Automated Radiological Analysis of Spinal MRI

D.Phil Thesis

Robotics Research Group
Department of Engineering Science
University of Oxford

Supervisors:
Professor Andrew Zisserman
Doctor Timor Kadir

Meelis Lootus
Christ Church

Hilary Term, 2015

Meelis Lootus
Christ Church

Doctor of Philosophy
Hilary Term, 2015

Automated Radiological Analysis of Spinal MRI

Abstract

This thesis addresses the problem of analysing clinical MRI using modern computer vision methods for a variety of clinical and research-related tasks. We use automated machine learning algorithms to develop a spinal MRI analysis framework for a number of tasks such as vertebrae detection, labelling; disc and vertebrae segmentation, and radiological grading, and we validate the framework on a large, heterogeneous dataset of 300 symptomatic back pain patients from multiple clinical sites and scanners. Our framework has a number of back pain research and other spine-related clinical applications and could hopefully find application in a clinical workflow in the future.

Our framework has five steps – detection, labelling, segmentation, support regions and features, and machine learning for radiological measurements. The framework works in full 3D and has currently been implemented on sagittal T2 slices. We use Deformable Part Models along with a chain model to detect and label vertebrae, and a powerful graph cuts based method for vertebrae and disc segmentation. The labelled detections and segmentations are used to place support regions for feature extraction, which are mapped into a number of radiological measurements – namely Pfirrmann grade, disc space narrowing, and herniation/bulge. The radiological ground truth was provided by a clinical radiologist with 25 years experience. We demonstrate a high performance in the measurement in each. The measurements are performed using support vector machines and support vector regressors learned on training data.

We next investigate the problem of what is the best method of obtaining support regions. We first used pixel intensity features to predict the Pfirrmann grade, narrowing and bulge/herniation, with vertebrae segmentation to localise their support regions. Since segmentation of spine images, especially intervertebral discs is an unsolved problem and algorithms are prone to failure, we then ask the question, to segment or not to segment. To answer the question, we compare results on Pfirrmann grade prediction with three different points on the no segmentation to full disc segmentation involving no segmentation, vertebrae segmentation, or disc segmentation and find that vertebrae segmentation suffices.

We finally show preliminary results in distinguishing between different radiological conditions related to the posterior side of the disc more finely than before in literature, taking information from both sagittal and axial slices to attempt to distinguish between herniated and bulged discs.

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Meelis Lootus, Christ Church

Acknowledgements

I am very grateful to my supervisors Professor Andrew Zisserman and Doctor Timor Kadir. I would like to thank them for the continuous support of my DPhil; for their passion, optimism, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my DPhil study.

I would like to also thank Dr. Vaclav Potesil for sharing his code and his guidance at the beginning stages of my DPhil. I am very thankful for Dr. Jill Urban, Professor Jeremy Fairbank, and Professor Iain McCall for their support as clinical collaborators, and their enthusiasm for the project.

The environment in the lab in Oxford has been fantastic due to the people in VGG. My thanks go to all the lab members, in particular Amr, Relja, Ken, Max, Elliot, Omkar, Tomas, Karel, Carlos, Chai, Yusuf, Mircea, Arpit, Victor, Eric, Lubor, Amir, Aravindh, and Varun. I would also like to thank my incredible housemates, Georg and Himadri as well as the amazing people of Christ Church MCR. Most importantly, my thanks go to my family for their years of support, motivation, and belief.

Contents

Chapter 1

Introduction

Lower back pain is a global cause of life-long disability and is one of the leading causes of lost productivity (Fairbank and Pynsent [2000]). Chronic lower back pain patients require significant resources and are responsible for 80% of the costs attributed to back pain. Imaging, principally through the use of spinal MRI, is routine and is a key element of management of such patients, though, its efficacy in informing patient stratification and prognosis has not been demonstrated and is being increasingly questioned (Lurie et al. [2013], Modic and Ross [2007], Steffens et al. [2013], Videman et al. [2003]). Despite this, there are clinical situations in which MRI is a useful diagnostic tool, for instance with patients with suspected spinal stenosis or disc herniation, or tumours.

A key challenge in the application of imaging in the management of back pain patients has been the lack of an established link between reported radiological image findings and clinical symptoms such as pain and mobility scores or patient outcome post-intervention. One reason for this is the variability and imprecision of such clinical data, which are typically collected through scored questionnaires. Another is the qualitative nature of the radiological report which might consist of a short description of the principal findings that the radiologist feels are abnormal or otherwise

worthy of note along with some indication of the disc health with perhaps a note of its grade. Such an approach is subject to a large degree of intra- and inter-radiologist variation (Mulconrey et al. [2006], Jarvik and Deyo [2009]) and this, combined with the variability of the clinical findings, has hampered efforts to utilise imaging more broadly in patient management.

However, despite its limitations, MRI is still considered a mainstay of clinical practice and it is still believed that MRI could be used to better stratify patients and predict outcome, hence there is on-going research in this, with some promising results (Jensen et al. [2013]).

One approach to overcoming spinal MRI's shortcomings, is to address radiologist intra- and inter-observer variability through the use of quantitative techniques. Indeed, in other areas of radiological reporting, so-called quantitative reading has become the mainstay at least in clinical research. For example, the RECIST guidelines (Eisenhauer et al. [2009]) utilise standard quantitative measures of changes in tumour size to report response to therapy. Within the spinal MRI specialty, Pfirrmann grades are routinely reported in many centres. Therefore in keeping with much of radiology there is a need for, and growing recognition that, quantitative techniques that extract multiple measurements from the images have the potential to significantly improve stratification and hence management.

Unfortunately, manual extraction of relevant measurements is impractically laborious and, notwithstanding the noted benefits over purely qualitative reading, is still subject to intra- and inter-reader variability. Automated and semi-automated computer vision techniques can be utilised to address both concerns and consequently there is also a growing interest in the medical vision community in analysis of such imaging including the organisation of specialist workshops and special journal issues.

Within the area of MRI spine imaging, there are a number of medical vision problems that are of current research interest including: detection, labelling, and

segmentation of vertebrae and discs, characterisation of local and global appearance, vertebrae position and configuration, and finally, mapping to clinical findings. However, the nature of the task is such that research groups have to either develop entire processing pipelines from detection to classification, potentially solving problems that are common to all groups and re-developing existing work, e.g. detection and labelling of vertebrae, or they solve specific problems, but rely on manual methods for the rest of the pipeline, e.g. manual placement of ROIs for disc grading. We believe that this is holding back the community. And this is the focus of this thesis – to automate the radiological analysis of spinal MRI. Rather than focus on just one task, we provide a framework to handle a number of spine analysis tasks – detection, labelling, segmentation, and radiological measurements.

1.1 Measurements & Challenges of Spinal MRI

In this section, we discuss a range of quantitative image parameters that are of interest and how they might be related to the underlying pathology. We also discuss the challenges in automatically extracting such measurements.

1.1.1 Definition/Review of Measurements

A broad range of radiological measurements have been proposed in the literature, and we present a non-exhaustive summary; a number of measurements are illustrated in Figure 1.1. Most measurements attempt to capture either the image signal or spine shape globally or locally. But many focus on different parts of the spine such as the disc, endplates, vertebral body (VB), facet joints, foramina, the spinal cord. Some are measured continuously, such as disc intensity, spinal cord cross section, and disc height; others, quantised in a number of gradings, for instance using the Pfirrmann grade (Pfirrmann et al. [2001]). In fact, the Pfirrmann grade, illustrated in Figure 1.2, combines changes in the intensity of the nucleus pulposus

simultaneously with disc space narrowing.

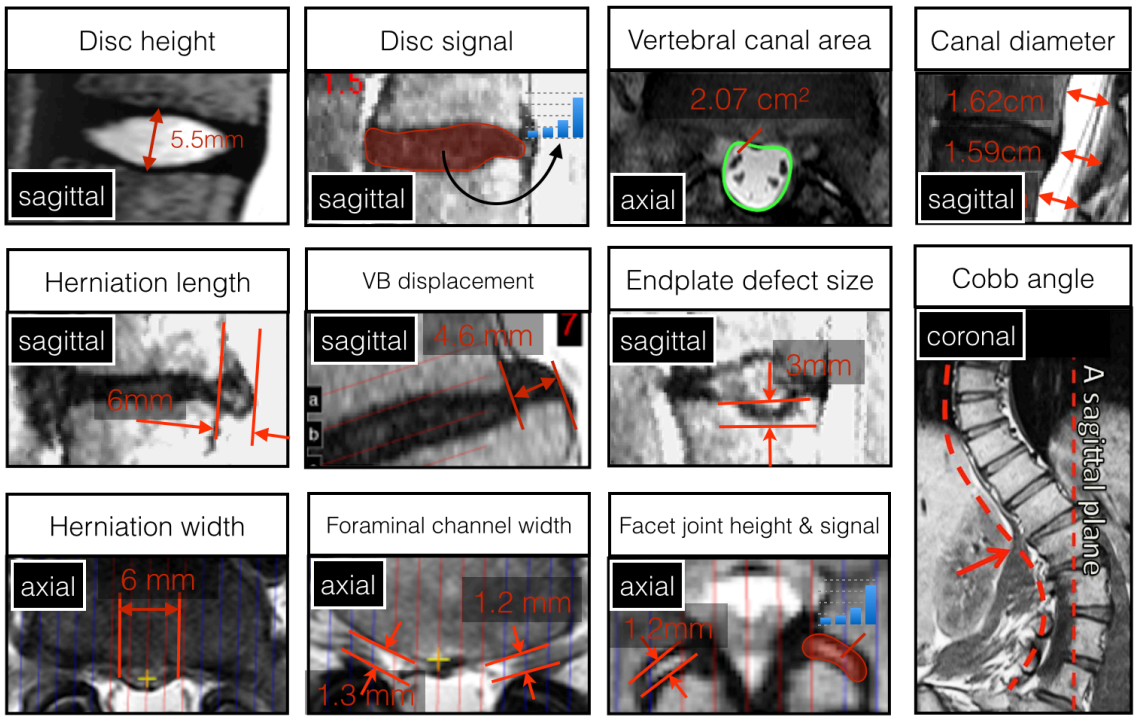


Figure 1.1: **An assortment of radiological MRI measurements**. All protocols are variations of T2 in multiple different machines.

Beyond the Pfirrmann grade, grading systems exist for so-called Modic Changes, pathological changes in the Vertebral Body (VB) marrow (Modic et al. [1988]), and the Facet Joints (Weishaupt et al. [1999]). The former characterises only *intensity*, whereas the latter both *intensity* features capturing tissue composition changes, and *shape* features capturing deformations (narrowing, etc.). The Cobb angle is used to capture scoliosis, and displacement of neighbouring vertebrae for spondylolisthesis.

Evidently, such measurements are performed to describe the condition of the spine and are intended, ultimately, to correlate with different disease processes – see Figure 1.3. We note that spinal degradation can occur due to a number of different reasons including age, trauma, repeated cyclic stress, genetics, or disc nutrition. Certain physiological processes, for example osteophyte growth, attempt to correct for sub-optimal spine geometry, to fix and support the spine.

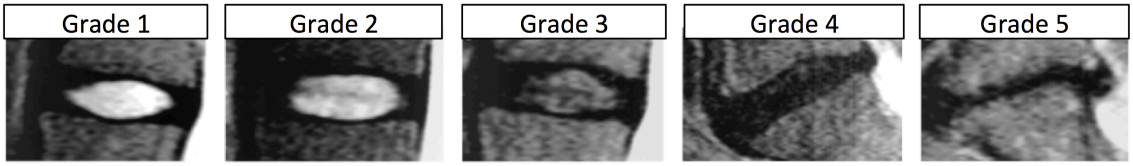


Figure 1.2: **Pfirrmann global disc grade and narrowing**. Disc degeneration process. Note that in the case of these particular disc examples, the degeneration process involves T2 signal loss, both posterior and anterior **bulging** (e.g. in Grade 5 example), and disc space **narrowing** (e.g. Grade 5 example), which are also measured separately.

Another class of disc related degenerations are Herniations (Figure 1.4), defined as a localised displacement of disc material beyond the normal margins of the intervertebral disc space (Fardon and Milette [2001]). Most herniations occur in the lower lumbar region, and are often related to back pain. The normal margins of the intervertebral space are defined, craniad and caudad, by the vertebral body endplates and peripherally by the edges of the VB ring apophyses, exclusive of osteophytic formations (Fardon and Milette [2001]). Several example herniations are shown in Figure 1.4. Herniations are categorised in a number of ways. First, according to type as protrusion, extrusion, and sequestration. Secondly, their location as central, posterolateral right/left, or foraminal right/left. Thirdly, according to whether they have displaced the nerve root, which is a predictor for pain. As a point of interest, we note that based on examination of sagittal slices only, bulges can be easily confused with herniations. The difference, according to Fardon et al. [2014], is morphologic, in that a disc bulge is a generalised protrusion. Clinically, as bulges point out less, they do not normally reach far enough to press against nerve roots and cause pain. In contrast, herniations are focal and may cause pain by pressing against nerves. Radiologically, the two can usually be differentiated if both sagittal and axial slices are assessed.

Finally, **Stenosis** is the narrowing of either the spinal canal, or the foraminal canals, and can be characterized by the canal area or width (Figure 1.1). The

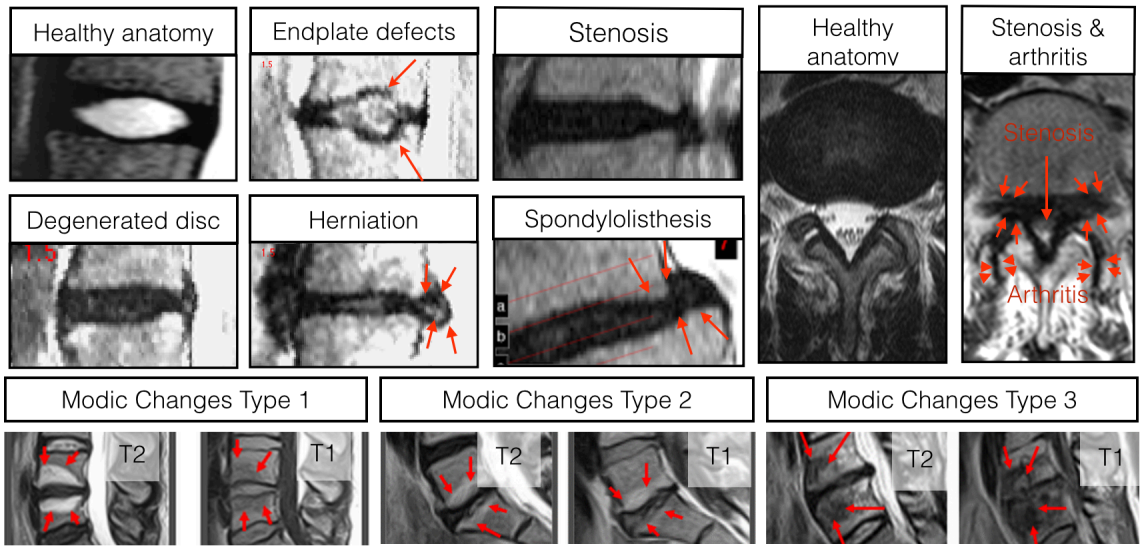


Figure 1.3: **A number of example conditions**. The measurements in Figure 1.1 are performed to capture these (and other) degradation processes. All protocols are variations of T2 in multiple different machines.

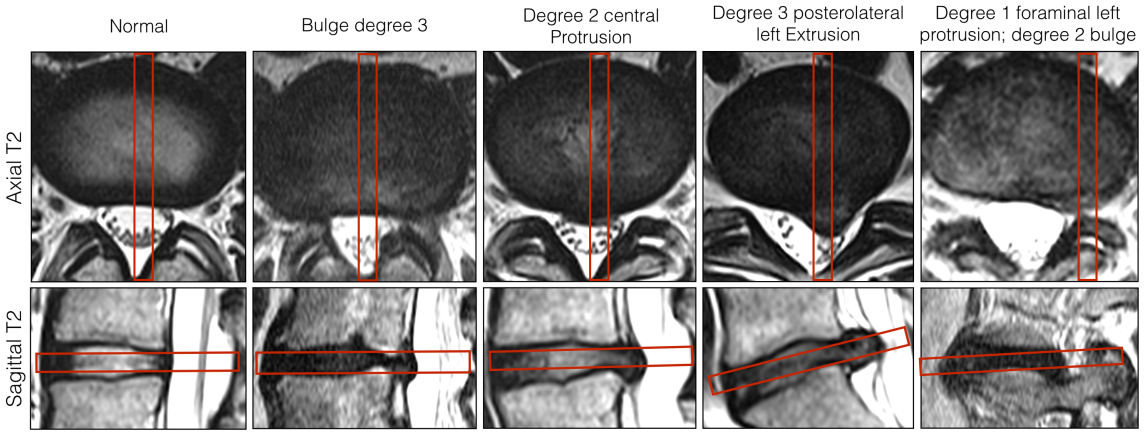


Figure 1.4: **Bulges and herniations**. Each column is a separate disc (from a number of patients). The red boxes on each axial slice note the intersecting sagittal slice, and vice versa. The difference between the bulge and assorted types of herniations is shown. In the case of a normal disc there is no posterior section deformation seen in either axial or sagittal slices. In the bulged case, the posterior side of the disc is seen displaced in a range of sagittal slices. In the herniated cases, the posterior side of the disc presses out in a localized manner. The herniations differ by type (protrusion-extrusion), location (central, posterolateral, foraminal) and degree (1-3), e.g. the mass displaced.

difference between stenosis and bulge/herniation is that with stenosis, the bone is deformed, whereas with bulge/herniation the disc material is deformed and may

cause nerve compression and thus pain.

1.1.2 Challenges of Spinal MRI

There are three principal types of challenge to spinal MRI analysis: first, imaging quality; second, anatomical and pathological patient variation; third, inter-observer variability. They are illustrated in Figure 1.5.

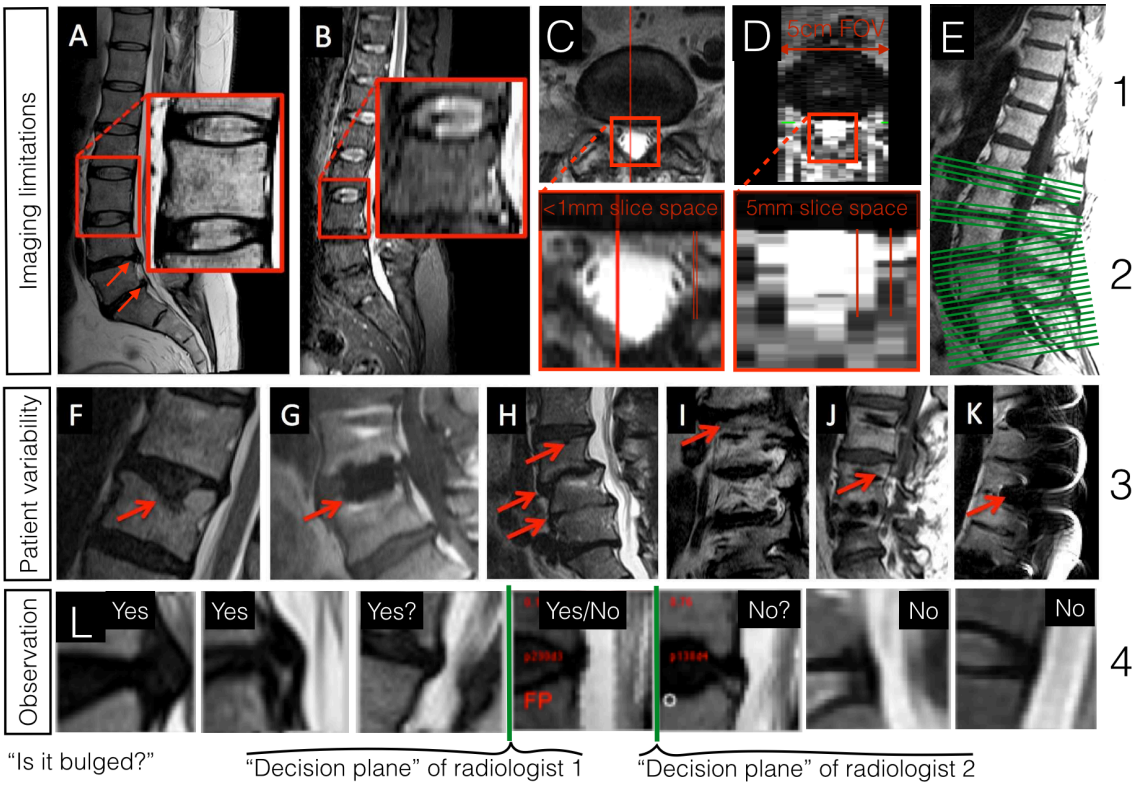


Figure 1.5: **Challenges of radiological spine measurements.** Abnormalities have been highlighted by the red arrows. (A-E) imaging limitations; (F-K) vertebrae shape variations; (L) inter-radiologist variability. Imaging limitations: (A) Normal resolution sagittal image, with a zoom on a vertebra. (B) A low-resolution sagittal image. (C) Axial view of a high-resolution 3D research scan (non-clinical, not in our dataset), along with a zoom on spinal cord. (D) Axial view of a typical standard clinical sagittal scan (in our dataset), along with a zoom on spinal cord. Note both the narrowness of the FOV, and the sparse slice spacing (typically 4-6mm). (E) Typical axial scan lines shown on a sagittal scan. (L) Seven discs of varying degree of bulge are shown, as graded by two hypothetical radiologists. Note that while on some of the discs both radiologists agree as either definitely bulged (yes) or not (no); others are often more open to interpretation – the middle disc in row 4 could either be classified as bulged or normal, depending on the radiologist.

Imaging limitations. Clinical MR images are typically of sparse slice spacing and limited left-to-right Field of View (FOV) (see row 1 in Figure 1.5). Additionally, they tend to comprise low in-slice resolution. This is due to the inherent trade-off between scan acquisition time, signal-to-noise ratio, and image resolution in MRI. As an example, a consequence often might be that, in the case of herniation, the displaced disc material may only be visible as a partial volume, due to the coarse slice spacing. This challenges both the delineation and diagnosis tasks. Another problem is that image appearances also vary widely across scanners and protocols.

Patient variability. The vertebrae, discs, and other anatomical structures in the spine show wide anatomical variation, complicating their detection and delineation. In addition, pathological abnormalities come in many variants and are thus also hard to model and recognise across patients. See row 2 in Figure 1.5 for example variations in vertebra shape.

Observation. Since many conditions exist as a continuous spectrum of severity, it can be unclear where to draw the line when classifying anatomies to normal vs. abnormal. See row 3 in Figure 1.5.

Those aforementioned challenges can impact every part of our proposed pipeline. In the rest of the thesis, those three principal challenges are referred back to when discussing its performance.

1.2 A Framework for Image Analysis of Spinal MRI

In this section, we present an overview of the image analysis framework explained further throughout this thesis. The proposed framework leads from clinical MRI to radiological measurements using methods from the modern computer vision community. The framework is also useful for various other tasks such as anatomy detection,

segmentation, and labelling.

The framework has five steps, as illustrated in Figure 1.6: **(1)** detection, **(2)** labelling, **(3)** segmentation, **(4)** support regions extraction and features, **(5)** mapping to measurements. In essence, steps 1–3 are performed in order to arrive at relevant image support regions. Steps 4–5 focus on the characterisation and classification of those regions, using machine learning.

The **first (detection)** step takes as input a sagittal MRI scan of the spine comprising a widely variable number of slices and resolution, and outputs detection candidates (tight bounding boxes) for the VBs, searching over position, scale, and angle. The detections are performed using a sliding window vertebrae detector in each of the 2D slices independently. The candidate detections from all slices are passed onto the **second (labelling)** step that takes them as input and finds the optimal layout of the spine in 3D according to the detector scores and vertebrae relative position, angle, and size, and labels the detections. From these, the **third (segmentation)** step then outputs vertebrae segmentations (delineated vertebrae) in each of the slices independently. This is performed to produce more precise localisations for the subsequent steps. At this stage the global layout of the spine becomes known. It is also possible to make 'global' measurements such as torsion and scoliosis, though we do not focus on these in this thesis. Instead, the subsequent steps target more local regions for measurement.

The **fourth (support region)** step defines a number of support regions based on the previous steps, and outputs image features suitable for predicting radiological measurements and gradings. Finally, the **fifth (prediction)** step takes as input the feature vector(s) from step four, and maps them to measurements and gradings of interest using machine learning techniques such as classification, regression and ranking. Importantly, to reflect clinical reality where no one strict image acquisition standard is adhered to across hospital sites, the framework is flexible to and tested

on data acquired from multiple scanners and protocols.

1.3 Thesis layout

- In **Chapter 2**, we provide background on spine anatomy and MRI, along with a literature review on previous image analysis techniques.

- In **Chapter 3**, we introduce the dataset on which all our experiments are performed. In the following three chapters, we introduce parts of our image analysis pipeline and explain them in more detail, along with experimental results on the dataset described in Chapter 3.

- In **Chapter 4**, we present a method for automated vertebral body detection and labelling, using the Deformable Part Model (DPM) sliding window detector along with a chain graphical model.

- In **Chapter 5**, we segment the vertebrae and discs using a powerful Graph Cuts method. The segmentations are initialised with the detections from Chapter 4.

- In **Chapter 6**, we use machine learning methods to automatically perform three different radiological measurements: Pfirrmann grade, disc narrowing, and herniation/bulge. The feature support regions for measurement prediction are initialised using the detections and segmentations from the previous Chapters. We compare those results to ones obtained using manually placed support regions. We also ask the question, "to segment or not to segment?", and we answer the question by experimenting with support regions obtained using various levels of segmentation.

- In **Chapter 7**, we conclude the thesis, listing our contributions and discussing potential future work.

1.4 Publications

The work presented in this thesis has been published in the following conferences:

- Lootus, Meelis, Timor Kadir, and Andrew Zisserman. "Vertebrae detection and labelling in lumbar MR images." In MICCAI workshop Computational methods and clinical applications for spine imaging, pp. 219-230. Springer International Publishing, 2014

- Lootus, Meelis, Timor Kadir, and Andrew Zisserman."Automated Radiological Grading of Spinal MRI." In MICCAI workshop Recent Advances in Computational Methods and Clinical Applications for Spine Imaging. Springer International Publishing, 2015. 119-130.

These papers are extended in a journal paper currently under review in Medical Image Analysis journal:

- Lootus, Meelis, Timor Kadir, and Andrew Zisserman. "A Framework for the Automated Analysis of Spinal MRI" In Medical Image Analysis (under review).

The extensions in the journal paper include: (i) a comprehensive survey of the literature; (ii) a more stringent success criteria for detection; (iii) new patch features used for the radiological measurement; and (iv) prediction of two additional measurements – herniation/bulge, and narrowing.

In addition, work on automated landmark localisation in CT scans, not presented in this thesis, has been published in ISBI 2013, in

- Potesil, V., Lootus, M., El-Labban, A., and Kadir, T. (2013, April). "Landmark localization in images with variable Field-Of-View". In Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on (pp. 1046-1049). IEEE.

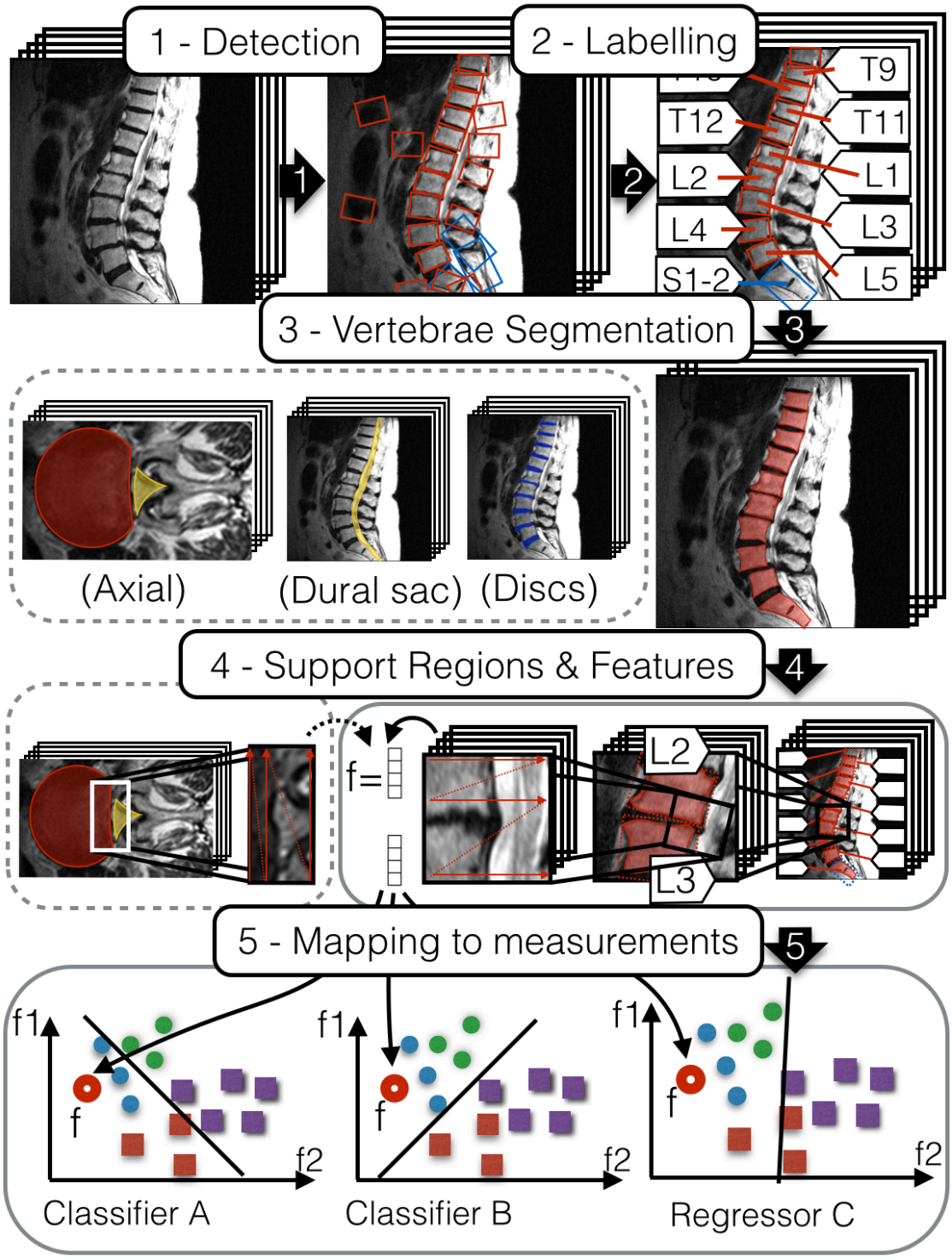


Figure 1.6: **Our fully automatic pipeline**. The pipeline has five core steps (boxes with solid boundaries), with some optional additions (boxes with dashed boundaries).

Chapter 2

Literature Review and Background

This Chapter has three parts. We first provide a general anatomical background to the spine in Section 2.1. Next, we discuss imaging background (radiograph, CT, MRI protocols) in Section 2.2, and finally a literature review on automation of spine image analysis in Section 2.3.

2.1 Background on Spine Anatomy

The spine anatomy is illustrated in Figure 2.1. In total, there are 33 **vertebrae** in the normal human spine. The upper 24 are articulating vertebrae, separated by **intervertebral discs** (IVD), while the lower 9 are fused: five in the sacrum, and four in the coccyx. The spine acts as a structural element in the body, and provides the **spinal canal**, a housing and protection to the spinal cord (SC).

The vertebrae are grouped according to the body regions as seven cervical (C1-C7), twelve thoracic (T1-T12), five lumbar (L1-L5), and nine sacral links (S1-S9), as indicated in Figure 2.1 A. The IVD-s are labelled according to their neighbouring vertebrae, e.g. "L1-L2" for the IVD between L1 and L2.

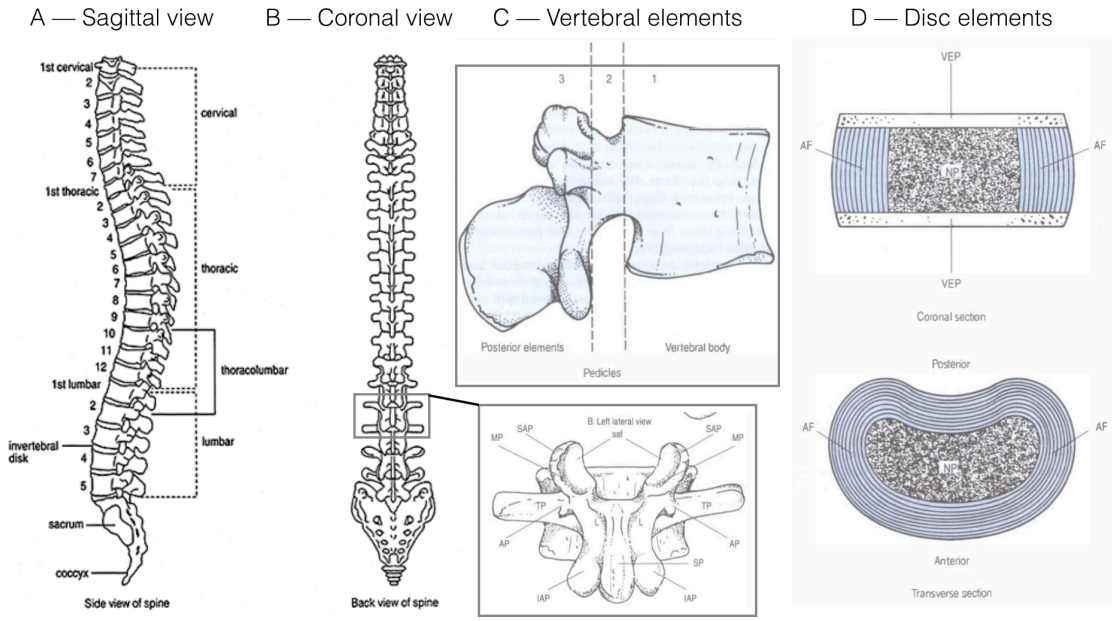


Figure 2.1: **Spine anatomy.** The sections of the spine (cervical, thoracic, lumbar, sacral) are shown on a global spine layout (A,B), along with detailed vertebra and disc anatomy for a single vertebra /disc in sagittal and coronal views in (B,C). The images in A, B are from the websiteful [2015], and the images in C, D are from Bogduk [2005].

The vertebra has a more complex shape than the IVD. It is made of three functional components: the vertebral body (VB), the pedicles, and the posterior elements. The VB subserves the weight-bearing function, the posterior elements collectively form a very irregular mass of bone and help control twisting and rotating motions, and the pedicles connect the VB to the posterior structures, forming a protective channel for the spinal cord. The vertebral anatomy is illustrated in Figure 2.1 C in sagittal and coronal views.

Any two consecutive articulated vertebrae form a three-joint complex: the IVD between the VB-s, and the two facet joints (FJ) between specific posterior elements. The IVD acts both as glue, a cushion, and a pivot point. The IVD is made of two components: a central nucleus pulposus (NP) surrounded by a peripheral annulus fibrosus (AF). Although they are quite distinct, there is no clear boundary between them in an MRI. The disc anatomy is illustrated in Figure 2.1 D in sagittal and

coronal views.

A third component of the IVD comprises two layers of cartilage which cover the top and bottom aspects of the IVD – each is called a vertebral endplate (EP). They separate the vertebrae from the IVDs, and hence can be considered either part of the vertebra or the IVD.

In IVD-s of young, healthy adults, the NP is a semifluid mass of mucoid material, with the consistency of toothpaste. Histologically, it consists of a few cartilage cells and some irregularly arranged collagen fibres, dispersed in a medium of semifluid ground substance. Biomechanically, the NP can deform under pressure but cannot be expanded or compressed. The AF surrounding the NP acts as a response to that pressure. The AF consists of collagen fibres in a highly ordered pattern arranged in concentric rings.

The **spinal cord** (SC) is the connection between the brain and the body, and continues from the skull through to around the L1 level, where it splits into multiple fibres, and is referred to as the dural sac (DS) below the split. In an MR image, the SC is easier to distinguish from the background than the DS.

Finally, ligaments connect to both the vertebrae and the IVDs, joining them and connecting to the muscles too to move them. The spine is surrounded by blood vessels, and other organs in the body.

2.2 Background on Spine Imaging

Imaging provides a window into the patient anatomy. It can be used clinically (e.g. to diagnose disease and select treatment), in research (e.g. to understand diseases and develop with new treatments), and for navigation (e.g. surgery planning).

Imaging modalities. In spine imaging, the most commonly used modalities are the radiograph (informally also called X-ray; in Spine Imaging DEXA in particular), Computed Tomography (CT), and MRI. These modalities have important

	X-ray	CT	MRI
Diagnostic test for	Bone issues	Bone issues	Soft tissue & bone issues
Cost	Low	Lower than MRI	High
Acquisition time	Fraction of a second	Tens of seconds	Several, up to tens of minutes
Resolution	Fine	Often fine (trade-off with radiation dose)	Often coarse (trade-off with acquisition time and SNR)
Radiation dose (equivalent time of background radiation)	Several months	Half, up to several years	None
Suitable for patients with metal implants?	Yes	Yes	No
Left-right FOV	Various	Full	Low (trade-off with acquisition time and SNR)

Table 2.1: **X-ray vs. CT vs. MRI**. Advantages and disadvantages of each modality.

differences overviewed in Table 2.1. In addition, PET is sometimes used to find tumors.

The **radiograph** (in particular DEXA in spine imaging) is the most common diagnostic test, largely due to its low price and ease of acquisition. It is also acquired in fraction of a second and at high resolution (voxel size in tens of micrometres). Its view into the anatomy is limited however, providing only a radiographic shadow of the anatomy according to a projection, rather than a 3D view, and resolves best the bones (though, with inadequate contrast). Thus, it is commonly used as the first diagnostic imaging test to check for serious bone issues. Radiographs are usually taken either projected in the sagittal, or sometimes in coronal plane, or at times both (Shen et al. [2013]). Compared to CT, a radiograph has very low dose but compared to MRI it has more dose. The radiograph dose is typically equivalent to multiple months of natural background radiation.

The **CT** also operates in the X-ray modality. The difference from the radiograph is that it is a 3D volume, usually acquired using a small field of view detectors moved in a helical pattern around the patient as they are pushed through the scanner. The CT is also lower resolution than radiograph. CT voxels are typically $1 \times 2 \times 2$mm (though, resolution can vary considerably and is in trade-off with radiation dose). CT scans are acquired in tens of seconds (thus, showing minimal motion artefacts

in comparison to MRI). Like the radiograph, a CT is limited in its diagnostic ability. It is useful mainly for assessment of bone issues such as vertebral fractures, however is not sensitive to soft and cartilaginous tissue issues (e.g. disc herniations, degeneration). However unlike X-ray, it provides a full 3D view of anatomy. In fact, bone (e.g. vertebrae) segmentation can be easier in CT than in MRI due to higher bone edge visibility, and due to a standard relationship between the tissue type and image intensity (Hounsfield units). Another disadvantage besides soft tissue diagnostic capabilities of the CT is a very high radiation dose in comparison to MRI and X-ray (a single several-second acquisition can be equivalent to several years of natural background radiation).

The **MRI** operates based on the excitation and measurement of radiofrequency pulses absorbed and emitted by hydrogen atoms in the body. The acquisition mechanism is much more complicated than for X-ray and CT, requiring sophisticated apparatus. It is also typically more time-consuming (protocol-dependent, usually tens of minutes per patient for spine scan protocols), and expensive. In MRI, there is no standard correspondence between the voxel intensity and the tissue type, and the image contrast can be dramatically manipulated by the acquisition parameters. The advantages of MRI are its ability to diagnose a broad range of soft-tissue conditions, and its safety to the patient – no radiation dose whatsoever. The only health and safety issue is that patients with metal implants cannot be placed in the machine due to the magnetic force exerted on metals in a strong magnetic field. Also since the acquisition time is in trade-off with resolution, signal-to-noise ratio (SNR), and the field of view (FOV), the MRI scans are often acquired as thick sagittal and axial slices, rather than dense 3D volumes, and cover only the minimal necessary left-to-right FOV, e.g. the disc or the VB width (leaving out the spinal processes), in order to keep the acquisition time reasonable. Long acquisition times both reduce the number of patients that can be diagnosed, and increase the likelihood of

artefacts from patient motion.

