University of Exeter
Department of Physics

Automatic Segmentation of the Lumbar Spine from Medical Images

Hugo Winfield Hutt

February, 2016

Submitted by Hugo Winfield Hutt, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Physics , February, 2016.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Declaration

Chapters 3 and 4 are based on work first published in the following papers:

1. H. Hutt, R. Everson, and J. Meakin. 3D Intervertebral Disc Segmentation from MRI using Supervoxel-Based CRFs. In *Proc. 3rd MICCAI Wksp. Comput. Meth. Clin. App. Spine Imag. (MICCAI-CSI 2015)*, pages 119–123, 2015b

2. H. Hutt, R. Everson, and J. Meakin. Segmentation of Lumbar Vertebrae Slices from CT Images. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 61–71. Springer, 2015a

3. G. Zheng, C. Chengwen, B. Ibragimov, R. Korez, T. Vrtovec, H. Hutt, R. Everson, J. Meakin, I. Lopez Andrade, B. Glocker, H. Chen, Q. Dou, P.-A. Heng, C. Wang, D. Forsberg, A. Neubert, J. Fripp, M. Urschler, D. Stern, M. Wimmer, A. A. Novikov, D. L. Belav, H. Cheng, G. Armbrecht, D. Felsenberg, and S. Li. Evaluation and Comparison of 3D Intervertebral Disc Localization and Segmentation Methods for 3D T2 MRI Data: A Grand Challenge. *Med. Imag. Anal.*, 2016. forthcoming

Hugo Winfield Hutt

Abstract

Segmentation of the lumbar spine in 3D is a necessary step in numerous medical applications, but remains a challenging problem for computational methods due to the complex and varied shape of the anatomy and the noise and other artefacts often present in the images. While manual annotation of anatomical objects such as vertebrae is often carried out with the aid of specialised software, obtaining even a single example can be extremely time-consuming. Automating the segmentation process is the only feasible way to obtain accurate and reliable segmentations on any large scale.

This thesis describes an approach for automatic segmentation of the lumbar spine from medical images; specifically those acquired using magnetic resonance imaging (MRI) and computed tomography (CT). The segmentation problem is formulated as one of assigning class labels to local clustered regions of an image (called *superpixels* in 2D or *supervoxels* in 3D). Features are introduced in 2D and 3D which can be used to train a classifier for estimating the class labels of the superpixels or supervoxels. Spatial context is introduced by incorporating the class estimates into a conditional random field along with a learned pairwise metric. Inference over the resulting model can be carried out very efficiently, enabling an accurate pixel- or voxel-level segmentation to be recovered from the labelled regions.

In contrast to most previous work in the literature, the approach does not rely on explicit prior shape information. It therefore avoids many of the problems associated with these methods, such as the need to construct a representative prior model of anatomical shape from training data and the approximate nature of the optimisation. The general-purpose nature of the proposed method means that it can be used to accurately segment both vertebrae and intervertebral discs from medical images without fundamental change to the model.

Evaluation of the approach shows it to obtain accurate and robust performance in the presence of significant anatomical variation. The median average symmetric surface distances for 2D vertebra segmentation were 0.27 mm on MRI data and 0.02 mm on CT data. For 3D vertebra segmentation the median surface distances were 0.90 mm on MRI data and 0.20 mm on CT data. For 3D intervertebral disc segmentation a median surface distance of 0.54 mm was obtained on MRI data.

Acknowledgements

I would first like to thank my supervisors Judith Meakin and Richard Everson for their support and many helpful discussions throughout the development of this work.

I am grateful to the numerous volunteers from the University of Exeter for providing MRI scans used throughout this thesis and also to Dr Jonathan Fulford, Dr Alan Barker, Mr Agris Liepa and Ms Vladimira Juroskova for allowing the use of data from their studies. I would also like to thank Dr Jonathan Fulford for helping to acquire the MRI scans. In addition, many of the results presented in this thesis would not have been possible without the existence of publicly available datasets. I therefore thank the organisers of the 2014 and 2015 MICCAI workshops on Computational Methods and Clinical Applications in Spine Imaging (CSI) for providing annotated MRI and CT datasets.

Finally, I would like to thank my parents for their great support and encouragement over the years.

Contents

1 Introduction **13**

1.1 Approach 14

1.2 Contributions and Overview 15

2 Background and Related Work **17**

2.1 Introduction 17

2.1.1 The Lumbar Spine 17

2.1.2 Segmentation of the Lumbar Spine 19

2.1.3 Applications 20

2.2 Imaging Modalities 22

2.2.1 Magnetic Resonance Imaging (MRI) 22

2.2.2 Computed Tomography (CT) 23

2.3 Review of Spine Segmentation Methods 24

2.3.1 Statistical Shape Models 25

2.3.2 Deformable Surfaces and Level Sets 28

2.3.3 Registration and Atlas-Based Methods 32

2.3.4 Variational Methods 34

2.3.5 Markov Random Fields 35

2.3.6 Conclusion 38

2.4 Evaluating Segmentation Quality 39

2.4.1 Evaluation Measures 40

2.5 Annotated Datasets 41

2.5.1 Comparison of MRI Protocols 41

2.5.2 Description of Datasets 43

2.5.3 Inter-Annotator Agreement 45

2.6 Conclusion 45

3 2D Segmentation of Lumbar Vertebrae **47**

3.1 Introduction 47

3.2 Segmentation Model 49

3.2.1 Markov Random Fields 49

3.2.2 Conditional Random Fields 51

3.3 Superpixels . . . 52
3.3.1 Quick Shift . . . 52
3.3.2 SLIC . . . 53
3.4 Characterising Superpixels . . . 55
3.4.1 Intensity and Texture . . . 56
3.4.2 Location . . . 58
3.4.3 Edge Response . . . 61
3.4.4 Performance of Combined Features . . . 63
3.5 Potential Functions and Max-Marginals . . . 64
3.5.1 First-Order Potential . . . 64
3.5.2 Second-Order Potential . . . 65
3.5.3 Max-Marginals . . . 68
3.6 Experiments . . . 69
3.6.1 Image Datasets . . . 69
3.6.2 Model Training . . . 70
3.6.3 Segmentation Results . . . 70
3.6.4 3D Reconstruction . . . 72
3.7 Conclusion . . . 73

4 3D Segmentation of the Lumbar Spine 74
4.1 Introduction . . . 74
4.2 Supervoxels . . . 77
4.2.1 Axis-Weighted SLIC . . . 78
4.3 Learned Supervoxel Features . . . 79
4.3.1 Multi-Scale Dictionary Learning . . . 80
4.3.2 Supervoxel Feature Encoding and Pooling . . . 83
4.4 Location Features . . . 84
4.4.1 Contour Matching . . . 85
4.4.2 Outlier Removal . . . 86
4.4.3 Feature Extraction . . . 87
4.5 Supervoxel Classification . . . 88
4.5.1 Mining for Hard Examples . . . 88
4.5.2 Evaluation . . . 89
4.5.3 Discussion . . . 91
4.6 CRF with Learned Potentials . . . 93
4.6.1 Potential Functions . . . 93
4.7 Smoothed Max-Marginals . . . 95
4.8 Experiments and Results . . . 96
4.8.1 Datasets . . . 96
4.8.2 Implementation Details . . . 96
4.8.3 Effect of Supervoxel Size . . . 96

4.8.4 Segmentation Results 98
4.8.5 Detailed Comparison 101
4.8.6 MICCAI-CSI 2015 Challenge Results 103
4.9 Influence of Sample Size 104
4.10 Conclusion 106

5 Summary and Conclusion **109**

References **113**

List of Tables

2.1 Overview of datasets used in the thesis. IPR is the in-plane resolution and CPR is the cross-plane resolution (slice thickness). The last column gives the number and dimension of the manually annotated images. 43

2.2 Inter-annotator agreement on a sample of 10 images from MRI dataset 1. The table shows the minimum, median and maximum values over all individual images. 45

3.1 Superpixel features (p_n denotes the n-th percentile). The right column gives the dimension of the superpixel feature vector. 57

3.2 Segmentation results on the MRI data. The table shows the minimum, median and maximum values over all individual volumes. . . . 70

3.3 Segmentation results on the CT data. The table gives the minimum, median and maximum values of the evaluation measures for each lumbar vertebra. 71

4.1 F-scores averaged over all leave-one-out iterations using different features for classification (see text for details). The number of pyramid levels used for dictionary learning is M. 90

4.2 Segmentation results on the MRI and CT datasets. The table shows the minimum, median and maximum values over all individual volumes. 99

4.3 Comparison of different segmentation methods on the two MRI datasets. The table gives the minimum, median and maximum values over all individual volumes. GAC is geodesic active contours, RC-SVM is region competition with SVM pre-segmentation and CRF is the proposed method. The best scores for the various criteria are highlighted in bold. 102

4.4 Results from the MICCAI-CSI 2015 localisation and segmentation challenge. A description of the different evaluation measures is given in the text. 103

4.5 Segmentation performance as a function of the training sample size. The rows give the median values of the evaluation measures for each sample size. 107

List of Figures

2.1 Anatomical outline of the human spine [Kurtz and Edidin, 2006]. The lumbar spine is shown in green with the vertebrae labelled from L1 to L5. The intervertebral discs are shown in yellow. 18

2.2 Anatomical features of an individual lumbar vertebra. 19

2.3 (Top row) Axial cross-sections of two example lumbar vertebrae acquired using T1-weighted MRI (left) and CT (right). The yellow dots are provided for reference and mark the location of the spinal canal in both images. (Bottom row) Sagittal views showing locations of intervertebral discs (indicated by yellow arrows) in a T2-weighted MR image and a CT image, respectively. 20

2.4 (a) Example graph construction for a 2-class segmentation problem on a 4×4 pixel image. In blue are the edge links between pixel nodes. In orange are the links between class nodes and pixels. (b) Segmentation resulting from cutting the graph through the dashed black line. Each of the 4 pixels are connected to a single unique class, which determines their class label assignment (i.e. $(x_1, x_2) \in l_1$ and $(x_3, x_4) \in l_2$). 37

2.5 Close-up of a right transverse process from an example T1-weighted axial MR image, showing merging between vertebra and background regions. 41

2.6 Axial slices showing vertebra on the left and intervertebral disc on the right. Images on the top row are T1-weighted. The bottom row shows the images obtained of the same subject using fat suppression. 42

3.1 The left figure shows a 2D axial MR slice with ground truth contour (yellow) for a section of the vertebra. The right figure shows boundaries for superpixels assigned to the vertebra class (cyan) and background class (magenta). The superpixels preserve the boundary detail of the vertebrae by clustering contiguous pixels with similar intensities. 48

3.2 (a) A graphical representation of a conditional random field (CRF) on a 2D grid. The class label x_i of node i is dependent on the class labels of its neighbours (shown in blue), given the observation (shown in green). (b) The factor graph representation for two nodes (superpixels) i and j, showing the relationship between the variables and the first- and second-order potentials of the energy function (3.6). . . 51
3.3 Illustration of the SIFT descriptor [Lowe, 2004]. The magnitudes and orientations of local image gradients are weighted by a Gaussian window (blue circle). The samples are aggregated into orientation histograms characterising sub-regions (shown on the right), where the length of the arrows is determined by the sum of gradient magnitudes near their directions. 58
3.4 ROC curves of SVM probability estimates for the individual and combined superpixel features. 59
3.5 (a) Example MRI vertebra slice. The cortical bone can be seen as the dark boundary surrounding the vertebra. (b) Example matched contour (magenta) found using (3.13). Also shown are the axes of the centred coordinates (cyan). (c) Gradient of the contour region distance transform in the horizontal and vertical directions, respectively. 60
3.6 Visualisation of a Laplacian of Gaussian filter. 61
3.7 (a) Shown top to bottom are MR images for which the CRF obtained the minimum, median and maximum Dice similarity score (0.81, 0.88 and 0.95), respectively. (b) Probability estimates using the combined superpixel features for the images in the first column. Darker regions indicate higher probability of belonging to the vertebra class. (c) Max-marginals computed from the CRF graph cut solutions. (d) Segmentation contours shown for both the ground truth annotations (magenta) and CRF model (cyan). 64
3.8 (a) Shown top to bottom are CT images corresponding to the minimum, median and maximum Dice similarity score (0.88, 0.97 and 0.98), respectively. (b) SVM probability estimates for the images in the left hand column. Darker regions indicate higher probability of belonging to the vertebra class. (c) Final segmentation contours from the CRF shown overlaid with the probability estimates (cyan). (d) Segmentation contours shown for both the ground truth annotations (magenta) and CRF model (cyan). 65
3.9 The left image shows the projection of a sample of 2000 superpixel features onto their first 2 principal components prior to metric learning (red are positive examples, blue are negative). The right image shows the projected features in the transformed space found by LMNN. 67

3.10 Singular values of an example transformation matrix learned by LMNN on the superpixel features. 68
3.11 The top figure shows a 3D segmentation of a lumbar vertebra (L2) constructed from segmentations of the constituent CT slices. The bottom figure shows the overlap between the CRF segmentation (cyan) and ground truth (magenta). 72

4.1 Annotated sections of the lumbar spine showing vertebrae (a) and intervertebral discs (b). Note that the two figures have different scale. 75
4.2 Overview of the proposed segmentation method. Features are extracted densely over the input volume at multiple scales and pooled within supervoxels. Estimates of the supervoxel class labels are obtained using an SVM with a generalised RBF kernel. A CRF model incorporating the SVM predictions and a learned pairwise metric is used for the final labelling of the supervoxels. 76
4.3 Figures show supervoxels belonging to an example vertebra, where the class of each supervoxel is determined by majority vote. (a) Boundaries of supervoxels in an axial slice of a CT volume. Note that the image shows a single 2D slice through the 3D supervoxels. (b, c) Surfaces of supervoxels where the vertebra supervoxels have been assigned random colours and background supervoxels are shown in grey. The supervoxels preserve the boundary detail of vertebrae, enabling an accurate voxel-level segmentation to be recovered. 78
4.4 Representation of filters at two successive levels of a pyramid. A learned $5 \times 5 \times 5$ filter $\mathbf{D}_{:j}^{(l)}$ at level l of the pyramid (cyan) captures a greater spatial area than a filter $\mathbf{D}_{:j}^{(l-1)}$ with equal dimensions at level $l-1$ (magenta). 81
4.5 Examples of $5 \times 5 \times 5$ features learned using sparse coding on 3D MRI data. The features shown correspond to a dictionary $\mathbf{D}^{(1)}$ learned over the first pyramid level. 82
4.6 Example supervoxel feature vector. The encoded responses at each of the three pyramid levels are shown in different colours. 84
4.7 (Left) Matched contour (magenta) for an example axial vertebra slice. (Right) The contour distance transform (darker regions are further from the contour). 86
4.8 (Left) Precision-recall curves for the IVD dataset using k-means (K) and sparse coding (S). (Right) Precision-recall curves for the LV dataset using location features (L), k-means (K), sparse coding (S) and sparse coding combined with location features (SL). 89

4.9 Rendered 3D views of example volumes from the LV and IVD datasets. The rows show the manual annotations (a), SVM probability estimates (b), smoothed CRF max-marginals (c) and the final thresholded segmentations (d). Darker regions in rows (b) and (c) indicate higher probability of belonging to the object class. 92
4.10 The left image shows the projection of a sample of 2000 supervoxel features onto their first 2 principal components prior to metric learning (red are positive examples, blue are negative). The right image shows the projected features in the transformed space found by LMNN. 94
4.11 Precision-recall curves for the voxel-level smoothed CRF max-marginals (red) and SVM predictions (blue). Results are shown for the IVD dataset (left) and LV dataset (right). 95
4.12 (Left) Effect of SLIC size parameter S on the voxel-level Dice score. The values are given for both the IVD dataset (blue) and LV dataset (red). (Right) Boundary recall as a function of the S parameter. . . . 97
4.13 (a, c) Example segmentation results on the IVD dataset overlaid onto a mid-sagittal slice from two subjects. (b, d) Overlap between the CRF segmentations (cyan) and manual annotations (magenta). 98
4.14 (a, c) Example segmentation results on the LV dataset showing the volume overlap between CRF segmentations (cyan) and manual annotations (magenta) for two subjects. (b, d) Segmentation boundaries in the axial slices marked by the grey plane. 100
4.15 (a, c) Example segmentation results on the CT dataset showing the volume overlap between CRF segmentations (cyan) and manual annotations (magenta). (b, d) Segmentation boundaries in the axial slices marked by the grey plane. 101
4.16 Learning curves for the SVM predictions (left) and the CRF max-marginals (right) showing the F-score as a function of the training sample size. 105
4.17 Box plots of the Dice scores and surface distance measures as a function of the training sample size. 106

1 Introduction

This thesis describes an approach for automatic segmentation of the lumbar spine from medical images. Automatic segmentation is one of the fundamental tasks in medical image analysis and forms a necessary step in numerous medical applications. In general terms the segmentation problem can be formulated as one of automatically assigning labels to the pixels or voxels of an image, which indicate the object classes to which they belong. This results in a segmentation mask from which the anatomical object(s) of interest can be extracted for further processing and analysis. Although much research has been devoted to this problem, it remains very challenging for automated computational methods to obtain accurate results. In addition, the highly complex and varied anatomy of the lumbar spine make it a particularly difficult problem in medical image segmentation.

Although manual annotation of anatomical objects such as vertebrae and intervertebral discs is often carried out with the aid of specialised software, obtaining even a single example can be extremely time-consuming [Cook et al., 2012]. In addition, manual delineation of anatomical objects can be highly subjective and prone to both intra- and inter-annotator variability [Suetens, 2009]. Automating the segmentation process is the only feasible way to obtain accurate and reliable segmentations on any large scale. This remains a difficult task for computational methods due to the complexity of the anatomy (see Figures 2.1 and 2.2) and the various imperfections often present in the images, such as noise and intensity inhomogeneity.

Previous work on spine segmentation has focused predominantly on computed tomography (CT), in part due to the generally higher contrast between bone and surrounding tissue. However, magnetic resonance imaging (MRI) is the preferred imaging modality for many applications as it does not use ionising radiation. Because of this, for most non-clinical applications MRI is likely to be the only way of acquiring scans of the spine from healthy subjects. Although automatic segmentation of MR images presents additional challenges due to the lower contrast and usually more limited resolution, there is a strong motivation for a method which can operate effectively on MRI data. A reliable method for automatic segmentation may also allow the use of MRI in certain clinical applications where CT is currently used predominantly.

Recent work on spinal segmentation has tended to focus on incorporating prior shape information into the model [Kadoury et al., 2013, Mizaalian et al., 2013, Kirschner et al., 2011, Lim et al., 2013, Ma and Lu, 2013, Klinder et al., 2009]. One major drawback of shape-based approaches is their reliance on accurate prior knowledge of anatomical variation to constrain the segmentation. Constructing a statistical model of shape usually involves the precise placement of anatomical landmarks across a representative training dataset, a step which is often carried out manually by an expert. In addition, a representative model of anatomical shape requires a large enough set of annotated training images in order to capture most of the global variation in shape. As the acquisition and manual annotation of medical images is a very laborious and potentially expensive process, this requirement is often prohibitive. Existing methods also tend to rely on accurate initialisation of the algorithm to ensure that the optimisation process does not become trapped at a local minimum, which can result in a poor segmentation.

This thesis presents an alternative approach to lumbar spine segmentation that makes extensive use of machine learning methods to both find effective representations of the image data that are robust to noise and intensity inhomogeneity and to use those representations for accurate prediction. An advantage of this approach to segmentation is that the resulting method is not completely task-dependent, and can therefore be applied to segment different anatomical structures without changing the underlying model. For example, the segmentation of intervertebral discs (IVDs) has typically been treated as an independent problem in the literature, employing specialised methods differing from those used for vertebra segmentation. The following chapters show that it is possible to obtain accurate solutions to both problems by employing a more general-purpose method. The direct applicability of the method to different medical imaging modalities is also demonstrated.

1.1 Approach

The primary goal of the thesis is an accurate and reliable method for segmentation of both vertebrae and intervertebral discs from 3D MR images. Although focused on MRI, the flexibility of the approach is also demonstrated by evaluating it on comparable CT data, showing that accurate results can be obtained without fundamental change to the model.

The segmentation problem is formulated as one of assigning class labels to local clustered regions of an image (called *superpixels* in 2D or *supervoxels* in 3D). Operating on superpixels or supervoxels reduces computational complexity and enables descriptive features to be extracted to characterise the different classes. Machine learning

techniques are used both to learn effective features for the superpixels/supervoxels and to train a classifier for estimating the class labels. Spatial context is introduced by incorporating the classifier estimates into a conditional random field along with a learned pairwise metric to promote spatial consistency. Inference over the resulting model can be carried out very efficiently using a class of algorithms called *graph cuts*, enabling an accurate pixel- or voxel-level segmentation to be recovered from the labelled superpixels or supervoxels.

One advantage of the proposed approach is that it does not rely on explicit prior knowledge of anatomical shape to constrain the optimisation. This minimises the amount of manual intervention required by human annotators and makes it possible to adapt the approach to other segmentation tasks without requiring any radical change to the model. The resulting method fully automates the segmentation process, with no additional manual initialisation or interaction required on the part of the user.

1.2 Contributions and Overview

The main contribution of the thesis is the introduction of a novel approach for automatic segmentation of the lumber spine from medical images. The approach is evaluated extensively on CT and MRI data consisting of scans of lumbar vertebrae and intervertebral discs and is shown to obtain state-of-the-art performance when compared to existing methods.

Chapter 2 starts by providing background material for the following chapters and reviewing a number of existing approaches to lumbar spine segmentation. The chapter also provides details on the annotated datasets and methodology used for the quantitative evaluation of segmentation results.

In Chapter 3 a method for automated segmentation of lumbar vertebrae from 2D image slices is introduced. The method is based on a conditional random field (CRF) operating on superpixels and incorporating learned models into the potential functions. A set of superpixel features is described which includes information characterising the intensity, texture, location and edge response of the superpixels. The features are used to train a classifier for estimating the class labels of the superpixels. It is shown that *distance metric learning* can be incorporated naturally into a graph-based segmentation framework by defining an appropriate second-order term for the CRF. The effectiveness of the method is demonstrated on both MRI and CT images of lumbar vertebrae, where the segmentation results are competitive with existing approaches.

In Chapter 4 the method introduced for 2D segmentation is developed and extended to the problem of 3D segmentation of the lumbar spine. More specifically, the chapter is focused on segmentation of vertebrae and intervertebral discs from 3D MR images. It is shown that unsupervised feature learning can be used as an alternative to standard hand-designed texture descriptors for characterising supervoxels in 3D. This enables features to be extracted rapidly over 3D volumes and used to train classifiers to discriminate between the different classes. To this end novel supervoxel features are introduced based on encoding the responses from dictionaries of filters learned over volume pyramids. By learning the features over multiple scales, larger-scale spatial structure can be represented while ensuring the dictionary learning procedure remains tractable. The method is evaluated on 3D MRI and CT datasets of lumbar vertebrae and intervertebral discs and in all cases is shown to obtain results competitive with existing approaches. The chapter also discusses additional experiments undertaken to evaluate the performance of the method. In particular, an investigation is carried out to determine how the size of the training dataset affects the resulting performance.

Finally, Chapter 5 concludes the thesis and summarises the results from the preceding chapters. Possible further extensions of the approach are discussed along with potential applications to other tasks in medical image analysis.

2 Background and Related Work

This chapter provides the necessary background material for the rest of the thesis. The initial sections describe the problem of lumbar spine segmentation and discuss issues related to the different modalities used to acquire the images. A number of existing approaches to the problem of spine segmentation are then reviewed. Finally, details are provided on the annotated datasets and the methodology used to evaluate the segmentation results presented in the rest of the thesis.

The chapter is organised as follows: Section 2.1 describes the anatomy of the lumbar spine and introduces the problem of automatic segmentation along with potential applications. Section 2.2 gives an overview of the two imaging modalities used in the thesis, namely magnetic resonance imaging (MRI) and computed tomography (CT), including issues associated with the imaging of bone. Section 2.3 reviews existing approaches to spine segmentation, focusing on those which are most relevant to the segmentation of vertebrae and intervertebral discs from magnetic resonance images. Section 2.4 considers the issue of how segmentation quality is to be evaluated and Section 2.5 describes the annotated datasets used throughout the following chapters of the thesis.

2.1 Introduction

2.1.1 The Lumbar Spine

The human lumbar spine consists of five separate vertebrae, named according to their location in the vertebral column [Bogduk, 2012]. As shown in Figure 2.1, they are labelled L1 to L5 and are positioned between the thoracic vertebrae above (T1–T12) and the sacrum in the lower region of the spine (S1–S5). The lumbar spine is the main weight-bearing section of the spine and consequently plays an important role in the biomechanics of the spine. The lumbar vertebrae themselves are relatively complex structures, composed of multiple anatomical parts. This is illustrated in Figure 2.2, which shows the anatomical features of an individual lumbar vertebra. The largest part of a lumbar vertebra is the vertebral body, which forms the anterior part of the vertebra. The interior of the vertebral body is composed of *trabecular*

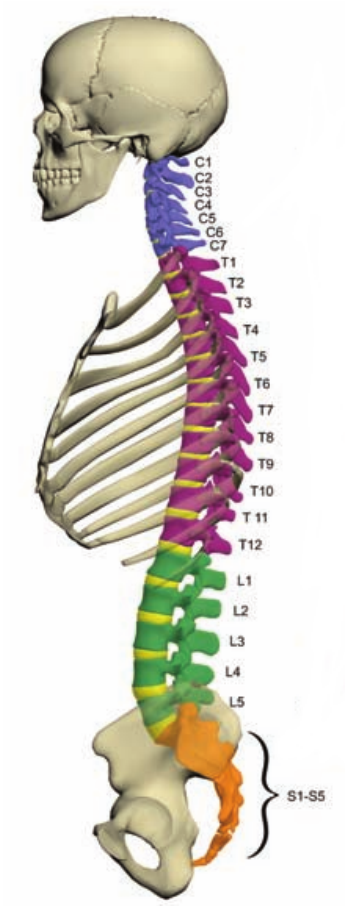


Figure 2.1: Anatomical outline of the human spine [Kurtz and Edidin, 2006]. The lumbar spine is shown in green with the vertebrae labelled from L1 to L5. The intervertebral discs are shown in yellow.

bone, which is an irregular lattice structure containing bone marrow. The interior of trabecular bone is surrounded by an outer layer of hard *cortical* bone. The vertebral body is joined to the pedicles and posterior elements of the vertebra, which include the thin transverse and spinous processes.

Located between the vertebral bodies of each pair of adjacent vertebrae are the intervertebral discs (IVDs), which also play a crucial role in the biomechanics of the lumbar spine by allowing flexibility [Bogduk, 2012]. Each individual IVD consists of a central fluid-containing mass called a *nucleus pulposus* surrounded by an *annulus fibrosus* consisting of an ordered arrangement of collagen fibres embedded in a highly-hydrated gel. Unlike the lumbar vertebrae, the individual IVDs are relatively uniform in terms of their shape and have a much simpler geometry. Note however that pathologies leading to degeneration of the IVDs can deform the shape and appearance considerably, for example due to herniation or displacement [Modic and Ross, 2007].

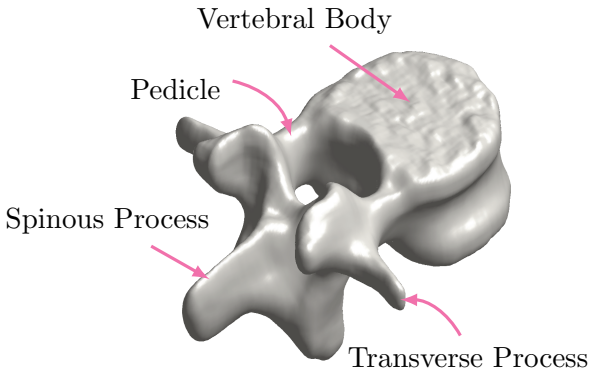


Figure 2.2: Anatomical features of an individual lumbar vertebra.

2.1.2 Segmentation of the Lumbar Spine

In order to study the properties of the lumbar spine using computational methods, a necessary step is to first segment the lumbar spine from scans of individual subjects. This involves assigning to each pixel or voxel of the image an associated *label*, which expresses the class to which it belongs (e.g. vertebra or background). Figure 2.3 shows images of lumbar vertebrae and intervertebral discs acquired using MRI and CT. Medical images are typically represented as stacks of such 2D images — called *slices* — which can later be rendered in three dimensions for visualisation. Although segmentation can be carried out manually, it is extremely time-consuming to obtain even a single example as it requires the careful delineation of the anatomy in each individual slice of the image stack. By automating the segmentation process using computational methods, this would make it possible to obtain segmentations from a much larger population of subjects.

Automating the lumbar spine segmentation process requires that the lumbar vertebrae and intervertebral discs be segmented from multi-slice MR or CT images of the spine.[1] Ideally, the segmentation framework should be as automatic as possible, minimising the amount of manual input required to obtain accurate and consistent results. This is a problem of considerable difficulty due to the varying shape and size of the lumbar spine between subjects and the overlapping intensity values of bone with other non-bone structures in MR images. Comparing the two images in the top row of Figure 2.3, it can be seen that CT provides a significantly higher contrast between bone and surrounding tissue when compared to MRI. However, due to the characteristics of CT imaging it is unsuitable for acquiring scans of the intervertebral discs of the spine, as can be seen by comparing the images in the bottom row of Figure 2.3. A more detailed description of the properties of images acquired using the two modalities of MRI and CT is provided after discussing potential applications

[1] Note that this is distinct from volumetric imaging, in which full (isotropic) 3D volume images are obtained [Hashemi et al., 2010]. For the images considered in the thesis, typically the in-plane resolution will differ from the slice thickness.

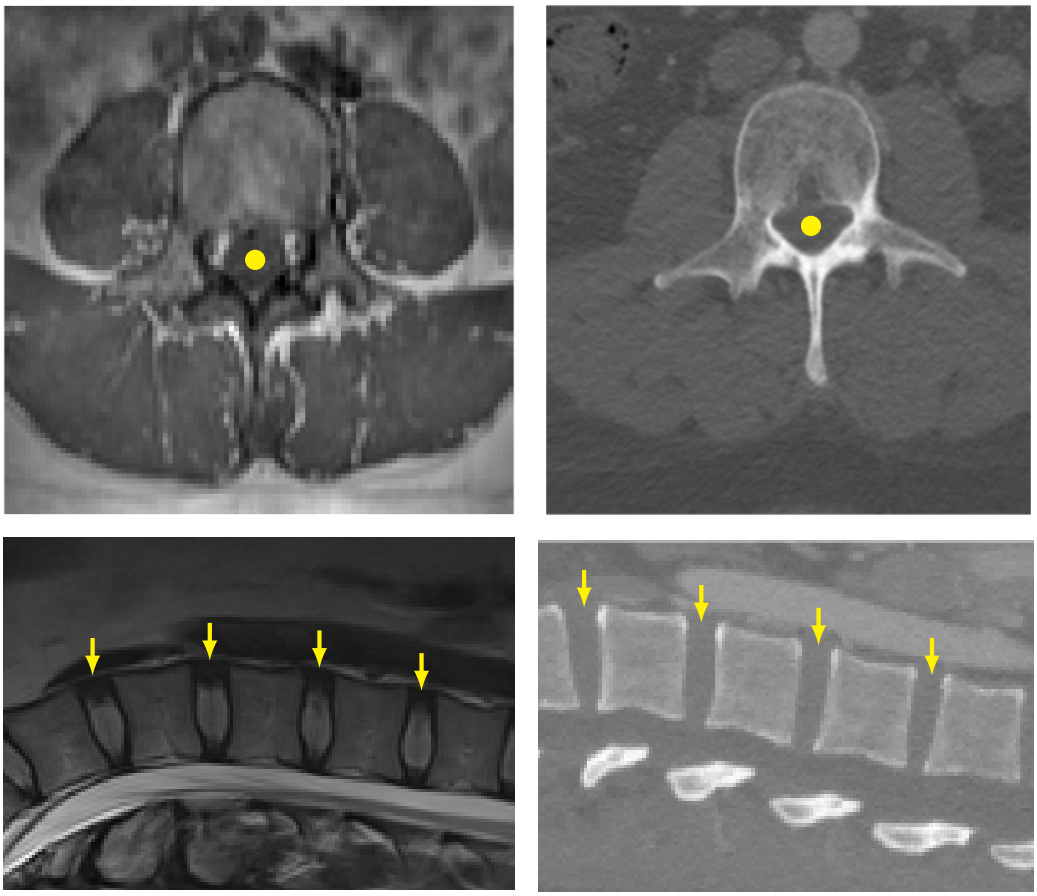


Figure 2.3: (Top row) Axial cross-sections of two example lumbar vertebrae acquired using T1-weighted MRI (left) and CT (right). The yellow dots are provided for reference and mark the location of the spinal canal in both images. (Bottom row) Sagittal views showing locations of intervertebral discs (indicated by yellow arrows) in a T2-weighted MR image and a CT image, respectively.

of lumbar spine segmentation.

2.1.3 Applications

One application area which would benefit greatly from automated segmentation is statistical modelling of the shape of the lumbar spine. Statistical models of anatomy have been applied to numerous problems in medical image analysis, such as detecting specific anatomical structures and measuring the variation in shape between subjects [Cootes and Taylor, 2004, Meakin et al., 2009]. Statistical models of shape enable an analysis of the anatomical differences between multiple subjects, as well as potentially providing insights into pathologies associated with the anatomy.

There are also clinical implications of inter-subject shape variation, which may imply that certain individuals are more prone to injury or pathology [Meakin et al., 2009].

As an example, the sagittal[2] shape of the spine is thought to be associated with low-grade spondylolisthesis, in which one or more vertebrae are displaced relative to the vertebrae below [Roussouly et al., 2006]. Understanding the shape variation of individual lumbar vertebrae can allow for a better diagnosis of vertebral fractures and other pathologies of the spine [Whitmarsh et al., 2012]. It is hypothesised that the shape of individual vertebrae may contribute to the overall shape of the spine. In constructing a 3D statistical model of the lumbar spine, this could help to explain the source of spinal shape variation between subjects. A fully automated segmentation method would also enable further biomechanical analysis of the lumbar spine by providing a surface model of the lumbar vertebrae and intervertebral discs. A similar approach has been taken in previous studies of bone structures, such as in Bryan et al. [2009] where a 3D statistical model of the whole femur bone was used to study fracture risk. This was carried out using finite element analysis, which enables numerical modelling of biomechanical properties such as stress [Kurtz and Edidin, 2006]. It has been recently noted in the literature that automatic segmentation is one of the major bottlenecks in constructing subject-specific finite element models of the spine [Jones and Wilcox, 2008].

In addition to providing a means for computational analysis of the lumbar spine, there are also direct clinical applications for automated lumbar spine segmentation. Computational tools for computer-aided diagnosis and surgical planning are increasingly being used to facilitate the routine work of clinicians [Suetens, 2009]. An important application of automated segmentation in this area is in aiding the diagnosis of intervertebral disc degeneration and subsequent treatment planning [Modic and Ross, 2007, Violas et al., 2007]. Automated computational methods have the potential to significantly reduce the time required for diagnosis, while at the same time reducing errors due to operator variability. One of the contributions of the thesis is the introduction of a fully automated method for 3D intervertebral disc segmentation from MR images, presented in Chapter 4.

Another potential clinical application of lumbar spine segmentation is in the automated detection of vertebral fractures, which could be achieved by using the information given by the automatic segmentation to infer the particular fracture status of the vertebrae. At present, the assessment of patients with traumatic spinal injuries is one of the most challenging tasks for a radiologist [Yao et al., 2012]. Due to the difficulty in visually assessing vertebral fractures and the variation between operators, they are often not recognised clinically [Hospers et al., 2009]. Computational methods could potentially help to reduce the time required to determine subsequent treatment planning; this would consequently reduce further patient suffering due to delays in detection and diagnosis.

[2]Sagittal refers to a side view of the spine, such as shown in Figure 2.1 [Hashemi et al., 2010].

2.2 Imaging Modalities

In this section the two different medical imaging modalities used in the thesis are described. The two modalities have specific advantages and limitations which lend themselves to certain applications and make them unsuitable for others.

2.2.1 Magnetic Resonance Imaging (MRI)

This thesis is primarily concerned with the segmentation of images obtained through magnetic resonance imaging (MRI). In contrast to alternative medical imaging techniques, such as computed tomography (CT), MRI does not involve potentially harmful exposure to ionising radiation and can therefore be used for obtaining images of healthy subjects. MRI also has technical advantages for certain applications, such as greater soft tissue detail and the ability to image in any plane [Withey and Koles, 2007]. The following is an outline of the basic principles behind MRI.

