Automatic Detection of the Lumbar Spine using Transfer Learning for Quantitative Fluoroscopy

Rantilini Samaratunga¹, Marcin Budka¹, Alexander Breen¹, and Alan Breen¹

¹Affiliation not available

March 04, 2024

Abstract

Quantitative assessment of spinal motion plays a pivotal role in diagnosing and understanding lower back pain. This paper utilises a Convolutional Neural Network for precise landmark localisation of bounding boxes encompassing the lumbar spine in sagittal plane lumbar fluoroscopy image sequences. The proposed methodology aims to automate spinal movement tracking and provide a benchmark for future research, thereby enhancing the efficiency and accuracy of low back pain diagnosis.
Automatic Detection of the Lumbar Spine using Transfer Learning for Quantitative Fluoroscopy

Rantilini Samaratunga, Marcin Budka, Alexander Breen and Alan Breen

Abstract—Quantitative assessment of spinal motion plays a pivotal role in diagnosing and understanding lower back pain. This paper utilises a Convolutional Neural Network for precise landmark localisation of bounding boxes encompassing the lumbar spine in sagittal plane lumbar fluoroscopy image sequences. The proposed methodology aims to automate spinal movement tracking and provide a benchmark for future research, thereby enhancing the efficiency and accuracy of low back pain diagnosis.

Index Terms—Deep-Learning, Medical Imaging, Lumbar Spine, Low Back Pain, Fluoroscopy.

I. INTRODUCTION

Low back pain (LBP) stands as the leading contributor to the number of days lost to disability worldwide, with much of this burden arising from neglected pain that was allowed to become chronic and debilitating [1]. Quantitative Fluoroscopy (QF) is the gold standard for the measurement of segmental biomechanics for LBP [26]. It has been shown to be highly repeatable for vertebral translation measurements [22] with proven intra patient repeatability [23]. At present, the world’s first and largest collection of healthy ‘normative’ lumbar spine QF sequences, was amassed by the authors during prior work, and was gathered using custom-built motorised bed equipment for deliberate and controlled bending of the the spine at precise and measureable angles [24].

Fluoroscopy is a low-dose imaging technique with the ability to capture images at high frame rates, making it ideal for the assessment of spine kinematics [21]. However, the low ionizing radiation dose and high temporal resolution comes at a cost to image quality. The reduction in signal-to-noise ratio means that the current vertebra tracking technique [24] is only reliable in high-quality imaged instances of bone, which excludes many older, obese, osteopenotic or scoliotic patients. Even amongst a normative dataset, spinal detection sensitivity is heavily dependant on the patient. Fig. 1 illustrates the effect on spinal visibility between participants (in the QF dataset) with relatively low BMI compared to one with higher BMI. Although both these BMI values are within the healthy range for adults, the vertebral edge contrast and overall visibility is severely reduced at higher BMI. Since LBP has shown to be related to age and high BMI, and can be used as an early indicator for more severe diseases [2], automated techniques should aim for robustness across the extremes of these factors.

To be fit for clinical use, the tracking technique is required not only to be invariant to this compounded image quality degradation, but also be time-efficient. At present the existing QF tracking takes approximately two hours per participant’s QF sequence, and requires the first frame to be annotated manually by a human observer [20]. This is not uncommon for brute-force matching techniques, where often, a template image is used and its features extracted and searched for, in the target image; a very tedious and time-consuming process. Prior literature has quoted similarly slow times for segmenting the lumbar spine in fluoroscopy [6], [7]. Reducing this time means faster processing of sequences and hence shorter turnaround for patient diagnoses. This would also mean a more cost-effective solution compared to using the current approach.

A. Deep Learning and Medical Imaging

The advent of Artificial Intelligence (AI) has led to much innovation in the field of Computer Vision (CV) in both the 2D and 3D imaging domains, often matching or even exceeding human-level accuracy.

CV techniques can be broken down into a variety of sub-domains such as object detection, image segmentation, and reconstruction to name a few. Object detection refers to the identification (classification) and localisation of specific objects within an image often by drawing bounding boxes around them. The process of determining the corners of these bounding-boxes is referred to as landmark localisation.

Convolutional Neural Network (CNN), a deep-learning architecture, is commonly used to solve such localisation tasks. Where, previously, brute-force techniques required hard-coded information on features to match, CNNs could now automatically and hierarchically learn features and patterns from the input image dataset. In general, the deeper the network, the more expressive power it has. However, the nature of calculation of weights for each layer in the network (through back propagation) meant that, as the network deepened, the gradient at which the weights were updated reduced (slowing down the learning) to the point where the network would eventually stop learning altogether.
The break-through came in the form of Residual Network (ResNet) architectures. ResNets were first proposed by He et al. in 2015 to solve this problem (known as the vanishing gradient problem) by utilising skip connections allowing the training of very deep (up to 152 layers) networks and making them a very attractive option for CV [14].