Patient position. In CT and MRI, the patient generally lies on their back, in supine position. In radiograph, the patient might be standing. Since the patient position affects the configuration of the spine, some conditions such as bulge/herniation may go unnoticed in lying patient position as noted and demonstrated in Alyas et al. [2008]. In addition, there may be a tendency to place the patient in such position as to minimise pain to enable them to stay still for the long acquisition times, and thus the causes for pain may not show in the acquisition.

The scanner coordinate system. Within the scanner, the directions used in medical imaging are defined according to the patient position as Superior-Inferior (S-I) head-to-toe, Left-Right (L-R), and Anterior-Posterior (A-P) front-to-back. These conventions will be used through the thesis. In case of multi-series studies, all the series in the study are generally expressed in the same coordinate system. If the patient is perfectly still between the acquisitions, image annotations can be transformed between the series using coordinate transformations.

MRI imaging protocols. Out of X-ray, CT, and MRI, MRI is arguably the most complex and involved imaging method, providing also the most opportunities for variation. An MR image's appearance is influenced by

- **Design of the scanner**, including the magnetic field strength (typically between 0.3 and 3 Tesla) and the imaging coils design. Stronger magnets can generally produce crisper images yet they have smaller enclosures and are thus less suitable for patients who suffer from claustrophobia, or are obese.

- **Selection of the imaging pulse sequences**. Given a specific MRI scanner model, the image appearance (signal scale for tissue types) can be influenced by the radiofrequency pulse sequence production and recording. Parameters such as TE, TR, and spin angle characterise the protocols. Pulse generation

is a scientific field of its own and many protocols exist but the most standard ones are the T1, T2, and PD (proton density). In T2, water is bright; in T1 it is dark. Fat is bright in both T1 and T2 (Weissmann [2009]). Note that the acquisition time also varies depending on the protocol (e.g. it generally takes longer to acquire a T2 scan compared to T1). The appropriate imaging sequence is selected based on the clinical task dealt with and the scanner.

- **Acquisition matrix selection: 2D slices vs. 3D volume matrix**. In MRI, there is a trade-off between the spatial resolution, signal-to-noise ratio, and acquisition time. Hence, clinical spine MRI scans are typically acquired as multiple thick slices (usually sagittal and axial) rather than (isotropic) 3D volume. See Figure 2.2 for illustration. This is motivated by the aim to achieve a sufficient SNR for diagnosis at a reasonable acquisition time (already around 20 minutes for a typical set of series for a patient). The disadvantage of this acquisition are the partial volumes from edges lateral to the image slice plane – note that the voxels are like long sticks (up to 10x length to width ratio). The advantage of 3D acquisition is the higher resolution (and thus, less partial volumes), and the disadvantage is the lower SNR, harder to detect edges, and longer acquisition times (Neubert et al. [2012]).

- **Selection of the imaging resolution and planes placement**. The planes are typically placed according to the task, to show the anatomy in the most informative way. Higher resolution will typically have lower SNR for a given scanner, however lower resolution scans will show more partial volumes. Planes placement similarly influences partial volumes – when placed lateral to tissue boundaries, more partial volumes are produced.

- **Artefacts**. MRI can suffer from a number of artefacts – motion, flow, wrap-around, chemical shift, susceptibility, and magic angle. These complicate all

the image analysis tasks, both manual and automatic.

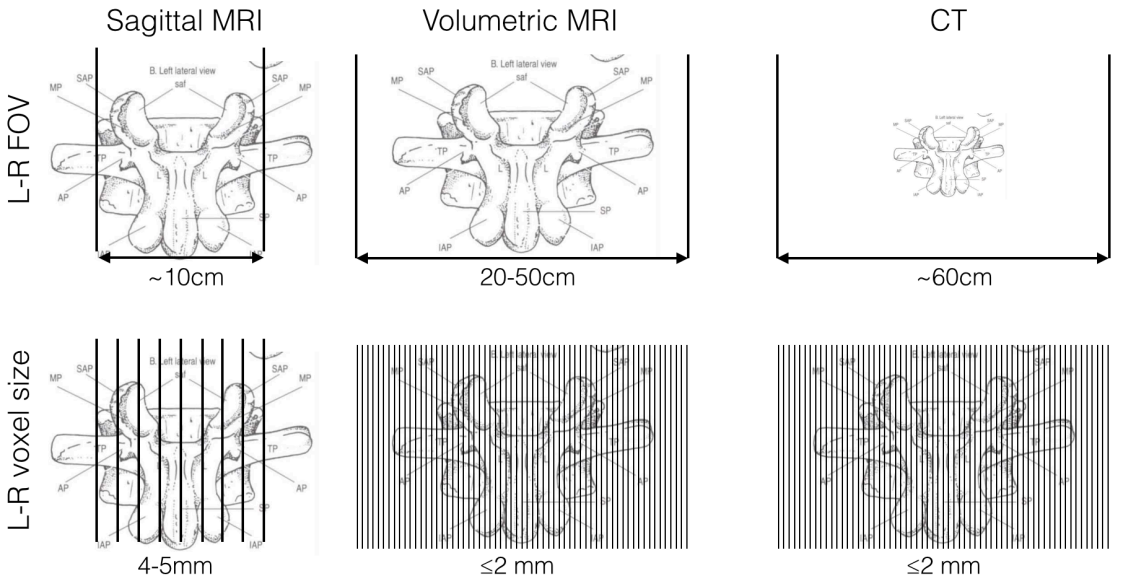


Figure 2.2: **Resolution and FOV comparison.** The typical FOV and left-right (L-R) voxel size for slice-based sagittal MRI, volumetric MRI, and CT are shown side-by-side to illustrate the differences. Note that the typical sagittal MRI is much narrower FOV and higher slice thickness than the other scan types.

In MRI for spine imaging, the most common protocols are the T1- and T2-weighted scans. T2 scan better visualises the water content of tissue, whereas T1 scan is used additionally for some conditions such as Modic changes. Note the difference between T1-weighted and T1 scans: T1-weighted scans only refer to the pulse sequence used to acquire the scan, whereas in T1 scans, the voxel intensity value equals the actual T1-relaxation time. The latter scans require reconstructions involving multiple acquisitions.

CT and radiograph acquisition protocols and appearance are fairly standard, varying mainly in the resolution (in trade-off with radiation dose and SNR) and the FOV. The CT left-to-right, anterior-posterior FOV tends to cover the entire patient, while the radiograph FOV can be alternated more precisely according to the clinical need.

Diagnosis of spine conditions requires specialised knowledge of the imaging pro-

tocols. The imaging protocols show great variation in both resolution, FOV (L-R, A-P, and S-I), and contrast profiles. The imaging protocol of a particular scan has impact on the vision algorithms applicable to this particular scan, as discussed in the following literature review.

2.3 Automated Spinal Image Analysis

There are two aspects to automated Spine Image Analysis: anatomy localisation, and its parametrisation for disease state prediction. We present a summary of literature in two tables related to the two: Table 2.2 for localisation and Table 2.3 for characterisation. In these, properties of both the task, the strategy (the method), its implementation (e.g. is it automatic?), and validation (e.g. dataset size, modality, FOV, etc.) are covered.

In a fully automated pipeline, the localisation stage generally starts with detection and disc/vertebrae labelling, sometimes involves segmentation initialised by the detections. The localisation of the anatomy can then be used to initialise region placement and feature extraction for disease state prediction using machine learning methods in the parametrisation stage for conditions such as disc herniation, degeneration, etc.

Segmentation sometimes precedes labelling as in Ma and Lu [2013]. And, sometimes the support regions are placed manually, as in Koh et al. [2012]. In those cases, only the disease state parametrisation is automated using machine learning methods. At other times, in semi-automated cases, somewhat more limited manual input may be required, such as point placement for localisation initialisation as in Štern et al. [2011]. Sometimes, experiments are performed on small datasets, as segmentation in Lu et al. [2012] on 4 scans, or degeneration in Unal et al. [2011] on 9 scans. A two-digit number of scans is common. The scans can also be of high quality, more reflective of research than everyday clinical protocols as in Neubert

No	Paper	Scans	Mod.	Approach	Structures		Shape	Label	S-I FOV constraints	Manual input
					Det	Seg				
1	Alomari et al. [2011a]	105	MR	Probabilistic model	D	–	–	✓	Fix to T12-S1	Auto
2	Corso et al. [2008]	21	MR	Probabilistic model	D	–	–	✓	Fix to T12-S1	Auto
3	Ghosh et al. [2012]	53	MR	HOG + Heuristics	D	–	–	✗	Axial heuristics	
4	Oktay and Akgul [2013]	80	MR	HOG + GM	V-D	–	–	✓	Contain T12-S1	Auto
5	Lootus et al. [2013]	371	MR	HOG + GM	V	–	–	✓	Contain S1-2	Auto
6	Wu et al. [2011]	619	CT	Orthogonal matching pursuit	V	–	–	✓		
7	Zhan et al. [2012]	15+	MR	Adaboost+local articulated model	V-D	–	–	✓	Contain anchor	Auto
8	Zhan et al. [2012]	300	MR	Adaboost+local articulated model	V-D	–	–	✓	Contain anchor	Auto
9	Vrtovec et al. [2005]	–	–	–	c-line	–	–	✗	–	–
10	Vrtovec et al. [2007]	–	–	–	c-line	–	–	✗	–	–
11	Stern et al. [2010]	13	MR	Hough transform	V-D	–	–	✗	Crop S	Crop S
12	Glocker et al. [2012]	200	CT	Random regression forest + GM	V	–	–	✓	Auto	
13	Stern et al. [2010]	29	CT	Edge detection	V-D	–	–	✗crop S		
14	THIS WORK	300	MR	HOG + GM + Graph Cuts	V	2D3-VB-D	–	✗Contain S1-2	Auto	
15	Huang et al. [2009]	22	MRI	Adaboost + normalized cut	V	2D-VB	–	✗	?	
16	Seifert et al. [2009]	9	MR	Hough transform + snake + ASM	V	2D3-D,S	2D-D-ASM	✗		
17	Law et al. [2012]	33	MR	Level set on anisotropic flux	2D-V	2D-D	–	✗		two points
18	Ghosh et al. [2011c]	50	CT	Intensity model	V	2D-VB	–			
19	Shi et al. [2007]	50	MRI	Hough transform + edge detection	D	D	–	✗		
20	Peng et al. [2005]	5	MRI	Intensity model+edge detection	V	VB	–			
21	Jerebko et al. [2007]	42	MR	Ellipses fitted to VB cross-sections	S	3D-VB	2D-D-ellipse	✗		
22	Davatzikos et al. [2002]	14	MR	Deformable shape model	–	3D-VB	3D-VB-DM	✗		
23	Kelm et al. [2012]	42	MRI,	Marginal space learning + graph cuts	D	3D-VB-D	–	✓	Fix Full FOV	*
24	Shen et al. [2008]	21	CT	Level set on global + local priors	*	3D-VF	3D-VF-MSM	✗		
25	Leener et al. [2014]	15	MR-3D	Hough transform + deformable model	S	S	–		✗	
26	Kelm et al. [2012]	30	CT	Marginal space learning + graph cuts	D	3D-VB-D	–	✓	Fix Full FOV	
27	Klinder et al. [2008]	18	CT	Prior shape & spatial knowledge	V	3D-VF	3D-VF-DM			
28	Klinder et al. [2009]	–	CT	Prior shape & spatial knowledge	V	3D-VF	3D-VF-DM			
29	Alomari et al. [2013]	65	MR	Probabilistic model + ASM + GVF snake	–	2D-D	2D-D-ASM	✗	✗	
30	Ayed et al. [2011]	10	MR	Graph cuts + object interaction prior	–	2D-D	2D-D-ellipse	✗	?	
31	Stern et al. [2011]	9	MR	3D superquadric model fitting	–	3D-VB	3D-VB-SQ	✗	VB-pt	
32	Lu et al. [2012]	4	MR	Normalized cut	–	2D-VB	–	✗	?	
33	Gamio et al. [2004]	6	MR	Normalized cuts	–	3D-VB	–	✗	User click in VBs	
34	Chevrefils et al. [2009]	11	MR	Watershed + nearest neighbor classifier	–	2D3-D	–	✗Manual		
35	Neubert et al. [2012]	28	MR	3D Statistical Shape Models	V	3D-D	3D-D-SSM	*	One pt per VB	
36	Neubert et al. [2014]	14	MR	3D-vbSeg-SSM	–	3D-D	3D-D-SSM	✗	One pt per VB	
37	Michopoulou et al. [2009]	34	MR	Probabilistic atlas + fuzzy clustering	–	2D-D	2D-D-prAtlas	–		
38	Hoad and Martel [2002]	*	MR+CT	Thresholding plus morphological operations	–	3D-VF	manual ellipse	✗	define ellipse	
39	Kadoury et al. [2013]	8	MR	Manifold embedding + higher-order MRFs	3D-V	3D-VF +	3D-VF-PDM	?	Identify L5	
40	Kadoury et al. [2013]	21	CT	Manifold embedding + higher-order MRFs	3D-V	3D-VF +	3D-VF-PDM	Manual	Place a point in L5	
41	Mahmoudi and Benjelloun [2005]	100	X-ray	Edge detection + polynomial fitting	–	2D-VB	–	no-lb	Click per VB	
42	Roberts et al. [2006]	250	X-ray	Active appearance model	–	2D-VB+	2D-VB+-AAM			
43	Mastmeyer et al. [2006]	10	CT	3D balloon snake	–	3D-VB	3D-VB-ellipse			
44	Aslan et al. [2009]	10	CT	Graph cuts	–	3D-VB	No			
45	Aslan et al. [2010]	11	CT	Graph cuts + shape prior	–	3D-VB	3D-VB-ellipse			
46	Wong et al. [2009]	3	CT	Level set	–	3D-D	–			
47	Korez et al. [2014]	20	CT	3D superquadric model fitting	–	3D-VB,D	3D-VB,D-SQ	✗	Point in L3	
48	Stern et al. [2011]	10	CT	3D superquadric model fitting	–	3D-VB	3D-VB-SQ	✗	VB-pt	
49	Ma and Lu [2013]	40	CT	Edge detection (steerable features) + SSM	–	3D-VF	3D-VB-SSM	✗	manual crop	

Table 2.2: **Localization literature**. Localization of anatomical elements: vertebrae and discs. Note that other elements (spinal cord, etc.) are not included in this table. Det - detection. Seg - segmentation. V - vertebrae. D - discs. S - spinal cord / dural sac. Labelling: ✓if it could work on (our) standard clinical dataset automatically. '*' – no information available.

No	Paper	Task	Scans	Scan type	Region type	Region error	Feature	Auto?			Acc	Sen	Spe
1	Tsai et al. [2002]	Herniation	16	Axial MRI or CT	VDS-seg	Manual	Shape	✗	✗	✗	*	*	*
2	Koh et al. [2010]	Herniation	68	Sagittal T2	VDS-seg	Manual	Shape	✗	✗	✗	97	99	96
3	Koh et al. [2012]	Herniation	70	Sagittal T2	VDS-seg	Manual	Shape	✗	✗	✗	99	98.9	100
4	Ghosh et al. [2011b]	Herniation	35	Sagittal T2-SPIR	D-box	*,*	Both	m	‡	✓	98.3	96.2	99.2
5	Ghosh et al. [2011a]	Herniation	35	Sagittal T2-SPIR	D-box	*,*	Both	m	‡	✓	94.9	92.5	95.9
6	Alomari et al. [2010a]	Herniation	33	Sagittal T2-SPIR	D-seg	*,*	Both	✗	*	*	91	94	91
7	Alomari et al. [2011b]	Herniation	65	Sagittal T2-SPIR	D-seg	*,*	Shape	a	*	✓	92.5	86.4	96.6
8	Alomari et al. [2013]	Herniation	65	Sagittal T1 & T2	D-seg	*,*	Shape	✗	*	✓	93.9	86.4	96.6
9	Alomari et al. [2010b]	Abnormality	80	Sagittal T2	D-reg	*,*	Signal	m	*	*	91	*	*
10	Alomari et al. [2009b]	Abnormality	80	Sagittal T2	D-reg	*,*	Signal	m	*	*	91	*	*
11	Ghosh et al. [2013a]	Abnormality	86	Axial T2	D-Box	*,*	Signal	✗	⋆	✓	80.8	85.3	75.6
12	Alomari et al. [2009a]	Desiccation	55	Sagittal T2	D-reg	*,*	Signal	m	*	*	96	*	*
13	Unal et al. [2011]	Degeneration	9	Sagittal T2	D-box	Manual	Signal	*	✗	✗	98.9	*	*
14	Hao et al. [2013]	Degeneration	27	Sagittal T2	D-seg	Manual	Both	*	✗	✗	91	87.4	97.9
15	Neubert et al. [2013]	Degeneration	42	3D & sagittal T2	D-seg	0.87 ± 0.04 ;*	Both	✓	*	✓	ROC AUC >0.98		
16	Oktay et al. [2014]	Degeneration	102	Sagittal T1&T2	D-seg	*,*	Both	m	✓	✓	89.5	96	82
17	Lootus et al. [2014]	Pf. Grade	285	Sagittal T2	D-seg	0.81*; *	Both	✓	✓	✓	85.9	*	*
18	Mateos et al. [2014]	Pf. Grade	30	Sagittal T2	D-seg	0.92; 0.82mm	Both	m	✗	✓	*	*	*
19	Helo et al. [2013]	V-Fracture	50	CT	D-seg	*,*	Shape	✓	*	✓	98	>99	87.5
20	Stern et al. [2013]	V-Morphometry	108	CT	V-segpar	*;1.2mm	Shape	✓	✗	✓	92.5	92.5	92.5
21	Wels et al. [2012]	V-Lesions	34	CT	V-box	Manual	Signal	✗	✗	✗	*	75	*
21	Jerebko et al. [2007]	V-Lesions	42	Sagittal T2	V-seg	Auto	Both	✓	✓	✓	*	84.6	80
22	Forsberg et al. [2013]	Scoliosis angles	4	CT	VDS-rot	N/A	angle	a	✓	✗	97	*	*
22	Shen et al. [2013]	Scoliosis	255	x-ray	D-segpar	N/A	Shape	a	?	?	*	*	*
23	Koh et al. [2011]	Stenosis	55	Sagittal T2+MRM	S-seg	*,*	Shape	a	†	✓	91.3	*	*
24	Koompairojn et al. [2010]	Stenosis	50	Axial T2	seg	*,*	Shape	✗	*	✓	92-7	*	*

Table 2.3: **Disease State Prediction literature.** Grouped by task. Asterix (*) means data unavailable. **V**-vertebra. **D**-disc. **S**-dural sac. **Region:** type of support region: D-seg – disc segmentation, D-reg – a region ensured to be inside a disc, D-box – a disc bounding box; V-box – vertebra bounding box. **Auto:** a 3-letter string given: (slice-detection-segmentation). Each letter indicates automation: ✓-automatic; ✗-manual. For slice: **m**-mid-scan; **a**-all slices. ✗– placement of one point per V/D. **Examples**: ✓✓✓–fully automatic. ✗✗✗- fully manual. *✗✓– info on slice selection unavailable; manual initialization; automatic segmentation. **Special symbols:** † – using special modality, MRM; ‡ – using axial intersections; ⋆ - cropped.

et al. [2013].

In the existing approaches mainly the vertebrae, the discs, and sometimes the spinal cord have been localised or parametrised in radiographs, CT, and MRI. We cover both MRI and CT, and some radiograph papers. While MRI is more relevant to us, there are interesting ideas applied only in CT so far which may be potentially applicable to MRI in the future.

Strategies that work in 2D and 3D exist. The strategies in 2D tend to be simple, fast, and most of the ones in 3D complex, slower. The complex, slower strategies often involve lots of prior information on the anatomical structures, e.g. they may segment the vertebral posterior elements as discussed in Section 2.3.2. As previously described in Section 2.2, the standard lumbar MRI protocol tends to be sparse, 4-5mm thick sagittal + axial T1 + T2 slices, with just over or under 1mm in-slice resolution, and with wide variation of resolution and scan parameters for each series (e.g. sagittal T2 covers a range of protocols) and thus may limit the use of 3D methods.

Essentially an analysis strategy is a mapping from images to information about the patients. The perfect strategy would take in images of any possible view from any modality, acquired under any protocol and any unknown FOV, containing a patient with any conditions, placed in the scanner in any position, and map them into clinically relevant understanding to aid with research, diagnosis or treatment.

The design of a suitable analysis strategy for a given use case should be informed by its utility: firstly by the patients' anatomy and secondly the images on which it should be functional. The influence factors are, **firstly** according to the patient, the patient's disease state. **Secondly**, the image FOV, its resolution and image protocol (CT / MRI T1 / MRI T2, etc.).

Conversely, for good performance evaluation of an analysis strategy, a dataset containing the following is required: **first**, a broad range of different anatomies,

including diseased ones. **Second**, an appropriate range of different image protocols. **Third**, high-quality gold standard for assessment. A large dataset is more likely, however not certain to cover more of those than a small dataset.

The localisation task is challenged by unusual anatomies and image quality, and its algorithm training and performance evaluation potentially stifled by low-quality ground truth. The disease state prediction task is similarly challenged by the same factors, and in addition to co-morbidities – if conditions other than the one it was designed for are present. For example, potential presence of bulged discs may make it more difficult to predict disc herniation.

We also note that the appropriate way to assess whether a localisation method is good enough might be to include it in an end-to-end pipeline and assess its effect on the following diagnosis step. Then the diagnosis accuracy could work as an indicator for the required localisation performance.

We first discuss the localisation strategies in Sections 2.3.1 (detection + labelling) and 2.3.2 (segmentation + shape modelling), and then disease state prediction strategies in Section 2.3.3. We then discuss separately the levels of automation in pipelines in Section 2.3.4, and finally datasets and scanning protocols in Section 2.3.5.

2.3.1 Detection and Labelling Strategies

In almost all approaches, the localization algorithms work in two stages: candidate detection, and layout fitting. First, some anatomical parts characteristic of the spine are detected (vertebrae in Aslan et al. [2010], Chwialkowski et al. [1989], Glocker et al. [2012] / disks in Alomari et al. [2011a], Ghosh et al. [2012], Kelm et al. [2012], Pekar et al. [2007] / both in Oktay and Akgul [2013], Zhan et al. [2012]). Second, a spine layout model is fitted to the candidates to determine the best hypothesis for the spine layout. Often, labelling is built into this two-step process.

In effect, the vertebrae/discs detection algorithm is developed on the observation of self-similarity and repeating structure of the spine. The best spine layout and labelling is generally found based on both the parts local appearance, and their location, size and orientation with relative to each other.

Candidate detection. The candidates are mostly detected using sliding window detectors. Most commonly, either Haar, as in Kelm et al. [2012], or HOG, as in Ghosh et al. [2012], Lootus et al. [2013], Oktay and Akgul [2013] features are used. The detector output tends to be either a point inside the vertebra or disc, as in Alomari et al. [2011a], Oktay and Akgul [2013], a tight bounding box around it as in Huang et al. [2009], Kelm et al. [2012], or landmarks on the vertebrae or discs Kadoury et al. [2013]. The bounding box may also capture pose as in Kelm et al. [2012], Lootus et al. [2013].

Layout fitting and labelling. Generally, the spatial configuration of the spine parts, and in some cases also their individual distinct characteristics (Glocker et al. [2012], Klinder et al. [2009], Zhan et al. [2012]), are taken into account to both label the disks and/or vertebra, and localise the spine.

The general detection & labelling problem would be the following: given any scan of any view, modality, acquisition protocol (patient placement, etc.) and any unknown FOV, containing a patient with any conditions, fully automatically localise and label all the vertebrae or discs in the scan.

The actual task dealt with in existing papers is usually more narrow: either constrained by the knowledge of the FOV, a set resolution, protocol, or only applied on given patient disease state. For example the scans in Kelm et al. [2012] all include the full spine, C1-S1, in Alomari et al. [2011a] contain T12-S and only T12-S, in Stern et al. [2010] exclude S, and no labelling is performed. Scans are constrained to a set of allowed FOV-s in some works (contain S – Lootus et al. [2013], contain at least one of four 'anchor vertebrae' – Zhan et al. [2012]) or manual assistance from

axial slices is used (Ghosh et al. [2012]). In some cases a point is placed inside one of the VB-s, and other V-s are found based on this initialisation Kadoury et al. [2013]. They are also specialised on a given modality, and sometimes a single acquisition protocol (especially significant in MRI).

In high-resolution, full-width L-R FOV scans, vertebrae can be distinguished from each other using specific detectors (Ma and Lu [2013], Zhan et al. [2012]). Zhan et al. [2012] introduce a model which works for most FOV-s, requiring the presence of one of four 'anchor' vertebrae (C2,T1,L1,S1). They first detect the distinct 'anchor' vertebrae, and then other self-similar 'bundle' vertebrae connected to it graphically. Although the method works very well within its domain, it requires isotropic 2.1mm resolution scans which limits its applicability severely as clinical MRI scans come in narrower L-R FOV and usually 4-5mm sagittal slice spacing. Our method is not limited to this isotropic domain and, in particular, does not require the high isotropic resolution. In lower-resolution CT, full-width L-R FOV additional context of other organs and bones is taken into account by Glocker et al. [2012]. In typical clinical MRI however, both the L-R FOV is only disc-wide, *and* the resolution is low. In such scans, adjacent vertebrae appear very similar to each other and are thus hard or impossible (see Figure 1.5) to label, even by radiologists. The task in such scans has been handled by using a Sacrum detector in lumbar scans, as in Lootus et al. [2013], or having more vertebrae-specific detectors (Oktay and Akgul [2013], Zhan et al. [2012]), using known (fixed) S-I FOV (Alomari et al. [2011a], Kadoury et al. [2013], Kelm et al. [2012]) or assuming the existence of suitably placed (through discs) sparse axial slices (Ghosh et al. [2012]) scans, or not do it altogether (Gamio et al. [2004]). However, labelling is necessary to match parts of spine with radiological reports. In this thesis, we focus on lumbar scans, where the sacrum is always present.

Criminisi et al. [2011] were the first to parametrise the anatomy localisation

task as a multi-variate, continuous parameter estimation problem – addressed via regression to place bounding boxes around a number of organs in 3D CT using long-range spatial features. Glocker et al. [2012] build on this approach to train a system to detect (points inside) and label vertebrae in arbitrary FOV CT scans of heterogeneous protocol (resolution, etc.), based on the broad context of organs in the scan. Their approach first runs a random forest regressor to simultaneously find vertebrae candidates and approximately label them based on the long-range spatial features, more specifically – mean intensities over displaced, asymmetrical cuboidal regions. In second step, they run template-based detectors trained individually for each vertebra of the spine, and finally, dynamic programming on a graphical spine layout model, achieving 81% correct identification rate.

Wu et al. [2011] uniquely identify thoracolumbar vertebrae based on orthogonal matching pursuit, using rib detectors as T12 vertebra is distinguishable from L1 as it has false ribs attached to it whereas L1 does not.

Ma and Lu [2013] segment the vertebrae before labelling in thoracic CT of 1-12 vertebrae FOV. They check the match between the segmented vertebrae and the learned mean and the decision is made dependent on the best match.

In some approaches, the **spinal cord/dural sac/spinal centerline** is detected or segmented, such as in Stern et al. [2010], Koh et al. [2014], Chevrefils et al. [2007], Horsfield et al. [2010], Nyúl et al. [2005], Seifert et al. [2009], Jerebko et al. [2007], Leener et al. [2014], Wu et al. [2013]. This is sometimes done before vertebrae/disc detection, like in Klinder et al. [2009] where it is followed by image planar reformation and then vertebrae detection.

Vrtovec et al. [2005, 2007] detect the spine curve (a continuous line passing through the centres of IVDs and VBs in 3D, described by a 3rd degree polynomial model) in CT and MRI. In CT, they fit the spine curve to a set of points extracted from a distance map that emphasizes the vertebral bodies. In MR images, VB cen-

tres are located by searching for circular regions of low homogeneous intensity. They also curved planar reformatted the spine to remove its curvature. They initialise the process using a single manually placed point.

Stern et al. [2010] first detect the spinal centreline as a 3D curve passing through the centre of each VB. They then detect the centres of the VBs and IVDs along the centreline. They use detection of opposing VB edge pairs in axial planes for finding the centreline, and period detection in the intensity signal along the centreline to find the VB centres – they break the localisation task down by detecting candidates in axial sections of a 3D image, and only then perform localisation in the up-down direction. A similar strategy is partly followed by Kim and Kim [2009] and Forsberg et al. [2013].

Selection of a suitable labelling strategy for a given use case depends on first the patient anatomy, and second, the scan. First, severely damaged discs, vertebrae or spinal cords require detectors trained on a broader range of abnormal inputs, and scoliotic patients require layout models allowing for abnormal bending of the spine. Second, clinically realistic scans may pose limitations on available scan quality and its FOV: spatial resolution and FOV in both left-right (L-R) and superior-inferior (S-I) directions.

Spine detection and labelling is nearly a solved problem. The best previously published algorithms are those of Glocker et al. [2012] and Zhan et al. [2012]. Further work in detection could involve further validation of the developed algorithms on more heterogeneous data, investigating robustness to various diseases and imaging protocols.

2.3.2 Segmentation Strategies

We start this section with a general overview of spinal segmentation, and continue into vertebra, disc, and spinal cord segmentation strategies separately. Finally in

Section 2.3.2.5 we discuss more universal multi-object-class segmentation methods.

2.3.2.1 General Overview

Many papers have performed segmentation of spinal anatomy either as an end-goal in itself or as part of a processing pipeline. It should be noted however that some methods including our own have explicitly avoided segmentation. Here we review the techniques that have been proposed.

Segmented objects and modalities. Segmentation of vertebrae, discs, and the spinal cord has been attempted in the past. The VB and discs share a boundary, they are separated by the common endplate. Similarly, both the vertebrae and discs share a vertical boundary with the spinal cord. Therefore, the tasks of disc and vertebra segmentation include redundancy over the endplate line. The task of spinal cord segmentation includes redundancy over its anterior boundary. Nevertheless, the tasks have been handled separately. **Vertebrae** have been segmented in each radiograph, CT, and MRI. The segmentations have been used to initialise further image analysis, e.g. vertebral fractures diagnosis (Štern et al. [2013]). **Discs** have been segmented usually in MRI since MRI provides soft tissue contrast making it a useful tool for disc disease diagnosis, whereas CT does not. In addition, disc anterior and posterior boundaries can be virtually invisible in CT. But some work has looked at CT disc segmentation – for example by Korez et al. [2014] for prosthetic disc mould design. The **spinal cord** has been segmented in both MRI and CT, and is for example useful for assessment of stenosis or aneurism conditions.

2D vs. 3D. Segmentations have been performed in both 2D, and in 3D datasets, and both 2D, and 3D algorithms have been used. In **CT**, almost all spinal segmentation methods work in 3D. As mentioned earlier, most of them focus on vertebrae segmentation, segmenting either the complex-shaped full vertebrae (VF) – illustration in Figure 2.1 – or just the cylinder-like vertebral body. In **MRI**, 3D and 2D

methods are more equally frequent. In MRI, full vertebrae are rarely segmented, possibly since the complex posterior elements lack clear boundaries with ligaments and adipose tissue in MRI. Instead, usually only the vertebral bodies are segmented. A large number of MRI disc segmentation methods exist as well – disc delineations can be used to initialise various disc diagnosis methods in MRI. The disc segmentation methods are usually 2D – largely since only one slice is required for the diagnosis of a number of conditions, but also due to greater simplicity and speed, and the left and right boundaries of the disc may be invisible so 3D methods are challenging as they would have to try to segment the left and right boundaries as well. The advantage of 3D methods over 2D in general is potentially greater accuracy and robustness by taking into account surrounding slices' context, whereas the disadvantage is increased computation time, along with possibly greater need for learning and evaluation data along with ground truth, and greater implementation complexity. Another important point to note with regards to 2D methods is that almost all of them work in the sagittal view, as this is the most commonly used view for lumbar spine diagnosis. Sometimes also axial, very rarely coronal views are used. Additionally, the dataset may be another reason to use 2D rather than 3D methods. These methods working in 2D slices are particularly applicable for clinical MRI such as the ones dealt with in this thesis due to the fact that datasets can be very non-standard in both anatomy and imaging protocol (particularly e.g. either the field of view may be too narrow for 3D methods, or the slice spacing too large).

Methods. A number of standard computer vision methods have been applied for each vertebrae, disc, and spinal cord segmentation. Some of them have been successfully applied directly from computer vision as Kelm et al. [2012], others adapted to work for the spine segmentation task such as Law et al. [2012]. The main method adaptations from computer vision have to do with learning the image appearance in the specific medical protocols and applying methods on volumetric

images not encountered in computer vision. The methods can be broadly classified into four categories. **First**, curve evolution methods (e.g. snake, ACM, level sets), applied mainly on disc segmentation in MRI (e.g. Alomari et al. [2010a], Law et al. [2012], Mateos et al. [2014]). **Second**, graph-based methods, e.g. graph cuts (Boykov and Jolly [2001]) or normalised cuts (Shi and Malik [2000]) have been used often for VB, sometimes for IVD segmentation, in both 2D and 3D, in both MRI and CT. **Third**, a number of deformable models have been applied, often shape constrained (e.g. Ma and Lu [2013]). Those methods have mainly been used for 3D VF segmentation in CT, sometimes instead for 3D VB, or rarely applied MRI in addition to CT (Kadoury et al. [2013]). **Fourth**, there is a large number of other kinds of methods, usually performing low-level image processing in 2D, such as Watershed (Chevrefils et al. [2007]), Hough transforms (Shi et al. [2007]), tree-based classification (Ghosh et al. [2013b]). The advantage of the first group is the intuitiveness, the advantage of the second group is their speed and global optimum, the advantage of the third category is that they can include complex priors on shape yet they are very slow. In contrast, Wang et al. [2014] provide a unified framework for a rapid regression segmentation approach for both VBs and discs in both CT and MRI in various views in 2D. In 3D, Kelm et al. [2012] similarly segment both VBs and IVDs in both MRI and CT using graph cuts, whereas Štern et al. [2011], Korez et al. [2014] do the same using shape-based parametric superquadric models.

Image cues. The image cues according to which the segmentations are performed are, essentially, region information (intensity/colour, texture), boundary information (gradient, or a more complex boundary, e.g. laplacian zero crossing, or a learned boundary as in Ma and Lu [2013], or neighbourhood information like in Lu et al. [2012]), context (e.g. ligaments, etc. could be learned into the background model), and the shape (e.g. a learned shape model as in Neubert et al. [2012]).

The simplest deformable model is the snake. In other words an active contour

model, this is a time-evolved model containing an image energy that seeks a good contour fit to the image and a smoothness energy term that seeks to keep the curve smooth. The approach relies on good initialisation and is known to get stuck to local minima. The snake has been used for spine segmentation in 2D by Alomari et al. [2010a, 2011b, 2013], Seifert et al. [2009], and in 3D by Mastmeyer et al. [2006]. In those works, the snake is initialised by automated detections in the 2D approaches, and manually by placing ellipses in 3D by Mastmeyer et al. [2006]. In those approaches, the only shape regularisation is the surface smoothness term.

Another deformable model is the active shape model (ASM) (Cootes et al. [1995]). These are statistical models which similar to snakes iteratively deform to fit. The shapes are constrained by the PDM (point distribution model) SSM (statistical shape model) to vary only in ways seen in a training set of labelled examples. The strong shape priors can either be an advantage (e.g. in case of low image quality and false positive double contours) or a disadvantage (e.g. in case of a new diseased variant not represented in the training set). The ASM have been used by Seifert et al. [2009], Alomari et al. [2013] for 2D MRI, by Helo et al. [2013] for 2D sagittal CT sections, by Kadoury et al. [2013] for 3D vertebrae in CT and MRI, and by Neubert et al. [2013] for 3D discs. In a comprehensive study of disc segmentation using level set optimisation on ASM, Law et al. [2012] involve a number of hand-crafted terms in order to stop the model getting stuck into local minima proving that high performance is achievable given lots of work on building suitable terms into the model. It should be noted they initialise the model fitting using two manually placed points, though.

Performance. For a comprehensive assessment of spinal segmentation performance of a given method, three things would be required. **First**, a broad range of different anatomies, including diseased ones, in a large number of scans. **Second**, a range of different image protocols. **Third**, high-quality gold standard for assess-

ment. Since most methods have been implemented on private, often small, datasets, it is hard to pick winners. In 3D CT, in terms of DCI coefficient, the best full vertebrae segmentation performance of around 92.5% was reported by Kadoury et al. [2013] on 353 vertebrae, who embedded spinal centrelines into a manifold that was incorporated into a MRF optimisation, and obtained the surface of each vertebra by using deformable models, whereas Ibragimov et al. [2014] acieved 93.6% performance on 50 lumbar vertebrae in CT. In MRI, a number of methods also quote DCI coefficients in the 90-s, e.g. Huang et al. [2009] on 52 vertebrae in 2D. In fact, it seems a large share of the papers for both vertebrae and disc segmentation achieve a 90% or greater DCI measure. While the performance of such methods seems to be very high on healthy patients in high-quality images, especially in the methods using strong shape priors, no evidence has been provided that it could work in the presence of pathology. Given that they achieve comparable results to the complex, the simpler, faster 2D methods may be preferable. It is also hard to assess what DCI scores are required for accurate radiological measurements later in the pipeline, as the measurement papers often use manual segmentations, e.g. Koh et al. [2010, 2012], Hao et al. [2013], or no segmentation at all at the end (Alomari et al. [2009a], Ghosh et al. [2011b,a]), or they do not disclose the performance of the segmentation algorithm (Alomari et al. [2010a, 2011b, 2013]).

Manual input. Note that all the segmentation approaches require some initialisation – usually either position of a point or a bounding box (Huang et al. [2009]), or a full position + pose (Ma and Lu [2013]). In some cases, fully automatic initialisation has been provided (Huang et al. [2009]) whereas at other times, either manually placed points (Štern et al. [2011], Kadoury et al. [2013]) or other structures such as ovals (Hoad and Martel [2002]) have been provided. Note that while works have been developed, as reviewed in Section 2.3.1 to automate initialisations, the segmentation performances in papers relying on manual clicks may not accurately

reflect equivalent results given automated (and potentially imperfect) initialisations.

2.3.2.2 Vertebrae Segmentation

There are three main groups of vertebral segmentation methods according to the object dimensionality and completeness. **First**, segmentation of VB-s in 2D, where they appear rectangle-like. **Second**, segmentation of VB-s in 3D, where they appear cylinder-like. **Third**, segmentation of VF-s in 3D, where they appear as complex objects, including the pedicles and posterior elements in addition to just the VB-s. The VF has not been segmented to our knowledge in 2D.

The methods in the **first** category are the simplest and fastest. They have almost all been applied on MRI, or some in radiographs, and include usually no prior model in MRI, sometimes a 2D shape model in radiographs. In MRI, 2D methods may be preferable for two reasons. First, it may be impractical to process in 3D as the MR scans may be anisotropic resolution, and only some slices may be available. Second, 3D methods may just be unnecessary as the end goal might be diagnosis in a single slice. The 2D methods in MRI have been mainly graph-based, and some by edge detection. The graph based methods are Huang et al. [2009], Lu et al. [2012] using Normalised Cuts and Lootus et al. [2014] graph cuts. The edge detection methods are, with intensity modelling Peng et al. [2005]. Differently, Ghosh et al. [2013b] solve the problem using decision trees, and Wang et al. [2014] by regression from a number of standard computer vision features to the image. In sagittal radiographs (X-rays), Roberts et al. [2006] use linked AAMs; Mahmoudi and Benjelloun [2005] use edge detection with polynomial VB model fitting. In single sagittal planes of CT, Ghosh et al. [2011c] use edge detection along with Hough transform, and Helo et al. [2013] use Active Shape Model + Gradient Vector Flow snake – a boundary-based model.

