightly lower than CNN-

Figure 5.7: Typical detection results on FDDB dataset (red: ground truth, blue: true positive).

Table 5.4: Recall rate and number of false positives of individual detection stage of the proposed method on FDDB dataset.

Stages	36×36 minimal face		48×48 minimal face	
	Dis. Recall	#FP	Dis. Recall	#FP
Hypothesis	95.16%	17843K	91.94%	16033K
ElmNet	90.17%	471K	87.16%	314K
S1-LocNet	88.05%	114K	83.52%	64K
S2-LocNet	85.48%	42K	81.63%	28K
S3-LocNet	83.10%	23K	79.37%	16K
Soft-LocNets	88.74%	117K	85.84%	78K
DetNet	82.38%	723	80.89%	450

Figure 5.8: Examples of false positives and false negatives on FDDB dataset (red: ground truth, blue: true positive, yellow: false positive, green: false negatives).

Figure 5.9: Examples of correct detections but counted as false positives (red: ground truth, blue: true positive, yellow: false positive).

Cascade [155] (97.97%, +0.54%), but outperforms other state of the art methods, such as DPM [158] (97.21%, -0.22%,), HeadHunter [176] (97.14%, -0.29%), Structured Models [179] (95.19%, -2.24%), Shen et al. [156] (89.03%, -8.4%), and TSM [178] (87.99%, -9.44%). Qualitative results are shown in Fig. 5.10, where square detection bounding boxes were used to match the original annotations.

5.3.2.4 Evaluation on CMU-MIT & GENKI Datasets

The proposed method was also evaluated on two early face detection benchmarks, CMU-MIT face dataset [180], and GENKI database [181]. Several examples of typical detection results are presented in Figs. 5.11 and 5.12. CMU-MIT dataset contains a total of 511 faces from

Figure 5.10: Examples of qualitative results on AFW dataset. (green: ground truth, blue: detection results of the proposed method).

130 grey-scale images. The top right image in Fig. 5.11 shows that our method is able to tolerate rotation variance, and there are only one false negative and two false positives in the top right image. Current release of GENKI database contains two subsets, where GENKI-4K subset contains 4,000 images, and GENKI-SZSL subset contains 3,500 images. Some detection examples with different poses and facial expressions are shown in Fig. 5.12.

5.3.2.5 Detection Speed

The proposed detector was implemented and evaluated on *Matlab 2016* using two different GPUs, GeForce GTX TITAN X (Maxwell) and Quadro K2000, which have 3,072 CUDA cores with 12GiB memory and 384 CUDA cores with 2GiB memory respectively. Table 5.5 shows the running speed of individual stages. It can be observed that TITAN X outperforms K2000 as more CUDA cores and GPU memory are available. The computation time increases as the complexity of the model increases. To processes one 640×480 VGA image with the size of minimum face of 80×80, our method takes 40.1ms using CPU only, whereas [155] takes 71ms on average.

We proposed an efficient multi-stage cascade method that is well suited for binary detection problems, where the number of positive samples is significantly smaller than negative samples. Instead of resorting to deep structures that are time consuming and laborious to train, the pro-

Figure 5.11: Examples of qualitative results on CMU-MIT dataset.

Figure 5.12: Examples of qualitative results on GENKI database.

Table 5.5: Speed of individual stage (hypotheses/second)

	ElmNet	LocNet	DetNet
TITAN Maxwell	102,380 ± 4,964	71,844 ± 574	17,599 ± 99
Quadro K2000	61,112 ± 2799	22,988 ± 415	2,383 ± 9

posed nested shallow CNN-cascade overcomes these difficulties by solving three sub-problems from easy to hard using models from weak to strong. In addition, a nested soft cascade is introduced to compensate the loss of recall when multiple classifiers are used to reject a large amount of negatives. The proposed method was evaluated on three datasets including FDDB and AFW. Quantitative and qualitative results show promising performances on detecting face in unconstrained environment with much improved efficiency compared to state of the art.

5.4 Detection-Regression Cascade

The performance of the soft-cascade method is bounded by two major factors. First, the feasibility of mining discriminative features is limited by the depth and architecture of the CNN that is used. Second, we found that there are some false positive detections which are very closed to the true positive. However, due to the IoU being lower than 0.5, they are counted as negatives. Feature aggregation and multi-resolution strategies were proved to be the efficient schemes for visual recognition tasks with hand-crafted features [153, 182]. In this section, we show that introducing such strategies into CNN architecture design also helps improving the accuracy of the challenging face detection problem. The proposed Multi-Resolution Feature Aggregation (MRFA) face detector embeds a fast elimination stage, and two verification stages into a cascade framework. A large amount of easy background patches from the whole hypothesis set generated by sliding window are eliminated at the very early stage using a shallow but fast net at a coarse scale. To precisely locate the face region, verification nets are designed with feature aggregation at multi-resolution via average pooling and channel-wise feature concatenation. The face-nonface binary decision is first made by the detection classifier, and then for all positive predictions, a regression procedure is applied to refine the locations, aspect ratios of major and minor axes, and angles of output bounding boxes.

5.4.1 Proposed Method

Fig. 5.13 shows the basic flowchart of the proposed MRFA face detector. It consists of two main phases: fast elimination, and precise verification. Window patches are firstly generated by densely scanning the input image at multiple scales using sliding windows. The majority of window patches are quickly eliminated as background by an *ElmNet* using a patch resolution of 12×12. Then, all retained candidates from *ElmNet* are verified by two *VefNets* using a patch

resolution of 48×48. At the end of cascade, the detection branch outputs the binary classification of face-nonface decision with confidence scores, meanwhile the regression branch refines the bounding box location by determining the optimal face center, angle, and aspect ratio. The final detections are obtained via removing redundant detections with a 2-step NMS.

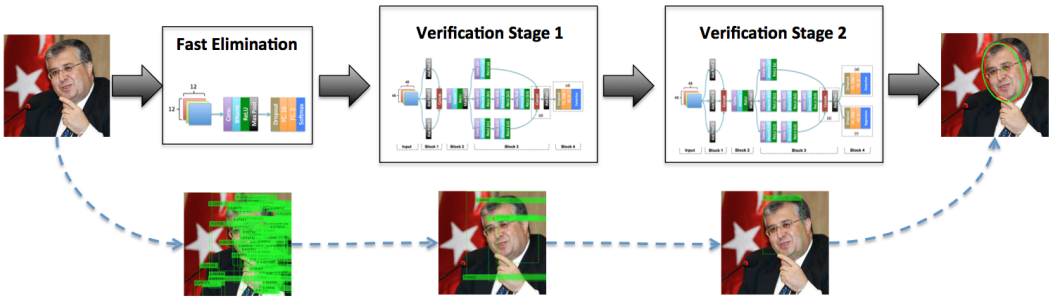


Figure 5.13: The pipeline of the proposed MRFA detector.

5.4.1.1 Sliding Window Elimination Net

A large amount of patch candidates are generated by the sliding window method. The *ElmNet* proposed in Section 5.3 are reused as fast elimination classifier, where a batch normalisation layer and a drop-out layer are added to regularise the training process for current method. Table 5.6 and Fig. 5.14 provide the details of the architecture for *ElmNet* with a batch normalisation layer and a drop-out layer. Since there is no hierarchical feature extraction within *ElmNet*, the discriminative power is limited. In order to retain most positive windows for the following stage, a high recall rate can be achieved by shifting the decision boundary of Softmax layer towards zero. For example, using a minimal face size of 48×48, 91.12% recall can be achieved on FDDB dataset by shifting the decision boundary to 0.01. Compared to the original *ElmNet* (87.16%), the regularised *ElmNet* has a higher recall rate.

5.4.1.2 Multi-Task Verification Net

A multi-task *VefNet* is designed to precisely locate face regions by verifying retained face candidates at a higher image resolution of 48×48. Table 5.7 and Fig. 5.15 provide the details of the architecture, where *VefNet* is divided into 4 main blocks.

Block 1 consists of 3 average pooling branches which use three filters (1×1, 3×3, and 5×5 respectively) with no spatial down-sampling. Three pooling branches joint together via

Table 5.6: The network architecture of *ElmNet* with a batch normalisation and a drop-out regularisations for fast elimination.