Since then, there has been a surge of innovation in this field, often building on the basic ResNet architecture, which has become a fundamental component in deep learning. However, training CNNs is still time-consuming and researchers have constantly been looking at ways to accelerate the process. This is where transfer learning proves to be a key technique.

Transfer learning is a method where models that have been pre-trained on other datasets (such as ImageNet [19]), are used as backbones and fine-tuned on the desired dataset. This is a very popular method, particularly in medical imaging [11] as, depending on the dataset size and imaging modality [16], it offers better performance, requires less data and is faster than training from scratch.

In spinal studies specifically, much work has gone into deploying more complex CNNs for automatic segmentation and reconstruction of the lumbar spine in CT and MRI imaging [5]. However, equivalent research in 2D imaging modalities, specifically X-ray imaging is relatively sparse [3], [4], [9] and studies in using AI in standard fluoroscopy are mainly targeted at the cervical spine [8], [10]. However, landmark localisation in cervical studies are not directly comparable to the lumbar spine task since differences in occlusions such as bowel gas obstructions are unique to the lumbar region. Therefore, further exploration into utilising CNNs in spinal fluoroscopy is not just necessary, but long overdue.

B. Aims

The QF dataset is the first of its kind, and therefore, no prior model performance bench-marking exists. This work will explore the use of ResNets as a starting point for benchmarking. We aim to take advantage of transfer learning to reduce training time. Through the performance analysis of these models, we hope to recommend the best direction for further research in vertebral motion in fluoroscopic images.

II. MATERIALS AND METHODS

A. Data

The dataset was manually annotated by a group of human observers. The sequences were imaged at 15fps. Dataset was anonymised and participants given a unique alphanumeric code. This data involved a repeatability study which meant that patients were imaged twice 6 weeks apart. Detailed participant demographic inclusion/exclusion criteria can be found in [23].

1) Inclusion Criteria: This research had access to data from 166 participants’ sagittal plane QF sequences. After excluding sequences which did not have a matching set of gold standard tracking points, and excluding those of different movement types (recumbent flexion REC_FLEX, recumbent extension REC_EXT and weight-bearing extension WB_EXT, 139 sagittal plane QF sequences of weight bearing flexion WB_FLEX movement were found to be suitable for training (Fig. 2).

2) Pre-processing: Each patient had raw DICOM image sequences of resolution 1024x1024 with an average number of frames of 330. Each raw DICOM file was converted to separate .jpeg files of 1024x1024 and 512x512 images with a compression quality of 99% with two types of interpolation, nearest neighbours (nn) and bi-linear (bn). The DICOM files contained four different look-up-table (LUTs) to scale the

![Fig. 1. Example High and Low BMI images, showing effect of contrast on vertebra](image)

![Fig. 2. Data Inclusion and Exclusion process with criterion error codes](image)
raw values of the pixels. Images were converted for these. The most common 2D image CNN architectures take 3-channel RGB images as inputs. Since this dataset contained grayscale images, for each resolution, images were composed with the first channel as LUT 1 with interpolation mn, the second channel with LUT 2 nn, and the third channel as LUT 1 interpolation ln. This selection was random. The corresponding gold standard template points for each DICOM sequence were contained in a MATLAB file. This file was flattened to give a master .csv file of all templates with bounding box in anti-clockwise direction: L2_TL_X, L2_TL_Y ... L5_TR_X, L5_TR_Y. Note that in our case each bounding box is described by four corner points, unlike typical object detection workflows.

As there may be more than four vertebrae visible in each image, there is ambiguity as to which four vertebrae are of interest. Therefore the top two points of the sacrum (SAC) were also included (SAC_TL, SAC_TR) to assist with anchoring the points. Therefore, a total of 18 x,y coordinate pairs were used.

3) Augmentations: To bolster robustness of the model, augmentations were applied to the data using fastai in-built data augmentations [18]. These included random brightness, contrast, rotation, scale, blur. However, to simulate occlusions that occur in this region of the body, a custom augmentation function was implemented. Prior literature often used a cropping technique involving replacing certain regions of the image with pixels of a uniform colour (usually black), to simulate non-organic items (such as implants and items in patients’ pockets). However, in the lumbar region of the body, the most common type of obstructions are bowel gas and organs. These often don’t show up as such ultra-high contrast regions. Therefore, the custom ‘obstructify’ function used the average pixel value of a region around a random vertebra bounding box corner as the occlusion function. The function was set to have a 50% probability of occurring in the dataset, and when active, had a 30% probability of obscuring the box points.