MRI focuses predominantly on the imaging of tissues containing high quantities of hydrogen nuclei, such as brain, organ and muscles [Hashemi et al., 2010]. When exposed to an external magnetic field, the spins of the protons in the nuclei align with the field. A radio frequency (RF) wave with a specific frequency (called the *Larmor frequency*) excites the protons. In returning to their original alignment after the RF pulse is turned off, they generate an MR signal. Spatial encoding is used to differentiate between distinct areas of the subject and to construct the image [Hashemi et al., 2010].

The protons return to their original alignment by relaxing back into their lowest energy state, referred to as the *equilibrium state*. This return to the equilibrium state is called *relaxation*, of which there are two different kinds: T1 (*Spin-lattice*) relaxation and T2 (*Spin-spin*) relaxation. The relaxation times of both T1 and T2 depend significantly on the tissue type, although T1 is always larger than T2. The process of T1- and T2-weighting allows the image contrast to be enhanced for a particular tissue type by exploiting the different relaxation times [Hashemi et al., 2010]. For example, T2-weighting is often used for acquiring images of the intervertebral discs as the presence of fluid inside of the discs normally causes them to show up as dark regions in T1-weighted images.

There are certain limitations associated with MRI, which have implications for image segmentation. These include the presence of intensity inhomogeneities in the resulting image, sometimes referred to as the *bias field* [Guillemaud and Brady, 1997]. The images can contain high levels of (random) noise caused by external RF interference or internal defects and imperfections. Motion artefacts can also be present in the

image, either due to small movements of the subject, or periodic changes (such as occur inside blood vessels). Any segmentation method dealing with MRI must take into account these issues, in the simplest case by applying pre- or post-processing to the image. However in certain cases artefacts may also exist, such as geometric distortion caused by gradient power drop-off, which are much more difficult to counter directly through image processing techniques [Hashemi et al., 2010].

MRI of Bone

As MRI is particularly effective at imaging soft tissue, most of the segmentation methods existing in the medical image analysis literature are applied to problems in this domain. An important example in this area is the separation of brain MR images into white matter, grey matter and cerebrospinal fluid [Zhang et al., 2001]. There are significantly fewer examples dealing with the segmentation of structures composed of bone, which can introduce additional challenges. Discussed next are some of these challenges and how they relate to lumbar spine segmentation.

As bone by itself appears dark in MR images, what is actually seen in the image is a mixture of bone, fat and water. Given that there are often other structures in the image containing fat and water, this can lead to overlapping intensity ranges. As a consequence, simpler approaches to segmentation, such as those based on intensity thresholding or basic edge detection, are not applicable as they result in too many errors. In addition, anatomical objects such as vertebrae have complex geometries which vary significantly between subjects. In particular, the long and thin posterior elements of the vertebrae can be extremely difficult to distinguish due to partial volume effects and intensity overlap with surrounding tissue. This difficulty is compounded by the often limited resolution of images acquired using most current MRI scanners.

2.2.2 Computed Tomography (CT)

Computed tomography (CT) is one of the predominant imaging modalities for obtaining scans of the spine. In addition, most existing approaches to spine segmentation in the literature are applied to CT in part due to the greater contrast between bone and surrounding tissue when compared to MRI. Although the physical and mathematical foundations of CT are beyond the scope of this chapter, a description is provided in this section of the basic concepts involved in CT imaging and how it differs from MRI.

Images are acquired in CT by sending X-rays through the subject from different directions to obtain a set of 2D slices, which are then used to reconstruct a 3D

representation of the object of interest [Suetens, 2009]. One or more X-ray sensors are placed against the subject to detect the incoming source X-rays after passing through the subject. The intensity of a point on the resulting image is proportional to the energy of the X-rays impacting on that point after having passed through the subject. For example, structures composed of bone mineral provide good contrast due to greater attenuation of the X-rays on subsequently reaching the sensors. Crucially, the contrast of the resulting image depends on the *dose* (i.e. the amount of radiation the subject is exposed to). In general, lowering the dose used to acquire the images leads to a degradation in image quality, making it harder to discern finer anatomical details. Due to concerns about the increasing and routine usage of CT imaging in a clinical setting, the recent trend has been towards imaging with minimum possible dosages [Suetens, 2009].

Although CT provides good contrast for bone and a shorter scanning duration compared to MRI, the main disadvantage is the exposure to potentially harmful ionising radiation. This makes it unsuitable for most of the non-clinical applications described earlier in this chapter, where the subjects are assumed to be healthy. As noted previously, another limitation of CT is that it is generally not adequate for acquiring images of the intervertebral discs, whereas T2-weighted MRI is able to provide good contrast due to being more capable of imaging structures composed of soft tissue. However, there do exist a number of important clinical applications of lumber spine segmentation for which CT is a suitable modality. These include the automated analysis and diagnosis of vertebral fractures and other traumas of the spine, as well as assisting in the planning of surgical procedures involving the lumbar spine.

2.3 Review of Spine Segmentation Methods

The general problem of image segmentation is fundamentally *ill-posed* [Hadamard, 1923]. This implies that there is no unique formulation of the problem offering a stable algorithmic solution; additional regularisation assumptions (such as smoothness) are instead required in order to constrain the set of possible solutions. As a consequence the type of approach used for medical image segmentation is often dependent on the particular problem under consideration, taking into account information such as the shape and appearance of the anatomy and the properties of the imaging modality used to acquire the images.

This section describes a number of existing methods in the literature that are specific to spine segmentation. As mentioned previously, existing methods for spine segmentation tend to focus on CT and are much less often evaluated on MRI data.

The focus of this review is on methods that are likely to be applicable to both modalities, although the methods may need to be adapted to cope with the additional challenges associated with MRI. The review is also focused predominantly on 3D segmentation, both to limit the scope and also because most methods for 2D segmentation are not directly applicable to 3D data without fundamental changes being made.

The review is organised according to the general segmentation framework to which the methods belong. Although this distinction is not always easy to make, most existing methods tend to fall into one of a few general classes with a common underlying approach, or are combinations thereof. In each section a brief description of the general framework is given, followed by a number of specific applications to spine segmentation and a discussion of their particular strengths and weaknesses.

2.3.1 Statistical Shape Models

Statistical shape models are commonly used for medical image segmentation as they provide a convenient way of incorporating prior knowledge of anatomy to guide the segmentation algorithm [Heimann and Meinzer, 2009, Castro-Mateos et al., 2014]. By enabling a global, top-down approach to segmentation, shape models can be more robust to local image variation than methods based purely on local properties of the images. The main difficulty associated with using shape models is how to construct an accurate and representative model of shape given a necessarily finite set of training images.

Although there exist numerous techniques for representing shape in a form amenable to automated processing, statistical shape models are the most common in medical image analysis [Heimann and Meinzer, 2009]. In the *point distribution model* (PDM), a shape is described by a set of N points placed on the object:

$$\mathbf{p} = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}. \tag{2.1}$$

These points are called *landmarks* and typically correspond to salient anatomical features which are preserved between all subjects in the population. Alternatively the landmarks may be defined mathematically or geometrically, for example as points of high curvature or other extrema.

In order to construct a statistical shape model, the landmarks from each example in the training data must be aligned so that they are in correspondence. The most common method for aligning the training examples is *generalised Procrustes analysis* (GPA) [Gower, 1975]. Essentially GPA transforms the training examples to minimise the Euclidean distance to the mean shape (i.e. the mean of the landmark points in

the training set). After alignment of the shapes using GPA, dimensionality reduction is applied to find the principal modes which best describe the observed variation. This is usually carried out using linear principal component analysis (PCA) [Jolliffe, 2002].

As the set of landmark points $\mathbf{p}$ describe the shape of an object, valid shapes can be approximated by a linear combination of modes (eigenvectors), written as

$$\mathbf{p} = \bar{\mathbf{p}} + \sum_i b_i \mathbf{v}_i \tag{2.2}$$

where $\bar{\mathbf{p}}$ is the mean shape vector, $\mathbf{v}_i$ are the orthogonal modes of shape variation and b_i are the shape parameters [Cootes and Taylor, 2004]. By changing the shape parameters, variations around the mean shape can be obtained. A common choice is to limit the range of each parameter b_i to $[-3\nu_i, 3\nu_i]$ where ν_i is the associated eigenvalue, which constrains it to be within ± 3 standard deviations of the mean. Note that this assumes the population of shapes follows a Gaussian distribution, which is not always a reasonable assumption [Cremers et al., 2003, Kirschner et al., 2011]. This issue is discussed in more detail after a description of how shape models can be used for segmentation.

The most widely used segmentation techniques incorporating statistical shape models are the so-called active shape models (ASMs) [Cootes et al., 1995]. These algorithms use statistical shape models learned from a representative training set of example shapes (expressed as landmark points) to constrain the segmentation to be within some plausible range of shape variation. The performance of ASM methods depends to a large extent on the quality of the learned shape model, which in turn requires a large enough training set of (usually manually) annotated data with corresponding landmarks. Active appearance models (AAMs) differ from ASMs by modelling the appearance properties of the region covered by the structure, rather than just the shape [Cootes and Taylor, 2004]. They are essentially an extension of ASMs which also incorporate models of the object appearance by taking into account intensity and textural properties of the image. These models have generally superseded purely ASM-based approaches to segmentation, mainly due to the performance increase obtained when additional features of the object are incorporated into the model [Cootes and Taylor, 2004]. A number of existing approaches to vertebra segmentation are based on active shape and appearance models or are extensions of the original techniques [Kadoury et al., 2013, Mizaalian et al., 2013, Kirschner et al., 2011, Pereanez et al., 2015, Castro-Mateos et al., 2015].

In Mizaalian et al. [2013] the authors present a method for 3D segmentation of vertebrae from CT using statistical shape models built from annotated training data. The statistical shape models were first generated using a training dataset

of 154 manually annotated single vertebra 3D volumes. At test time, a boundary detector trained on features extracted using orientation-selective filters was applied to guide the registration of the shape model to obtain the segmentation. Combining the top-down shape model with local boundary cues was shown to be effective on the CT images from 7 subjects considered as test cases by the authors.

It is well known that the standard point distribution model applied to image segmentation suffers from a number of limitations arising from the linearity of PCA applied to find the modes of variation. The standard formulation of PCA is sensitive to outliers in the population due to the least squares estimation of the principal modes, which has motivated alternatives to be put forward in the literature [Heimann and Meinzer, 2009]. In addition, the assumption that the shape variation is Gaussian distributed is not always appropriate. Shape models learned using linear PCA cannot capture more complex deformations of the vertebrae, such as those that may occur in the presence of pathology [Kirschner et al., 2011].

In Kirschner et al. [2011] the authors attempt to address these limitations by using nonlinear shape models based on *kernel principal component analysis* (KPCA) [Scholkopf et al., 1998]. The idea behind KPCA is to first map the original data to a feature space using the so-called *kernel property* [Hastie et al., 2009]. Standard linear PCA in then applied in this transformed space defined by the choice of (centred) kernel function $\tilde{K}(\mathbf{p}, \mathbf{p}')$ (see Kirschner et al. [2011] for a more detailed description). The energy function for the shape model is given by

$$E(\mathbf{p}) = \sum_i \frac{\beta_i^2}{\nu_i} + \frac{1}{\epsilon}\left(\tilde{K}(\mathbf{p}, \mathbf{p}) - \sum_i \beta_i^2\right) \tag{2.3}$$

where ν_i is an eigenvalue of the kernel (Gram) matrix and $\beta_i = \sum_j b_{ij}\tilde{K}(\mathbf{p}_i, \mathbf{p})$ where b_{ij} is the j-th element of eigenvector $\mathbf{b}_i$. Larger values of the parameter ϵ allow larger deviations from the shapes in the training set. Minimisation of the shape energy function was carried out iteratively using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Press et al., 2007]. The shape energy term was combined with an appearance-based energy term by fitting Gaussian distributions to values of the gradient and gradient magnitude around the landmark points. The authors apply the model to vertebra segmentation from CT images and show that the nonlinear shape model leads to improvements in accuracy compared to the standard (linear) statistical shape model. While obtaining promising results on CT data, the method did however rely on manual interaction to initialise the optimisation parameters.

It is interesting to note that the same shape energy as equation (2.3) was previously formulated within a variational framework in Cremers et al. [2003]. In this work

the authors described a level set segmentation process based on the Mumford-Shah functional (discussed in Section 2.3.4), which integrated a nonlinear shape term into the energy. The training data was transformed non-linearly into a feature space and the resulting mapped data was modelled by a Gaussian distribution. The corresponding density in the original space was highly non-Gaussian enabling a more flexible shape representation. This variational approach has not yet been applied to problems in medical image segmentation.

Nonlinear shape representations such as those obtained using KPCA are a promising area for future research in shape-based medical image segmentation. This is especially true when the method must operate effectively on potentially abnormal anatomy; for example in clinical applications, where the presence of pathology may require a more flexible representation of shape.

An obvious disadvantage of using statistical shape models for segmentation is the need for appropriately labelled training sets, which often requires manual placement of the landmark points. In 3D this can be extremely time consuming and prone to annotator variability, although some attempts have been made to automate the landmark generation procedure [Campbell and Petrella, 2015, Davies et al., 2010, Cootes and Taylor, 2004, Souza and Udupa, 2005]. In addition, the construction of a representative prior model of shape may require a prohibitively large set of annotated training images. Methods that focus instead on lower-level properties of the images are likely to be better suited to smaller datasets where there is consequently only limited ability to accurately model global properties such as shape. Another limiting factor in the use of most shape modelling techniques is that they are generally unsuitable for segmentation in the presence of pathology (such as tumours) [Heimann and Meinzer, 2009]. This is because the pathology often deforms the shape of the object so that it is far outside the range of the statistical shape model. More generally, any situation causing the shape of the anatomy to deviate substantially from the training set can be problematic. This is true for example if parts of the anatomy are missing, additional parts are present, or if they are connected differently to surrounding structures.

2.3.2 Deformable Surfaces and Level Sets

Active contours have been very influential in general image segmentation, where they are also referred to as *snakes* [Kass et al., 1988, Blake and Isard, 1998]. The general idea is to minimise an energy functional, which describes a parametric contour subject to various internal and external forces. As the functional is minimised iteratively, the contour "moves" to fit the edges of the region of interest.

Let $m : [0,1] \mapsto \mathbb{R}^2$ be a parameterised contour. The energy functional itself is given by the sum of the internal and external energy of the contour, which can be written as

$$E(m) = \underbrace{\alpha \int_0^1 \left| \frac{\partial^2 m}{\partial u^2} \right|^2 \mathrm{d}u + \beta \int_0^1 \left| \frac{\partial m}{\partial u} \right|^2 \mathrm{d}u}_{\text{Internal energy}} + \underbrace{\gamma \int_0^1 V\left(m(u)\right) \mathrm{d}u}_{\text{External energy}} . \tag{2.4}$$

The internal energy enforces smoothness of the contour, while the external energy forces the contour towards the boundary of the object to be segmented [Wirjadi, 2007]. The two terms of the internal energy can be seen as modelling the tension and rigidity of the contour, respectively. The external energy is usually defined heuristically so as to take on lower values at object boundaries, with a common choice being some function of the gradient (e.g. $V\left(m(u)\right) = -|\nabla I\left(m(u)\right)|^2$). The coefficients α, β and γ control the relative importance of each term and are dependent on the particular problem under consideration. Many active contour formulations additionally incorporate a so-called "balloon" term [Cohen, 1991] that provides an outward normal force to the contour, enabling the contour to be initialised inside of the object before expanding outward towards the boundary.

The analogue of active contour methods for three dimensional images is provided by *deformable surfaces* [McInerney and Terzopoulos, 1996]. As the name suggests, rather than considering one dimensional contours, these approaches fit a 2D surface to the 3D object using a similar energy formulation. In this case the evolution of the surface typically corresponds to the deformation of the vertices of a discrete triangular or tetrahedral mesh.

Despite being widely used in medical image analysis, the original snakes model suffers from several major limitations. The standard snakes model uses an *explicit* representation of the curve or surface and is evolved towards a local optimum of the search space. The local nature of the optimisation means that the curve or surface is prone to becoming stuck at local minima, resulting in a poor segmentation. As the explicit representation of the snake forms a single unbroken curve or surface, segmentation of objects with holes (such as a simple torus) is not possible in a single step. Application to more complex topology requires multiple initialisations near to the boundaries of the object and thus is usually carried out in an interactive manner. These limitations motivate more flexible *implicit* representations of the surface, which have the additional advantages of being less sensitive to noise and more numerically stable [Younes, 2010].

Implicit representations are usually defined in terms of *level set* methods [Osher and Sethian, 1988, Cremers et al., 2007]. In level set methods a function $\phi : \Omega \mapsto \mathbb{R}$ is

used to define the implicit representation of a contour C as the following

$$C = \{p \in \Omega \mid \phi(p) = 0\} \tag{2.5}$$

where p denotes the coordinates of a point in the image domain. The level set function ϕ is typically defined to be a *signed distance function* between p and the contour, i.e.:

$$\phi(p) = \pm \mathrm{dist}(p, C) \tag{2.6}$$

where ϕ is negative inside C and positive outside (or *vice versa*). Optimisation is carried out by evolving the function ϕ, which propagates the contour. The main advantage of the implicit representation is that the topology of the curve or surface is not fixed, enabling segmentation of disjoint structures that are otherwise problematic using an explicit representation.

One level set formulation which has become very popular in medical image analysis is known as *geodesic active contours* [Caselles et al., 1995, 1997]. In geodesic active contours the contour evolves by gradient descent with respect to the level set function:

$$\frac{\partial \phi}{\partial t} = g(I)\, |\nabla \phi|\, \mathrm{div}\left(\frac{\nabla \phi}{|\nabla \phi|}\right) + \nabla g(I)\, \nabla \phi \tag{2.7}$$

where $\frac{\partial \phi}{\partial t}$ denotes the derivative with respect to time and $g(I)$ is some function of the image I that attracts the contour towards the boundary of the object. Note that the divergence of the normal vector is equivalent to the local curvature.

Based on the geodesic active contour formulation, in Huang et al. [2013] the authors introduced a variational[3] level set method for segmentation of vertebrae slices in CT images. The method used an energy functional integrating both edge and region terms that is able to better cope with intensity inhomogeneities and blurred boundaries compared to the standard geodesic active contour formulation. Initialisation of the level set function was carried out using Otsu's method to find a global intensity threshold, enabling automated segmentation without user interaction. The authors demonstrated that 3D segmentations could also be obtained from CT images by reconstructing the volumes from individually segmented slices.

In Lim et al. [2013] the authors describe a deformable model for vertebrae segmentation from CT images based on the *Willmore flow*. Optimisation was carried out within a level-set framework which also incorporated prior shape information in the form of a kernel density estimator [Cremers et al., 2006]. The *Willmore energy* is formulated as

$$E(M) = \int_M h^2 \, \mathrm{d}A \tag{2.8}$$

[3]Variational methods are discussed in detail later in the review.

where h is the *mean curvature* of the surface M. The Willmore energy can be viewed as a measure of the degree to which the surface M deviates from a sphere, and so effectively encodes a prior preference for spherical shapes; the Willmore flow refers to the geometric (gradient) flow of the Willmore energy. The authors incorporated the Willmore flow into a level set framework as a geometric functional along with a gradient-based edge indicator function to better preserve the vertebra boundary. The method was shown to obtain superior performance when compared with a number of other segmentation methods, both with and without shape constraints. The authors note however that the improvement in accuracy does come at a computational cost due to the energy functional incorporating multiple terms. In addition, segmentation was only performed semi-automatically, with the user required to select a seed point within each vertebra prior to processing. A legitimate question with all semi-automatic approaches to segmentation is how dependent the results are on the manual initialisation. This is an obvious concern when the method is aimed at clinicians and other end users where the required manual interaction with the method is to be minimised. However, it could be possible to automate the placement of the initial seed points by using a separate method to detect the vertebrae prior to segmentation.

In Kim and Kim [2009] the authors introduced a method using deformable surfaces (or "fences") to separate and label individual vertebrae in CT images. The authors started by pre-processing the image using Gaussian filtering and greyscale morphological operations to obtain a "valley-emphasised" image. The intervertebral discs were then detected by searching along rays initialised from seed points placed inside the vertebrae. A deformable surface was then fitted through the intervertebral discs between adjacent vertebrae, which enabled a separate labelling of each individual vertebrae in the volume. The authors used an energy function for the deformable surface based on first- and second-order derivatives of the contour, as in the original snakes model. The valley-emphasised image was also used in the energy function to promote alignment of the segmentation along the boundaries of the vertebrae. The authors only reported a qualitative evaluation of the method by radiologists, so direct comparison of the segmentation performance with other techniques is not possible. Although the resulting method is fully automated, by tailoring the method to the properties of CT images it is not likely to be effective when applied to MRI data.

A disadvantage of deformable models of the kind described in this section is that they are highly sensitive to the initial conditions, meaning that the final segmentation result is dependent on an accurate initialisation [Withey and Koles, 2007]. As such, they have a tendency to become stuck at local optima if the initialisation is not chosen correctly. Although in certain cases it is possible to automate the

initialisation, this is often assisted by manual interaction with the user and thus is time-consuming and subject to human error. Another primary drawback of deformable models is that they are generally slow to converge, especially in the case of 3D segmentation where the discretisation of the surface may lead to a heavy computational load during optimisation.

2.3.3 Registration and Atlas-Based Methods

Image registration is a central topic in medical image analysis with many applications in areas such as multi-modality image fusion and population modelling [Sotiras et al., 2013]. Registration-based approaches to segmentation typically start with a labelled reference image (or *atlas*) constructed from training data, which is then "warped" onto a new image using some class of transformation. Once the atlas has been registered to the target image, the class labels can simply be transferred from the atlas to obtain the segmentation. The registration of a template can also be used to propagate information on the global shape (for example of individual vertebrae or the lumbar spine), which can then guide the segmentation on a pixel or voxel level.

In medical image segmentation the class of transformation is most commonly a *non-rigid registration* [Sotiras et al., 2013], which is better able to deal with the complex deformations arising due to anatomical variation. Mathematically, an atlas $\mathcal{A}$ is a mapping from the n-coordinates of the image to a set of class labels:

$$\mathcal{A} : \mathbb{R}^n \mapsto \mathcal{L}. \tag{2.9}$$

The estimated class label $l_p \in \mathcal{L}$ for a location $p \in \mathbb{R}^n$ in the target image is found by the mapping

$$l_p \mapsto \mathcal{A}(T(p)). \tag{2.10}$$

The transformation function $T(p)$ is parameterised and the process involved in finding the optimal set of parameters is referred to as *image registration* [Rohlfing et al., 2005]. In common with other image segmentation approaches, image registration is an inherently ill-posed problem and many different methods have been proposed in the literature using various regularisation techniques to constrain the solution [Sotiras et al., 2013].

Atlas-based methods have a long history in medical image segmentation and have recently been applied to segment the spine from CT data [Forsberg, 2015]. In this work the authors applied a non-rigid registration procedure which involved minimising the local phase difference between the reference and target images [Knutsson and Andersson, 2005]. The authors used multiple grey-level atlases to improve perfor-

mance, with the final labelling obtained by taking a majority vote over all of the individually registered atlases. As the resulting registration procedure is computationally intensive, it was implemented on a graphics processing unit (GPU) to reduce the run time complexity. The method was shown to obtain high accuracy when evaluated on a CT dataset consisting of 10 subjects.

The strength of atlas-based approaches is a result of their ability to consider the global anatomical structure of objects, rather than simply the intensity or textural properties of the image. Although they can often obtain accurate results, atlas-based methods suffer from the limitations associated with image registration in general. In particular the optimisation is often only guaranteed to find a local optimum and is dependent on initialisation. To obtain an accurate initialisation, additional steps are often necessary to roughly locate the object of interest. For example, in Forsberg [2015] the authors carry out extensive pre-processing involving intervertebral disc detection and vertebrae pose estimation prior to registration of the atlases. Constructing a representative atlas is also a challenging problem due to both the limited data available and the computational cost incurred when using multiple atlases. Care must be taken to make sure the atlas is not biased towards a specific set of subjects, which can only be countered by increasing the number of subjects and/or atlases. It has been suggested that these issues make atlas-based approaches best suited to segmenting structures with relatively low variability [Rohlfing et al., 2005]. These problems are similar to those discussed previously in the context of statistical shape models; namely the inability of the methods to cope with normal variants such as extra or missing anatomical parts, or various pathologies causing the anatomy to differ widely from the training examples.

A number of recent methods for vertebra or spine segmentation incorporate registration as one component of a larger, multi-stage segmentation framework [Kadoury et al., 2013, Klinder et al., 2009, Hammernik et al., 2015]. For example, in Klinder et al. [2009] the authors propose a multi-stage framework for segmentation of the spine from CT data. A rigid registration procedure is used to identify the individual vertebrae after they have been detected using an algorithm based on the generalised Hough transform. The final segmentation is obtained by adapting triangulated mesh shape models of the individual vertebrae using a collision detection algorithm. While the method was shown to obtain good results on CT data, the multi-stage architecture is very computationally intensive, which could limit its practical application. Adapting the method for segmentation of MR images also poses difficulties as certain pre-processing steps (such as spinal cord extraction), along with the gradient-based mesh adaptation algorithm rely on the higher contrast provided by CT.

2.3.4 Variational Methods

Variational methods for image segmentation have their origin in the work of Mumford and Shah [Mumford and Shah, 1989], who first proposed minimising the following energy functional to obtain a piecewise smooth approximation to the input image f:

$$E(u) = \lambda \int_{\Omega} (f-u)^2 \,\mathrm{d}x + \int_{\Omega \setminus S_u} |\nabla u|^2 \,\mathrm{d}x + \nu H^1(S_u) \tag{2.11}$$

where $u : \Omega \mapsto \mathbb{R}$ is a piecewise smooth function describing the sub-regions (segments) of the image, H^1 is the one-dimensional Hausdorff measure and S_u is a discontinuity set (segmentation boundary) [Pock et al., 2009]. The first two terms of the functional impose smoothness of u, while the third term regularises the boundary in terms of its one-dimensional Hausdorff measure (i.e. the length of the contour). A popular extension of the Mumford-Shah functional is the Chan-Vese model [Chan and Vese, 2001], which includes an additional term penalising the enclosed area and restricts the range of u to take on only two values [Getreuer, 2012].

Although the Mumford-Shah model has a sound mathematical basis, the lack of efficient algorithms for numerical implementation has limited its direct application to medical image segmentation (particularly 3D segmentation). The difficulty in finding efficient numerical approximations of the Mumford-Shah model is due to the non-regularity of the edge term [Pock et al., 2009]. Recently, more efficient *primal-dual* algorithms have been devised which allow for a globally optimal solution after convex relaxation [Chambolle and Pock, 2011]. The resulting optimisation problem is known as a "saddle point" problem and is more amenable to efficient minimisation algorithms.

A variational approach to 3D vertebra segmentation from CT data was recently described in Hammernik et al. [2015], which incorporates both shape and intensity priors into an energy functional. Letting Ω denote the image domain as before and $u : \Omega \mapsto \{0, 1\}$ denote the segmentation, the energy functional defined by the authors can be written as

$$E(u) = \mathrm{TV}_g + \lambda_1 \int_{\Omega} u f_s \,\mathrm{d}x + \lambda_2 \int_{\Omega} u f_b \,\mathrm{d}x \tag{2.12}$$

where f_s is a pre-registered mean shape model and f_b is a bone probability map. TV_g is the weighted *total variation* (TV) norm described in Reinbacher et al. [2010], which is particularly suitable for the segmentation of thin structures such as the vertebral processes. The weighted TV norm can be written as

$$\mathrm{TV}_g = g(p) n n^\top + n_0 n_0^\top + n_1 n_1^\top \tag{2.13}$$

where $n = \nabla I / \|\nabla I\|$, n_0 is the tangent of n and $n_1 = n \times n_0$. The edge function $g(p)$ is defined as $g(p) = \exp\left(-\alpha \|\nabla I(p)\|^{\beta}\right)$.

The shape term f_s of the energy functional was obtained by first registering a mean shape model to the thresholded bone probability map. By using only coarse mean shape models, the authors avoid the requirement for manually annotated landmark points prior to registration. The bone probability map f_b was estimated using intensity histograms of foreground and background regions from the training data. After discretisation, minimisation of the energy can be carried out using the primal-dual algorithm described in Chambolle and Pock [2011]. Numerical implementation of the method made use of a GPU to exploit parallelisation and improve run time performance. The method was shown to obtain accurate results when evaluated on a dataset of CT volumes from 10 different subjects, although manual interaction was required to initialise a single point at the centre of each vertebrae.

One of the main advantages of spatially continuous variational approaches to segmentation is the avoidance of *metrication artefacts*, which can occur with grid-based random field models due to the discrete approximation of the Euclidean boundary [Nieuwenhuis et al., 2013]. However, as mentioned previously one of the main challenges in using variational methods is in finding efficient algorithms for minimising the resulting energy functionals [Pock et al., 2009]. The development of more efficient algorithms for numerical approximation of variational methods, in combination with the decreasing cost of GPU hardware, should enable more widespread application of these methods to 3D medical image segmentation.

2.3.5 Markov Random Fields

Markov random fields (MRFs) are a type of undirected graphical model which have proven to be successful for a wide range of tasks in image processing and computer vision [Geman and Geman, 1984, Blake et al., 2011]. An MRF represents an image by an undirected graph where the nodes correspond to the individual elements of the image (e.g. pixels or voxels) and the local neighbours of a node are linked by edges. One of the main advantages of MRF models is that long range correlations across the image are obtained implicitly by considering only local interactions between pairs of neighbouring nodes, which ensures that the computations remain tractable. A brief description is first given of the main concepts behind MRFs. A more in-depth discussion of MRFs is provided in the next chapter, where they form the basis for an approach to vertebra segmentation from 2D images.

In an MRF the state of each variable i depends on the state of its neighbours in the graph, where $\mathcal{N}_i$ is used to denote the indices of the sites neighbouring i. Pixels

or voxels are usually defined to be neighbours if they are immediately adjacent (in cardinal or diagonal directions). A *configuration* of a random field X is denoted $\mathbf{x} = \{x_1, \ldots, x_N\}$ and corresponds to a particular assignment of states to the variables. For image segmentation, the variables are associated with the image sites and the states correspond to the class labels assigned to those sites. Formally, letting $\mathcal{X}$ denote the set of all configurations, a random field X is an MRF if

$$P(\mathbf{x}) > 0 \quad \forall \mathbf{x} \in \mathcal{X} \tag{2.14}$$

and

$$P(x_i \mid x_{j \neq i}) \equiv P(x_i \mid x_{\mathcal{N}_i}). \tag{2.15}$$

The latter condition is known as the *Markov property*. The Hammersley-Clifford theorem shows that $P(\mathbf{x})$ is characterised by a Gibbs distribution:

$$P(\mathbf{x}) = Z^{-1} \exp\left(-E(\mathbf{x})\right) \tag{2.16}$$

where Z is a normalisation constant called the *partition function*. The term $E(\mathbf{x})$ is an energy function, given by a sum over the neighbourhoods (or *cliques*) of the MRF:

$$E(\mathbf{x}) = \sum_c V_c(\mathbf{x}) \tag{2.17}$$

where $V_c(\mathbf{x})$ is a potential function whose value depends on the configuration of the clique c. A popular class of MRF models are *conditional random fields* (CRFs), which define a posterior distribution for the labels $\mathbf{x}$ given an associated set of observations (features) [Blake et al., 2011]. CRFs allow for the representations of complex dependencies between the labels and observations and are powerful models for image segmentation problems. A more detailed discussion of CRFs is given in the next chapter of the thesis.

The increasing popularity of MRFs in computer vision has been driven by a class of efficient algorithms for inference known as *graph cuts* [Boykov and Funka-Lea, 2006]. In specific cases, such as the problem of binary segmentation, graph cuts can find the globally optimal solution (i.e. the global minimum energy) in polynomial time provided that the energy function is *submodular* [Kolmogorov and Zabih, 2004, Szeliski et al., 2008]. Figure 2.4 is a simple illustration of how a cut through a graph representation of an image results in a binary segmentation, expressed as an assignment of class labels. The successful results obtained in the general field of computer vision has led to increasing interest in these models for various tasks in medical image analysis, particularly with respect to segmentation.

Despite the widespread use of MRF models for general image segmentation problems, relatively few attempts have been made to apply them to 3D segmentation

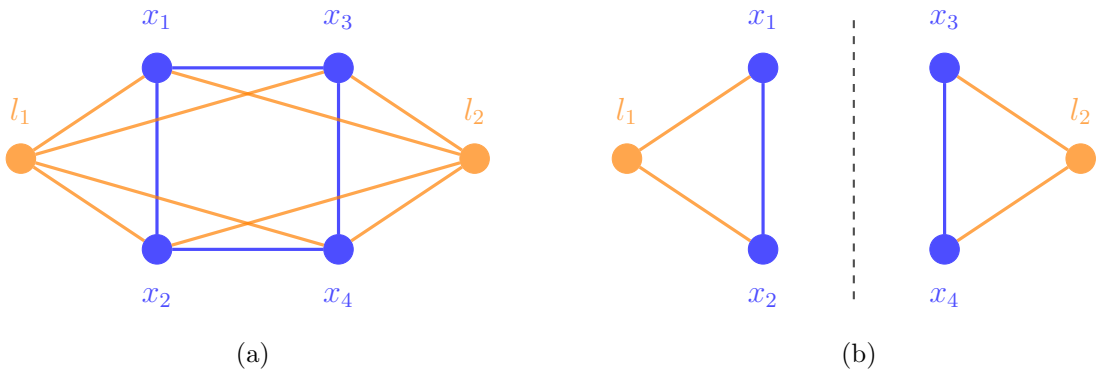


Figure 2.4: (a) Example graph construction for a 2-class segmentation problem on a 4×4 pixel image. In blue are the edge links between pixel nodes. In orange are the links between class nodes and pixels. (b) Segmentation resulting from cutting the graph through the dashed black line. Each of the 4 pixels are connected to a single unique class, which determines their class label assignment (i.e. $(x_1, x_2) \in l_1$ and $(x_3, x_4) \in l_2$).

of the spine from medical images. One reason for this is that the sheer size of the 3D volumes often prohibits the direct application of MRFs on a voxel level using standard hardware. Recently, effective algorithms for clustering voxels into larger groups (called *supervoxels*) have made it more feasible to apply various forms of MRF models to these problems. Operating on supervoxels dramatically reduces the computational complexity and enables the globally optimal solution to be obtained in polynomial time using graph cuts. An MRF model based on supervoxels is presented in this thesis for 3D segmentation of the lumbar spine and is discussed at length in Chapter 4.

Another possibility for applying MRF models in 3D is to first detect the regions of interest for initialisation prior to segmentation on a voxel level. This is the approach taken in Kelm et al. [2013], who used a Markov random field to segment vertebral bodies and intervertebral discs from CT images after fast detection using an algorithm based on *marginal space learning* [Zheng et al., 2008]. The proposed MRF formulation modelled the distribution of voxel intensities for each class using standard histograms estimated from randomly sampled voxels within detected regions of the image. The authors used a standard pairwise term to model the intensity difference between voxels that was first introduced in Boykov and Funka-Lea [2006]. The method was shown to obtain good results on CT data in terms of vertebral body segmentation and was also able to segment intervertebral discs from MR images, although no quantitative evaluation of the segmentation results was presented in this case. For application of the method to vertebra segmentation from MRI data, purely intensity-based information is likely to fail due to the lower contrast and intensity inhomogeneity of the images. More robust models are therefore required to obtain accurate and consistent results, such as those discussed in the following

chapters of the thesis.

Recent work has also investigated the use of MRF models with higher-order potential functions, which can lead to improvements over standard pairwise formulations in certain cases [Wang et al., 2013, Komodakis and Paragios, 2009, Kohli et al., 2008]. A recent application of higher-order MRFs to the problem of vertebra segmentation from CT and MRI was described in Kadoury et al. [2013], which enabled global pose and local shape parameters to be incorporated into a discrete optimisation framework. The higher-order potentials of an MRF model were used to encode geometric modes of variation of the vertebrae. In order to address some of the limitations associated with standard shape models, the authors used a nonlinear manifold embedding to capture global shape variations from a set of training images. The shape manifold was learned from a large training set of 711 scoliotic spines, each of which had been manually annotated with landmark points. Optimisation of the model was carried out using linear programming and the fast primal-dual algorithm of Komodakis et al. [2011]. The method obtained promising segmentation results on both CT and MRI images of the lumbar vertebrae. As the authors relied on a large annotated training set to learn the manifold, it is unclear how successful this approach would be on the much smaller datasets common in medical image analysis.