No	Type	Parameter
0	Input	12x12x3 images scaled to [0,1]
1	Conv.	16 3x3 filters with stride 1
2	BNorm.	$\varepsilon = 0.0001$
3	ReLU	Rectified linear unit
4	Max Pool	3x3 filter with stride 2
5	Dropout	0.20 dropout rate
6	F.C.	Fully connected with 16 outputs
7	F.C.	Fully connected with 2 outputs
8	Softmax	Softmax probability for binary classes

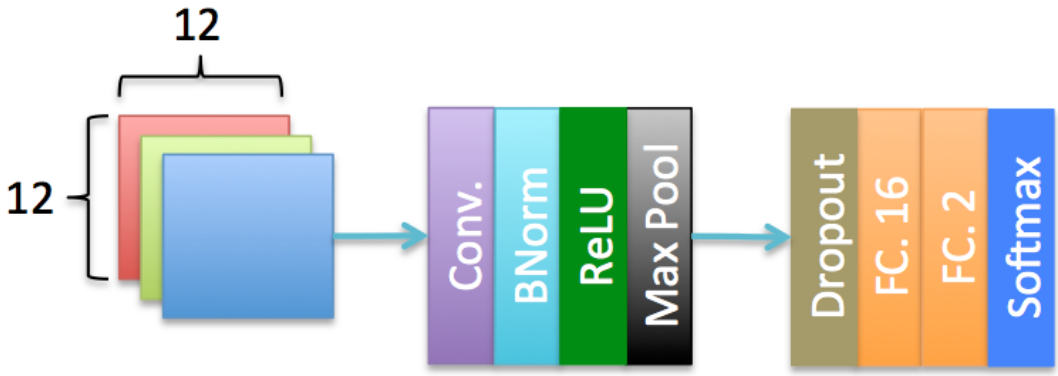


Figure 5.14: Network architecture of *ElmNet* with batch normalisation and drop-out regularisations.

concatenating the outputs across channels. In contrast to traditional multi-scale methods that construct Gaussian pyramid as network input, such structure embeds a simple average blurring scheme into network itself, which greatly helps the later computational blocks to identify scale-invariant features.

Block 2 extracts the first level of visual features via sequentially passing the multi-resolution images through a convolutional layer, a batch normalisation layer, a ReLU non-linear transform layer, and a max pooling layer. In order to gain high speed efficiency, we aggressively reduce the spatial resolution by setting the strides of convolutional layer and max pooling layer both to 2. Batch normalisation layer is inserted between convolutional layers and ReLU layers (same for other blocks) to enforce regularisation to internal co-variate shift caused by weight updates

during back-prop. Inspired by GoogLeNet [183], a simplified inception module which contains three feature extraction branches, is used to generalise discriminative power further.

Each branch in Block 3 starts with a dimensionality reduction module with a 1×1 convolutional layers which removes redundant feature channels, and improves computational efficiency. Block 3.b and 3.c consist of two 3×3, and one 5×5 feature extraction modules respectively. It is worth noting that although a 5×5 filters has the same reception field as two consecutive 3×3 filters, the later could generalise even deeper structures. The outputs of three branches in Block 3 are concatenated across channels, and then followed by an average pooling layer to reduce the spatial resolution. Yang *et al.* [153] shows that aggregating hand-crafted features improves the detection accuracy. In our method Block 3 embeds such multi-level feature interfusion into a learnable framework.

Block 4 contains two fully connected objective branches, detection branch and bounding box regression branch (top row and bottom row of Block 4 in Fig. 5.15 respectively). Previous blocks are trained with detection branch using Softmax loss, whereas regression branch is trained using smooth $\ell 1$ loss. Binary classification is carried out by the detection branch without shifting the decision boundary at the last stage.

As the face candidates are generated by sliding window, the optimal locations of faces may not be in the hypothesis set. The detection performance can be further boosted by refining the locations of output bounding boxes. The regression target is a quintuple defined by two coordinates of face center offset to top left corner, lengths of major and minor axes with respect to the size of bounding box, and the angle of major axis with vertical axis. A positive value of face angle indicates an anti-clock rotation with respect to vertical axis. The bounding box calibration procedure only applies to the positive response given by detection branch. Then a 2-step NMS is followed to remove redundancies. For the detections at the same scale, we iteratively select the detection with highest confidence score and remove the detections that has the IoU ratio larger than 0.50 with selected window. For the detections at different scales, the redundancies can be found by measuring the Intersection over Minimum (IoM) ratio, where the threshold is set to 0.75. The first step removes the redundant detections that are spatially offset to the correct location, and the second step enables removing redundancies in scale.

Table 5.7: The network architecture of multi-task *VefNet* for face detection and bounding box regression.

BK.	No	Type	Parameter
	0	Input	48x48x3 images to [0,1]
1	1.1	Ave. Pool 1	1x1 filter with stride 1
	1.2	Ave. Pool 3	3x3 filter with stride 1
	1.3	Ave. Pool 5	5x5 filter with stride 1
	1.4	Concat.	Concat. 2.1, 2.2, 2.3
2	2.1	Conv.	96 3x3 filter with stride 2
	2.2	BNorm.	$\varepsilon = 0.0001$
	2.3	ReLU	Rectified linear unit
	2.4	Max Pool	3x3 filter with stride 2
3.a	3.a.1	Conv.	32 1x1 filter with stride 1
	3.a.2	BNorm.	$\varepsilon = 0.0001$
	3.a.3	ReLU	Rectified linear unit
3.b	3.b1.1	Conv.	32 1x1 filter with stride 1
	3.b1.2	BNorm.	$\varepsilon = 0.0001$
	3.b1.3	ReLU	Rectified linear unit
	3.b2.1	Conv.	48 3x3 filter with stride 1
	3.b2.2	BNorm.	$\varepsilon = 0.0001$
	3.b2.3	ReLU	Rectified linear unit
	3.b3.1	Conv.	48 3x3 filter with stride 1
	3.b3.2	BNorm.	$\varepsilon = 0.0001$
	3.b3.3	ReLU	Rectified linear unit
3.c	3.c1.1	Conv.	24 1x1 filter with stride 1
	3.c1.2	BNorm.	$\varepsilon = 0.0001$
	3.c1.3	ReLU	Rectified linear unit
	3.c2.1	Conv.	32 5x5 filter with stride 1
	3.c2.2	BNorm.	$\varepsilon = 0.0001$
	3.c2.3	ReLU	Rectified linear unit
3.d	3.d.1	Concat.	Concat. outputs of 3.a-3.c
	3.d.2	Ave. Pool	2x2 filter with stride 2
4.c	4.d.1	Dropout	0.20 dropout rate
	4.d.2	F.C.	with 256 outputs
	4.d.3	F.C.	with 2 outputs
	4.d.4	Softmax	Binary classification
4.r	4.r.1	Dropout	0.20 dropout rate
	4.r.2	F.C.	with 256 outputs
	4.r.3	F.C.	with 5 outputs
	4.r.4	Smooth $\ell 1$	Bounding box regression