4) Stratification: The data was split at random with a 80%-10%-10% train-val-test split. However, image quality is heavily dependent on the individual being imaged. For example, a high BMI individual is likely to have their vertebra appear at a lower contrast than someone with a lower BMI. Therefore, to ensure that training, validation and test data was matched, participants were categorised by age range, gender and BMI range. The training validation and test sets were stratified, to ensure at least one individual from each age/BMI range and gender group was present in all data sets as shown in Fig. 4.

B. Neural Network Architecture

1) Models: CNN Model selection was based on what prior literature utilised - ResNet50, Resnet101 and Resnet152. To compare the mature ResNet model, a more modern architecture was required. ConvNeXt was an ideal candidate, as it is a morphed architecture with a base ResNet configuration that was shown to meet (and surpass) the performance of vision transformers [28], yet having all the qualities of CNNs which makes them straightforward to implement [25]. For each of these models, the fully connected (fc) layer was modified to have 36 outputs (for the 18 bounding-box points).

2) Variant Configurations: To compare effects of common modifications to the CNN architectures and the effects of different loss functions, and pooling layers, four main configurations for each architecture were used: (1) base architecture with a Root Mean Squarer Error loss (rmse), (2) base architecture with a Weighted Mean squared Error loss (rmse-w) with emphasis on the two SAC keypoints to anchor the predictions, (3) Active Shape Model [27] layer on top of the base architecture with rmse loss, and (4) Active Shape Model layer but with rmse-w loss. For each of these four configurations, 3 types of pooling layers were used: (1) global average pooling, (2) 1x1 attention-weighted average pooling, and (3) 2x2 attention-weighted average pooling (see Table I). Each component of the variant configurations is described in detail in the following sections.

3) Loss Function: In our initial approach, each point of the vertebrae bounding boxes was treated independently, without any form of anatomical relationship between the vertebrae.
<table>
<thead>
<tr>
<th>MODEL</th>
<th>LOSS</th>
<th>POOLING</th>
</tr>
</thead>
<tbody>
<tr>
<td>rn50</td>
<td>rmse</td>
<td>adapt</td>
</tr>
<tr>
<td>m101</td>
<td>rmse-w</td>
<td>atn2x2</td>
</tr>
<tr>
<td>m152</td>
<td>shape</td>
<td>atn1x1</td>
</tr>
<tr>
<td>convnext</td>
<td>shape-w</td>
<td></td>
</tr>
</tbody>
</table>

Therefore, the simple rmse loss function was used. After initial experimentation, it was found that, since some images had more than four vertebra visible, there was confusion as to which four vertebrae correspond to L2-L5. This is seen in Fig. 5 where the yellow points depict the ground truth and the blue points are the predictions. Therefore, using SAC_TL and SAC_TR the rmse function was modified to accommodate weights (rmse-w, Eq. (1)) and the effect of the weighting can be seen in Fig. 6.

\[
\text{rmse-w} = \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]}{\sum_{i=1}^{n} w_i}}
\]

where

- \((x_i, y_i)\) are the ground truth x and y coordinates for the \(i\)-th bounding-box corner point in the image.
- \((\hat{x}_i, \hat{y}_i)\) are the predicted x and y coordinates for the \(i\)-th bounding-box corner point in the image.
- \(w_i\) is the weight assigned to the \(i\)-th bounding-box
- \(n\) is the total number of vertebra bounding-box corner points in the image.

The loss was weighted in such a way that the Euclidean distance between L5’s and the SAC’s ground-truth and predicted points contributed 5 and 10 times more respectively, to the overall loss than other bounding-box points. The weight values were selected manually through trial and error to ensure that predicted points were correctly anchored.

4) Active Shape Model Layer: To add more anatomical context to the individual points, the Active Shape Model layer was considered. Shape Models take into account the relationship between points by treating them as a fully connected whole entity. It was hypothesised to help with predictions where occlusions occurred in the image, since the shape model need not require all points to be present. A shape model layer was constructed that took the individual bounding box points which were transformed to a shape model with 36 different components. Each component corresponded to a modification of the base shape (the mean shape of all the training data in the WB_FLEX dataset) in different modes, such as curvature, distance between vertebrae (horizontally and vertically) etc. The CNN was then trained to predict the \(x,y\) coordinates via predicting weightings of these components.

