Automated Lumbar Vertebral Level Identification using Ultrasound

by

Jorden Hicklin Hetherington

B.A.Sc., The University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Biomedical Engineering)

The University of British Columbia
(Vancouver)

August 2017

© Jorden Hicklin Hetherington, 2017
Abstract

Spinal needle procedures require identification of the vertebral level for effectiveness and safety. E.g. in obstetric epidurals, the preferred target is between the third and fourth lumbar vertebra. The current clinical standard involves “blind” identification of the level through manual palpation, which only has a 30% reported accuracy. Therefore, there is a need for better anatomical identification prior to needle insertion. Ultrasound provides anatomical information to physicians, which is not obtainable via manual palpation. However, due to artifacts and the complex anatomy of the spine, ultrasound is not commonly used for pre-puncture planning.

This thesis describes two machine learning based systems that can aid physicians to utilize ultrasound for lumbar level identification.

The first system, LIT, is proposed to identify vertebrae, assigning them to their respective levels and tracking them in a sequence of ultrasound images in the para- median plane. A deep sparse auto-encoder network learns to extract anatomical features from pre-processed ultrasound images. A feasibility study (n=15) evaluated performance.

The second system, SLIDE, identifies vertebral levels from a sequence of ultrasound images in the transverse plane. The system uses a deep convolutional neural network (CNN) to classify transverse planes of the lower spine. In conjunction, a novel state-machine is developed to automatically identify vertebral levels as the transducer moves. A feasibility study (n=20) evaluated performance. The CNN achieves 88% accuracy in discriminating images from three planes of the spine. As a system, SLIDE successfully identifies all lumbar levels in 17 of 20 test scans, processed at real-time speed.

A clinical study with 76 parturient patients was performed. The study com-
pares level identification accuracy between manual palpation, versus SLIDE, with both compared to freehand ultrasound. SLIDE’s level identification outperformed palpation with an odds ratio of nearly 3. A subset of recorded ultrasound (n=60) was labeled and used to retrain the CNN, improving classification accuracy to 93%.

The systems showcase the utility of machine learning in spinal ultrasound analysis, with varied approaches to automatically identifying vertebral levels. The systems can be used to improve the accuracy of vertebral level identification compared to manual palpation alone.
Lay Summary

Needles are inserted into the spine for various medical procedures. One common procedure is epidural analgesia, which a majority of laboring women in North America receive to reduce birthing pain. For safety and effectiveness, the needle must be inserted into a prescribed location of the patient’s spine. Currently, physicians identify the target insertion location by feeling a patient’s back for bony landmarks. However, the accuracy of this method is as low as 30%. Ultrasound is a safe technology that can aid physicians in identifying patient anatomy, but is challenging to interpret and is not commonly used in practice. This thesis describes the development of two systems that can aid physicians in utilizing ultrasound to identify the correct needle insertion location. These systems use machine learning for automatic image analysis. The real-time success of the second system led to a test with 76 pregnant patients, outperforming expert manual palpation.
Preface

This thesis is predominantly derived from three manuscripts, two of which have been published [25, 26], and the third of which will be submitted for publication in 2018. The presented work involves the collaboration of students, professors, sonographers, and anesthesiologists between The University of British Columbia in the Department of Electrical and Computer Engineering and the Department of Mechanical Engineering, as well as the BC Women’s Hospital in the Department of Sonography and the Department of Anesthesiology.

Chapter 3 is a modified version of the following published work:


The author collected all of the data involved in this work. The author contributed to the design and development of the software components involved in the system, with the technical guidance and aid of Mehran Pesteie, while using some external software components as described in the manuscript. Specifically, Mehran Pesteie suggested the sparse auto-encoder network architecture used in the system, and was consulted on technical design decisions of the framework. The author was the primary contributor in writing the manuscript. Victoria Lessoway is an expert sonographer. She provided sonographic guidance and performed some image labeling. Professor Purang Abolmaesumi and Professor Robert Rohling provided technical guidance, and insight into the problem being addressed.