The methods in the **second** category, for 3D VB segmentation, have been im-

plemented more equally in both CT and MRI. Some of them can still be relatively fast however require a number of slices to work. Some methods work on only CT, whereas others work on MRI only, and some work on both MRI and CT. In CT, the methods are Mastmeyer et al. [2006], Tan et al. [2006], Aslan et al. [2009, 2010], Štern et al. [2011], Kelm et al. [2012]. In MRI, the methods are Davatzikos et al. [2002], Gamio et al. [2004], Jerebko et al. [2007], Kelm et al. [2012], Štern et al. [2011], Neubert et al. [2012]. Some of them involve strong shape priors, such as Davatzikos et al. [2002], Štern et al. [2011], Kelm et al. [2012], Neubert et al. [2012], whereas other are relatively simple, prior-free, and likely fast, e.g. Gamio et al. [2004], Aslan et al. [2009, 2010], Kelm et al. [2012] using graph/normalised cuts or Tan et al. [2006] using level sets or as Jerebko et al. [2007] using ASM. Lots of those methods involve manual initialisation.

The methods in the **third** category have been mostly applied on CT, where the vertebrae are better distinguishable, and include Shen et al. [2008], Klinder et al. [2008, 2009], Kim and Kim [2009], Kadoury et al. [2013], Ma and Lu [2013], Ibragimov et al. [2014], Korez et al. [2015]. Of those, Kadoury et al. [2013] also apply their method on MRI, and to our knowledge uniquely Hoad and Martel [2002] apply their method on MRI only. These methods tend to use very strong shape priors, generally learned from the VF shapes on a set of training images, incorporated into 3D deformable models. The methods in this category are generally very slow – e.g. Ma and Lu [2013], Klinder et al. [2009] spend tens of minutes per image. The other possible disadvantage of these methods is inapplicability on severely deformed vertebrae. These approaches are generally more suitable in images where the edges are clearly visible, and the images of isotropic resolution, with wider L-R FOV, and this is generally the case with CT.

We now pick a number of methods from each category, and explain them in more detail.

In the **first** category, Huang et al. [2009] and Lu et al. [2012] segment the VB in 2D MRI using normalised cuts, whereas Peng et al. [2005] use canny edge detector.

In both Peng et al. [2005] and Huang et al. [2009], the segmentations are initialised by vertebrae detections. In Peng et al. [2005], a canny edge operator is used to extract all edging points. Next, hand-crafted connectivity tracing and corner point finding procedures are applied to provide a rough boundary and to locate the four corner points. Finally, gap filling and line segment linking operations are applied between the corner points. In Huang et al. [2009], the segmentation is performed using iterative normalised cuts, whereby the edge energy is the standard gradient strength. No spatial priors are applied.

In Lu et al. [2012], the initialisation strategy is not disclosed, however a novel affinity matrix of pairwise affinities is used to calculate the pairwise costs for the normalised cuts, resulting in around 97% overlap measure. The method is tested on only 4 patients.

In the **second** category, a set of works perform segmentations of 3D VB structures. Here, in both CT and MRI, Štern et al. [2011] use a parametric model, and Kelm et al. [2012] use 3D graph cuts. In MRI, Neubert et al. [2012] use a statistical shape model.

In both CT and MRI, Štern et al. [2011] use parametric superquadric VB models (initialised by user clicks) to impose heavy shape constraints on the segmentation. To model the basic 3D shape of the VB, they generate a 38-dimensional superquadric in the form of an elliptical cylinder, which is then gradually deformed by introducing transformations that yield a more detailed representation of the vertebral body shape. The optimisation is performed greedily in 3 successive steps, using a downhill simplex algorithm. The advantage of this approach is the clean, low-dimensional representation that can be used for parameter measurement directly, along with high smoothness of the segmentation. The disadvantages are the long computation time,

and likely misfits to VB-s / IVD-s that do not conform to the regularities built into the relatively regular, low-dimensional models.

Kelm et al. [2012] use a simple, fast, globally optimal computer vision method – graph cuts (initialised by vertebrae detection in full-body-FOV scans) to segment both the VB-s and IVD-s in both MRI-s and CT-s. They perform experiments in only full-body isotropic scans, and initialise the segmentation using automated detection. They do not quote numerical performances.

Neubert et al. [2012] segment the vertebral bodies in 3D MRI using statistical shape models trained on a subset of their images. They use grey level matrix as a similarity measure, and achieve 91% DCI coefficient on a dataset of 42 scans.

In the **third** category, Ma and Lu [2013] propose an interesting two-step hierarchical deformation scheme to potentially speed up high-dimensional model fitting, and in addition, learn steerable features for edge detection in CT. They divide the deformable vertebrae models up into 12 different spatial regions. Given an initialisation (position, scale, angle, provided by automated proprietary Siemens FAST 3D detection + pose search) for the mean V-model (as learned on a training set) they first greedily search over a restricted set of position + scale + angle deformations for each region in turn, not allowing shape change for any of the regions. The similarity metric is match of the model boundaries to a steerable feature detector output. Having repeated this three times, in the second step, they fix the subregions and apply patch-based deformations across the whole model, repeating the second process four times. They apply Gaussian smoothing to the model and recalculate the steerable feature map (which is dependent on gradients parallel to the model surfaces) after each iteration. Despite the speedup scheme, the process still takes 5-10 minutes per vertebrae in their published experiments with 1000 vertices per V, though they say in private communications that the number of model vertices could potentially be reduced. For highly pathological V-s, the approach would have to

be combined with a bottom-up approach. In this approach, the shape constraints are encoded into the limitations to the search space for both the subregion and the patch deformations search.

Uniquely in just MRI, Hoad and Martel [2002] segment the full vertebrae using thresholding plus sequences of morphological operations, for the VB and the posterior components of the V separately. The segmentations are initialised by precisely manually placed ellipses (1-2 per vertebra) in axial slices. The ellipses are required to be placed in a way that their boundaries overlap the anterior edges of the vertebrae.

In both MRI and CT, Kadoury et al. [2013] segment the full vertebrae using a point distribution model. Their shape constraints for the individual vertebrae are dependent on the global curvature of the spine, adapting for scoliotic patients.

A big challenge with deformable models is finding strategies to fit them in a way that would be fast and not get stuck in local minima. Štern et al. [2011] and many others fit their models in a greedy manner, searching first for the best cylinder fit, then fix those parameters, and deform the next set of parameters.

2.3.2.3 Disc Segmentation

The intervertebral discs segmentation has almost always been performed in MRI, motivated by the unique usability of MRI for soft-tissue diagnosis. It has usually been done in 2D slices, and in some rarer cases, in full 3D. While good progress has been made, it remains an unsolved problem, especially in diseased discs.

Disc segmentation may be more challenging than vertebrae segmentation because (i) discs are soft-tissue objects and show more signal and shape variation, particularly due to diseases compared to the vertebrae, (ii) the lateral boundaries of discs are more invisible than those of the vertebrae, since the AF is of the same intensity in T2 images as blood vessels, ligaments, and muscles.

Also of note is that there can be considerable inter- and even intra-observer

variation in manual segmentation (e.g. for ground truth). For example Haq et al. [2014] found mean 0.265mm intra-rater, and 0.285 inter-rater variation in mean surface distance, and over 3 mm mean Hausdorff distance for both intra-rater and inter-rater difference, where most of the variation was found in the lateral regions of the disc with ambiguous boundaries.

While a number of papers have achieved around 90% DCI coefficient, they have often not explicitly demonstrated good results on significantly diseased discs, leaving particularly herniated disc segmentation relevant for radiological measurements an unsolved problem. See Figure 2.3. In addition, in some cases, strong prior terms and learned shape priors may hinder segmentation of herniated disc parts Neubert et al. [2012], Haq et al. [2014], Wang et al. [2014].

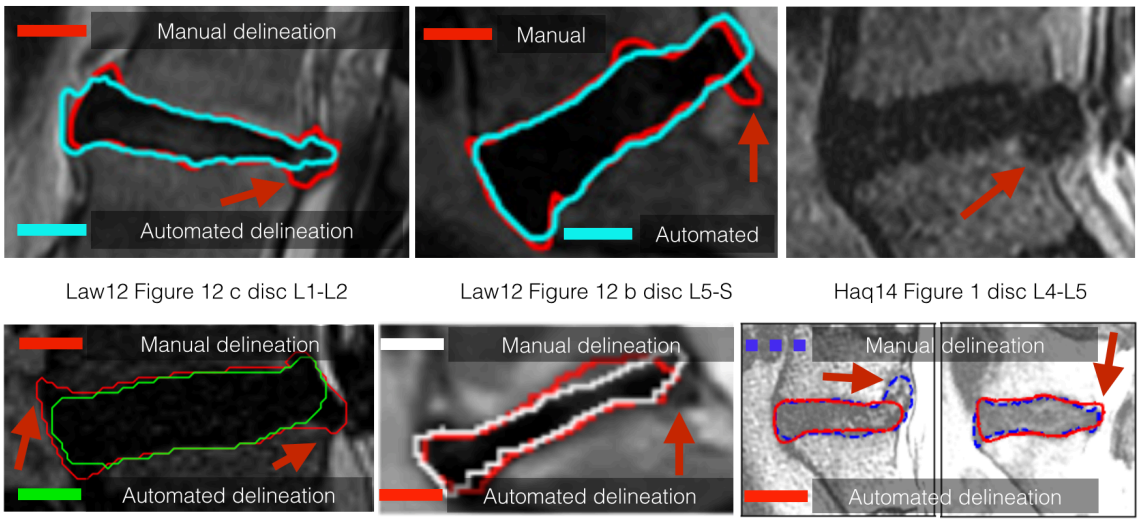


Figure 2.3: **Disc segmentation is an unsolved problem on herniated cases.** Example disc segmentation cases are shown extracted from Figures in literature on herniated discs, from better performing methods which achieve high mean performances of DCI score around 90%. Law et al. [2012], Wang et al. [2014], Ayed et al. [2011], Mateos et al. [2014], Haq et al. [2014] . Note also that the manual segmentation can have high rate of uncertainty. The issues are largely partial volumes and invisible boundaries due to ligaments. And hence the task is often ill-defined as boundaries are simply invisible due to low image quality, making even accurate ground truth definition problematic.

In **2D MRI slices**, discs have been segmented by Michopoulou et al. [2008, 2009], Ayed et al. [2011], Lootus et al. [2014], Alomari et al. [2010a, 2011b, 2013], Ghosh et al. [2011b,a, 2013b], Mateos et al. [2014], Oktay et al. [2014], Chevrefils et al. [2007, 2009], Shi et al. [2007], Wang et al. [2014], Seifert et al. [2009].

In the following subset of the 2D approaches, the segmentations have been performed as part of a diagnosis pipeline, without disclosing segmentation performance. Alomari et al. [2010a, 2011b, 2013] use contour evolution ASM and snake models to segment the disc for herniation prediction, whereas Ghosh et al. [2011b,a] use similar methods for disc bounding box refinement, also preceding herniation prediction. Out of the last set of papers, Alomari et al. [2010a], Ghosh et al. [2011b,a] use ASM of Cootes et al. [1995]; Alomari et al. [2011b] use Gradient Vector Flow (GVF) Snake of Xu and Prince [2000]; whereas Alomari et al. [2013] use first ASM for the whole disc, and then GVF Snake for the posterior section of the disc. Note also that they work in scans with fixed, known FOV (containing all of and no more V-s than T12-S). Oktay et al. [2014] perform segmentation on 102 subjects, 612 discs (349 normal, 263 with degenerative disc disease) in a fully automatic measurement pipeline using Active Appearance Models (Cootes et al. [1995]) preceding automated Pfirrmann grading.

Out of the rest of the 2D MRI disc segmentation approaches, for which quantified results are provided, the five top-performing ones are Ayed et al. [2011], Seifert et al. [2009], Law et al. [2012], Mateos et al. [2014], Wang et al. [2014], all citing performance around DCI of 90%. While they score uniformly highly in DCI, they have not necessarily demonstrated the ability to perform diseased discs, especially herniated. In fact, high DCI might not be an appropriate measure to evaluate the segmentation of diseased discs. See Figure 2.3 with some example cases extracted from the papers, showing inaccuracies in herniated disc areas.

Those five aforementioned top-performing 2D papers are now discussed in more

detail.

Law et al. [2012] use level sets on top of anisotropic flux, in a hand-crafted pipeline of a large number of steps involving morphological operations and terms to the cost function to specifically address the various anatomical structures in and around the disc. They evaluate their method on 22 subjects, 69 slices, 455 disc cross sections using manual initialisation with placement of a point into the top and bottom vertebra in the image. Under evaluation, they categorise discs according to their disease state into four categories: (i) 110 normal, (ii) 109 degeneration, (iii) 310 extrusion, herniation, protrusion, and bulging, (iv) 74 both ii and iii, with DCI coefficients (i) 0.92; (ii) 0.91 ; (iii) 0.92; (iv) 0.91 however with significantly lower mean-square boundary errors for normal (0.8mm) than for categories ii-iv (1.2,1.0,1.0). Since their method is evaluated in the mid-sagittal slices yet the disease state may appear in other slices, the influence of disease may be masked.

Ayed et al. [2011] on 10 subjects, 60 disc cross-sections, each from a unique disc using manual initialisation with three points per disc using graph cuts with elliptical shape priors and obtaining mean DCI score of 0.88;

Wang et al. [2014] use a regression approach, regressing from image features (WI-SIFT, WI-SURF, GIST, HOG) to 100-point contours using multi-class SVR, with 0.87 DCI for discs in sagittal images on 465 disc cross-sections from 113 subjects. Each segmentation takes around 0.1 seconds on a laptop. They initialise the segmentation based on manually cropped images as seen in Figure 2.3. As evident in Figure 2.3, the segmentation can still be inaccurate on herniated areas.

Seifert et al. [2009] segment both the discs and the spinal cord in the cervical spines of nine patients. They first use four open snakes for approximate, and then ASM-s combined with fuzzy connectedness algorithm for more precise segmentations. Their ASM-s for discs include a mean disc shape model along with grey-level profiles. They achieve 91% average DCI score. Their segmentation is initialised

using automated disc detection.

Mateos et al. [2014] on 30 subjects, 110 disc cross-sections, each from a unique disc including Pfirrmann grades 1-5 using manual initialisation with one point roughly in the centre of each disc using active contour models along with fuzzy C-means allowing each contour point to move up to two pixels for the final result. They achieve a mean DCI score of 91.7%, with 90.9, 92.5, 92.6, 90.2, 87.5 for the five Pfirrmann grades respectively.

Shi et al. [2007] introduce another fast segmentation method based on Hough transforms and run it on 50 patients however do not provide quantified results other than to say it was satisfactory on 48 out of the 50 patients, 96%. Chevrefils et al. [2007] propose a Watershed segmentation approach and run it on 4 patients with various levels of scoliosis however do not provide quantified results. Based on their Figures with example results, the method seems to suffer from over-segmentation.

Disc pathology, especially herniation, generally varies widely across patients and thus cannot be represented well by strong shape prior models. The approaches in literature that have been applied on herniated discs have been Active Shape Models as in Alomari et al. [2010b, 2011b], in the picked herniated slice, without mentioning segmentation performance. Alomari et al. [2013] perform separate segmentation for the main part of the disc and the herniated back-side using ASM and GVF-snake respectively.

In **3D MRI**, Neubert et al. [2013] segment discs in 3D using a statistical shape model, initialised by manually placed points. Their model is learned on the training set and tested on 42 scans, achieving a DCI score of 89% on normal and degenerated discs. In Haq et al. [2014], herniated discs are explicitly considered. A shape model is used only for healthy discs, and switched off manually for deformed discs, in order to still allow for accurate segmentation. In their method, they initialise an ellipsoidal simplex template within the disc image boundary through affine landmark-based

registration and allow it to deform according to image gradient forces. They segment 16 healthy discs automatically, and 5 pathological discs 'with minimal supervision' as required, e.g. placing points to guide the segmentation.

Uniquely in **CT**, Korez et al. [2014] use a 3D parametric superquadric disc shape model to segment the discs, initialised by segmentations of the vertebrae (obtained according to Štern et al. [2011]).

Kelm et al. [2012] segment discs in both 3D MRI and CT using Graph Cuts, initialised by seeds obtained by vertebrae segmentations however they do not disclose quantified segmentation performances.

All in all, disc segmentation remains an unsolved, and somewhat ill-defined problem. Most of the methods only work in a single 2D slice and involve manual initialisation. Since the disc boundaries smoothly morph into their surrounding ligaments which are often indistinguishable from the annulus of the disc, the task remains very challenging. However as we demonstrate in this thesis, only vertebral segmentation may suffice for radiological measurement.

2.3.2.4 Spinal Cord Segmentation

Leener et al. [2014] segment the spinal cord in isotropic 3D MRI using propagation of a low-resolution deformable model. Their approach has three steps: detection of the cord position using a circular Hough transform for initialisation within slice; secondly, the deformable model is propagated across slices with local contrast-to-noise adaption; thirdly, a refinement process.

Horsfield et al. [2010] introduce a semi-automatic method to segment the spinal cord in the cervical region in 3D MRI (voxel 1-by-1-by-1.09mm) based on an active surface model which is manually initialised.

Nyúl et al. [2005] segment the spinal cord/spinal canal in the thoracic region in CT, initialised by a single point placement. They use region growing for spinal

canal, and active contour model for spinal cord segmentation.

Koh et al. [2011] perform spinal cord segmentation in sagittal slices of Magnetic Resonance Myelography, using morphological operations.

Chevrefils et al. [2007] segment the cord in both sagittal and axial slices of multi-slice MRI.

McIntosh and Hamarneh [2006] segment the cord in 3D MRI (voxel size 0.98-by-0.96-by-1 mm) using 'living organisms', the spine crawlers. They model the organisms to have brain and body. They initialise the segmentations using two seed points.

2.3.2.5 Multi-Structure Segmentation

These methods all work in 2D slices and are especially applicable for clinical MRI such as the ones dealt with in this thesis due to the fact that datasets can be very non-standard.

A particularly interesting approach applicable to those datasets is taken by Wang et al. [2014] who segment both discs and vertebrae in both CT and MRI slices regressing 100 boundary points directly to ROI-s (cropped according to GT segmentations in their paper) in less than 0.1 second per structure.

Ghosh et al. [2013b] use decision trees on HOG features to simultaneously segment vertebrae, discs, and the spinal cord in a multi-class problem.

Koompairojn et al. [2010] apply a multi-layer perceptron algorithm for segmentation, trained on example cases, for segmentation of a number of anatomical components in axial images (facets, ligamentum falvum, the disc, the spinal canal). Also, there are more partial volumes in axial slices, especially between the disc and the vertebrae.

2.3.2.6 Summary

While a number of works have been published, spine segmentation on clinical MRI remains a challenging problem. Like other spine image analysis tasks, it involves heterogeneities – both anatomical, and imaging – which have not been addressed in existing work. The best papers so far are those of Law et al. [2012] and Wang et al. [2014]. Law et al. [2012] perform disc segmentation on using a hand-crafted method that carefully accounts for the anatomy, using manual initialisation however. Wang et al. [2014] present a regression method to segment all spine structures in a single framework however seem to fail on herniated cases based on their Figures. Further work on segmentation should specifically address segmentation on diseased cases, and in axial slices.

2.3.3 Measurement Strategies

Measurement strategies lead from localised anatomical structures to their parametrisations. The localisations required can be in the form of organ detections (e.g. VB, disc) (Lootus et al. [2013]), segmentations (Wang et al. [2014]), or landmark detections (Koh et al. [2012]). Those can be provided either manually, or by automated methods. While inclusion of manual localisation to avoid first developing localisation methods can provide a quicker estimate to the parametrisation performance, it might lead to overly optimistic estimate of the system performance (as evident in our results in Chapter 6).

2.3.3.1 General Overview

See Table 2.3. In MRI, binary classifications (normal or diseased) of the following disease states of the disc have been attempted so far: disc herniation in sagittal (Alomari et al. [2010a, 2011b, 2013], Ghosh et al. [2011b,a], Koh et al. [2010, 2012]) and axial MRI scans (Tsai et al. [2002]); general disc abnormality in sagittal multi-

slice MRI by Alomari et al. [2009b, 2010a], Ghosh et al. [2013a]; desiccation (drying out) by Alomari et al. [2009a]; degeneration by Hao et al. [2013], Neubert et al. [2013], Oktay et al. [2014], Unal et al. [2011]. Automatic Pfirrmann grading has been performed by Lootus et al. [2014] using regression, and by Mateos et al. [2014] using classification.

Prediction of central spinal cord canal stenosis has been attempted by Koh et al. [2011] in sagittal Magnet Resonance Myelography (MRM) images, and prediction of both foraminal and central stenosis conditions in axial MR images by Koompairojn et al. [2010]. Jerebko et al. [2007] detect vertebral lesions in vertebrae in sagittal MRI.

In CT, prediction of scoliosis (Forsberg et al. [2013], Shen et al. [2013]); vertebral fracture (Helo et al. [2013]), osteophytes (in PET/CT) (Yao et al. [2013]) have been attempted.

2.3.3.2 Herniation Measurement

In the literature, automated herniated vs. normal disc classification has been tackled in sagittal MRI using 3 principal sets of methods, in 2D sagittal slices: (i) Alomari et al. [2010a, 2011b, 2013] based on the shape of the full segmented disc boundary in 2D slices; (ii) Koh et al. [2010, 2012] based on the percentage of manually segmented vertebrae, disc, and sacrum voxels in a manually placed box at the posterior side of the disc, and (iii) Ghosh et al. [2011b,a] based on a number of image features extracted from a rectangle covering the disc.

In the first set, Alomari et al. [2010a] measure herniation based on two shape features (distance between ASM points approximately corresponding to the sum of the height and width of the disc, and the height of the disc at the posterior third), and two intensity features (the mean and standard deviation values of a Gaussian fitted to the disc intensities in the segmented disc). Alomari et al. [2011b] use very

similar shape features, in T2-SPIR, except they are extracted from snake segmentation that according to their figures segments the nuclear disc boundary. They do not use intensity features. The snake segmentations are initialised using a point automatically placed into the disc using the method of Corso et al. [2008]. Alomari et al. [2013] once again use only two shape features however those are one corresponding to disc width from the ASM segmentation, and another is the segmented herniated boundary length from the Gradient Vector Flow snake segmentation of the posterior side of the disc.

In the second set, Koh et al. [2010, 2012] perform extensive, fully manual annotation: they segment the discs, the vertebrae, and the dural sac, and place 27 landmark points per disc to define a ROI – a box in the posterior region of the disc. They use as features the percentage of vertebra, disc, and spinal cord voxels in the box. They do not specify, which slice(s) are used in the analysis, nor how they are selected (automatically or manually).

In the third set, Ghosh et al. [2011b,a] use a probabilistic model for automatic disc localisation and labelling in all sagittal slices (according to Corso et al. [2008], resulting in a point inside each disc). Next, the disc orientation is found according to the angle of the corresponding axial images. Note that this approach will not work in studies where axial images do not follow the angle of the spine, or where axial images are unavailable. They then initialise an ASM segmenter based on these localisations, and place a rectangular support region by eroding the tight bounding box of the segmentations. Ghosh et al. [2011a] use the GLCM (Haralick [1979]) features: divide the region into 8 spatial bins and extract intensity (mean, min, max, etc.), and texture features (contrast, correlation, energy, homogeneity, entropy) from the bins, and in addition the height-width ratio of the box as shape feature. They find the highest classification accuracy with a majority vote classification scheme. Ghosh et al. [2011b] additionally to GLCM experiment with raw (similar to ours),

LBP, Gabor, and additional intensity and shape features extracted from the whole region (rather than each bin individually).

All those approaches classify discs based on sagittal slice(s) into two categories: herniated vs. normal, based on analysis of one slice (Alomari et al. [2013] manually select the slice where herniation is both visible) or more slices (Alomari et al. [2011b] run the classifier independently in all slices and if at least one slice is found 'herniated', the disc is classified as 'herniated') independently. However, often a disc may be bulged instead of herniated, and yet be indistinguishable from herniation based on such classification. Alomari et al. [2010a] do not disclose which slice(s) is the analysis performed.

In none of the above approaches is bulge considered as an option. It is not mentioned whether bulged discs are included in the studies. Unlike, Tsai et al. [2002] consider a number of axial slices (one or more per disc), segmenting the disc manually, and proposing features to describe the 3D geometry of the disc, to distinguish normal, herniated, and bulged discs. They fit a B-spline model to approximate the normal disc curvature, and measure the local deviation from the normal curvature to detect herniations. However they do not learn the model parameters nor report classification performances in comparison to a radiologist, and note that the method as is does not apply to discs which are not roughly round, or belong to aged patients. They do however discuss a number of relevant issues to the task.

2.3.3.3 Degeneration and Abnormality Measurement

Alomari et al. [2009a] predict desiccation based on the median intensity in a 5-by-5 pixel square that has to be manually ensured to be fully in the disc. Unal et al. [2011], Hao et al. [2013], Neubert et al. [2013], Oktay et al. [2014] predict the degeneration of a disc. Unal et al. [2011] use neural networks on GLCM on a number of features (autocorrelation, contrast, etc.), and AAD on intensity, extracted from 200-by-200

pixel patches manually centred on the discs. Hao et al. [2013] classify manually segmented discs into normal and degenerated using SVM on intensity and carefully parametrised shape features, introducing active learning into disc diagnostics. Oktay et al. [2014] classify degenerated discs using a large number of different features building on Ghosh et al. [2011b]. Mateos et al. [2014] classify discs according to the Pfirrmann grade, and Neubert et al. [2013] classify degenerated discs in 3D using GMM fit and disc height to width ratio.

Alomari et al. [2009b, 2010b] use the same features as Alomari et al. [2009a] however for a different task – abnormality measurement. Ghosh et al. [2013a] use CNN features to detect abnormalities in axial slices localised by cropping according to the sagittal slice intersections.

2.3.3.4 Other Measurements

Vertebral fractures. Helo et al. [2013] 2D, and Štern et al. [2013] 3D vertebral morphometry for vertebral fracture. Helo et al. [2013] use measurements from 2D segmentations, and Štern et al. [2013] use measurements from the fit of parametric superquadric models of the VB explained in more detail under the segmentation literature review section.

Vertebral lesions. Jerebko et al. [2007] and Wels et al. [2012] in CT detect lesions in the vertebral bodies.

Scoliosis. In pairs of X-ray, Shen et al. [2013] parametrise scoliosis by a torsion estimator; and in CT, Forsberg et al. [2013] by axial vertebral rotation.

Stenosis. In sagittal T2 MRI + MRM, Koh et al. [2011] estimate scoliosis by a combination of spinal cord width measurements; and in axial T2 MRI, Koompairojn et al. [2010] using neural nets for segmentation and then similarly spatial dimensions of the cord and other objects to predict stenosis.

2.3.3.5 Summary

Automated measurements, mainly of degeneration and herniation, have been performed in simplified cases – with in small datasets, with homogenous imaging protocol; perhaps excluding co-morbidities. The best works so far are those of Alomari et al. [2011b], Alomari [2011] for herniation (although highly simplified experimental setting), and Oktay et al. [2014] for disc degeneration. Further work on automated measurements should address measurements in more heterogeneous settings (both imaging and anatomy), on larger datasets, and involving less manual steps (e.g. slice selection).

2.3.3.6 Support regions & Features

To capture the conditions listed earlier, a number of signal, texture, and shape features have been proposed. For disc conditions, these are extracted from regions covering the disc and its context for normalisation, and range from rectangular bounding box around the disc (Ghosh et al. [2011b,a]) through a non-specified (details not provided in paper) region required to be fully in the disc (Alomari et al. [2009a,b, 2010b]), to disc segmentation (Alomari et al. [2010a]).

The signal features extracted from the disc regions range from mean intensity for desiccation (Alomari et al. [2009a]) and abnormality (Alomari et al. [2009b, 2010b]) through Gaussian Mixture Model coefficients (Neubert et al. [2013]) for degeneration and histograms and statistical features (Lootus et al. [2013]) for Pfirrmann grade to Gray Level Co-Occurrence Matrix (GLCM) for herniation (Ghosh et al. [2011b,a]). Some approaches such asGhosh et al. [2011b] have combined a number of features (e.g. GLCM, Gabor, Patch) to improve performance. The signal features are usually either normalised to the CSF in the spinal cord (Videman et al. [2003]), vertebrae intensity (Lootus et al. [2014], Neubert et al. [2013]), or also to other discs in the scan (Oktay et al. [2014]).

The shape features are either dimensions of the segmented discs, e.g. mid-height to width ratio among other features for degeneration (Lootus et al. [2014], Neubert et al. [2013]), or directly the fitted shape model parameters for herniation: effectively capturing disc height and width (Alomari et al. [2011b]) or its overall boundary and herniated material boundary length (Alomari et al. [2013]). Koh et al. [2010, 2012] count the disc, vertebrae and voxels in a precisely placed square around the back of the disc to predict herniations.

2.3.4 Full pipeline automation: error propagation, and levels of performance

Note that every part in a pipeline might have imprecisions in their outputs, and that errors from earlier stages may propagate into the later stages. For example, detection errors may cause segmentation errors, and segmentation errors may cause measurement errors. Thus, it is important to calculate outputs for later links, given imprecise inputs by the earlier links.

Segmentation performance achieved by initialising an algorithm using manually placed seeds (e.g. points or bounding boxes) may exceed the performance achievable using seeds placed by automated detections, as a human detector may rely on prior knowledge on the segmentation algorithm not available to the detection algorithm. And manual vs. automatic localisations may have a similar effect on measurement results.

The sensitivity to initialisation may be heavily dependent on the task (e.g. vertebra vs. disc segmentation) and the cost function (e.g. deformable models cost function for which algorithm may get stuck to local minima vs. cost function globally optimisable by graph cuts).

See Tables 2.2 (column 'Manual input') and 2.3 (column 'Auto?'). There are two levels to automation. **First**, automated localization of image support regions,

given an image of a patient. And **second**, automated prediction of disease given the relevant image support regions, using machine learning techniques.

It should be noted that many methods require (or rather, use in their experiments, and have not proven to work with automated initialisation) manual initialisation. This could be either selection of the slice where a herniation is most visible (Alomari et al. [2013]), manual placement of points to initialise a segmentation algorithm (Korez et al. [2014], Neubert et al. [2013], Štern et al. [2011]), or fully manual initialisation by drawing bounding boxes (Wels et al. [2012]) or segmentation (Hao et al. [2013]). The approaches that appear to require the most manual work per scan are Koh et al. [2010, 2012], Tsai et al. [2002] for herniation prediction. Koh et al. [2010, 2012] require manual placement of 27 anatomical landmarks per disc plus manual segmentation of the disc, the vertebrae, and the spinal cord. Tsai et al. [2002] require manual segmentation of multiple structures in axial slices. Some other methods initialise a segmentation using manually placed points (Mateos et al. [2014], Štern et al. [2013]). Our method requires no manual intervention.

Ghosh et al. [2012] assume axial slices through discs at known locations and angles to detect and label discs. Kelm et al. [2012], Kadoury et al. [2013], Alomari et al. [2011a] perform experiments on fixed FOV, treating which vertebrae are in the scan as prior knowledge.

The fully automatic detection approaches are Glocker et al. [2012] on CT; Zhan et al. [2012] on wide-FOV, isotropic resolution scout MRI scans, and our approach (Lootus et al. [2013]) on narrow-FOV, multi-sliced sagittal clinical scans of thick slices, sometimes with slice gaps. Huang et al. [2009] are also fully automatic, however they do not label. Note also that not all methods include a labelling step, which is necessary to match radiological reports to vertebrae in measurement experiments.

In comparison, our framework provides a fully automatic pipeline from a hetero-

geneous dataset to predictions, with no manual interventions.

Manual Intervention in Detection and Segmentation Automatic systems have been proposed to localise either a point inside the disc (Alomari et al. [2009b, 2010a]) or a bounding box around the disc or the vertebra (Lootus et al. [2014], Oktay et al. [2014]) (see Table 2.2).

In many approaches, it is not made clear in the paper, whether any initialisation to the segmentation was required. In some cases, it is mentioned that a point in VB (Korez et al. [2014], Štern et al. [2011]) or in disc (Mateos et al. [2014]) was required to initialise. In measurement papers, a localization approach is often mentioned, but it is sometimes unclear whether the localisations actually used in the prediction experiments were automatic, or whether any manual steps were involved (Alomari et al. [2010b, 2011a], Ghosh et al. [2011b,a]).

Conditional Automation Note that some of the detection approaches promise only strictly conditional automation – e.g. the conditions may be having clusters of axial slices which cut through and angled as the discs (Alomari et al. [2011a]) or presence of six and only six discs (T12/L1-L5/S1) in the scan (Ghosh et al. [2012]) simplifying the problem.

Note that some of the detection approaches are restrictive on the scan type, or use a small number of scan protocols, scanner types, or sites, sometimes just one (Alomari et al. [2011a], Gamio et al. [2004], Ghosh et al. [2012]), and in some other papers, this information is not available (Ayed et al. [2011], Lu et al. [2012]).

Patient Conditions in the Dataset Note that the segmentation success can be dependent on the scan quality, and patient disease state. Mateos et al. [2014] measure segmentation success for discs of each Pfirrmann grade 1-5 separately, finding lowest overlap measure for grade 5. As evident in Table 2.2, column (radiological) 'Conditions' and 'Pain', in most detection and segmentation papers, the disease

state of the data is not cited. The exceptions are Alomari et al. [2011a], Ghosh et al. [2012], Lootus et al. [2013], Neubert et al. [2013], Oktay and Akgul [2013] where the radiological conditions present are specified.

Manual inputs requirements Note that some manual inputs (e.g. cropping the scan – are under "datasets"). Here only consider cropping discs out, and providing some points for initialisation of the segmentation.

May include manual slice selection, given the scan. Slice selection, can just take the middle slice because of the way it's placed, in most cases. But may want all pedicle-free slices.

2.3.5 Datasets and scanning protocol requirements

A general point is that most papers report results on relatively small datasets, with a two-figure number of patients/scans being typical. Exceptions in MRI are Oktay and Akgul [2013] who used 102, and ours (Lootus et al. [2014]) where we used 285 scans. In CT/X-ray, the exceptions are Štern et al. [2013] who used 105 CT scans, and Shen et al. [2013] who used 255 X-rays.

Some papers are specialized to specific scan quality or protocols. Neubert et al. [2013] use only dense 3D scans. Shen et al. [2013] assume the acquisition of two coplanar X-rays. Koh et al. [2011] use Magnet Resonance Myelography (MRM) images which are easy to segment (possibly by just thresholding) for stenosis prediction. Most databases are not representative of different machines and protocols. An exception is Neubert et al. [2013] who use scans with a range of TE and TR parameters, however with only 42 scans in total.

According to disease state, there are two principal groups of papers. Firstly, papers that quote automated anatomy localization results on (mostly) healthy patients (though, 3D disc detections have been performed on patients with serious pathology in 3D MRI by Zhan et al. [2012]). Second, papers that quote prediction

results on semi-automated localization regions. Since localization of diseased discs may be more difficult than on healthy ones, the localization errors from the first group do not directly translate to the prediction papers.

In contrast to many previous works, we perform both automated localization and prediction to test our framework on a large, heterogeneous set of 300 patients, with 1800 labelled discs containing a broad range of conditions, scan types and disease states, providing a more realistic evaluation of the clinical reality.

2.3.6 Summary

In summary, a large number of works have been published for spine detection, labelling, segmentation, and measurement. While high performance levels have been quoted (over 90% success rate for each), certain compromises have been made in almost all cases to simplify the very heterogeneous problem. Results have not been presented so far on datasets covering the full variation in terms of both anatomy and image protocol representing the clinical realities. Further work in spine imaging should address this issue, and we see our work as a step in that direction with our large, heterogeneous dataset described in more detail in the next Chapter.

Chapter 3

Dataset

In this Chapter we describe the dataset that we will use to illustrate and evaluate the framework throughout this thesis. The dataset consists of scans of 300 symptomatic back pain patients, in total 1800 discs with expert radiological annotation. The patients and images are described in Section 3.1, along with radiological labels in Section 3.2, and expert image pixel labels in Section 3.3.

3.1 The Patients and Images

Our clinically representative dataset contains clinical lumbar MRI scans of 300 symptomatic back pain patients (all experienced pain), each annotated by an expert radiologist. The patients exhibit a number of radiologically defined characteristics including degeneration, disc space narrowing, bulging, annular tears, central & foraminal stenosis, spondylolisthesis, endplate defects, Modic changes, facet joint arthropathy, herniation, and scoliosis. Of the patients, 132 were male, 162 female (6 unknown), with ages 10-88 years.

Each patient had one or more MRI studies associated with it. An imaging study is a set of series (scans) taken at a given time, one after the other, with the patient lying on their back, in supine position. A series is a scan containing a set of slices

acquired at given scanner settings. The mean acquisition of all the scans in the series is 16 minutes. Thus there is at times patient motion between the series acquisitions, or sometimes possibly between slice acquisitions.

Most patients had been involved in only one imaging study. Each imaging study contained either a T2, or STIR/TIRM/SPIR sequence sagittal scan (series). In all but one case, some axial T2 slices are available. In most cases, T1 sagittal and axial slices are available too. The number of slices in both sagittal and axial scans is typically 10-20.

The scans have isotropic in-slice resolution varying from 0.34 to 1.64 mm with mean at 0.78, median at 0.84 mm; and varying slice spacing from 3mm to 5 mm, with 4mm in almost all scans. Note that there are gaps between the slices, often around 0.4 mm.

The sagittal scans range in superior-inferior field of view (FOV), containing 7 to 23 vertebrae starting from the Sacrum, with median at 10 per scan. The most superior vertebra present in the scans is C4 (cervical level 4). While the scans contain a variable number of vertebrae, the first two sacral links are always present.

Out of the patients, approximately 90% were non-scoliotic, and 10% were scoliotic. For the non-scoliotic patients, the L-R FOV of the sagittal scans was usually disc-wide (excluding transverse processes – see Figure 1.5), excluding the vertebral processes, and was acquired with the spine centred in the scan. In the scoliotic cases, the scans were acquired to cover the discs fully at all lumbar levels within S-I FOV.

The axial slices were usually acquired around the bottom three lumbar discs (L3/L4,L4/L5,L5/S1), typically with 3-5 slices per disc. In some cases, all the slices were acquired in a block, e.g. parallel to each other; whereas in other cases, their orientations were made to follow the spine curve. It is important to note that the axial slices often do not necessarily intersect the disc exactly but contain both

vertebral bone and disc voxels.

In contrast to much previous work where images were acquired using the same scanner and protocol, the database is very heterogeneous, including scans from 25 different sites, from 14 different scanner models of various field strengths (0.6-3T). The sagittal scans were acquired under a wide gamut of T1, T2, and other sequences, whereas one scanner's T1 may not be the same as that of another (similarly for T2 and any other sequence). Where available, a T2 sequence was picked for the analysis; where not, either a STIR, SPIR, or a TIRM sequence. The sequence scan parameters ranged as follows: TE 69-139ms, TR 1180-1210ms, flip angle 90-180 degrees, echo train length 8-40.

Note that an implicit assumption in this work is that the image parameters have been chosen so that the image contrast is sufficient for whatever condition is specified/suspected. The contrasts have not been tailored for that – we had no control over this.

3.2 Radiological Labels

Each patient had associated with it a number of different radiological annotations for the six lumbar discs T12/L1...L5/S1. In total, there were 1800 annotated lumbar discs. The radiological annotations were provided by an expert spine radiologist with 25 years of clinical experience (among others).

To annotate the images with the radiological labels, the radiologist considered all the slices of all the image series available, both sagittal and axial; T1 and T2. Thus the labels were based on a 3D understanding of the spine.

Each patient was annotated with the following labels:

- **Pfirrmann grade** 1-5 (Pfirrmann et al. [2001]).
- **Narrowing** 0-3 (normal, slight, moderate, collapsed).

- **Annular tears (HIZ)** 0-2 (normal, present but not touching disc boundary, present and touching disc boundary).

- Separate scores for **Anterior disc bulging**, and **Posterior disc bulging**: 0-3 (below 0.5mm, 0.5-2.5, 2.5-4.5, over 4.5mm) Roughly, the convention of Fardon et al. [2014] was followed for assessment, with the exception that bulge was accounted to be more than 90, rather than more than 180 degrees.