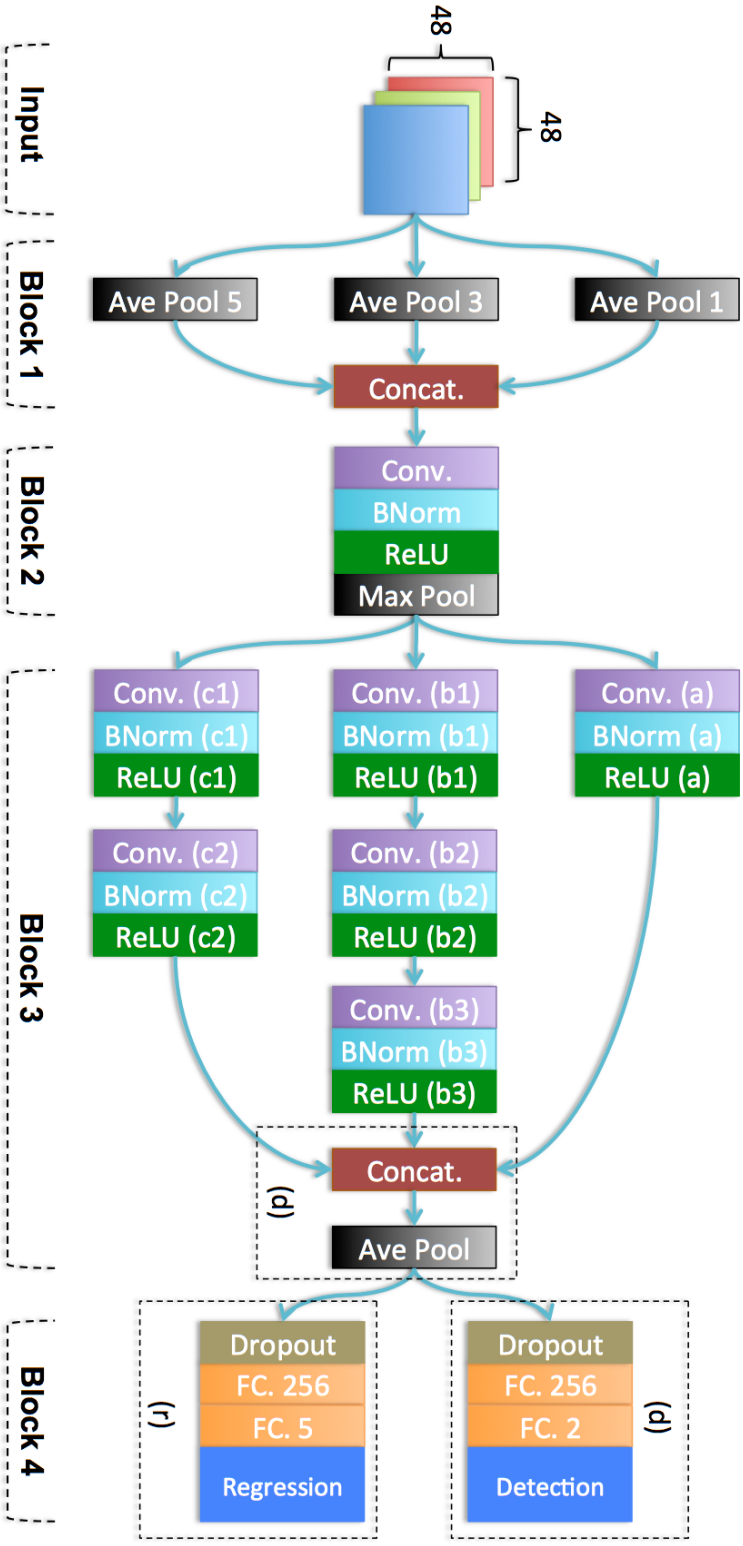


Figure 5.15: Network architecture of *VefNet*.

5.4.2 Experiment and Discussion

5.4.2.1 Detector Training

The Annotated Facial Landmarks in the Wild (AFLW) [173] dataset was used to train the face detector. To train *ElmNet*, non-face samples were cropped randomly from negative images, and then resized to 12×12. The negative-positive ratio of *ElmNet* was set to 10:1. With cascading set-up, the negative samples for training the detection branch of *VefNet* were the residuals (false positives) generated by densely scanning the negative image set using previous stages. Since a large number of negative samples were generated by shifting the decision boundary of *ElmNet* to 0.01, in order to keep the negative-positive ratio of *VefNet* under 5:1, additional positive samples were added by randomly blurring the original face images. The maximum sigma value for Gaussian blurring was set to 2. The networks were trained using MatConvNet [184]. The number of epoch was set to 50, and a size of mini-batch was 128, and momentum of 0.9 were used. The learning rate gradually drops from $1e^{-2}$ to $1e^{-5}$. Regression branch of *VefNet* was trained independently, and it converges in 2 epochs.

5.4.2.2 Evaluation on FDDB

The proposed face detector was quantitatively evaluated on the FDDB [174] dataset. The quantitative results were generated following the standard evaluation procedure with the software provided by the authors. For discrete score evaluation, detections with over 0.50 IoU with annotations are counted as true positive. Since the ground truth faces are labeled using ellipses, for fair comparison, we also fitted ellipses to our bounding boxes given the outputs of regression branch of *VefNet*. Table 5.8 shows the discrete metrics of individual stages of the proposed method using minimal face sizes of 36×36 and 48×48. Over 90% of hypotheses were eliminated, but reasonable recall rate was achieved by *ElmNet* at the first stage, which ensures deeper network can be computed effectively in the following cascade without overwhelming computational cost. The results of different stages show that higher resolution and hierarchical feature abstraction are the key to build discriminative models. At the last stage of cascade, majority of retained hypotheses are distributed around the face regions. A simple bounding box regression is used in the proposed method, which helps to tie those hypotheses together, and then NMS retains the ones with highest confidence scores while removing the redundancies. We also compared proposed method with state-of-the-art representative methods which are evaluated

on the same dataset. Table 5.9 shows the comparison of discrete and continuous detection rates given the number of false positives, and the discrete ROC curves are shown in Fig. 5.16. DPM based methods, Faceness [154] is leading the performance, mainly because the variations of facial parts are relatively small, thus detecting facial parts are more robust than detecting face as a whole. However, DPM methods require training part detectors, and searching optimal configuration, which makes building the detector a laborious, time-consuming task, and are known to be much slower than cascade based methods. CasCNN [155] refines the bounding boxes between each stage, and then re-fetches the image patches for the next stage. Such procedure does improve both discrete and continuous scores, however it is non-trivial. Our method only applies simple location calibration at the last stage, and no re-fetching is required. Compared to Deep Dense Face Detector (DDFD) where a deeper structure is used, the proposed MRFA achieved higher True Positive (TP) rate after 100 false positives, and outperformed it by a significant margin (6.9% higher) at 500 false positives. ACF-Multiscale [153] method aggregates multiple features, such as colour, gradient, local histogram, into a rich representation, and then trains multiple soft cascade with depth-2 decision tree for different views. It shows that combining multiple models and features outperforms a single model. The computational cost of aggregating feature channels is considerably more expensive. Significantly, Koestinger [161] shows that without rich features, the performance of multi-view based method drops by a significant margin. In addition, sophisticated post-processing is required to combine the multiple detection outputs given by detectors of different views. The proposed method embeds feature aggregation and multi-resolution strategies into the network architecture. The features are self-learned through training, and it outperforms the traditional methods which use the cascade framework, and hand-crafted features, such as NPDFace [177], ACF-Multiscale [153], and Koestinger [161]. For image retrieval based methods, such as Boosted Exemplar [157], they generally have higher recall rate compared to those traditional methods, however, our method outperformed [157] it in all aspects. Those methods also suffer from severe efficiency issues. For example, to process an image of size 1480×986 with minimal face size 80×80, Boosted Exemplar [157] takes 900ms, whereas our methods only requires 537ms using a non-optimised Matlab implementation.

Qualitative results on the FDDB dataset are shown in Figs. 5.17, 5.18, and 5.19. Red and blue ellipses represent ground truth and true positives, whereas yellow and green ellipses represent false positives and false negatives respectively. Fig. 5.17 illustrates some examples of

Table 5.8: Recall rate and number of false positives of individual detection stage of the proposed method on FDDB dataset.

Stages	36×36 mini face		48×48 mini face	
	Dis. TP	#FP	Dis. TP	#FP
Hypotheses	*N.A.*	30.42M	*N.A.*	15.43M
ElmNet	94.64%	2.69M	91.12%	1.39M
VefNet S1	90.99%	76.52K	88.70%	58.97K
VefNet S2	85.34%	6140	83.56%	5873
Reg. & NMS	86.09%	246	83.93%	186

Table 5.9: Comparison of detection rates (%) with both discrete and continuous metrics on FDDB.

	Discrete Metric			
	FP=25	FP=50	FP=100	FP=500
Proposed MRFA	75.22	78.92	82.77	87.89
Faceness [154]	85.81	86.87	87.64	89.38
CasCNN [155]	81.26	83.48	85.07	*N.A.*
DDFD [84]	75.40	79.23	80.99	83.40
BoostedExampler [157]	74.14	77.66	80.82	83.89
NPDFace [177]	74.53	76.50	77.97	80.89
ACF-Multiscale [153]	78.21	80.00	81.65	84.45
Koestinger [161]	36.34	47.22	57.03	69.70
	Continuous Metric			
	FP=25	FP=50	FP=100	FP=500
Proposed MRFA	59.46	62.27	65.13	68.85
Faceness [154]	68.18	69.01	69.70	71.38
CasCNN [155]	63.49	65.11	66.29	*N.A.*
DDFD [84]	60.38	63.12	64.45	66.41
BoostedExampler [157]	52.14	54.64	56.87	59.20
NPDFace [177]	55.54	56.96	58.04	60.25
ACF-Multiscale [153]	57.90	59.20	60.43	62.49
Koestinger [161]	25.96	33.65	40.55	49.49