Chapter 4 is a modified version of the following published work:

Hetherington, J., Lessoway, V., Gunka, V., Abolmaesumi, P., & Rohling, R.
In this work, the author organized the collection of data with equal effort from Mehran Pesteie and sonographer Victoria Lessoway, who performed the scans. The author entirely designed and developed the new software components involved in the system, using some external software components as described in the manuscript. The author was the primary contributor in writing the manuscript. Dr. Vit Gunka, Professor Purang Abolmaesumi, and Professor Robert Rohling provided technical guidance, and insight into the problem being addressed.

Chapter 5 contains clinical data from work which will be submitted as a manuscript to a clinical journal. It will be co-first authored, with the following author information, and draft title:


In this work, the author performed the study alongside an anesthesiologist, Dr. Janette Brohan, as described in the chapter. The automated software used was that of Chapter 4. Statistical analyses of primary and secondary study metrics were performed by BC Women’s Hospital statistician, Dr. Arianne Albert. The author performed all further training and analysis of SLIDE in this chapter. Anesthesiologists, Dr. Simon Massey, Dr. Vit Gunka, and Dr. Anton Chau provided insight into the problem being addressed, and helped develop the protocol of the study. Professor Purang Abolmaesumi and Professor Robert Rohling provided technical guidance and support.

Chapters 3 and 4 involved ultrasound imaging of participants, all of whom provided informed written consent after approval of the study by the Clinical Research Ethics Board (Ultrasound imaging of the musculoskeletal system: laboratory study of ultrasound imaging algorithms - Certificate H15-01310). Chapter 5 involved palpation and ultrasound imaging of parturient patients at the BC Women’s Hospital, all of whom provided informed written consent after approval of the study by the Clinical Research Ethics Board (Automated UltraSound System Identification
of vertebral level in obstetric patients Study - Certificate H16-02858).
## Table of Contents

Abstract ................................................................. ii

Lay Summary ...................................................... iv

Preface ................................................................. v

Table of Contents ............................................... viii

List of Tables .................................................. xi

List of Figures ................................................... xii

Glossary ............................................................ xviii

Acknowledgments ............................................... xx

Dedication ......................................................... xxii

1 Introduction ...................................................... 1
   1.1 Motivation .................................................. 1
   1.2 Thesis objectives ........................................... 2
   1.3 Contributions ............................................... 4
   1.4 Outline .................................................... 5

2 Background .................................................... 6
   2.1 Spinal anatomy ............................................. 6
   2.2 Obstetric analgesia and anesthesia ....................... 8
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.1 Complications</td>
<td>12</td>
</tr>
<tr>
<td>2.2.2 Manual palpation</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Ultrasound in spinal analgesia</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Introduction of ultrasound</td>
<td>15</td>
</tr>
<tr>
<td>2.3.2 Ultrasound technologies for spinal anatomy understanding</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Machine learning in ultrasound</td>
<td>20</td>
</tr>
<tr>
<td>2.4.1 Classification</td>
<td>20</td>
</tr>
<tr>
<td>2.4.2 Segmentation and localization</td>
<td>23</td>
</tr>
<tr>
<td>2.4.3 Transfer learning</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Summary</td>
<td>24</td>
</tr>
<tr>
<td>3 LIT: Lamina Identification and Tracking using deep networks with</td>
<td>26</td>
</tr>
<tr>
<td>unsupervised feature learning</td>
<td></td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>26</td>
</tr>
<tr>
<td>3.2 Methods and materials</td>
<td>27</td>
</tr>
<tr>
<td>3.2.1 Data acquisition</td>
<td>27</td>
</tr>
<tr>
<td>3.2.2 Ultrasound bone enhancement</td>
<td>29</td>
</tr>
<tr>
<td>3.2.3 Vertebrae identification</td>
<td>30</td>
</tr>
<tr>
<td>3.2.4 Vertebrae level labeling and tracking</td>
<td>34</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>38</td>
</tr>
<tr>
<td>3.4 Discussion and conclusion</td>
<td>39</td>
</tr>
<tr>
<td>4 SLIDE: automatic Spine Level IDENTification system using a deep</td>
<td>41</td>
</tr>
<tr>
<td>convolutional neural network</td>
<td></td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>4.2 Methods and materials</td>
<td>42</td>
</tr>
<tr>
<td>4.2.1 System overview</td>
<td>42</td>
</tr>
<tr>
<td>4.2.2 Data acquisition</td>
<td>44</td>
</tr>
<tr>
<td>4.2.3 Plane classification</td>
<td>45</td>
</tr>
<tr>
<td>4.2.4 State machine for improved vertebral level identification</td>
<td>48</td>
</tr>
<tr>
<td>4.2.5 Real-time system</td>
<td>51</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>53</td>
</tr>
<tr>
<td>4.3.1 Transfer learning training</td>
<td>53</td>
</tr>
</tbody>
</table>
4.3.2 Vertebral level identification ........................................ 54
4.3.3 Real-time implementation ........................................... 56
4.4 Discussion and conclusion ............................................ 56