- A number of characteristics for **disc herniation**: location C, PLL, PLR, FL, FR (central, posterolateral left, posterolateral right, foraminal left, foraminal right), **degree** 0-3 (below 2mm, 2-5mm, 5-10mm, over 10mm), **type** (protrusion, extrusion, sequestration). Associated nerve root compression 0-3 (none, touching, displaced, compressed). Roughly, the convention of (Fardon et al. [2014]) was followed for assessment.

- **Foraminal stenosis** (left and right separately) 0-3 (absent, mild, moderate, severe). Associated nerve root compression 0-3 (none, touching, displaced, compressed).

- **Central canal stenosis** 0-3 (absent, mild, moderate, severe).

- **Spondylolisthesis**: slippage of neighbouring VB-s with respect to each other: 0-4 (0%, 25%, 50%, 75%, 100% of VB width).

- **Endplate defects** 0-3 (absent, slight 1-5mm, moderate 5-10mm, severe over 10mm).

- **Modic changes** 0-1, upper and lower endplate separately for each disc, and types 1, 2, 3, and 1-2 mix given separately (Modic et al. [1988]).

- **Facet joint arthropathy** – rating 0-3, according to severity.

- **Other** - various, typically scoliosis, cysts, generalized endplate irregularity, remarks on surgery, transitional vertebrae.

Note that while intervals of quantified measurement sizes (e.g. below 0.5mm, 0.5-2.5mm, 2.5-4.5mm, over 4.5mm for bulge) are given for a number of the measures, the measurements were performed by eye according to an approximate assessment and may not match the given intervals. The intra-observer consistencies for the measurements were around 60-90% depending on the measurement. The intra-observer consistencies were estimated from a subset of 120 patients that were annotated twice by the same radiologist, with variable time delay between the two annotations (usually several months or more). Since the measurements were performed over a period of several years, there might be drifts in the ways of assessing some of the properties listed.

Some of the observations were based on sagittal slices only – such as Pfirrmann grade and narrowing, where the middle slice of the T2 scan was assessed, or Modic changes, spondylolisthesis, scoliosis or endplate defects – where all sagittal slices of either only T2, or both T1 and T2 modality were assessed. Other observations – including annular tears, disc bulging and herniation, stenosis, and facet joint defects – were based on both axial and sagittal slices. Nerve root compression was generally assessed based on axial slices only.

3.3 Image Labels

For each patient, I produced additional pixel-wise manual annotation based on repeated experience with a 2000-patient imaging database. To keep consistency, a rigid set of rules was followed. The types of annotations were both landmark points, segmentations, and vertebral level labels. The landmark points were the vertebrae corners, and posteriorly most protruding points of the disc in sagittal T2 slices. The segmentations were in T2 sagittal slices, vertebrae and disc segmentations, and in

T2 axial slices, spine contour outline (e.g. VB + disc) delineations. In addition, the migrated disc material beyond the vertebral edges was annotated separately, for assessment of bulge/herniation.

Segmentations. The ground truths were marked and stored as closed polygons in a tool developed in MATLAB, in a slice-by-slice basis (cross-sections annotated). The number of points used depended on the complexity of the shape. In total, 592 vertebrae, 1092 disc sagittal cross-sections were manually delineated, as illustrated with examples in Figures 3.1 and 3.2 respectively. The vertebrae delineations cover 300 unique patients (347 unique vertebrae) and the disc delineations 85 unique patients (87 unique discs). In addition, herniated mass is delineated in 902 cross-sections, covering 102 unique patients (105 unique discs). These numbers, along with segmentation results, are summarised in Table 3.1. All the annotated vertebrae were at levels 4-6 (L3,L4,L5), as those are the ones near the discs that are most commonly diseased. In total, 21 sections at level 4, 436 at level 5, and 135 at level 6. An L5 section was segmented in every patient.

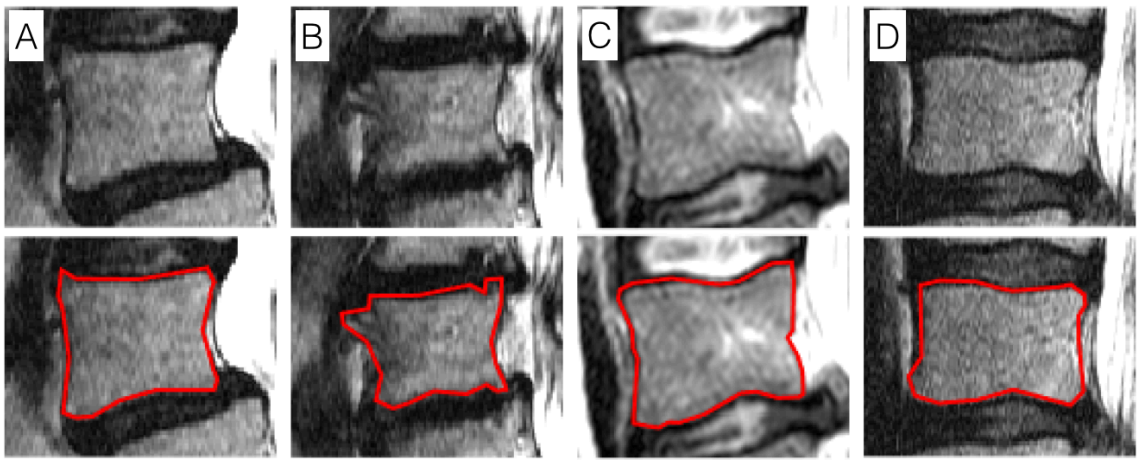


Figure 3.1: **Vertebrae segmentation Ground Truth.**

For **vertebrae**, the following procedure was followed. Delineate the vertebral body. The epidural fat at the back of the VB, and the dark image artefact line along with ligaments surrounding the VB were left out if it was sharp, or the line

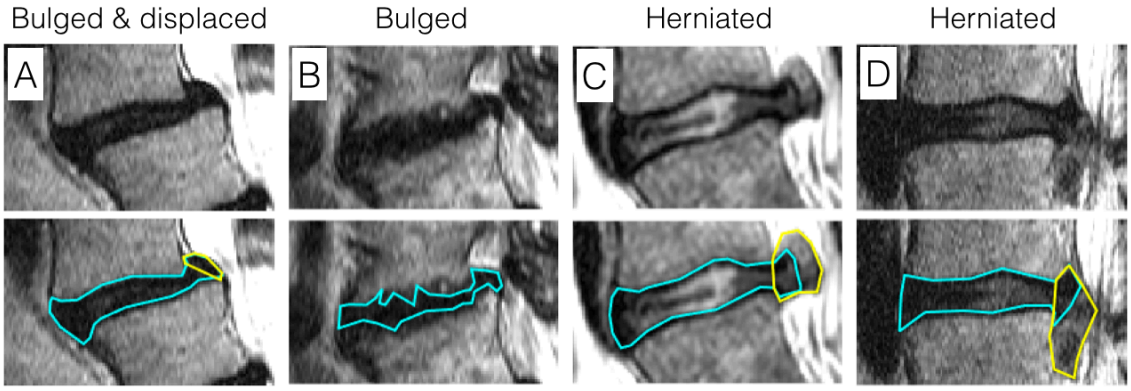


Figure 3.2: **Disc segmentation Ground Truth.** The cyan lines mark the disc border enclosed by annulus; the yellow borders the disc mass displaced across the vertebrae edges.

Anatomy	**Cross-sections**	**Unique VBs/IVDs**	**Unique patients**
Vertebra	$592/10^5$	347/1800	300/300
Disc	$1092/10^5$	128/1800	116/300
Hern.	$989/10^5$	120/1800	103/300
Axial	207/5000	*/900	43/300

Table 3.1: **Ground truth segmentations summary.** The number of vertebrae, discs and herniated disc masses in sagittal slices are listed.

was drawn in the middle of the image gradient, if it was a smooth gradient. Partial volumes, e.g. in case of Schmorl's node, on VB-IVD border, were usually left out. The delineation was performed by marking the corners first, and then points between the corners as necessary. The number of points used to define the polygon varied from 4 to 55, with mean at 21 and median at 21.

For **discs**, the following procedure was followed. Delineate the part of the disc surrounded by the annulus fibrosus outer shell. Herniated disc mass that had left the annulus was marked with a separate label, as illustrated by the yellow line in Figure 3.2. The following protocol was followed. First, mark the most anterior and most posterior points. Then, mark points between each pairs of corners, to get closer to the boundary. The number of points used to define the polygon varied from 4 to

52, with mean at 20 and median at 20.

Based on segmenting a randomly selected subset of twelve vertebrae and six discs cross-sections twice, approximately two months apart, the mean intra-observer variability for vertebrae in slices excluding pedicles was found to be 0.96 (median 0.96, range 0.94-0.98), and 0.88 (median 0.89, range 0.83-0.93) for discs. The greater intra-observer variability in the disc segmentations likely lends itself to the greater uncertainty around the anterior and posterior sides of the disc. This is largely due to the invisible boundaries between the annulus and surrounding ligaments, muscles, and vessels. Thus, the anterior boundary is often invisible. In addition, the VBs are over two times bigger than the IVDs (found as the ratio of the median areas of the samples), and thus the overlap measure for discs is around doubly more significant to errors of the same area at the shared IVD-VB boundary than for vertebrae.

Chapter 4

Detection and Labelling

The task dealt with in this chapter is the following: given an MRI scan of the lumbar spine, localise and label all the vertebrae present in that scan. In more detail, the input scan is a (sparsely spaced) stack of 2D sagittal images (slices), of variable slice count, in-slice resolution, and slice spacing, making up a 3D volume. The output consists of labelled tight bounding boxes around all the vertebrae in the scan. The pipeline is illustrated in Figure 4.1. Each bounding box is specified in 3D by its slice number, position, orientation, and scale, as illustrated in Figure 4.2.

Our method brings together two strong algorithms – the Deformable Part Model of Felzenszwalb et al. [2010] based on Histogram of Oriented Gradients (HOG) image descriptors (Dalal and Triggs [2005]) and efficient inference on graphical models (Fischler and Elschlager [1973], Felzenszwalb and Huttenlocher [2005]) – making the algorithm accurate, robust, and efficient on challenging spine datasets. The algorithm is also tolerant to varying MRI acquisition protocols, image resolutions, patient position, and varying slice spacing. It localises all the vertebrae present in a scan, and labels them correctly as long as the sacrum is present. Importantly, the method is applicable to standard MRI protocols.

The method has two distinct stages. First, vertebrae and Sacrum candidates are

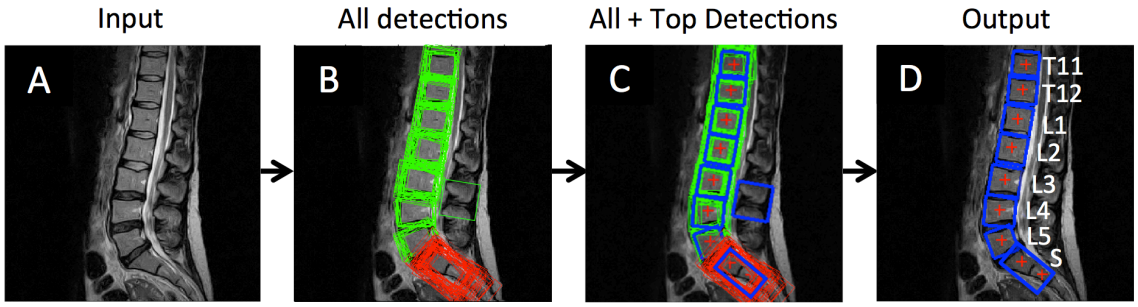


Figure 4.1: **Vertebrae Detection & Labelling Pipeline.** Note that a single slice is shown for visualisation purposes, whereas in reality detections from all slices are considered. (A) Input image. (B) All detections at all rotation angles and scales. The green rectangles are generic vertebrae, and the red rectangles are Sacrum candidates. (C) All detections, with top detections shown in thick blue line, and the "+" mark the ground truth vertebrae centre locations. (D) Output detection bounding boxes along with the ground truths and labels.

detected using a sliding window detector searching over position, scale, and angle (section 4.2; Figure 4.1). Second, a graphical model is fitted to the set of candidate detections to find the optimal spine layout and labelling based on the unary soft output score of the detector for each part, and a spatial cost between each pair of connected parts (section 4.3). The HOG descriptor captures the near rectangular shape of the vertebrae. We detect vertebrae rather than discs since the vertebrae shape is more consistent than the disc shape as the lumbar spine studies are more often aimed at targeting disc deformations, and more suitable to be modelled with HOG. Disc locations can easily be found after detecting vertebrae.

4.1 Why detect and label?

Detections localise anatomy, and can be used both to initialise further processing, and semantic scan navigation. Labelling allows the matching of vertebrae/disc level to radiological reports, and written communication of the results to physicians. The detections are also used to initialise further processing, e.g. segmentations in the next chapter.

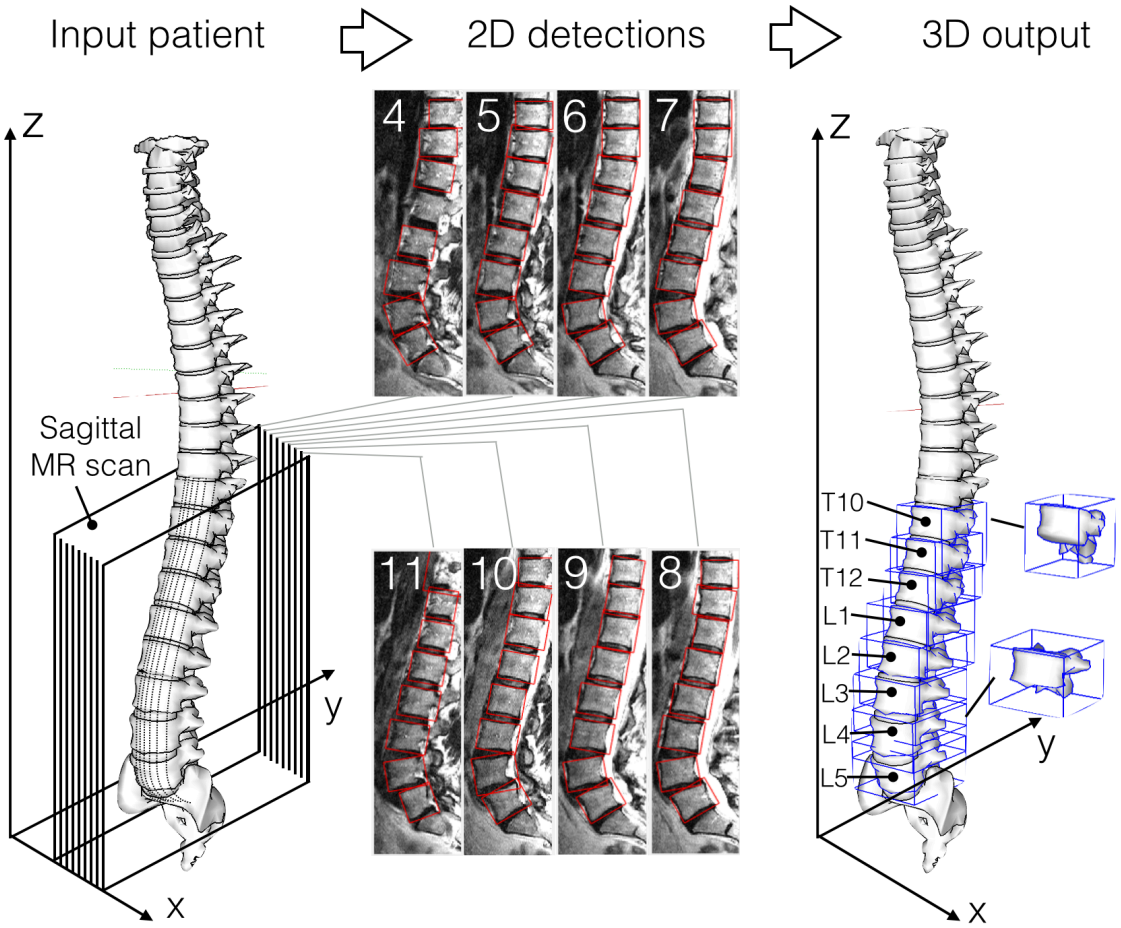


Figure 4.2: **The detection and labelling method operates in 2D, but reaches 3D in output.** An example input spine, a 15-slice sagittal MR-scan (crops of slices 4-11 shown) of it with 2D detections. The 2D slice-based detections are considered to arrive at 3D output.

4.2 Detection

The first step in the framework is vertebrae detection. We choose to detect vertebrae rather than discs because they are large, well defined objects with quite clear boundaries in T2 images and not typically subject to some of the degeneration processes that we are attempting to characterise.

Model. The vertebrae detection is implemented using two sliding window detectors. We learn one generic 2D detector for VBs, and another more specific 2D detector for the Sacrum part, comprising the VBs of the first two links of the Sacrum. Both the

detectors are visualised along with a set of training samples in Figure 4.3. The sliding window detectors are implemented using a Deformable Part Model (DPM) (Felzenszwalb et al. [2010]) over Histogram of Oriented Gradients (HOG) features, as described below.

We pick the DPM for detection as it has been proven to deal well with object detection in computer vision, being able to handle shape and intensity variations and different lightning conditions.

The detectors are run in all the 2D sagittal slices, and all the detections passed onto a non-maxima suppression step. Thus, if a vertebra is visible in any of the slices, it can be detected. In this way, the framework is able to handle scoliotic images as illustrated in Figures 4.4 and 4.5, and the detection process is invariant to slice spacing. The learnt HOG templates capture the rectangular shape of the vertebrae, with variations due to deformation, and the trapezoid shape of the first two links of the Sacrum.

4.2.1 Training

Both the sliding window detector for the VBs and the one for the Sacrum are trained on example vertebrae and Sacrums respectively from a training set. The positive training examples for the VB detector are tight bounding boxes around the VBs of T10...L5 vertebrae with the bounding box sides parallel to the vertebral facets as shown in Figure 4.3 A. The positive training examples for the Sacrum detector are tight bounding boxes around the first two links of the Sacrum, with one side parallel to the posterior side of the Sacrum as shown in Figure 4.3 B. The bounding boxes for both the VB and the Sacrum are defined by fitting a minimum bounding rectangle to landmarks on them – four for the VB and eight for the Sacrum. Each training sample is extracted from the slice intersecting the middle of the respective VB.

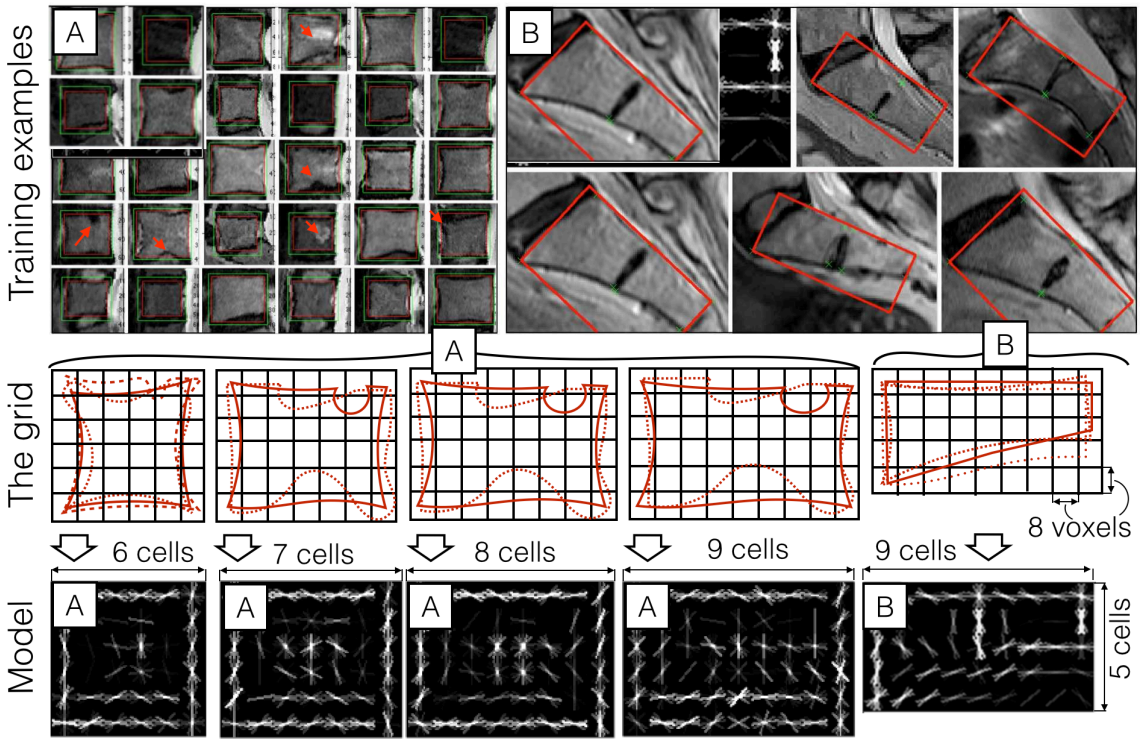


Figure 4.3: **Training the object detectors.** The training procedure and the HOG implementation of our vertebra detector are illustrated. Some training data samples and a learned HOG template are shown for both the generic vertebrae body (VB) detector (A) and for the Sacrum detector (B). The data samples have been hand-annotated with tight ground truth bounding boxes as shown above and explained in Figure 4.6. The model contains sub-models (components) of four different aspect ratios. Each is divided into a grid of cells. Outlines of three training example vertebrae are overlaid on each HOG grid. The lines in each cell of the model show the dominant directions and strengths (brightness) of gradients. Note that the same procedure can be followed to train detectors for axial slices and other objects (facet joints, say).

The detectors are trained using the DPM framework of Felzenszwalb et al. [2010]. For the VB detector, four HOG templates are trained, each with a different aspect ratio. The HOG templates are each 6 cells high, and 6, 7, 8, and 9 cells wide, corresponding to aspect ratios between 1 and 1.5. The HOG cell size for the VB model is 8×8 pixels. The HOG template for the Sacrum detector is 9 cells high by 5 cells wide, with 8×8 pixel HOG cell size. The HOG feature vectors are 31-dimensional, with 18 contrast-sensitive, 9 contrast-insensitive direction bins; and 4

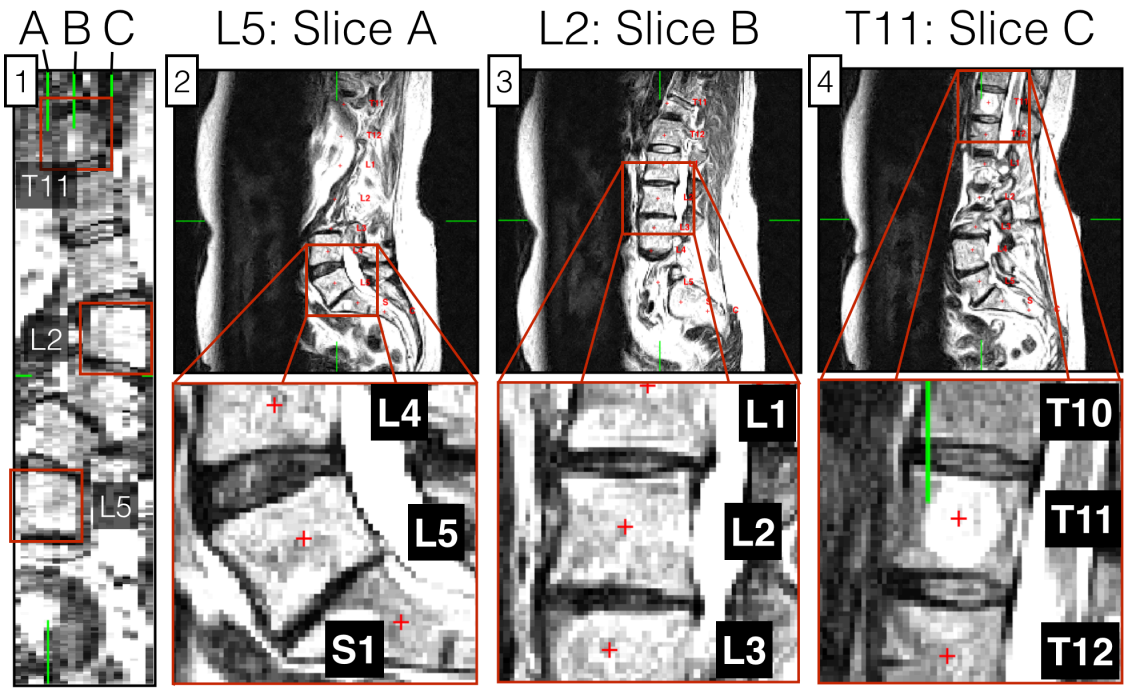


Figure 4.4: **Dealing with scoliosis during detection.** A sagittal T2 scan of a scoliotic spine is shown: in 1, a coronal section of the scan, and in 2-3, three different sagittal slices of the scan are shown. The vertebra L5 is best visible in slice A, vertebra L2 in L2, and vertebra T11 in slice C. Note that not all vertebra are visible in any one slice, unlike in the case of a non-scoliotic spine. The green lines A, B, C in the coronal section 1 mark the sagittal slices A, B, C shown in 2-3. In our dataset, around 90% patients are normal, and around 10% scoliotic.

dimensions capturing the block normalisations.

The model is learned iteratively in several steps, with latent positive samples mined by running the detector on the initial positive samples, collecting the strongest detections as latent positives, and re-training the detector using the latent positives. The negative samples for the vertebrae detector are first picked randomly from mid-slices with the vertebrae masked using a manually annotated polygon. Next, an iterative learning procedure is employed to pick hard negatives as false positive detections on the negative training images as detailed in Felzenszwalb et al. [2010].

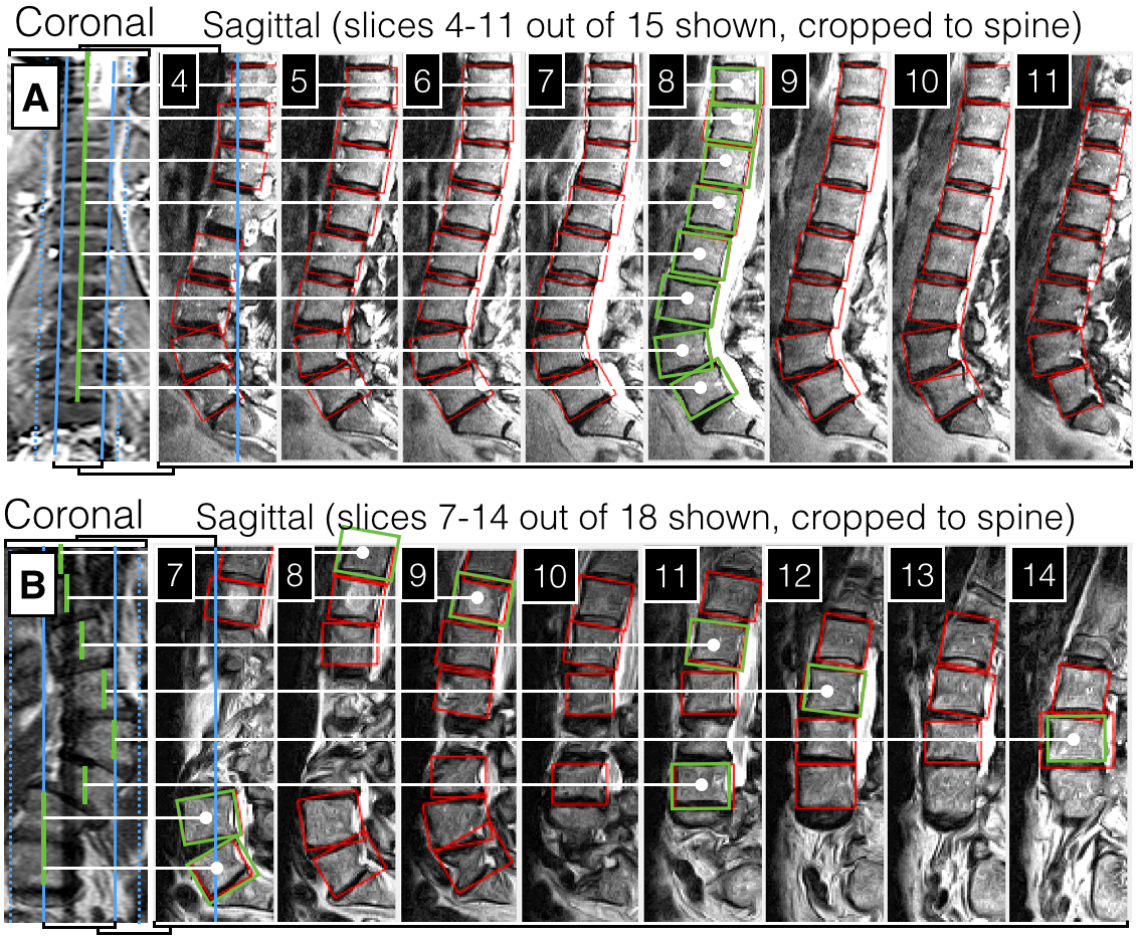


Figure 4.5: **The detection and labelling method operates in 2D, with its output reaching 3D.** For two spines A (straight) and B (scoliotic), one coronal MRI slice, and eight sagittal slices for two spines are shown, along with vertebrae detector outputs (red and green bounding boxes). The red boxes show all the detections after non-maxima suppression, and the green boxes show the automatically picked detections (by Graphical Model) constituting the optimal spine layout. Note that in spine A, the layout is in a single slice plane, whereas in spine B, across 8 slices, and that the slices have all been cropped for the visualisation.

4.2.2 Inference

There are two steps to vertebrae detection: first, candidate detection. Second, candidate selection by non-maxima suppression (NMS).

During the candidate detection step at test time, a previously unseen sagittal scan is taken as input, and tight bounding boxes around vertebrae candidates are returned as output. The candidate search is performed in all slices of the scan. The

VB and Sacrum detector are run on each slice, searching over position, scale, and orientation. In the search over orientation, the scan rotated by $-20°$ to $20°$ for vertebrae, and $-60°$ to $0°$ for the Sacrum, in 10 degree increments. A feature pyramid is calculated for each angle, with HOG cells placed densely next to each other. The feature pyramid has 10 levels per doubling of resolution (10 levels per octave), with the image resized and resampled to 2x the original size to 0.5x the original size from the finest to coarsest scale. All the detections at all positions, scales, orientations are collected and transformed onto the original test image coordinate system as shown in Figure 4.1.

A greedy non-maxima suppression algorithm is employed to remove most of the false positive detections in each slice as follows. First, the top-scoring bounding box is retained, and all bounding boxes overlapping it more than 50% are discarded. Next, the second-highest scoring remaining bounding box is retained, and the discarding and retention process continues until all the remaining bounding boxes have at most 50% overlap.

Next, the remaining bounding boxes from all the slices are collected, and the non-maxima suppression process is repeated to retain only highest-scoring bounding boxes across all the slices that have at most 50% overlap between any two boxes. These bounding boxes are next passed as input to the Graphical Model as described in Section 4.3 in order to eliminate any remaining false positives, and to label the vertebrae.

4.2.3 Evaluation Protocol & Performance

Dataset. Our clinically representative dataset contains clinical lumbar MRI scans of 300 symptomatic back pain patients (all experienced pain) with each annotated by an expert radiologist. The overall database is described in Chapter 3.

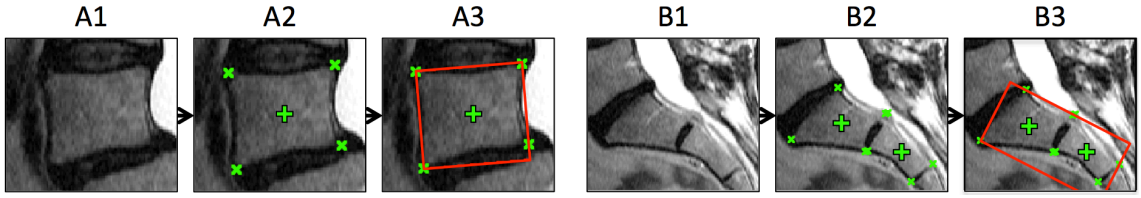


Figure 4.6: **The vertebrae ground truth annotation process.** A1-A3 show the generic vertebral, and B1-B3 the Sacrum annotation process. There are two types of annotation: single point (the green "+" in the Figure – used for testing) and bounding box (the red rectangle – used for training). Given an input (A1, B1), the points ("+" and "x") are hand-placed (A2,B2). The bounding box annotation is found as the minimal bounding rectangle to the "x" points around the vertebra / Sacrum boundary. There are four boundary points for vertebrae (A) and eight for the Sacrum (B).

Annotation. The scans were hand-annotated with two types of ground truth as illustrated in Figure 4.6: (i) All the vertebrae centres in all the scans are marked with a point ("+" in Figure 4.6), and labelled with the vertebrae name; and (ii) all the training scans plus some test scans are annotated with a tight bounding box around each vertebra (Figure 4.6 A3, B3). The tight bounding boxes were defined by points ("x" in Figure 4.6) along the vertebrae boundaries as shown.

Evaluation protocol. The detections are evaluated against vertebrae-centre and the Sacrum-centre ground truth points. A positive detection for the Sacrum is counted if a detected Sacrum bounding box contains the Sacrum ground truth point and does not contain any vertebrae centre ground truth points. A positive detection for the vertebrae is counted if a detected vertebra bounding box contains one and only one ground truth point for a VB, including the Sacrum. Note, this evaluation protocol ensures that the case where a large detection covers several vertebrae is not counted as positive.

Performance. The detection errors are plotted by vertebrae type in Figure 4.7. The mean detection error between the ground truth centre of the vertebrae and the centre of the detected bounding box is 3.3mm, with standard deviation 3.2mm.

Measure	T12	L1	L2	L3	L4	L5	S	Lumbar	Overall
Centre: Mean error (mm)	2.4	2.5	2.3	2.3	2.2	2.4	S	2.3	3.8
Centre: Std in error (mm)	1.8	1.9	1.3	1.2	1.3	1.7	S	1.5	1.6
Corner 1, S-P (mm)	4.5	5.3	5.0	4.7	4.7	5.4	8.2	4.9	10.7
Corner 2, I-P (mm)	6.2	5.9	5.8	6.1	6.5	7.2	7.5	6.3	11.8
Corner 3, I-A (mm)	5.2	4.9	4.3	4.1	4.0	3.8	18.0	4.4	10.5
Corner 4, S-A (mm)	4.5	4.2	4.0	3.9	4.0	3.6	4.7	4.0	9.8
Detection rate	99.3%	100%	100%	100%	100%	100%	100%	100%	Overall
Labelling rate	83%	83%	86%	88%	87%	87%	91%	86.9%	84.1%
Labelling rate (±1)	93%	93%	96%	98%	98%	92%	91%	94.7%	92.9%

Table 4.1: **Localisation and labelling evaluation.** The mean and standard deviation (std) of localisation errors are shown for the correctly detected and labelled vertebrae (identification rate 84% overall and 87% for lumbar). In addition the "count" – the number of vertebrae detected of each type – is provided, along with the mean width of each of the vertebrae in training set. By allowing the labelling to be correct to +/-1 vertebrae, the identification rates become 93% and 95% for all and lumbar vertebrae respectively. Note that some thoracic vertebrae results are omitted to save space.

Independent Sacrum detection (without graphical model) with local non-maxima suppression shows 98.1% recall at 48% precision. Independent general vertebrae detection (without graphical model) shows 97.1% recall at 9.1% precision. This reflects that the threshold for vertebrae detection was set lower than for sacrum detection, in order to still keep detection candidates for severely deformed vertebrae. This likely has to do with greater similarity of vertebrae with surrounding structures, in comparison to the Sacrum. The numerical detection results are listed by vertebra type in Table 4.1. The median localisation error does not show a strong dependence on the vertebrae type. More than half of all vertebrae have detection errors between one and three mm.

Our method works well on very challenging examples with various imaging and anatomical anomalies. The identification results compare favourably to other approaches in the literature, although direct comparison is not possible since the algorithms have been evaluated on different datasets. Glocker et al. [2012] report

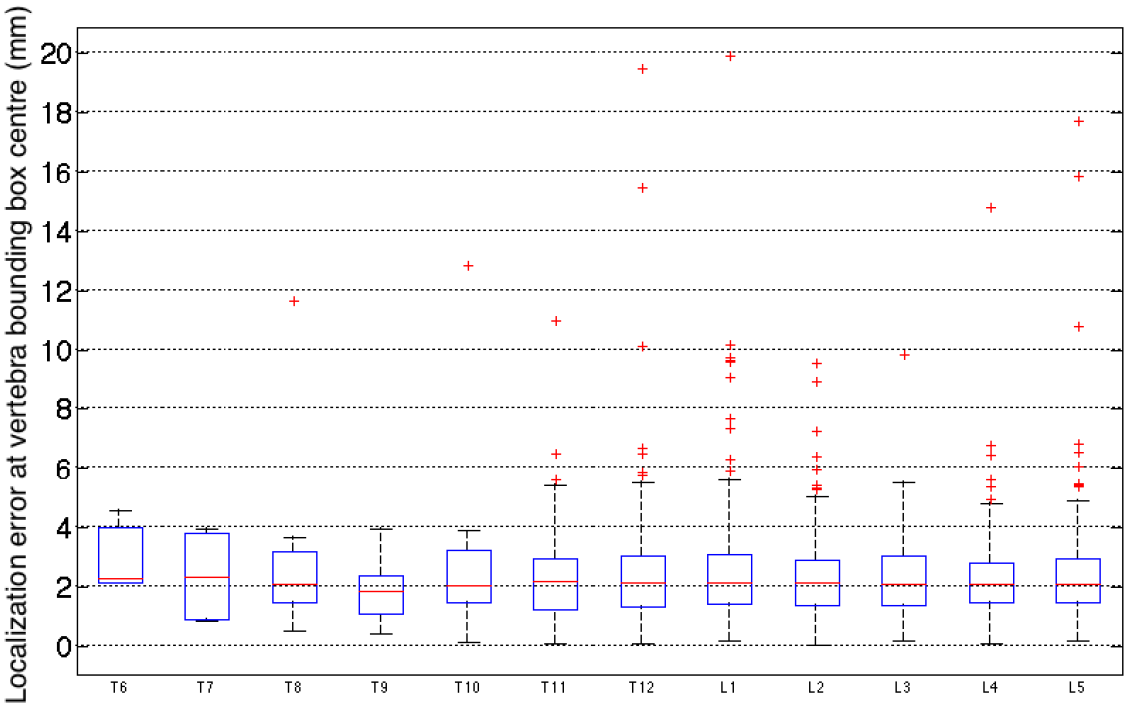


Figure 4.7: **Localisation error by vertebrae type.** Boxplots representing detection errors are shown. The error for a given vertebra type is calculated as the distance between the centre of the detected bounding box and the ground truth vertebra centre, divided by the mean width of that vertebra. The mean vertebrae widths are evaluated based on the bounding boxes on the training set. The horizontal line in the middle of each box is the median error, and the bottom and top of each box are the 25 and 75 percentile errors respectively. The bottom and top error bar end are the 5 and 95 percentile errors respectively, and the '+' denote statistical outliers.

median identification error of 81% with median localisation error below 6mm on CT images. Zhan et al. [2012] detect discs and vertebrae in isotropic MRI scans with 97.7% "perfect" labelling rate as assessed by a medic but do not report detection errors. Pekar et al. [2007] report 83% correct labelling rate on 30 lumbar MRI scans. Our method correctly localises the centres of vertebrae out of the mid-sagittal slice in scoliotic cases such as scan (f) in Figure 4.10.

Two typical failure cases are shown in Figure 4.8.

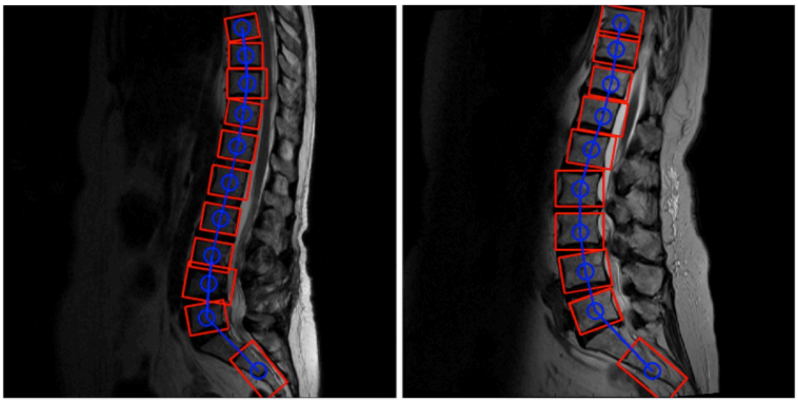


Figure 4.8: **Typical detection failure examples.** Note how the Sacrum detector fails, detecting the sacral links S2-S3 rather than S1-S2.