A promising area of future research is the possible combination of MRF models with deformable registration techniques, such as those described earlier in the review. This presents the possibility of using top-down shape constraints provided by a registration procedure, alongside low-level cues from MRFs in a joint optimisation framework. A similar approach has recently been explored in the context of brain tumour segmentation from 3D MR images [Parisot et al., 2013].

2.3.6 Conclusion

It is evident from the various methods covered in this review that a completely general purpose method for automatic lumbar spine segmentation does not yet exist. This is perhaps not surprising given that medical image segmentation is very challenging even for human experts. Previous results from the literature suggest that some notion of spatial context needs to be present in order to achieve accurate and reliable segmentation, as discrimination of structures is usually not possible using only information on the intensity distribution of the images. This is especially true of MRI, where the lower contrast often leads to overlapping intensity values at the boundaries of objects.

Approaches based on Markov random fields have the advantage of enabling a globally optimal solution to be obtained in polynomial time using graph cuts. They are also

very flexible, allowing multiple constraints to be incorporated into the potentials to enhance performance. The main drawback is the difficulty in incorporating into the model high-level information on global properties such as the shape or pose (orientation) of the anatomy, due to the local nature of the potentials. In the context of medical image registration, this type of global information is often provided by a single manually-annotated atlas or multiple atlases. However, relying on atlas-based techniques alone could pose problems due to the large variation between spines, which can make it difficult to construct a representative atlas.

The choice of approach also depends on the particular type and amount of annotated training data available. It is well known that methods depending on learned global representations of shape, such as atlases and statistical shape models, are particularly sensitive to the properties of the training set [Heimann and Meinzer, 2009]. Successful application of these methods requires a large enough dataset to capture most of the anatomical variation likely to occur in a population, which is clearly problematic in cases where only small annotated datasets are available.

A promising area of future research is in the combination of registration-based methods with graph-based models such as Markov random fields. This could allow the incorporation prior knowledge of anatomical shape into the model alongside local cues based on low-level image appearance. In the simplest case this could take the form of local constraints on the shape and appearance which are encoded into the energy function. The development of efficient optimisation algorithms for higher-order MRFs also presents opportunities for more expressive segmentation models.

2.4 Evaluating Segmentation Quality

In order to evaluate the quality of segmentations obtained using an automated method, a standard approach is to compare them with a set of manually annotated images. This provides a direct measure of the agreement between the automated method and a human annotator or annotators. This section describes the evaluation measures used throughout the thesis to evaluate the segmentation method, while the next section provides details on the annotated datasets themselves. In order to facilitate comparison with other existing segmentation methods, the evaluation measures considered in this thesis are well known and widely used in medical image analysis for assessing the quality of segmentation results.

2.4.1 Evaluation Measures

The Dice similarity coefficient (DSC) is used to measure the segmentation quality in terms of the overlap between two segmentations and is equivalent to the well-known F-score. Given two segmentations $\mathbf{x}$ and $\mathbf{x}'$, the Dice score is defined as

$$\mathrm{DSC}(\mathbf{x}, \mathbf{x}') = \frac{2|\mathbf{x} \cap \mathbf{x}'|}{|\mathbf{x}| + |\mathbf{x}'|}. \tag{2.18}$$

The score is in the range $[0, 1]$ with 0 indicating no overlap and 1 indicating maximum overlap (i.e. a perfect match between the two segmentations). Note that the Dice score is sometimes reported in the literature as a percentage value between 0% and 100%. In order to avoid confusion, all Dice scores reported in this thesis are in the range from 0 to 1 and are converted from percentages where necessary.

Three standard surface distance measures are also used in the thesis to evaluate the segmentation quality:

1. The average symmetric absolute surface distance (ASD) is determined by finding for each set of boundary points of both the segmentation and corresponding manual annotation, the closest boundary points of the other set. The mean of the Euclidean distances to the closest points gives the score, with 0 indicating a perfect segmentation. This can be written as

$$\mathrm{ASD}(\mathcal{B}, \mathcal{B}') = \frac{1}{|\mathcal{B}| + |\mathcal{B}'|} \left(\sum_{i \in \mathcal{B}} \min_{j \in \mathcal{B}'} d_{ij} + \sum_{j \in \mathcal{B}'} \min_{i \in \mathcal{B}} d_{ij} \right) \tag{2.19}$$

   where $\mathcal{B}$ and $\mathcal{B}'$ are the sets of boundary points for the two segmentations and d_{ij} is the distance between boundary points i and j.

2. The RMS symmetric surface distance takes the squared distances between the two sets of boundary points, with the final score defined as the square root of the average squared distances:

$$\mathrm{RMS}(\mathcal{B}, \mathcal{B}') = \left(\frac{1}{|\mathcal{B}| + |\mathcal{B}'|} \left(\sum_{i \in \mathcal{B}} \min_{j \in \mathcal{B}'} d_{ij}^2 + \sum_{j \in \mathcal{B}'} \min_{i \in \mathcal{B}} d_{ij}^2 \right) \right)^{1/2}. \tag{2.20}$$

3. Finally, the maximum symmetric absolute surface distance (MSD) is similar to the first measure but takes the maximum of the distances instead of the mean:

$$\mathrm{MSD}(\mathcal{B}, \mathcal{B}') = \max \left\{ \max_{i \in \mathcal{B}} \min_{j \in \mathcal{B}'} d_{ij}, \max_{j \in \mathcal{B}'} \min_{i \in \mathcal{B}} d_{ij} \right\}. \tag{2.21}$$

   This is mathematically equivalent to the Hausdorff distance between the two sets of boundary points.

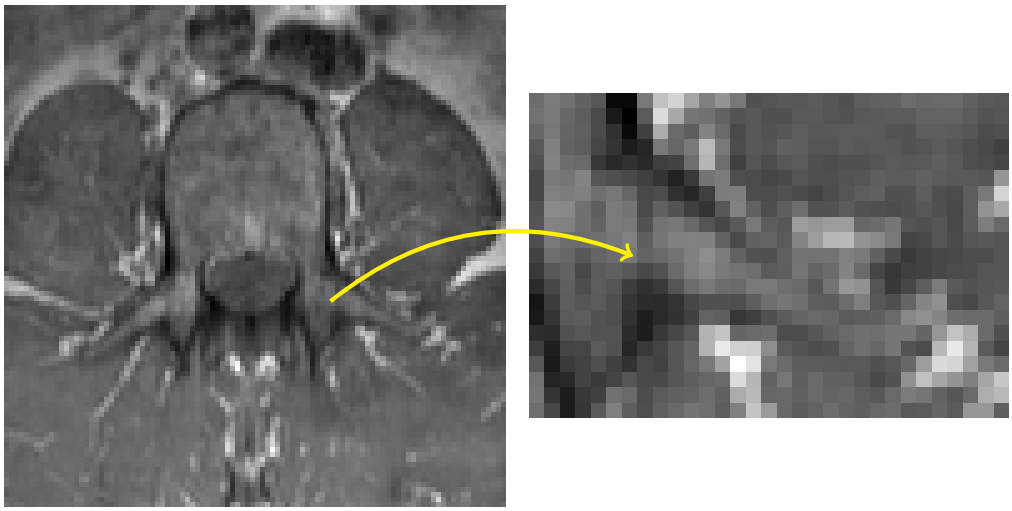


Figure 2.5: Close-up of a right transverse process from an example T1-weighted axial MR image, showing merging between vertebra and background regions.

Further discussion of these evaluation measures is provided in Gerig et al. [2001]. Note that the three surface distance measures are usually scaled based on the image resolution and reported in terms of physical distances in millimetre units (mm).

2.5 Annotated Datasets

Although the goal of this thesis is a method for fully-automated lumbar spine segmentation, manual annotation is still required in order to provide a means for evaluating the performance of the method. In addition, annotated data is required in order to form a training dataset for learning model parameters.

The quality of a segmentation is usually quantified by comparison with a manual annotation of the same dataset, often by measuring the degree of overlap between the two. The objective is for the method to obtain a high level of agreement with the manual annotations, ideally obtaining comparable performance to a human annotator. This section provides details on the manually annotated datasets used in the following chapters of the thesis.

2.5.1 Comparison of MRI Protocols

A common issue with axial MR images is the blending of intensity and texture of the vertebral regions with other non-vertebra regions at the boundary, as shown in Figure 2.5. This may introduce difficulties for automated segmentation, especially when the method is dependent on local properties of the images.

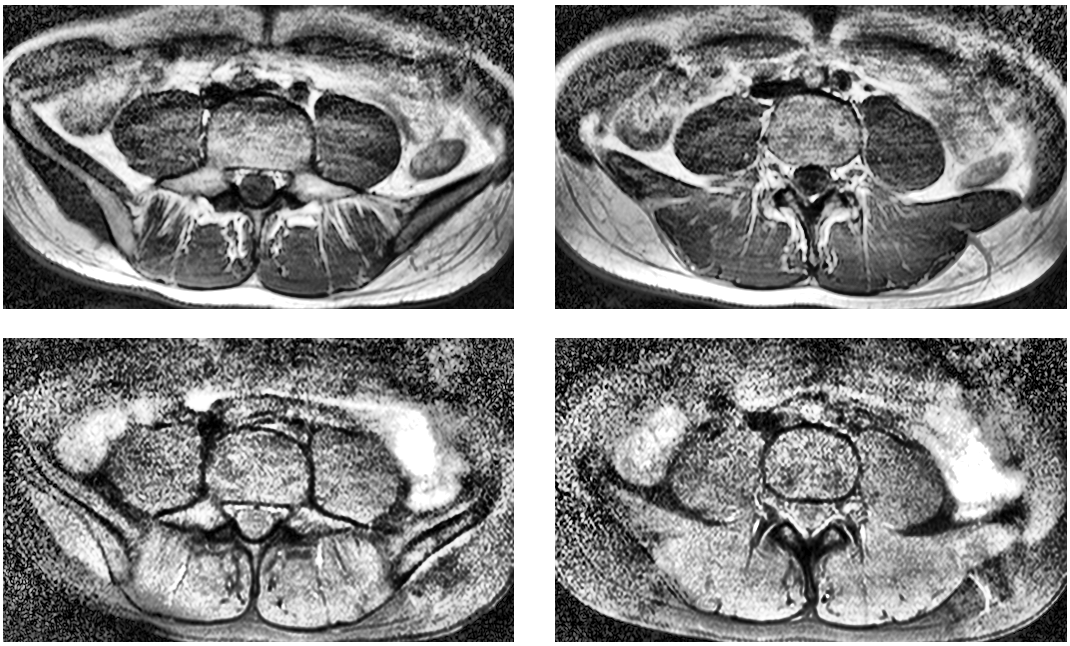


Figure 2.6: Axial slices showing vertebra on the left and intervertebral disc on the right. Images on the top row are T1-weighted. The bottom row shows the images obtained of the same subject using fat suppression.

Prior to obtaining a 3D MRI dataset of lumbar vertebrae (described in the next section under MRI Dataset 2), a number of different scanning protocols were also considered and evaluated with respect to their suitability for segmentation. The images were obtained on 29 August 2013 using the MRI scanner at St Luke's Campus, University of Exeter. Scans were carried out on two subjects under different parameter settings. It was decided that T1-weighted images were the most suitable, as loss of anatomical detail (particularly the merging of object boundaries) was a concern when using the alternative MRI protocols.

Fat suppression techniques [Hashemi et al., 2010] were also tested as a way of enhancing the cortical boundary surrounding the vertebrae. Figure 2.6 shows the result of using fat suppression alongside a standard T1-weighted image of the same subject. In particular, the boundary around the vertebral body is noticeably easier to distinguish in the fat suppressed slice. A major limitation of images acquired in this way is that the intensity properties of the resulting images make it difficult to distinguish between vertebra and intervertebral disc regions of the same subject, which will affect segmentation performance in 3D. This is illustrated in the bottom row of Figure 2.6, which shows a vertebra and intervertebral disc slice obtained using fat suppression. As the ultimate aim of the thesis is 3D segmentation of both vertebrae and intervertebral discs as distinct objects, fat suppression was rejected as an enhancement technique for the MRI data.

Table 2.1: Overview of datasets used in the thesis. IPR is the in-plane resolution and CPR is the cross-plane resolution (slice thickness). The last column gives the number and dimension of the manually annotated images.

Dataset	Modality	Subjects	IPR (mm)	CPR (mm)	Annotated
MRI Dataset 1	T1 MRI	21	0.49–0.50 mm	4.00 mm	63 (2D)
MRI Dataset 2	T1 MRI	8	1.02 mm	1.20 mm	8 (3D)
MRI Dataset 3	T2 MRI	15	1.25 mm	2.00 mm	15 (3D)
CT Dataset	CT	10	0.31–0.36 mm	1.00 mm	10 (3D)

2.5.2 Description of Datasets

Table 2.1 provides an overview of the different datasets used in the thesis. The following sections discuss the properties of each dataset in detail and also describe how manual annotation of the images was carried out.

MRI Dataset 1

The first MRI dataset used for the experiments consists of 2D axial cross sections of lumbar vertebrae from 21 healthy subjects. The images were acquired on a 1.5 T MR scanner (Intera, Philips) using a receive-only spine coil (Synergy, Philips). T1 weighted turbo-spin-echo scan sequences were used with a repetition time, TR, of either 400 ms or 497 ms, an echo time, TE, of 8 ms and 4 signal averages. Slices were obtained with an in-plane resolution of between 0.49 mm and 0.50 mm (depending on TR) and a slice thickness of 4 mm. The dataset was collected as part of a study carried out at the University of Exeter. All subjects gave informed consent prior to taking part in the study.

Manual annotation of the images was carried out by the author, which required delineation of the vertebrae in each individual slice. A total of 63 manually annotated images were obtained by selecting 3 central slices from each of the subjects. A subset of the annotated images were compared with a second annotator[4] in order to measure the inter-annotator agreement (see Section 2.5.3).

MRI Dataset 2

The second MRI dataset consists of 3D scans of a section of the lumbar spine from 8 different subjects, encompassing the lumbar vertebrae from L3 to L5. The images were acquired on a 1.5 T MR scanner (Intera, Philips) using a receive-only spine coil (Synergy, Philips). T1-weighted turbo-spin-echo scan sequences were used with a repetition time, TR, of 1000 ms and an echo time, TE, of 8 ms. The images were

[4]Dr Judith Meakin from the Department of Physics, University of Exeter, UK.

obtained with an in-plane resolution 1.02 mm and a slice thickness of 1.2 mm. Ethical approval was granted by the College of Engineering, Mathematics and Physical Sciences at the University of Exeter. All subjects gave informed consent prior to taking part in the study.

Annotation of the MRI data was carried out by the author, which involved manual delineation of the vertebrae boundaries in individual 2D slices of the 3D volumes. For manual annotation of the 3D MRI dataset the ITK-SNAP [Yushkevich et al., 2006] application was used, which enables volumes to be viewed from different anatomical planes for additional context during annotation. The voxels belonging to the separate vertebrae and sacrum in each scan were assigned unique labels to enable identification of the different structures.

MRI Dataset 3

The third MRI dataset used for the evaluation consists of T2-weighted turbo-spin-echo MR images from 15 different subjects provided for the MICCAI 2015 intervertebral disc localisation and segmentation challenge [Chen et al., 2015].[5] Each image consists of intervertebral discs of the lower spine from T11 to L5. A total of 7 intervertebral discs in each image have been manually annotated by the providers of the data. The images were acquired using a 1.5 T MRI scanner (Siemens Magnetom Sonata) and have an in-plane resolution of 1.25 mm and a slice thickness of 2 mm. An evaluation of the inter-annotator agreement on the dataset is given in Zheng et al. [2016], where it was found that the manual segmentations are consistent.

CT Dataset

The CT dataset consists of 3D scans of the lumbar spine from 10 different subjects. The dataset was provided for the MICCAI 2014 spine and vertebrae segmentation challenge and is publicly available (see footnote 5). All vertebrae in the images have been manually annotated by the providers of the data. The images were acquired with Philips or Siemens multidetector CT scanners using an in-plane resolution of between 0.31 mm and 0.36 mm with a slice thickness of 1 mm [Yao et al., 2012]. As the sacrum has not been annotated, it was cropped from the volumes below the middle axial slice of lumbar vertebra L5 to remove it from the evaluation. The volumes were additionally cropped above L1 to remove the thoracic vertebrae from evaluation.

[5] The dataset is available from `http://spineweb.digitalimaginggroup.ca`.

Table 2.2: Inter-annotator agreement on a sample of 10 images from MRI dataset 1. The table shows the minimum, median and maximum values over all individual images.

Sample size	Measure	Min	Median	Max
10 images	Dice score	0.88	0.93	0.97
	Avg. surf. dist. (mm)	0.02	0.11	0.33
	RMS surf. dist. (mm)	0.13	0.56	1.91
	Max. surf. dist. (mm)	2.12	6.61	21.15

2.5.3 Inter-Annotator Agreement

To assess the reliability of the manual annotations of the MRI data, the inter-annotator agreement between the author and a separate annotator was measured on a randomly sampled subset of the images from the dataset described in Section 2.5.2. MRI Dataset 1. The manual segmentations were carried out independently by both annotators and then compared using the same evaluation measures as described in Section 2.4. The evaluation measures essentially quantify the overlap and surface distances between the two sets of segmentations and together provide a measure of agreement between the annotators.

The results from the analysis are summarised in Table 2.2 for each individual evaluation measure. The mean Dice score between the two annotators was 0.93 ± 0.03 taken over all images in the sample. The values indicate a generally high level of agreement between the two annotators on the MRI data, suggesting that the manual segmentations are consistent. The maximum surface distance obtained over all of the comparisons was due to disagreement between annotators on the presence of a thin spinous process in one of the images.

2.6 Conclusion

In this chapter the problem of lumbar spine segmentation was introduced and a number of existing methods in the literature were reviewed. The chapter also provided details on the annotated datasets used in the remaining chapters of the thesis for evaluating segmentation results and investigated the inter-annotator agreement on a sample of the data.

Although the segmentation methods covered in the review have their own particular strengths, they tend to share certain problems; these include the inherent limitations of the optimisation procedure, sensitivity to initialisation and reliance on accurate prior information to constrain the solution. In addition, most of the approaches discussed are not suitable for small datasets due to the dependence on a representative

prior model of shape, which requires a large enough dataset to capture most of the anatomical variation in a population.

A comparison between manual segmentations carried out independently by two separate annotators showed a generally high level of inter-annotator agreement on a randomly sampled subset of MR images. A separate comparison between manual segmentations of the IVD dataset also found a high level of agreement among the annotators [Zheng et al., 2016]. As one of the main goals of automatic segmentation is to match the performance of human annotators, the manually annotated data can be used as a reference for evaluating the performance of automated segmentation methods.

The next chapter introduces a method for fully-automated 2D segmentation of lumbar vertebrae from MR and CT images. Although segmentation of vertebrae in 2D has direct application in a number of areas, the main motivation for initially focusing on 2D segmentation is that it provides a clearer picture of which path to take towards full 3D segmentation of the lumbar spine.

3 2D Segmentation of Lumbar Vertebrae

This chapter describes an approach for segmentation of lumbar vertebrae from 2D axial slices.[1] Although the focus of this chapter is predominantly on segmentation from MR images, results obtained on CT data using the same approach are also presented.

3.1 Introduction

Segmentation of vertebrae in axial cross sections has potential applications for clinical research in orthopaedics. This includes the measurement of axial vertebral rotation, where identification of anatomical landmarks is often carried out manually by multiple annotators [Vrtovec et al., 2010, Janssen et al., 2010]. Segmentation of axial slices also facilitates the analysis of specific parts of the vertebra which are either not present or only partially visible in sagittal cross sections, such as the neurocentral junction which bilaterally connects the pedicles to the vertebral body [Schlosser et al., 2013].

With respect to segmentation of vertebrae from MR images, a number of methods have been reported in the literature for automated segmentation of vertebral bodies in the sagittal plane [Carballio-Gamio et al., 2004, Huang et al., 2009, Zukic et al., 2012]. However, as these techniques only consider the vertebral bodies, the pedicles and posterior elements of the vertebrae are not taken into account. Attempting to segment all of the vertebra parts represents a greater challenge due to certain parts appearing disconnected in individual slices, which often include thin transverse and spinous processes (see left panel of Figure 3.1). The different intensity and textural properties of the vertebra parts also means that these features cannot be relied upon alone to discriminate the vertebra from the other background structures in the image. A method for segmentation of vertebrae in multiple anatomical planes was recently introduced in Wang et al. [2015], although manual cropping of the vertebrae was required in order to obtain accurate results.

[1] This chapter is based on work first published in Hutt et al. [2015a].

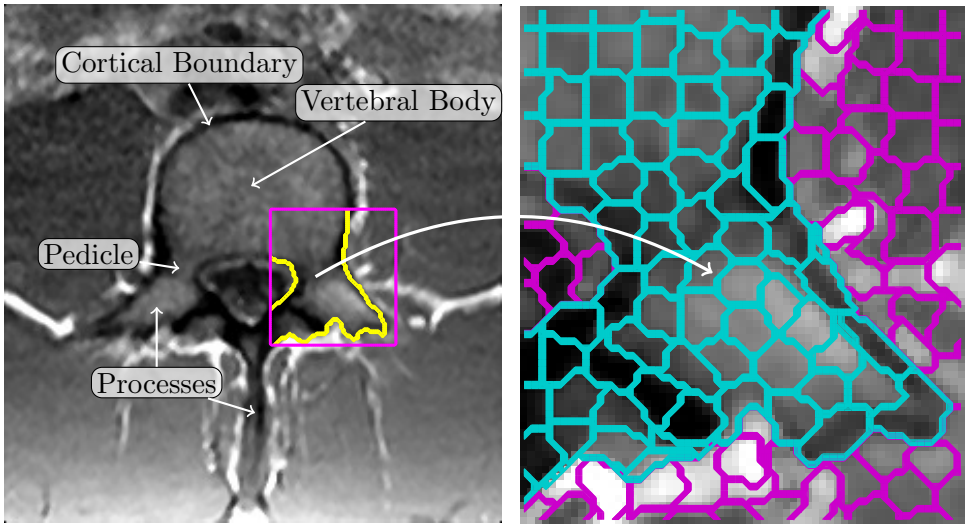


Figure 3.1: The left figure shows a 2D axial MR slice with ground truth contour (yellow) for a section of the vertebra. The right figure shows boundaries for superpixels assigned to the vertebra class (cyan) and background class (magenta). The superpixels preserve the boundary detail of the vertebrae by clustering contiguous pixels with similar intensities.

Many existing approaches to vertebra segmentation rely heavily on prior shape information, often in the form of an explicit shape template or statistical shape model [Peng et al., 2005, Kirschner et al., 2011, Kadoury et al., 2013]. While such methods have been applied successfully to segment a wide range of structures, there are a number of issues which make the application to vertebra segmentation problematic. The shape of the vertebra structure itself varies widely, both between different subjects and between different vertebrae within the same subject. As mentioned previously, when considering 2D slices various parts of the vertebra may appear disconnected in a single slice due to overlapping tissue and the presence of partial volume effects. These issues make it difficult to construct representative models of shape in 2D, especially when very large datasets of manually annotated images are not readily available.

This chapter describes a novel fully automated method for segmentation of lumbar vertebrae from 2D axial slices. The method uses a conditional random field (CRF) model on superpixels (groups of contiguous pixels with similar intensities). Operating on superpixels reduces computational complexity and enables more descriptive features to be extracted to characterise the separate classes, while the CRF relates the underlying class labels of the superpixels to the observed features and promotes spatial consistency. Supervised learning is used to train a classifier on labelled superpixel features and obtain probability estimates expressing the likelihood of belonging to either the vertebra or background class. Distance metric learning [Weinberger and Saul, 2009] is also used to find an appropriate dissimilarity measure between superpixel pairs. The probability estimates and learned distance metric are incor-

porated into the CRF model in the form of first- and second-order clique potentials of the CRF energy function. This formulation enables minimisation of the energy function to be carried out efficiently using graph cuts [Boykov and Funka-Lea, 2006].

The performance of the method is evaluated on MR data consisting of 2D axial slices of lumbar vertebrae from a range of subjects (details are given in Section 2.5.2. MRI Dataset 1). The method is shown to obtain consistently high segmentation performance when applied to vertebrae encompassing significant anatomical variation.

The main contributions of the chapter can be summarised as the following:

- A method is proposed for reliable automatic segmentation of axial vertebra slices from MR images. This is seemingly the first approach to deal specifically with this problem, which places demands on the generalisation performance of the segmentation method to account for the highly varied nature of the images.
- A novel way of deriving features is described which encodes the relative location of superpixels based on the output of a contour matching algorithm. This effectively translates global location information into local features which can be used to characterise the individual superpixels.
- It is shown how distance metric learning can be incorporated into the CRF model, offering potential improvements over the standard Euclidean-based measures often used in the literature.
- The method is evaluated extensively on MR images encompassing a diverse range of shape and textural properties and is shown to perform consistently well. It is also shown that the method is directly applicable to CT images without requiring any fundamental change to the model.

The rest of the chapter is organised as follows: Section 3.2 outlines the segmentation model. Sections 3.3–3.4 discuss superpixels, including feature extraction and classification. Section 3.5 describes the final form of the CRF potential functions and discusses probability estimates obtained from graph cuts. Section 3.6 describes the experimental setup and segmentation results. Section 3.7 concludes the chapter.

3.2 Segmentation Model

3.2.1 Markov Random Fields

Markov random fields were first introduced in Section 2.3.5; for completeness, some of the initial material is repeated here. Let $\mathcal{S} = \{1, \ldots, N\}$ be a set of indices into

the image sites (pixels or superpixels). A random field is a set of random variables $X = \{x_i \mid i \in \mathcal{S}\}$ where each of the variables X_i takes a value x_i in its state space. A *configuration* of X is denoted $\mathbf{x} = \{x_1, \ldots, x_N\}$ and corresponds to a particular assignment of states to the variables. For image segmentation, the variables are associated with the image sites and the states correspond to the class labels assigned to those sites. For example, in binary segmentation each pixel or superpixel has an associated label (0 or 1) that signifies background or foreground, respectively. The configuration $\mathbf{x}$ is then a segmentation of the image (i.e. a particular assignment of background or foreground labels). Furthermore, let $\mathcal{X}$ denote the set of all possible configurations:

$$\mathcal{X} = \{\mathbf{x} = \{x_1, \ldots, x_N\} \mid x_i \in \mathcal{L}, i \in \mathcal{S}\} \tag{3.1}$$

where $\mathcal{L}$ is the set of class labels. This corresponds to the set of all possible segmentations (assignments of class labels) for an image.

In a Markov random field (MRF), the state of each variable depends on the state of its neighbours $\mathcal{N} = \{\mathcal{N}_i \mid i \in \mathcal{S}\}$, where $\mathcal{N}_i$ are the indices of the sites neighbouring i. Pixels are usually defined to be neighbours if they are immediately adjacent (in cardinal or diagonal directions), whereas two superpixels are neighbours if they share a common boundary. Formally, a random field X is an MRF if

$$P(\mathbf{x}) > 0 \quad \forall \mathbf{x} \in \mathcal{X} \tag{3.2}$$

and

$$P(x_i \mid x_{j \neq i}) = P(x_i \mid x_{\mathcal{N}_i}). \tag{3.3}$$

The latter condition is known as the *Markov property*. The Hammersley-Clifford theorem shows that $P(\mathbf{x})$ is characterised by a Gibbs distribution:

$$P(\mathbf{x}) = Z^{-1} \exp\left(-E(\mathbf{x})\right) \tag{3.4}$$

where Z is a normalisation constant called the *partition function*. The term $E(\mathbf{x})$ is an energy function, given by a sum over the neighbourhoods (or *cliques*) of the MRF:

$$E(\mathbf{x}) = \sum_c V_c(\mathbf{x}) \tag{3.5}$$

where $V_c(\mathbf{x})$ is a potential function whose value depends on the configuration of the clique c. For the model considered here, a clique can be defined as a subset of sites in which every pair are neighbours.

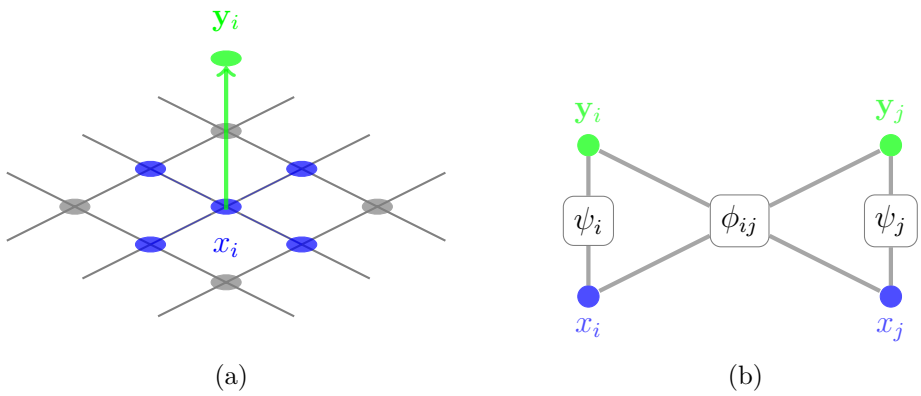


Figure 3.2: (a) A graphical representation of a conditional random field (CRF) on a 2D grid. The class label x_i of node i is dependent on the class labels of its neighbours (shown in blue), given the observation (shown in green). (b) The factor graph representation for two nodes (superpixels) i and j, showing the relationship between the variables and the first- and second-order potentials of the energy function (3.6).

3.2.2 Conditional Random Fields

In a conditional random field (CRF) model, the energy function is generalised to incorporate the observed data (e.g. pixel intensities or superpixel features) into the potential functions [Blake et al., 2011]. The energy function then defines a posterior probability distribution $P(\mathbf{x} \mid \mathbf{y})$ for the variables $\mathbf{x}$ given the observed data $\mathbf{y}$ and can be written as a sum of first- and second-order potential functions in the form

$$E(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{S}} \underbrace{\psi(\mathbf{y}_i \mid x_i)}_{\text{Data term}} + \lambda \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} \underbrace{\phi(\mathbf{y}_i, \mathbf{y}_j \mid x_i, x_j)}_{\text{Smoothness term}} \tag{3.6}$$

where the constant λ controls the relative importance of the data and smoothness terms. The data term of the CRF gives the cost of assigning the label x_i at site i. The smoothness term gives the cost of assigning the labels x_i and x_j at the neighbouring sites i and j and is defined so as to promote spatial consistency of the labels. Note that the form of the CRF is similar to the hidden Markov random field of Zhang et al. [2001], except that the second-order potential also has a dependency on the data. Section 3.5 describes how the potential functions used in the model are learned from data. Figure 3.2 shows a graphical representation of a CRF alongside its factor graph representation [Bishop, 2006] for two nodes in the graph.

The CRF formulation enables maximum *a posteriori* (MAP) inference over the model to be carried out efficiently using graph cuts [Boykov and Funka-Lea, 2006]. Finding the MAP estimate of $P(\mathbf{x} \mid \mathbf{y})$ is equivalent to finding a labelling $\hat{\mathbf{x}}$ that minimises the energy:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}, \mathbf{y}). \tag{3.7}$$

For the problem of binary segmentation, graph cuts can find the globally optimal solution (i.e. the global minimum energy) in polynomial time provided that the energy function is *submodular* [Kolmogorov and Zabih, 2004, Szeliski et al., 2008].[2] The min-cut/max-flow algorithm of Boykov and Kolmogorov [2004] is used to find the optimal solution.

3.3 Superpixels

While most graph-based image representations tend to be constructed from the individual pixels of the image, recent work has demonstrated the advantages of using local aggregates of similar pixels (or *superpixels*) rather than unary pixels as the image primitives [Fulkerson et al., 2009, Lucchi et al., 2010]. The advantages of this approach are twofold: firstly, as the number of nodes in the graph tends to decrease significantly, there is a corresponding reduction in computational complexity. Secondly, multiple features can be extracted from the superpixel regions which can help to discriminate between the classes more effectively.

This section first considers two different methods for generating superpixels to determine which is the most appropriate for the segmentation problem. The two superpixel algorithms discussed have both been demonstrated to obtain very good performance on natural images [Achanta et al., 2012] and are also suitable for application to the MR and CT images considered in this chapter. They are therefore a natural choice for further investigation.

3.3.1 Quick Shift

Quick shift [Vedaldi and Soatto, 2008] is a clustering method which can be used for superpixel generation. It is an example of a *mode seeking* algorithm, which attempts to associate each data point with a mode of the underlying probability density function. Mode seeking algorithms start by computing the *kernel density estimate* [Hastie et al., 2009] of the data:

$$P(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{y}_i - \mathbf{y}) \tag{3.8}$$

where K is a suitable kernel function (e.g. Gaussian) and the features are usually the image intensity values (for greyscale images). Each of the data points is then moved

[2] The energy function is submodular if the pairwise term of (3.6) satisfies $\phi(\mathbf{y}_i, \mathbf{y}_j \mid 0,0) + \phi(\mathbf{y}_i, \mathbf{y}_j \mid 1,1) \leq \phi(\mathbf{y}_i, \mathbf{y}_j \mid 0,1) + \phi(\mathbf{y}_i, \mathbf{y}_j \mid 1,0)$ for every pair [Kolmogorov and Zabih, 2004]. It is easy to verify that this condition holds from the definition of equation (3.23).

towards a mode of the density by following the direction of highest gradient from the current point. The points that converge on the same mode form a cluster. Mode seeking algorithms differ primarily in the scheme they use to evolve the gradient trajectories.

Rather than using a gradient evolution, quick shift moves each data point to the nearest neighbour for which there is an increment of the density function. The algorithm constructs a tree of data points with the branches expressing the distance between those points. Branches that have a greater distance than a threshold value τ are cut, forming the superpixels of the image. In addition to its computational efficiency, one of the main advantages over mode seeking algorithms is its ability to effectively balance under- and over-segmentation of the image through the choice of the τ parameter. Despite its simplicity, the segmentation performance is comparable to the slower mean shift procedure [Vedaldi and Soatto, 2008].

A major limitation of quick shift is the inability to directly control the size of the superpixels, which can lead to some superpixels in an image having a much larger or smaller size compared to the majority. This is undesirable for segmentation as it will affect the statistics of the features computed within the superpixel regions. As the segmentation model is a graph over the superpixels in an image, allowing the superpixels to take on any size also results in some superpixels having a significantly larger number of neighbours than others, which could present problems during inference.

The next section describes a different method which has the ability to constrain the average size of the superpixels in an image, making it more suitable for the segmentation framework described in this chapter.

3.3.2 SLIC

The method presented in this chapter uses the Simple Linear Iterative Clustering (SLIC) [Achanta et al., 2012, Vedaldi and Fulkerson, 2008] algorithm to partition the image into superpixels. The SLIC algorithm generates superpixels by clustering pixels based on their intensity values and spatial proximity in the image. The algorithm starts by initialising the cluster centres on a regularly spaced grid. In order to avoid placing the centres on edges or noisy regions in the image, the 3×3 pixel neighbourhood around each centre is searched to find the position with the lowest gradient. The algorithm then assigns each pixel in the image to the cluster centre with the smallest distance within a local region. After all pixels have been assigned, the updated cluster centres are computed and the procedure is repeated until convergence. As a final step, connectivity of the superpixel regions is enforced

Algorithm 1 SLIC superpixels

Require: Image I

```
Initialise cluster centres c_k on a regular grid with step size S
Move centres to lowest gradient position in 3 × 3 neighbourhood
d_i ← ∞    for all pixels
l_i ← −1    for all pixels
repeat
    for each cluster centre c_k do
        for each pixel i in 2S × 2S region around c_k do
            dist(c_k, i) ← distance between c_k and i
            if dist(c_k, i) < d_i then
                d_i ← dist(c_k, i)
                l_i ← k
            end if
        end for
    end for
    Compute new cluster centres
    Compute residual error r between previous and updated centres
until r ≤ threshold
```

Ensure: Connectivity of superpixels

by detecting any disjoint segments sharing the same label and assigning the smallest segment to its largest neighbouring cluster. Algorithm 1 gives pseudocode for the SLIC superpixel generation process. As shown in Figure 3.1, boundaries of SLIC superpixels have the property of adhering to object boundaries, enabling an accurate pixel-level segmentation to be recovered from the classified superpixels.