5 Clinical evaluation of SLIDE ............................................ 59
  5.1 Introduction .......................................................... 59
  5.2 Experimental design ................................................ 61
    5.2.1 Statistical analysis methods .................................... 64
    5.2.2 Data acquisition protocol ...................................... 65
    5.2.3 US data storage ................................................ 67
    5.2.4 SLIDE CNN analysis methods ................................. 68
  5.3 Results .................................................................... 68
    5.3.1 Primary clinical outcome: accuracy of L3-L4 ............... 69
    5.3.2 Secondary clinical outcomes: other levels .................. 70
    5.3.3 Other clinical outcomes ........................................ 70
    5.3.4 SLIDE CNN analysis ............................................ 72
  5.4 Discussion and conclusions ........................................ 73

6 Conclusion and future work ............................................. 75
  6.1 Summary of contributions .......................................... 75
  6.2 Limitations and future work ....................................... 77

Bibliography ............................................................... 80
List of Tables

Table 4.1  CNN 20-fold cross validation accuracy and CNN classification speed. .......................................................... 54
Table 5.1  Demographic data of 76 included patients in SLIDE study . . . 64
Table 5.2  Summary of accuracy and regression analyses for all levels, between manual palpation and automatic US analysis with SLIDE. 72
Table 5.3  SLIDE transition-accuracy. The number of correct transitions are compared to the number of incorrect transitions that occurred due to falsely transitioning away and back to a plane (bouncing) or missing a plane. The number of transitions should ideally sum to 76 (the total number of patients) in each column. However, in a few cases, SLIDE failed to transition above a certain level (e.g., L1-L2 column sums to 72, indicating in 4 patients, this metric was not recorded). ......................... 73
List of Figures

Figure 1.1 US scan in the paramedian (left) and transverse (right) plane. Either plane can be used to identify the vertebral level, although the paramedian is conventionally used first to find the vertebral level, and the transverse plane is used second to find the anatomical midline of the spine. 3

Figure 2.1 Dermatome chart. Anatomical mapping of skin sensation related to spinal nerve location. Artist: Mikael Häggström, used with permission. 7

Figure 2.2 Spinal cord tapering at the conus medullaris, resulting in the cauda equina. Public domain historical image: Carter & Gray, Anatomy of the Human Body, 1858. 8

Figure 2.3 Lumbar vertebrae with prominent ananomies labeled. Artist: Vickie Earle, used with permission. 9

Figure 2.4 Sagittal view of the lumbar vertebrae with prominent anatomies labeled. In this illustration, the needle is inserted into the epidural space via the L3-L4 intervertebral gap, as is convention in epidural injections. Artist: Vickie Earle, used with permission. 10