4.3 Labelling

The next step is to select and label the detections from Step 1. For this, a simple chain graphical model is employed. Given the quality of the predictions from the vertebrae detector, it is not necessary to use a more complete graphical model, such as a fully connected graph, and this means that fewer parameters need be learnt than in previous approaches (e.g. in Kadoury et al. [2013]).

Model. The vertebrae are connected in a chain modelled as a graph (Felzenszwalb and Huttenlocher [2005]). The spine layout is given as a configuration $L = (l_1, l_2, ..., l_{n-1}, l_n)$ where l_i are the vertebra locations, with l_1 the C1 and $l_n = l_{25}$ the Sacrum. The optimal configuration L^* of the graphical model is

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{v_{i,j} \in G} d_{ij}(l_i, l_j) \right) \tag{4.1}$$

where l_i and l_j denote the vertebrae locations $l = (x_i, y_i, height_i, width_i, \theta_i)$ given by their location (x, y), size $(height, width)$, and orientation θ_i. The best model fit minimises the sum of the unary appearance mismatch terms m_i from the part detectors output and the spatial deformation cost d_{ij} for connected pairs ij of parts, laid at l_i and l_j respectively. The last appearance term value, m_{25}, comes from

the Sacrum detector, and the rest of the appearance term values come from the universal vertebra detector. The terms m_i are equal to the negative of the classifier output. Since there might be fewer vertebrae visible in the scan than the model has, an additional "out-of-FOV" state is available for those vertebrae that indicates if they are not included in the scan. The appearance term m_i value for those vertebrae takes a constant penalty value learned on the training set as described in Potesil et al. [2013].

The spatial deformation cost is a sum of four box functions S, T, U, V on pairs of adjacent vertebrae in the chain:

$$d_{ij}(l_i, l_j) = S(A_i/A_j) + T(x_i - x_j) + U(y_i - y_j) + V(\theta_i - \theta_j) \tag{4.2}$$

where A_i, A_j are the areas, x_i, x_j & y_i, y_j the positions, and θ_i, θ_j the angles of the adjacent vertebrae i and j. The box functions F take a low constant value if their argument values k are within favourable interval $[k_{min}, k_{max}]$ and a higher constant value if their arguments are outside that interval:

$$F(k) := \begin{cases} 0, & k(\in k_{min}, k_{max}). \\ c > 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

where c is a constant to be learned for each of the four box functions (S, T, U, V).

4.3.1 Training

The edges k_{min} and k_{max} for the box functions S, T, U, and V are found as the minimum and maximum argument values of those functions on the training set (e.g. the minimum and maximum x-distance between L1 and L2 for T, etc.).

4.3.2 Inference

The model-fitting algorithm takes as input the detections after non-maxima suppression described in the previous section, and gives as output the placement and labels of all vertebrae in the volume (including slice index). Three-dimensionality is dealt with by picking the slice with the strongest detector score for each part. The method deals with detections in multiple slices by ignoring the slice index in inference. However, detections in all slices are considered, and the slice index of the picked detection is returned in output. To speed up the fitting process, a Viterbi message passing scheme from Felzenszwalb and Huttenlocher [2005] for fast inference in $O(nh^2)$ time is employed where n is the number of parts and h the number of candidates per part. Typically, there are around $h = 100$ candidate positions per part, plus an "out-of-FOV" state for each part.

To go from the 2D detections to 3D bounding boxes, for each labelled bounding box, detections from all neighbouring slices are considered, and at most one per slice picked to build up the 3D box around the VB. For each slice, the strongest detection in that slice which has at least 50% overlap with the labelled box is picked.

4.3.3 Evaluation Protocol & Performance

Dataset. The dataset consists of 371 MRI T2-weighted lumbar scans of 300 unique patients, acquired under various protocols. The scans contain normal and various abnormal cases as illustrated in Figure 4.9. This is a subset of the full dataset presented in Chapter 3. The dataset is split into 80 training and 291 testing images. The scans have isotropic in-slice resolution varying from 0.34 to 1.64 mm with mean at 0.78, median at 0.84 mm; and varying slice spacing from 3mm to 5 mm, with 4mm in almost all scans. The scans range in fields of view, containing 7 to 23 vertebrae starting from the sacrum, with median at 10 per scan.

The algorithm is evaluated on the 291 lumbar spine test images, which have

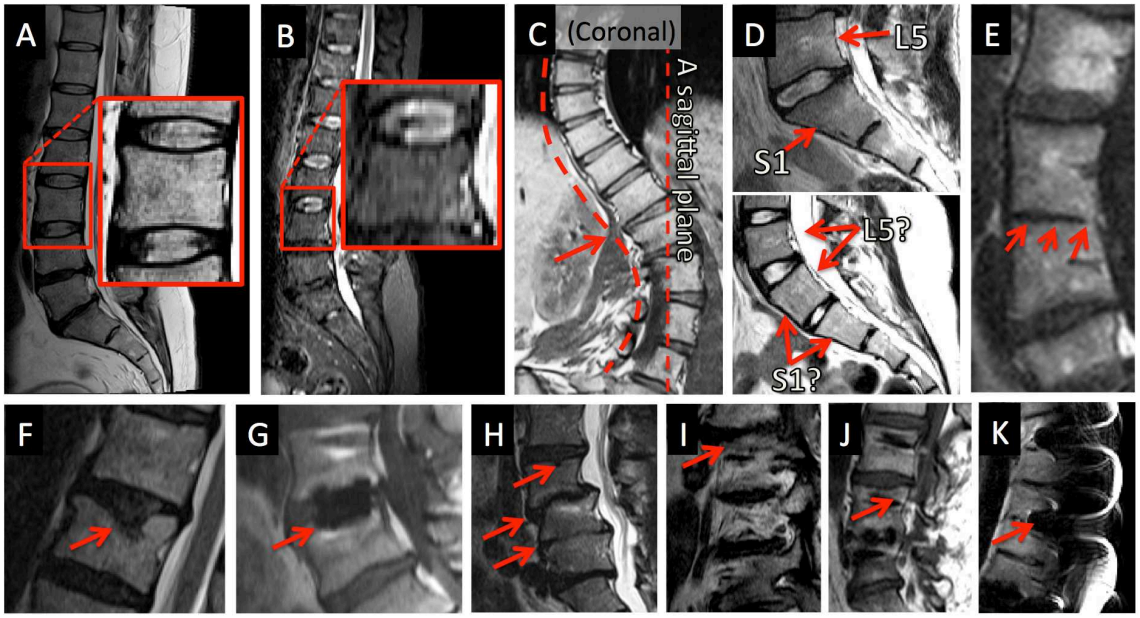


Figure 4.9: **Spine variation in our data.** A collection of example images showing assorted image, anatomical and pathological modes of global variation of the spine shape, and local variation of the vertebrae, and the disks. Our algorithm is robust to all those variations. Abnormalities have been highlighted by the red arrows. (A) Normal spine with a zoom on a normal vertebra. (B) A low-resolution image. (C) A coronal view of a scoliotic spine, resulting in the spine not being cut by a single sagittal slice. (D) Top: a normal sacrum, with unambiguous L5, S1 labelling based on shape and S1 and L5 orientation. Bottom: a sacrum with ambiguous L5, S1 labelling based on their shape and orientation. (E) Joined vertebrae. (F-J) Pathologically deformed vertebrae and disks. (K) Magnetic susceptibility imaging artefacts.

variable number of vertebrae visible. Example outputs are shown in Figure 4.10, and statistical results on localisation+labelling error over the test set are tabulated in Table 4.1 by vertebrae type.

Results. We achieve 84.1% correct identification rate overall, and 86.9% for the lumbar vertebrae. If the assigned labels are allowed to be shifted by one vertebra in either direction, the rates are 92.9% and 94.7% respectively. Typically, the full detection and labelling process from input to output takes less than a minute, with the majority of time spent on candidate detection.

4.4 Discussion

We have presented a HOG-based algorithm to localise vertebrae in lumbar MRI scans of the spine that is simple, accurate and efficient. We demonstrate robustness to severe deformations due to diseases, image artefacts, and a wide range of resolution, patient position, and acquisition protocols on a challenging clinical dataset. It is possible to extend the method to completely general FOVs if required, by taking other anatomical context into account (Glocker et al. [2012]).

Our contribution is in the combination of two powerful algorithms to create a highly robust vertebrae detection and labelling system and its evaluation on a dataset showing various heterogeneities.

Application to CT images. Although the method has been principally designed for MR images, it is directly applicable to CT images as shown in Figure 4.11. No retraining is required for detection on CT due to the high generalisation of HOG detectors. Future work might include rigorous evaluation on CT images.

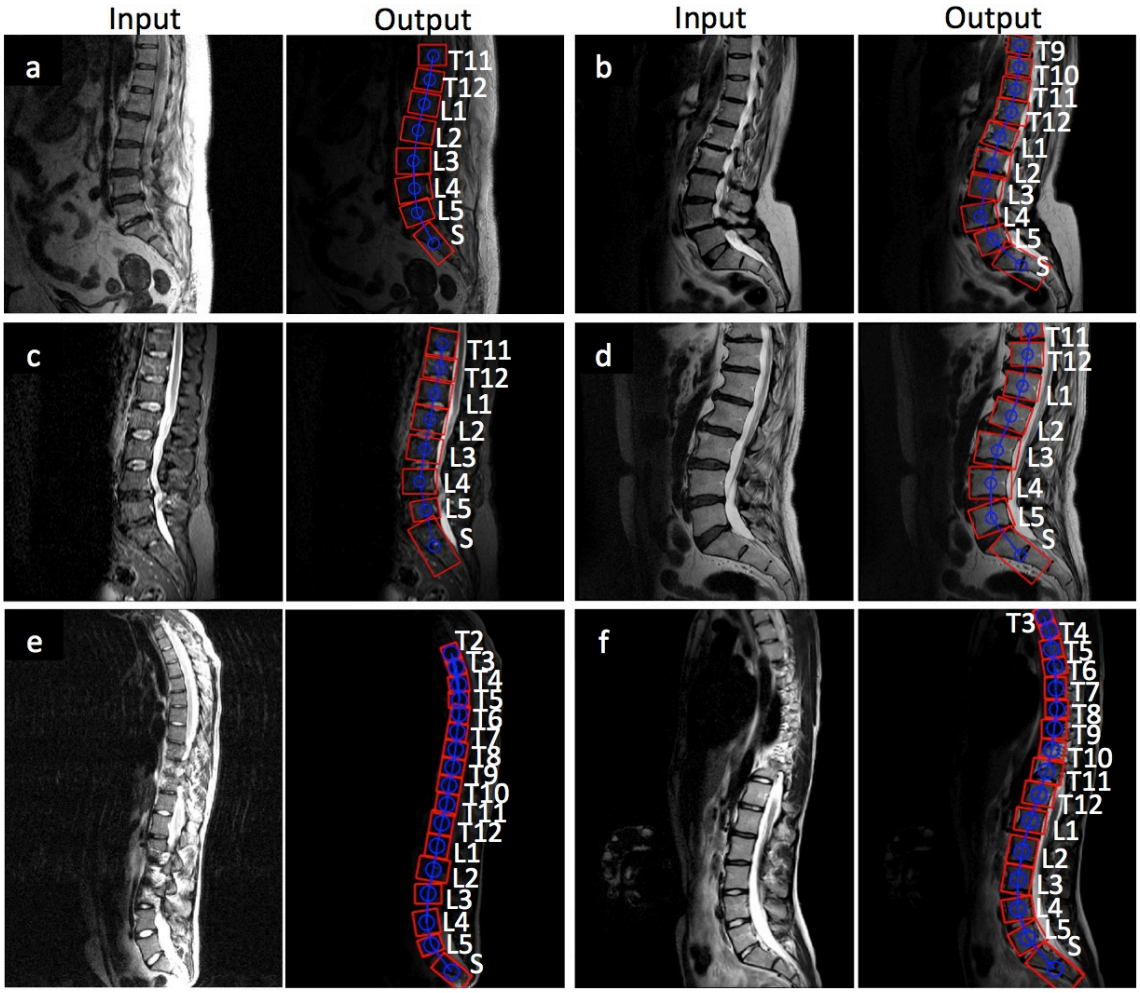


Figure 4.10: **Example results: detection+labelling.** Input and output are shown for six different scans a-f. The thick solid line rectangles show the detections for each vertebrae, along with their anatomical labels. Note how the algorithm is robust to varying FOV, resolution, and anatomy. Note, for visualisation purposes, only the mid-sagittal slice is shown for each scan, and all bounding boxes are projected onto it.

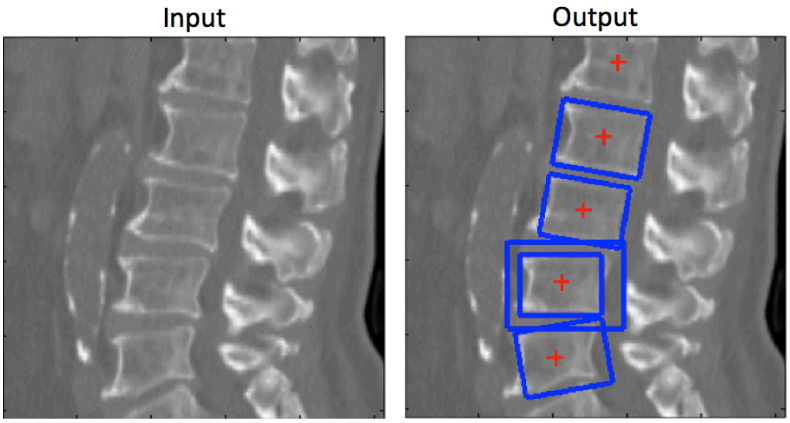


Figure 4.11: **Detection on CT images with detectors trained on MRI.** Detectors trained on MR images can also successfully localise vertebrae in CT scans, indicating the robustness of the method to varying image appearance.

Chapter 5

Segmentation

The task dealt with in this Chapter is the voxel-wise semantic segmentation of sagittal spine images. The goal, given an image, along with labelled vertebrae bounding boxes from the previous Chapter, is to return a label indicating whether the voxel is part of a vertebra, a disc, or background (other tissue), for each voxel in the image. The process is illustrated in Figure 5.1. In more detail, the input scan is a (sparsely spaced) stack of (usually thick, 3-5mm) 2D sagittal slices, of variable slice count, in-slice resolution, and slice spacing, making up a 3D volume. The output consists of pixel labels for vertebrae and discs where they are visible.

Our method builds on a powerful algorithm from the Computer Vision community – Graph Cuts (Boykov and Jolly [2001]) – an efficient framework to rapidly achieve globally optimal solution to an energy minimisation problem defined on a Markov Random Field (MRF) representing the pixel labelling. The energy function includes both boundary and regional terms measured from the image. To deal with the challenge of heterogeneous dataset of various resolutions, left-right FOVs, and slice spacings, we perform the segmentations in a slice-by-slice basis, in 2D. Combined afterwards, the full stack of 2D segmentations constitutes an approximation to the 3D shapes of the VBs and IVDs.

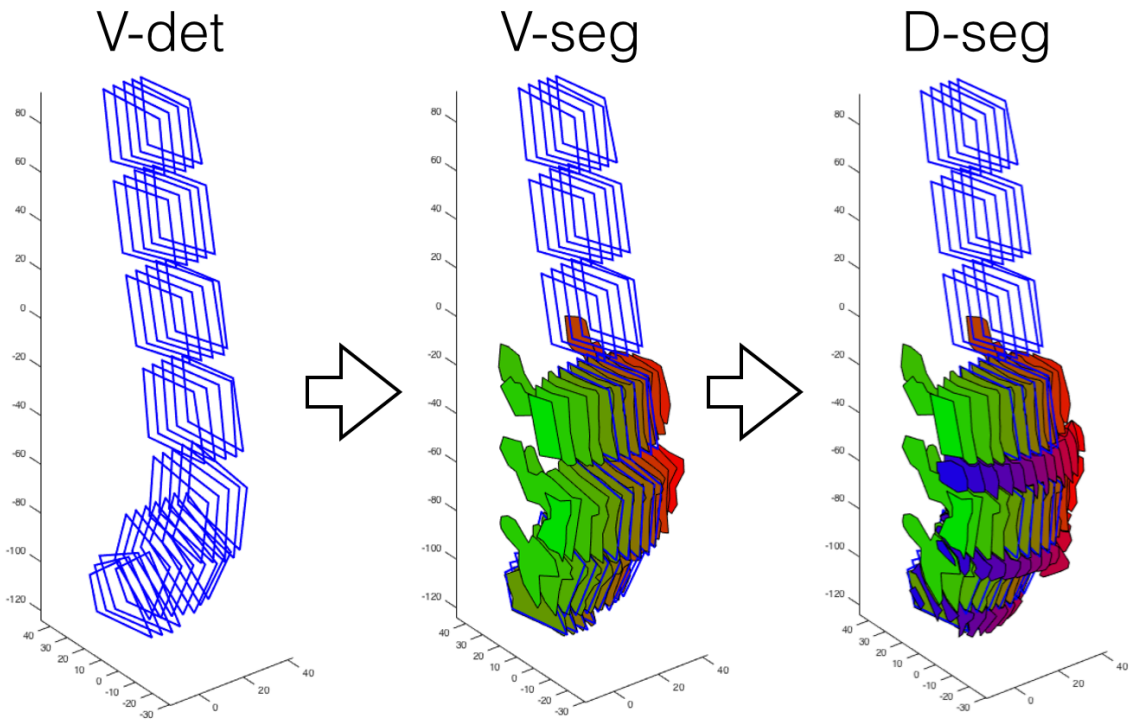


Figure 5.1: **The segmentation process**. The process is initialised using the vertebrae bounding boxes (blue rectangles under V-det) in sagittal slices. The vertebrae are segmented first (V-seg) and then the discs (D-seg), using the vertebrae segmentations for initialisation. Note that only a subset of all segmentations are shown for visualisation purposes, with the colour gradients between slices to help distinguish slices.

The vertebrae are segmented first, followed by the discs. The vertebrae segmentations are initialised, for each VB individually, with seeds derived from VB detections from Chapter 4, and the disc segmentations similarly, with seeds derived from the vertebrae segmentations.

The rest of this Chapter is organised as follows. In Section 5.1, we motivate the segmentation task clinically, and within the pipeline, discussing the challenges in segmentation in Section 5.2. We motivate and explain our segmentation method in detail in Section 5.3, and assess its performance in Section 5.4. Finally, a discussion and conclusion is provided in Section 5.5.

5.1 Why segment?

Fundamentally, segmentation helps to crystallise on a fine-grained level an observer's understanding of the tissue distributions in the scan, to characterise the anatomy. It is a more detailed task than detection, helping to externalise a finer internal understanding, forcing to define anatomical units more strictly.

Clinically, segmentation can thus help with clear communication between physicians, and in addition, extract information further to what the measurements discussed in this thesis may provide.

Within our spine measurement pipeline, it can provide spatial cues further to detections to define support regions for feature extraction. Both signal features can be extracted from segmented areas, and sizes of and distances between segmented objects measured. In the pipeline applied clinically, interactive segmentation can be useful to allow for debugging cases where segmentation has failed.

While there can be clear benefits to high quality segmentation, it is often an ill-posed and challenging problem in Computer Vision, with Ground Truth hard to define, as explained in the next Section in the context of spine MRI, and algorithms brittle. We still choose to attempt it, and assess its performance both on its own in this Chapter, and its effect and importance on the radiological measurements in Section 6.5 of the next Chapter.

5.2 Why is spine MRI segmentation difficult?

The definition of ground truth in both disc and vertebrae segmentation is extremely challenging. In disc segmentation, this is mainly due to (i) **intensity similarity** (and lack of visible boundary) between the ligamentous annulus in discs, and the surrounding anatomy, either adjacent or connecting to the disc (particularly ligaments, vessels, muscles, nerves), (ii) **partial volumes** due to high slice thickness

(median across dataset of 4mm, vs. median intra-slice pixel size of 0.7mm) – see illustration in Figure 5.2, (iii) **large shape and signal/texture variation** (developmental variants, and diseases such as degeneration, herniations, endplate defects, etc.), (iv) **confusion** between image features manifestation reasons (e.g. a bright spot could be either a HIZ disc or limbus vertebra) or an imaging artefact, and (v) **other image quality problems** such as imaging noise, contrast variation, and other artefacts.

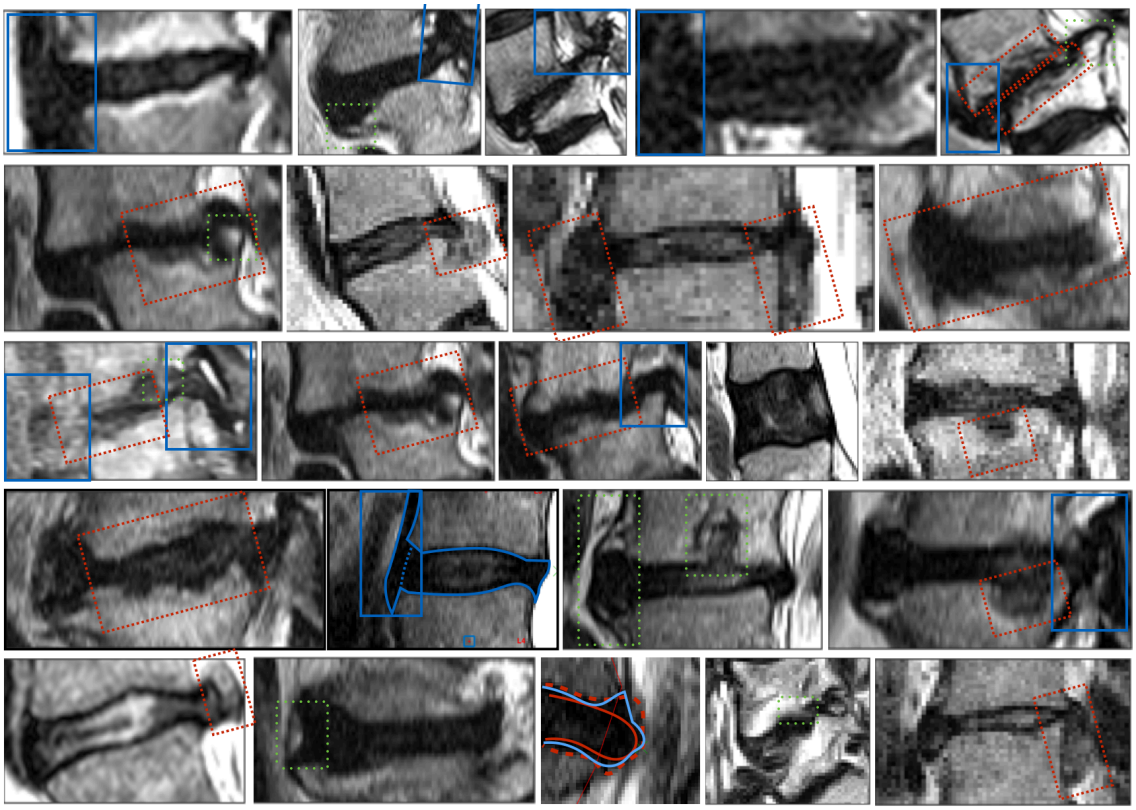


Figure 5.2: **Challenges**. The principle challenges in disc segmentation – partial volumes (red angled dotted box), invisible boundaries between tissue classes (blue solid boxes) and pointy deformations of the vertebrae (green dotted box). The partial volumes happen due to the high thickness of the slices, especially at places of pointy anatomical structures (e.g. herniations); the invisible boundaries commonly between the disc annulus and either ligaments, muscles, or blood vessels.

The above reasons render the disc segmentation task both hard to define (e.g. ground truth) and execute (e.g. automatic segmentation). The challenges involved

are exemplified with a number of cases in Figure 5.2. The vertebrae segmentation is challenging due to similar reasons, except that in (i) most commonly, the posterior epidural fat can be confused with marrow.

The definition used to define Ground Truth segmentations in this work is explained in detail in Section 5.4.

5.3 Segmentation Method

5.3.1 Graph Cuts in Computer Vision

Many vision problems, especially in early vision, can naturally be formulated in terms of energy minimisation. The classical use of energy minimisation is to solve the pixel-labelling problem, which is a generalisation of such problems as stereo, motion, and image restoration.

The attraction of graph cuts in computer vision is that they give a guarantee to the global minimum of the energy function. For example, imperfections in a globally optimal solution are guaranteed to directly relate to the cost function rather than to a numerical problem during the optimisation. Thus, global methods can be more reliable and robust.

Let $\mathbf{I}$ be an image and $\mathbf{L}$ be a partition of the image into foreground (object) and background components. In particular, let I_i denote the intensity of the i-th pixel and let L_i be equal to $+1$ if the pixel belongs to the object and to 1 otherwise.

A standard form of the energy function is

$$E(\mathbf{I}, \mathbf{L}) = \sum_{i \in \mathcal{P}} U(L_i|\mathbf{I}) + \sum_{(i,j) \in \mathcal{E}} P(L_i, L_j|\mathbf{I}) \tag{5.1}$$

There are two terms to this energy function: a unary term $\sum_{i \in \mathcal{P}} U(L_i|\mathbf{I})$, and a pairwise term $\sum_{(i,j) \in \mathcal{E}} P(L_i, L_j|\mathbf{I})$. Each unary term is a function of label L_i of a single pixel i, summed over the set of all pixels $\mathcal{P}$. The pairwise term measures the

edges in the image, and the edge price is paid if the label changes. Each $P(L_i, L_j|\mathbf{I})$ is a function of two pixel labels L_i, L_j, summed over all pairs of pixels connected in a neighbourhood system $\mathcal{E} \subset \mathcal{P} \times \mathcal{P}$[1]. Each $P(L_i, L_j|\mathbf{I})$ is a penalty for discontinuity between positions i and j in the image. Thus the segmentation energy combines region-based properties with boundary regularisation in a single, global optimisation framework. The neighbourhood, the pairwise, and the unary terms are next explained in more detail, along with examples.

The **unary potential** $\sum_i U_i(L_i|\mathbf{I})$ a function derived from the observed data that measures the cost of assigning the label L_i to the pixel i. For each pixel i, the element $U(L_i|\mathbf{I})$ takes a low value if the pixel value I_i is close to the model, and high value if it is far from the model. In the simplest case, the unary potential $U(L_i|\mathbf{I})$ is calculated based on just the pixel I_i value, e.g. it simplifies to $U(L_i|I_i)$, as in Boykov and Jolly [2001]. More generally, it can be a value of the pixel itself and its neighbours, e.g. a weighted combination of the pixel and its 5-pixel-radius neighbourhood (Lu et al. [2012]). Boykov and Jolly [2001] calculate the unary term from the intensity histograms under the seeds. Rother et al. [2004] calculate them from GMMs fitted to colour models. In Lucci et al. [2010], the unary term is calculated based on a probability $P(L_i|\mathbf{f}(\mathbf{I}))$ computed from the output of an SVM $U(L_i|\mathbf{I}) = 1/(1 + P(L_i|\mathbf{f}(\mathbf{I})))$.

The **edge system** $\mathcal{E} \subset \mathcal{P} \times \mathcal{P}$ typically in the 2D case connects the pixel to its 4 or 8 nearest neighbours, or in 3D, to its 6 or 26 neighbours (Boykov and Jolly [2001]).

The **pairwise** term measure edges and should be interpreted as a penalty for label discontinuity between i and j. The penalty is large when pixels are similar (e.g. in intensity) and close to zero if very different. The costs can be based on local intensity gradient, Laplacian zero-crossing, gradient direction, geometric or other criteria

[1]Without loss of generality, we can assume that $\mathcal{E}$ contains only ordered pairs p, q for which $p < q$ since we can combine two terms $V_{p,q}$ and $V_{q,p}$ into one term.

(e.g. Mortensen and Barrett [1995]). In the simplest case, the term is $P(L_i, L_j|I_i, I_j)$, e.g. based on the two pixel intensities only, $\gamma \sum_{(i,j)\in\mathcal{E}} \exp(-\beta|I_i - I_j|^2)$(Boykov and Jolly [2001]), or their distance in RGB colour space, $\gamma \sum_{(i,j)\in\mathcal{E}} \exp(-\beta||z_i - z_j||^2)$ (Rother et al. [2004]). In other cases, it can be from a feature detector (Lucci et al. [2010], Moschidis and Graham [2010]), $P(L_i, L_j|\mathbf{I}) = \exp(-\epsilon\dot{(}R_i+R_j)^2)$ where R_i and R_j are the responses of the edge detector on pixel i and j respectively. For directed edges, the penalty depends on the sign of the edge energy, e.g. the cost depends on whether it is directed from Obj to Bkg, or Bkg to Obj: $P(L_i, L_j|\mathbf{I}) \neq P(L_j, L_i|\mathbf{I})$. Typically, $P(L_j, L_i|\mathbf{I})$ is a non-decreasing function of $||L_i - L_j||$ where I_i might be a vector Boykov and Veksler [2006].

The segmentation is defined as the minimiser $\bar{\mathbf{L}} = \arg\min_{\mathbf{L}} E(\mathbf{I}, \mathbf{L})$ of the energy using GraphCut Boykov et al. [2001]. Note that extra constraints may be included, for example star convexity Gulshan et al. [2010].

The energy function has to be submodular in its pairwise terms to be solvable by graph cuts. According to Kolmogorov and Zabih [2004], the energy in Eq. 5.1 can be optimised exactly with a graph cut if all the pairwise terms are submodular, where a binary function g of two variables is submodular if $g(0,0)+g(1,1) \leq g(1,0)+g(0,1)$.

The segmentations are initialised based on "hard constraints". These set up the problem, fixing a subset of $\mathcal{P}$ as "Obj", and another subset as "Bkg". The value of the rest of $\mathcal{P}$ is decided in the optimisation based on the interplay of U and P.

Schemes have been developed to segment more than one object class, e.g. with more than two label values: alpha-expansion, beta-swap Boykov et al. [2001], and for iterative solution called Grabcut Rother et al. [2004] where the unary terms are iteratively updated based on the improving segmentation.

5.3.2 Model – Graph Cuts for MRI Segmentation

In our case, the system is an MRI slice, e.g. a 2D image matrix. For MRI segmentation we use a standard energy minimization formulation Boykov et al. [2001] defined over the pixels (x, y) of the slice. The edge system $\mathcal{E}$ is here the standard eight-way connectivity scheme. The pairwise potential $P(L_i, L_j|\mathbf{I})$ favours neighbour pixels to have the same label unless an edge separates them:

$$P(L_i, L_j|\mathbf{I}) = \gamma \exp(-e_j(\mathbf{I})/\beta) \tag{5.2}$$

where $e_j(\mathbf{I})$ is the edge intensity at pixel j and $\beta = \langle e_j(\mathbf{I}) \rangle$ is the average edge intensity in the image. Note that the edge is measured only at pixel j, as defined by the edge system $\mathcal{E}$ (here j is the pixel more on the right/south). The parameter γ is learnt on the validation data. The unary terms U_i are given by negative log-likelihoods of the foreground-background spectral model: $U_i = -\log p(I_i)$. The segmentation is defined as the minimiser $\arg\min_{\mathbf{L}} E(\mathbf{I}, \mathbf{L})$ of the energy using Graph-Cut Boykov et al. [2001].

5.3.3 Implementation

In this section, the implementation of the Graph Cut algorithm discussed in the previous section is discussed first for vertebrae, and then for disc segmentations.

The VBs are segmented first since their larger size than IVDs allows more confident placement of hard constraints to initialise the segmentations. The IVD segmentation is easily initialised based on the VB segmentation.

The decision to segment on a slice-by-slice basis, and combine into 3D later is motivated by the variation in image resolutions, and coarse slice spacing.

5.3.3.1 Vertebra Segmentation

In sagittal scans the vertebrae segmentation is performed sequentially, independently for each vertebra in the scan, independently for each slice. The segmentations are performed in a rotated coordinate system where the vertebra endplates are horizontal. The segmentation for a given vertebra is automatically initialised according to its bounding box detection. In a patient-adaptive graph cuts (Boykov and Jolly [2001]) framework, the object (Obj.) and background(Bkg.) seeds and three-component Gaussian Mixture Models (GMMs) for region modelling are set according to the bounding boxes. The seeds are obtained as follows: the foreground seed by eroding the bounding box of the vertebra to half its size, and the background seeds by dilating the bounding box to 125% its size. The GMMs are trained on the fly according to intensities in the given patient: for Obj. it is taken from the foreground seeds, and for Bkg., from rectangles placed on the neighbouring discs. In more detail, for the vertebra v_i with bounding box $B(v_i)$, let its neighbouring vertebrae be v_{i-1} and v_{i+1} with bounding boxes $B(v_{i-1})$ and $B(v_{i+1})$. Then the two disc-rectangles are placed between $B(v_{i-1})$ and $B(v_i)$, and $B(v_i)$ and $B(v_{i+1})$ respectively, at the arithmetic mean position of the corners of the adjacent bounding box edges. Each seed is a quarter the height of $B(v_i)$. The seeding process is illustrated in Figure 5.3.

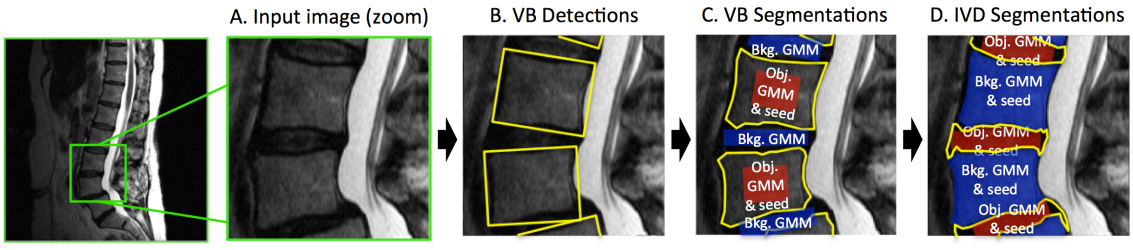


Figure 5.3: **Segmentation seeding process.** The segmentation process is initialised by vertebrae detections and consists of sequential vertebrae segmentation, and intervertebral disc segmentation. The yellow lines show the detection and segmentation outputs, and the Obj. (object) and Bkg. (background) seeds show the initialisations used for the segmentations as the red and blue areas.

5.3.3.2 Disc Segmentation

After vertebrae segmentation explained above, the discs are segmented. The disc segmentations are automatically initialised using seeds placed according to the automated vertebrae segmentations. Similarly to vertebrae, the segmentations are performed independently for each disc in each slice. For each disc, the process is automatically initialised by seeding the foreground as the space between, and background as the area of the neighbouring vertebra segmentations. The disc segmentations are performed for discs T12/L1 to L5/S1. The region terms are modelled as three-component Gaussian Mixture Models (GMMs) according to the image intensities in the Obj. and Bkg. seeds. Three components were picked as best performing at earlier experiments. The seeding process is illustrated in Figure 5.3

5.4 Evaluation Protocol & Performance

In this section, the ground truth annotation, the evaluation measures, and the results are presented and discussed.

Ground truth annotation. In total, 592 vertebrae, 1092 disc sagittal cross-sections were manually delineated, as illustrated with examples in Figures 5.4 and 5.5 respectively. The vertebrae delineations cover 300 unique patients (347 unique vertebrae) and the disc delineations 85 unique patients (87 unique discs). In addition, herniated mass is delineated in 902 cross-sections, covering 102 unique patients (105 unique discs). These numbers, along with segmentation results, are summarised in Table 5.1. All the annotated vertebrae were at levels 4-6 (L3,L4,L5), as those are the ones near the discs that are most commonly diseased. In total, 21 sections at level 4, 436 at level 5, and 135 at level 6. An L5 section was segmented in every patient.

The ground truth annotation was performed by an imaging expert with repeated experience with a 2000-patient imaging database. To keep consistency, a rigid set of

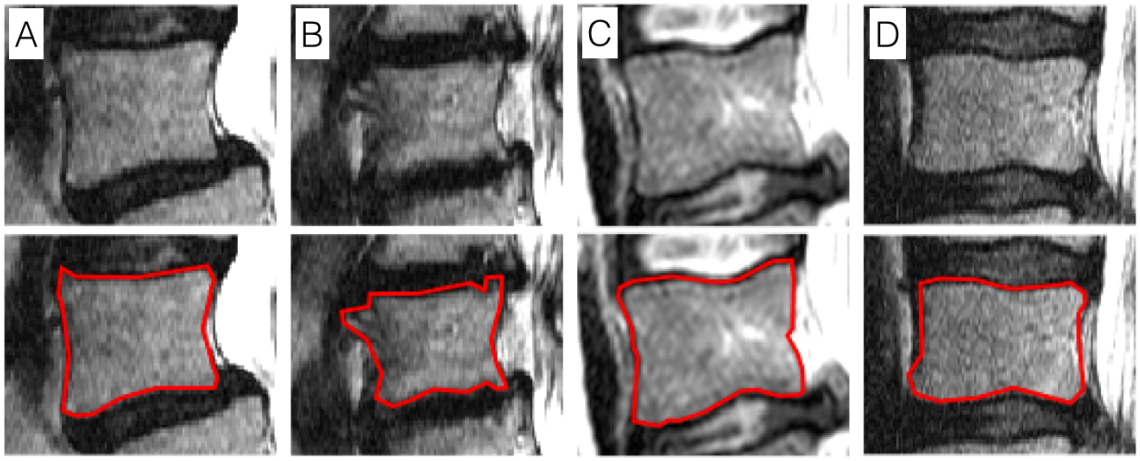


Figure 5.4: **Vertebrae segmentation Ground Truth.**

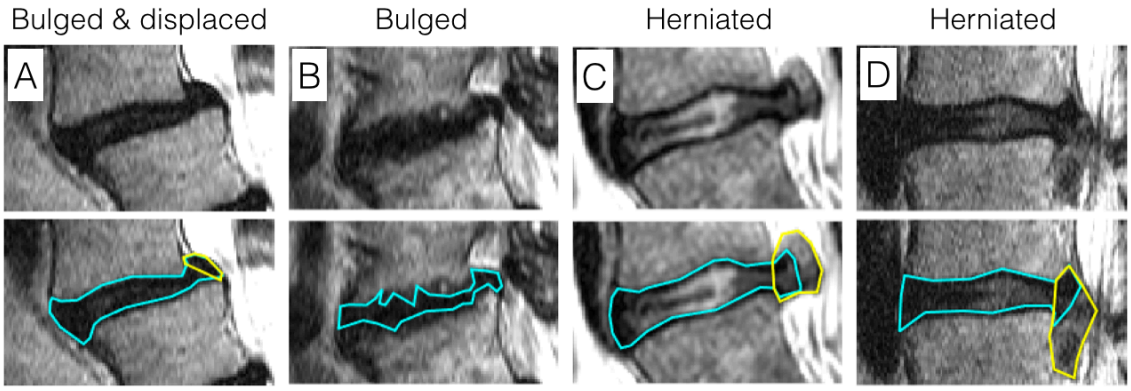


Figure 5.5: **Disc segmentation Ground Truth.** The cyan lines mark the disc border enclosed by annulus; the yellow borders the disc mass displaced across the vertebrae edges.

rules was followed. The ground truths were marked and stored as closed polygons in a tool developed in MATLAB, in a slice-by-slice basis (cross-sections annotated). The number of points used depended on the complexity of the shape.

For **vertebrae**, the following procedure was followed. Delineate the vertebral body. The epidural fat at the back of the VB, and the dark image artefact line along with ligaments surrounding the VB were left out if it was sharp, or half-in, if it was a smooth gradient. Partial volumes, e.g. in case of Schmorl's node, on VB-IVD border, were usually left out. The delineation was performed by marking the corners first, and then points between the corners as necessary. The number of

points used to define the polygon varied from 4 to 55, with mean at 21 and median at 21.

