Radiological Grading of Spinal MRI

Meelis Lootus¹, Timor Kadir², and Andrew Zisserman¹

Oxford University¹,
Mirada Medical²
meelis@robots.ox.ac.uk
timork@mirada-medical.com
az@robots.ox.ac.uk
http://www.robots.ox.ac.uk/~vgg/

Abstract. This paper describes a fully automatic system for obtaining the standard Pfirrmann degeneration grading of individual intervertebral spinal discs in T2 MRI scans. It involves detecting and labelling all the vertebrae in the scan and then learning a regression from the disc region to the grading.

In developing the regression function we investigate a spectrum of support regions which involve differing degrees of segmentation of the scan: our intention is to ascertain to what extent segmentation is necessary or detrimental in obtaining robust and accurate measurements.

The methods are assessed on a heterogeneous clinical dataset containing 1710 Pfirrmann-graded discs, from 285 symptomatic back pain patients. We are able to predict the grade to ±1 precision at 85.8% accuracy. Our novel method proposes new image features that outperform previous features and utilizes techniques to improve robustness to MR imaging variations.

1 Introduction

Our primary goal in this paper is to automate radiological measurements in multi-slice clinical Magnetic Resonance Imaging (MRI) spinal scans, and to this end we describe a system to extract the standard clinical Pfirrmann disc grading that is used in the diagnosis and management of back pain patients, exploring accuracy and robustness in the process. The task is defined in Figure 1.

Our secondary goal is to investigate the need for a complete segmentation of the disc in order to accomplish this task. On the one hand, voxel wise segmentation can help better define the grading task. On the other hand, anatomical units (discs and vertebrae in this case) may be inseparable due to pathological changes, rendering the task ill-defined. Also, in practice, segmentation is often prone to failure so avoiding it can possibly improve the overall results. Thus, we ask the question: to segment or not to segment?

We answer the question by formulating the task as one of regressing between an image support region and the Pfirrmann disc grading, and then investigate a spectrum of ways of obtaining the support region which cover: no segmentation, segmentation of only the vertebrae, and finally segmentation of the disc.
Fig. 1: **The task:** Given a clinical MRI volume of the lumbar spine (A) as input, fully automatically localize, label, and radiologically measure (B) the six lumbar discs in that volume, according to a standard radiological grading system (C) [1]. The radiological measurement is the ‘degeneration grade’ describing drying out of the disc (darkening in T2 MRI), and disc space collapse, in terms of four radiological features as defined in the main text. Note that in evaluation, the grade is considered correct if predicted to ±1 accuracy, due to the ground truth intra-observer variability.

The method is evaluated over a large heterogeneous clinical dataset, and this adds to the challenge of the task since T2 images of the same anatomy and pathology look different in different MRI machines and under different protocols (different “tissue contrasts”). We introduce a normalization scheme to address this problem. The task is also challenging because MR imaging artifacts can be confused with pathology.

**Background.** The normal intervertebral disc is composed of a soft liquid central part, the nucleus pulposus (NP), and a hard ligamentous surrounding, the annulus fibrosus. It is interfaced to the vertebral bodies above and below by cartilaginous endplates. The disc acts as both a cushion and pivot point in the spine. Disc problems are a common cause of back pain.

One common disc problem is *Degeneration*, the drying out and collapse of the disc space, and this abnormality is clinically measured using the standard radiological Pfirrmann grade [1], illustrated in Figure 1. Pfirrmann defines the categorical five-score grade in terms of sequential changes to four MRI features:
brightness of the NP, uniformity of the NP, distinction between the NP and the annulus fibrosus, and the disc height.

Since our ground-truth labelling is not perfect – the intra-observer grade agreement in our database is only 71% based on grading 121 patients twice, while agreement to ±1 is 98% – we assess our scores to within ±1 grade accuracy. Note, Pfirrmann [1] measured 88-92% intra-observer agreement over measurements in a single day. Our database was annotated by one radiologist over several years – achieving similar variability to inter-observer variability witnessed by Pfirrmann.

While the Pfirrmann grade is widely used in clinical practice to assess the overall disc quality, conflicting accounts have been presented in research studies regarding the correlation between the grade, back pain, and surgical outcomes [2–4].

Given the quite high intra- and inter-observer variability of radiological measurements, one advantage of automating measurement is that it should lead to an improvement in consistency. In turn, this consistency may lead to improvements in both clinical research studies on the correlation of back pain with radiological measurements, and communication between radiologists.

1.1 Related Work

Recently, multiple medical imaging papers have been published attempting to automatically diagnose a number of spinal conditions [5–9].

The existing publications on Disc Degeneration deal with the binary classification task – e.g. the presence or absence of desiccation/degeneration [5, 6, 8, 7] – rather than measuring the standard radiological Pfirrmann grade or a similar radiological quantity. In addition, they are generally restricted to homogeneous data collected from the same scanner, using the same [5, 6], or a relatively narrow range [7] of protocols.