The number and regularity of the resulting superpixels is controlled by two parameters S and m, which determine the average size of the regions and their spatial regularity. These are set to $S = 6$ and $m = 0.05$ for all of the experiments described below, resulting in approximately 1700 superpixels per image. These parameters were found by searching for the maximum average superpixel size that obtained an adequate level of overlap with the ground truth segmentations on a set of 63 manually labelled images (see Section 2.5.2. MRI Dataset 1). For each setting of the parameters, the class of each superpixel was first determined by taking the majority vote of the pixel-level ground truth labels within the superpixel. The Dice score was then computed between the pixel-level ground truth labels and the superpixel-level labels. Compared to the superpixels generated by Quick Shift and a number of other competing methods, the advantages of SLIC superpixels are their spatial regularity (they tend to be approximately convex) and the ability to constrain the average size of the superpixels in an image. Experiments have also shown that SLIC superpixels tend to be better at preserving boundary details than competing methods [Achanta et al., 2012].

Note that unlike graph-based image representations on standard pixels, which tend

to use an isotropic neighbourhood system of a constant size, superpixel neighbourhoods vary in connectivity depending on their spatial regularity. In the following sections two superpixels are defined to be neighbours if they share a common boundary (i.e. they contain at least one adjacent pixel). For the parameters and images used here, the superpixels have an average neighbourhood size of 6.

3.4 Characterising Superpixels

The set of pixels comprising a superpixel can characterise the region in many ways, providing a greater range of descriptive features than a single pixel value. Recent work has investigated the use of superpixels within a supervised learning context by training a classifier on a labelled set of features [Fulkerson et al., 2009, Lucchi et al., 2010]. In particular, it is possible to learn estimates of class probabilities $P(\mathbf{y}_i \mid x_i)$, where $\mathbf{y}_i$ is a feature vector for superpixel i. Incorporating these learned probability estimates into a graph cut framework can significantly improve the segmentation performance compared to simpler parametric models (such as a mixture of Gaussians).

The aim is to characterise the superpixels by extracting multiple features from them that incorporate information about intensity, texture, location and edge detection response. The features are used to discriminate between the vertebra and background superpixels by training a classifier on a set of ground truth images. Note that this training occurs only once, after which the trained classifier can be used to provide probability estimates for any further images. The pixel-level ground truth labels are first converted into superpixel-level labels by assigning each superpixel to the class with the majority vote. In the vast majority of cases this vote is unanimous. The superpixel feature/label examples are then used to train a support vector machine (SVM) [Chang and Lin, 2011] using a radial basis function (RBF) kernel, given by

$$K(\mathbf{y}_i, \mathbf{y}_j) = \exp\left(-\gamma ||\mathbf{y}_i - \mathbf{y}_j||_2^2\right) \tag{3.9}$$

where γ is a kernel width parameter found using cross-validation on the training data. Probability estimates for the vertebra and background classes are obtained from the SVM using the method of Wu et al. [2004] and incorporated into the data term of the CRF. To do this the data term of equation (3.6) is defined as the negative log likelihood of an observation (feature vector) given the class label (i.e. vertebra or background):

$$\psi(\mathbf{y}_i \mid x_i) = -\log\left(P(\mathbf{y}_i \mid x_i)\right) \tag{3.10}$$

where the likelihood term $P(\mathbf{y}_i \mid x_i)$ is found by computing the SVM probability estimates for each of the superpixels. The probability estimates are obtained using

an improved version of Platt scaling [Platt, 1999], which approximates the posterior distribution over the classes by fitting a logistic regression model to the raw classifier scores. The improved version (as detailed in Wu et al. [2004]) extends this approach in a number of ways to provide more reliable estimates of the class probabilities. In Section 3.5 the potential functions of the CRF model are described in full.

A number of different features for classification are evaluated, which are summarised in Table 3.1 and discussed below. The data used for the experiments consists of 63 2D axial MR images of the lumbar spine from 21 different subjects, details of which are given in Section 2.5.2. MRI Dataset 1. To train the SVMs, leave-one-out cross-validation was performed by leaving out one subject (i.e. 3 images) on each iteration and training on the remaining 60 images. The model was then tested on the 3 images from the held out subject and the process was repeated for all 21 subjects. Thus the training and test images were always from separate subjects.

To evaluate the performance of the SVM using the different features, Receiver Operating Characteristic (ROC) [Fawcett, 2006] curves are calculated from the probability estimates of the trained SVMs, taken over all leave-one-out iterations. As the ground truth has pixel-level granularity, the superpixel-level probability estimates are first converted to pixel-level before calculating the ROC curves. This was carried out by assigning the probability estimate for each superpixel to the individual pixels comprising the superpixel. The area under the ROC curve (AUC) provides a measure of the overall classification performance of the SVM trained on different sets of features.

The following sections give a detailed description of the features and assess their impact on the classification performance on the SVM. Section 3.6.1 describes the experimental setup in more depth.

3.4.1 Intensity and Texture

Intensity Histograms

The feature vector $\mathbf{y}_i^{T_1}$ provides intensity information in the form of histograms over the superpixel regions. The histogram $\mathbf{h}_i$ for superpixel i is defined to be a 10-bin normalised image intensity histogram over the pixels in i. Specifically, for each superpixel i a concatenation is used of the intensity histogram over the superpixel and an average histogram computed from the neighbouring superpixels in $\mathcal{N}_i$:

$$\mathbf{y}_i^{T_1} = \left[\mathbf{h}_i, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{h}_j \right]^\top . \tag{3.11}$$

Table 3.1: Superpixel features (p_n denotes the n-th percentile). The right column gives the dimension of the superpixel feature vector.

Feature	Description	Dim.
$\mathbf{y}_i^{T_1}$	Concatenation of intensity histogram from superpixel i and average histogram from neighbours $\mathcal{N}_i$.	20
$\mathbf{y}_i^{T_2}$	SIFT descriptor calculated at the centroid of superpixel i.	128
$\mathbf{y}_i^{L_1}$	Mean, p_{10} and p_{90} of the row and column pixel coordinates in the superpixel, centred on the matched contour region.	6
$\mathbf{y}_i^{L_2}$	Mean, p_{10} and p_{90} of the matched contour distance transform gradient in the superpixel, in both the horizontal and vertical directions.	6
$\mathbf{y}_i^{E_1}$	Mean, p_{10} and p_{90} of the LoG response within the superpixel, taken over 4 scales.	12
$\mathbf{y}_i^{E_2}$	Mean, p_{10} and p_{90} of the structure tensor eigenvalues of the superpixel, taken over 4 scales.	24

The motivation for including the average histogram over the neighbours of i is that the range of intensities within an individual superpixel tends to be small, meaning that the histogram is not always distinctive. Including the average histogram of neighbouring superpixels acts to counter this effect and improve classification performance [Fulkerson et al., 2009]. Figure 3.4 shows the ROC curve for the SVM trained on intensity histogram features, which obtained a test AUC of 0.63. Other bin sizes were also tested for the histograms, but it was found in practice that the performance of the SVM was not sensitive to the chosen number of bins.

SIFT Descriptors

The scale-invariant feature transform (SIFT) [Lowe, 2004] is a method for feature description and detection widely used within computer vision for tasks such as object recognition and tracking. Although originally proposed as both a keypoint detector and descriptor, the descriptors can be used independently to characterise the textural properties of regions at arbitrary locations within an image. A brief description of the SIFT descriptor is next provided, before discussing how it can be effectively used to characterise superpixels.

The SIFT descriptor is essentially a weighted spatial histogram of image gradient orientations computed around a specified keypoint (see Figure 3.3 for an illustration). A histogram is first formed from the orientations around the keypoint[3] and each of the histogram samples is then weighted by its corresponding gradient magnitude. The orientations corresponding to the highest peak of the histogram and the local peaks within 80% of the highest peak are used to assign the orientation

[3]The term keypoint here refers to an arbitrary location in the image.

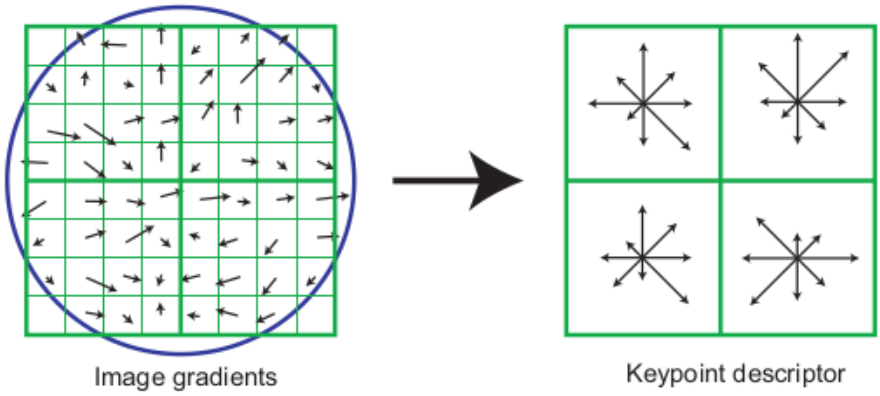


Figure 3.3: Illustration of the SIFT descriptor [Lowe, 2004]. The magnitudes and orientations of local image gradients are weighted by a Gaussian window (blue circle). The samples are aggregated into orientation histograms characterising sub-regions (shown on the right), where the length of the arrows is determined by the sum of gradient magnitudes near their directions.

to the keypoint. To maintain orientation invariance, the coordinates of the descriptor and gradient orientations are rotated relative to the keypoint orientation. Each individual SIFT descriptor is a vector constructed by concatenating the orientation histogram entries. The standard SIFT descriptor computes multiple histograms from a 4×4 sample region with 8 orientation bins each; the resulting descriptor is then a vector with $4 \times 4 \times 8 = 128$ elements.

In order to more effectively characterise the local textural properties of the superpixel regions, 128-dimensional SIFT descriptors of a fixed size are computed at the centroid of each superpixel (denoted by $\mathbf{y}_i^{T_2}$). This results in a set of highly distinctive features that are robust to noise and changes in scale and intensity. The ROC curve for the SIFT features is shown in Figure 3.4. Using the SIFT features, the SVM obtained a test AUC value of 0.96.

3.4.2 Location

By reliably detecting specific regions of the vertebrae, it is possible to obtain information on the location of superpixels relative to these detected regions. In particular, it is noticeable that the upper part of the vertebral body is surrounded by a boundary of cortical bone with a distinctive, roughly semi-circular shape (see Figure 3.5a). Locating the upper part of the vertebra in an axial image makes to possible to correctly classify a large section of the image, as the superpixels above the region are known to belong to the background class. Described next is a contour matching approach that can be used to detect these boundary regions consistently.

A set of partial segmentation contours is first extracted from the ground truth images

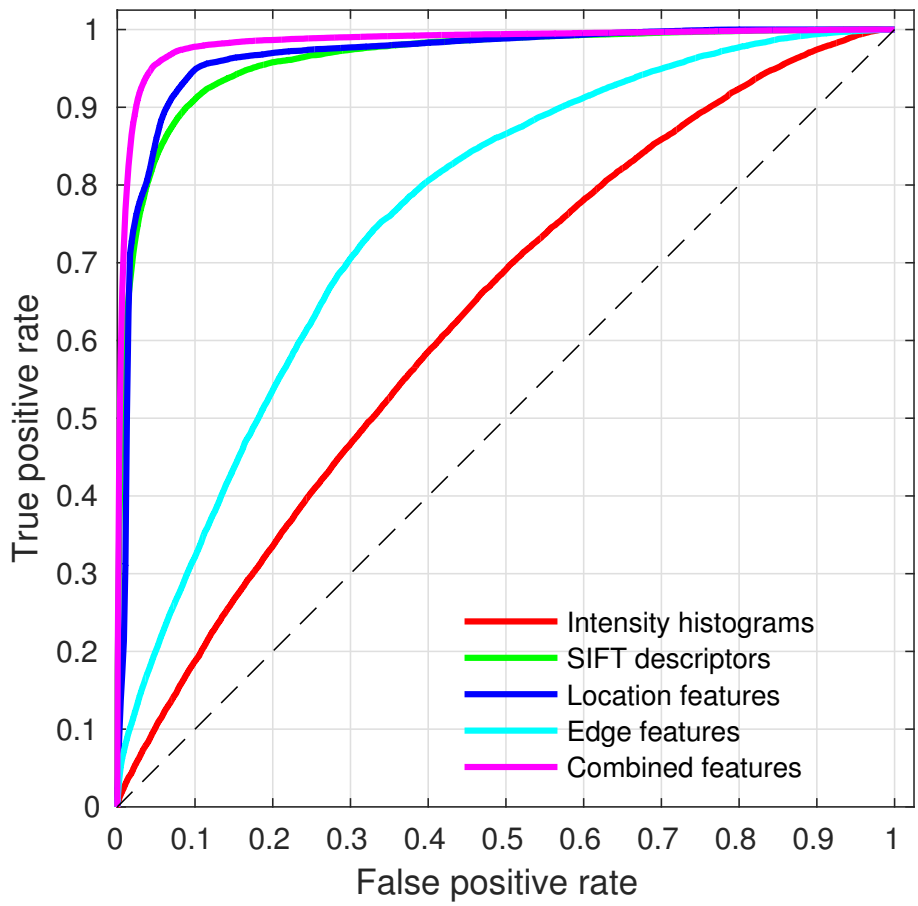


Figure 3.4: ROC curves of SVM probability estimates for the individual and combined superpixel features.

by cropping the original contours below their centroids, so that the resulting contour set $\mathcal{C}$ corresponds to the contours around the upper part of the vertebral body. Each ground truth image is therefore associated with a single contour $C \in \mathcal{C}$ and the goal is to find the best matching contour of the set for a new image. A Laplacian of Gaussian (LoG) filter is used to detect the outer boundary of the vertebra (see the next section for a detailed description of the LoG filter). Given an image $I(r, s)$, the function $L_\sigma(r, s)$ is defined as the convolution of the LoG with the image:

$$L_\sigma(r, s) = \nabla^2 G_\sigma(r, s) * I(r, s) \tag{3.12}$$

where ∇^2 is the Laplacian operator and $G_\sigma(r, s)$ is a Gaussian kernel with standard deviation σ. For each contour $C \in \mathcal{C}$, a search is then performed over the convolved image to find the point where the average LoG response along the contour is greatest. The best match is the contour with the maximum response of the set:

$$\arg\max_{C \in \mathcal{C}} \left[\max_{(r,s) \in \Omega} \left(\frac{1}{|C|} \sum_{(u,v) \in C} L_\sigma(r + u, s + v) \right) \right] \tag{3.13}$$

where Ω denotes the set of all (r, s) coordinate pairs for the image and $(u, v) \in$

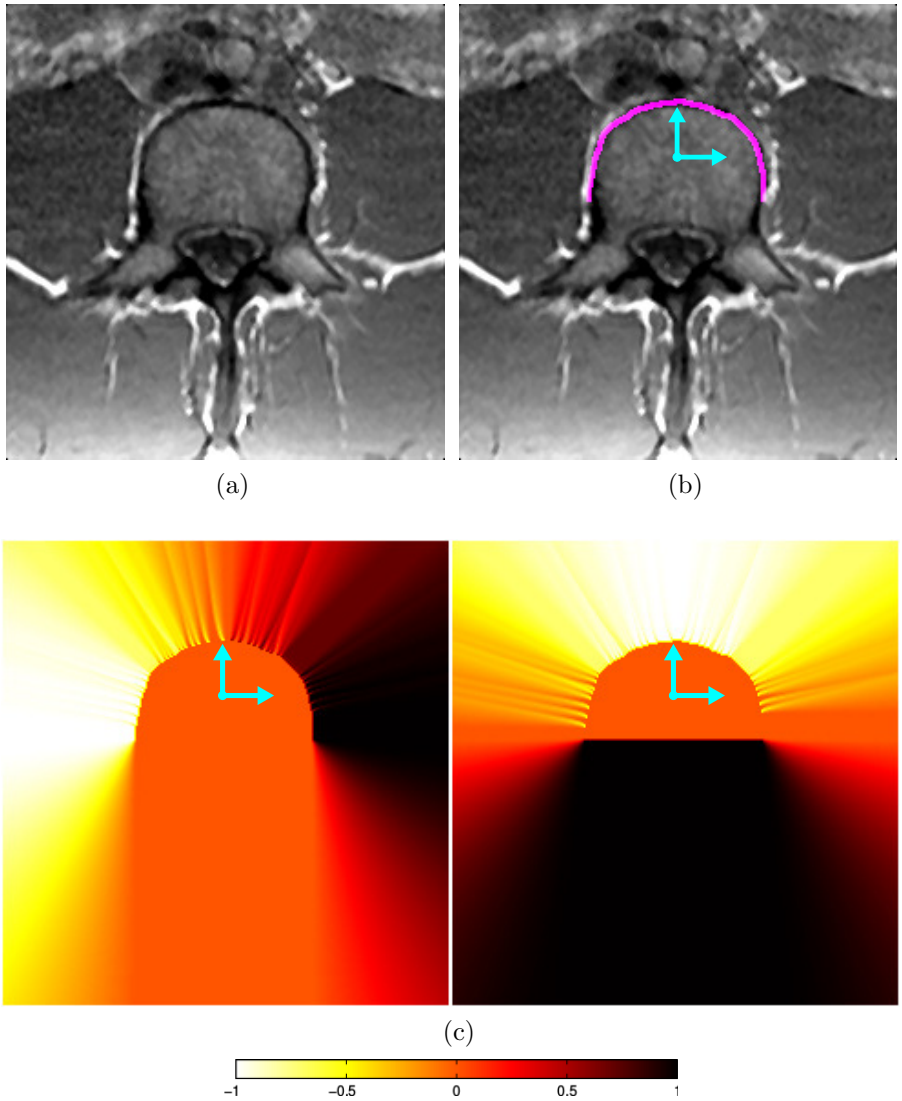


Figure 3.5: (a) Example MRI vertebra slice. The cortical bone can be seen as the dark boundary surrounding the vertebra. (b) Example matched contour (magenta) found using (3.13). Also shown are the axes of the centred coordinates (cyan). (c) Gradient of the contour region distance transform in the horizontal and vertical directions, respectively.

C are the coordinates of the contour. Note that the matching procedure can be implemented very efficiently using convolution operations. An example matched contour is shown in Figure 3.5b.

The matched contours are used to extract location features for the superpixels. For the first set of location features $\mathbf{y}_i^{L_1}$, the pixel coordinates are set to have their origin at the centroid of the matched contour region, as shown in Figure 3.5b. Letting p_n denote the n-th percentile the mean, p_{10} and p_{90} of the row and column pixel coordinates within the superpixel are then taken to form the first 6-dimensional feature vector.

The second set of location features $\mathbf{y}_i^{L_2}$ is obtained from the distance transform

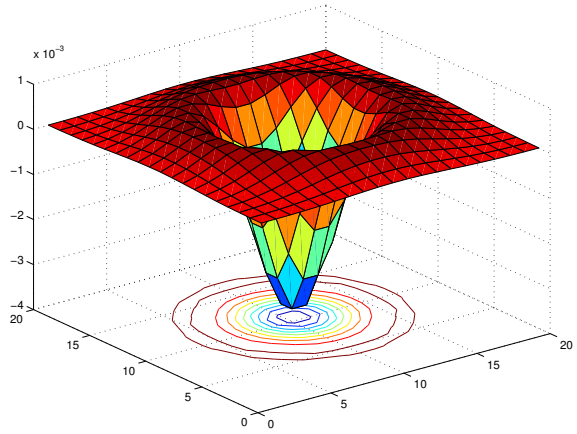


Figure 3.6: Visualisation of a Laplacian of Gaussian filter.

[Maurer et al., 2003] of the matched contour region. The features are formed by taking the gradient of the distance transform in both the horizontal and vertical directions, as illustrated in Figure 3.5c. The feature vector is then formed by taking the mean, p_{10} and p_{90} of these gradients within the superpixel. Note that the gradients of the distance transform have unit norm almost everywhere (i.e. they are unit vectors) [Osher and Sethian, 1988]. However, the horizontal and vertical gradients encode local information defined over the space of *orientations*, the distribution of which can be seen as providing a general representation of object shape [Gurumoorthy et al., 2011]. The combined 12-dimensional feature vector $\mathbf{y}_i^L = [\mathbf{y}_i^{L_1}, \mathbf{y}_i^{L_2}]^\top$ therefore provides important information on the superpixel location relative to the upper part of the vertebra. Figure 3.4 shows the ROC curve for the SVM trained on the combined location features, which obtained an AUC of 0.97.

3.4.3 Edge Response

The feature vector $\mathbf{y}_i^E$ incorporates information on the "edgeness" of the superpixels. These features are distinctive of superpixels at the edges and corners of the vertebrae and help to separate the vertebra and background classes around the boundary.

Laplacian of Gaussian

In order to detect the boundary, convolution of the image with a Laplacian of Gaussian (LoG) filter is carried out to highlight areas corresponding to rapid changes of intensity. Unlike filters based on first-order derivatives of the image, the Laplacian is a second-order operation. As a consequence, it is isotropic (rotation invariant) and obviates the need for multiple filter masks to highlight edges at different angles [Gonzalez and Woods, 2008].

Convolution with the LoG filter can be viewed as applying the Laplacian operator to

an image which has been first smoothed with a Gaussian filter. The representation of image I at scale σ is defined as the convolution of I and a Gaussian kernel with standard deviation σ:

$$I_\sigma(r,s) = G_\sigma(r,s) * I(r,s) \tag{3.14}$$

where

$$G_\sigma(r,s) = \frac{1}{2\pi\sigma^2} e^{-\frac{r^2+s^2}{2\sigma^2}} \tag{3.15}$$

To obtain the LoG, the Laplacian operator is then applied to the resulting image:

$$L_\sigma(r,s) = \frac{\partial^2 I_\sigma(r,s)}{\partial r^2} + \frac{\partial^2 I_\sigma(r,s)}{\partial s^2} \tag{3.16}$$

The value of σ is chosen based on the size of the regions to be detected. The zero crossings of the LoG occur at $r^2 + s^2 = 2\sigma^2$, defining a circle centred on the origin of radius $\sqrt{2}\sigma$. Thus, applying the LoG filter to an image will result in a positive response for low intensity regions of extent $2\sqrt{2}\sigma$ [Gonzalez and Woods, 2008]. Figure 3.6 shows a visualisation of a 20×20 LoG filter with $\sigma = 3$.

The first set of features $\mathbf{y}_i^{E_1}$ is obtained from the LoG response over the superpixel region. The LoG at 4 different scales is calculated by setting the standard deviation to $\{2, 4, 6, 8\}$ pixels. The mean, p_{10} and p_{90} over the superpixel is then taken at each scale to form a 12-dimensional feature vector for the superpixel. Using multiple scales enables the features to capture edge regions of different width and extent, which are generally not known *a priori*.

Structure Tensor

The second set of features $\mathbf{y}_i^{E_2}$ is obtained from the structure tensor of the image [Forstner and Gulch, 1987, Knutsson, 1989]. For an image $I(r,s)$, the structure tensor can be written as

$$J_\rho(\nabla I) = G_\rho * \left(\nabla I \nabla I^\top\right) = \begin{bmatrix} G_\rho * I_r^2 & G_\rho * (I_r I_s) \\ G_\rho * (I_r I_s) & G_\rho * I_s^2 \end{bmatrix} \tag{3.17}$$

where I_r and I_s are the partial spatial derivatives of the image in the horizontal and vertical directions and G_ρ is a Gaussian kernel with standard deviation ρ. The parameter ρ is called the outer scale and determines the extent of spatial averaging, with larger scales acting to reduce noise present at smaller scales. The partial derivatives of the image are computed by convolution with Gaussian derivative filters of standard deviation τ, referred to as the inner scale.

The two eigenvectors and eigenvalues of the structure tensor at each pixel characterise the gradient of the image at a given scale, providing information on the local

structure of the image. Specifically, the eigenvalues describe the average contrast in the corresponding directions given by the eigenvectors [Weickert, 1999]. The eigenvalues of the structure tensor are computed at each pixel within the superpixel region, taken over outer scales $\rho \in \{2, 4, 6, 8\}$ with the inner scale set in proportion $\tau = \rho/2$. The mean, p_{10} and p_{90} of the eigenvalues at each scale are combined to form a 24-dimensional feature vector for the superpixel.

The ROC curve for the edge features is shown in Figure 3.4. An AUC of 0.76 was obtained using the SVM trained on edge features. The relevance of these superpixel features comes from the performance increase when used with other features in combination, which is discussed in the next section.

3.4.4 Performance of Combined Features

Finally, this section considers the performance of the SVM trained on the combined superpixel features. The features are concatenated into a single feature vector $\mathbf{y}_i$ for each superpixel:

$$\mathbf{y}_i = \left[\mathbf{y}_i^T, \mathbf{y}_i^L, \mathbf{y}_i^E\right]^\top . \tag{3.18}$$

Figure 3.4 shows the ROC curve for the SVM trained on the combined features, which obtained an AUC of 0.98. Other subsets of features were exhaustively tested but it was found that the best results were obtained by using a combination of all the features. Comparing this with the ROC curve using only the histogram features shows the importance of incorporating multiple features that are not based solely on intensity information. While the ROC curves of the individual SIFT and location features obtain much higher performance, the improvement provided by the combined features translates into a significant increase in the amount of overlap with the ground truth segmentations.

Figure 3.7b shows example images of the class probabilities $P(\mathbf{y}_i \mid x_i)$ obtained from the SVM using the combined features. Note that all pixels within a given superpixel are assigned the same probability, so the figure shows the superpixel-wise probability estimates. Example images of the SVM probability estimates obtained on CT data are shown in Figure 3.8b. As these probability estimates correspond to the data term of the CRF without the influence of the smoothness term, they are not fully representative of the final segmentations. The next section describes the full CRF model used to obtain the segmentations.

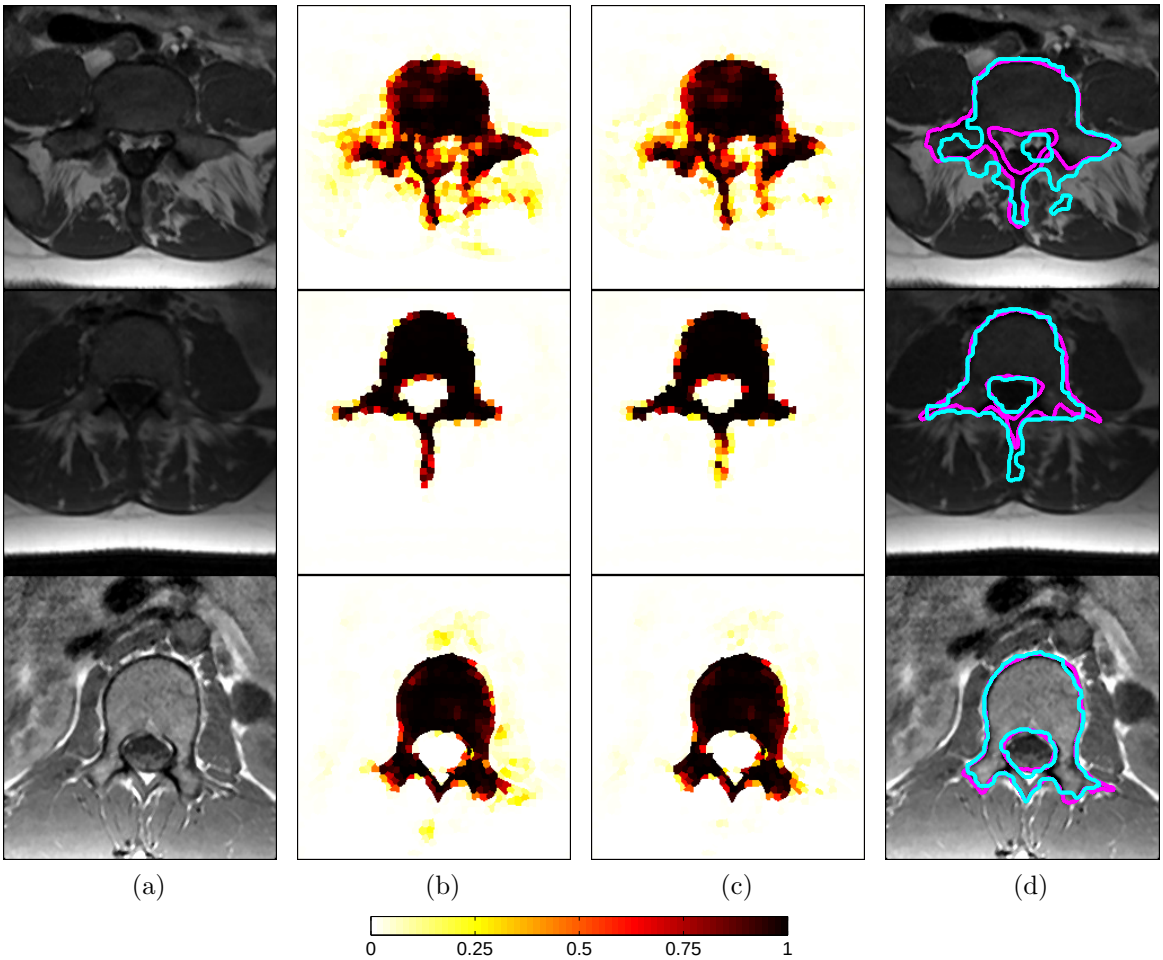


Figure 3.7: (a) Shown top to bottom are MR images for which the CRF obtained the minimum, median and maximum Dice similarity score (0.81, 0.88 and 0.95), respectively. (b) Probability estimates using the combined superpixel features for the images in the first column. Darker regions indicate higher probability of belonging to the vertebra class. (c) Max-marginals computed from the CRF graph cut solutions. (d) Segmentation contours shown for both the ground truth annotations (magenta) and CRF model (cyan).

3.5 Potential Functions and Max-Marginals

This section defines the potential functions used for the CRF of (3.6), each of which incorporates information learned from the superpixel training examples. Also discussed are the probability estimates for the final superpixel labels obtained using graph cuts.

3.5.1 First-Order Potential

The first-order potential function incorporates the SVM probability estimates into the CRF as the negative log likelihood, defined in equation (3.10). The superpixel

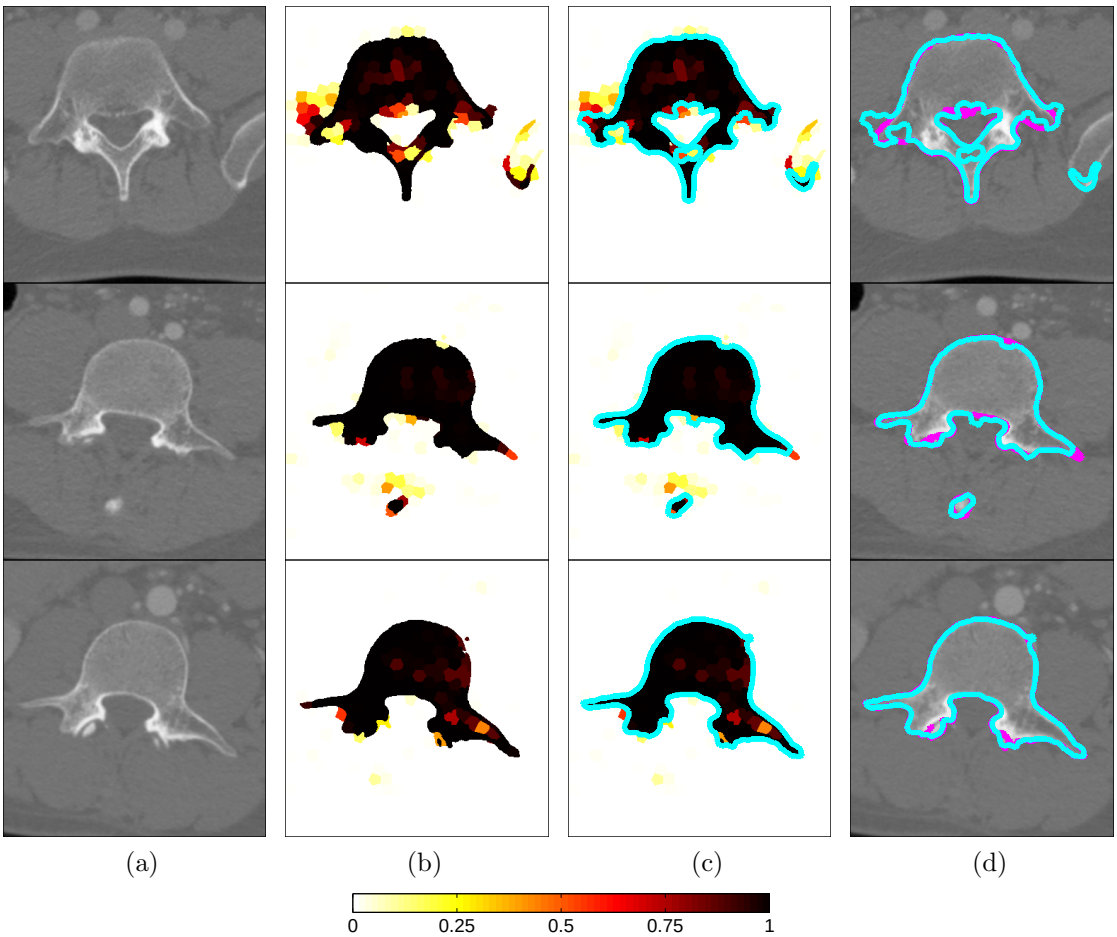


Figure 3.8: (a) Shown top to bottom are CT images corresponding to the minimum, median and maximum Dice similarity score (0.88, 0.97 and 0.98), respectively. (b) SVM probability estimates for the images in the left hand column. Darker regions indicate higher probability of belonging to the vertebra class. (c) Final segmentation contours from the CRF shown overlaid with the probability estimates (cyan). (d) Segmentation contours shown for both the ground truth annotations (magenta) and CRF model (cyan).

likelihoods given by the data term are highly discriminative and localised to the vertebrae regions, as can be seen in the examples shown in Figures 3.7b and 3.8b. While in many cases accurate segmentations can be obtained by simply thresholding the probability estimates of the data term, the smoothness term of the CRF (described in the next section) can improve the accuracy further by promoting spatial consistency.

3.5.2 Second-Order Potential

Many graph cut formulations incorporate a penalty based on a Euclidean distance measure between the features of neighbouring sites, such as the one proposed in Boykov and Funka-Lea [2006]. However, using a standard Euclidean distance disre-

gards any regularities that may be present in the data and which can be exploited to improve performance [Weinberger and Saul, 2009].

In order to address these issues, the focus can instead be placed on using *distance metric learning* to learn an appropriate distance metric for the second-order potential. In particular, the Large Margin Nearest Neighbour (LMNN) [Weinberger and Saul, 2009] algorithm is used to learn a pseudometric[4] of the form

$$D_{\mathbf{M}}(\mathbf{y}_i, \mathbf{y}_j) = (\mathbf{y}_i - \mathbf{y}_j)^\top \mathbf{M}(\mathbf{y}_i - \mathbf{y}_j) \tag{3.19}$$

where $\mathbf{M}$ is a positive semidefinite matrix. This matrix can be expressed in terms of a linear transformation $\mathbf{L}$ such that $\mathbf{L}^\top\mathbf{L} = \mathbf{M}$. The objective is to learn the metric such that the k-nearest neighbours of examples in the transformed space (determined by $\mathbf{L}$) belong to the same class while those belonging to different classes are separated by a large margin. There are a number of parallels between the LMNN algorithm and SVMs as a consequence of their shared focus on margin maximisation, with both methods involving a convex optimisation procedure using similar loss functions [Weinberger and Saul, 2009].

The first term of the LMNN loss function penalises large distances between the input and target neighbours and can be seen as having a "pulling" effect on the target neighbours. It can be written as

$$\ell_{\text{pull}}(\mathbf{L}) = \sum_{j \to i} \|\mathbf{L}(\mathbf{y}_i - \mathbf{y}_j)\|_2^2 \tag{3.20}$$

where $j \to i$ denotes that i is a target neighbour of j. The second term of the loss function penalises small distances between examples with different class labels (termed *impostors*). The term can therefore be seen as exerting a "pushing" force on the examples and can be written as

$$\ell_{\text{push}}(\mathbf{L}) = \sum_{i,j \to i} \sum_{l} (1 - [x_i = x_l]) \max\left\{0, 1 + \|\mathbf{L}(\mathbf{y}_i - \mathbf{y}_j)\|_2^2 - \|\mathbf{L}(\mathbf{y}_i - \mathbf{y}_l)\|_2^2\right\} \tag{3.21}$$

where $[x_i = x_l]$ is 1 if $x_i = x_l$ and 0 otherwise. The combination of the two competing terms of the LMNN loss function is analogous to the formulation of the loss function used for SVM learning [Scholkopf and Smola, 2002]. The optimisation problem can be solved efficiently by formulating the problem as an instance of semidefinite

[4]Unlike a metric, a pseudometric allows two distinct points to have a distance of zero.