Figure 2.5 Anatomical planes. Sagittal plane located at the anatomical midline. Paramedian plane also referred to as parasagittal plane. This edited image is licensed by Rice University, at cnx.org/contents/F-TuqKAF@3/Anatomical-Terminology, under a Creative Commons Attribution License (by 4.0): creativecommons.org/licenses/by/4.0/ 14
Figure 2.6 Exemplary US machine. ........................................... 15
Figure 2.7 Visualization of basic US principal function. (a) Shows an idealized wave path to construct an A-mode signal. (b) Shows an idealized multi-crystal wave propagation that is used to construct a 2D B-mode image. ........................................... 16
Figure 2.8 Example 2D US B-mode image. Basic anatomy is labeled and acoustic shadowing beneath the highly reflective bone is identified. ........................................... 17
Figure 2.9 Example fully-connected neural network with 3 hidden layers. This image is licensed by Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015, at neuralnetworksanddeeplearning.com/chap6.html, under a Creative Commons Attribution-NonCommercial 3.0 Unported License: creativecommons.org/licenses/by/4.0/ ................. 21
Figure 2.10 Example convolutional neural network architecture. ........... 22
Figure 3.1 Data flow of LIT. Original US images are augmented with the bone enhanced images by a weighted sum, and downsized to $83 \times 110$. A $27 \times 27$ sliding window generates image patches for pixel-level classification. Unsupervised features are extracted from the image patch by two sparse auto-encoders which were previously trained, and classified by a fully-connected layer, to identify the vertebrae in the image. Each vertebra is then labeled with its respective level and is tracked in the sequential images. ........................................... 28
Figure 3.2 Paramedian US of the lumbar and sacral vertebrae, with anatomies labeled. The left figure shows a typical starting image, with the intervertebral gap between the sacrum and the L5 (the L5-S1 gap) approximately centered in the image. It also highlights the peak of the L5 lamina, and a typical reflection from the sacrum. The right figure shows a typical paramedian image of only the lumbar vertebra, often referred to as a wave or sawtooth image, with prominent anatomies indicated. ................. 29
Figure 3.3  (a) Collection path on the left and right paramedian planes of all participants, moving cephalid from the L5-S1 intervertebral gap up to the L1 vertebrae. (b) The paramedian imaging plane intersects the vertebrae as indicated by the purple plane. Artist: Vickie Earle, used with permission. (c) Exemplary US image in the paramedian plane showing the wave-like pattern of laminae. In this image, the reflection of four lamina are visible. .

Figure 3.4  LIT bone enhanced image. A weighted sum of the initial raw B-mode image and a bone response estimation. .

Figure 3.5  Example SAE configuration. $x$ and $\hat{x}$ both have the same dimensionality, which is higher than the encoded dimensionality of the hidden layer, $h$. .

Figure 3.6  Horizontal position of lamina peaks vs frame number, with labels and LIT outputs compared. Examples of lamina tracking successes in (a) vs. certain level mismatches shown in (b) and (c). (a) Coloured points indicate the outputs of LIT, and black “X”s are the manually identified peak position labels. The first label in time, that has the highest position (y-axis) corresponds to L5, and the next label corresponds to L4 and so on. Label levels always correspond to the closest label in the next frame (the closest “X”s). (b) An off-by-one tracking level mismatch, where all system outputs track labels, but indicate the wrong level because the L5 was missed (indicated by the grouping of “X”s on the left without matching LIT outputs). Low contrast in many initial frames caused L5 to be missed. (c) An example of two level mismatches (red points jump from L4 labels to L3 labels, but converge back to L4 labels).