Image features for Pfirrmann grading have been proposed before [10], however their computation is not fully automatic and they have not been used to automatically measure the grade. Alomari et al. [5] and Neubert et al. [7] automatically predicted a binary grading to high accuracy in MRI scans over a homogeneous dataset; and we compare to their features here. Often, the methods require a segmentation step [7] to accurately delineate the discs, or to find an exact square in the disc [5]. The robustness of the methods to segmentation has not been explicitly studied, however it is an important point.

A number of vertebra and disc segmentation algorithms have been proposed [11–14]. Vertebra segmentation methods have been more successful than disc segmentation ones. This is largely because vertebrae in MRI have well-defined edges and consistent appearances across patients. In contrast, discs have variable appearance, lack clear boundaries, and vary considerably due to degradation (the very process we are assessing).

1.2 Contributions & Paper Layout

Contributions. We fully automatically measure the Pfirrmann grade to ±1 accuracy, and are the first to present results generalizing across clinical data
from a number of sites, scanner types & imaging sequences. We train and assess our results based on ground truth annotations on clinical data, by an expert radiologist with 25 years of experience. We experimentally assess the effect of varying the amount of segmentation to define the feature support region, and compare our features to those proposed in previous work on binary auto-grading [5, 7].

**Paper layout.** The method is presented in full in Section 2, explaining in detail the three steps of the pipeline: feature support region definition, image feature extraction, and regression. The dataset is described in Section 3.1, the evaluation protocol in Section 3.2, and the experimental results and discussion in Section 3.3.

## 2 Regression of Radiological Measures

In this section, we describe a method for predicting the Pfirrmann grading. It is formulated as a regression task, and we use standard machine learning methods to learn the regressor from expert-annotated training scans, and then apply the regressor to previously unseen clinical scans.

Although Pfirrmann grading is categorical, the underlying fundamental disc degeneration process is continuous over time, and that is why we choose to formulate it as regression rather than classification.

The pipeline from a raw multi-slice MRI scan to radiological measurement of disc degeneration has three steps, described in more detail in the following sections: first, finding the support region; second, extracting image features; third, predicting the Pfirrmann grade. In Section 2.1, we explore three alternative methods of obtaining the feature support region, named **V-det** (vertebra detection), **V-seg** (vertebra segmentation), and **D-seg** (disc segmentation), illustrated in Figure 2. This explores three points on the ‘no segmentation’ to ‘full disc segmentation’ spectrum.

### 2.1 Step 1: Three Alternative Support Regions

Taking a clinical MRI scan as input, this step outputs the feature support region for the six lumbar discs (annotated in Figure 1) in three different ways, as contrasted in Figure 2. 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 initializing the next. Both the vertebrae and the discs are segmented using the standard graph cuts algorithm of Boykov-Jolly [15] with region and boundary terms. 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.
Fig. 2: **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, and the Obj. (object) and Bkg. (background) seeds show the initializations used for the segmentations as the red and blue areas. 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 2.1 for more detail.

**Implementation details.** The detections are performed in all slices using the method of Lootus *et al.* [16], picking – for each vertebra $v_i$ – the tight labelled bounding box $B(v_i)$ from the slice with the most confident detector output. This way, the system is robust to scoliotic spines where not all lumbar vertebrae may be clearly visible in the same sagittal slice. Note that the labelling starts from the sacrum and is performed in the same manner and is consistent with the ground truth labelling process. Therefore any labelling errors due to the natural variability in vertebra number will not affect the results presented here, but will need to be considered in the future work. The segmentations and gradings are performed in the mid-sagittal slice (as clinical standard), initializing the support region selection based on the automatically labelled bounding boxes. The segmentation of the vertebrae T12 ($v_1$) to S1 ($v_7$) is initialized automatically for each $v_i$ independently, by placing object (Obj.) and background (Bkg.) seeds according to the bounding boxes $B(v_{i-1})$, $B(v_i)$, and $B(v_{i+1})$. The Obj. seed for $v_i$ is placed in an area obtained by eroding $B(v_i)$ to half its size. The two Bkg. seeds are placed as rectangles 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 disc segmentation is performed for discs T12/L1 to L5/S1. For each disc, it is automatically initialized by seeding the foreground as the space between, and background as the area of the neighboring vertebra segmentations. 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 and disc segmentations. Three components were picked as best performing at earlier experiments. To arrive at the final box regions for V-det, the box is placed at the angle of its lower neighboring 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 grid search). In V-seg, the vertical ‘walls’ of the region are placed at the same positions as for the V-det case.

2.2 Step 2: Image Features

To characterize the disc, a number of intensity and shape features are extracted from the feature support region, as described below. As mentioned in the introduction, one of the challenges is the variation of MRI contrast across different protocols. Therefore, before extraction, the image is normalized to 1.25 times the median vertebral intensity as found from the bounding boxes (for V-det) or the segmented vertebrae (for V-seg/D-seg). In early experiments, we tested a number of intensity normalization techniques including normalizing to CSF and found this approach performed best and provided sufficient robustness to variations in protocol and MRI scanners. The conventional CSF-based normalization [17] also suffers from inconsistent CSF signal due to various deformations in the dural sac. After normalization 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 vertebral intensity (e.g. for grade 1 or 2 discs), unlike in the case if the vertebra intensity was normalized to one.