5) Pooling: Pooling in CNNs is an operation performed to reduce computations and improve translation invariance, and refers to the stage of condensing the spatial information of the input image to a down-sampled representation, whilst attempting to retain the most pertinent features for the task at hand. Common types are max pooling and average pooling which use a fixed window size and when applied, the maximum activation value within this window or the average of all activations are taken, respectively. Adaptive Pooling allows for a resize-able window depending on the input resolution of the image. Therefore, Adaptive Max and Adaptive Average Pooling are better at handling images of different resolutions. When applied in a way resulting in 1x1 output volume,
these are equivalent to Global Adaptive Pooling (GAP) [30]. Attention Pooling is a more sophisticated technique that allows for weighted averaging, where the weights are derived from normalised attention scores. Attention scores are often the cosine similarity between the element of interest (Query) and overall/neighbouring elements (Key). A SoftMax function is then applied to obtain (normalized) attention weights [29].

To investigate the effect of different pooling methods on the models’ performance, three variations were used. Adaptive pooling, Attention Pooling 2x2, Attention Pooling 1x1. The change in window size in the attention pooling variants was hypothesised to affect granularity of the calculated scores and therefore affect overall accuracy of the model.

C. Training

Training was split into three stages. The first stage (STAGE00) used ResNet 50 (rn50), ResNet 101 (rn101) and ConvNeXt with 512x512 images, with all variations to loss functions, shape model layers, pooling layers and with and without the custom obstruction function. It was hypothesised that the performance of a larger ResNet should yield lower loss values, at increased computational cost, but would be outperformed by the newer ConvNeXt architecture. A batch size of 64 was used.

Based on the results of these, the second stage (STAGE01) took the top five val loss and top five test loss variants with obstructions and top five val loss and top five test loss variants without obstructions. These were then trained on 1024x1024 resolution images, with a batch size of 32. The final stage (STAGE02) took the top three best models for val loss and top three best variants for test loss for obstruction and not obstructed and applied them to ResNet 152 (rn152) with images of 1024x1024 resolution and batch size 24.

Batch sizes were swept until a value was found that fit on the GPU with the largest model for each training stage. Smaller models (whom otherwise could be run with much larger batch sizes) were then run with this value so that they could be fairly compared to the large models. To avoid over-fitting, the model was trained with data augmentations to diversify the dataset. All models were trained for 20 epochs. For the first two epochs, all bar the last two layers (which were randomly initialised and had not been pre-trained on ImageNet) were frozen to aid in fine-tuning and reduce computation time. Each variant was trained three times to ensure repeatability and it is the means of these, that are reported.

D. Inference

Inference time was measured for one QF sequence of 351 frames (the maximum seen in this dataset). The batch size was swept to find the maximum allowable size for each variant and image resolution. Inference was conducted for the best models for all stages (repeated 3 times). Inference times are reported as mean times in seconds.

III. RESULTS

For fair comparison all loss results were reported in pixels with respect to 512x512 resolution. Any loss values for variants using the $\text{rmse-w}$ loss function underwent inference using the (unweighted) $\text{rmse}$ function to ensure that that their validation and test losses could be compared with those variants that used unweighted loss functions during training.

A Shapiro-Wilks test for normality was applied to the test losses for each stage. To quantify statistically significant differences between groups, either one-way ANOVA (when distribution was normal) or Kruskal–Wallis test was conducted. Based on the outcomes of these tests, whenever possible, a paired t-test was then performed. Otherwise, its non-parametric equivalent (Wilcoxon signed rank) was used when the dataset did not follow a normal distribution.
Fig. 9. Effect of Different CNN Model Architectures on Test Loss (px) in STAGE00

A. STAGE00

To identify generalised effects of each variant, pooling layers, loss functions and CNN models were varied individually. Pooling layer changes were found to have no statistically significant effect on test losses (p=0.88) when taken across all model types, loss functions, and with and without occlusions. Varying the loss function however had a significant effect on test loss, with the rmse-w + shape model producing the highest losses. The rmse, rmse-w and shape (without rmse-w) variations had significant effects on test losses (Fig. 8).

When model architectures were varied, it was found that although there was no statistical significance (p=0.099) seen when they operated on un-obstructed data, there was a significant performance improvement with rn101 on obstructed data compared to rn50 and convnext (p=0.0004) test losses with a decrease in mean loss by 1.22 px. Overall, both ResNet architectures performed better than ConvNext as shown in Fig. 9. Therefore, all ConvNext variants were eliminated as candidates for STAGE01. The top five architectures consisted mainly of rmse and rmse-w loss functions and continued on to STAGE01 training based on their validation losses. The overall best model in STAGE00 across obstructed and unobstructed experiments was rn50 with rmse-w loss and atn2x2 pooling achieving validation loss of 3.94 ± 0.12 px (Table II).