Figure 3.7  Confusion matrix for the classification of test image patches from paramedian US as either lamina (0) or not lamina (1). .
Figure 4.1  Anatomy and representative US images of the lower vertebrae. Overlaid on the lumbar spine is the desired path of a operator. The operator begins at the sacrum, with the US transducer oriented in the transverse plane, and moves cephalad up to the L1 vertebra. The associated US images on the right show typical images from three distinct planes. Modified work from artist: Vickie Earle, used with permission.  

Figure 4.2  SLIDE displays. The left figure shows the desktop display integrated into 3D Slicer, indicating the current predicted level and current predicted plane, along with the associated confidence from the deep network output (showing the gap classification in green with > 99% confidence). The right figure shows the AR display projected onto the patient’s skin (enlarged for visibility in the figure). In this diagram, the predicted level, which is being projected, matches the gold standard markings on the participant’s skin.  

Figure 4.3  System architecture, with data flow and hardware configuration.  

Figure 4.4  Inter-frame NCC comparison between all pairs of frames in a given data set. The diagonal shows the NCC with a given frame compared to itself, having value of 1. A given row or column (mirrored along the diagonal) shows the NCC comparison between one frame and all others. A square wave in a row or column would be expected if such a strategy could identify planar transitions (which is not present).  

Figure 4.5  Transition frame classification ambiguity due to US slice-thickness artifacts. The AC and AP are expected in intervertebral gap images. The lamina is expected in vertebral bone images. In the transition image between gap and bone, the AC, AP, and lamina are present, which make its true class ambiguous.
Figure 4.6 Vertebral plane identification state machine. Averaged classification probabilities for the sacrum, gap, and bone planes are input into the state machine, which transitions between identified planes, thereby counting up a patient’s spine. Certain cases will revert the system into the Unknown state from any other state.

Figure 4.7 (a) Exemplary output from the softmax layer of the CNN softmax layer, $P$. (b) Averaged values of $P$, $P^*$. (c) Various outputs, compared to the labeled level transitions. The SLIDE output shows a successful result, as there is overlap at all vertebral levels from the sacrum to L1. Latency between plane transitions in SLIDEs output versus the labels is due to a difference in protocol for identifying transition locations. With $t_{stable} = 0$ and $t_{unstable} = 0$, removing the utility of the thresholds in the state machine, a transition-error occurs at L5-Sacrum.

Figure 4.8 AR projection hardware. Custom 3D printed mount to attach a PicoPro projector to the Analogic C5-2/60 curvilinear transducer.

Figure 4.9 SqueezeNet CNN architecture. The left figure shows a high level 10 layer CNN. This architecture includes simple (direct) and complex (convolutional) bypass connections between layers. The right figure shows an expanded view of SqueezeNet’s Fire module. Numbers beneath a layer indicate that layer’s depth.

Figure 5.1 Transferring markings from the patient’s back to a fixed transparency sheet. Reference lines are first matched from the corners and sides of the transparency sheet to markings on the patient’s back, and then results are transferred onto the transparency sheet. Marks on the skin are erased to blind results between methods.
Figure 5.2  Example populated transparency sheet. The left figure shows the coloured markings, without annotations. The right figure shows the full sheet, with annotations describing the markings. Horizontal dashed lines are extended from the identified locations of L1 through L5. This example shows success in SLIDE and palpation for all 5 identified gaps. L3-L4 markings are pointed to, and both reside in the “bin” indicated by the L3 and L4 freehand US markings. The column of the palpation markings is further to the right, because it was marked close to the patient’s midline, whereas the SLIDE and freehand US markings are more distal because the transducer and gel cover the midline.  

Figure 5.3  Identified level by palpation or SLIDE relative to freehand US for the primary outcome: L3-L4 identification. The size of the circles is proportional to the number of cases in each category as indicated by the numbers in the circles.  