For **discs**, the following procedure was followed. Delineate the part of the disc surrounded by the annulus fibrosus outer cell. Herniated disc mass that had left the annulus was marked by a separate label, as illustrated by the yellow line in Figure 5.5. First, mark the most anterior and most posterior points. Then, mark points between each pairs of corners, to get closer to the boundary. The number of points used to define the polygon varied from 4 to 52, with mean at 20 and median at 20.

Based on segmenting a randomly selected subset of twelve vertebrae and six discs cross-sections twice, a couple of months apart, the mean intra-observer variability for vertebrae in slices excluding pedicles was found to be 0.961 (median 0.964, range 0.94-0.98), and 0.884 (median 0.890, range 0.834-0.927) for discs. The greater intra-observer variability in the disc segmentations likely lends itself to the greater uncertainty around the anterior and posterior sides of the disc (invisible boundaries between the annulus and surrounding ligaments, muscles, and vessels, thus the anterior boundary is defined imprecisely by a mental shape prior and context of other slices), and the fact that by area, the VBs are over two times bigger than the IVDs (found as the ratio of the median areas of the samples), and thus the Jaccard similarity measuring overlap for discs is around doubly more significant to errors of the same area at the shared IVD-VB boundary than for vertebrae.

Performance measures. The following four quantitative performance measures were used to compare the automatic segmentation results (A) to the manual ones (M): Jaccard overlap (O), DICE Similarity Coefficient (DSC), the Hausdorff distance, and the mean absolute shape distance (MASD) in millimetres. The Jaccard overlap measure is the intersection over union, defined as the number of correctly labelled pixels of a class, divided by the number of pixels labelled with that class in either the ground truth labelling or the automatic labelling. Equivalently, the

accuracy is given by the equation

$$O(A, M) = |A \cap M|/|A \cup M| \tag{5.3}$$

whereas the DSC is defined as

$$DSC(A, M) = 2|A \cap M|/(|A| + |M|) \tag{5.4}$$

To obtain the MASD and Hausdorff distance, for each point i in A the minimum distance d_i^{AM} to all points in M is determined. The mean directed boundary distance ($ASD_{AM} = average(d_i^{AM})$) and the directed Hausdorff distance ($H_{AM} = max(d_i^{AM})$) are calculated. The same same calculation is performed for each point in M to all points in A, with the mean average boundary distance $MASD = (ASD_{AM} + ASD_{MA})/2$ and Hausdorff distance $HD = \max(H_{AM}, H_{MA})$ then calculated.

Both the Jaccard overlap measure, and the DSC increase with improving segmentation, taking the maximal value of 1 for perfect segmentation, and 0 for complete failure for each vertebra and disc. The MASD and the Hausdorff distance decrease with improving segmentation, taking the value 0 for perfect segmentation.

The motivation for using MASD and the Hausdorff distance in addition to the overlap measures is that they allow for further comparison with works in literature, and correspond more directly to the requirements for accurate localisation of the support regions in Chapter 6. In addition, they are less sensitive to the object size, making the vertebrae and disc segmentation results more comparable to each other (as the median vertebra is around 2 times larger in area than the disc, however only 1.1 times longer perimeter length). E.g. the same error in the VB-IVD shared boundary will show up as the same Hausdorff distance, and similar mean distance between boundaries.

Results. Some example segmentations for vertebrae are shown in Figure 5.6; and for discs in Figure 5.7. The statistics of the overlap measure, between the ground truth and automatic segmentation, over the cross-sections, are cited for all slices, for mid-range slices, and for side-range slices. The mid-range slices are defined as at maximum 2 slices away from the mid-disc slice (with the middle slice defined according to the pedicle locations), and the side-range slices as the rest of the slices. We give two statistical measures for both vertebrae and discs overlap in every case – the median, and the mean. There were 511 annotated vertebrae cross-sections in the mid-range zone (81 out of the zone), and 486 annotated disc cross-sections in the zone (606 out of the zone).

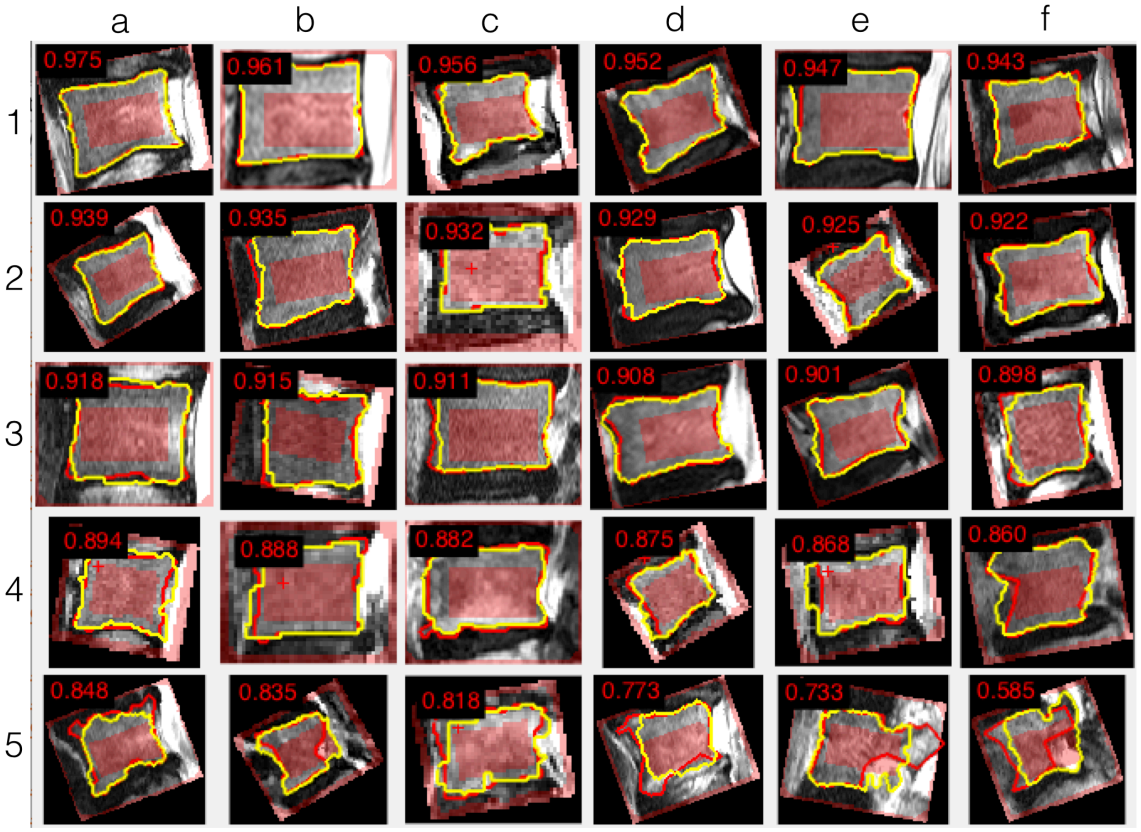


Figure 5.6: **Vertebrae segmentation example results.** The yellow lines mark the automatic, and the red lines the manual "ground truth" results.

The statistical results are shown in Table 5.1, and plotted, ordered by increasing

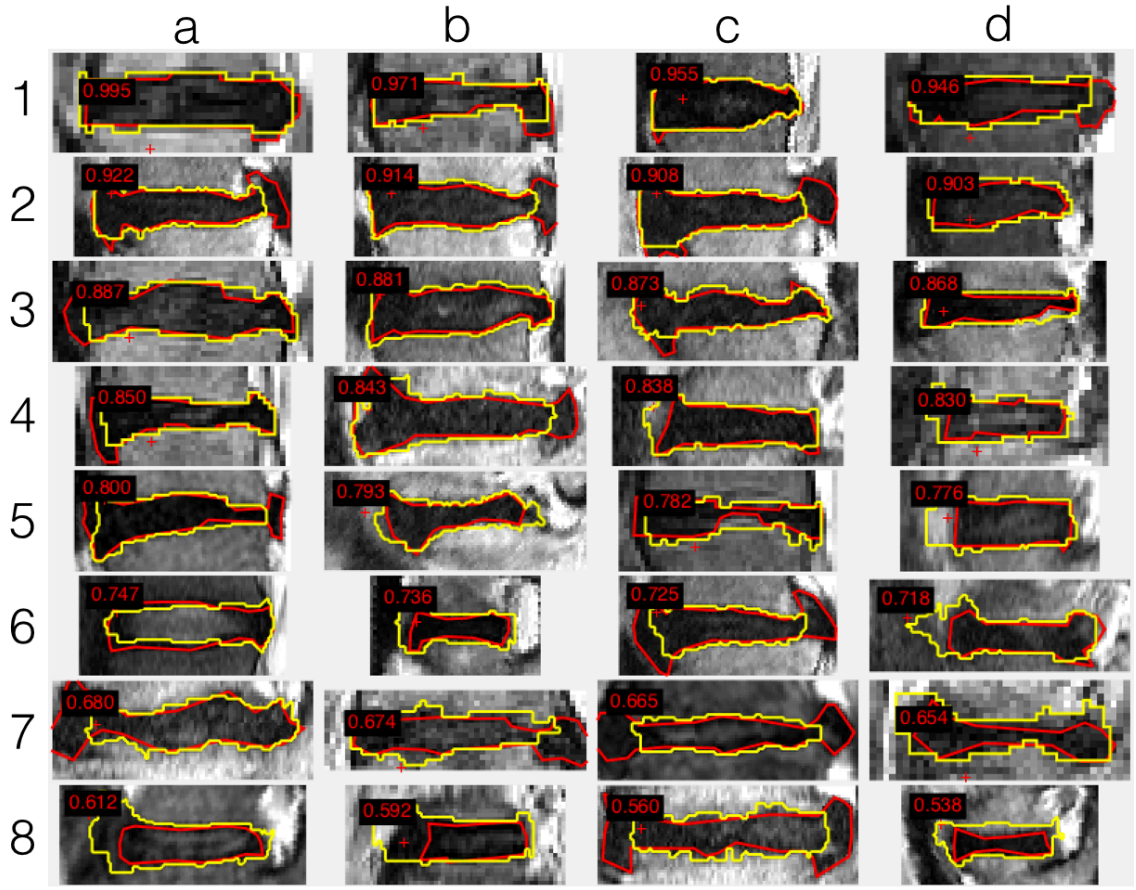


Figure 5.7: **Disc segmentation example results.** The red lines are ground truth; the yellow lines are graph cut segmentation results. The examples are ranked by decreasing overlap measure, which is indicated as the number on the top left corner of each subfigure.

overlap measure, in Figure 5.8. Across all slices, the mean Jaccard measure for vertebrae was 0.839 (median 0.905), and for discs 0.717 (median 0.773). Across the mid-disc zone slices, the mean for vertebrae was 0.8666 (median 0.9109), and for discs 0.7808 (median 0.8323). For out-of-zone slices, the mean for vertebrae was 0.6678 (median 0.8233), and for discs 0.667 (median 0.7246).

The failures are possibly caused in some cases by segmentation seeds misplacement. The foreground seeds cross the ground truth VB boundary in 277 cases out of 592, and in 49 cases the leaked seed area is 5% or more, in 22 cases 10% or more, and in 14 cases 20% or more of the area of the ground truth VB. These failures will

Anatomy	Cross-sections	Unique V/D	Unique patients	Overlap	DSC
Vertebra	$592/10^6$	347/1800	300/300	0.839	0.892
Disc	$1092/10^6$	128/1800	116/300	0.784	0.866
Hern.	$989/10^6$	120/1800	103/300	*	*
Axial	207/5000	*/900	43/300	*	*

Table 5.1: **Ground truth and segmentation results.** The number of vertebrae, discs and herniated disc masses in sagittal slices are listed. Segmentation results along with the overlap measures on the set of patients with segmented vertebrae (V) and disc (D) sagittal cross-sections.

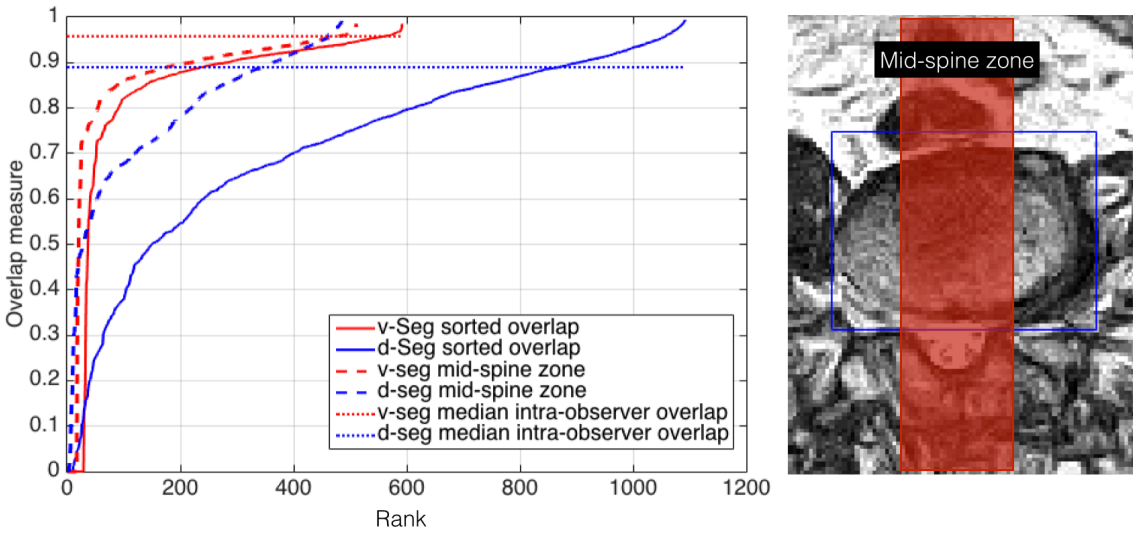


Figure 5.8: **Segmentation results sorted by overlap** for vertebrae (v-seg) and discs (d-seg) for sagittal cross-sections overall, and in the mid-spine-zone.

cause error by at least the leaked seed area (b5, e5, f5 showing a large seed leak in Figure 5.6), or possibly the effect could be greater at times (c2, d2, d3 posterior area in Figure 5.6). Other seed leaks in Figure 5.6 are e1, f1.

The large drop in the mean, but smaller in the median from mid-disc zone to out of the zone shows that most of the failures occurred in the out-of-zone slices. The drop is more dramatic in discs, since there are more disc ground truth segmentations available out of the zone, than for vertebrae. Segmentation is more challenging out of the zone due to increasing number of partial volume voxels, and invisible boundaries, and the presence of pedicles (so the VB cross-sections no longer appear rectangular)

and similarly more partial volumes for vertebrae. In addition, the area of both the disc and vertebrae cross-sections is smaller out of the zone, making the segmentation more sensitive to precise initialisation, and also the overlap measure more sensitive to the same absolute area error in segmentation.

The increased failures in segmentations in discs versus vertebrae for mid-zone cross-sections come from the intensity similarity and invisible boundary between the disc anterior side, and the anterior blood vessels, ligaments, and muscles.

No significant difference was observed in segmentation quality between the disc levels. However categorising by the ground truth radiological measurements (Pfirrmann grade, narrowing, Modic changes,), differences in the segmentation quality were observed. On severely narrowed discs, the performance according to Jaccard measure in the mid-spine zone was significantly worse for both vertebrae (0.813 vs. 0.883) and discs (0.668 vs. 0.819). This is because the disc disappears in those cases, making its ground truth very small area and thus hard to initialise for discs, and cutting out the dark buffer area between adjacent vertebrae, encouraging leakage from one vertebra to the next. On vertebrae with Modic changes, the vertebrae segmentation was slightly improved on Modic type 1 and 3 (brightened VB: Jaccard 0.849 vs. 0.839) but decreased on Modic type 2 (darkened VB: Jaccard 0.805 vs. 0.853).

Although our results are not directly comparable to works in literature due to difference in datasets, and we do not propose our method as direct competition to alternatives (as the context of segmentation in the thesis was to investigate its importance as input for further processing rather than target segmentation perfection), the numbers are similar to what has been cited in the past as evident in the comparison Table 5.2. It has to be noted that first, comparison to other similar works is hard, as the ground truth is ambiguously defined, and the datasets in literature are usually of higher resolution, might not contain pathologies, and can be of only

Authors	**Anatomy**	**Modality**	**Measure**	**Mean values**
Davatzikos et al. [2002]	VB-3D	MR	DSC	0.815
Štern et al. [2011]	VB-3D	MR	Landmark dist.	1.85mm
Hoad and Martel [2002]	V-3D	MR		1.11mm
Klinder et al. [2009]	V-3D	CT	Point-surf. dist.	1.12mm
Ma et al. [2010]	V-3D	CT	Point-surf. dist.	0.95mm
Chevrefils et al. [2009]	IVD-2D	MR	DSC	0.85
Michopoulou et al. [2009]	IVD-2D	MR	DSC	0.92
Michopoulou et al. [2009]	IVD-2D	MR	MASD	0.61mm
Seifert et al. [2009]	IVD-3D	MR	DSC	0.84-0.98
Seifert et al. [2009]	IVD-3D	MR	MASD	1.86-2.5mm
Seifert et al. [2009]	IVD-3D	MR	Hausdorff dist.	3.12-6.62mm
Neubert et al. [2012]	IVD-3D	MR	DSC	0.89
Neubert et al. [2012]	IVD-3D	MR	MASD	0.55mm
Neubert et al. [2012]	IVD-3D	MR	Hausdorff dist.	3.55mm
Neubert et al. [2012]	VB-3D	MR	DSC	0.91
Neubert et al. [2012]	VB-3D	MR	MASD	0.67mm
Neubert et al. [2012]	VB-3D	MR	Hausdorff dist.	4.08mm
Ayed et al. [2011]	IVD-2D	MR	DSC	0.88
This work	IVD-2D	MR	Jaccard overlap	0.784
This work	IVD-2D	MR	DSC	0.866
This work	IVD-2D	MR	Haudsdorff dist.	7.43mm
This work	IVD-2D	MR	MASD	1.76mm
This work	VB-2D	MR	Jaccard overlap	0.867
This work	VB-2D	MR	Haudsdorff dist.	4.6mm
This work	VB-2D	MR	MASD	0.77mm

Table 5.2: **Comparison to literature.** (V-whole vertebrae, surf.-surface, dist. - distance.). Adapted from Neubert et al. [2012]

around ten patients.

Our DSC measure of 0.892 for VB-2D compares well to values presented in the literature: 0.815 for VB-3D by Davatzikos et al. [2002]; 0.89 for VB-3D by Neubert et al. [2012], while the Hausdorff distance is slightly greater (worse) – 4.6 mm for our VB-2D segmentations, versus 4.08mm for VB-3D by Neubert et al. [2012]. We think this could be due to the fact that they assess their segmentations in 3D, with cubic voxel (higher resolution). At the same time, the MASD of 0.77mm for our

VB-2D segmentations is reasonably close to results in other works (Štern et al. [2011] 1.85mm landmark distance for VB-3D, Klinder et al. [2009] 1.12mm, Ma et al. [2010] 0.95mm

Our DSC measure of 0.866 for 2D-IVD segmentation is similar to a slice-based segmentation approach of Chevrefils et al. [2009] (0.85), however slightly lower than that of Michopoulou et al. [2009] who initialise their segmentations with two manually placed points per disc. Our result is also similar to that of Seifert et al. [2009] – DSC of 0.84-0.98 – who take a very similar strategy to us, performing segmentations slice-by-slice in 2D, and combining them into a 3D volume later, however in cervical spine. They achieve a MASD of 1.86-2.5mm, which is higher (worse) than our 1.76mm.

5.5 Discussion

Automated vertebrae and disc segmentation were performed on a large dataset using a powerful graph-cuts algorithm, and evaluated on a large manually annotated subset of the dataset. The segmentation quality was examined across the slice range, against spine level, and its dependence on a number of spine conditions investigated.

As a function of slice number, both the manual segmentation variability and the automated segmentation quality degrades with departure from the mid-range of the disc. This is due to the increased number of partial volumes due to a geometrical effect, and more invisible boundaries due to the signal similarity of the tissue of the disc annulus to the surrounding tissues.

The selection of the segmentation method is heavily motivated by the dataset (particularly, variably sparse slice spacing, and variable left-to-right FOV). To keep it applicable to a wide variety of scanning protocols and image qualities, the method works slice-by-slice in 2D, allowing for 3D reconstruction of the volumes of the spine parts afterwards.

The segmentation quality could be improved by learning improved features for segmentation from the current slice (e.g. SIFT, etc.), and incorporating 3D context from other slices (e.g. to distinguish a black-white tissue partial volume from a solid gray voxel), and features from other series (e.g. geometrically different – axial, coronal – and different in terms of the radio-frequency spin excitation sequences – e.g. T1). The extra challenge this introduces is patient motion between the scans, which is challenging particularly if it occurs in the direction perpendicular to the wide slice spacing.

All in all, the segmentation of spinal MRI-s remains an inherently challenging problem due to wide variations in anatomy, pathology, and image quality (resolution, particularly slice spacing, artefacts, etc), and inter-observer variability in ground truth definition.

Chapter 6

Learning and Radiological Measurements

This chapter focuses on the application of machine learning techniques to automatically predict clinically relevant radiological spine measurements. We present and validate a framework for learning, which is the last part in our image analysis pipeline. It takes as input anatomy localisations from previous chapters (detected, labelled and segmented vertebrae / discs) and outputs disease state predictions. The chapter has three parts. **First**, the learning framework is overviewed in Section 6.1 discussing radiological measurements, support regions, features, and machine learning methods. **Second**, the utility of the framework is demonstrated in automatically predicting three different measurements from spinal MRI scans, given the detections and segmentations from previous chapters, and using common, universal patch features across all measurements. The particular measurements we predict are: Pfirrmann grade and narrowing in Section 6.2; disc herniation/bulge in Section 6.3. The performance of the framework is examined on the 300-patient dataset presented in Chapter 3. **Third**, the importance and influence of segmentation in the framework is investigated on the measurement of Pfirrmann grade in Section 6.5.

Three different, more hand-crafted features are used, that can be extracted from regions of any shape, rather than just rectangles, and are thus more suitable for the study.

6.1 Learning Framework Overview

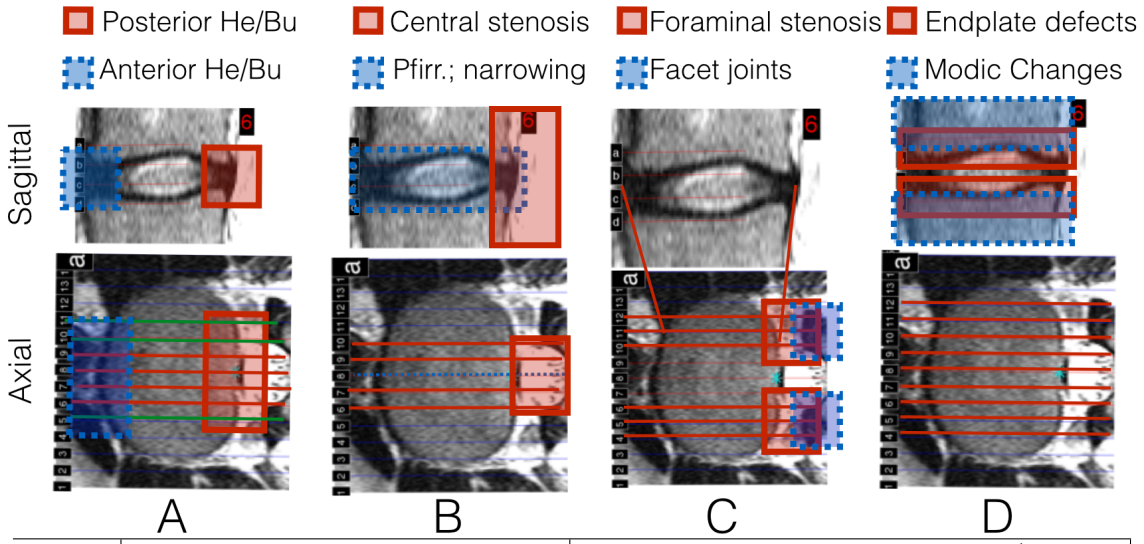
The learning framework has two steps: **first**, features and support regions explained in Section 6.1.1; **second**, learning and mapping to measurements explained in Section 6.1.2.

6.1.1 Support Regions and Features

The aim of this step is to automatically generate support regions for feature computations. These features are then used in the learning algorithms of the following subsection. The support regions target relevant anatomical locations, to provide radiological measurements and gradings. Figure 6.1 illustrates possible region placements for a number of different disc conditions.

The regions are placed independently for each disc in turn. In typical clinical data, the slice spacing is much sparser (often up to 10 times) than the in-slice pixel spacing. Therefore, the support region localization problem naturally decomposes into two sub-problems: **(1)** slice selection, and **(2)** region localization within the selected slice(s). These sub-problems are explained in the following sections.

The slice range is selected automatically for each disc independently according to the vertebrae detections. First, the mid-disc slice is determined. For a given disc, this is automatically found as the median of all slices, where both the superior and inferior vertebra of the disc were detected. See Figure 6.2 for an example – here, it is slice 7, e.g. the median of slices 5-9. In these slices, the VB appears as a rectangle. In slices further off to right or to left, the VB cross-section is no longer rectangular, as the pedicles start to appear (top of pedicles are marked as points P_L and P_R in



Col.	Measure	Feature support size in pixels/slices					
		w_{sag}	h_{sag}	n_{Sag}	w_{ax}	h_{ax}	n_{Ax}
A	aBu/**pBu/cHe**	31	31	5-9	61	31	3-6
B	**Pfirr./Narrow.**	61	21	1-3	–	–	–
B	Central stenosis	41	61	1-5	41	41	3-6
C	Foraminal stenosis/herniation	41	41	1-5	41	41	1-3
C	Facet joints	41	41	1-2	41	41	1-2
D	Endplate defects	61	21	5-12	–	–	–
D	Modic changes	61	31	5-12	–	–	–

Figure 6.1: **Support regions** are shown for a number of radiological measurements in sagittal and axial slices. Each of the columns A-D shows two sets of suggested support regions marked by the semi-transparent rectangles – e.g. posterior/anterior herniation/bulge (he/bu) under A, or central stenosis, Pfirrmann grade (Pf.), and narrowing under B. The lines in sagittal slices mark intersections with axial slices. The lines in axial slice mark intersections with sagittal slices. In each case, the thick lines indicate the slices included in the support region. The sagittal slice spacing is typically 4-5mm. The axial slice spacing within a slice group is similar, but between slice groups can be up to 3cm. In the **table**, each row is for a measurement as marked by A-D also used above. Under the column 'feature support size in pixels/slices', w denotes the width, h the height & n the number of slices for the respective support region. The w, h, n are given for both sagittal (subscript 'sag') and axial (subscript 'ax') support regions.

Figure 6.2).

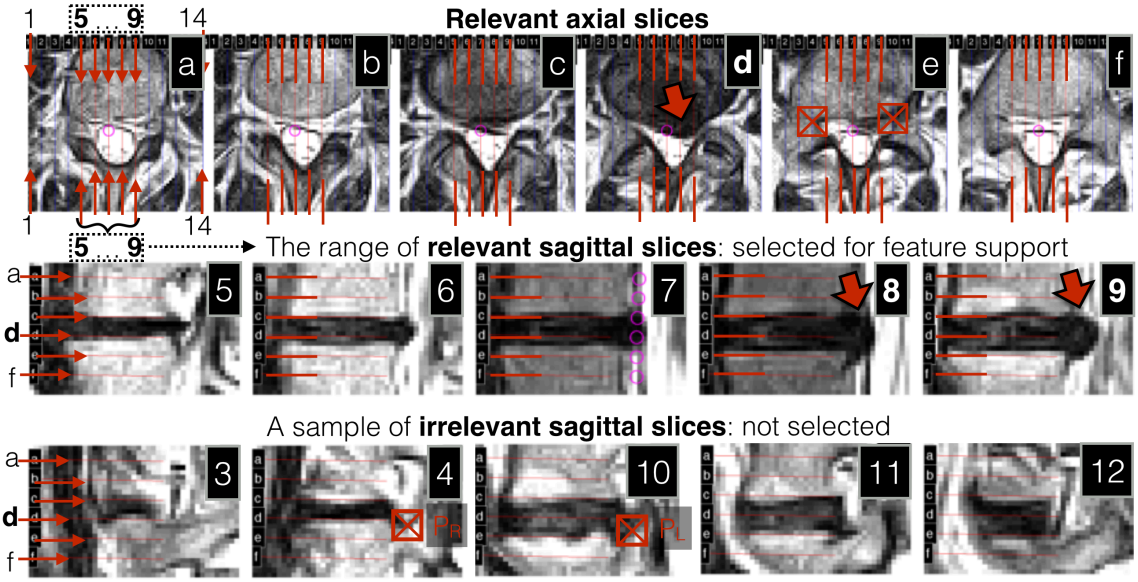


Figure 6.2: **Slice selection**, illustrated on an example L3-L4 disc, for the task of central herniation measurement, in typically available clinical T2 slices (sagittal and axial). For the given patient, there were 14 sagittal, and 16 lumbar axial slices available in total, however only a subset of all slices are relevant for herniation measurement. Notice that the given disc happens to be herniated, with the displaced disc material (pointed at with large red arrows) best visible in axial slice d, and sagittal slices 8-9. Note also that the range of relevant slices – both axial and sagittal – depends on the type of radiological measurement.

Example Support Regions

Example (hypothetical) support regions for a number of measurements are given in Figure 6.1 for illustration. For Pfirrmann Grade, Narrowing, and herniation/bulge, the regions are discussed in more detail in the following sections. For **anterior bulge**, the region is the same box around the anterior region of the disc, in sagittal slices only, as for herniation/bulge. For **central stenosis**, the region is a box around the spinal cord. For **Schmorl's nodes & Modic changes**, the box is around the vertebra area near the disc, covering all sagittal slices. For **foraminal stenosis**, the region is a box around the foramina – that is, left-right sagittal slices, in the posterior region of the disc. For **Schmorl's nodes (endplate defects) & Modic changes**, the box is around the vertebra area near the disc, covering all sagittal

slices. For **facet joint problems**, the region is in the left and right sagittal slices, and in the axial slices around the facet joints.

Note that the table shows hypothetical potential region placements. The optimal sizes for the support regions used in the experiments are determined by grid-search on the validation set.

Image Normalisation

As mentioned in the introduction, one of the challenges is the variation of MRI contrast across different protocols. Therefore, before extraction, the image is normalised to 1.25 times the median vertebral intensity as found from the segmented vertebrae. In early experiments, we tested a number of intensity normalisation techniques, including normalising to CSF. We found the median vertebral intensity normalisation approach performed the best and provided sufficient robustness to variations in protocol and MRI scanners. The conventional CSF-based normalisation method (Battie et al. [1995]) also suffers from inconsistent CSF signal, due to various deformations in the dural sac. After normalisation, the new median vertebral intensity becomes 0.8; intensities above one are truncated to one. This way, there is still dynamic range kept above the vertebrae intensity (e.g. for grade 1 or 2 discs), unlike in the case if the vertebra intensity was normalised to one. The disc histogram ranges between the CSF/water level, and zero. By normalising to 1.25, the relevant dynamic range of the disc histogram is scaled to be between 0 and 1. This normalisation technique proved to work across various protocols as illustrated in Figure 6.3.

Features

Intensity and shape features are extracted from the feature support region, as described in the following sections. In the following experiments, raw patch features are used for all measurements, along with disc height and level for some of the measurements.

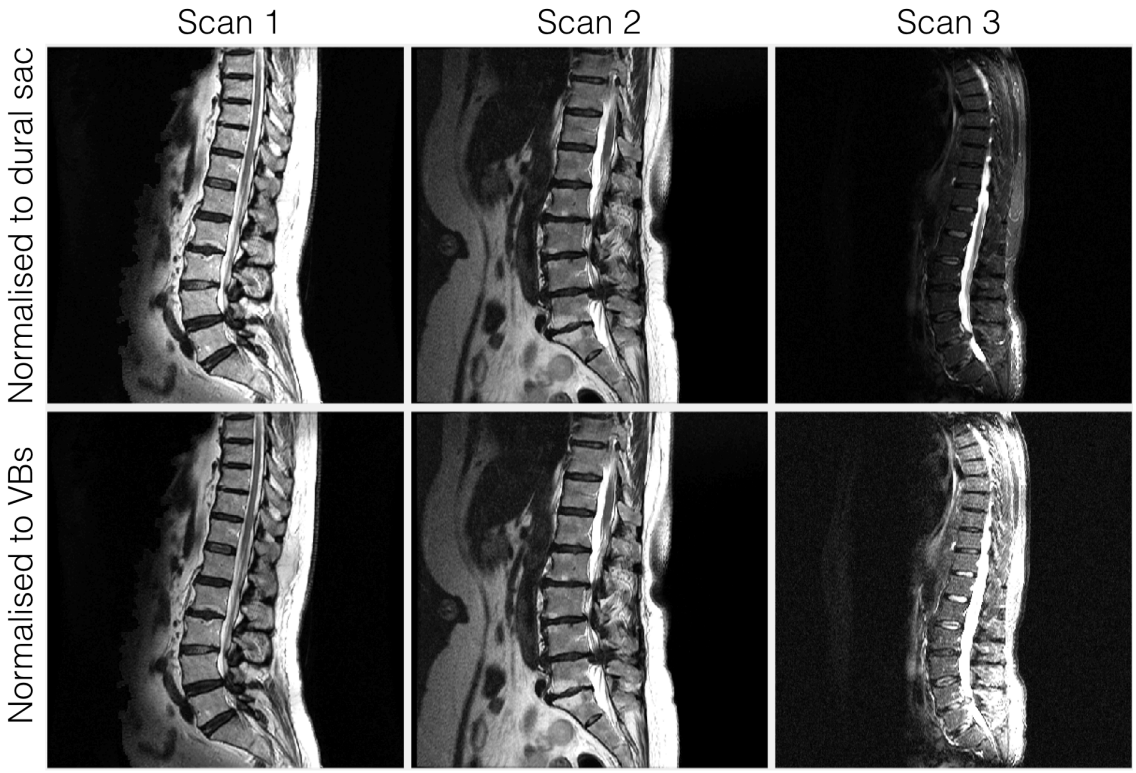


Figure 6.3: **Image normalisation.** in three different scans: (Scan 1) General Electric Genesis Signa 1.5T Spin Echo, (Scan 2) General Electric Signa HDxt 1.5T Spin Echo, (Scan 3) Siemens Symphony 1.5T Inversion Recovery / Spin Echo using two different methods. Notice how the vertebral body (VB) normalisation results in more even contrast, whereas the conventional dural sac normalisation results in too dark vertebrae for scan 3, and slightly too bright for scan 1.

6.1.2 Learning: Mapping to Measurements

In this step, the feature vector extracted from the support region(s) is mapped to radiological measurements, performing regional analysis with linear regressors and classifiers. In the following experiments, support vector machines, and support vector regressors were used as robust learners.

A regressor is learnt for Pfirrmann grade and narrowing, and a classifier for herniation/bulge, in each case from the feature vector and annotated measurements. The regressor is an epsilon Support Vector Regressor implemented using the LIBSVM (Chang and Lin [2001]), and the classifier a C-SVM using the LIBLINEAR

(Fan et al. [2008]) package.

We choose to define the Pfirrmann grading and narrowing prediction tasks as regression problems as they correspond to continuous processes, whereas herniation is a more abrupt event. We use the ϵ-SVR for the regression, as it is able to deal well with label noise and its intuitive interpretation, as the distance from the decision plane. It is an attractive way to relate the disc degeneration process with a linear scale, which is what the radiologists are looking for.

The **ϵ-SVR** maps the feature vector $\vec{x}$ to a continuous output variable f as:

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{6.1}$$

where the vector $\vec{w}$ and bias b are learnt on the training set. The fitting process is influenced by two unitless parameters: $C > 0$, and $\epsilon > 0$. The parameter C determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ (margin width around the decision plane) are tolerated. This corresponds to dealing with a so called ϵ-insensitive loss function:

$$|\xi|_\epsilon := \begin{cases} 0, & \leq \epsilon. \\ |\xi| - \epsilon, & \text{otherwise.} \end{cases} \tag{6.2}$$

where ϵ is the maximum penalty-free deviation of $f(\vec{x})$ from the actually obtained targets y_i.

The **linear C-SVM** (Smola and Schölkopf [2004]) is learned to map the feature vector $\vec{x}$ into classifier score f as:

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{6.3}$$

where the vector $\vec{w}$ and bias b are learnt on the training set. The fitting process is influenced by a unitless parameter $C > 0$, which determines the margin width of

the hyperplane (small C for large margin; large C for small margin).

Feature set augmentation

The feature set can be augmented either at training or testing, or both times, to build some robustness to region localization errors into the system. In data augmentation, extra training or testing samples are generated by applying transformations (such as scaling, rotation, translation). We augment the training set, adding displacements of ± 2 voxels for x and y, and of up to ± 5 degrees in in-slice angle to the training set samples. Augmentation during testing time could be performed by perturbing the position of the detected region by small amounts in various directions, applying the predictor on each version of the region, and taking e.g. the mean, the median, or the max value as the prediction result.

6.1.3 Dataset Details

The dataset used in the experiments is the one described in Chapter 3.

Train/Test splits. The patients were split into a 114-scan training (684 discs), 57-scan validation (342 discs), and 129-scan (774 discs) testing set.

Additionally, within each set, a 'picked' subset is introduced for Herniation/Bulge (He/Bu), as illustrated in Figure 6.4. This consists of hand-picked examples of clearly bulged/herniated discs as positives and clearly non-bulged, non-herniated discs as negatives. This is done to remove ambiguous or borderline cases affecting the classifier training. In addition, extra care is taken to remove any ambiguous ground truth in the 'picked' set. The 'picked' training, validation, and testing subsets contain 117, 61, and 71 discs, respectively. They are used to sandbox-experiment on how well the system works if the samples are perfect, and to study the effect of measurement accuracy purely as a function of displacement.

For the further experiments with Pfirrmann grade, described in Section 6.5, the dataset was split into a 114-scan training (684 discs), 57-scan validation (342 discs),

and 114-scan (684 discs) testing set.

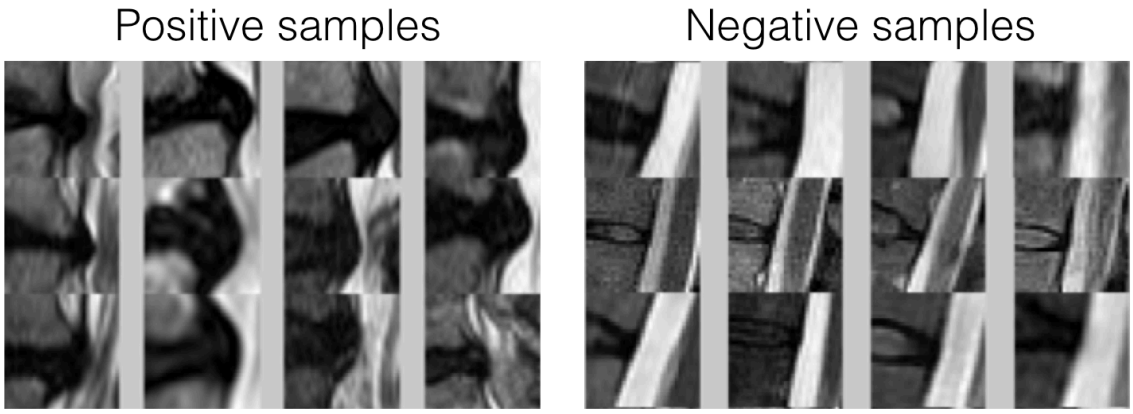


Figure 6.4: **Herniation/Bulge.** Samples of the 'picked' subset of disc posterior sides.

6.2 Pfirrmann Grade and Narrowing

See Figure 6.5 for illustration of the Pfirrmann grading system, including narrowing. A healthy Pfirrmann grade disc has a well hydrated nucleus, appearing bright white in a T2 scan. The height of lumbar discs increases slightly from disc to disc across the normal lumbar spine.

The Pfirrmann grade and narrowing measures have been designed to capture departures from those trends. Radiologically and spatially, they fit together as (i) they both relate to the disc globally, and (ii) narrowing is itself a feature that is used for Pfirrmann grade prediction radiologically.