B. STAGE01

The effect of varying pooling layers for STAGE01, like STAGE00, was insignificant on both validation and test losses. However, rn50 was clearly outperformed by its larger version, with rn101 also showing better robustness to occlusions (Fig. 11). The best configurations for each model in STAGE01 are summarised in Table III.

C. STAGE02

Unlike STAGE01, the final stage of training in fact, showed increased losses with the best architectures shown in Table IV.

IV. DISCUSSION

The outcomes of this research are the first step towards an automated tracking method for Quantitative Fluoroscopy. The implementation of common architecture modifications that were thought to hypothetically improve performance, yielded
counter-intuitive results and therefore, was useful in informing and understanding what future methods to pursue.

The effect of weighting on the rmse loss function, for example was thought to allow for better anchoring of the predictions, to avoid the ambiguity seen in Fig. 6. However, though this performed well on the first frame of sequences, its failure in successive frames shows that further fine-tuning of the weights may be required. However, doing this effectively can become an exceedingly manual task and therefore other methods should be explored first.

Statistical shape models have long been used in spinal literature for many purposes, be it assessing spinal curvature for diagnosis [15] or other complex applications such as reconstruction of 3D vertebra from 2D images [13]. Their use, was hypothesised to overcome the problems of occlusions, as now point relationships could include anatomical context. However, their poor performance suggests that the variation in spinal shape and inter-vertebral position was too great in the model is now prioritising the positioning of the SAC points, extreme intra-vertebral planar shifts and other exaggerated shape modifications can occur in order to fit the SAC (Fig. 6).

Although both shape, and shape-w configurations showed signs of marginally increased robustness to occlusions in comparison to rmse and rmse-w’s loss (Fig. 10), the gains

![Graph showing validation loss vs. inference time](image1)

![Graph showing test loss vs. inference time](image2)

**Fig. 12.** Validation Loss vs. Inference

**Fig. 13.** STAGE00 Heatmaps showing the Effects of Different Model Configurations on rmse Results (in px) for Un-obstructed (Top) and Obstructed (Bottom) Image Sequences
compared to the increased complexity of the model may not be enough to warrant further research down this avenue.

Instead, looking at what literature in lumbar spine tracking offers might provide a guide to better solutions at less complexity. Many studies propose a multi-step approach, starting with rough localisation followed by refinement on a cropped and enlarged area [3], [4], [9].

Comparing the results of our base ResNet Model (rn50 rmse adapt) to An et al.’s [3] part affinity model, there was a significant loss reduction, with their best loss at 11.37 ± 7.21 pixels, and ours at 7.85 ± 0.33 pixels. This was surprising, as it was hypothesised that the QF’s poorer image quality would have a stronger negative effect on accuracy and would not be able to match the performance of current literature. However, this may not be a true one-to-one comparison as our QF dataset has only healthy adult participants whereas, the spread of the datasets used in [3] included unhealthy patients. Though our training tried to account for this with data augmentations, these were best at simulating occlusions and variations in spinal contrast, rather than intra-vertebral shifting, and vertebra shape variations which are common in unhealthy patients. Therefore, for a fairer comparison, a QF dataset for unhealthy participants would have to either be acquired or synthesised.

Increasing the resolution of the images and the size of the ResNet model showed significant improvement in loss performance up to a point as seen in Fig. 12, and started to return higher losses in STAGE02. However, the logarithmic relationship between loss and inference time must be taken into account, since a 50% decrease in loss would come at the cost of 250% the inference time.

A. Future Work

Conversion of the predicted points into inter-vertebral angles is the next step to verifying whether the current tracking method, as-is, is suitable enough. The use of the basic ResNet architecture shows very promising results, and further research into its transferability to other modes of motion (e.g. REC_FLEX), other imaging planes (namely coronal), and other parts of the spine, can also be conducted, using this pre-trained model.

V. Conclusion

Using a basic configuration of a ResNet model, we were able to benchmark the QF dataset, and the results can be used to inform future research direction. The ResNet architecture performed remarkably well, in comparison to current literature. Modifications, to the basic architecture, such as changing attention pooling layers were found to have no significant effect. On the other hand, using weighted loss functions and shape models yielded significantly undesirable results. Only increasing image resolution and increasing ResNet layer depth yielded performance gains at the cost of inference time. We recommend looking at these avenues in more depth, and moving towards a two-stage process.

REFERENCES