Figure 5.4  Identified level by palpation or SLIDE relative to freehand US for all intervertebral gaps. The size of the circles is proportional to the number of cases in each category. Larger size along the diagonal shows higher performance. As shown along the diagonal, SLIDE outperforms manual palpation at identifying all lumbar interspaces correctly. Also shown is that manual palpation has a higher tendency to incorrectly identify levels higher than the true location, compared to SLIDE. This adds higher risk to the patient than incorrectly identifying levels lower than the true location.
Glossary

**1D** one-dimensional

**2D** two-dimensional

**3D** three-dimensional

**AC** anterior complex, spinal anatomy

**ANOVA** analysis of variance, a set of statistical techniques to identify sources of variability between groups

**AR** augmented reality

**AUSSI** Automated UltraSound Software for Identification of lumbar vertebral levels, the formal clinical study name presented in this thesis

**BMI** body mass index

**CI** confidence interval

**CNN** convolutional neural network

**CPU** central processing unit

**CSA** continuous spinal analgesia

**CSE** combined spinal epidural

**CSF** cerebrospinal fluid

**CT** computed tomography, used for 3 dimensional medical imaging with X-rays
EM  electromagnetic
FPS  frames per second
GPU  graphics processing unit
ILSVRC  ImageNet Large Scale Visual Recognition Challenge
LF  ligamentum flavum, spinal anatomy
LIT  Lamina Identification and Tracking system, presented in this thesis
ML  machine learning
MRI  magnetic resonance imaging, a form of non-ionizing 3 dimensional medical imaging
NCC  normalized cross-correlation
OR  odds ratio, a statistical measure representing the odds that an outcome will occur with a given technique, compared to the odds that the outcome will occur without that technique
PDPH  post-dural puncture headache
PLUS  Public software Library for UltraSound imaging research
RNN  recurrent neural network
SAE  sparse auto-encoder
SLIDE  Spine Level IDENTification system, presented in this thesis
SVM  support vector machine
UI  user interface
US  ultrasound
Acknowledgments

I would like to give my utmost gratitude to Professor Robert Rohling and Professor Purang Abolmaesumi for their dedicated role as my co-supervisors. Your insight into our field and guidance throughout my program has been invaluable. I would additionally like to thank all of the instructors that I have had the good fortune to learn from over the last few years.

Thank you to all of the amazing clinical colleagues I have had the pleasure of working with. Thank you Victoria Lessoway for your countless hours teaching me about sonography. Thank you Dr. Janette Brohan for all of our early starts in the hospital that included anesthesiology lessons over coffee.

The Engineers in Scrubs fellowship furthered my appreciation of biomedical innovation. Throughout this program, I got to work with remarkable colleagues, all of whom helped cultivate a fantastic environment and deserve thanks.

I have had the pleasure of being part of the Robotics and Control Lab. Thank you to every single member that I have interacted with for making it a fantastic place to work. I am lucky to have worked with and learned from many of you.

To my friends who took time out of their busy lives to help edit this thesis, Alexander Vlasev, Gorden Larson, Manyou Ma, and Rohit Singla, thank you again.

I would like to whole-heartedly thank my entire family, amazing girlfriend, and supportive friends for their continual love and support. Thank you dad for everything. Thank you mom, Quang, Daisy, and Weili, for your care and making sure I didn’t eat instant noodles every night of my degree. Thank you to Tayler and Shadi for all of your support and acting as guides through the unknown aspects of Grad School. Thank you Shayna for all of your dedication and love, and keeping me sane with our dancing. Thank you to the rest of my family throughout Canada,
especially my grandma and grandpa who are always happy to chat. Thank you to all of my friends inside and outside of academia for all of the good that you bring to my life.

Finally, I would like to acknowledge the Natural Science and Engineering Research Council (NSERC), Canadian Institutes of Health Research (CIHR), and Collaborative Research and Training Experience (CREATE), who helped fund this work.
Dedication

This thesis is dedicated to my father, Darrin Ray Hetherington, who inspired me to become an Engineer. This work would not be possible without him.
Chapter 1

Introduction

The position of woman in any civilization is an index of the advancement of that civilization; the position of woman is gauged best by the care given her at the birth of her child.
— Howard W. Haggard

1.1 Motivation

Percutaneous spinal needle insertions are common and are involved in a wide range of procedures. For example, neuraxial analgesia (epidurals), spinals, and facet joint injections involve the injection of analgesic agents for pain relief. There are millions of these procedures done every year in the United States alone. E.g., in obstetric anesthesia, there were approximately 2.4 million epidurals given to labouring women in the United States in 2012 [60].