**Intensity features.** Two groups of intensity features are extracted: first, a histogram normalized 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 intensity features: Disc Mean Intensity [5] and a GMM fitted to the support region [7].
2.3 Step 3: Regression

A linear epsilon Support Vector Regressor [18] is learned to map the 26-dimensional feature vector \( \mathbf{x} \) described above to the grading as:

\[
f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b
\]  

(1)

where \( f \) is the predicted (continuous) disc grade, and the vector \( \mathbf{w} \) 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 \( \epsilon \) are tolerated.

**Implementation details.** The values of \( \epsilon \) and \( C \) for the Support Vector Regression cost function are learnt by a grid search on a hold-out validation set, with a range of 0.01 to 1.0, and 0.1 to 10000 respectively, using the LIBSVM package [19].

3 Evaluation and Comparison

3.1 Dataset

We evaluate our algorithm using T2 sagittal MRI scans from a large, clinical, heterogeneous dataset of 1710 radiologically annotated lumbar discs in 285 symptomatic back pain patients, exhibiting all the challenges highlighted in the introduction including degeneration, herniation and scoliosis. Of the patients, 116 were male, 151 female (30 unknown), with ages 10-88 years. The patients were split into a 114-scan training (684 discs), 57-scan validation (342 discs), and 114-scan (684 discs) testing set.

In contrast to previous work where images were acquired using the same scanner and protocol, our database includes scans from 25 different sites acquired using a wide gamut of T2 sequences, fields-of-view, and 14 different scanner models. The scan parameters ranged as follows: magnetic field 0.6-3T, TE 69-139ms, TR 1180-1210ms, voxel size 0.34-1.68mm, slice spacing 3.85-15mm.

Each patient had each of their six lumbar discs (T12-L1, L1-L2, L2-L3, L3-L4, L4-L5, and L5-S) Pfirrmann-graded by an expert spine radiologist with 25 years of clinical experience. There were 550 grade 1, 194 grade 2, 379 grade 3, 379 grade 4, and 208 grade 5 discs in the dataset.

3.2 Evaluation Protocol

We assess vertebrae detection accuracy by distance between the detected bounding box centre from ground truth bounding box centre in millimeters, and segmentation accuracy according to an overlap measure – the ratio of intersection to union between manually segmented vertebrae/discs, and the automatic segmentation results.

We assess grading accuracy by measuring the proportion of discs that are graded correctly, as proposed by [20] for eye cataract grading. However, as discussed in the introduction, our ground-truth labelling is not perfect due to
### Table 3.1: Numerical results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean Int.</th>
<th>GMM</th>
<th>Hist+</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-det</td>
<td>73.4%</td>
<td>77.5%</td>
<td>79.7%</td>
</tr>
<tr>
<td>V-seg</td>
<td>81.1%</td>
<td>84.6%</td>
<td>85.2%</td>
</tr>
<tr>
<td>D-seg</td>
<td>79.6%</td>
<td>83.3%</td>
<td>85.8%</td>
</tr>
</tbody>
</table>

Fig. 3: **Numerical results.** Percentage of discs with Pfirrmann grade predicted to ±1 accuracy using each of the features (columns) on each of the support regions (rows).

intra-observer variability, and thus we evaluate our regression performance as the fraction of scores which are predicted to ±1 accuracy.

### 3.3 Results and Discussion

A summary of the overall results and the D-seg confusion matrix is presented in Figure 3, with example grading result shown for two patients in Figure 4, and for two more in Figure 5. The median detection error was 2.0mm; the mean vertebrae segmentation overlap measure was $0.808 \pm 0.132$ 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, there were 3.5% discs with detection/segmentation failure, 1.5% with low imaging quality, 1.3% with failed slice selections. This sums to 6.3% which is 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 minimizes 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.

We investigated the role of segmentation in our processing pipeline by testing three different approaches to defining the feature support region – vertebra detection only, vertebra segmentation and direct disc segmentation. This spectrum represents a trade-off between specificity of the feature support region and the
likelihood of failure in that step. In principle, though subject to failure, a support region that encompasses only the object of primary interest, i.e. the disc, will outperform more generic support regions.

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.

4 Conclusion

We have presented the first fully automated system to predict all five Pfirrmann grades. In a large dataset of 1710 discs from 285 patients using standard clinical MRI scans acquired on different scanners, protocols and sites, our system correctly predicted to within ±1 Pfirrmann grade in 85.8% of discs. Our novel method proposes new features that outperform previous ones, improves robustness to MR imaging variations, and shows that disc segmentation is not essential for for this level of performance.

References

Fig. 4: **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. 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.
Fig. 5: **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. 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.