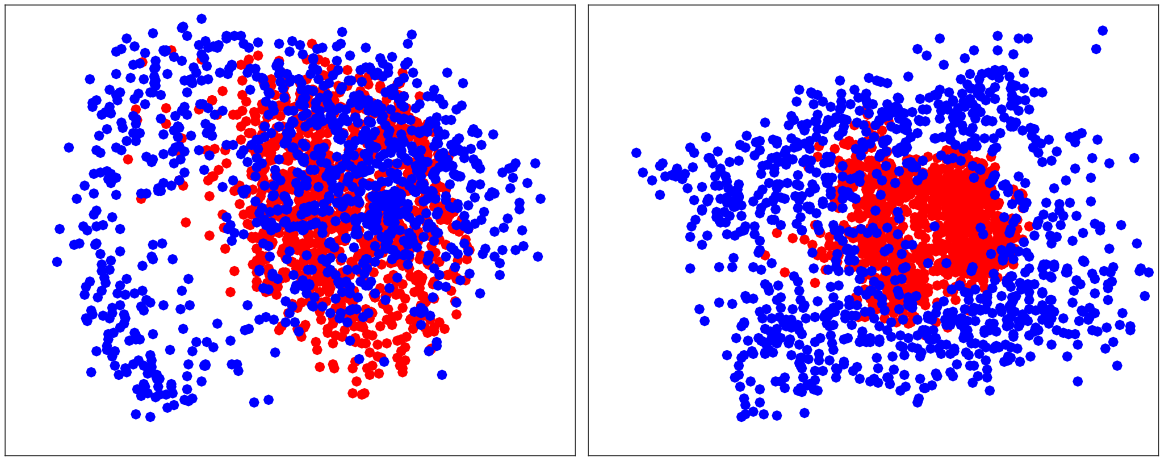


Figure 3.9: The left image shows the projection of a sample of 2000 superpixel features onto their first 2 principal components prior to metric learning (red are positive examples, blue are negative). The right image shows the projected features in the transformed space found by LMNN.

programming [Boyd and Vandenberghe, 2004]. The objective then becomes

$$\begin{aligned} &\min_{\mathbf{M}} \sum_{i,j\to i} D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{subject to } D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l) - D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_{ijl} \\ &\qquad \xi_{ijl} \geq 0 \\ &\qquad \mathbf{M} \succeq 0 \end{aligned} \tag{3.22}$$

where ξ_{ijl} are slack variables for target neighbours i, j and impostors l, which are introduced to control violations of the large margin inequality.

A demonstration of the application of LMNN to the superpixel features is given in Figure 3.9, which shows the projection of a sample of superpixel features onto their first 2 principal components, both in the original and transformed space found by LMNN. In general, the two classes are better separated in the transformed space, with the nearest neighbours of the superpixels tending to belong to the same class. The singular values of a transformation matrix $\mathbf{L}$ learned using LMNN are shown in Figure 3.10. Notice that there is a significant number of singular values near zero, indicating the subset of the features that are collapsed by the transformation. Under this view, the transformation can be interpreted as carrying out a form of feature selection on the original superpixel vectors.

The learned metric is incorporated into the second-order potential function as follows

$$\phi(\mathbf{y}_i, \mathbf{y}_j \mid x_i, x_j) = \begin{cases} \exp\left(-D_{\mathbf{M}}(\mathbf{y}_i, \mathbf{y}_j)\right) & \text{if} \quad x_i \neq x_j \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

which penalises neighbouring superpixels which have similar feature vectors and are

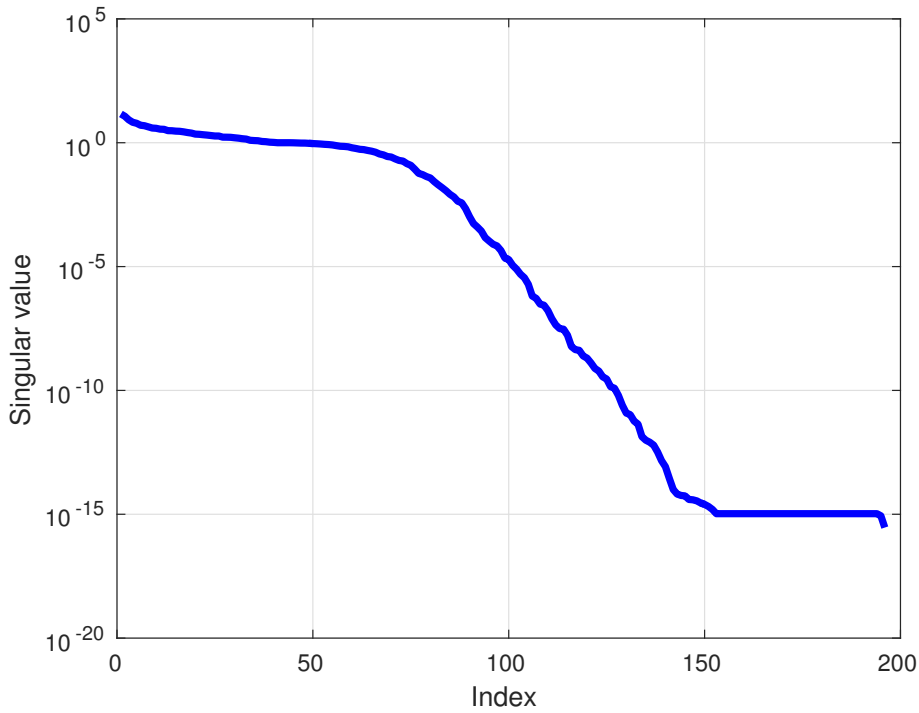


Figure 3.10: Singular values of an example transformation matrix learned by LMNN on the superpixel features.

assigned to different classes. Note that, as with the SVM used in the data term, the metric $\mathbf{M}$ need only be learned once on the training data.

The LMNN algorithm can be related to more traditional statistical methods such as *Fisher's linear discriminant* [Fisher, 1936], which can also be viewed as inducing a metric via a linear transformation of the original feature space. The objective of Fisher's discriminant is essentially to find a linear transformation that maximises the ratio of the inter-class variance to the intra-class variance [Bishop, 2006]. In common with LMNN, Fisher's discriminant tries to maximise the separation of different classes and operates in a supervised setting, but seeks to minimise the variance of all examples within a class rather than a subsample of target neighbours.

3.5.3 Max-Marginals

In order to obtain confidence measures for the CRF segmentations, the *max-marginal* probability estimates $P(x_i \mid \mathbf{y}_i)$ are computed from the graph cut solution. The min-marginal energies are defined as the graph cut solution of the CRF where a single variable x_i is clamped to take the label k:

$$\xi(x_i, k) = \min_{\mathbf{x} \in \mathcal{X}, x_i = k} E(\mathbf{x}, \mathbf{y}). \tag{3.24}$$

The min-marginals can be computed very efficiently using the method of Kohli and Torr [2008]. By taking advantage of *dynamic programming* [Felzenszwalb and

Zabih, 2011], the time required to compute the min-marginals is only slightly more than required for a single graph cut solution. After the min-marginals have been computed, the max-marginals are then obtained using a softmax function over the negative min-marginals:

$$P(x_i = 1 \mid \mathbf{y}_i) = \frac{\exp\left(-\xi(x_i, 1)\right)}{\exp\left(-\xi(x_i, 1)\right) + \exp\left(-\xi(x_i, 0)\right)}. \tag{3.25}$$

Example images of the resulting max-marginal probabilities for the superpixels are shown in Figure 3.7c. Compared with the estimates obtained using just the data term (shown in Figure 3.7b), the CRF probabilities tend to be more localised to the region outlined by the ground truth, which reduces the likelihood of false positives occurring outside the vertebra region.

3.6 Experiments

The performance of the method is next assessed on both MR and CT images from a number of different subjects. A description is first given of the datasets used for the experiments and the training procedure for the CRF model. The segmentation results obtained on each dataset are then presented and discussed.

3.6.1 Image Datasets

MRI Dataset

The MRI dataset used for the experiments consists of 2D axial images of the lumbar spine from 21 healthy subjects with an in-plane resolution of between 0.49 mm and 0.50 mm and a slice thickness of 4 mm. A detailed description of the dataset is given in Section 2.5.2. MRI Dataset 1. It is emphasised that the acquisition protocol used to obtain the images was not tuned for segmentation and is typical of those used for clinical work.

Each of the 512×512 pixel training images were cropped to 251×241 using a minimum bounding box calculated from the set of ground truth segmentations. Prior to processing, the MR images were contrast enhanced by saturating 1% of the intensities (using the MATLAB `imadjust` function). This prior contrast enhancement step was carried out mainly to improve the boundary adherence properties of the SLIC superpixels.

Table 3.2: Segmentation results on the MRI data. The table shows the minimum, median and maximum values over all individual volumes.

Measure	Min	Median	Max
Dice score	0.81	0.88	0.95
Avg. surf. dist. (mm)	0.06	0.27	0.93
RMS surf. dist. (mm)	0.33	1.26	3.41
Max. surf. dist. (mm)	3.50	12.82	29.47

CT Dataset

The CT data consists of central 2D axial slices of lumbar vertebrae from 10 different subjects obtained from the dataset detailed in Section 2.5.2. CT Dataset. The images have an in-plane resolution of between 0.31 mm and 0.36 mm with a slice thickness of 1 mm.

A total of 50 ground truth images were obtained by selecting the middle vertebral slice from each of the 5 lumbar vertebrae of each manually annotated subject. The 512×512 pixel images were cropped to 391×371 using a global bounding box around the vertebrae regions.

3.6.2 Model Training

The SVMs were trained using the procedure described in Section 3.4. Note that the training data is unbalanced, as there are many more negative (background) superpixel examples in an image than positive (foreground) examples. This was addressed by training on a fixed proportion of randomly sampled positive and negative examples at each leave-one-out iteration. An alternative approach to handling unbalanced data is to weight the separate classes in proportion to their frequency in the training data. However, for the data considered here the number of negative examples is prohibitively large, so subsampling is necessary to reduce computation. The same leave-one-out testing approach was used for the LMNN algorithm, with the distance metric learned on the training images for each leave-one-out iteration and applied on the 3 images from the held out subject.

3.6.3 Segmentation Results

The Dice Similarity Coefficient (or Dice score) was used to evaluate the segmentation quality, as defined in Section 2.4.1. Leave-one-out testing was used to evaluate the segmentation performance of the method, with the summary statistics of the Dice score taken over all leave-one-out iterations. A morphological closure opera-

Table 3.3: Segmentation results on the CT data. The table gives the minimum, median and maximum values of the evaluation measures for each lumbar vertebra.

Measure		L1	L2	L3	L4	L5
Dice score	Min	0.92	0.96	0.95	0.94	0.88
	Median	0.97	0.97	0.97	0.97	0.97
	Max	0.98	0.98	0.98	0.98	0.98
Avg. surf. dist. (mm)	Min	0.01	0.01	0.01	0.01	0.01
	Median	0.02	0.03	0.02	0.03	0.02
	Max	0.32	0.05	0.04	0.09	0.58
RMS surf. dist. (mm)	Min	0.08	0.06	0.07	0.07	0.06
	Median	0.15	0.19	0.18	0.20	0.12
	Max	2.43	0.88	0.55	0.62	3.75
Max. surf. dist. (mm)	Min	1.25	0.99	2.28	1.77	1.88
	Median	3.40	3.88	3.50	3.92	2.60
	Max	28.68	22.46	11.76	13.22	34.91

tion was applied as a final post-processing step to smooth the boundaries of the segmentations.

The average Dice score for the MR images was 0.88 with standard deviation 0.03. Table 3.2 summarises the segmentation results using the evaluation measures that were described in Section 2.4.1. Figure 3.7d shows example segmentation contours for both the ground truth and CRF model, corresponding to the minimum, median and maximum Dice score. Although benchmark datasets for axial vertebra segmentation are not publicly available, it is noted here that the worst case Dice score of 0.81, obtained on images which include the pedicles and posterior elements, is still higher than the average score reported in Zukic et al. [2012] for vertebral body segmentation. Most of the disagreement with the ground truth segmentations tends to occur around the posterior elements of the vertebrae, although lower inter-annotator agreement is also expected in these regions due to the low resolution and presence of partial volume effects.

On the CT data the average Dice score was 0.97 with standard deviation 0.01. Table 3.3 summarises the segmentation results obtained on each lumbar vertebra. Figure 3.8d shows example segmentation contours for both the ground truth and CRF model, corresponding to the minimum, median and maximum Dice score (0.88, 0.97 and 0.98). As the figure suggests, in most cases the automatic segmentation is very close to the manually determined region. The results obtained by the method compare favourably with those recently presented in Huang et al. [2013], who reported an average Dice score of 0.94 ± 0.02. In the same work, the authors showed that their method obtained superior results compared with two other recent approaches to vertebra segmentation [Lim et al., 2013, Kim and Kim, 2009].

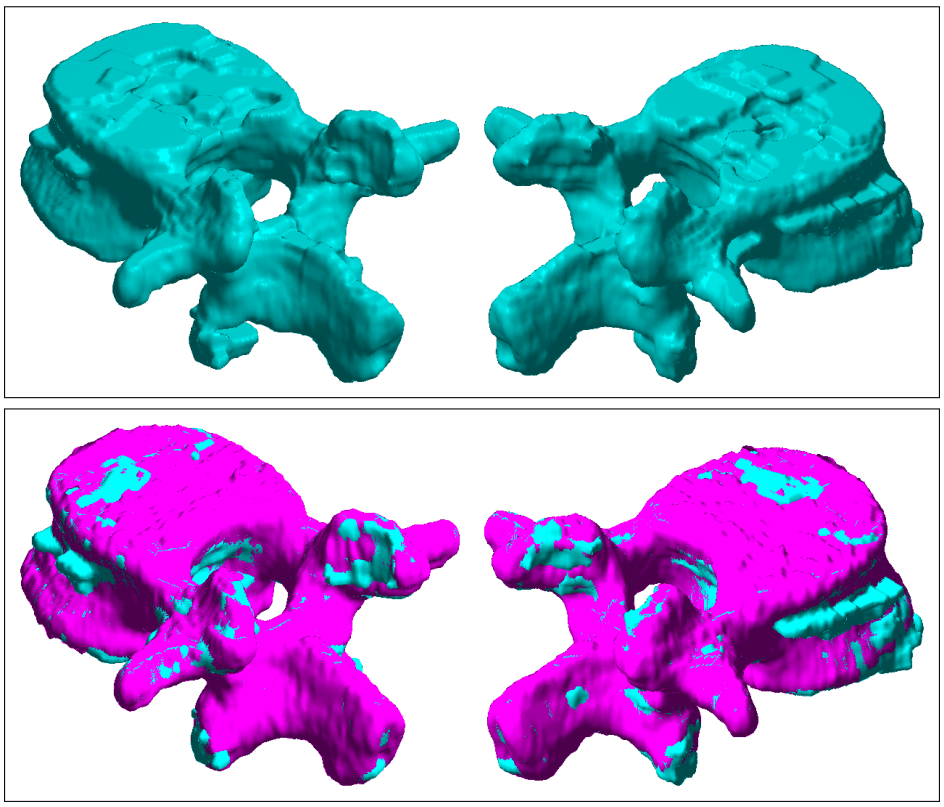


Figure 3.11: The top figure shows a 3D segmentation of a lumbar vertebra (L2) constructed from segmentations of the constituent CT slices. The bottom figure shows the overlap between the CRF segmentation (cyan) and ground truth (magenta).

3.6.4 3D Reconstruction

Although the focus of this chapter is on 2D segmentation, the method can also be used to reconstruct 3D segmentations of vertebrae from individually segmented slices by modifying the way the location features are derived. To do this, the contour matching is first carried out on each slice of the image stack. The M-estimator sample consensus (MSAC) [Torr and Zisserman, 2000] algorithm is then used to remove poor contour matches by detecting and eliminating outliers. Outliers are determined based on the distance to their k-nearest neighbours in the set of matched contours and removed by fitting a line through the set of inliers. Location features analogous to the 2D case can then be derived from the correctly matched contours by computing the distance transform in 3D. This method of obtaining 3D location features is described more completely in the next chapter. Figure 3.11 shows an example 3D vertebra segmentation constructed from segmentations of the constituent slices. Note that while this makes it possible to construct a 3D segmentation from CT data, it is not suitable for MRI due to the much greater difficulty in discriminating between the vertebra and background structures in the individual axial slices.

3.7 Conclusion

This chapter presented an automatic approach for segmentation of vertebrae from MR images. The method avoids the requirement of explicit prior shape information and can therefore deal with a range of normal anatomical variation. An advantage of the method is that it can be applied to images acquired using standard clinical protocols and does not require specialised scanning sequences.

The experimental results show that the method achieves very good segmentation performance on vertebrae from a variety of subjects. The segmentations themselves are highly localised to the vertebra region, with most of the disagreement with the ground truth occurring around the thin posterior elements where lower inter-annotator agreement is expected. The time to segment a single image was approximately 50 seconds using an Intel Core i5 2.50 GHz machine with 8 GB of RAM running Linux (64-bit).

The efficacy of the method relies on the contribution of several methods novel to medical image segmentation. The use of superpixels provides computational efficiency and allows us to more fully characterise an image region in terms of texture than is possible with single pixels. Although similar in spirit to the hidden Markov random field approach introduced by Zhang et al. [2001], the use of conditional random fields allows the similarity between the hidden (super)pixel states to be conditioned on the observed data. Furthermore the use of metric learning tailors the similarity measure between superpixel features to the data itself, rather than relying on a metric chosen *ad hoc*; this promotes spatial consistency and reduces the number of false positives. In contrast to methods that use a parametric form for assessing likelihood of an observed intensity conditioned on the (super)pixel state (e.g. a Gaussian or mixture of Gaussians), an important contribution to the model's accuracy is made by the learning of this likelihood with SVMs. Finally, the combination of descriptive superpixel features, namely the SIFT, location and edge features significantly enhances the segmentation accuracy.

Although this work has focused on the segmentation of vertebrae from MR images, the approach that has been described is applicable to general segmentation problems. MR images have low resolution, low contrast and are noisy compared with CT images and the experiments carried out on CT data demonstrate that very accurate segmentations can be achieved without modification of the underlying algorithm. It is reasonable to assume that the learning of both first- and second-order potential functions will be effective for general medical image segmentation problems given a large enough dataset of labelled images. The formulation of the segmentation problem as one of assigning class labels to local regions of the image also enables a natural extension to 3D segmentation, which is described in the next chapter.

4 3D Segmentation of the Lumbar Spine

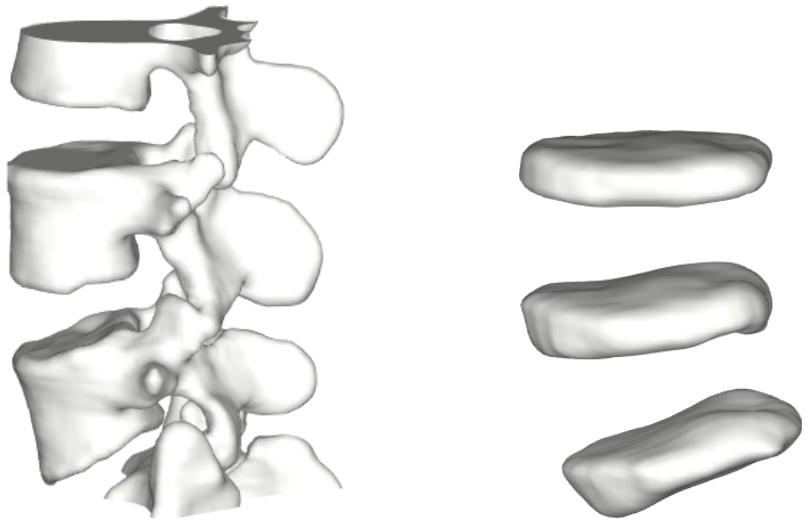
The previous chapter demonstrated that accurate segmentation is possible in 2D using descriptive image features in combination with a superpixel-based conditional random field model. This chapter extends the approach described in the previous chapter for 2D segmentation to enable segmentation of the lumbar spine from 3D MRI data.[1] The method is shown to obtain accurate 3D segmentations of both lumbar vertebrae and intervertebral discs. Additional results obtained on CT data are also presented along with a report of the MICCAI-CSI 2015 challenge on automatic IVD localisation and segmentation from 3D T2 MRI data [Zheng et al., 2016].

In order to extend the approach to 3D, one of the main issues is in finding descriptive features that are able to capture the properties of local voxel regions (as opposed to pixel regions). This chapter demonstrates that very effective features can be learned from data, thus avoiding the need to hand-design features analogous to those used for 2D segmentation.

4.1 Introduction

Segmentation of the lumbar spine remains a difficult task for automated methods due to the complexity of the anatomy (see Figure 4.1) and various imperfections that are often present in the images, such as noise and intensity inhomogeneity. Recent work on vertebra segmentation in 3D has tended to focus on incorporating prior shape information into the model. This is achieved using methods such as statistical shape models [Kadoury et al., 2013, Mizaalian et al., 2013, Kirschner et al., 2011] or shape-constrained deformable surfaces [Lim et al., 2013, Ma and Lu, 2013, Klinder et al., 2009]. One drawback of shape-based approaches is their reliance on prior knowledge of anatomical variation to constrain the segmentation. Construction of statistical shape models requires the accurate placement of anatomical landmarks across a representative training dataset, a step which is often carried out manually by an expert. As discussed in Section 2.3.2, deformable models also rely on accurate

[1] This chapter is based on work first published in Hutt et al. [2015b] and Zheng et al. [2016].



(a) (b)

Figure 4.1: Annotated sections of the lumbar spine showing vertebrae (a) and intervertebral discs (b). Note that the two figures have different scale.

initialisation to ensure the optimisation process does not become trapped at a local minimum, which can result in a poor segmentation. A few alternative methods have recently been introduced which avoid explicit shape constraints [Huang et al., 2013, Kim and Kim, 2009]. However, these methods are tailored specifically to the properties of CT images and are not likely to be effective when applied to lower-contrast MRI data. Furthermore, existing methods that are designed for automatic segmentation of 2D images are likely to become either intractable or ineffective in 3D due to the demands of operating on much larger and more complex images.

Unlike the proposed approach, the segmentation of intervertebral discs (IVDs) has typically been treated as an independent problem in the literature, employing specialised methods differing from those used for vertebra segmentation. Existing methods for IVD segmentation are predominantly focused on 2D images, recent examples being the atlas-based approach of Michopoulou et al. [2009] and the level set active contour approach of Law et al. [2013]. A 3D approach was proposed by Kelm et al. [2013] using marginal space learning to estimate the location of the IVDs, followed by a graph-based segmentation. The recent work of Chen et al. [2015] obtained state-of-the-art results in 3D using a two-step localisation and segmentation approach.

In this chapter, a novel approach is described for segmentation of the lumbar spine from 3D MR images based on a conditional random field (CRF) operating on supervoxels. Basing the model on supervoxels reduces computational complexity and enables more descriptive features to be extracted to characterise the separate classes. Features are obtained by learning dictionaries of filters from multi-scale volume pyramids. By learning the features over multiple scales, larger-scale spatial structure

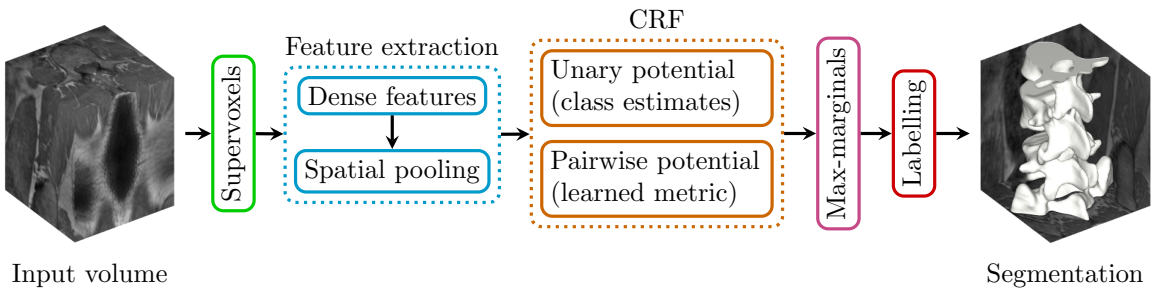


Figure 4.2: Overview of the proposed segmentation method. Features are extracted densely over the input volume at multiple scales and pooled within supervoxels. Estimates of the supervoxel class labels are obtained using an SVM with a generalised RBF kernel. A CRF model incorporating the SVM predictions and a learned pairwise metric is used for the final labelling of the supervoxels.

can be represented while ensuring the learning process remains tractable. Additional features are described that encode the relative location of supervoxels using an extension of the contour matching procedure introduced in the previous chapter for 2D segmentation. Closely following the approach used in the previous chapter, supervised learning is used to train a support vector machine (SVM) on labelled supervoxel features and obtain probability estimates expressing the likelihood of belonging to either the object or background class. Distance metric learning is used to find an appropriate dissimilarity measure between supervoxel pairs. The probability estimates and learned metric are incorporated into a CRF model in the form of first- and second-order clique potentials of the CRF energy function. This formulation enables minimisation of the energy function to be carried out efficiently using graph cuts [Boykov and Funka-Lea, 2006]. Figure 4.2 illustrates the main components of the proposed segmentation method.

The method is evaluated extensively on 3D MRI datasets of lumbar vertebrae and intervertebral discs. An initial report of the IVD segmentation results obtained by the method was given in Hutt et al. [2015b].

The main contributions of the chapter can be summarised as the following:

- A method is introduced for automatic segmentation of the lumbar spine from 3D MR images. It is shown that the method is able to obtain accurate segmentations from scans of lumbar vertebrae and intervertebral discs.

- A multi-scale dictionary learning approach is used to obtain descriptive features for the supervoxels, which are subsequently used to train a classifier for estimating the supervoxel class labels. Additional features are described which encode the relative location of supervoxels.

- A CRF model is introduced with learned potential functions incorporating

the classifier label estimates in addition to a learned metric between pairs of supervoxel features.

- The method is evaluated on 3D MRI datasets of vertebrae and intervertebral discs from a number of different subjects.

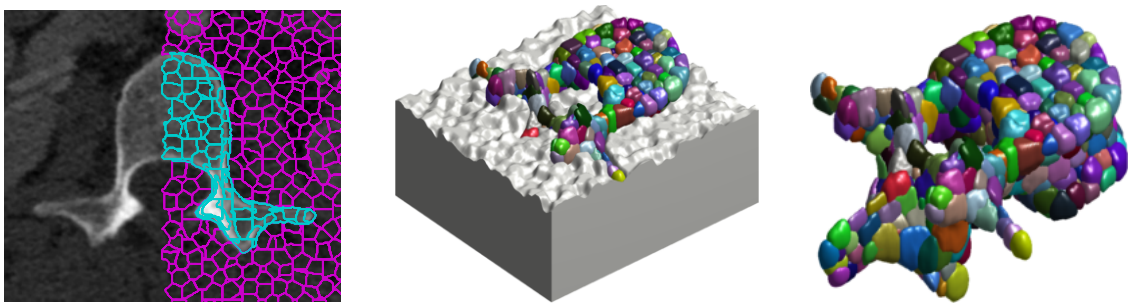
The first MRI dataset consists of 3D scans of a section of the lumbar spine from 8 different subjects, encompassing the lumbar vertebrae from L3 to L5. The details of the dataset are given in Section 2.5.2. MRI Dataset 2. The dataset is referred to in this chapter by the abbreviation LV.

The second dataset used for the evaluation consists of T2-weighted turbo-spin-echo MR images from 15 different subjects provided for the MICCAI 2015 intervertebral disc localisation and segmentation challenge. A detailed description of the dataset is given in Section 2.5.2. MRI Dataset 3. The dataset is referred to in this chapter by the abbreviation IVD.

In the first part of the chapter supervoxel classification is discussed as follows: Section 4.2 discusses supervoxels. Section 4.3 describes the learned set of supervoxel features. Section 4.4 describes the location features. Section 4.5 discusses supervoxel classification and evaluation. The second part of the chapter describes the segmentation model and presents the final results as follows: Sections 4.6 and 4.7 discuss the final form of the CRF model with learned potential functions. Section 4.8 presents additional experiments along with the final segmentation results. Section 4.9 investigates the influence of sample size on the resulting performance of the method and Section 4.10 concludes the chapter.

4.2 Supervoxels

The 3D segmentation problem is formulated as one of assigning class labels to *supervoxels* (groups of similar voxels). The advantages of using supervoxels are twofold: firstly, multiple features can be extracted from the supervoxel regions to discriminate between the separate classes more effectively [Lucchi et al., 2012]. Secondly, for graph-based models such as CRFs the number of nodes in the graph decreases dramatically from the millions of individual voxels in a volume to a much smaller set of supervoxels. This leads to a corresponding reduction in computational complexity, enabling segmentation of a volume to be carried out very efficiently.



(a) (b) (c)

Figure 4.3: Figures show supervoxels belonging to an example vertebra, where the class of each supervoxel is determined by majority vote. (a) Boundaries of supervoxels in an axial slice of a CT volume. Note that the image shows a single 2D slice through the 3D supervoxels. (b, c) Surfaces of supervoxels where the vertebra supervoxels have been assigned random colours and background supervoxels are shown in grey. The supervoxels preserve the boundary detail of vertebrae, enabling an accurate voxel-level segmentation to be recovered.

4.2.1 Axis-Weighted SLIC

The Simple Linear Iterative Clustering (SLIC) [Achanta et al., 2012] algorithm is used to partition a volume into supervoxels. As discussed previously in Section 3.3.2, the advantages of SLIC supervoxels compared to those obtained by similar algorithms are their spatial regularity (they tend to be approximately convex) and the ability to constrain the average size of the supervoxels in a volume. The number and regularity of the resulting supervoxels is controlled by two parameters, which determine the average size of the regions and their spatial regularity, respectively. The parameters are determined empirically by searching for the maximum supervoxel size that still preserves almost all object boundaries in the training images. The effect of the supervoxel parameters is discussed in more detail in Section 4.8. As shown in Figure 4.3, boundaries of SLIC supervoxels have the property of adhering to object boundaries, enabling an accurate voxel-level segmentation to be recovered from the classified supervoxels.

The SLIC algorithm generates supervoxels by clustering voxels based on their intensity values and spatial proximity in the volume. One issue that needs to be accounted for is that the volumes may have an in-plane resolution which is different to the slice thickness (i.e. the voxel grid may be anisotropic). This is achieved by adapting the SLIC algorithm to assign the appropriate weight to distances computed along the depth dimension, where the weight is given by the ratio of the in-plane image resolution to the slice thickness. This modification to the algorithm leads to supervoxels with approximately equal *physical* extent in all directions. More formally, letting (r, s, t) denote the spatial coordinates of a voxel and τ denote the ratio

of the in-plane resolution to the slice thickness, the spatial distance is defined as

$$\text{spatial distance} = \sqrt{(r - r')^2 + (s - s')^2 + \tau^2(t - t')^2} \tag{4.1}$$

where r', s' and t' are the coordinates of a cluster centre. In addition, the SLIC algorithm begins with an initialisation step where an initial set of cluster centres are sampled on a regularly spaced grid. In order to account for anisotropic volumes, a further adaptation of the algorithm is made so that the initial grid of cluster centres is scaled based on the resolution. For example, this means that a volume with a slice thickness of 1 mm results in twice the number of initial centres along the depth dimension of the grid compared with a volume of equal dimensions having a slice thickness of 2 mm.

Note that unlike graph-based image representations on standard pixels or voxels, which tend to use a fixed neighbourhood system of a constant size, supervoxel neighbourhoods vary in connectivity depending on their spatial regularity. For the rest of the chapter two supervoxels are defined to be neighbours if they share a common boundary (i.e. they contain at least one adjacent voxel).

The aim is to characterise the supervoxels by extracting descriptive features from them which can then be used to learn a model from training data to estimate the class label (i.e. object or background). In the next few sections the supervoxel features are described in detail.

4.3 Learned Supervoxel Features

Many techniques for obtaining descriptive local image features have been introduced in the computer vision literature (such as SIFT [Lowe, 2004] and HOG [Dalal and Triggs, 2005]) and have proven to be effective on 2D natural images. Densely extracted SIFT descriptors have also been used in combination with other features to segment vertebrae in 2D image slices (see the previous chapter and Hutt et al. [2015a]). Generalising these low-level descriptors to 3D is however non-trivial and greatly increases the computational cost of feature extraction. In addition, designing comparable features by hand for imaging modalities such as MRI is a very difficult task requiring that accurate assumptions are made about the data based on prior knowledge.

Recent work has focused on addressing the limitations associated with standard descriptors such as SIFT by instead attempting to learn features from data, often in an unsupervised fashion [Bengio et al., 2013]. Approaches based on learned features have proven to be successful in a wide range of computer vision tasks, in many

cases obtaining superior performance when used in place of standard descriptors. Motivated by this, an approach is next described for learning descriptive multi-scale features from data which can be used to characterise supervoxels.

4.3.1 Multi-Scale Dictionary Learning

In order to learn features an unsupervised approach is used based on encoding randomly sampled image patches into a bank of linear filters, also known as a *dictionary* [Elad, 2010]. First, a set of N patches of fixed dimension (in this case $5 \times 5 \times 5$ voxels) are sampled from the training images and reshaped into vectors $\{\mathbf{v}_i\}_{i=1}^N$ of dimensionality $d = 125$. The patches are standardised to zero mean and unit standard deviation and then whitened (decorrelated) using the ZCA transform [Bell and Sejnowski, 1997].[2] After these pre-processing steps, *sparse coding* [Olshausen and Field, 1996] is applied to obtain a dictionary of k filters $\mathbf{D} \in \mathbb{R}^{d \times k}$. To do this, optimisation is carried out according to the following

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{s}} \sum_{i=1}^{N} \|\mathbf{D}\mathbf{s}_i - \mathbf{v}_i\|_2^2 + \beta \|\mathbf{s}_i\|_1 \\ &\text{subject to } \|\mathbf{D}_{:j}\|_2 = 1 \quad \forall j \end{aligned} \tag{4.2}$$

where $\mathbf{s}$ are the "code vectors" and $\mathbf{D}_{:j}$ denotes the j-th column of $\mathbf{D}$. The regularisation parameter β controls the sparsity of the solution. The resulting optimisation problem is both convex in $\mathbf{s}$ (with $\mathbf{D}$ held fixed) and convex in $\mathbf{D}$ (with $\mathbf{s}$ held fixed) and can be solved by alternating the optimisation of both sub-problems using the fast *feature-sign search* algorithm and Lagrange dual proposed in Lee et al. [2006].

An alternative method for learning dictionaries of filters is *spherical k-means*, as described in Coates and Ng [2012]. Specifically, spherical k-means replaces the optimisation of (4.2) with the following

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{s}} \sum_{i=1}^{N} \|\mathbf{D}\mathbf{s}_i - \mathbf{v}_i\|_2^2 \\ &\text{subject to } \|\mathbf{s}_i\|_0 \leq 1 \quad \forall i \\ &\qquad\qquad \|\mathbf{D}_{:j}\|_2 = 1 \quad \forall j \end{aligned} \tag{4.3}$$

where the code vectors $\mathbf{s}$ now simply indicate which k-means cluster (column of $\mathbf{D}$) each patch belongs to. An advantage of this approach is that learning the dictionaries is much more efficient, although this comes at a small cost in terms of representational power. A comparison of the performance of both dictionary

[2] If $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ is the eigendecomposition of the data covariance matrix $\mathbf{\Sigma}$, the whitened patches are given by $\mathbf{V}(\mathbf{\Lambda} + \epsilon\mathbf{I})^{-1/2}\mathbf{V}^\top\mathbf{v}$, where ϵ is a small constant that is set to 0.1.