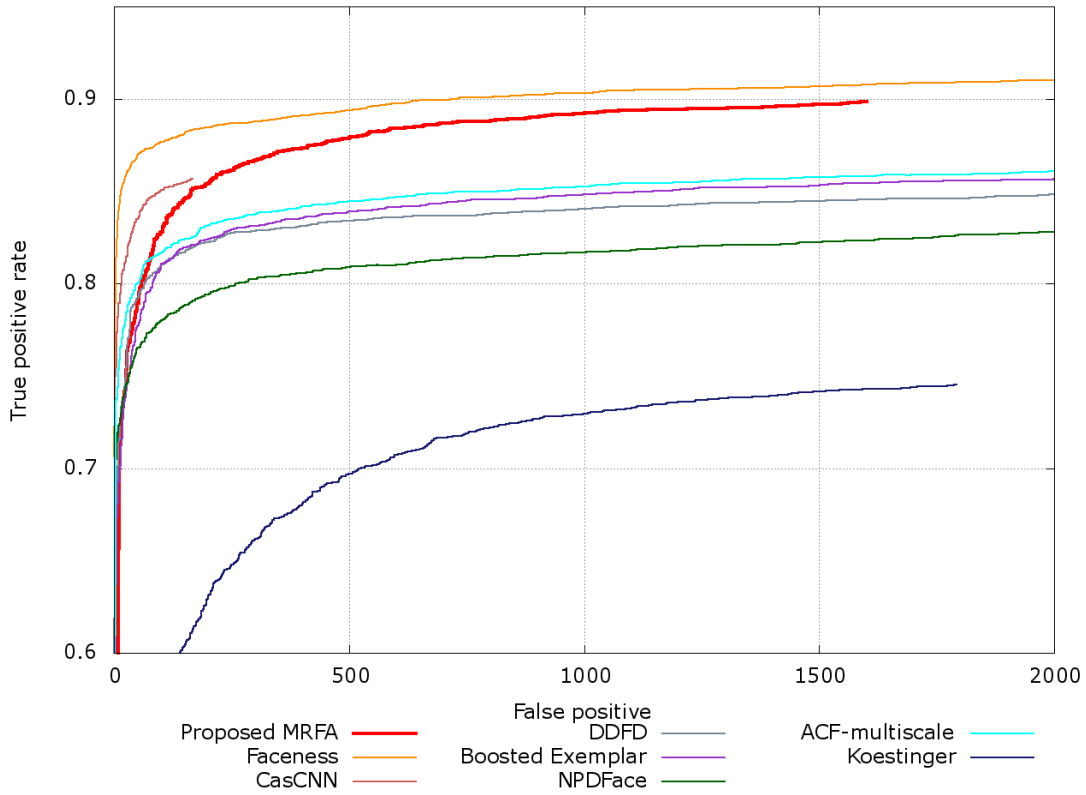


Figure 5.16: ROC curves of the proposed detector and recent methods on FDDB with the discrete score metric.)

typical detection results with large pose and facial expression variations, blurring, and severe occlusion and clutter. Fig. 5.18 shows some examples of false positives and false negatives. The false positives are usually observed at the region that contains partial face, and false negatives are mainly caused by severe blurring and faces in small scale. Fig. 5.19 shows some interesting detections in yellow, which are counted as false positives since there are no annotations to match. However, they are in fact correct detections.

5.4.2.3 Evaluation on CMU-MIT & GENKI

The proposed method was also evaluated on CMU-MIT [180], and GENKI [181]. Fig. 5.20 shows some typical detection results on CMU-MIT and GENKI datasets, in (a) and (b) respectively. In Fig. 5.20(a), there is no false positive observed, and the only false negative can be found is in the right bottom, at the edge of the image, mainly because more than half of the face

Figure 5.17: Typical detection results on FDDB dataset (red: ground truth, blue: true positive).

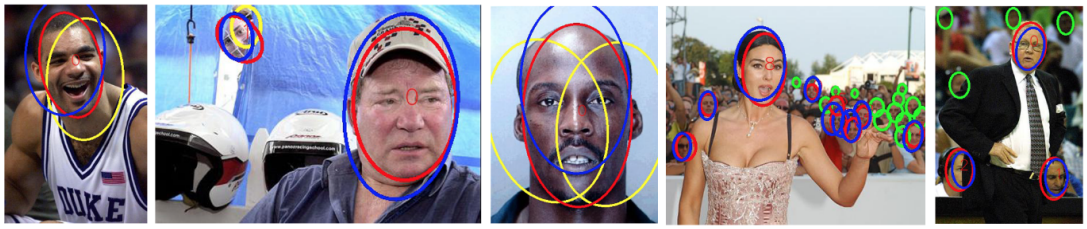


Figure 5.18: Examples of false positives and false negatives on FDDB dataset (red: ground truth, blue: true positive, yellow: false positive, green: false negatives).

Figure 5.19: Examples of correct detections but counted as false positives (red: ground truth, blue: true positive, yellow: false positive).

is occluded. In Fig. 5.20(b), we show that the proposed method is able to tolerate exaggerated facial expressions.

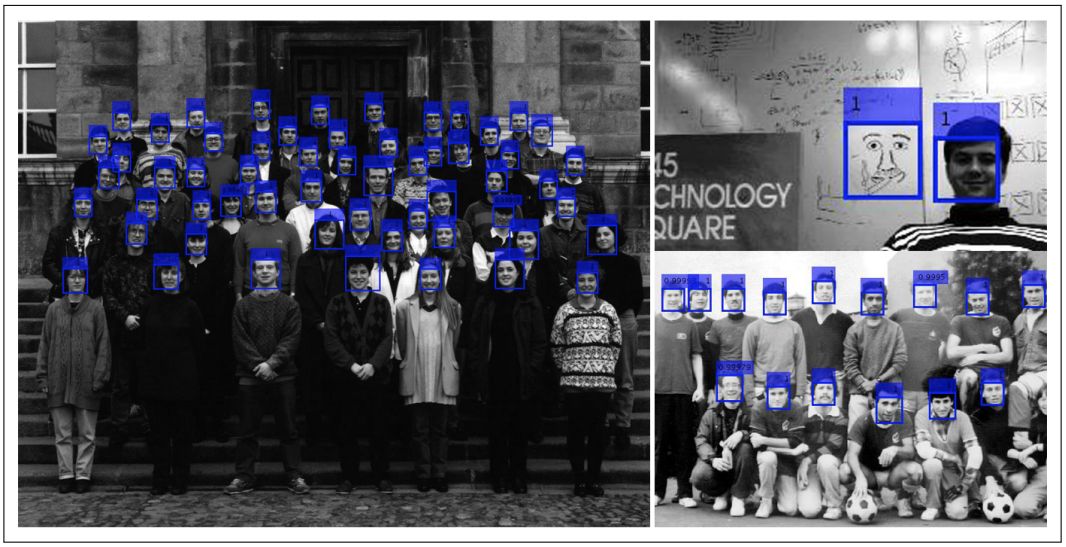
5.4.2.4 Computational Efficiency

The proposed detector was implemented and evaluated on MatConvNet *Matlab 2016 version* using GeForce GTX TITAN X (Maxwell), which has 3,072 CUDA cores with 12GiB memory. Table 5.10 shows the training and testing speed, and corresponding GPU load for individual stage. The training and testing speeds of regression branch are negligible, since it shares the feature computational blocks with detection branches. The computation time increases as the complexity of the model increases. However, the GPU is not fully loaded, and detection speed can be further boosted via parallelization and optimizing execution order of individual branches in *VefNet*.

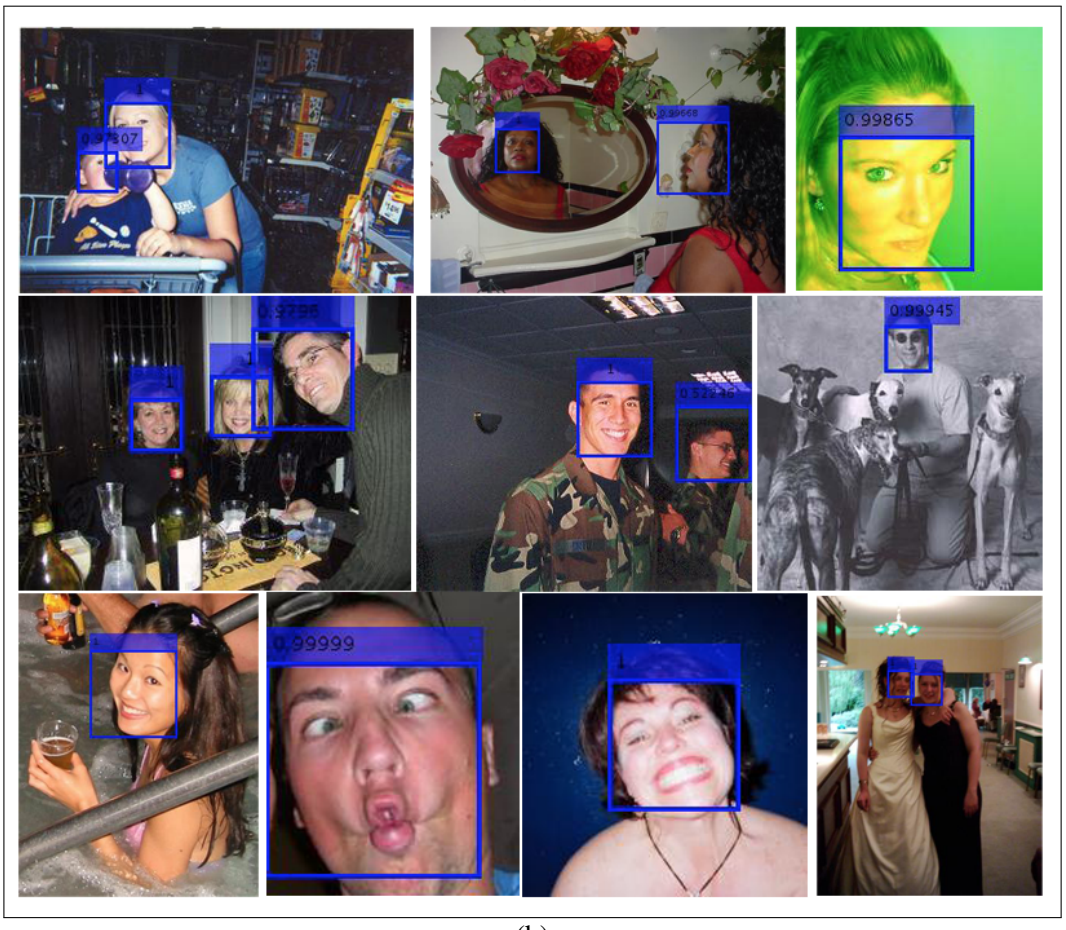
Table 5.10: Speed (samples/second) and GPU load (usage percentage and memory consumption) of individual stages.