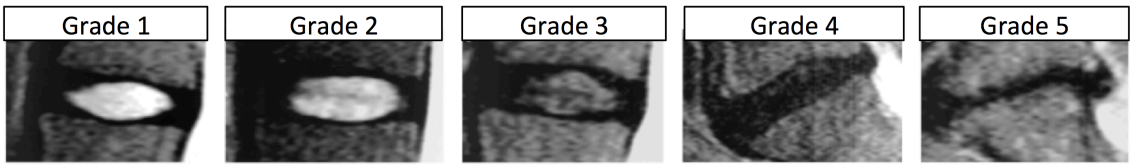


Figure 6.5: **Pfirrmann global disc grade and narrowing**. Disc degeneration process. Note that in the case of these particular disc examples, the degeneration process involves T2 signal loss, both posterior and anterior **bulging** (e.g. Grade 5 example), and disc space **narrowing** (e.g. Grade 5 example), which are also measured separately.

6.2.1 Measurement, Dataset and Ground Truth Details

Experiments were performed on the full 300-patient dataset described in Chapter 3. The discs were labelled with the following relevant labels:

- **Pfirrmann grade**: There are 550 grade 1, 194 grade 2, 379 grade 3, 379 grade 4, and 208 grade 5 discs in the dataset.
- **Disc space narrowing**: There are 1009 grade 0 (normal), 219 grade 1 (slightly narrowed), 348 grade 2 (moderately narrowed), and 223 grade 3 (collapsed/severely narrowed) discs in the dataset.

6.2.2 Support Regions and Features

Support regions.

For both **Pfirrmann grade & narrowing**, the region is placed in only the mid-disc sagittal slice, and is a 21-by-61 voxel rectangle covering the disc. The optimal size was found by grid search on the validation set, with height 11 to 31 voxels in 5-voxel steps, and width 41 to 71 voxels with 5-voxel steps.

The sensitivity of the results for the grid search was reasonably low. This indicates that the results were robust to the parameters.

The region is found as follows. First, a ROI is selected around the disc area, according to the bounding boxes of its neighbour vertebrae. The disc angle is determined as the mean angle of the vertebrae, and the image is rotated by that angle. The image is resized according to the anterior-posterior width w_{AP} of the disc, making the vertebral anterior-posterior width 60 pixels. Next, the rectangular-shaped support regions are placed, in each slice independently.

Features. Voxel intensity, the mean disc height, and the disc level i are extracted from the support regions as features. Each pixel's intensity in the support region makes an entry into the feature vector. The disc level, and the mean disc height

add two more features. Thus, the sagittal feature vector for Pfirrmann/narrowing has $21 \times 51 + 2 = 1073$ elements.

6.2.3 Mapping to Measurements

A linear epsilon Support Vector Regressor (Smola and Schölkopf [2004]) is learned to map the feature vector to the grading, as explained in Section 6.1.2.

Implementation details. The values of ϵ and C for the Support Vector Regression cost function are learnt by a grid search on a 57-patient hold-out set, with a range of (on exponential scale) 0.01 to 1.0, and 0.01 to 10000 respectively, using the LIBSVM package (Chang and Lin [2011]). The best values were 0.5 for ϵ, and 1 for C.

For **Pfirrmann grade**, the final result is the regressor output, rounded to the nearest integer from one to five; for **narrowing**, a binary integer (1 for narrowed, 0 for normal).

6.2.4 Evaluation Protocol

Pfirrmann Grading. We assess accuracy by measuring the proportion of discs that are graded correctly, as proposed by Xu et al. [2013] for eye cataract grading. However, our ground-truth labelling is not perfect, due to intra-observer variability, and thus we evaluate our regression performance as the fraction of scores which are predicted to ± 1 of the radiologist grade.

Narrowing. Narrowing is assessed just as herniation/bulge (explained in the next section).

6.3 Herniation/Bulge

In this Section, we turn to a different measure, characterising deformations of the posterior region of the disc. The bulge and herniation are shown for a number of example cases in Figure 6.6. There is some ambiguity and disagreement in the definition of herniation, and its distinction from bulge in the spine research and

clinical community. Bulge is usually clinically considered a 'normal' phenomenon, whereas herniation 'abnormal'. Fardon et al. [2014] distinguish herniation from bulge solely based on the the morphological shape of the back side of the disc only, saying whereas other people have defined it as the rupture of the annulus of the disc with the soft annulus flowing out. In some other works of automated herniation prediction, it is not clear whether bulge was considered as herniation or not (Alomari et al. [2010a, 2011b, 2013], Ghosh et al. [2011b,a], Koh et al. [2010, 2012]).

Due to this ambiguity, and to have a more balanced classes in the dataset, we group bulge and all types of herniation together as one class to start with, and revisit the finer herniation vs. bulging problem in Chapter 7, taking into consideration axial slices. In sagittal slices, the bulge and herniation look very similar.

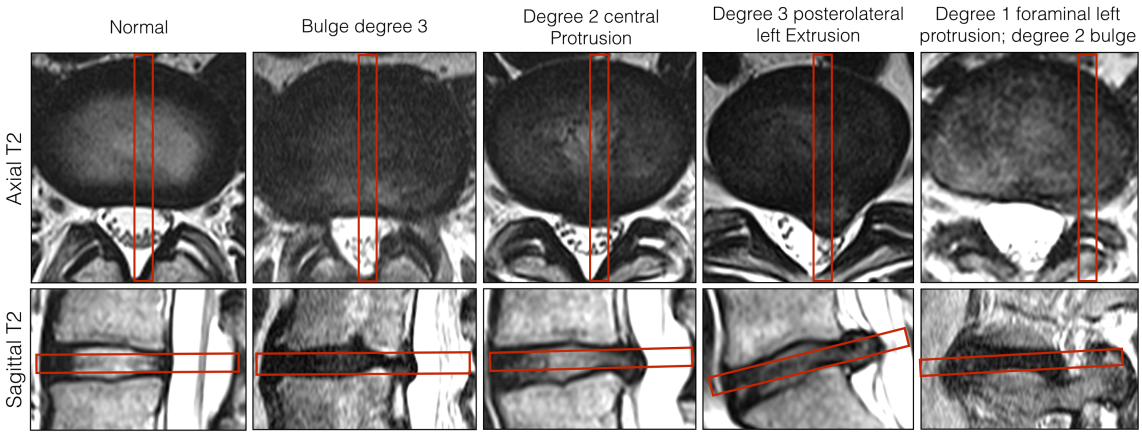


Figure 6.6: **Bulges and herniations**. Each column is a separate disc (from a number of patients). The red boxes on each axial slice note the intersecting sagittal slice, and vice versa. The difference between the bulge and assorted types of herniations is shown. In the case of a normal disc there is no posterior section deformation seen in either axial or sagittal slices. In the bulged case, the posterior side of the disc is seen displaced in a range of sagittal slices. In the herniated cases, the posterior side of the disc presses out in a localised manner. The herniations differ by type (protrusion-extrusion), location (central, postero-lateral, foraminal) and degree (1-3), e.g. the mass displaced.

6.3.1 Measurement, Dataset and Ground Truth Details

Experiments were performed on the full 300-patient dataset described in Chapter 3. The discs were labelled with the following relevant labels by a radiologist, as already described in Chapter 3:

- **Disc bulge**: There are 1219 grade 0 (normal), 398 grade 1 (slightly bulged), 153 grade 2 (moderately bulged), 29 grade 3 (severely) discs in the dataset.
- **Herniation**: There are 1634 non-herniated, and 165 herniated discs in the dataset. The herniations are characterised by their degree (0 to 3), location (central/postero-lateral left/postero-lateral right/foraminal left/foraminal right), type (protrusion/extrusion/sequestration), and associated nerve root compression (none, touching, pressing, displaced).

6.3.2 Support Regions and Features

The support regions are placed according to the segmentations and detections as follows. The region position and angle are fixed at posterior edge of the disc, at point P_{disc}, at the angle of the disc (from VB detections). The point P_{disc} is found according to the vertebrae segmentations, and the disc angle from vertebrae bounding boxes, as for Pfirrmann grade and narrowing. Prior to extraction, the image is resized according to the anterior-posterior width w_{AP} of the disc, making the vertebral anterior-posterior width 60 pixels, as for Pfirrmann grade/narrowing.

The size of the region for bulge/herniation is learned by grid search, and the optimal size is found to be 31-by-31 voxels. Thus, the sagittal feature vector for herniation/bulge has $31 \times 31 + 2 = 963$ elements, where the extra two elements are the disc level, and the disc height.

6.3.3 Mapping to Measurements

For successful measurement prediction, precise support region localization and suitable features are required. To evaluate the effect of localization quality (both detection and slice selection) and dataset variability, we perform experiments with two types of testing sets – 'picked' and full (as explained in the next section) – and compare hand-placed region results to fully automatic ones.

See Figure 6.7 for the improvement from using just the original dataset, to using the augmented dataset for training. Note how if the training set is augmented with extra samples in training, area under the ROC curve (for herniation/bulge classification) on the test data drops off slower with degrading localization of region.

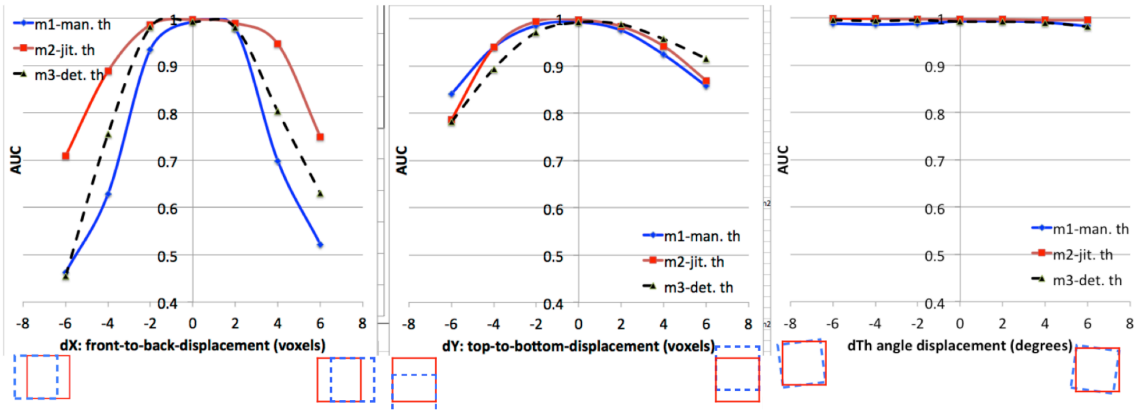


Figure 6.7: **Effect of training set augmentation.** The area under herniation/bulge vs. normal ROC curve (higher is better) is plotted against displacement in support region localization. The three curves correspond to three different classifiers. The first, m1, is trained on manually marked regions, the second, m2 on manually marked and augmented regions, and the third, m3, on automatically detected regions. Note how m2 is most immune to localization inaccuracies, for displacement in the x direction (anterior-posterior).

Training

The classifiers for herniation/bulge are trained on a slice-by-slice basis. That means, each bulged/herniated slice provides a training sample. At training time, slices showing sufficient herniation/bulge are picked for training manually and based

on radiological Ground Truth. The C-SVM is used, with the parameter C learned on a hold-out validation set.

Implementation details. The value of C for the SVM cost function is learnt by a grid search on a 57-patient hold-out set, with an exponential range of 10^{-2} to 10^4 respectively, using the LIBLINEAR package (Fan et al. [2008]). The picked best value was 0.1.

Inference. The result is the thresholded mean of the classifier scores for the block of five slices centred on mid-disc.

6.3.4 Evaluation Protocol

Dataset: The dataset is split into training and testing sets as described in Section 6.1.3.

Herniation/Bulge Prediction. We assess by classification accuracy (percentage of discs that were classified correctly), area under the ROC curve (AUC), and the ROC curve Equal Error Rate (EER) – the point on ROC curve where the false accept rate and false reject rate are equal.

6.4 Discussion and Comparison

A summary of the overall results for the three measures is given in Table 6.1. Example Pfirrmann grading result are shown for two patients in Figure 6.9. On the full dataset, we observe the following results. On herniation/bulge we achieve 80.3% accuracy, with AUC value of 0.863, EER 0.21. On Pfirrmann grading, we achieve 87.4% accuracy to +/- 1 of the radiologist grade. For narrowing, 83.7% accuracy, with AUC value of 0.911, EER 0.16.

6.4.1 Results by measurement type

For **herniation/bulge**, we achieved 84.9% (80.3%) accuracy at 0.13 (0.21) EER for manual (automatic) region placement on the *full*, and 98.9% (87.8%) accuracy at

Localis.	Test set	Hern./bulge			Pf±1	Narrowing		
		Acc.	AUC	EER	Acc.	Acc.	AUC	EER
Manual	Picked	**98.9**	99.95	0.01	**97.3**	**91.9**	97.3	0.09
Auto	Picked	87.8	92.6	0.21	89.2	87.8	86.9	0.13
Manual	Full	**84.9**	93.4	0.13	**90.1**	**85.1**	92.8	0.13
Auto	Full	80.3	86.3	0.21	87.4	83.7	91.1	0.16
Intra-κ	122 pat.	0.72			0.91	0.89		

Table 6.1: **Prediction performance.** The radiological measurement prediction results are given for both manual and automatic localization on both full and the hand-picked 'picked' subset (Figure 6.4). Along with the prediction results, the intra-rater variability is given as the standard κ-value. The manual prediction accuracies are highlighted by bold typeface.

0.01 (0.21) EER on the *'picked'* dataset. The method works well on cases where the herniation/bulge is clearly visible in the sagittal slices. It works less well on images with low resolution, and ambiguous cases. Our fully automatic herniation/bulge prediction performance on the full dataset is lower than presented in literature for herniation (Alomari et al. [2011b, 2013, 2010a], Ghosh et al. [2011b,a], Koh et al. [2010, 2012], Tsai et al. [2002]), however with the important difference that our system is fully automatic from start to finish, and on a clinically representative dataset. This is ignoring labelling errors (around 10% of all cases). In addition, our dataset includes herniations and bulges of various degrees, and the boundary cases are quantified as 'herniated/bulged' or 'not herniated nor bulged' in the current implementation. In future work, a continuous regression may be more appropriate.

Our results on the 'picked' dataset with manual region placement are similar to those quoted by Koh et al. [2012]. However with manual placement on our 'full' dataset, they drop to levels similar to Alomari et al. [2010a]. We believe this to be due to the difference between our datasets, our features, or because only herniated, and not bulged patients, were considered in their dataset.

For **Pfirrmann grade**, we achieved 90.1% (87.4%) accuracy of ±1 prediction

for manual (automatic) region placement on the *full*, and 97.3% (89.2%) accuracy on the *'picked'* dataset. This compares well with results cited in the literature, however, we predict the (collapsed) 5-step grade instead of degeneration/non-degeneration as in previous literature (Alomari et al. [2009a], Hao et al. [2013], Lootus et al. [2014], Mateos et al. [2014], Neubert et al. [2013], Oktay et al. [2014], Unal et al. [2011]).

For **narrowing**, we achieved 85.1% (83.7%) accuracy at 0.13 (0.16) EER for manual (automatic) region placement on the *full*, and 91.9% (87.8%) accuracy at 0.09 (0.13) EER on the *'picked'* dataset. This compares well with the intra-observer kappa of 89%. No prior literature for automated narrowing prediction exists to our knowledge.

6.4.2 Effect of Localization Accuracy by Measurement Type

Based on these results, the drop on the full dataset performance from manual to automatic is 4.5%, 2.8% and 1.4% in accuracy for herniation/bulge, Pfirrmann grade, and narrowing, respectively, and 0.07 and 0.03 in EER for herniation/bulge and narrowing respectively. Therefore, herniation/bulge prediction appears more sensitive to localization than the others. This could be since the herniation/bulge features capture the displacement of the posterior disc region, whereas the former capture overall disc shape and appearance change.

Size of the region For herniation/bulge, the optimal size of the region is centred on the posterior edge of the vertebrae corners, and contains the vertebrae corners. For both Pfirrmann grade and narrowing, the regions cover the whole disc section.

Dataset Heterogeneity Note that our dataset contained scans from a number of sites; different scanner manufacturers and models; acquired on a number of different protocols. The pulse sequence (TE, TR, flip angle), scanner (magnetic field strength, manufacturer, model) and resolution parameters (in-slice, slice spacing) varied.

6.4.3 Causes of Failure

The fully automatic predictions can fail for the following five broad reasons. First, due to support region placement error, due to detection or segmentation inaccuracy/failure. Second, due to failure to pick the mid-sagittal slice. Third, due to feature failure. Fourth, due to low image quality. And fifth, due to possible ground truth inconsistency.

Based on qualitative analysis on herniation/bulge failure modes on discs 1-5 (1500 in total), there were 60 discs with detection/segmentation failure (3.5%), 25 discs (1.5%) with low imaging quality, 23 discs (1.3%) failed slice selections. This accounts for 6.3% of all measurements, which is approximately half of all failure cases, leaving the other half due to possible feature failures. These should be investigated in more detail in a future study.

The third category of reasons – feature failure – can be for a number of subcauses. Cause **A**, due to normalisation failure, which can be due to vertebrae segmentation failure, or discrepancy between the radiological 'normalisation' and our normalisation scheme. Cause **B**, due to a unusual disc variant not represented in the training set. Cause **C**, due to inability of the feature set to capture a variation.

6.5 To Segment or Not to Segment?

In this section we pose the question, how important is segmentation in a computer vision pipeline for radiological measurement to investigate the problem of what is the best method of obtaining support regions.

On the one hand, accurate segmentations can target the measurements more precisely to just the relevant areas. On the other hand, as segmentation algorithms are often brittle, and thus accurate segmentations may not be available, methods which do not require segmentations may produce more reliable results. Furthermore,

performing hard segmentations in images with large amount of partial volumes may compromise prediction results, since precise organ boundaries cannot be determined in those cases.

To answer the question, we assess Pfirrmann grade prediction results using feature support regions localised based on three different strategies named **V-det** (vertebra detection), **V-seg** (vertebra segmentation), and **D-seg** (disc segmentation), illustrated in Figure 6.8. This explores three points on the 'no segmentation' to 'full disc segmentation' spectrum.

Note that the segmentation analysis for support region placement applies better on Pfirrmann but not as well on e.g. bulge/herniation since for Pfirrmann grade the disc signal and height, whereas for bulge/herniation, disc material displacement with respect to vertebrae corners is evaluated.

6.5.1 Three Alternative Support Regions

Taking a clinical MRI scan as input, this step outputs the feature support region for the six lumbar discs in three different ways, as contrasted in Figure 6.8. In **V-det**, the region is defined as a rectangle between vertebrae bounding boxes. In **V-seg**, the region is defined as a rectangle based on vertebrae segmentations, and excluding any vertebrae voxels. In **D-seg**, the region is defined as the disc segmentation result, the disc mask. The full algorithm from image to vertebrae and disc segmentation is sequential: (1) vertebrae detection, (2) vertebrae segmentation, (3) disc segmentation, with each step automatically initialising the next. The vertebrae detections are performed as described in Chapter 4, and the segmentations as described in Chapter 5. As might be expected, each step in the process has some degree of failure rate. So the more steps we employ, the greater the potential for failure.

Implementation details. The segmentations and gradings are performed in the mid-sagittal slice (as clinical standard), initialising the support region selection based

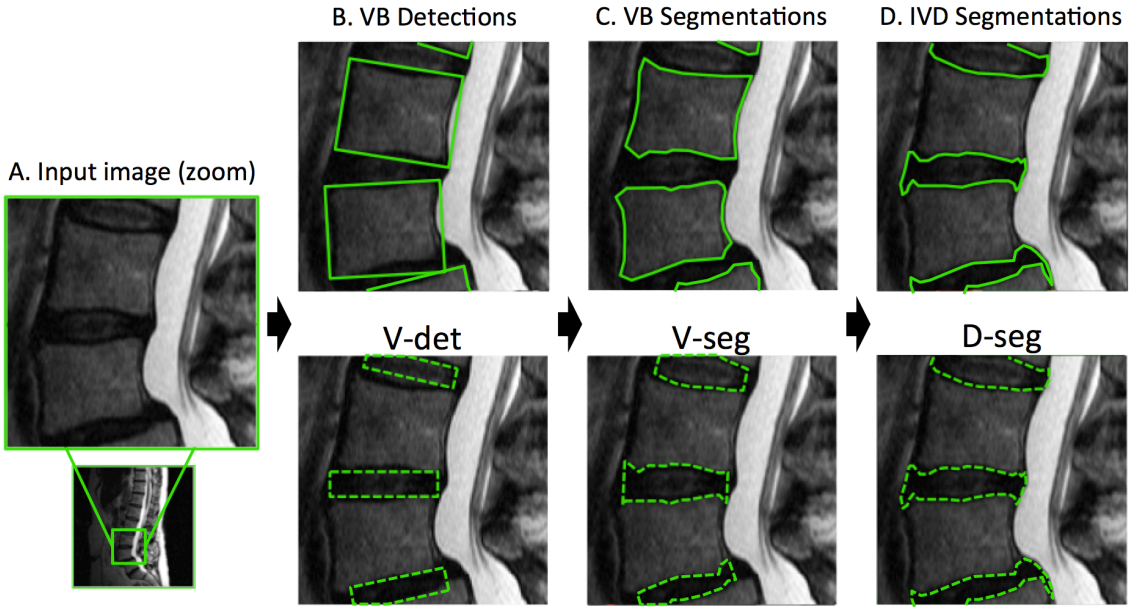


Figure 6.8: **Computation of the feature support regions.** The top row illustrates the detection and segmentation steps; the bottom row shows the three corresponding feature support regions for the (B) **V-det**, (C) **V-seg**, and (D) **D-seg** methods. The full segmentation pipeline consists of vertebrae detection, vertebrae segmentation, and intervertebral disc segmentation. In the top row, the green lines show the detection and segmentation outputs. The resulting support regions are shown as green dashed lines in the bottom row. The V-det pipeline is the shortest, involving no segmentation, and the D-seg pipeline is the longest, involving both vertebrae and disc segmentation. See Section 6.5.1 for more detail.

on the automatically labelled bounding boxes. To arrive at the final box regions for V-det, the box is placed at the angle of its lower neighbouring vertebra, with its height equal to a quarter of the mean heights, and its width equal to the mean width of the neighbouring vertebrae bounding boxes (optimal size found by gird search). In V-seg, the vertical 'walls' of the region are placed at the same positions as for the V-det case, and in D-seg the region is simply the segmented disc.

6.5.2 Image Features

We perform the experiments on features that can be extracted from regions of any shape (unlike the features used previously in this Chapter). Thus, the results can be more directly influenced by the level of segmentations available.

To characterise the disc, a number of intensity and shape features are extracted from the feature support region, as described below. The image intensities are normalised to the median vertebrae intensities as explained in Section 6.1.1;

Intensity features. Two groups of intensity features are extracted: first, a histogram normalised so that the highest entry is one; second, four global statistical features: standard deviation, kurtosis, skewness, and entropy.

Shape features. The mid-height to width ratio h/w of the feature support region, approximating that of the disc, is used as the shape measure.

Implementation details. The histogram feature is modelled with 21 bins, making up a 26-dimensional feature vector for each disc. The mid-height to width ratio h/w is measured in the middle of the feature support region – the rectangle for V-det, the disc space for V-seg, and the segmented disc for D-seg.

Baseline Features. In addition to the above features, as a baseline we assess the performance of our system using two previously proposed feature vectors: first, a 2-dimensional vector similar to Alomari et al. [2009a], containing Disc Mean Intensity and disc heigh to width ratio; secondly, a 6-dimensional vector as in Neubert et al. [2013], containing the five coefficients (means, variances, and relative weight) of a 2-peak GMM fitted to the intensities from the support region, along with the disc height to width ratio as in Neubert et al. [2013].

6.5.3 Mapping to Measurements

The feature vectors are mapped to measurements using the same types of Support Vector Regressors as described earlier in the Chapter under Section 6.2. However here, nine regressors are trained in total, for comparison against each other: one for each feature vector, on each type of support region.

Implementation details. The values of ϵ and C for the Support Vector Regression cost function are learnt by a grid search, as described earlier.

	Mean Int.	**GMM**	**Hist+**
V-det	73.4%	77.5%	79.7%
V-seg	81.1%	84.6%	85.2%
D-seg	79.6%	83.3%	**85.8**%

Table 6.2: **Numerical results.** Percentage of discs with Pfirrmann grade predicted to $\pm$ 1 accuracy using each of the features (columns) on each of the support regions (rows).

6.5.4 Evaluation and Comparison

A summary of the overall results is presented in Table 6.2, with example grading result shown for two patients in Figure 6.9, and for two more in Figure 6.10. The median detection error was 2.0mm; the mean vertebrae segmentation overlap measure was 0.808 ± 0.132 on fifty randomly selected patients. Overall, the best performing method is D-seg (85.8%) using the Hist+ features. For all three support region methods, the Hist+ features outperforms the baseline GMM and the Mean Int. features. Also, the segmentation methods (D-seg and V-seg) outperform detection only (V-det). For both Mean Int. and GMM, V-seg outperforms D-seg by a small margin, but performs similarly with Hist+ features. The D-seg with Hist+ features confusion matrix shows that the greatest errors are in predicting a grade of '3'.

Based on qualitative analysis on all discs from levels 1-5 (1425 in total), there were 50 discs with detection/segmentation failure (3.5%), 21 discs (1.5%) with low imaging quality, 19 discs (1.3%) failed slice selections. This sums to 6.3%, to approximately half of all failure cases in the test set, and covers the principal causes of error.

The Hist+ features provide a clear improvement. This may be because the Mean Int. is insufficiently discriminative, while the GMM parameters might vary

significantly between discs (since GMM fitting minimises error to the underlying distribution, but does not constrain the component centres). The Hist+ features provide improved discriminative power, are repeatable, and also include the global descriptors of the distribution shape (that the GMM can capture). In early experiments, we found the global descriptors to improve performance.

There was a 5–7% difference in performance between the methods employing segmentation and detection only, while there was little difference between the two segmentation approaches. That vertebrae are easier to segment than discs due to their more consistent and distinct appearance, as noted in previous work, may explain the small margin between V-seg and D-seg.

An interesting question is whether the lower performance of V-det is due its box-shaped support region or sub-optimal setting of its parameters. To answer this, we replaced the simple vertebra height based adaptation with a box height set from the height of the V-seg region. In other words, the support region is still a box but its height is based on the V-seg support region. The regression performance for this variation was 80.2%, 82.6%, 84.3% for the Mean Intensity, GMM and Hist+ features, which is very similar to the V-seg and D-seg methods. So indeed it seems that it is the size of the support region more than its exact shape that is responsible for the lower performance.

A final question is whether the sub-optimal size setting is affecting all the discs, or just a subset. One might expect that the support region for the L5-S disc to be problematic due to variability in the curvature of the spine at that location. Indeed, by excluding all L5-S discs the average V-det performance improves to 80.4%, 82.2% and 83.3% for the three features (computed on the box of original size).

One limitation of our study is that the grades could only be evaluated to within ± 1 Pfirrmann grade, in effect reducing the five grades to three in the evaluation step. This was due to the variability in the ground-truth mark-up. Nevertheless,

the system could still output the full range, and prior work has only reported results on a binary classification of healthy versus degraded.

6.6 Discussion

In this Chapter, we discussed learning for radiological measurements, going from localised anatomy (detections, labels, and segmentations from previous chapters) to radiological measurements.

We used simple patch features for prediction of three different radiological measures – Pfirrmann grade, disc space narrowing, and bulge/herniation. We investigated the effect of localisation accuracy in the process, comparing results in two cases – first, using manually localised support regions, and second, using automatically localised regions, e.g. according to both vertebrae detections and segmentations.

We finally asked the question, whether to segment or not to segment, and answered the question by comparing Pfirrmann grade prediction results from experiments with support regions obtained from three different levels of segmentation – ranging from no segmentation (vertebrae detection only) to full disc segmentation.

The results show that radiological measurements can be made using automated computer vision techniques. While useful in some cases, segmentation can possibly be avoided as it is prone to failure in other cases. The results seem to indicate that vertebrae segmentation suffices for localisation for Pfirrmann grade prediction and possibly for other measurements.

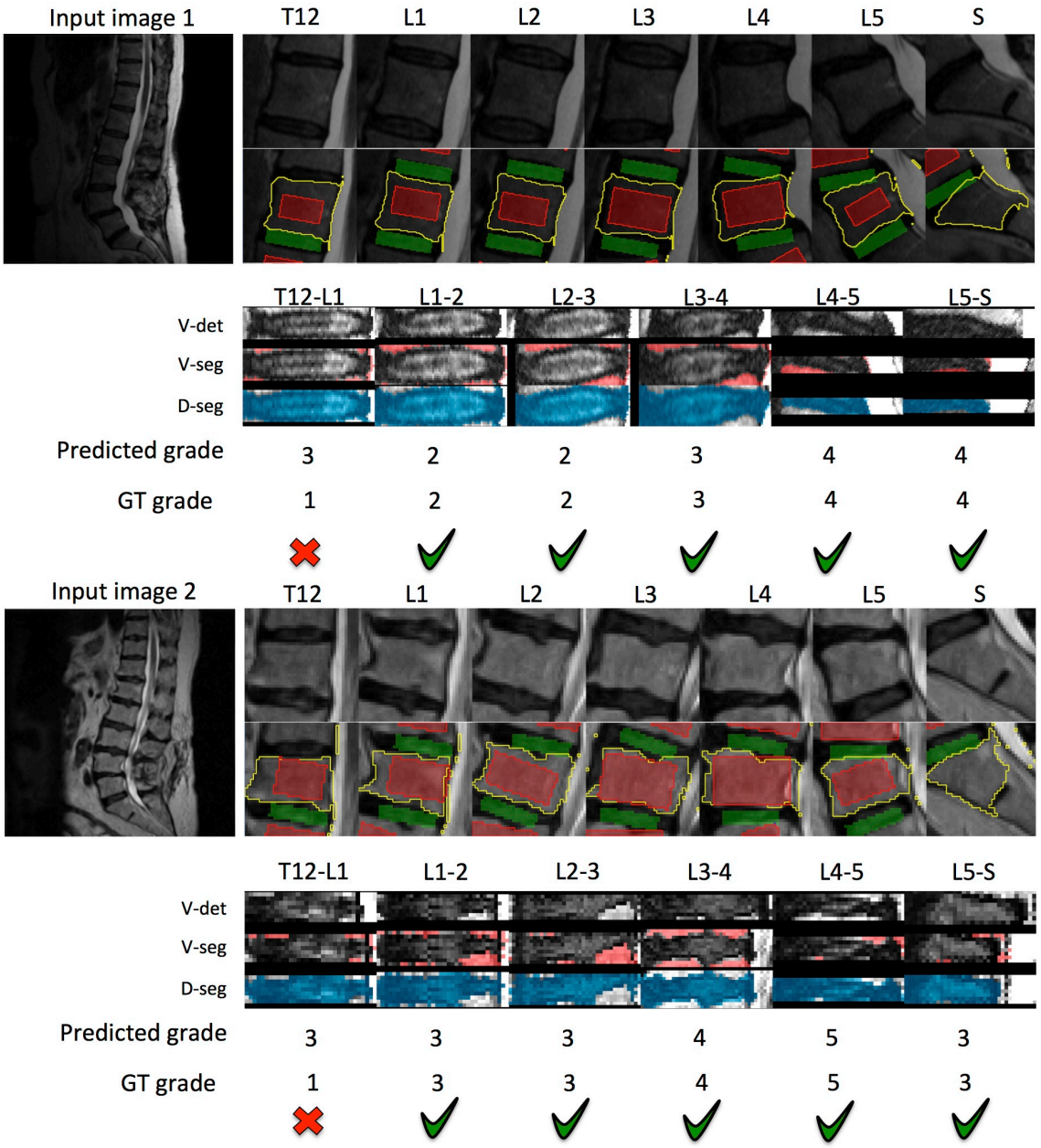


Figure 6.9: **Example results 1-2.** The full segmentation and grading results are given for two example input images. For each patient, in the second row from the top, the vertebrae segmentations are shown as yellow lines. In the bottom three rows, the extracted regions for feature extraction, along with the segmented areas for the vertebrae and the discs are given in the three bottom rows for V-det, V-seg, and D-seg methods (V-det based on vertebrae detections; V-seg on vertebrae segmentations, and D-seg based on disc segmentations). At the bottom, the grading result is given for the D-seg method. For both patients 1 and 2, five disc gradings succeed, and one fails. Note that the fact that both the discs T12-L1 are predicted two grades too high is not a systematic error, but randomly present in those two cases.

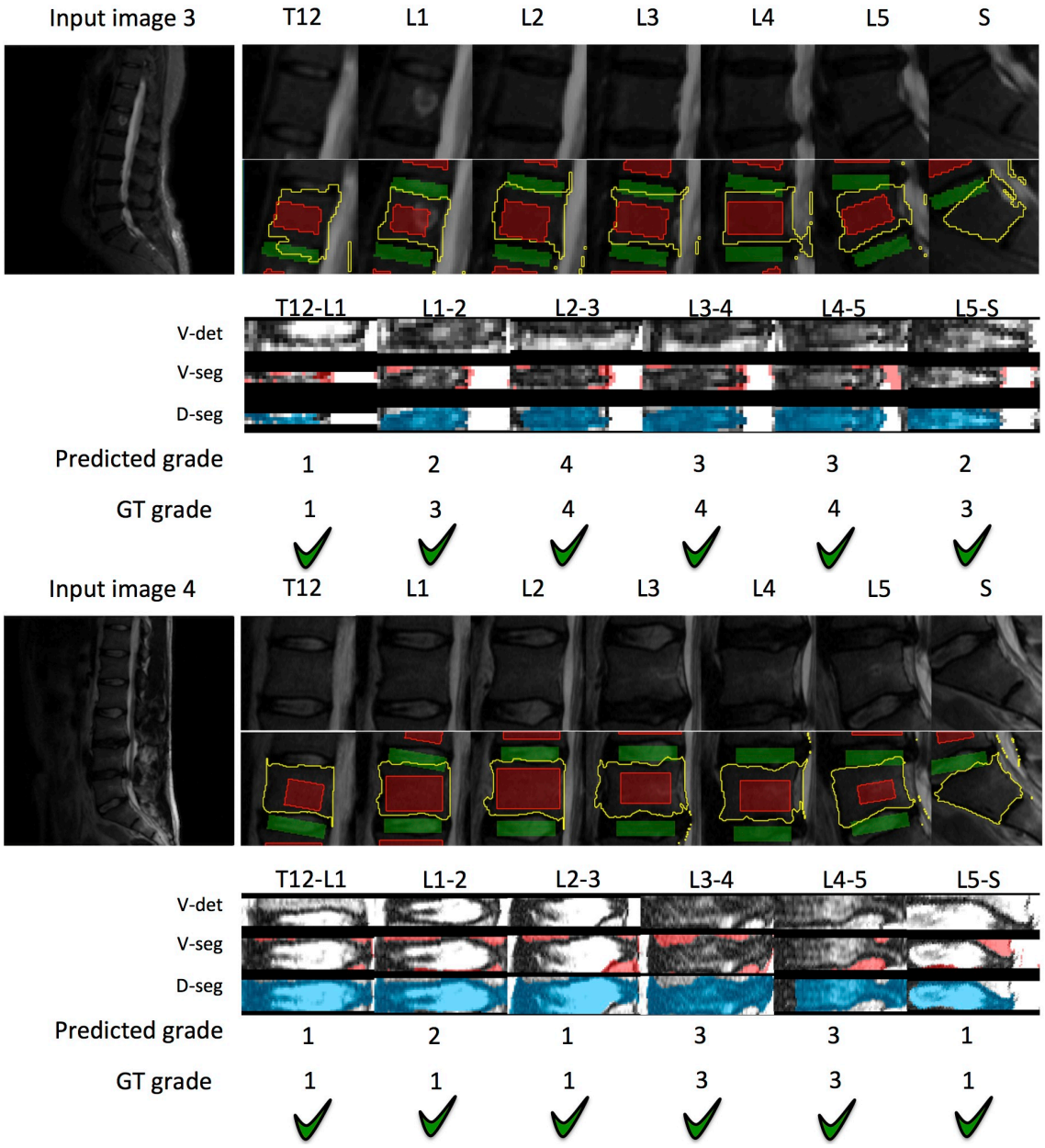


Figure 6.10: **Example results 3-4.** The full segmentation and grading results are given for two example input images. For each patient, in the second row from the top, the vertebrae segmentations are shown as yellow lines. In the bottom three rows, the extracted regions for feature extraction, along with the segmented areas for the vertebrae and the discs are given in the three bottom rows for V-det, V-seg, and D-seg methods (V-det based on vertebrae detections; V-seg on vertebrae segmentations, and D-seg based on disc segmentations). At the bottom, the grading result is given for the D-seg method. For both patients 3 and 4, all the six discs succeed. Note that in patient 3, discs L1-2, L3-L4, L4-5, and L5-S, and in patient 4, disc L1-L2, are off by one grade, but still considered correct predictions according to our ± 1 criterion.

Chapter 7

Summary and Future Work

In this Chapter, we summarise our achievements and contributions along with potential applications of the work in Section 7.1, and discuss some work in progress in Section 7.2, and potential future work in Section 7.3.

7.1 Achievements and Contributions

In this Section, we first discuss our contributions chapter by chapter, and then discuss the potential bigger picture application of the system.

In this thesis, we introduce a state-of-the-art, robust, completely automatic image analysis framework for spinal images with particular focus on spinal MRI. Our focus through the experimental work is on three aspects: (i) application of modern computer vision methods on spine analysis, (ii) usage of bottom-up 2D instead of more demanding top-down 3D methods, (iii) the question of the necessity of segmentation in an anatomy characterisation pipeline. We utilise modern and efficient computer vision and machine learning techniques and, critically, our framework can analyse standard clinical MRI datasets and does not require specialised protocols that might otherwise limit its application. In contrast to existing approaches, we process in 2D slices and combine the results, showing that complex 3D models are

not necessary for high performance.

Our framework has been tested on a large 300-patient, 1800-disc heterogeneous dataset of spinal MRI datasets from multiple centres.

In Chapter 4, we introduced a robust method for the detection and labelling of vertebrae in lumbar MRI. This system is able to handle lumbar scans of various lumbar field of view, in a wide range of scanning protocols.

In Chapter 5, as a next step in our pipeline, we developed a powerful Graph Cuts based segmentation method for automated delineation of vertebrae and discs in the spine, also capable of dealing with a wide range of MRI protocols.

In Chapter 6, we focused on the characterisation of spine anatomy using machine learning techniques. Experimentally, we automatically predicted three different radiological measurements – Pfirrmann grade, narrowing, and bulge/herniation. We also investigated the importance of segmentation in the assessment of Pfirrmann grade.

The framework could be applied to aid the analysis of patient anatomy for both research studies and clinical diagnostics purposes. The computerised methods could help to establish standards in the spine research community, who are actively seeking for systems better than the non-standard measurement protocols currently in use. The key improvements over the existing measurements are analysis speed, lower cost, and consistency. The framework introduced in this thesis could be a stepping stone to further developments in standardised spine analysis. We hope that the methods and tools introduced could be used to tackle the back pain management problem in the clinic, and help further spinal research efforts.

7.2 Work in Progress

7.2.1 Axial analysis: Distinguishing Herniations and Bulges

To provide more extensive analysis of spines, axial slices can be included. Since the sagittal slice thickness is high (4-5mm typically), complementary information on the 3D structure and hence the diagnosis of a number of conditions can be extracted from axial slices. The analysis of axial slices is challenging similar to sagittal slices – they are also thick, and sometimes they are placed in a block structure at a constant angle; at other times, they follow the spine.

We have made some preliminary progress with axial image analysis, implementing a detection, segmentation and classification method on axial images.

Detection. Detection in axial slices is performed using a method similar to the one used on sagittal slices as described in Chapter 4. Detection in axial slices is more complicated because of extensive partial volumes between disc and vertebra voxels within slice, especially if the slice is cutting the disc at an angle. This also makes it more difficult to define the Ground Truth. The learned axial detectors are shown in Figure 7.1.