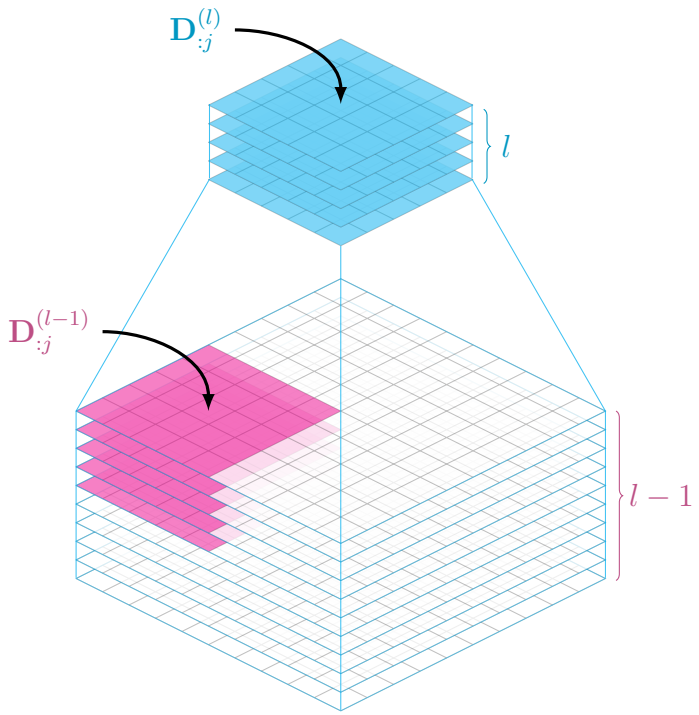


Figure 4.4: Representation of filters at two successive levels of a pyramid. A learned $5 \times 5 \times 5$ filter $\mathbf{D}_{:j}^{(l)}$ at level l of the pyramid (cyan) captures a greater spatial area than a filter $\mathbf{D}_{:j}^{(l-1)}$ with equal dimensions at level $l-1$ (magenta).

learning algorithms is given in Section 4.5. Unless stated otherwise, it is assumed in this chapter that the dictionaries have been learned using sparse coding.

Although the features learned using sparse coding are effective at representing local image structure, the small and fixed size of the sampled patches severely limits the amount of spatial context that can be captured when the dictionaries are learned over a single scale. In order to obtain features which can be used for accurate and robust classification, larger-scale spatial context needs to be taken into account. In theory this could be achieved by simply expanding the size of the sampled patches, but in practice this quickly becomes intractable as the number of samples required to learn effective sparse projections of the data tends to increase dramatically with larger patch sizes [Elad, 2010, Coates and Ng, 2012].

To address this issue, an alternative approach is taken based on learning features over multiple scales using volume pyramids. A Gaussian pyramid is first constructed by successive smoothing and downsampling along each axis by a factor of 2. A separate set of features is then learned over each level of the pyramid using sparse coding, where the patch size is held constant over all levels (see Figure 4.4 for an illustration). This results in a set of dictionaries $\{\mathbf{D}^{(l)}\}_{l=1}^{M}$ corresponding to M different scales which are able to capture larger-scale 3D structure in the volume, but are also efficient to learn due to the small size of the sampled patches. Figure 4.5

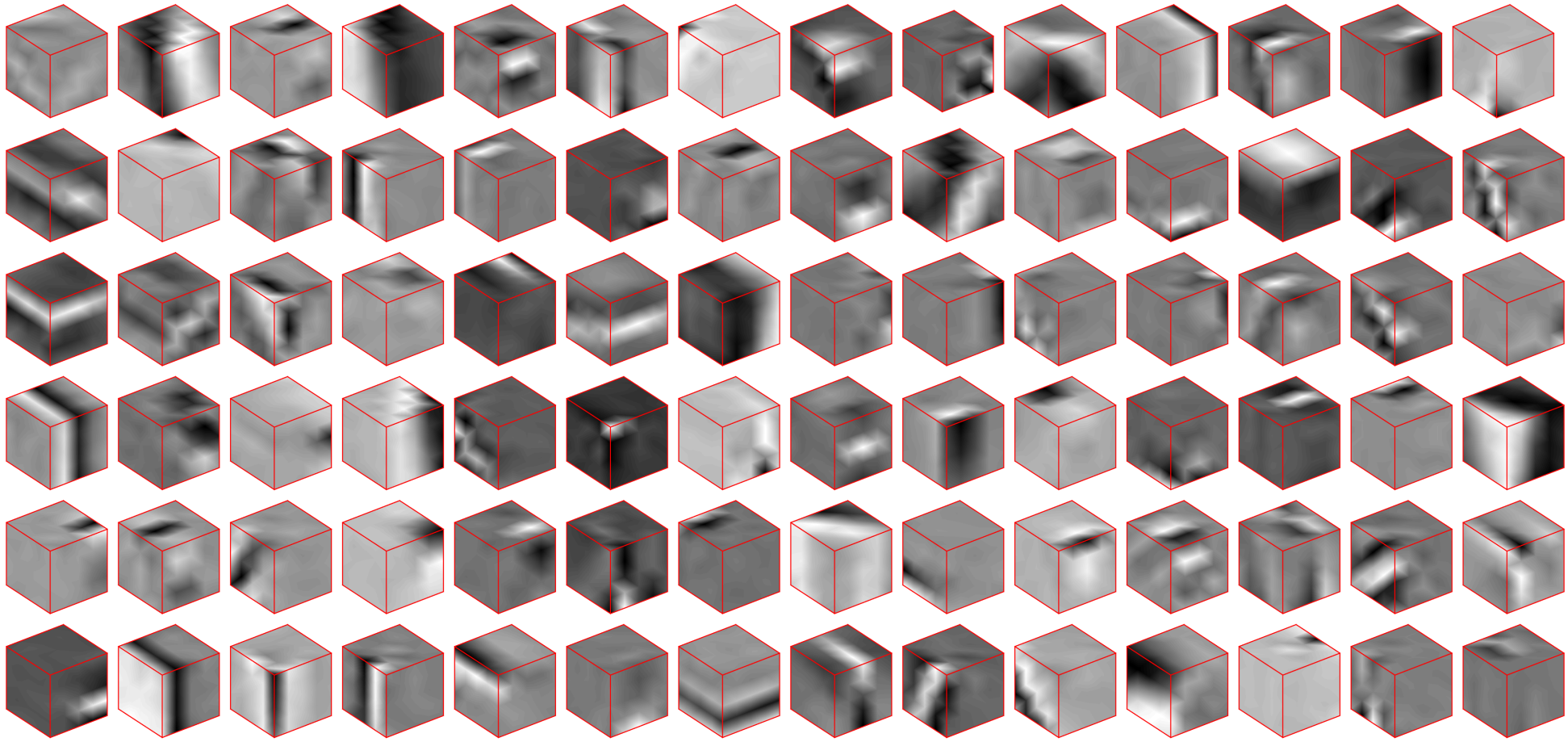


Figure 4.5: Examples of $5 \times 5 \times 5$ features learned using sparse coding on 3D MRI data. The features shown correspond to a dictionary $\mathbf{D}^{(1)}$ learned over the first pyramid level.

shows examples of first-level learned features, corresponding to the dictionary $\mathbf{D}^{(1)}$. Consistent with dictionaries learned from 2D images, the learned filters are sparse and composed mainly of oriented edges in 3D.

For the experiments described below, a separate dictionary of 128 filters is learned from 100 000 randomly sampled patches at each level of the training pyramids. A patch size of $5 \times 5 \times 5$ is used and the dictionaries are learned over 3-level pyramids. More information on these parameters is provided in Section 4.5, where a comparison is given of the classification performance under different settings.

4.3.2 Supervoxel Feature Encoding and Pooling

Given a new image patch $\mathbf{v}_r^{(l)}$ at level l of the pyramid and the learned dictionary $\mathbf{D}^{(l)}$, an encoding function is then used to map the input patch to a vector of features. A nonlinear function of the filter responses is used that results in a sparse feature vector for the patch:

$$\mathbf{u}_r^{(l)} = \max\left\{0, \left[-\mathbf{D}^{(l)}, \mathbf{D}^{(l)}\right]^\top \mathbf{v}_r^{(l)}\right\} \tag{4.4}$$

where $\left[-\mathbf{D}^{(l)}, \mathbf{D}^{(l)}\right]$ is a $d \times 2k$ matrix formed by column-wise concatenation. This effectively splits the positive and negative components of the filter responses into separate features. Splitting the features in this way addresses the ambiguity associated with the sign of the filter responses (i.e. it is not known *a-priori* if the positive or negative components are more important). The split encoding also maintains sparsity of the vector, which is not the case if for example the absolute value of the responses is used. The encoding is very fast, enabling features to be computed over the entire pyramid.

To extract features for the supervoxels, patches are sampled densely by stepping over the volume with a step-size of 2 voxels and encoding filter responses using (4.4) over all levels of the pyramid.[3] The features from all M levels are concatenated into a single vector at each location:

$$\mathbf{u}_r = \left[\mathbf{u}_r^{(1)}, \ldots, \mathbf{u}_r^{(M)}\right]^\top. \tag{4.5}$$

As more than one patch is sampled from each supervoxel, the local features extracted from a supervoxel region are then *pooled* to aggregate them into a single, fixed-size feature vector. A commonly used method for pooling features is to take

[3]Note that the coordinates of a sampled location are reduced in proportion to the reduction of the volume due to downsampling. If (r, s, t) are the coordinates of a patch centre at the first level of the pyramid, the corresponding coordinates at level l of the pyramid are given by $\left(r/2^{l-1}, s/2^{l-1}, t/2^{l-1}\right)$.

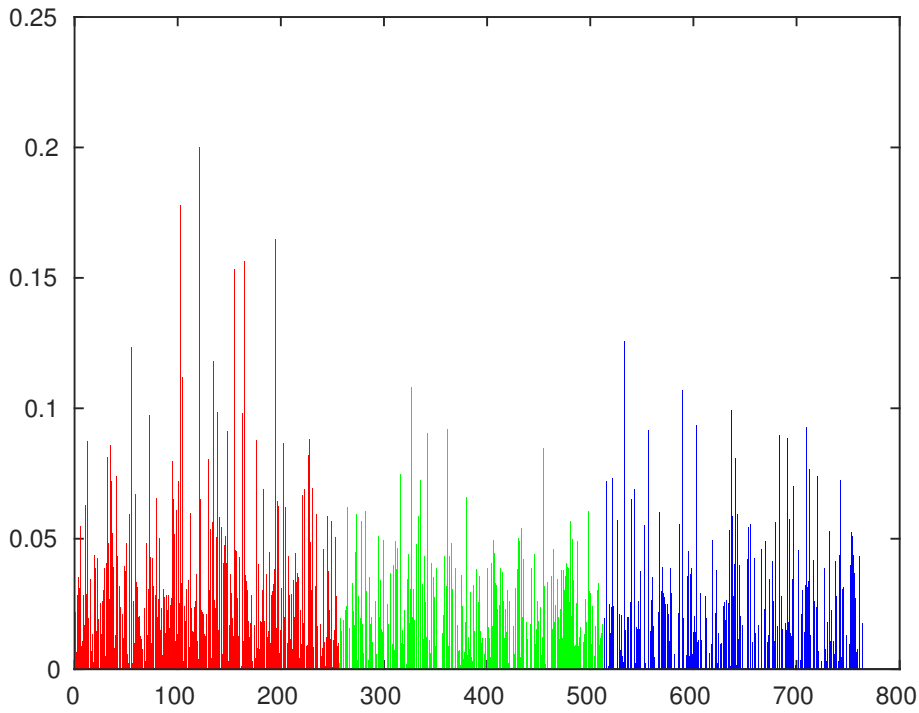


Figure 4.6: Example supervoxel feature vector. The encoded responses at each of the three pyramid levels are shown in different colours.

the average of each feature within a specified region. However, it has been demonstrated that *max* pooling consistently outperforms average pooling in numerous vision tasks [Boureau et al., 2010]. Initial experiments also confirmed that max pooling outperforms average pooling for supervoxel classification, details of which are given in Section 4.5. Max pooling is therefore used to aggregate the features within the supervoxels. Letting $\mathcal{R}_i$ be the locations of the subset of features (4.5) within the supervoxel region i, the pooled representation is formed by taking the maximum of each feature over all locations:

$$\mathbf{y}_{ij}^S = \max_{r \in \mathcal{R}_i} \mathbf{u}_{rj} \quad \forall j \tag{4.6}$$

where $\mathbf{y}_{ij}^S$ is the j-th element of feature vector $\mathbf{y}_i^S$ (the superscript denotes the feature type). The final pooled features obtained in this way are robust to local spatial variation within the supervoxel regions. Figure 4.6 shows an example pooled supervoxel feature vector.

4.4 Location Features

In addition to the learned set of sparse features described in the previous section, features are next introduced for the supervoxels which encode their relative location in the volume. The location features are similar to those introduced in the previous chapter and are based on computing the distance transform from a set of elliptical

contours matched to the anterior part of the vertebral bodies.

4.4.1 Contour Matching

The contour matching procedure is adapted from the algorithm introduced in Section 3.4.2 to enable reliable and efficient matching over an entire volume. The main difference with respect to the original algorithm is that the template contours are obtained by generating a set of partial ellipses, rather than cropping the contours from a set of manual segmentations. This increases the robustness of the matching algorithm as it no longer relies on the manually segmented training images to provide a suitable set of template contours. The resulting algorithm shares similarities with the generalised Hough transform introduced by Ballard [1981], but is more efficient and does not use a hard edge detection prior to template matching.

A set of template contours is first obtained by generating ellipses of differing scales and eccentricities, where the range of scales used to generate the ellipses is based on the resolution of the data. The ellipses are then cropped below the major axis, so that the resulting set of contours $\mathcal{C}$ can be used as templates for matching to the boundary regions. The goal is then to find the best matching contour of the set for each axial slice in a volume. Following the procedure that was previously discussed in Section 3.4.2, a Laplacian of Gaussian (LoG) filter is used to detect the outer boundary of the vertebra. Given a 2D axial slice $I(r, s)$ of the volume, the function $L_\sigma(r, s)$ is defined as the convolution of the LoG with the image:

$$L_\sigma(r, s) = \nabla^2 G_\sigma(r, s) * I(r, s) \tag{4.7}$$

where ∇^2 is the Laplacian operator and $G_\sigma(r, s)$ is a Gaussian kernel with standard deviation σ. For each contour $C \in \mathcal{C}$, a search is then performed over the convolved image to find the point where the average LoG response along the contour is greatest. The best match is the contour with the maximum response of the set:

$$\arg\max_{C \in \mathcal{C}} \left[\max_{(r,s) \in \Omega} \left(\frac{1}{|C|} \sum_{(u,v) \in C} L_\sigma(r + u, s + v) \right) \right] \tag{4.8}$$

where Ω denotes the set of all (r, s) coordinate pairs for the image and $(u, v) \in C$ are the coordinates of the contour.

Note that the matching procedure can be implemented very efficiently using convolution operations, enabling a relatively large set of contours to be used. For the results given in this chapter, ellipses of 10 different scales and eccentricities were generated resulting in a total of 100 template contours. An example matched contour

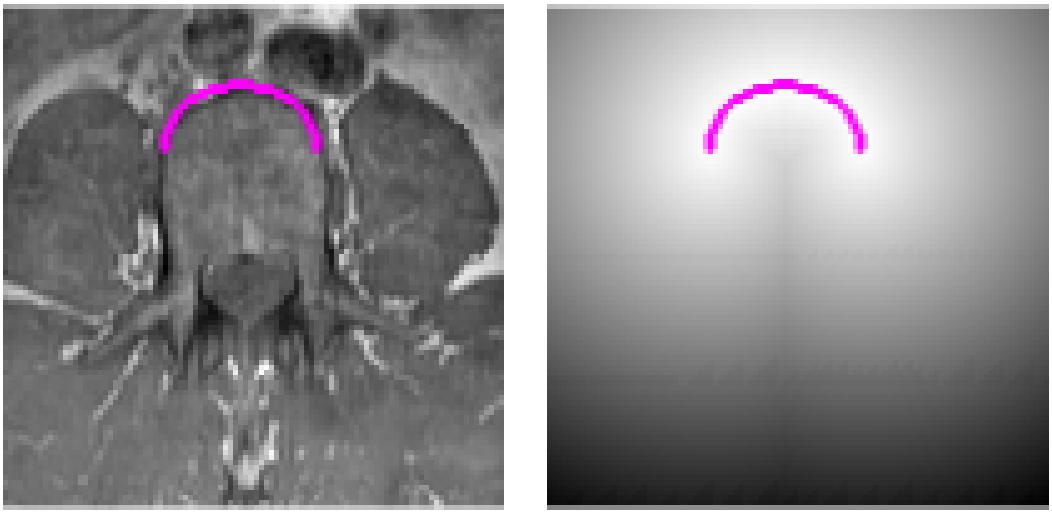


Figure 4.7: (Left) Matched contour (magenta) for an example axial vertebra slice. (Right) The contour distance transform (darker regions are further from the contour).

is shown in Figure 4.7.

Outliers can result from the matching procedure described above due to the matching taking place at every slice of the volume. The outliers occur predominantly at the locations of the intervertebral discs, due to lower edge response in these areas. The following section describes how to remove the outliers from the set of matched contours.

4.4.2 Outlier Removal

Let $\mathbf{c}_t$ be the 3D vector of coordinates of the centroid of the matched contour at axial slice t of a volume. A matched contour is identified to be an outlier if the maximum distance to the centroids of its k-nearest neighbours is greater than a specified threshold. More formally, the maximum distance among the neighbours of centroid $\mathbf{c}_t$ is written

$$d_t = \max_{k \in \mathcal{K}_t} \|\mathbf{c}_t - \mathbf{c}_k\|_2^2 \tag{4.9}$$

where $\mathcal{K}_t$ is the set of indices for the k-nearest neighbours (by distance) of $\mathbf{c}_t$. Intuitively, the t-th matched contour is an outlier if $d_t > \delta$ for some threshold δ estimated from the data. Based on this observation, an algorithm is next described for removing the outliers from the set of matches.

The m-estimator sample consensus (MSAC) [Torr and Zisserman, 2000] algorithm is an improved modification of the well known random sample consensus (RANSAC) [Fischler and Bolles, 1981] method for fitting a model in the presence of outliers. MSAC is used to model the inliers in the set of centroid distances $\mathbf{d} = \{d_1, \ldots, d_T\}$ corresponding to each of the T axial slices of the volume as a linear function of the

axial slice coordinate, t. That is, the inlier distances are modelled as $\hat{d}_t = mt + b$ and $\mathbf{w}$ denotes the parameters (m, b).

To do this, the following overall cost is sought to be minimised

$$R(\mathbf{w}) = \sum_{t=1}^{T} \rho(d_t, \hat{d}_t(\mathbf{w})) \tag{4.10}$$

where the loss function ρ is defined as

$$\rho(d_t, \hat{d}_t) = \begin{cases} (d_t - \hat{d}_t)^2 & \text{if} \quad (d_t - \hat{d}_t)^2 < \delta^2 \\ \delta^2 & \text{otherwise.} \end{cases} \tag{4.11}$$

The best fit $\mathbf{w}$ is found by random search. At each iteration of the algorithm, 2 samples are selected at random from $\mathbf{d}$ and used to determine a candidate line through the data. The line is scored according to the cost assigned by (4.10) and the candidate with the smallest score is retained. This procedure is repeated until convergence of the score, after which outliers are deemed to be those $\mathbf{c}_t$ for which $|d_t - \hat{d}_t(\mathbf{w})| > \delta$. The threshold δ determines the cost assigned to outliers, which is set to $\delta = 0.1\sigma$ where σ is the standard deviation of the centroid distances $\mathbf{d}$. The number of nearest neighbours used to compute $\mathbf{d}$ is set to $k = 12$. Note that this outlier removal algorithm ensures that the matching procedure is robust, as the only required assumption is that some majority of contours are accurately matched to the vertebrae.

4.4.3 Feature Extraction

Given the set of matched contours, the location features for the supervoxels are obtained from the Euclidean distance transform [Maurer et al., 2003] of the union of the contours (see Figure 4.7). The features for a supervoxel take the form of the mean and the 10-th and 90-th percentile values of the distance transform within the supervoxel. Also computed are the mean and 10-th and 90-th percentile values of the vertical gradient of the distance transform and the absolute horizontal gradient, which encode local orientation information (see Section 3.4.2). The values are concatenated to form a 9-dimensional vector $\mathbf{y}_i^L$ for the supervoxel. These features provide detailed information on the location of the supervoxels relative to the anterior part of the spine, which helps to isolate the vertebrae from other structures in the volume when combined with the learned sparse features.

4.5 Supervoxel Classification

The features are used to discriminate between the object and background supervoxels by training a classifier on a set of manually annotated volumes. The features are first standardised by subtracting the mean and dividing by the standard deviation of each feature in the training set. Each instance is then normalised so that $\|\mathbf{y}_i^S\|_2 = \|\mathbf{y}_i^L\|_2 = 1$. For the remainder of this chapter superscripts are not used to denote different feature types, as this will be clear from the context.

The classifier training procedure closely follows the approach described in the previous chapter (Section 3.4) for superpixel classification. To obtain training examples, the original voxel-level class labels are first converted into supervoxel-level labels by assigning each supervoxel to the class with the majority vote. In almost all cases this vote is unanimous. To estimate the class labels for the supervoxels a support vector machine (SVM) [Chang and Lin, 2011] is trained on the supervoxel feature/label examples. A generalised RBF kernel is used, given by

$$K(\mathbf{y}_i, \mathbf{y}_j) = \exp\left(-\gamma(\mathbf{y}_i - \mathbf{y}_j)^\top \mathbf{M}(\mathbf{y}_i - \mathbf{y}_j)\right) \tag{4.12}$$

where γ is an overall kernel width parameter found using cross-validation on the training data. The positive semidefinite matrix $\mathbf{M}$ defines a pseudometric between supervoxel features, which is learned from data using the Large Margin Nearest Neighbour (LMNN) algorithm [Weinberger and Saul, 2009]. Metric learning using LMNN was discussed in detail in Section 3.5.2. Note that setting the matrix $\mathbf{M}$ to the identity $\mathbf{I}$ results in the standard Euclidean RBF kernel that was previously discussed in Section 3.4 in the context of superpixel classification. Probability estimates for the object and background classes are obtained from the SVM using the method of Wu et al. [2004]. The probability is denoted by $P(\mathbf{y}_i \mid x_i)$ where $x_i \in \{0, 1\}$ is the class label for the supervoxel i.

4.5.1 Mining for Hard Examples

The data for training the classifiers is unbalanced, as there are many more negative (background) supervoxel examples in a volume than positive (foreground) examples. Furthermore, the sheer number of examples means that training a single SVM simultaneously on all the data is prohibitively expensive. To address this, a retraining approach is instead used that works by iteratively mining the data for hard examples and learning a new model on the updated set of examples.

The model is first learned on a small, randomly sampled subset of $N = 400$ examples from the training data and hard examples generated by the model are added to a

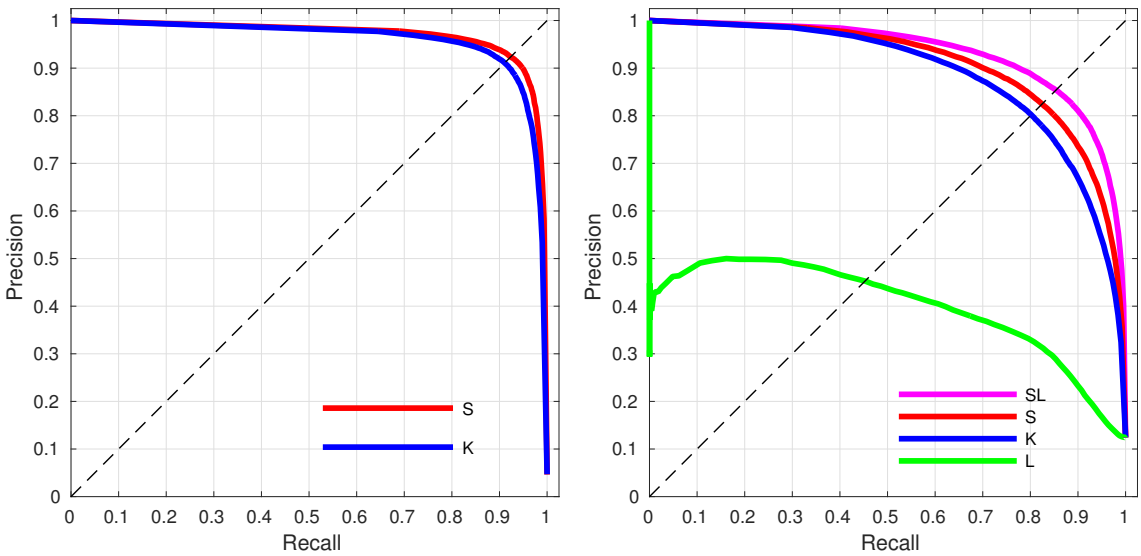


Figure 4.8: (Left) Precision-recall curves for the IVD dataset using k-means (K) and sparse coding (S). (Right) Precision-recall curves for the LV dataset using location features (L), k-means (K), sparse coding (S) and sparse coding combined with location features (SL).

cache. The hard examples are defined for a learned model as those assigned the highest probability $P(\mathbf{y}_i \mid x_i)$ of belonging to the opposite of the true class (i.e. the most confusing for the classifier). The process is then iterated by learning a new model using the updated cache and adding new hard examples found at the current iteration. The mining process is repeated until a specified cache size limit is exceeded, which is set to 20 000.

4.5.2 Evaluation

To evaluate the classification performance under different settings, leave-one-out cross-validation was used to estimate the generalisation performance. By leaving out one subject on each iteration and training on the remaining subjects, maximal use is made of the data for learning, while at the same time ensuring that the training and test data are always from separate subjects. The performance measures given in the following sections are taken over all leave-one-out iterations. Prior to processing, each volume was standardised by subtracting from each voxel the mean intensity of the volume and dividing by the standard deviation.

To measure the SVM classification performance, precision-recall curves are calculated from the supervoxel probability estimates of the trained SVMs, taken over all leave-one-out iterations. The F-score is also calculated to measure the overall classification performance, defined as:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.13}$$

Table 4.1: F-scores averaged over all leave-one-out iterations using different features for classification (see text for details). The number of pyramid levels used for dictionary learning is M.

Features	IVD dataset	LV dataset
Location	N/A	0.49
k-means ($M = 3$)	0.91	0.80
k-means + Location	N/A	0.84
Sparse coding ($M = 1$)	0.79	0.54
Sparse coding ($M = 2$)	0.90	0.75
Sparse coding ($M = 3$)	**0.92**	0.83
Sparse coding + Location	N/A	**0.86**

For the data considered here the F-score is a more appropriate measure of classification performance than accuracy, which can be misleading given that the classes are imbalanced (i.e. there are many more negative than positive examples in the data).

Comparison of Sparse Coding and Spherical k-Means

To justify the use of sparse coding for dictionary learning, the classification performance is next investigated in comparison with spherical k-means. Note that the execution time after learning is essentially identical using both methods, as the resulting dictionaries are the same size and the same encoding function (4.4) is used. Figure 4.8 shows the precision-recall curves for the SVMs trained on the IVD and LV datasets and Table 4.1 gives the F-scores.

The results show a consistent decrease in performance when using spherical k-means in place of sparse coding. A Wilcoxon signed-rank test for pairwise comparisons also showed a significant difference in the classification performance between the two dictionary learning methods ($p < 0.001$ for both datasets). Apart from the faster learning procedure, there appears to be no compelling advantages to using spherical k-means over sparse coding for learning the dictionaries.

Note that although it is also possible to evaluate other low-level features for supervoxel classification, such as 3D Gabor filters or Haar-like features, the focus in this work is on learning features from data in an unsupervised fashion. As mentioned previously, one of the main advantages of using dictionary learning is the avoidance of having to explicitly design and tune the features and parameters by hand.

Effect of Multiple Scales

The use of multi-scale dictionaries is next justified by comparing the classification performance on both datasets as the number of pyramid levels is varied. Table 4.1

shows the F-scores as the number of pyramid levels increases from $M = 1$ (no pyramid) to $M = 3$.

The results show a dramatic improvement in performance when using pyramids, demonstrating the importance of multiple scales for accurate classification. For all other results reported in this chapter, the number of pyramid levels for dictionary learning was set to 3 for both datasets. The number of levels can be approximated based on the resolution of the data, as a higher resolution is more likely to capture structure over a wider range of scales. Using more than 3 levels is not practical for the datasets considered here as the resolution becomes too coarse after downsampling to provide any useful information.

Effect of Location Features

For the LV dataset the location features provide important information on relative location when combined with the learned sparse features. Figure 4.8 shows the precision-recall curve for the location features on the LV dataset and Table 4.1 gives the F-score. Note the sharp decrease during the initial section of the precision-recall curve due to an increase in false positives. As precision does not necessarily decrease when recall increases, this can cause the precision-recall curve to fluctuate. Also shown are the results from combining (by concatenation) the location features with those learned using k-means and sparse coding. Combining the two feature types improves performance in both cases, with the best performance on the LV dataset obtained using location features in combination with sparse coding.

4.5.3 Discussion

Example SVM probability estimates for the best performing features are visualised in Figure 4.9. Note that all voxels within a given supervoxel are assigned the same probability, so the figure shows the supervoxel-wise probability estimates. While the outputs from the classifier provide generally good predictions of the supervoxel class labels, each supervoxel is only considered in isolation (i.e. there is no spatial context from surrounding supervoxels). In particular this means that isolated errors can occur due to similarity between the object and background supervoxels. In the following section the segmentation model is described, which incorporates spatial context from surrounding supervoxels enabling a more accurate prediction of the class labels.

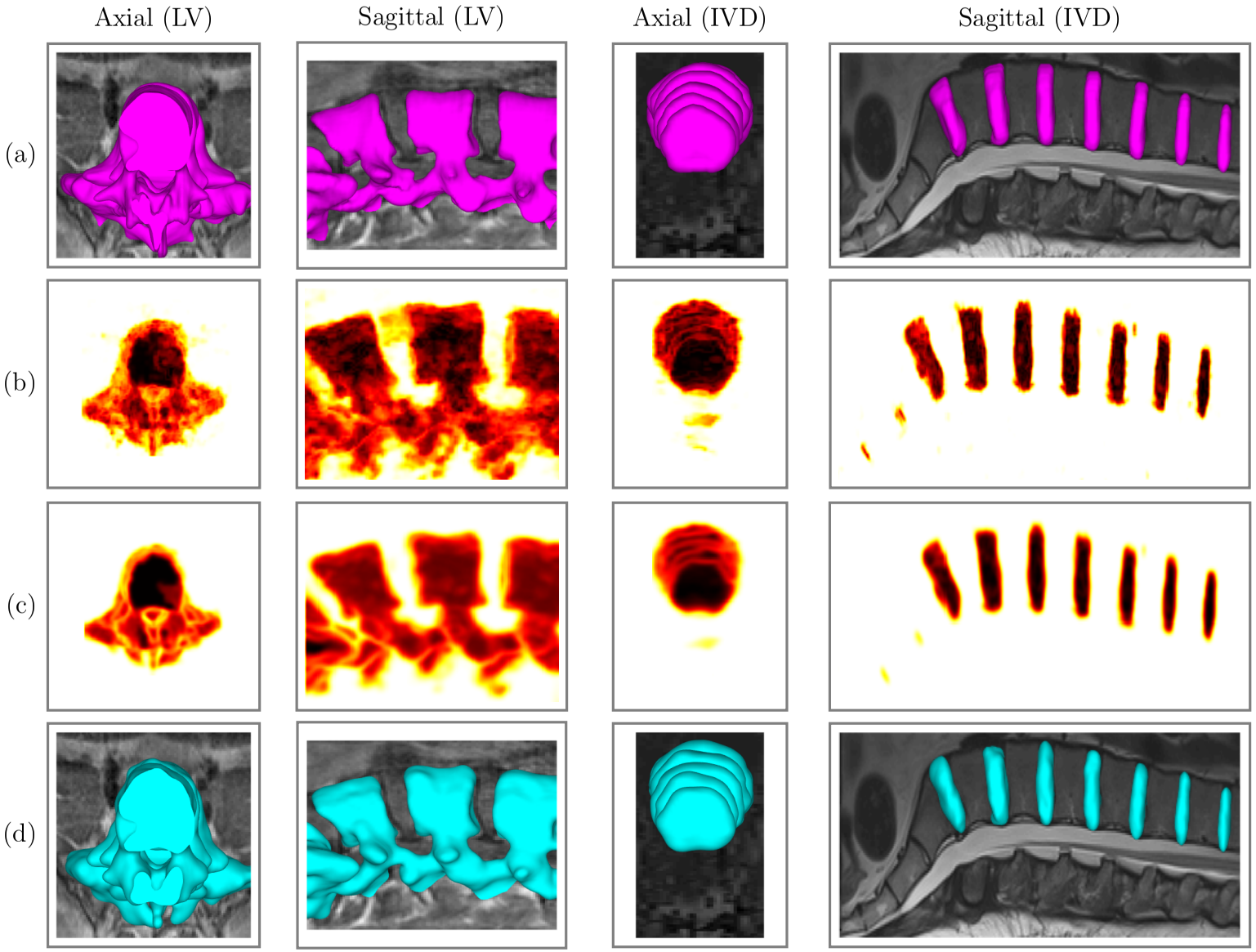


Figure 4.9: Rendered 3D views of example volumes from the LV and IVD datasets. The rows show the manual annotations (a), SVM probability estimates (b), smoothed CRF max-marginals (c) and the final thresholded segmentations (d). Darker regions in rows (b) and (c) indicate higher probability of belonging to the object class.

4.6 CRF with Learned Potentials

In order to introduce spatial context into the segmentation method, a CRF is defined over the supervoxels that incorporates the SVM predictions in addition to promoting spatial consistency of the labels using a learned metric.

A detailed description of Markov random fields and conditional random fields was given in Section 3.2 of the previous chapter. The CRF model used in this chapter differs only in that it is defined on a 3D supervoxel graph as opposed to a 2D superpixel graph. Letting $\mathcal{S}$ denote the set of supervoxel sites and $\mathcal{N}_i$ denote the neighbours of supervoxel i, the energy function can once again be written

$$E(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{S}} \underbrace{\psi(\mathbf{y}_i \mid x_i)}_{\text{Data term}} + \lambda \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} \underbrace{\phi(\mathbf{y}_i, \mathbf{y}_j \mid x_i, x_j)}_{\text{Smoothness term}} \tag{4.14}$$

where the constant λ controls the relative importance of the data and smoothness terms. The data term of the CRF gives the cost of assigning the label x_i at site i. The smoothness term gives the cost of assigning the labels x_i and x_j at the neighbouring sites i and j and is defined so as to promote spatial consistency of the labels.

As with the CRF model presented in the previous chapter for 2D segmentation, graph cuts can find the optimal solution (global minimum) of the energy function. The min-cut/max-flow algorithm of Boykov and Kolmogorov [2004] is used to find the optimal solution.

4.6.1 Potential Functions

The definition of the CRF potential functions follows Section 3.5 and is repeated here for completeness. The first-order potential of the CRF (4.14) is defined as the negative log likelihood of an observation (feature vector) given the class label (i.e. object or background):

$$\psi(\mathbf{y}_i \mid x_i) = -\log\left(P(\mathbf{y}_i \mid x_i)\right) \tag{4.15}$$

where $P(\mathbf{y}_i \mid x_i)$ is the SVM probability estimate for the supervoxel.