	ElmNet	*VefNet*
Train Speed	13,602 $\pm$ 286	2578 $\pm$ 22 %
Train GPU Load	24%, 469Mib	40%, 717MiB
Test Speed	193,060 $\pm$ 19,153	872 $\pm$ 156
Test GPU Load	7%, 497Mib	52%, 4,040Mib

We proposed an efficient multi-stage cascade method that is well suited for binary detection problems, where the number of positive samples is significantly smaller than negative samples. Instead of resorting to deep structures that are time consuming and laborious to train, the proposed MRFA overcomes these difficulties by embedding multi-resolution and feature aggregation into shallow networks. Considering computational efficiency, it combines fast elimination and precise verification into a cascade framework. The proposed method was evaluated on three datasets including FDDB. Quantitative and qualitative results show promising performances on detecting face in unconstrained environment with much improved efficiency while introducing multi-resolution feature aggregation into the network architecture.



(a)



(b)

Figure 5.20: Examples of qualitative results on CMU-MIT and GENKI datasets, showing in (a) and (b) respectively.

5.5 Summary

Face detection in the wild is a challenging vision problem due to large variations and unpredictable ambiguities which commonly exist in real world images. Whilst introducing powerful but complex models is often computationally inefficient, using hand-crafted features is also problematic. Some recent works on adapting pre-trained large scale recognition models to face detection problem often requires excessive resource expenditure. In this chapter, we proposed two CNN based cascade detection methods to overcome the difficulties that are introduced by the human face detection problems under an unconstrained environment.

In section 5.3, we propose a nested CNN-cascade learning algorithm that adopts shallow neural network architectures that allow efficient and progressive elimination of negative hypotheses from easy to hard via self-learning discriminative representations from coarse to fine scales. The face detection problem is considered as solving three sub-problems: eliminating easy background with a simple but fast model, then localising the face region with a soft-cascade, followed by precise detection and localisation by verifying retained regions with a deeper and stronger model. The face detectors are trained on the AFLW dataset following the standard evaluation procedure, and the method is tested on four other public datasets, i.e. FDDB, AFW, CMU-MIT and GENKI. Both quantitative and qualitative results on FDDB and AFW are reported, which show promising performances on detecting faces in unconstrained environment.

Feature aggregation and multi-resolution are two efficient strategies for traditional visual recognition methods. In section 5.4, we show that such strategies can be integrated into the architecture design of CNN via average pooling and channel-wise feature concatenation. Shallow networks with feature aggregation at multi-resolution enables the traditional cascade framework to tackle the challenging detection problems efficiently. The proposed method is tested on three public benchmarks with across dataset evaluation. Both quantitative and qualitative results show promising performance improvements on detecting faces in unconstrained environment. It leverages recent advances in CNNs for efficient face detection, where deep structure and large scale model adapting that require excessive resources, such as training data and time on both pre-trained and adapted models, are avoided. Our proposed solution is not overwhelmed by the model complexity.

Chapter 6

Conclusion and Future Work

Contents

6.1 Conclusion 128
6.2 Future Work 130

6.1 Conclusion

In this thesis, we have investigated the feasibility of applying adaptive learning to both medical image analysis and generic computer vision problems, more specifically, medical image segmentation and face detection. For medical image segmentation, we tried to fill the knowledge gap between computer scientists and experienced radiologists via introducing semi-automatic segmentation schemes. The adaptive learning leads to an interactive image segmentation methods where the user's interpretation of image and intermediate result can be progressively integrated via learning an incremental classification model, or fine-tuning existed model. In addition, adaptive learning can also lead to a problem subdivision scheme, where a difficult classification problem can divided into several simple sub-problems that can be solved adaptively using a set of weaker classifiers. The final solution can then be found via combining individual sub-solvers in a cascade fashion. In this thesis, we have shown that both schemes that derived from adaptive learning strategies are efficient to address coronary artery and aorta segmentation problems, and the face detection problem. The main contributions are summarized as follows.

- **An interactive segmentation method with minimal elastic user input.** This robust

method was developed for use in segmenting the coronary artery from CTA volumetric image. We first proposed a multi-scale vessel feature based on the eigen system of the Hessian matrix, where an effective classifier can be built. An initial vessel classification is given by a RF classifier which is trained on a few user strokes: the foreground stroke labels the coronary artery and the background stroke indicates the other tissues. Based on the label population in the leaf nodes of the randomised decision trees, we formulated the final segmentation as an MRF based optimisation with local consistency constraints. Promising segmentation results were achieved with just a few user strokes.

- **Feature awareness adaptive learning for complex anatomy segmentation.** This robust method was developed for use in segmenting aorta arch and root from CTA volumetric images. A cascade detector was proposed to efficiently delinearate the foreground objects and background regions. It consists of an intensity-based Naive-Bayesian classifier for fast elimination, and a pseudo-3D CNN classifier for precise classification. The representative features for region-based detection are automatically learnt in a supervised fashion together with the decision boundary for binary classification, hence, no hand-feature-crafting is needed. Adaptive learning and localised refining strategies were introduced which further improve the detection result and boosts the accuracy with help of user intervention.

- **Topology awareness implicit shape representation and deformation.** We presented a novel parametric implicit representation method that blends the level set of a shape using locally supported B-spline patches. The control knots are placed according to the complexity density that is estimated using wavelet coefficient. It is able to adapt according to the local topology, where highly curved regions are blended using more compact patches to avoid over-smoothing or adding unnecessary knots. We also derived the formulation of shape deformation based on regional data support that can then be used to impose the piecewise constant for segmentation purpose.

- **Depth adaptive CNN cascade detection.** We proposed an adaptive cascade scheme that the depth of CNN model is progressively increased with the increase of stages. Nested soft decision method, feature aggregation via average pooling and channel-wise feature concatenation, and multi-task training schemes for CNN-based cascade were investigated which have proved to be effective to boost the detection accuracy.

- **Prototype of integrated segmentation platform, *SVMIST*.** We developed a segmentation platform which contains the following essential functionalities: image dataset management, 3D viewing, volumetric rendering, and interactive labelling. The coronary segmentation method proposed in this thesis has been integrated into *SVMIST*. In addition, it is also an practical software for data labelling, for example, the segmentation ground-truth for aorta segmentation was labelled using *SVMIST*.

6.2 Future Work

There are three key areas which we believe this work can be built upon.

- **Dataset:** At the time of this work, we had a relatively limited medical dataset to evaluate our proposed methods. We had 36 labelled CTA volumes for evaluating aorta segmentation method, whereas there is no labelled data to carry out quantitative evaluation for coronary artery segmentation method due to small size and poor connectivity. Labelling medical data requires related background knowledge and clinical experience, and it is also very time consuming. The labelling process generally involves identifying the locations of pathological changes, labelling the ROI with opened or closed contours, and then assigning with semantic annotation, where cross validation is required to ensure the labelling quality. The process of labelling a 3D volumetric data is often performed in a slice by slice fashion, sometimes it has to be applied on multiple separated views independently. Hence, large labelled 3D medical datasets are very rare and not readily available for researchers in general. Fortunately, the situation is improving due to the joint effort of computer scientists and radiologists. For example, the grand challenges hosted by two well-known medical imaging communities, IEEE Signal Processing Society (ISBI) and Medical Image Computing and Computer Assisted Intervention Society (MICCAI) start to provide public access of biomedical images that are labelled by specialists, where the most of labelled datasets are 2D images. We would like to keep on putting effort into data preparation for volumetric medical images, and sharing our resources with the other researchers. There are still over 100 unlabelled CTA volumes that can be used for both aorta and coronary artery segmentation studies. We are planing to work closely with relevant clinical experts and radiologists to provide public access to the labelled dataset in the future.