Segmentation. Segmentation in axial slices is performed using a method similar to the one used on sagittal slices as described in Chapter 5. Example axial segmentation results are shown in Figure 7.2, and visualised in 3D along with sagittal results in Figure 7.3. The spine segmentation in axial images is performed for each axial slice according to its spine bounding box detection, using the same patient-adaptive graph cuts Boykov and Jolly [2001] framework, as in sagittal slices. The seeds are obtained as follows: the foreground seeds by eroding the bounding box to half its size, and the background seeds by dilating the bounding box to 125% its size. The region terms are modelled as three-component Gaussian Mixture Models (GMMs) according to the image intensities in the Obj. and Bkg. seeds for both the vertebrae

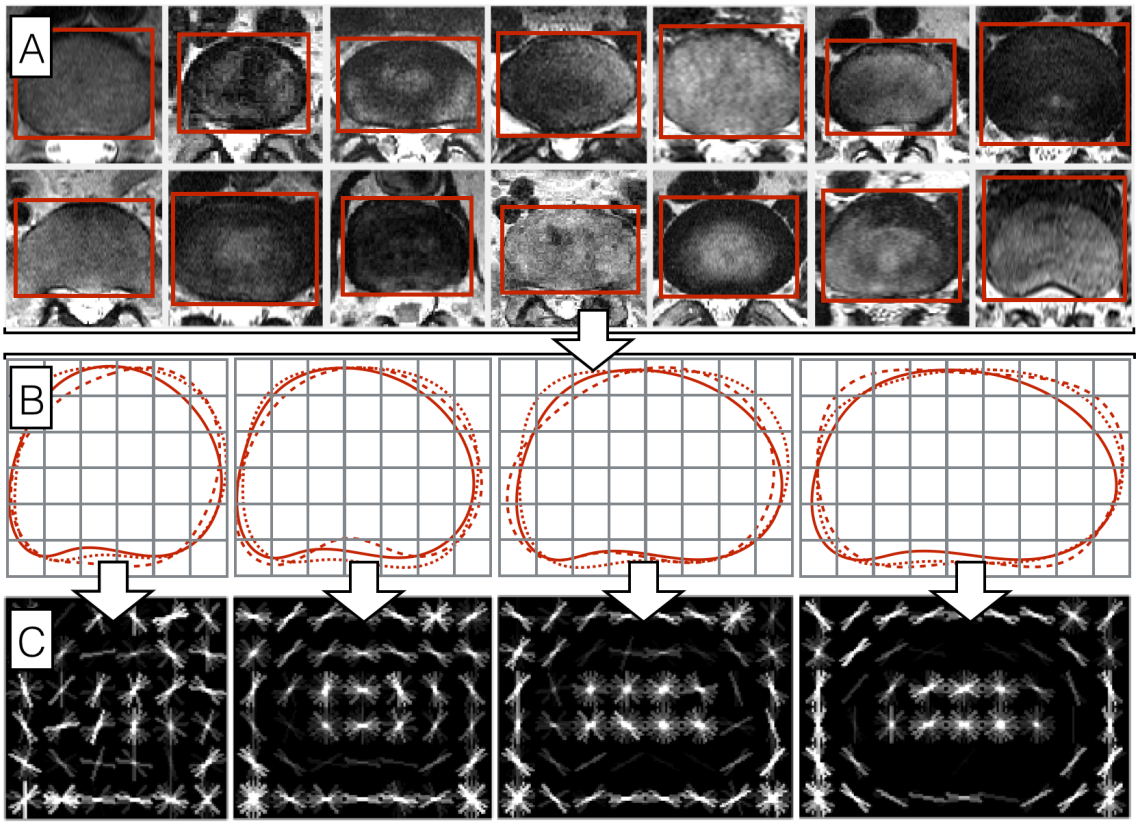


Figure 7.1: **Axial HOG models.** The HOG model is shown trained on axial slices: annotated training samples shown in (A) were used to train HOG templates with four different aspect ratios of (6,7,8,9)-by-6 HOG cells. Example disc contours are shown overlapped with the HOG cells in (B), and the resulting HOG templates in (C).

and disc segmentations. Three components were picked as best performing at earlier experiments.

Mapping to Measurements and Learning. The axial slices contain extra information for herniation prediction. They can help to distinguish between different types of herniation. The shape of the back-side of herniated discs is parametrised in Figure 7.4. Evidently, and based on initial experiments the shape feature is not enough to distinguish herniations from bulges. Some encouraging initial experiments were performed with SVM classifiers on features such as in Section 6.3 in Chapter 6, except that pixel values from histogram-normalised axial slices were concatenated to the feature vector. The support regions for the features were obtained manually.

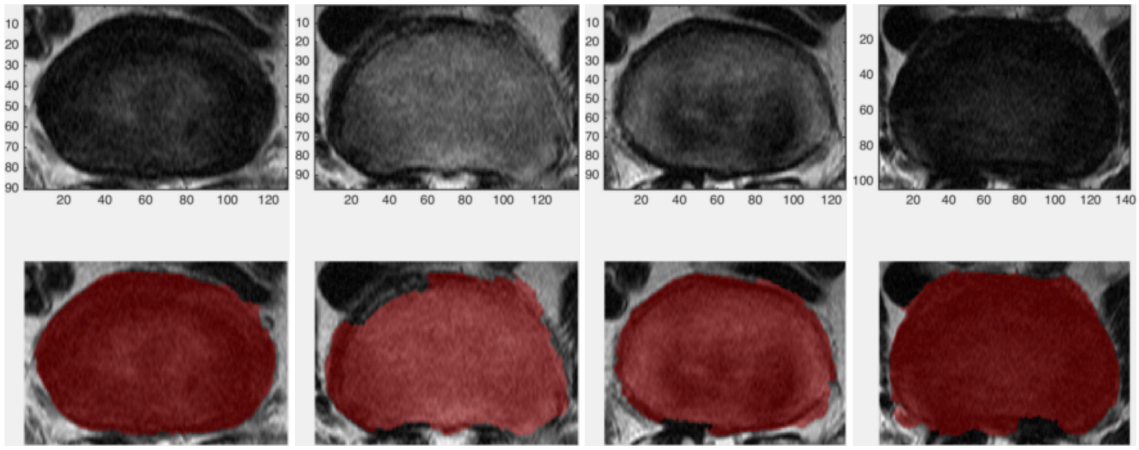


Figure 7.2: **Example axial segmentations** for the lumbar spine of a number of patients. In the top row, axial images of the disc, and in the bottom row, the segmentations are shown.

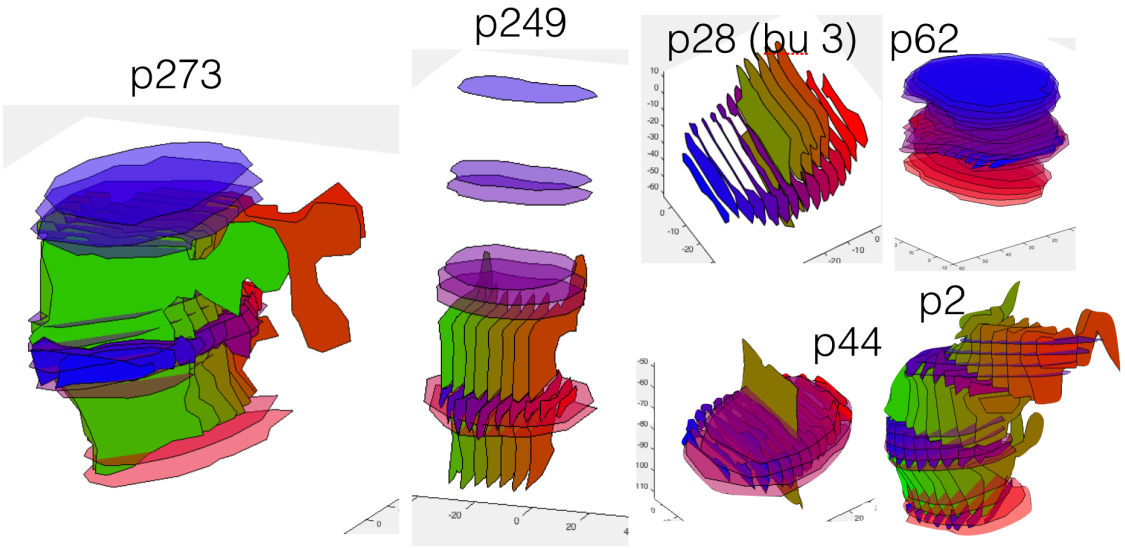


Figure 7.3: **Segmentations in 3D.** Some example axial and sagittal segmentation results are shown in 3D. The colour coding is used to distinguish slices from each other, with no other meaning.

They were extracted from the manually selected axial and sagittal slices (one of each per disc) where the disc deformation in both normal, bulged and herniated discs looked most 'herniation-like'. This initial experiment assumed the as based on an interpretation of Fardon et al. [2014] who propose a radiological consensus for herniation evaluation. The results from these experiments were poor however, which is likely because the herniation might not be parametrisable only by the changes to

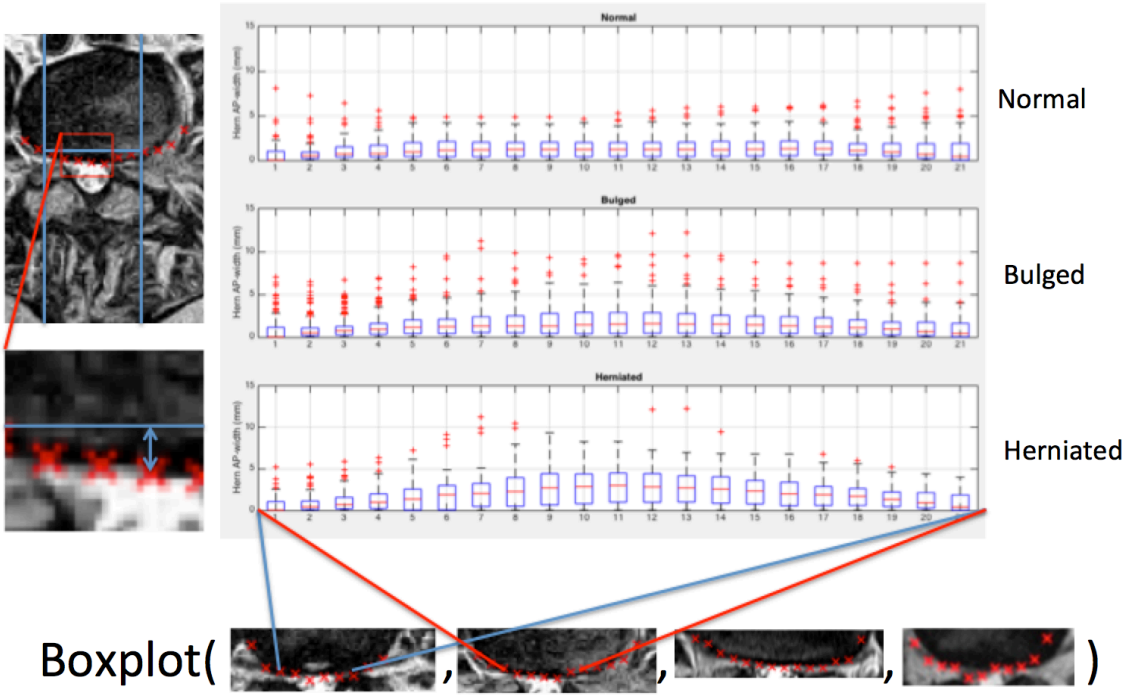


Figure 7.4: **Shape features for herniation measurements.** Box plots of the protruding distance at points across the disc for normal, bulged, and herniated discs are shown. This experiments demonstrated that discs are not necessarily distinguishable from each other based on only the shape of the disc posterior boundary.

the outer shell of the disc, but also to the internal changes, e.g. the migration of the nucleus pulposus from the centre of the disc. Usage of image intensity and texture features in combination with shape features, and extraction of information from more slices could result in higher performance.

7.2.2 Spinal Cord Segmentation

The Spinal Cord can act as a predictor to a number of spinal conditions such as Stenosis possibly related to back pain.

In sagittal slices, as visible in Figure 7.5, the spinal tissues appear in several curved bands: tissues anterior to the spine, the ladder-like vertebral column, the bright dural sac, the vertebral posterior elements, and the posterior fat. Because of this ordered curve structure, we can use a dynamic programming solution as

previously applied for tiered scene segmentation in computer vision (Felzenszwalb and Veksler [2010]). We represent the vertical curved boundaries in the spine as chain graphical models. The solutions will be curved lines funning from top to the bottom of the image.

We segment the spinal cord in sagittal slices using a dynamic programming approach where the unary cost is the gradient in the image, and the pairwise cost is related to the left-right step, with five steps (-2,-1,0,1,2) allowed with costs $(2, 1, 0, 1, 2)$, and the number of nodes is equal to the number of pixel rows in the image. Additionally, the costs were made cheaper around the back side of the vertebrae according to the vertebrae bounding boxes in a region of 2 cm around the back side of the boxes. First, the anterior line was found (red in the Figure) and then the posterior line. Finally, the blue back side line of the patient was found as well. The Viterbi algorithm was used to solve the dynamic programming problem (Viterbi [1967]). Some preliminary results in spinal cord segmentation are shown in Figure 7.5. The preliminary results show reasonable segmentation performance for the anterior side of the spinal cord yet fail at the back side.

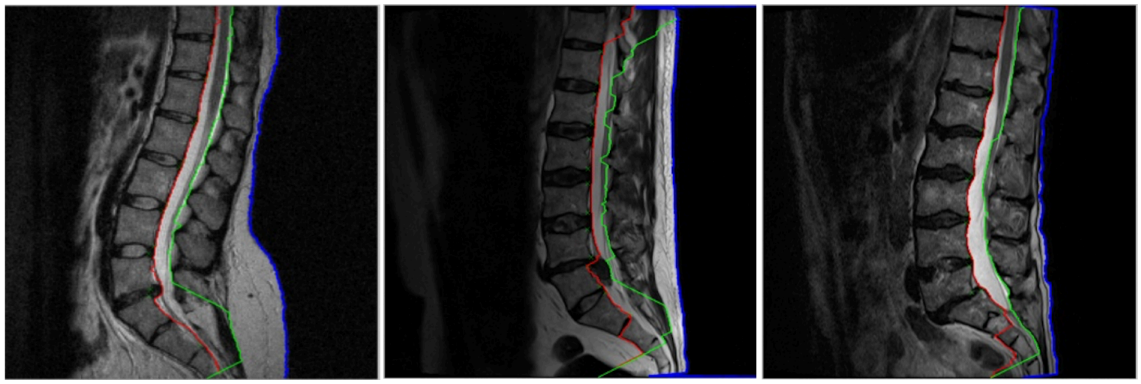


Figure 7.5: **Spinal cord Segmentation.** Some preliminary spinal cord segmentatiton results are shown. The red and green lines show the front and back of the cord respectively, and the blue line shows the body outline segmentation.

7.3 Future Work

In this Section, we discuss some potentially promising directions for future work. They are clustered into three themes: **first**, improvements and extensions to the pipeline; **second**, other use cases.

7.3.1 Improvements and Extensions to Pipeline

- **Extra image data**. Extra image data could be included in the analysis, and more clinical awareness built into the models. The extra image data could be both from extra slices, and from extra series, and the. In more detail, the extra data could be from all available sagittal, axial and coronal T1 and T2 slices, including scout scans. Usage of the axial and sagittal slices in combination could help produce sub-voxel accuracy segmentations over partial volumes. Note that the sagittal and axial; T1 and T2 slices could be combined either as discussed in Section 7.2, or by resampling the two (aligned) volumes. This could be done to construct an approximately isotropic volume in places where both sagittal and axial slices are available.

- **Extra clinical awareness**. This could constitute consideration of more anatomical variants (e.g. transitional vertebrae) and a broader ranges of abnormalities/measurements (e.g. Schmorl's nodes, Modic changes, stenosis, etc.).

- **Extra anatomical structures**. In addition to just discs and vertebrae, the analysis could be extended to include further structures related to the spine – e.g. the spinal cord, ligaments, muscles. A broader range of radiological measurements could be automated. Each detection, labelling, segmentation, and measurements could be performed based on those (further) structures for diagnosis of a broader range of conditions.

- **More refined models** In addition to just vertebral bounding boxes, anatomically relevant landmarks could be localised. Information could be extracted from these extra data using broader ranges of image representations from computer vision (e.g. features and descriptors) and integrated into (possibly 3D models) modelling the physiology more directly as needed by the clinic. This could result in a high-quality 3D representation of the disc and the vertebrae that could help both spine research and clinical practice. And, the full vertebra, rather than just VB-s, might be possible to be segmented to some degree, although with huge partial volume effects due to thinness of the processes.

- **Performance improvements**. The broader ranges of image representations (e.g. learned vertebrae, disc edges) combined with broader anatomy modelling (e.g. ligaments, vessels, muscles) and data (sagittal, axial, coronal T1, T2) could result in more discriminative detection, segmentation and radiological measurement methods, e.g. higher quality information. In particular, modelling anatomical elements such as the vessels neighbouring the discs may help with determining the uncertain anterior disc boundaries by inference with more confidence, and be a direct input to further processing for radiological measurements in the pipeline. This higher quality information could turn into knowledge that could be applied on a broader range of problems such as neurosurgery planning.

- Extension of the labelling algorithm in particular (e.g. as in Glocker et al. [2012] or Ma and Lu [2013]) could make the framework directly applicable on an even broader range of images beyond just lumbar containing the sacrum.

- In order to learn more comprehensive diagnostic capabilities into the machine learning models, more variants of various disc abnormalities could be simulated in a system with simulated anatomy, and simulated MR imaging process.

Then models could be trained in virtual reality which are invariant to varying anatomy, imaging protocol, and pathology. Neural networks could be used to learn deeper representations of the conditions on the additional data.

- An iterative segment-diagnose-segment-diagnose-segment loop could be performed whereby the segmentation algorithm's performance improves iteratively based on improving diagnosis accuracy. In addition to boundaries, some uncertainty measures, perhaps based on the gradient sharpness could be output, perhaps in tri-map format, or by alpha-matting. Tri-maps could also be used in annotations for assessment of segmentation results.

7.3.2 Further Use Cases

- Correlation to pain: The pipeline could be used to in experiments to find correlations between patient symptoms (e.g. back pain) and image appearance directly. For those studies, additional populations of pain-free patients would be required. Additionally for those studies, for soft tissue issues, the patient position in the scanner plays an important role. Thus, data should be acquired under different placements of the patient in the scanner (e.g. standing, laying in different positions) for a fuller picture.

- Scan acquisition planning: In fact, detections in scout scans could be used to automatically propose most optimal ways to acquire the slices in the clinic. Perhaps some automated pathology analysis could be performed in the scout scans that could suggest variable resolution acquisitions (e.g. higher resolution around pathologies).

- Spine surgery planning: The pipeline could be used to plan spine surgeries.

Bibliography

04 2015. URL http://www.jamesdisabilitylaw.com/back-injuries.htm.

R. S. Alomari. Computer aided diagnosis system for lumbar spine. In *ISABEL*, 2011.

R. S. Alomari, J. J. Corso, and V. Chaudhary. Desiccation diagnosis in lumbar discs from clinical MRI with a probabilistic model. In *ISBI*, 2009a.

R. S. Alomari, J.J. Corso, V. Chaudhary, and G. Dhillon. Automatic diagnosis of lumbar disc herniation with shape and appearance features from MRI. *SPIE*, 2010a.

R. S. Alomari, J. J. Corso, V. Chaudhary, and G. Dhillon. Lumbar spine disc herniation diagnosis with a joint shape model. In *Proceedings of MICCAI Workshop: Computational Spine Imaging*, 2013.

R.S. Alomari, J.J. Corso, and V. Chaudhary. Abnormality detection in lumbar discs from clinical MR images with a probabilistic model. *Int J Comput Assist Radiol Surg.*, 2009b.

R.S. Alomari, J.J. Corso, V. Chaudhary, and G. Dhillon. Computer-aided diagnosis of lumbar disc pathology from clinical lower spine MRI. *Int J Comput Assist Radiol Surg.*, 2010b.

R.S. Alomari, J.J. Corso, and V. Chaudhary. Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. *IEEE TMI*, 2011a.

R.S. Alomari, J.J. Corso, V. Chaudhary, and G. Dhillon. Toward a clinical lumbar cad: herniation diagnosis. *Int J CARS*, 6:119–126, 2011b.

F. Alyas, D. Connell, and A. Saifuddin. Upright positional mri of the lumbar spine. *Clinical Radiology*, 63, 2008.

M.S. Aslan, A. Ali, H. Rara, B. Arnold, A.A. Farag, R. Fahmi, and P. Xiang. A novel 3d segmentation of vertebral bones from volumetric ct images using graph cuts. In *ISVC*, pages 519–528, 2009.

M.S. Aslan, A. Ali, H. Rara, and A.A. Farag. An automated vertebra identification and segmentation in ct images. In *ICIP*, 2010.

I.B. Ayed, K. Punithakumar, G. Garvin, W. Romano, and S. Li. Graph cuts with invariant object-interaction priors: Application to intervertebral disc segmentation. *IPMI*, 2011.

M.C. Battie, T. Videman, L.E. Gibbons, L.D. Fisher, H. Manninen, and K. Gill. 1995 volvo award in clinical sciences - determinants of lumbar disc degeneration. *Spine*, 20(24):2601–2612, 1995.

N. Bogduk. *Clinical Anatomy of the Lumbar Spine and Sacrum*. Elsevier, 2005.

Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, volume 2, pages 105–112, 2001.

Y. Boykov and O. Veksler. Graph cuts in vision and graphics: Theories and applications. *Math. Models of C.Vision: The Handbook*, 2006.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (11):1222–1239, 2001.

C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

C. Chevrefils, F. Cheriet, G. Grimard, and C.-E. Aubin. Watershed segmentation of intervertebral disk and spinal canal from mri images. *ICIAR*, 2007.

C. Chevrefils, F. Cheriet, C.-E. Aubin, and G. Grimard. Texture analysis for automatic segmentation of intervertebral disks of scoliotic spines from mr images. *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE*, 2009.

M.P. Chwialkowski, P.E. Shile, D. Pfeifer, R.W. Parkey, and R.M. Peshock. Automated localization and identification of lower spinal anatomy in magnetic resonance images. *Computers and Biomedical Research*, 24(2), 1989.

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61 (1):38–59, 1995.

J.J. Corso, R.S. Alomari, and V. Chaudhary. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. In *MICCAI*, 2008.

A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in CT studies. In *MICCAI workshop on Probabilistic Models for Medical Image Analysis*, 2011.

N. Dalal and B Triggs. Histogram of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 886–893, 2005.

C. Davatzikos, D. Liu, D. Shen, and E.H. Herskovits. Spatial normalization of spine mr images for statistical correlation of lesions with clinical symptoms. *Radiology*, pages 919–926, 2002.

E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, pages 228–247, 2009.

J.C.T Fairbank and P.B. Pynsent. The oswestry disability index. *Spine*, 25(22): 2940–2953, 2000.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

D.F. Fardon and P.C. Milette. Nomenclature and classification of lumbar disc pathology. *Spine*, 26(5):E93–E113, 2001.

D.F. Fardon, A. L. Williams, E.J. Dohring, F.R. Murtagh, S.L.G. Rothman, and G.K. Sze. Lumbar disc nomenclature: version 2.0 recommendations of the combined task forces of the north american spine society, the american society of spine radiology and the american society of neuroradiology. *Spine*, 14:2525–2545, 2014.

P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 2005.

P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

P.F. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *CVPR*, pages 3097–3104, 2010.

M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, c-22(1):67–92, January 1973.

D. Forsberg, C. Lundström, M. Andersson, L. Vavruch, H. Tropp, and H. Knutsson. Fully automatic measurements of axial vertebral rotation for assessment of spinal deformity in idiopathic scoliosis. *Phys. Med. Biol.*, 58:1775–1787, 2013.

J. C. Gamio, S. J. Belongie, and S. Majumdar. Normalized cuts in 3D for spinal MRI segmentation. *TMI*, 2004.

S. Ghosh, R. S. Alomari, and V. Chaudhary. Computer-aided diagnosis for lumbar MRI using heterogeneous classifiers. In *ISBI*, 2011a.

S. Ghosh, R. S. Alomari, V. Chaudhary, and G. Dhillon. Composite features for automatic diagnosis of intervertebral disc herniation from lumbar MRI. In *Eng Med Biol Soc*, 2011b.

S. Ghosh, R.S. Alomari, V. Chaudhary, and G. Dhillon. Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis. In *SPIE*, 2011c.

S. Ghosh, M.R. Malgireddy, V. Chaudhary, and G. Dhillon. A new approach to automatic disc localization in clinical lumbar MRI: Combining machine learning with heuristics. In *ISBI*, 2012.

S. Ghosh, V. Chaudhary, and G. Dhillon. Exploring the utility of axial lumbar MRI for automatic diagnosis of intervertebral disc abnormalities. *SPIE*, 2013a.

S. Ghosh, M.R. Malgireddy, V. Chaudhary, and G. Dhillon. A supervised approach towards segmentation of clinical mri for automatic lumbar diagnosis. In *Proceedings of MICCAI 2013 Workshop: Computational Spine Imaging*, pages 167–177, 2013b.

B. Glocker, J. Feulner, A. Criminisi, D.R. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *MICCAI*, 2012.

V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

S. Hao, J. Jiang, Y. Guo, and H. Li. Active learning based intervertebral disk classification combining shape and texture similarities. *Neurocomputing*, 2013.

R. Haq, R. Aras, D.A. Besachio, R.C. Borgie, and M.A. Audette. 3D lumbar spine intervertebral disc segmentation and compression simulation from mri using shape-aware models. *Int J CARS*, 2014.

R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of IEEE*, 67(5):786–804, May 1979.

S.A. Helo, R.S. Alomari, S. Ghosh, V. Chaudhary, G. Dhillon, M.B.A. Zoubi,

H. Hiary, and T.M. Hamtini. Compression fracture diagnosis in lumbar: a clinical CAD system. *Int J CARS*, 8:461–469, 2013.

C.L. Hoad and A.L. Martel. Segmentation of mr images for computer-assisted surgery of the lumbar spine. *Physics in Medicine and Biology*, 47:3503–3517, 2002.

M.A. Horsfield, S. Sala, M. Neema, M. Absinta, A. Bakshi, M.P. Sormani, M.A. Rocca, R. Bakshi, and M. Filippi. Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: Application in multiple sclerosis. *Neuroimage*, 2010.

S.-H. Huang, Y.-H. Chu, S.-H. Lai, and C.L. Novak. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE TMI*, 28 (10):1595–1605, 2009.

B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec. Shape representation for efficient landmark-based segmentation in 3-d. *TMI*, 2014.

J.G. Jarvik and R.A. Deyo. Moderate versus mediocre: The reliability of spine mr data interpretations. *Radiology*, 250(1):15–17, 2009.

R.K. Jensen, T.S. Jensen, P. Kjaer, and P. Kent. Can pathoanatomical pathways of degeneration in lumbar motion segments be identified by clustering MRI findings. *BMC Musculoskeletal Disorders*, 14(198), 2013.

A.K. Jerebko, G.P. Schmidt, X. Zhou, J. Bi, V. Anand, J. Liu, S. Schoenberg, I. Schmueking, B. Kiefer, and A. Krishnan. Robust parametric modeling approach based on domain knowledge for computer aided detection of vertebrae column metastases in MRI. *Information Processing in Medical Imaging*, 2007.

S. Kadoury, H. Labelle, and N. Paragios. Spine segmentation in medical images using manifold embeddings and higher-order MRFs. *IEEE TMI*, pages 1227 – 1238, 2013.

B.M. Kelm, M. Wels, K.S. Zhou, S. Seifert, M. Suehling, Y. Zheng, and D. Comaniciu. Spine detection in CT and MR using iterated marginal space learning. *Medical Image Analysis*, 2012.

Y. Kim and D. Kim. A fully automatic vertebra segmentation method using 3d deformable fences. *Computerized Medical Imaging and Graphics*, 33:343–352, 2009.

T. Klinder, R. Wolz, C. Lorenz, A. Franz, and J. Ostermann. Spine segmentation using articulated shape models. In *MICCAI*, pages 227–234, 2008.

T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. Automated model-based vertebra detection, identification, and segmentation in ct images. *Medical Image Analysis*, 13(3):471–482, 2009.

J. Koh, V. Chaudhary, and G. Dhillon. Diagnosis of disc herniation based on classifiers and features generated from spine MR images. *SPIE*, 7624, 2010.

J. Koh, R. S. Alomari, V. Chaudhary, and G. Dhillon. Lumbar spinal stenosis CAD from clinical MRM and MRI based on inter- and intra-context features with a two-level classifier. In *SPIE*, volume 7963, 2011.

J. Koh, V. Chaudhary, and G. Dhillon. Disc herniation diagnosis in MRI using a CAD framework and a two-level classifier. *Int J CARS*, 7:861–869, 2012.

J. Koh, V. Chaudhary, E. K. Jeon, and G. Dhillon. Automatic spinal canal detection in lumbar MR images in the sagittal view using dynamic programming. *Computerized Medical Imaging and Graphics*, 2014.

V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2): 147–159, 2004.

S. Koompairojn, K. Hua, K. A. Hua, and J. Srisomboon. Computer-aided diagnosis of lumbar stenosis conditions. 2010.

R. Korez, B. Likar, F. Pernus, and T. Vrtovec. Parametric modeling of the intervertebral disc space in 3d: Application to ct images of the lumbar spine. *Computerized Medical Imaging and Graphics*, 2014.

R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec. Framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. *TMI*, 2015.

M. W. K. Law, K. Tay, A. Leung, G. Garvin, and Shuo Li. Intervertebrald disc segmentation in MR images using anisotropic oriented flux. In *MIA*, 2012.

B.D. Leener, S. Kadoury, and J. Cohen-Adad. Robust, accurate and fast automatic segmentation of the spinal cord. In *NeuroImage*, 2014.

M Lootus, T. Kadir, and A. Zisserman. Vertebrae detection and labelling in lumbar mr images. In *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2013.

M. Lootus, T. Kadir, and A. Zisserman. Automated radiological measurement of spinal MRI. In *MICCAI CSI Workshop*, 2014.

Z. Lu, Q. Zheng, W. Yang, Q. Feng, and W. Chen. Adaptive image segmentation based on local neighborhood information and gaussian weighted chi-square distance. In *ISBI*, 2012.

A. Lucci, K. Smith, R. Achanta, V. Lepetit, and P. Fua. A fully automated approach to segmentation of irregularly shaped cellular structures in em images. In *MICCAI*, 2010.

J. D. Lurie, R.A. Moses, A.N.A. Tosteson, T.D. Tosteson, E.J. Carragee, J.A. Carrino, J.A. Kaiser, and R.J. Herzog. Magnetic resonance imaging predictors of surgical outcome in patients with lumbar intervertebral disc herniation. *Spine*, pages 1216–1225, 2013.

J. Ma and L. Lu. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. *Computer Vision and Image Understanding*, 2013.

J. Ma, L. Lu, Y. Zhan, X.S. Zhou, M. Salganicoff, and A. Krishnan. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. In *MICCAI*, volume 6361, pages 19–27, 2010.

S. Mahmoudi and M. Benjelloun. A new approach for cervical vertebrae segmentation. *Progress in Pattern Recognition, Image Analysis and Applications*, pages 753–762, 2005.

A. Mastmeyer, K. Engelke, C. Fuchs, and W.A. Kalender. A hierarchical 3d segmentation method and the definition of vertebral body coordinate systems for qct of the lumbar spine. *Medical Image Analysis*, 2006.

I.C. Mateos, J.M. Pozo, A. Lazary, and A.F. Frangi. 2d segmentation of intervertebral discs and its degree of degeneration from t2-weighted magnetic resonance images. In *SPIE*, 2014.

C. McIntosh and G. Hamarneh. Spinal crawlers: Deformable organisms for spinal cord segmentation and analysis. In *MICCAI*, volume 4190, pages 808–815, 2006.

S. Michopoulou, L. Costaridou, E. Panagiotopoulos, R. Speller, and A Todd-Pokropek. Segmenting degenerated lumbar intervertebral disks from MR images. In *IEEE Nuclear Science Symposium Conference Record*, pages 4536–4539, 2008.

S. 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. *Transactions on Biomedical Engineering*, 56(9), 2009.

M.T. Modic and J.S. Ross. Lumbar degenerative disc disease. *Radiology*, 245(1), 2007.

M.T. Modic, P.M. Steinberg, J.S. Ross, T.J. Masaryk, and J.R. Carter. Degenerative disc disease: Assessment of changes in vertebral body marrow with MR imaging. *Radiology*, 166:193–199, 1988.

E. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 191–198, 1995.

E. Moschidis and J. Graham. Interactive differential segmentation of the prostate using graph-cuts with a feature detector-based boundary term. In *Medical Image Computing and Analysis*, 2010.

D.S. Mulconrey, R.Q. Knight, J.D. Bramble, S. Paknikar, and P.A. Harty. Interobserver reliability in the interpretation of diagnostic lumbar mri and nuclear imaging. *Spine*, pages 177–184, 2006.

A. Neubert, J. Fripp, C. Engstrom, R. Schwarz, L. Lauer, O. Salvadao, and S. Crozier. Automated detection, 3D segmentation and analysis of high resolution spine MR images using statistical shape models. *Physics in Medicine and Biology*, 57(24), 2012.

A. Neubert, J. Fripp, C. Engstrom, D. Walker, M.-A. Weber, R. Schwarz, and S. Crozier. Three-dimensional morphological and signal intensity features for detection of intervertebral disc degeneration from magnetic resonance images. *J Am Med Inform Assoc*, pages 1082–1090, 2013.

A. Neubert, J. Fripp, C. Engstrom, Y. Gal, S. Crozier, and M.I.C. Kingsley. Validity and reliability of computerized measurement of lumbar intervertebral disc height and volume from magnetic resonance images. *Spine*, 2014.

L.G. Nyúl, J. Kanyó, E. Máté, G. Makay, E. Balogh, M. Fidrich, and A. Kuba. Method for automatically segmenting the spinal cord and canal from 3d ct images. In *CAIP*, pages 453–463, 2005.

A.B. Oktay and Y.S. Akgul. Simultaneous localization of lumbar vertebrae and intervertebral discs with SVM based MRF. *IEEE TMI*, 2013.

A.B. Oktay, N.B. Albayrak, and Y.S. Akgul. Computer aided diagnosis of degenerative intervertebral disc diseases from lumbar MR images. *Computerized Medical Imaging and Graphics*, 2014.

V. Pekar, D. Bystrov, H. S. Heese, S. P. M. Dries, S. Schmidt, R. Grewer, C.J.d. Harder, R.C. Bergmans, A.W. Simonetti, and A.M.v. Muiswinkel. Automated planning of scan geometries in spine MRI scans. In *MICCAI*, 2007.

Z. Peng, J. Zhong, W. Wee, and J.-H. Lee. Automated vertebra detection and segmentation from the whole spine mr images. In *Engineering in Medicine and Biology*, 2005.

C. W. A. Pfirrmann, A. Metzdorf, M. Zanetti, J. Hodler, and N. Boos. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*, 26, 2001.

V. Potesil, M. Lootus, A. El-Labban, and T. Kadir. Landmark localization in images with variable field of view. In *ISBI*, 2013.

M.G. Roberts, T.F. Cootes, and J.E. Adams. Automatic segmentation of lumbar vertebrae on digitised radiographs using linked active appearance models. *MICCAI*, 2006.

C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 23(3):309–314, 2004. ISSN 0730-0301. doi: http://doi.acm.org/10.1145/1015706.1015720.

S. Seifert, I. Wächter, G. Schmelzle, and R. Dillmann. A knowledge-based approach to soft tissue reconstruction of the cervical spine. *IEEE TMI*, 28(4), 2009.

H. Shen, A. Litvin, and C. Alvino. Localized priors for the precise segmentation of individual vertebras from ct volume data. In *MICCAI*, pages 367–375, 2008.

J. Shen, S. Parent, and S. Kadoury. Classification of spinal deformities using a parametric torsion estimator. In *MICCAI CSI Workshop*, 2013.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

R. Shi, D. Sun, Z. Qiu, and K.L. Weiss. An efficient method for segmentation of mri spine images. In *International Conference on Complex Medical Engineering*, 2007.

A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 2004.

D. Steffens, M.J. Hancock, C.G. Maher, C. Williams, T.S. Jensen, and J. Latimer. Does magnetic resonance imaging predict future low back pain? a systematic review. *European Journal of Pain*, 2013.

D. Stern, B. Likar, F. Pernus, and T. Vrtovec. Automated detection of spinal centrelines, vertebral bodies and intervertebral discs in CT and MR images of lumbar spine. *Physics in Medicine and Biology*, 55:247–264, 2010.

D. Štern, B. Likar, F. Pernus, and T. Vrtovec. Parametric modelling and segmentation of vertebral bodies in 3D CT and MR images. *Phys. Med. Biol.*, 2011.

D. Štern, V. Njagulj, B. Likar, F. Pernus, and T. Vrtovec. Quantitative vertebral morphometry based on parametric modeling of vertebral bodies in 3D. *Osteoporos Int*, 24:1357–1368, 2013.

S. Tan, J. Yao, M.M. Ward, L. Yao, and R.M. Summers. Computer aided evaluation of ankylosing spondylitis. In *ISBI*, 2006.

M.-T. Tsai, S.-B. Jou, and M.-S. Hsieh. A new method for lumbar herniated intervertebral disc diagnosis based on image analysis of transverese sections. In *CMIG*, 2002.

Y. Unal, H. E. Kocer, and H. E. Akkurt. A comparison of feature extraction techniques for diagnosis of lumbar intervertebral degenerative disc disease. *INISTA*, 2011.

T. Videman, M.C. Battie, L.E. Gibbons, K. Maravilla, H. Manninen, and J. Kaprio. Associations between back pain history and lumbar MRI findings. *Spine*, 28(6): 582–588, 2003.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum

decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, April 1967.

T. Vrtovec, B. Likar, and F. Pernuš. Automated curved planar reformation of 3d spine images. *Physics in Medicine and Biology*, 50:4527–4540, 2005.

T. Vrtovec, S. Ourselin, L. Gomes, B. Likar, and F. Pernuš. Automated generation of curved planar reformations from mr images of the spine. *PHYSICS IN MEDICINE AND BIOLOGY*, 52:2865–2878, 2007.

Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li. A unified segmentation framework for m3 spinal images. *TMI*, 2014.

D. Weishaupt, M. Zanetti, N. Boos, and J. Hodler. MR imaging and CT in osteoarthritis of the lumbar facet joints. *Skeletal Radiology*, 28:215–219, 1999.

B.N.W. Weissmann. *Imaging of Arthritis and Metabolic Bone Disease*. Mosby Elsevier, 2009.

M. Wels, B.M. Kelm, A. Tsymbal, M. Hammon, G. Soza, M. Sühling, A. Cavallaro, and D. Comaniciu. Multi-stage osteolytic spinal bone lesion detection from ct data with internal sensitivity control. In *SPIE*, 2012.

A. Wong, A. Mishra, J. Yates, P. Fieguth, D.A. Clausi, and J.P. Callaghan. Intervertebral disc segmentation and volumetric reconstruction from peripheral quantitative computed tomography imaging. In *IEEE Transactions on Biomedical Engineering*, volume 56, 2009.

J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *ICMI*, 2013.

T. Wu, B. Jian, and X.S. Zhou. Automated identification of thoracolumbar vertebrae using orthogonal matching pursuit. In *MLMI*, pages 126–133, 2011.

C. Xu and J.L. Prince. *Handbook of medical imaging*, pages 159–169. Academic Press, Baltimore, 2000.

Y. Xu, X. Gao, S. Lin, D. W. K. Wong, J. Liu, D. Xu, C.-Y. Cheng, C. Y. Cheung, and T. Y. Wong. Automatic grading of nuclear cataracts from slit-lamp lens images using group sparsity regression. In *MICCAI*, 2013.

J. Yao, H. Munoz, J.E. Burns, L. Lu, and R. Summers. Computer aided detection of spinal degenerative osteophytes on sodium fluoride pet/ct. In *MICCAI CSI Workshop*, 2013.

Y. Zhan, D. Maneesh, M. Harder, and X.S. Zhou. Robust MR spine detection using hierarchical learning and local articulated model. In *MICCAI*, 2012.