Many graph cut formulations incorporate a second-order penalty based on a Euclidean distance measure between the features of neighbouring sites, such as the one proposed by [Boykov and Funka-Lea, 2006]. However, using a standard Euclidean distance disregards any regularities which may be present in the data and which can be exploited to improve performance. In order to address these issues, the focus is

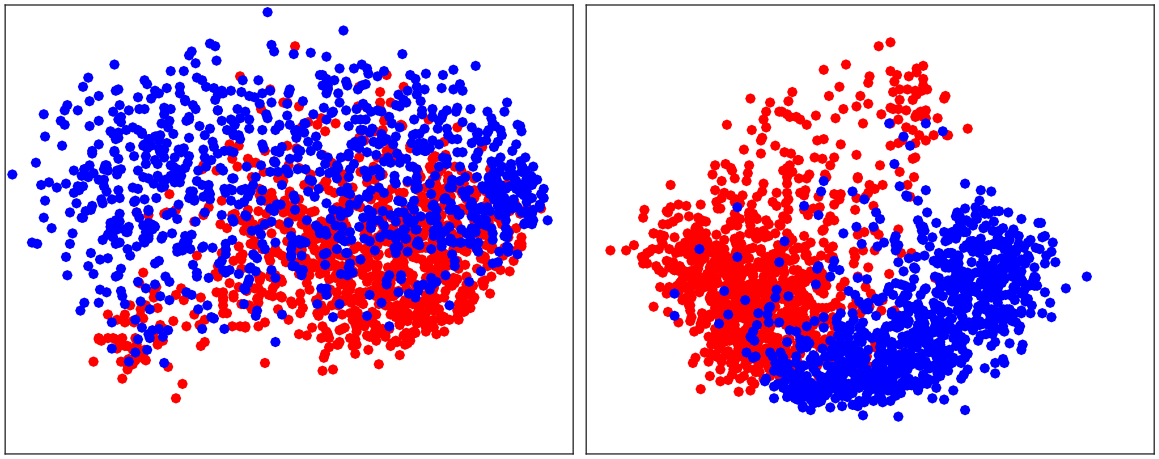


Figure 4.10: The left image shows the projection of a sample of 2000 supervoxel features onto their first 2 principal components prior to metric learning (red are positive examples, blue are negative). The right image shows the projected features in the transformed space found by LMNN.

instead placed on using *distance metric learning* to learn an appropriate distance metric for the second-order potential. More specifically, the Large Margin Nearest Neighbour (LMNN) [Weinberger and Saul, 2009] algorithm is used to learn a pseudometric of the form

$$D_{\mathbf{M}}(\mathbf{y}_i, \mathbf{y}_j) = (\mathbf{y}_i - \mathbf{y}_j)^\top \mathbf{M} (\mathbf{y}_i - \mathbf{y}_j) \tag{4.16}$$

where $\mathbf{M}$ is a positive semidefinite matrix. This matrix can be expressed in terms of a linear transformation $\mathbf{L}$ such that $\mathbf{L}^\top \mathbf{L} = \mathbf{M}$. The objective is to learn the metric such that the k-nearest neighbours of examples in the transformed space (determined by $\mathbf{L}$) belong to the same class while those belonging to different classes are separated by a large margin. There are a number of parallels between the LMNN algorithm and SVMs as a consequence of their shared focus on margin maximisation, with both methods employing a convex optimisation procedure using similar loss functions. Figure 4.10 shows the projection of a sample of supervoxel features onto their first 2 principal components in the original and transformed space. The learned metric is incorporated into the second-order potential function as follows

$$\phi(\mathbf{y}_i, \mathbf{y}_j \mid x_i, x_j) = \begin{cases} \exp\left(-D_{\mathbf{M}}(\mathbf{y}_i, \mathbf{y}_j)\right) & \text{if} \quad x_i \neq x_j \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

which penalises neighbouring supervoxels that have similar feature vectors and are assigned to different classes. Note that, as with the SVM used in the data term, the metric $\mathbf{M}$ need only be learned once on the training data. The same leave-one-out testing approach was used for the LMNN algorithm, with a metric learned on the training volumes for each leave-one-out iteration and then applied on the volume

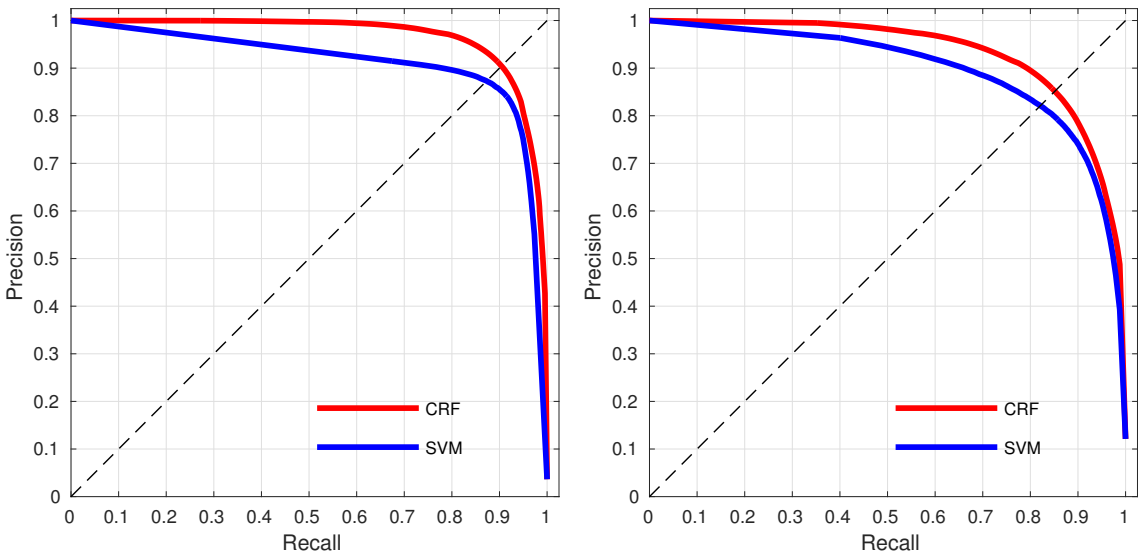


Figure 4.11: Precision-recall curves for the voxel-level smoothed CRF max-marginals (red) and SVM predictions (blue). Results are shown for the IVD dataset (left) and LV dataset (right).

from the held out subject.

4.7 Smoothed Max-Marginals

Soft estimates of the supervoxel class labels are obtained from the CRF by computing the max-marginals using the same method that was described previously in Section 3.5.3. As the max-marginals are on a supervoxel-level, before thresholding to obtain the final segmentations the probability maps are filtered on a voxel-level to smooth any boundary errors which may occur from inaccuracies of the supervoxel layout. A Gaussian filter is used with a kernel size of 5 and $\sigma = 1.5$ to smooth the probabilities. Example probability maps after Gaussian filtering are shown in Figure 4.9.

To evaluate the CRF in comparison with the raw classifier predictions, precision-recall curves are calculated using the smoothed max-marginals and the SVM probability estimates on the voxel-level (shown in Figure 4.11). The additional spatial context provided by the CRF significantly improves the performance over the raw classifier predictions. The final hard label assignments to the individual voxels are obtained by thresholding the smoothed max-marginals, where the threshold is chosen to maximise the F-score on the training data. As a final post-processing step, any isolated components of the segmentation that are below a small threshold size are removed.

4.8 Experiments and Results

In this section implementation details are discussed along with further experiments investigating the effect of the supervoxel parameters. The final segmentation results are subsequently presented and discussed in detail.

4.8.1 Datasets

The two MRI datasets used for the experiments are as described in Sections 2.5.2. MRI Dataset 2 and 2.5.2. MRI Dataset 3. To recap, the LV dataset consists of manually annotated T1-weighted turbo-spin-echo MR images of a section of the lumbar spine from 8 different subjects, encompassing the lumbar vertebrae from L3 to L5. The IVD dataset consists of manually annotated T2-weighted turbo-spin-echo MR images of intervertebral discs from 15 different subjects. The CT dataset is described in Section 2.5.2. CT Dataset.

4.8.2 Implementation Details

Learning the dictionaries using sparse coding (4.2) requires setting the parameter β, which controls the sparsity of the solution. For all results given in this chapter the parameter is simply fixed to $\beta = 0.4$. It was found that the method is not very sensitive to the exact choice of β, with values in the range 0.2-0.6 yielding consistently good results.

When extracting and pooling the learned sparse features within supervoxels, the boundary conditions need to be determined. One option is to allow the non-centre voxels of the sampled patches to cross the supervoxel boundary; the other is to constrain them to always be within the interior of the supervoxel boundary. A Wilcoxon signed-rank test showed that there was no significant difference in classification performance using either option ($p = 0.36$ for the IVD dataset and $p = 0.63$ for the LV dataset). The first option is therefore used as not having to check for supervoxel membership results in a slight speedup during feature extraction.

4.8.3 Effect of Supervoxel Size

The supervoxels generated using SLIC are controlled by two parameters, S and m, which control their average size and spatial regularity. This section evaluates the effect that varying the size of the supervoxels has on the voxel-level overlap with the manual annotations. Note that for a given value of the parameter S, the approximate

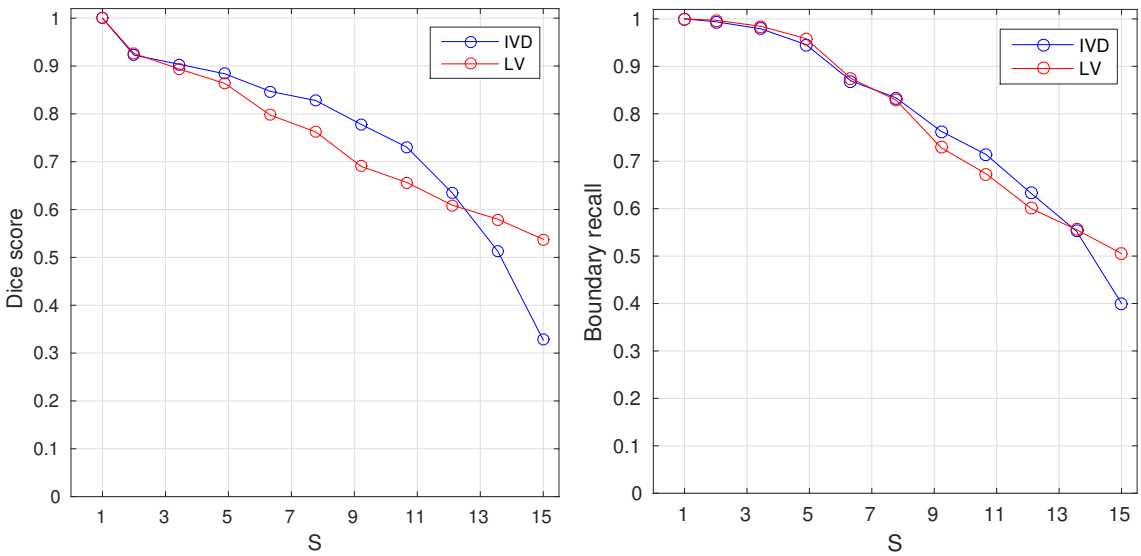


Figure 4.12: (Left) Effect of SLIC size parameter S on the voxel-level Dice score. The values are given for both the IVD dataset (blue) and LV dataset (red). (Right) Boundary recall as a function of the S parameter.

volume of a supervoxel is S^3. In general terms larger supervoxels are favoured as they are more descriptive and reduce computational complexity. However, smaller supervoxels are less likely to reduce the voxel-level accuracy of the segmentation by crossing object boundaries. The correct choice of the supervoxel parameters is therefore a trade-off between size and boundary adherence.

The voxel-level accuracy of the supervoxels is quantified using two standard measures: the Dice score and boundary recall. The Dice score measures the degree of overlap between the voxel-level manual annotations and the supervoxels (see Section 2.4.1 for the definition). Boundary recall measures the fraction of boundary voxels in the manual annotations that are within a small threshold distance (2 voxels) from at least one voxel of a supervoxel boundary. Figure 4.12 shows the effect of varying the size parameter for both the IVD and LV datasets; for all results the spatial regularity parameter is kept fixed at $m = 1$. It can be seen that increasing the supervoxel size results in an essentially monotonic decrease in both the Dice score and the boundary recall.

It is clear that the limited contrast and resolution of the MR images necessitates a small supervoxel size to preserve the object boundaries. Note however that although the supervoxel size needs to be relatively small, this still translates into a large reduction in time and memory complexity compared to operating on the individual voxels. For example, setting $S = 3$ reduces the number of image regions that need to be processed on average by a factor of 27. This reduction in complexity is a necessity for implementation on standard hardware, in terms of both time and memory requirements.

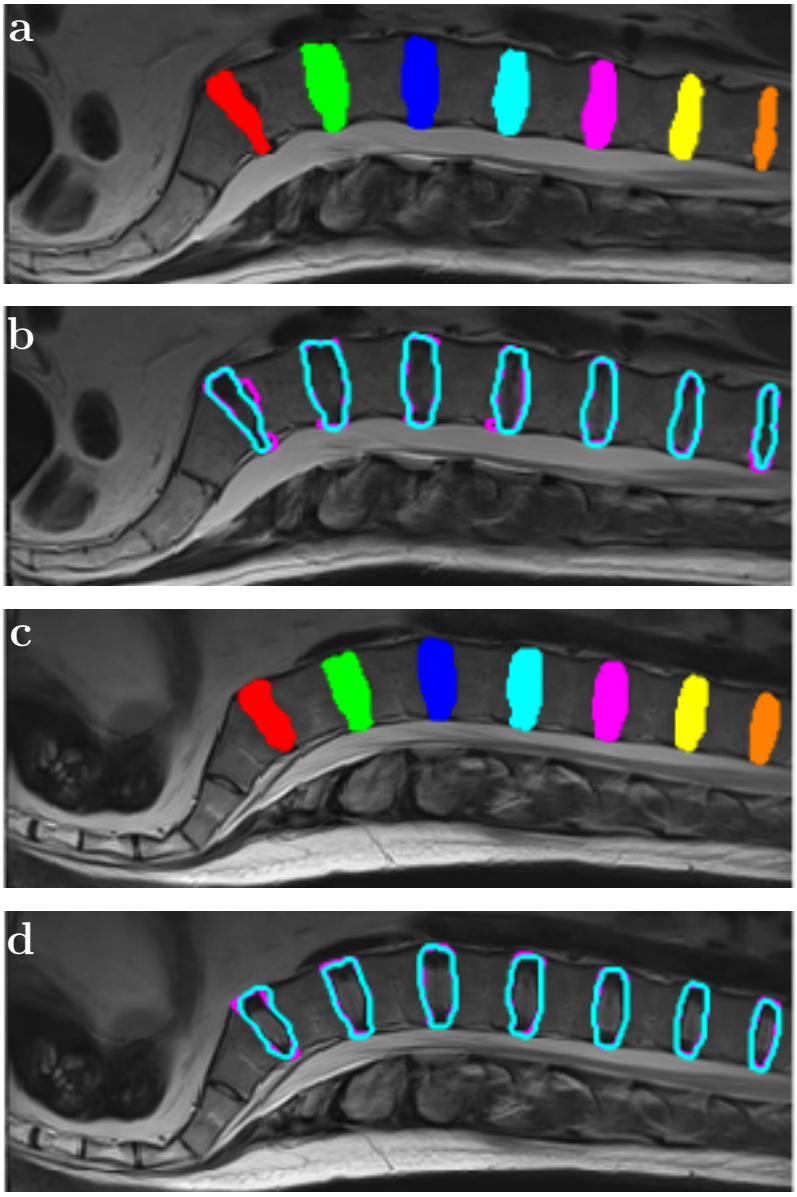


Figure 4.13: (a, c) Example segmentation results on the IVD dataset overlaid onto a mid-sagittal slice from two subjects. (b, d) Overlap between the CRF segmentations (cyan) and manual annotations (magenta).

4.8.4 Segmentation Results

The voxel-level segmentation performance of the method is next evaluated on the annotated MRI datasets. The Dice similarity coefficient (or Dice score) was used to measure the segmentation quality, which is equivalent to the F-score (see Section 2.4.1 for details). Leave-one-out testing was used to evaluate the segmentation performance of the method, with the summary statistics of the Dice score taken over all leave-one-out iterations. The segmentations were also evaluated using the three surface distance measures described in Section 2.4.1.

The results on the IVD and LV datasets using the different evaluation measures are summarised in Table 4.2. Figure 4.9 shows 3D views of the final segmentation

Table 4.2: Segmentation results on the MRI and CT datasets. The table shows the minimum, median and maximum values over all individual volumes.

Dataset	Measure	Min	Med.	Max
IVD	Dice score	0.84	0.91	0.94
	Avg. dist. (mm)	0.34	0.54	1.37
	RMS dist. (mm)	0.71	0.98	4.51
	Max. dist. (mm)	4.00	6.56	33.00
LV	Dice score	0.83	0.86	0.86
	Avg. surf. dist. (mm)	0.71	0.90	1.16
	RMS surf. dist. (mm)	1.15	1.60	2.28
	Max. surf. dist. (mm)	12.36	17.36	19.88
CT	Dice score	0.84	0.96	0.97
	Avg. surf. dist. (mm)	0.18	0.20	1.55
	RMS surf. dist. (mm)	0.42	0.47	2.99
	Max. surf. dist. (mm)	5.66	8.25	17.77

results for example volumes from each dataset.

On the IVD dataset the mean Dice score was 0.90 ± 0.03 and the mean average surface distance was 0.63 ± 0.32. The method outperforms the state-of-the-art results recently reported by Chen et al. [2015] on the same dataset, who obtained a mean Dice score of 0.85-0.88 and a mean average surface distance of 1.3-1.4 mm. Note that although obtained on a different dataset, the results are also significantly higher than the 3D IVD segmentation results of Neubert et al. [2011], who reported a mean Dice score of 0.76-0.80. Figure 4.13 provides a visual comparison between the automatic segmentations and manual annotations.

On the LV dataset the mean Dice score was 0.85 ± 0.01. The Dice scores are expected to be lower on the LV dataset due to the low contrast between bone and surrounding tissue, which makes it extremely challenging for automated segmentation algorithms to obtain accurate results. Despite this the method shows promising results, with the segmentations in most cases comparable to the manual annotations (see Figure 4.14). The Dice scores are similar to those reported for vertebral body segmentation by Neubert et al. [2011], who obtained a mean Dice score of 0.83-0.85. However, the method is able to segment the entire vertebrae (i.e. the vertebral body and posterior elements) while still maintaining a high level of accuracy. The average symmetric surface distances obtained by the method are very good, exceeding the surface-to-surface accuracy reported by Kadoury et al. [2013] for lumbar vertebrae. Note however that in this case the evaluation was not carried out on the same MRI dataset, so they are not directly comparable. The median average surface distance and median RMS surface distances are comparable with the in-plane resolution of 1.02 mm. The maximum surface distances obtained by the method are due to the segmentation occasionally missing a very thin transverse or spinous process. A limitation of the method when applied to this dataset is the fusion of the facet joints

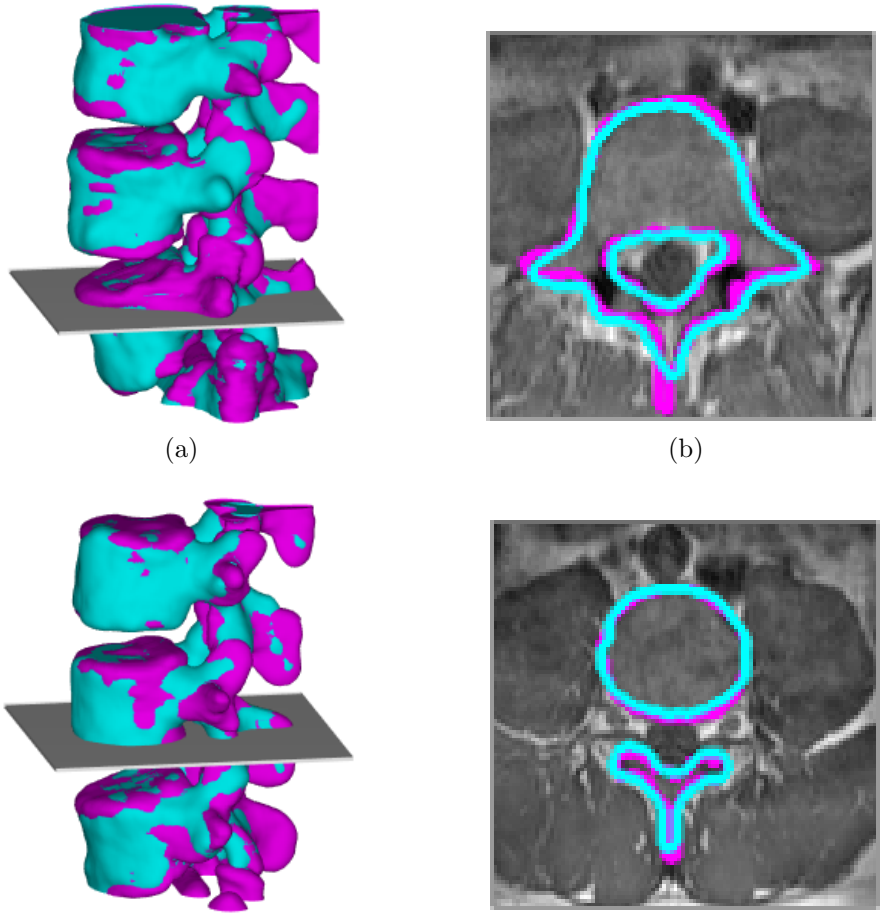


Figure 4.14: (a, c) Example segmentation results on the LV dataset showing the volume overlap between CRF segmentations (cyan) and manual annotations (magenta) for two subjects. (b, d) Segmentation boundaries in the axial slices marked by the grey plane.

connecting adjacent vertebrae. This occurs in most cases due to the low contrast and limited resolution of the MR images, which make it extremely difficult to distinguish the facet joint regions from other surrounding structures.

The average Dice score on the CT data was 0.95 ± 0.04. Figure 4.15 is a visual comparison between the automatic segmentations and manual annotations. The results on CT compare favourably with state-of-the-art methods recently evaluated using the same dataset. For example, Korez et al. [2015] use a shape-constrained deformable model and report an average Dice score of 0.94 ± 0.02 on lumbar vertebrae. Also on the same dataset, Forsberg [2015] obtain a Dice score of 0.94 ± 0.03 using an atlas-based registration approach. A variational framework incorporating shape and intensity priors is used in Hammernik et al. [2015], obtaining a Dice score of 0.96 ± 0.02. Note that all of these methods rely on an explicit prior model of vertebral shape. The lowest Dice score the method obtained was on subject number 6, which was also noted as problematic in Hammernik et al. [2015]. The reason given by the authors was the similar appearance of trabecular bone to soft tissue,

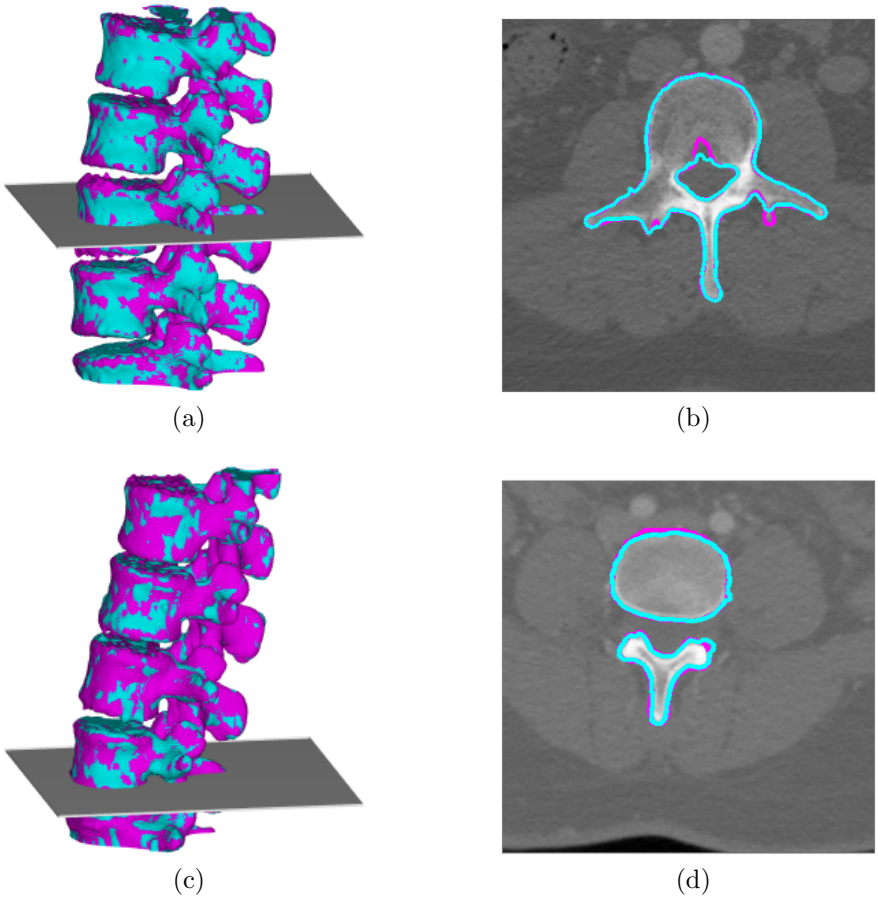


Figure 4.15: (a, c) Example segmentation results on the CT dataset showing the volume overlap between CRF segmentations (cyan) and manual annotations (magenta). (b, d) Segmentation boundaries in the axial slices marked by the grey plane.

leading to registration errors. The maximum surface distance was also obtained on subject 6 as a result of the segmentation missing a transverse process. Note that the median average surface distance and median RMS surface distances are comparable with the in-plane resolution for these data (between 0.31 mm and 0.36 mm).

4.8.5 Detailed Comparison

To further demonstrate the advantages of the proposed method, a comparison is next given with two alternative approaches to medical image segmentation based on the level set method [Osher and Sethian, 1988].[4] Level set methods have been used extensively for medical image segmentation tasks and form a baseline for segmentation performance. Two different variants of level sets are evaluated using the implementation provided in the popular ITK-SNAP application [Yushkevich et al., 2006]. The first uses an edge attraction term based on gradient magnitude and corresponds to the widely-used geodesic active contours as originally described in

[4]Level set methods for segmentation were discussed in Section 2.3.2.

Table 4.3: Comparison of different segmentation methods on the two MRI datasets. The table gives the minimum, median and maximum values over all individual volumes. GAC is geodesic active contours, RC-SVM is region competition with SVM pre-segmentation and CRF is the proposed method. The best scores for the various criteria are highlighted in bold.

Dataset	Measure	GAC			RC-SVM			CRF		
		Min	Med.	Max	Min	Med.	Max	Min	Med.	Max
IVD	Dice score	0.15	0.48	0.72	**0.85**	0.89	0.92	0.84	**0.91**	**0.94**
	Avg. dist.	2.04	5.41	11.16	0.44	0.56	**0.77**	**0.34**	**0.54**	1.37
	RMS dist.	3.35	7.40	13.91	0.85	1.04	**1.38**	**0.71**	**0.98**	4.51
	Max. dist.	15.91	18.69	39.23	5.62	6.87	**11.79**	**4.00**	**6.56**	33.00
LV	Dice score	0.49	0.57	0.61	0.81	0.82	0.84	**0.83**	**0.86**	**0.86**
	Avg. dist.	7.17	8.85	10.34	0.76	0.93	**1.03**	**0.71**	**0.90**	1.16
	RMS dist.	11.18	13.45	15.64	1.21	**1.45**	**2.12**	**1.15**	1.60	2.28
	Max. dist.	42.05	52.35	59.13	**11.00**	**14.42**	22.92	12.36	17.36	**19.88**

Caselles et al. [1997]. For the second variant the SVM predictions for the supervoxels are taken as an initial pre-segmentation to initialise a level set method based on the region competition model of Zhu and Yuille [1996]. As level set methods are a continuous space alternative to discrete CRF models, providing the same initialisation (via the SVM) enables a fair comparison between the relative merits of the two competing approaches. Note that both level set variants require interactive placement of seed points for initialisation, which was carried out in ITK-SNAP using a single seed point at the centre of each individual vertebra or IVD. Table 4.3 provides a detailed comparison of the segmentation results obtained by the different methods on both the IVD and LV datasets.

In the case of geodesic active contours, the low contrast between the different structures in the images leads to the object boundaries being violated by the segmentation. This results in severe under- and over-segmentation errors due to the lack of a clear boundary around the objects and the inhomogeneous intensity properties. As a consequence of these problems the geodesic active contour method failed to obtain an adequate segmentation on either dataset, with the best scores in most cases being significantly lower than the worst scores of the other two methods.

The region competition method initialised with the SVM predictions achieves much better results, which demonstrates the effectiveness of the supervoxel features and classification procedure. However, it can be seen that the final segmentation performance is significantly worse overall than the proposed CRF model on both datasets. The one area where the region competition model obtained better results is in terms of the worst-case performance as measured by the maximum surface distances (see Table 4.3). This may be explained by the fact that the level set method operates on a voxel-level, whereas the CRF model is supervoxel-based and can occasionally miss some of the finer details around the object boundaries.

Table 4.4: Results from the MICCAI-CSI 2015 localisation and segmentation challenge. A description of the different evaluation measures is given in the text.

Dataset	Dice	ASD (mm)	MLD (mm)	$R_{2\,\mathrm{mm}}$	$R_{4\,\mathrm{mm}}$	$R_{6\,\mathrm{mm}}$
Training	0.90 ± 0.03	0.63 ± 0.32	N/A	N/A	N/A	N/A
Test 1	0.90 ± 0.04	1.24 ± 0.24	1.05 ± 0.69	94.3	100	100
Test 2	0.91 ± 0.03	1.19 ± 0.20	0.89 ± 0.63	91.4	100	100

4.8.6 MICCAI-CSI 2015 Challenge Results

The method presented in this chapter was a winning entry into the MICCAI-CSI 2015 challenge on automatic IVD localisation and segmentation from 3D T2 MRI data [Zheng et al., 2016].[5] In this section an overview of the challenge is given and the results obtained by the method are presented.

The dataset used for the challenge consists of 25 different subjects, which is split into a training dataset of 15 subjects (constituting the IVD dataset used in previous sections) and two additional test datasets each consisting of 5 subjects. Participants in the challenge were invited to submit the results from their methods on the test data, which were then independently evaluated and compared with a number of competing methods. For the first stage of the challenge the participants submitted their results obtained on the first test dataset of 5 images. The second stage of the challenge involved an on-site evaluation where each participant was required to apply their method on 5 additional test images and submit the results within the time limit of one hour.

The segmentations were evaluated using the Dice score and average absolute surface distance (ASD). The mean Dice score on the training dataset was 0.90 ± 0.03 and the mean ASD was $0.63\,\mathrm{mm} \pm 0.32\,\mathrm{mm}$. On the two test datasets the mean Dice scores were 0.90 ± 0.04 and 0.91 ± 0.03; the mean ASD values were $1.24\,\mathrm{mm} \pm 0.24\,\mathrm{mm}$ and $1.19\,\mathrm{mm} \pm 0.20\,\mathrm{mm}$. A summary of the segmentation results for each stage of the challenge is given in Table 4.4.

The challenge entries were also evaluated on the test data in terms of localisation accuracy. This was measured by comparing the centroids of the automatically detected discs with the ground truth centroids. Assuming a total of n IVDs and letting x_i, y_i and z_i denote the spatial coordinates of the detected IVD centroids, the mean localisation distance is defined as

$$\mathrm{MLD} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2} \tag{4.18}$$

[5] A description and short report on the challenge can be found at `http://ijoint.istb.unibe.ch/challenge/index.html`

Algorithm 2 Model testing as function of sample size

Require: Set of all volumes $\mathcal{D}$

1: **for** $k = 1, \ldots, |\mathcal{D}| - 1$ **do**
2: **for** $n = 1, \ldots, \min\left\{19, \binom{|\mathcal{D}|}{k}\right\}$ **do**
3: $\mathcal{D}_{\text{train}} \leftarrow$ sample k-combination (without replacement)
4: Train model on $\mathcal{D}_{\text{train}}$
5: $\mathcal{D}_{\text{test}} \leftarrow \mathcal{D} \setminus \mathcal{D}_{\text{train}}$
6: Test model on $\mathcal{D}_{\text{test}}$
7: **end for**
8: **end for**

where x'_i, y'_i and z'_i are the coordinates of the ground truth centroids. The results are reported as physical mm distances.

The successful localisation rate R_t is defined as the percentage of correctly localised IVDs within t mm of the ground truth:

$$R_t = \frac{\text{number of IVD locations within } t\,\text{mm}}{\text{total number of IVDs}} \cdot 100. \tag{4.19}$$

A summary of the localisation results is given in Table 4.4. Note that evaluation of the localisation results was only carried out on the two test datasets. Out of a total of 10 participating teams, the method was unique in obtaining a top 3 ranking in both the localisation and segmentation stages of the challenge.

4.9 Influence of Sample Size

This section investigates the influence of the sample size (i.e. the number of individual subjects) on the resulting performance of the method. This is an important consideration as a primary goal of any segmentation method is to generalise well beyond the finite training set to arbitrary new images. An evaluation of the performance in terms of sample size indicates whether further improvement could be gained by expanding the training set with more examples. Due to the relatively small size of the LV dataset and given that the method has been compared on the IVD and CT datasets with competing methods, attention is focused on the LV dataset for the remainder of this section.

A random sampling procedure is used to evaluate the performance characteristics using different subsets of the training data. At each iteration of the procedure a subset of k subjects is randomly selected without replacement from the dataset to train the model. The trained model is then tested on the remaining subjects and the process is repeated n times for each iteration. Algorithm 2 gives pseudocode for

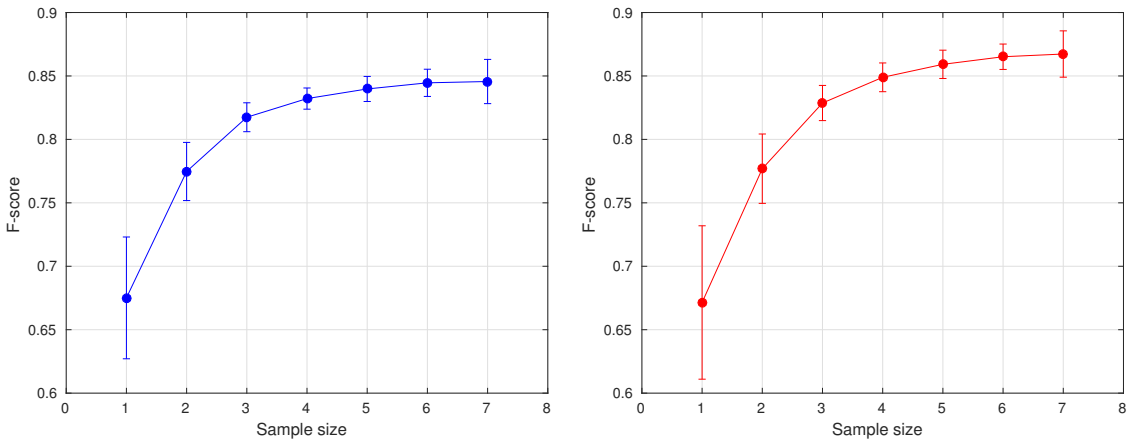


Figure 4.16: Learning curves for the SVM predictions (left) and the CRF max-marginals (right) showing the F-score as a function of the training sample size.

the general procedure. Note that the exhaustive enumeration of all k-combinations at step 2 is feasible due to the LV dataset only containing 8 volumes. This explicit enumeration step prevents the results from being biased due to sampling the same combinations of volumes multiple times. As the number of k-combinations can be very large even for small k, the maximum number of combinations is restricted to 19 in step 2. This ensures that an adequate number of combinations are sampled at each iteration, but prevents the algorithm from becoming prohibitively expensive.

Figure 4.16 shows plots of the F-scores (4.13) obtained by the SVM classifier and CRF as the sample size of the training set is increased (known as the *learning curves*). The error bars show the standard deviation over the random samples at each iteration. It can be seen that most of the improvement is obtained as the sample size is increased from 1 to 4, after which the learning curves begin to converge in both cases. This suggests that despite the relatively small size of the dataset, adding more data will not necessarily lead to an improvement in the classification performance. By comparing the two plots it is also evident that the CRF only starts to outperform the SVM after the second iteration. This may be due to the fact that the λ parameter of the CRF can be more accurately tuned with a larger training sample size.