- **Improvement to Proposed Method:** There are several potential improvement that can be made to the proposed methods in this thesis.
  - Efficient RF Training and Testing: In Chapter 3, the adaptive learner for voxel-wise foreground-background classification is built on a completely re-trained RF in each round of interaction. Although the speed difference between off-line RF and on-line RF is negligible due to small amount of training data, it is worth investigating on-line models that are more suitable for larger and higher dimensional dataset. The testing speed is the efficiency bottle neck of proposed method. In addition to on-line training strategy, feature ranking, importance selection and tree pruning could be potential ways to increase the overall speed efficiency, as the model complexity can be further reduced by removing the weak feature and unnecessary branches.
  - End-to-End Cascade Training: Cascade classifier was used in Chapter 4 and Chapter 5, which subdivide a big and difficult problem into a set of smaller and simpler sub-problems. Then, the solution to the original problem can be found via combining individual solvers of sub-problems consecutively to form a cascade classifier. However, training a cascade classifier is a non-trivial task, where the training data for current stage is the residues retained from previous stage. Hence, any modification to previous stage will lead to the rest of stages needing to be re-trained. An end-to-end learning method is required to avoid unnecessary intermediate manual training adjustment. There are a couple of strategies that we are going to investigate the feasible of end-to-end training method for cascade classifier, such as instance training weighting and multiple-pass training data generation.
  - Shape Prior Regularisation: Shape prior as an image segmentation regularisation has been widely studied, and proved to be efficient for anatomy specified segmentation. In Chapter 4, the proposed NU-IBS is feasible to cooperate with shape prior regularisation, and formulate it into a joint energy minimisation function. Given sufficient labelled data for certain anatomy, the shape prior can be learnt in the parameter space using the proposed implicit representation, where the regularisation can be imposed directly as a likelihood of statistical distribution on NU-IBS parameters. We would expect a promising accuracy boost for an anatomy specified segmentation task.

  - Implicit Shape Manipulation: The proposed NU-IBS was used to represent the geometrical structure of target object, where the region based deformation was derived based on level set PDE and Chan-Vese model. We believe that the interaction interface between proposed segmentation method and user can be further extended to geometry manipulation, for example level set surface editing operator [185], and Laplacian surface editing method [186].

- **System:** Our plans for further development are to aim to integrate both proposed medical image segmentation methods into our platform software, *SVMIST*, that can be delivered to radiologist and clinician to carry out realistic studies. Speed efficiency is the major issue of current machine learning based method for 3D volumetric data segmentation, where the methods generally involve voxel-wise feature extraction and prediction. Compared to 2D image, the number of hypotheses of a volumetric data are huge, hence, real-time interaction on a large size 3D scan is extremely challenging. Parallel implementation with dedicated hardware architecture could be a promising solution, especially for random forests and CNN based methods that are very suitable to be parallelised on both CPU and GPU. In addition, there are many standard but effective segmentation algorithms and functionalities need to be integrated, such as geometrical measurement, rich annotation tools, and standardised labelling output format. Fig. 6.1 shows the new interface of *SVMIST* that uses a cross-platform GUI library, QT.

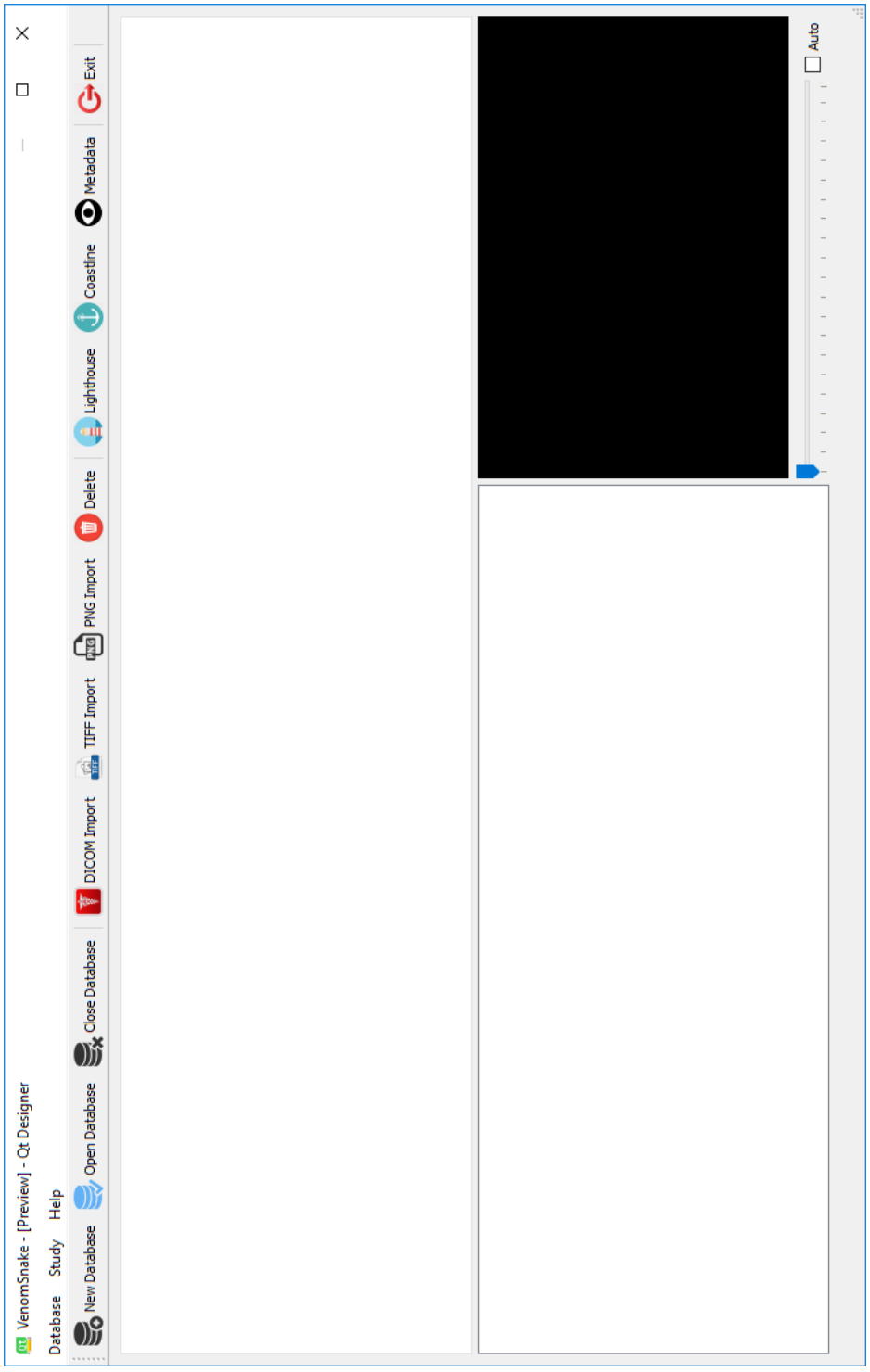


Figure 6.1: The new interface of *SVMIST* that uses a cross-platform GUI library, QT.

Bibliography

[1] A. Criminisi, J. Shotton, E. Konukoglu, *et al.*, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, pp. 81–227, 2012.

[2] X. Xie and M. Mirmehdi, "MAC: Magnetostatic active contour model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 4, pp. 632–646, 2008.

[3] A. Paiement, M. Mirmehdi, X. Xie, and M. C. Hamilton, "Registration and modeling from spaced and misaligned image volumes," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4379–4393, 2016.

[4] H. M. Salinas and D. C. Fernández, "Comparison of PDE-based nonlinear diffusion approaches for image enhancement and denoising in optical coherence tomography," *IEEE Transactions on Medical Imaging*, vol. 26, no. 6, pp. 761–771, 2007.

[5] "Volume rendering." https://en.wikipedia.org/wiki/Volume_rendering#/media/File:CTSkullImage.png. Accessed: 2017-02-28.

[6] "Snakes: Active contour model." https://en.wikipedia.org/wiki/Active_contour_model#/media/File:Snake-contour-example.jpg. Accessed: 2017-02-28.

[7] "Geodesic active contour." https://en.wikipedia.org/wiki/Level_set_method#/media/File:Level_set_method.jpg. Accessed: 2017-02-28.

[8] "Circulatory system." https://en.wikipedia.org/wiki/Circulatory_system#/media/File:Circulatory_System_en.svg. Accessed: 2017-02-28.

[9] "The human heart viewed from the front." https://en.wikipedia.org/wiki/Heart#/media/File:Blausen_0451_Heart_Anterior.png. Accessed: 2017-02-28.

[10] "The human heart viewed from behind." https://en.wikipedia.org/wiki/Heart#/media/File:Blausen_0456_Heart_Posterior.png. Accessed: 2017-02-28.

[11] "The heart, showing valves, arteries and veins." https://en.wikipedia.org/wiki/Heart#/media/File:Diagram_of_the_human_heart_(cropped).svg. Accessed: 2017-02-28.

[12] "The heart with the atria and major vessels removed." https://en.wikipedia.org/wiki/Heart#/media/File:2011_Heart_Valves.jpg. Accessed: 2017-02-28.

[13] "Illustration of the aorta root." http://my.clevelandclinic.org/health/articles/heart-blood-vessels-aorta/aortic-valve-root. Accessed: 2017-02-28.

[14] "Illustration of the coronary arteries." https://en.wikipedia.org/wiki/Coronary_circulation#/media/File:Blausen_0256_CoronaryArteries_02.png. Accessed: 2017-02-28.

[15] "The progression of atherosclerosis." https://en.wikipedia.org/wiki/Atherosclerosis#/media/File:Endo_dysfunction_Athero.PNG. Accessed: 2017-03-01.

[16] "Coronary artery disease." https://en.wikipedia.org/wiki/Coronary_artery_disease#/media/File:Blausen_0259_CoronaryArteryDisease_02.png. Accessed: 2017-03-01.

[17] "Transfemoral and transapical approaches for transcateher aortic valve replacement." http://www.heart-valve-surgery.com/heart-surgery-blog/2013/09/18/tavr-transfemoral-transapical-approaches/. Accessed: 2017-03-01.

[18] "Coronary artery disease treatment." http://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/diagnosis-treatment/treatment/txc-20165340. Accessed: 2017-03-01.

[19] "Brain PET." https://www.reference.com/health/pet-scan-cancer-28bd930602dd3766. Accessed: 2017-03-05.

[20] "Fetal ultrasound." https://en.wikipedia.org/wiki/Ultrasound#/mediaFile:CRL_Crown_rump_lengh_12_weeks_ecografia_Dr._Wolfgang_Moroder.jpg. Accessed: 2017-03-05.

[21] "Modern CT scanner." https://en.wikipedia.org/wiki/CT_scan#/media/File:UPMCEast_CTscan.jpg. Accessed: 2017-03-04.

[22] "X-ray computed tomography." http://quakerecnankai.blogspot.co.uk/2014/12/. Accessed: 2017-03-04.

[23] "CTA coronary artery." http://www.dei.org.mx/wp-content/uploads/2012/02/mid.jpg. Accessed: 2017-03-05.

[24] "Sagittal multiplanar reformation (SPR) of an abdominal aortic aneurysm (AAA) (arrows)." https://en.wikipedia.org/wiki/Computed_tomography_angiography#/media/File:SagitalAAA.jpg. Accessed: 2017-03-05.

[25] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, no. 34–47, 2001.

[26] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[29] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[30] 3Dim Laboratory s.r.o., "3DimViewer." http://www.3dim-laboratory.cz/software/3dimviewer. Accessed: 2017-03-01.

[31] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, pp. 23–37, Springer Berlin Heidelberg, 1995.

[34] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

[35] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[36] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[39] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, pp. 98–136, Jan 2015.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[41] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "CUDNN: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[42] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[43] J. Markoff, "Google cars drive themselves, in traffic," *The New York Times*, vol. 10, no. A1, p. 9, 2010.

[44] M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, and S. C. Strother, "Machine learning in medical imaging," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 25–38, 2010.

[45] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[46] A. Blake, P. Kohli, and C. Rother, *Markov random fields for vision and image processing*. Mit Press, 2011.

[47] N. Komodakis, *Optimization Algorithms for Discrete Markov Random Fields, with Applications to Computer Vision*. PhD thesis, University of Crete, 2006.

[48] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, pp. 1–58, 2006.

[49] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[50] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[51] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[52] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural network design*. Martin Hagan, 2014.

[53] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International conference on machine learning*, pp. 807–814, 2010.

[54] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[55] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[57] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 60–65, IEEE, 2005.

[58] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[59] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[60] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[61] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[62] L. G. Brown, "A survey of image registration techniques," *ACM computing surveys (CSUR)*, vol. 24, no. 4, pp. 325–376, 1992.

[63] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[64] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[65] M. Mirmehdi, *Handbook of texture analysis*. Imperial College Press, 2008.

[66] M. Lai, "Deep learning for medical image segmentation," *arXiv preprint arXiv:1505.02000*, 2015.

[67] F. Zhao and X. Xie, "An overview of interactive medical image segmentation," *Annals of the BMVA*, vol. 2013, no. 7, pp. 1–22, 2013.

[68] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[69] C. Xu, D. L. Pham, and J. L. Prince, "Image segmentation using deformable models," *Handbook of medical imaging*, vol. 2, pp. 129–174, 2000.

[70] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: a survey," *Medical image analysis*, vol. 1, no. 2, pp. 91–108, 1996.

[71] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on image processing*, vol. 7, no. 3, pp. 359–369, 1998.

[72] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International journal of computer vision*, vol. 22, no. 1, pp. 61–79, 1997.

[73] S. Z. Li, *Markov random field modeling in computer vision*. Springer Science & Business Media, 2012.

[74] Y. Boykov and V. Kolmogorov, "Computing geodesics and minimal surfaces via graph cuts.," in *ICCV*, vol. 3, pp. 26–33, 2003.

[75] B. Appleton and H. Talbot, "Globally minimal surfaces by continuous maximal flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 106–118, 2006.

[76] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[77] S. Mendis, P. Puska, B. Norrving, *et al.*, *Global atlas on cardiovascular disease prevention and control.* World Health Organization, 2011.

[78] C. J. Murray, K. F. Ortblad, C. Guinovart, S. S. Lim, T. M. Wolock, D. A. Roberts, E. A. Dansereau, N. Graetz, R. M. Barber, J. C. Brown, *et al.*, "Global, regional, and national incidence and mortality for hiv, tuberculosis, and malaria during 1990–2013: a

systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 384, no. 9947, pp. 1005–1070, 2014.

[79] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Computer vision, 1998. sixth international conference on*, pp. 555–562, IEEE, 1998.

[80] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 1491–1498, IEEE, 2006.

[81] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[82] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[84] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *ACM on International Conference on Multimedia Retrieval*, pp. 643–650, ACM, 2015.

[85] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, pp. 3431–3440, 2015.

[86] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *ACM Multimedia*, pp. 516–520, ACM, 2016.

[87] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[88] Z. Yang and R. Nevatia, "A multi-scale cascade fully convolutional network face detector," *arXiv preprint arXiv:1609.03536*, 2016.

[89] Y. Bai, W. Ma, Y. Li, L. Cao, W. Guo, and L. Yang, "Multi-scale fully convolutional network for fast face detection," in *BMVC*, September 2016.

[90] Y. Li, B. Sun, T. Wu, Y. Wang, and W. Gao, "Face detection with end-to-end integration of a ConvNet and a 3D model," *arXiv preprint arXiv:1606.00850*, 2016.

[91] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *arXiv preprint arXiv:1604.02878*, 2016.

[92] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874v3*, 2015.

[93] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[94] C. Kirbas and F. Quek, "A review of vessel extraction techniques and algorithms," *ACM Computing Survey*, vol. 36, no. 2, pp. 81–121, 2004.

[95] H. Li and A. Yezzi, "Vessels as 4-D curves: Global minimal 4-D paths to extract 3-D tubular surfaces and centerlines," *IEEE Transactions on Medical Imaging*, vol. 26, pp. 1213–1223, 2007.

[96] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Medical Image Analysis*, vol. 13, no. 6, pp. 819–845, 2009.

[97] S. Esneault, C. Lafon, and J.-L. Dillenseger, "Liver vessels segmentation using a hybrid geometrical moments/graph cuts method," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 2, pp. 276–283, 2010.

[98] C. Bauer, T. Pock, E. Sorantin, H. Bischof, and R. Beichel, "Segmentation of interwoven 3D tubular tree structures utilizing shape priors and graph cuts," *Medical Image Analysis*, vol. 14, no. 2, pp. 172–184, 2010.

[99] N. Zhu and A. C. Chung, "Graph-based optimization with tubularity markov tree for 3D vessel segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2219–2226, June 2013.

[100] X. Xie and M. Mirmehdi, "Magnetostatic field for the active contour model: A study in convergence.," in *Proceedings of the British Machine Vision Conference*, pp. 127–136, 2006.

[101] X. Xie and M. Mirmehdi, "Implicit active model using radial basis function interpolated level sets.," in *Proceedings of the British Machine Vision Conference*, pp. 1–10, 2007.