A summary of the segmentation results from the experiment is presented in Table 4.5, which gives the median Dice scores and surface distance values obtained on each sampled subset of the training data. Figure 4.17 shows box plots for the segmentation results taken over each sample. As with the learning curves for the SVM and CRF, most of the improvement in segmentation performance is obtained by increasing the sample size from 1 to 4. Although the best results were obtained using the maximum sample size of 7, the median evaluation scores are consistently good once the sample size reaches 4. It is important to note that while the perfor-

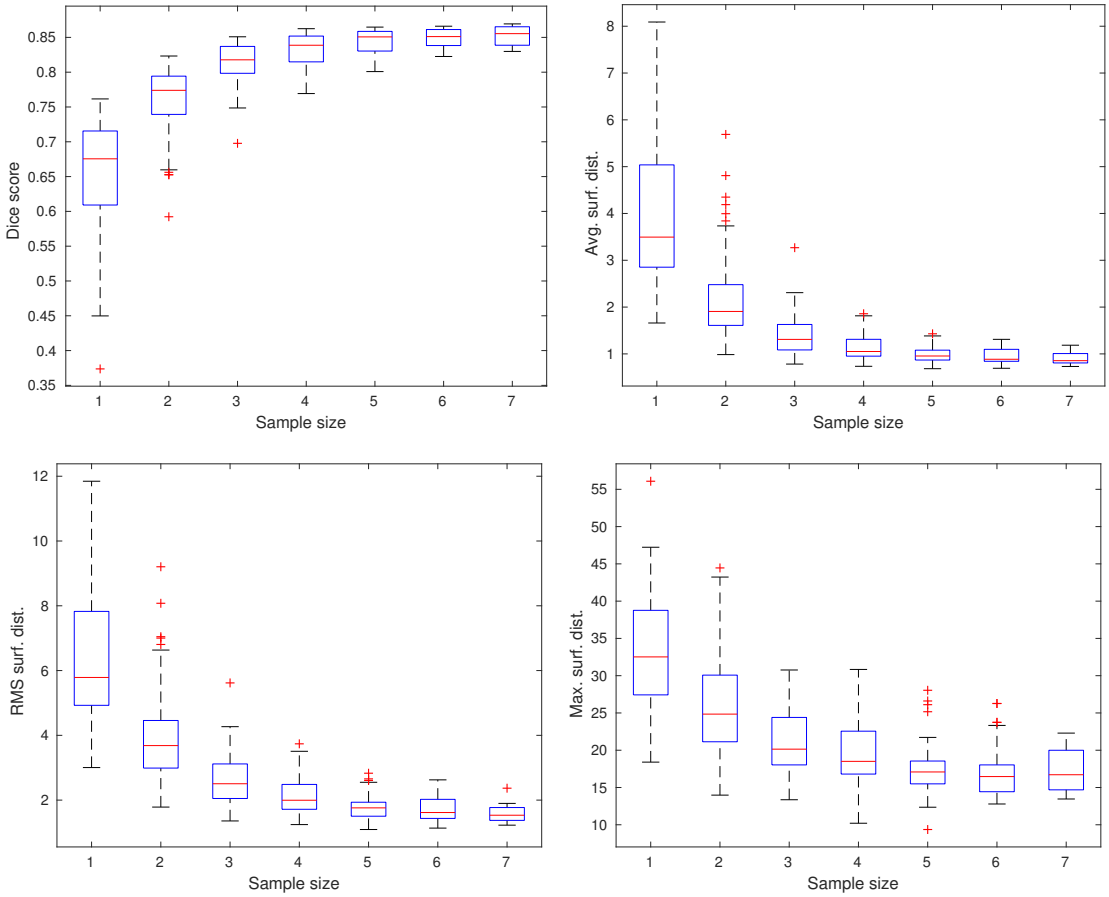


Figure 4.17: Box plots of the Dice scores and surface distance measures as a function of the training sample size.

mance increase for the last few iterations is small, this could still translate into a noticeable improvement in the anatomical accuracy of the segmentation. The results suggest that despite the relatively small size of the dataset, the model is still able to learn and generalise effectively after training on only a subset of the examples. One explanation for this ability is that while the number of individual subjects in the sample is small, the number of supervoxels is large enough to enable the model to capture most of the variation in the local features of the images. This would suggest that the method is less sensitive to the number of subjects in the sample than those relying on more global features such as shape.

4.10 Conclusion

This chapter described a fully automated approach to segmentation of the lumbar spine from 3D MR images. The method is able to obtain accurate segmentations of lumbar vertebrae and intervertebral discs, which previously have been approached as largely independent problems employing specialised techniques. Although this chapter has focused on segmentation from MRI data, the method is not dependent

Table 4.5: Segmentation performance as a function of the training sample size. The rows give the median values of the evaluation measures for each sample size.

Sample size	1	2	3	4	5	6	7
Dice score	0.68	0.77	0.82	0.84	0.85	0.85	**0.86**
Avg. surf. dist. (mm)	3.50	1.91	1.31	1.05	0.96	0.89	**0.86**
RMS surf. dist. (mm)	5.79	3.68	2.51	2.00	1.76	1.62	**1.53**
Max. surf. dist. (mm)	32.52	24.84	20.14	18.51	17.10	**16.47**	16.71

on a particular imaging modality. Specifically, the method can also obtain accurate lumbar vertebra segmentations from CT data, which is facilitated by the greater contrast between bone and surrounding tissue.

The use of unsupervised feature learning as an alternative to hand-designed features was shown to be very effective for obtaining high classification performance. Furthermore, learning features over multiple scales using pyramids incorporates larger-scale spatial context whilst ensuring that the dictionary learning process remains tractable. For vertebra segmentation, combining the learned features with information on the relative location of supervoxels further improves performance by helping to localise the segmentation. Incorporating the learned classifier and metric into a CRF model over the supervoxels was shown to promote spatial consistency and enable accurate and robust segmentation.

The method is efficient in terms of computational cost, with each individual 3D MR image taking approximately 6 min to segment using an Intel Core i5 2.50 GHz machine with 8 GB of RAM running Linux (64-bit). The implementation is written in MATLAB with external C++ code for computationally intensive tasks including supervoxel generation, SVM optimisation and computation of the CRF max-marginals. The performance of the algorithm can be improved by initially classifying a random sample of the supervoxels. A minimum bounding box around the object class supervoxels can then be computed automatically, with subsequent processing constrained to this cropped region. The final segmentation is then padded at the edges of each axis to recover the original volume dimensions.

Although basing the model on supervoxels is required in order to make the implementation computationally feasible, this means that the final segmentation performance is dependent on the voxel-level boundary adherence of the initial supervoxels. This could be addressed by adopting a coarse-to-fine hierarchical procedure, whereby an initial (coarse) set of supervoxels are first classified before refining the segmentation boundary using a finer set of supervoxels. As the finer set of supervoxels need only be considered in the local area around the segmentation boundary, computational complexity would remain tractable.

The general approach outlined in this chapter for 3D MRI segmentation is likely to be applicable to other segmentation tasks in medical image analysis and related areas. The absence of object-specific priors means that the method can be trained to segment structures from different imaging modalities without fundamental change to the model. The results in this chapter show that although the method was originally developed for 3D lumbar vertebra segmentation from MR images, it can be applied with minimal change to the 3D segmentation of intervertebral discs and is directly applicable to CT data.

5 Summary and Conclusion

This thesis presented an approach for automatic segmentation of the lumbar spine from medical images. The efficacy of the approach was demonstrated on the problems of vertebra and intervertebral disc segmentation from images acquired using MRI and CT. One of the main advantages of the method is its generality, enabling segmentation of both vertebrae and intervertebral discs in 3D without fundamental change to the model. This is in contrast to most previous work in the literature, which has tended to treat the two problems as largely independent and requiring specialised techniques.

The segmentation results were shown to be accurate on MRI and CT data in both 2D and 3D; in all cases the evaluation measures are comparable with the corresponding manual annotations. Although segmentation performance is typically quantified by comparison with manually annotated data, simply maximising the values of the overlap measures (e.g. the Dice score) is not necessarily advantageous. For example, manual annotation of anatomical objects is known to be subjective and some degree of intrinsic error can be assumed to exist. Thus, the best possible segmentation performance does not in general correspond to a perfect Dice score. A more reliable estimate of the segmentation performance could be obtained by evaluating the method against multiple annotators to assess the inter-annotator agreement. This would give a better estimate of the maximum possible performance that can be expected from the automated method, in addition to providing a more reliable measure of the segmentation error by comparing the automatic and consensus annotations. It should be noted however that as pixel- or voxel-level manual annotation is extremely time-consuming even for the relatively small datasets used in this thesis, obtaining segmentations from multiple expert annotators is often prohibitively expensive for larger datasets.

A central theme of the method introduced in this thesis is the extensive use of both supervised and unsupervised machine learning techniques. While most previous approaches to spine segmentation have focused on incorporating top-down prior shape information, the preceding chapters have shown that it is possible to obtain accurate segmentations without any explicit shape constraints provided that an effective representation can be learned that takes into account information across multiple scales. An interesting question is whether a significant improvement in

accuracy could be obtained by the introduction of more global shape information into the segmentation model. Related to this is the possibility of using a sophisticated model of anatomical shape to enable a decomposition of the segmented object in terms of its component parts. This type of contextual, part-based information is much more difficult to obtain using only local features of the images. However, as discussed in Section 2.3.1, one of the main limitations associated with standard statistical shape models is the difficulty in constructing a representative prior model of anatomical shape, which requires the presence of a sufficiently large dataset with corresponding landmark points. This is compounded by the introduction of bias into the model due to potentially inaccurate preconceptions of the human annotators, in addition to variation in both the intra- and inter-annotator agreement.

There are a number of aspects of the current model which could be extended in future. Although the second-order CRF formulations used in the thesis were shown to be effective, they are limited in the type of penalties that can be encoded into the potential functions. Using higher-order CRF models could help to improve the segmentation by encoding more complex interactions between the supervoxels. These could be designed so as to improve the segmentation of long and thin structures such as the transverse and spinous processes of vertebrae, where most of the segmentation errors occur. As mentioned in the conclusion of the previous chapter, adopting a coarse-to-fine approach could also potentially improve the voxel-level accuracy by effectively increasing the resolution at the segmentation boundary. Another possibility is the introduction of prior statistical shape information into the model by incorporating global parameters into the CRF. Recent work has shown that globally optimal solutions of such models are feasible and can lead to improvements on a number of 2D segmentation tasks [Lempitsky et al., 2012]. A potential application of this approach is in assigning unique labels to the individual vertebrae of the lumbar spine, which requires separating the adjacent vertebrae at the facet joints. This is particularly important for applications in the area of biomechanical modelling, as it enables the model to incorporate the articulation properties of the spine. The limited resolution of the MR images and presence of partial volume effects at the facet joints means that this is only feasible when additional prior information is incorporated into the model.

Due to constraints on the scanning acquisition time, the 3D MR images used in this thesis have a relatively low in-plane resolution. This makes accurate segmentation of specific areas of the vertebrae, such as the thin transverse and spinous processes, extremely challenging due to the small number of voxels comprising these regions. Increasing the resolution could provide more accurate estimates by enabling a larger supervoxel size and more descriptive features. A higher resolution could also increase the boundary adherence of the supervoxels, which is one of the limiting factors in

obtaining highly accurate voxel-level performance. Another limitation of the 3D datasets used in this thesis is that they consist of a small number of subjects (i.e. between 8 and 15). Although the experimental results presented in Section 4.9 suggest that increasing the number of subjects is not necessary for improving the performance of the method, the question remains as to whether enough inter-subject variation is captured when using small sample sizes. For more robust estimates of generalisation performance, additional experiments should ideally be carried out on much larger datasets. Further investigation could also be carried out into combining images acquired from multiple modalities to provide information which may not be captured when using a single modality (e.g. T1- or T2-weighted MRI).

Although the focus of this thesis has been on automatic segmentation, it is also possible to adapt the method for use in a semi-interactive setting. For example, one possibility is to enable the user to visually inspect an initial segmentation in order to check for segmentation errors. Regions of potential errors could be indicated by the user clicking on specific locations in the image, which would subsequently be incorporated into the CRF potential functions as additional spatial priors. The initial segmentation could then be updated with minimal cost using dynamic graph cuts [Kohli and Torr, 2007]. While this approach could potentially further improve the segmentation result, the manual search for errors would still be a very laborious process subject to the same limitations associated with manual annotation in general. It is also unlikely that user-provided spatial information alone can provide enough additional context to improve certain regions of the segmentation, such as the transverse and spinous processes of the vertebrae which are most susceptible to under-segmentation errors.

The multi-scale dictionary learning approach introduced for 3D lumbar spine segmentation is directly applicable to a wide range of tasks in medical image analysis. This is especially relevant in the case of MR image analysis, where it is often difficult to design descriptive features by hand that are robust to image artefacts such as noise and intensity inhomogeneity. In addition, the results suggest that features which enable high classification performance on MRI data are also likely to obtain high performance on other modalities such as CT. Although in this work the dictionaries were learned in an unsupervised fashion, it is also possible to adapt the optimisation objective for a supervised setting [Mairal et al., 2009]. This enables both the dictionaries and classifier parameters to be learned jointly, which could lead to improvements in performance. Another area for future research is in learning higher level representations for classification that can take into account more complex features of the images. An example within the sparse modelling domain is hierarchical convolutional sparse coding [Mairal et al., 2014], which performs successive applications of sparse coding and spatial pooling to extract higher-level representations.

Autoencoders and their extensions such as stacked denoising autoencoders [Vincent et al., 2010] are another example in this direction and have proven to be effective on natural images. Recent work has investigated their application to classification problems in medical image analysis [Thong et al., 2015], where they show promising results. The more general area of unsupervised learning is particularly relevant for medical image analysis as acquiring sufficient labelled data can be expected to remain a bottleneck in the use of purely supervised machine learning algorithms.

The approach described in this thesis is likely to have applications beyond the segmentation of vertebrae and intervertebral discs, particularly in the area of MR image analysis. For example, in the context of cardiac MRI automatic segmentation of the left and right ventricles is known to be a very challenging problem and existing methods require further improvement to provide results suitable for medical applications [Petitjean and Dacher, 2011, Petitjean et al., 2015]. The large inter-subject variation in the shape of the ventricles means that methods relying on global shape information suffer from the same problems encountered when using shape models for spine segmentation. The challenges involved in developing automated computational methods in these areas are therefore similar to those encountered in lumbar spine segmentation, including the complexity of the anatomy and the limited contrast with other structures in the image.

References

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recogn.*, 13(2):111–122, 1981.

A. J. Bell and T. J. Sejnowski. Edges are the 'Independent Components' of Natural Scenes. In *Adv. Neural Info. Process. Syst. 9*, pages 831–837. The MIT Press, 1997.

J. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

C. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, NY, 3rd edition, 2006.

A. Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion.* Springer-Verlag, Secaucus, NJ, 1st edition, 1998. ISBN 3540762175.

A. Blake, P. Kohli, and C. Rother, editors. *Markov Random Fields for Vision and Image Processing.* The MIT Press, Cambridge, MA, 1st edition, 2011.

N. Bogduk. *Clinical and Radiological Anatomy of the Lumbar Spine.* Churchill Livingstone, Edinburgh, UK, 5th edition, 2012.

Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning Mid-Level Features For Recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2559–2566, 2010.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 1st edition, 2004.

Y. Boykov and G. Funka-Lea. Graph Cuts and Efficient N-D Image Segmentation. *Int. J. Comput. Vis.*, 70(2):109–131, 2006.

Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

R. Bryan, P. B. Nair, and M. Taylor. Use of a statistical model of the whole femur in a large scale, multi-model study of femoral neck fracture risk. *J. Biomech.*, 42(13):2171–2176, 2009. URL `http://dx.doi.org/10.1016/j.jbiomech.2009.05.038`.

J. Q. Campbell and A. J. Petrella. An Automated Method for Landmark Identification and Finite-Element Modeling of the Lumbar Spine. *IEEE Trans. Biomed. Eng.*, 62(11):2709–2716, 2015.

J. Carballio-Gamio, S. J. Belongie, and S. Majumdar. Normalized Cuts in 3-D for Spinal MRI Segmentation. *IEEE Trans. Med. Imag.*, 23(1):36–44, 2004.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 694–699, 1995.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. J. Comput. Vis.*, 22(1):61–79, 1997.

I. Castro-Mateos, J. M. Pozo, T. F. Cootes, J. Mark Wilkinson, R. Eastell, and A. F. Frangi. Statistical Shape and Appearance Models in Osteoporosis. *Curr. Osteoporos. Rep.*, 12(2):163–173, 2014.

I. Castro-Mateos, J. M. Pozo, M. Pereanez, K. Lekadir, A. Lazary, and A. F. Frangi. Statistical Interspace Models (SIMs): Application to Robust 3D Spine Segmentation. *IEEE Trans. Med. Imag.*, 34(8):1663–1675, 2015.

A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Trans. Imag. Process.*, 10(2):266–277, 2001. ISSN 1057-7149. doi: 10.1109/83.902291. URL `http://dx.doi.org/10.1109/83.902291`.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Tech.*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C. Chen, D. Belavy, W. Yu, C. Chu, G. Armbrect, M. Bansmann, D. Felsenberg, and G. Zheng. Localization and Segmentation of 3D Intervertebral Discs in MR Images by Data Driven Estimation. *IEEE Trans. Med. Imag.*, 2015. doi: 10.1109/TMI.2015.2403285. in press.

A. Coates and Y. Ng, A. Learning Feature Representations with K-means. In *Neural Networks: Tricks of the Trade (2nd ed.)*, volume 7700 of *Lecture Notes in Computer Science*, pages 561–580. Springer, 2012.

L. D. Cohen. On active contour models and balloons. *CVGIP: Imag. Underst.*, 53 (2):211–218, 1991.

D. J. Cook, D. A. Gladowski, H. N. Acuff, M. S. Yeager, and B. C. Cheng. Variability of manual lumbar spine segmentation. *Int. J. Spine Surg.*, 6(1):167–173, 2012.

T. F. Cootes and C. J. Taylor. Anatomical statistical models and their role in feature extraction. *The British journal of radiology*, 77(2):S133–S139, 2004. URL `http://bjr.birjournals.org/cgi/doi/10.1259/bjr/20343922`.

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active Shape Models—Their Training and Application. *Comp. Vis. Imag. Underst.*, 61(1):38–59, 1995.

D. Cremers, T. Kohlberger, and C. Schnorr. Shape statistics in kernel space for variational image segmentation. *Pattern Recogn.*, 36(9):1929–1943, 2003.

D. Cremers, S. J. Osher, and S. Soatto. Kernel Density Estimation and Intrinsic Alignment for Shape Priors in Level Set Segmentation. *Int. J. Comput. Vis.*, 69 (3):335–351, 2006.

D. Cremers, M. Rousson, and R. Deriche. A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape. *Int. J. Comput. Vis.*, 72(2):195–215, 2007.

N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 886–893, 2005.

R. H. Davies, C. J. Twining, T. F. Cootes, and C. J. Taylor. Building 3-D Statistical Shape Models by Direct Optimization. *IEEE Trans. Med. Imag.*, 29(4):961–981, 2010.

M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer, New York, NY, 1st edition, 2010.

T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

P. F. Felzenszwalb and R. Zabih. Dynamic Programming and Graph Algorithms in Computer Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):721–740, 2011.

M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, 1981.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188, 1936.

D. Forsberg. Atlas-Based Segmentation of the Thoracic and Lumbar Vertebrae. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 215–220. Springer, 2015.

W. Forstner and E. Gulch. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *Proc. ISPRS Conf. Fast Processing of Photogrammetric Data*, pages 281–305, 1987.

B. Fulkerson, A. Vedaldi, and S. Soatto. Class Segmentation and Object Localization with Superpixel Neighborhoods. In *Proc. Euro. Conf. Comput. Vis. (ECCV)*, pages 670–677, 2009.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6): 721–741, 1984.

G. Gerig, M. Jomier, and M. Chakos. Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation. In *Med. Imag. Comput. Comput.-Assist. Interv. (MICCAI)*, volume 2208, pages 516–523. Springer, 2001.

P. Getreuer. Chan-Vese Segmentation. *Image Processing On Line*, 2012, 2012. doi: 10.5201/ipol.2012.g-cv. `http://dx.doi.org/10.5201/ipol.2012.g-cv`.

R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Pearson Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 2008.

J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

R. Guillemaud and M. Brady. Estimating the Bias Field of MR Images. *IEEE Trans. Med. Imag.*, 16(3):238–251, 1997.

K. S. Gurumoorthy, A. Rangarajan, and A. Banerjee. The Complex Wave Representation of Distance Transforms. In *Energy Min. Meth. Comput. Vis. Pattern Recog. (EMMCVPR)*, volume 6819 of *Lect. Notes Comput. Sci.*, pages 413–427. Springer, 2011.

J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Yale University Press, New Haven, CT, 1st edition, 1923.

K. Hammernik, T. Ebner, D. Stern, M. Urschler, and T. Pock. Vertebrae Segmentation in 3D CT Images Based on a Variational Framework. In *Recent Advances in*

Computational Methods and Clinical Applications for Spine Imaging, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 227–233. Springer, 2015.

R. H. Hashemi, W. G. Bradley, Jr., and C. J. Lisanti. *MRI: The Basics*. Wolters Kluwer, Philidelphia, PA, 3rd edition, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2nd edition, 2009.

T. Heimann and H.-P. Meinzer. Statistical shape models for 3D medical image segmentation: A review. *Med. Imag. Anal.*, 13(4):543–563, 2009.

I. C. Hospers, J. G. van der Laan, C. J. Zeebregts, P. Nieboer, B. H. Wolffenbuttel, R. A. Dierckx, H. G. Kreeftenberg, P. L. Jager, and R. H. Slart. Vertebral Fracture Assessment in Supine Position: Comparison by Using Conventional Semiquantitative Radiography and Visual Radiography. *Radiology*, 251(3):822–828, 2009.

J. Huang, F. Jian, H. Wu, and H. Li. An improved level set method for vertebra CT image segmentation. *BioMed. Eng. OnLine*, 12(48), 2013.

S.-H. Huang, Y.-H. Chu, and S.-H. Lai. Learning-Based Vertebra Detection and Iterative Normalized-Cut Segmentation for Spinal MRI. *IEEE Trans. Med. Imag.*, 28(10):1595–1605, 2009.

H. Hutt, R. Everson, and J. Meakin. Segmentation of Lumbar Vertebrae Slices from CT Images. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 61–71. Springer, 2015a.

H. Hutt, R. Everson, and J. Meakin. 3D Intervertebral Disc Segmentation from MRI using Supervoxel-Based CRFs. In *Proc. 3rd MICCAI Wksp. Comput. Meth. Clin. App. Spine Imag. (MICCAI-CSI 2015)*, pages 119–123, 2015b.

M. M. A. Janssen, K. L. Vincken, B. Kemp, M. Obradov, M. de Kleuver, M. A. Viergever, R. M. Castelein, and L. W. Bartels. Pre-existent vertebral rotation in the human spine is influenced by body position. *Eur. Spine. J.*, 19(10):1728–1734, 2010.

I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 2nd edition, 2002.

A. C. Jones and R. K. Wilcox. Finite element analysis of the spine: Towards a framework of verification, validation and sensitivity analysis. *Med. Eng. Phys.*, 30 (10):1287–1304, 2008.

S. Kadoury, H. Labelle, and N. Paragios. Spine Segmentation in Medical Images Using Manifold Embeddings and Higher-Order MRFs. *IEEE Trans. Med. Imag.*, 32(7):1227–1238, 2013.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. Comput. Vis.*, 1(4):321–331, 1988.

B. M. Kelm, M. Wels, S. K. Zhou, S. Seifert, M. Suehling, Y. Zheng, and D. Comaniciu. Spine detection in CT and MR using iterated marginal space learning. *Med. Imag. Anal.*, 17(8):1283–1292, 2013.

Y. Kim and D. Kim. A fully automatic vertebra segmentation method using 3D deformable fences. *Comput. Med. Imag. Graph.*, pages 343–352, 2009.

M. Kirschner, M. Becker, and S. Wesarg. 3D Active Shape Model Segmentation with Nonlinear Shape Priors. In *Med. Imag. Comput. Comput.-Assist. Interv. (MICCAI)*, volume 6892, pages 492–499. Springer, 2011.

T. Klinder, J. Osterman, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. Automated model-based vertebra detection, identification, and segmentation in CT images. *Med. Imag. Anal.*, 13(3):471–482, 2009.

H. Knutsson. Representing Local Structure Using Tensors. In *Proc. Scand. Conf. Imag. Anal.*, pages 244–251, Oulu, Finland, 1989.

H. Knutsson and M. Andersson. Morphons: segmentation using elastic canvas and paint on priors. In *Proc. IEEE Int. Conf. Imag. Process. (ICIP)*, volume 2, pages 1226–1229, 2005.

P. Kohli and P. H. S. Torr. Dynamic Graph Cuts for Efficient Inference in Markov Random Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2079–2088, 2007.

P. Kohli and P. H. S. Torr. Measuring Uncertainty in Graph Cut Solutions. *Comput. Vis. Imag. Underst.*, 112:30–38, 2008.

P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2008.

V. Kolmogorov and R. Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, 2004.

N. Komodakis and N. Paragios. Beyond Pairwise Energies: Efficient Optimization for Higher-order MRFs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2985–2992, 2009.

N. Komodakis, N. Paragios, and G. Tziritas. MRF Energy Minimization and Beyond via Dual Decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552, 2011.

R. Korez, B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec. An Improved Shape-Constrained Deformable Model for Segmentation of Vertebrae from CT Lumbar Spine Images. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 85–94. Springer, 2015.

S. M. Kurtz and A. A. Edidin, editors. *Spine Technology Handbook*. Elsevier Academic Press, Amsterdam, NL, 1st edition, 2006.

M. W. K. Law, K. Tay, A. Leung, G. Garvin, and Shuo Li. Intervertebral disc segmentation in MR images using anisotropic oriented flux. *Med. Imag. Anal.*, 17 (1):43–61, 2013.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Adv. Neural Info. Process. (NIPS)*, volume 19, pages 801–808, 2006.

V. Lempitsky, A. Blake, and C. Rother. Branch-and-Mincut: Global Optimization for Image Segmentation with High-Level Priors. *J. Math. Imag. Vis.*, 44(3):315–329, 2012.

P. H. Lim, U. Bagci, and L. Bai. Introducing Willmore Flow Into Level Set Segmentation of Spinal Vertebrae. *IEEE Trans. Biomed. Eng.*, 60(1):115–122, 2013.

D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

A. Lucchi, K. Smith, R. Achanta, V. Lepetit, and P. Fua. A Fully Automated Approach to Segmentation of Irregularly Shaped Cellular Structures in EM Images. In *Med. Imag. Comput. Comput.-Assist. Interv. (MICCAI)*, pages 463–471, 2010.

A. Lucchi, K. Smith, G. Achanta, R. Knott, and P. Fua. Supervoxel-Based Segmentation of Mitochondria in EM Image Stacks With Learned Shape Features. *IEEE Trans. Med. Imag.*, 31(2):474–486, 2012.

J. Ma and L. Lu. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. *Comp. Vis. Imag. Underst.*, 117(1):1072–1083, 2013.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised Dictionary Learning. In *Adv. Neural Info. Process. (NIPS)*, volume 21, pages 1033–1040, 2009.

J. Mairal, F. Bach, and J. Ponce. Sparse Modeling for Image and Vision Processing. *Found. Trend. Comput. Graph. Vis.*, 8(2–3):85–283, 2014.

C. R. Maurer, R. Qi, and V. Raghavan. A Linear Time Algorithm for Computing Exact Euclidean Distance Transforms of Binary Images in Arbitrary Dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):265–270, 2003.

T. McInerney and D. Terzopoulos. Deformable Models in Medical Image Analysis: A Survey. *Med. Imag. Anal.*, 1(2):91–108, 1996.

J. R. Meakin, J. S. Gregory, R. M. Aspden, F. W. Smith, and F. J. Gilbert. The intrinsic shape of the human lumbar spine in the supine, standing and sitting postures: characterization using an active shape model. *J. Anat.*, 215(2):206–211, 2009.

S. K. Michopoulou, L. Costaridou, E. Panagiotopoulos, R. Speller, G. Panayiotakis, and A. Todd-Pokropek. Atlas-Based Segmentation of Degenerated Lumbar Intervertebral Discs from MR Images of the Spine. *IEEE Trans. Biomed. Eng.*, 56(9), 2009.

H. Mizaalian, M. Wels, T. Heimann, M. Kelm, and M. Suehling. Fast and Robust 3D Vertebrae Segmentation using Statistical Shape Model. In *Proc. 35th IEEE Conf. Eng. Med. Biol. Soc.*, pages 3379–3382, 2013.

M. T. Modic and J. S. Ross. Lumbar degenerative disk disease. *Radiology*, 245(1): 43–61, 2007.

D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

A. Neubert, J. Fripp, R. Schwartz, L. Lauer, C. Engstrom, and S. Crozier. Automated 3D Segmentation of Vertebral Bodies and Intervertebral Discs from MRI. In *Int. Conf. Digit. Imag. Comput. Techniq. App. (DICTA)*, pages 19–24, 2011.

C. Nieuwenhuis, E. Toppe, and D. Cremers. A Survey and Comparison of Discrete and Continuous Multi-label Optimization Approaches for the Potts Model. *Int. J. Comput. Vis.*, 104(3):223–240, 2013.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

S. Osher and J. A. Sethian. Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *J. Comput. Phys.*, 79(1): 12–49, 1988.

S. Parisot, W. Wells, S. Chemouny, H. Duffau, and N. Paragios. Uncertainty-Driven Efficiently-Sampled Sparse Graphical Models for Concurrent Tumor Segmentation and Atlas Registration. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 641–648, 2013.

Z. Peng, J. Zhong, W. Wee, and J. Lee. Automated Vertebra Detection and Segmentation from the Whole Spine MR Images. In *Proc. 27th IEEE Annu. Int. Conf. Eng. Biol. Soc.*, pages 2527–2530, 2005.

M. Pereanez, K. Lekadir, I. Castro-Mateos, J. M. Pozo, A. Lazary, and A. F. Frangi. Accurate Segmentation of Vertebral Bodies and Processes Using Statistical Shape Decomposition and Conditional Models. *IEEE Trans. Med. Imag.*, 34(8):1627–1639, 2015.

C. Petitjean and J.-N. Dacher. A review of segmentation methods in short axis cardiac MR images. *Med. Imag. Anal.*, 15(1):169–184, 2011.

C. Petitjean, M. A. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. Ben Ayed, M. Jorge Cardoso, H.-C. Chen, D. Jimenez-Carretero, M. J. Ledesma-Carbayo, C. Davatzikos, J. Doshi, G. Erus, O. M. O. Maier, C. M. S. Nambakhsh, Y. Ou, S. Ourselin, C.-W. Peng, N. S. Peters, T. M. Peters, M. Rajchi, D. Rueckert, A. Santos, W. Shi, C.-W. Wang, H. Wang, and J. Yuan. Right ventricle segmentation from cardiac MRI: A collation study. *Med. Imag. Anal.*, 19(1):187–202, 2015.

J. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 1999. The MIT Press.

T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An Algorithm for Minimizing the Mumford-Shah Functional. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1133–1140, 2009.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, New York, NY, 3rd edition, 2007.

C. Reinbacher, T. Pock, C. Bauer, and H. Bischof. Variational Segmentation of Elongated Volumetric Structures. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 3177–3184, 2010.

T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, Jr. Quo Vadis, Atlas-Based Segmentation? In J. Suri, D. L. Wilson, and S. Laxminarayan, editors, *The Handbook of Medical Image Analysis – Volume III: Registration Mod-*

els, chapter 11, pages 435–486. Kluwer Academic/Plenum Publishers, New York, NY, 2005.

P. Roussouly, S. Gollogly, E. Berthonnaud, H. Labelle, and M. Weidenbaum. The sagittal alignment of the spine and pelvis in the presence of L5-S1 isthmic lysis and low-grade spondylolisthesis. *Spine*, 31(21):2484–2490, 2006.

T. P. C. Schlosser, K. L. Vincken, H. Attrach, H. J. Kuijf, M. A. Viergever, M. M. A. Janssen, and R. M. Castelein. Quantitative analysis of the closure pattern of the neurocentral junction as related to preexistent rotation in the normal immature spine. *Spine J.*, 13(7):756–763, 2013.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, MA, 1st edition, 2002.

B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.*, 10(5):1299–1319, 1998.

A. Sotiras, C. Davatzikos, and N. Paragios. Deformable Medical Image Registration: A Survey. *IEEE Trans. Med. Imag.*, 32(7), 2013.

A. Souza and J. K. Udupa. Automatic landmark selection for active shape models. In M. J. Fitzpatrick and J. M. Reinhardt, editors, *Medical Imaging 2005: Image Processing*, volume 5747 of *Proceedings of the SPIE*, pages 1377–1383. SPIE, 2005. doi: 10.1117/12.595463.

P. Suetens. *Fundamentals of Medical Imaging.* Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A Comparative Study of Energy Minimization Methods for Markov Random Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6): 1068–1080, 2008.

W. E. Thong, H. Labelle, J. Shen, S. Parent, and S. Kadoury. Stacked Auto-encoders for Classification of 3D Spine Models in Adolescent Idiopathic Scoliosis. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lect. Notes Comput. Vis. and Biomech.*, pages 13–25. Springer, 2015.

P. H. S. Torr and A. Zisserman. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Comp. Vis. Imag. Underst.*, 78(1):138–156, 2000.

A. Vedaldi and B. Fulkerson. VLFeat: An Open and Portable Library of Computer Vision Algorithms. `http://www.vlfeat.org/`, 2008.

A. Vedaldi and S. Soatto. Quick Shift and Kernel Methods for Mode Seeking. In *Proc. Euro. Conf. Comput. Vis. (ECCV)*, pages 705–718, 2008.

P. Vincent, H. Larochelle, I. Lojoie, Y. Bengio, and P. Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

P. Violas, E. Estivalezes, J. Briot, J. Sales de Gauzy, and P. Swider. Objective quantification of intervertebral disc volume properties using MRI in idiopathic scoliosis surgery. *Magn. Reson. Imag.*, 25(3):386–391, 2007.

T. Vrtovec, F. Pernus, and B. Likar. Determination of axial vertebral rotation in MR images: comparison of four manual and a computerized method. *Eur. Spine J.*, 19(5):774–781, 2010.

C. Wang, N. Komodakis, and N. Paragios. Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey. *Comput. Vis. Imag. Underst.*, 117(1):1610–1627, 2013.

Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li. Regression Segmentation for M^3 Spinal Images. *IEEE Trans. Med. Imag.*, 34(8):1640–1648, 2015.

J. Weickert. Coherence-Enhancing Diffusion Filtering. *Int. J. Comput. Vis.*, 31 (2–3):111–127, 1999.

K. Q. Weinberger and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.*, 10:207–244, 2009.

T. Whitmarsh, L. Del Rio Barquero, S. Di Gregorio, J. Sierra, L. Humbert, and A. Frangi. Age-Related Changes in Vertebral Morphometry by Statistical Shape Analysis. In *Mesh Processing in Medical Image Analysis*, volume 7599 of *Lect. Notes Comput. Sci.*, pages 30–39. Springer, 2012.

O. Wirjadi. Survey of 3D image segmentation methods. Technical Report 123, Fraunhofer-Institut fur Techno- und Wirtschaftsmathematik, 2007.

D. J. Withey and Z. J. Koles. Medical Image Segmentation: Methods and Software. *2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*, pages 140–143, 2007.

T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.*, 5:975–1005, 2004.

J. Yao, J. E. Burns, H. Munoz, and R. M. Summers. Detection of Vertebral Body

Fractures Based on Cortical Shell Unwrapping. In *Med. Imag. Comput. Comput.-Assist. Interv. (MICCAI)*, volume 7512, pages 509–516. Springer, 2012.

L. Younes. *Shapes and Diffeomorphisms*. Springer, New York, NY, 1st edition, 2010.

P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and Gerig G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006. Software available at `http://www.itksnap.org`.

Y. Zhang, M. Brady, and S. M. Smith. Segmentation of Brain MR Images through a Hidden Markov Random Field Model and the Expectation Maximization Algorithm. *IEEE Trans. Med. Imag.*, 20(1):45–57, 2001.

G. Zheng, C. Chengwen, B. Ibragimov, R. Korez, T. Vrtovec, H. Hutt, R. Everson, J. Meakin, I. Lopez Andrade, B. Glocker, H. Chen, Q. Dou, P.-A. Heng, C. Wang, D. Forsberg, A. Neubert, J. Fripp, M. Urschler, D. Stern, M. Wimmer, A. A. Novikov, D. L. Belav, H. Cheng, G. Armbrecht, D. Felsenberg, and S. Li. Evaluation and Comparison of 3D Intervertebral Disc Localization and Segmentation Methods for 3D T2 MRI Data: A Grand Challenge. *Med. Imag. Anal.*, 2016. forthcoming.

Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu. Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. *IEEE Trans. Med. Imag.*, 27(11): 1668–1681, 2008.

S. C. Zhu and A. Yuille. Region competition: unifying snakes, region growing, energy/Bayes/MDL for multi-band image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(9):884–900, 1996.

D. Zukic, A. Vlask, T. Dukatz, J. Egger, D. Hornek, C. Nimsky, and A. Kolb. Segmentation of Vertebral Bodies in MR Images. In *Vis. Model. Vis. (VMV)*, pages 135–142. Eurograph. Assoc., 2012.