[102] X. Xie and M. Mirmehdi, "Radial basis function based level set interpolation and evolution for deformable modelling," *Image and Vision Computing*, vol. 29, no. 2, pp. 167–177, 2011.

[103] R. Gonzalez and P. Wintz, "Digital image processing," 1977.

[104] G. Sivewright and P. Elliott, "Interactive region and volume growing for segmenting volumes in MR and CT images," *Medical informatics*, vol. 19, no. 1, pp. 71–80, 1994.

[105] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Proceedings of Medical Image Computing and Computer-Assisted Interventation*, vol. 1496, pp. 130–137, 1998.

[106] R. Manniesing, M. A. Viergever, and W. J. Niessen, "Vessel enhancing diffusion: A scale space representation of vessel structures," *Medical Image Analysis*, vol. 10, no. 6, pp. 815–825, 2006.

[107] T. K. Ho, "Random decision forests," in *IEEE International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[108] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[109] J. Santner, M. Unger, T. Pock, C. Leistner, A. Saffari, and H. Bischof, "Interactive texture segmentation using random forests and total variation," in *Proceedings of the British Machine Vision Conference*, pp. 66.1–66.12, 2009.

[110] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, 2013.

[111] S. Candemir and Y. S. Akgul, *Statistical Significance Based Graph Cut Segmentation for Shrinking Bias*, pp. 304–313.

[112] P. Kohli, J. Rihan, M. Bray, and P. H. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 285–298, 2008.

[113] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, Aug. 2004.

[114] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, 2003.

[115] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies," *Computer Vision and Image Understanding*, vol. 112, no. 1, pp. 14–29, 2008.

[116] C. Chekuri, S. Khanna, J. S. Naor, and L. Zosin, "Approximation algorithms for the metric labeling problem via a new linear programming formulation," in *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pp. 109–118, 2001.

[117] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1436–1453, 2007.

[118] N. Komodakis, N. Paragios, and G. Tziritas, "Mrf energy minimization and beyond via dual decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 531–552, 2011.

[119] A. Blum, "On-line algorithms in machine learning," in *Online algorithms*, pp. 306–325, Springer, 1998.

[120] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.

[121] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *European conference on computer vision*, pp. 428–441, Springer, 2004.

[122] M. Unger, T. Pock, W. Trobin, D. Cremers, and H. Bischof, "TVSeg: Interactive total variation based image segmentation.," in *BMVC*, vol. 31, pp. 44–46, Citeseer, 2008.

[123] J. Santner, M. Unger, T. Pock, C. Leistner, A. Saffari, and H. Bischof, "Interactive texture segmentation using random forests and total variation.," in *BMVC*, pp. 1–12, 2009.

[124] M. Jian and C. Jung, "Interactive image segmentation using adaptive constraint propagation," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1301–1311, 2016.

[125] J.-L. Jones, X. Xie, and E. Essa, "Combining region-based and imprecise boundary-based cues for interactive medical image segmentation," *International journal for numerical methods in biomedical engineering*, vol. 30, no. 12, pp. 1649–1666, 2014.

[126] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *IEEE International Conference on Computer Vision*, vol. 1, pp. 105–112, IEEE, 2001.

[127] S. Han, W. Tao, D. Wang, X.-C. Tai, and X. Wu, "Image segmentation based on GrabCut framework integrating multiscale nonlinear structure tensor," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2289–2302, 2009.

[128] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," in *ACM Transactions on Graphics*, vol. 23, pp. 303–308, ACM, 2004.

[129] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3161–3168, IEEE, 2010.

[130] J. Feng, B. Price, S. Cohen, and S.-F. Chang, "Interactive segmentation on RGBD images via cue selection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–164, 2016.

[131] T. Sahin and M. Unel, "Fitting globally stabilized algebraic surfaces to range data," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1083–1088, IEEE, 2005.

[132] O. Bernard, D. Friboulet, P. Thévenaz, and M. Unser, "Variational B-spline level-set: a linear filtering approach for fast deformable model evolution," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1179–1191, 2009.

[133] M. Rouhani and A. D. Sappa, "The richer representation the better registration," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5036–5049, 2013.

[134] M. Rouhani, A. D. Sappa, and E. Boyer, "Implicit B-spline surface reconstruction," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 22–32, 2015.

[135] B. S. Morse, W. Liu, T. S. Yoo, and K. Subramanian, "Active contours using a constraint-based implicit representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 285–292, IEEE, 2005.

[136] A. Paiement, M. Mirmehdi, X. Xie, and M. C. Hamilton, "Integrated segmentation and interpolation of sparse data," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 110–125, 2014.

[137] X. Xie and M. Mirmehdi, "Radial basis function based level set interpolation and evolution for deformable modelling," *Image and Vision Computing*, vol. 29, no. 2, pp. 167–177, 2011.

[138] A. Gelas, O. Bernard, D. Friboulet, and R. Prost, "Compactly supported radial basis functions based collocation method for level-set evolution in image segmentation," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1873–1887, 2007.

[139] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: a level set approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 2, pp. 158–175, 1995.

[140] H. Wendland, "Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree," *Advances in computational Mathematics*, vol. 4, no. 1, pp. 389–396, 1995.

[141] M. Botsch, D. Bommes, and L. Kobbelt, "Efficient linear system solvers for mesh processing," in *Mathematics of Surfaces XI*, pp. 62–83, Springer, 2005.

[142] K. Kamnitsas, C. Ledig, V. F. 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*, vol. 36, pp. 61–78, 2017.

[143] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks.," in *International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011.

[144] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[145] T. A. Davis, *Direct methods for sparse linear systems*. Society for Industrial and Applied Mathematics, 2006.

[146] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[147] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, vol. 3. Cambridge university press, 1999.

[148] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher, "A variational level set approach to multiphase motion," *Journal of computational physics*, vol. 127, no. 1, pp. 179–195, 1996.

[149] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, p. 29, August 2015.

[150] Intel Corporation, "Intel Threading Building Blocks." https://www.threadingbuildingblocks.org/. Accessed: 2017-03-01.

[151] T. Davis, "SUITESPARSE." http://faculty.cse.tamu.edu/davis/suitesparse. Accessed: 2017-03-01.

[152] M.-T. Pham, Y. Gao, V.-D. D. Hoang, and T.-J. Cham, "Fast polygonal integration and its application in extending Haar-like features to improve object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 942–949, 2010.

[153] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *IEEE International Joint Conference on Biometrics*, pp. 1–8, 2014.

[154] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *IEEE International Conference on Computer Vision*, pp. 3676–3684, 2015.

[155] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325–5334, 2015.

[156] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3460–3467, 2013.

[157] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1843–1850, 2014.

[158] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[159] H. Fang, J. Deng, X. Xie, and P. W. Grant, "From clamped local shape models to global shape model," in *IEEE Conference on Image Processing*, pp. 3513–3517, Sep 2013.

[160] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 236–243, June 2005.

[161] M. Koestinger, *Efficient Metric Learning for Real-World Face Recognition*. PhD thesis, Graz University of Technology, Faculty of Computer Science, 2013.

[162] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[163] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[164] I. Endres and D. Hoiem, "Category independent object proposals," in *European Conference on Computer Vision*, pp. 575–588, Springer, 2010.

[165] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 328–335, 2014.

[166] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1312–1328, 2012.

[167] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[168] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[169] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1408–1423, 2004.

[170] M. Edwards and X. Xie, "Graph based convolutional neural network," *arXiv preprint arXiv:1609.08965*, 2016.

[171] J. Deng, X. Xie, and M. Edwards, "Combining stacked denoising autoencoders and random forests for face detection," in *Advanced Concepts for Intelligent Vision Systems*, pp. 349–360, 2016.

[172] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," *arXiv preprint arXiv:1511.06523*, 2015.

[173] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *EEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[174] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Tech. Rep. UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[175] J. Yan, Z. Lei, L. Wen, and S. Li, "The fastest deformable part model for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2497–2504, 2014.

[176] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*, pp. 720–735, Springer, 2014.

[177] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 211–223, Feb 2016.

[178] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2879–2886, 2012.

[179] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[180] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 38–44, 1998.

[181] S. D. MPLab, University of California, "The MPLab GENKI Database, GENKI-SZSL Subset," 2009. Accessed: 2016-05-12.

[182] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*, pp. 828–837, Springer, 2007.

[183] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[184] A. Vedaldi and K. Lenc, "MatConvNet – convolutional neural networks for MATLAB," in *ACM Multimedia*, 2015.

[185] K. Museth, D. E. Breen, R. T. Whitaker, and A. H. Barr, "Level set surface editing operators," in *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 330–338, ACM, 2002.

[186] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, "Laplacian surface editing," in *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pp. 175–184, ACM, 2004.