In spinal needle insertions, it is important for the physician to insert the needle at a prescribed location to ensure efficacy of the procedure and safety of the patient. The insertion site for epidurals is between the third and fourth lumbar vertebrae (L3-L4), which is often administered to women during labour [53]. Most frequently, epidural insertions are performed using either manual palpation and loss-of-resistance [59] or under fluoroscopy guidance [18]. Fluoroscopy can be used to help needle placement in certain procedures, such as facet joint injections for pain management. However, due to the exposure of ionizing radiation, fluoroscopy-guided injections are prohibited for pregnant patients without other medical justifi-
cation. Thus, in obstetrics the current standard clinical method involves using only manual palpation of bony landmarks, including the spinous processes and ilium, for the identification of vertebral levels [3]. Manual palpation alone is considered “blind” identification, and Furness et al. showed that this technique is only 30% accurate in identifying vertebral levels [19]. This demonstrates the need for better anatomical identification prior to needle insertion.

The use of ultrasound (US) in neuraxial techniques improves injection success rates, and reveals anatomical information that is not obtainable by manual palpation alone [6, 54]. When used for epidural placement, US is typically first used to find the appropriate vertebral level in a paramedian view and mark it on the skin. The midline is then identified and marked in a transverse view [58]. The transducer placement in the paramedian and transverse planes are shown in Figure 1.1, and further expanded upon in Chapter 2. The intersection of the two marks provides the injection site on the skin. US imaging to visualize and identify the vertebral anatomy is continually being investigated as a solution by anesthesiologists. However, US is a complex imaging modality that produces images with many different artifacts, and requires the operator to have a strong understanding of cross-sectional anatomy. Thus, image interpretation of complex spinal anatomy remains a problem for these typically inexperienced users of US [14].

To help anesthesiologists with the interpretation of US images and to identify the vertebral level, previous work has explored the identification, tracking, and labeling of vertebrae in US images taken from the paramedian plane [31, 68] and the transverse plane [65–67]. Visualization and identification tasks often include using external tracking equipment [48, 50, 61]. Such tracking systems allow more freedom for anesthesiologists in the choice of their US scanning path, but add cost and additional steps to clinical workflow. A solution is needed that can enable identification and labeling of vertebrae in standard two-dimensional (2D) US images without external tracking hardware.

1.2 Thesis objectives

The work presented in this thesis proposes to aid physicians in utilizing US for spinal needle insertions, with an emphasis on obstetric spinal procedures. The
Figure 1.1: US scan in the paramedian (left) and transverse (right) plane. Either plane can be used to identify the vertebral level, although the paramedian is conventionally used first to find the vertebral level, and the transverse plane is used second to find the anatomical midline of the spine.

benefits of using US in neuraxial techniques when compared to palpation alone [6, 54] indicate that it should be used more frequently, but the standard of care is still manual palpation. This work attempts to reduce the barriers of using US by automatically analyzing anatomy in spinal US images, such that non-experts in sonography can utilize US. The main hypothesis of this thesis is that machine learning (ML), when applied to standard US, can provide sufficiently accurate guidance information of vertebral level to be considered for clinical use.

Two systems to automatically identify lumbar vertebral levels are proposed. The systems are developed with the constraints of requiring no external tracking hardware, but rather utilizing pre-defined clinical scan paths and analyzing stan-
standard US images. Further, the systems must fit into the existing clinical workflow. The potential clinical benefit that this work aims to achieve in the long term is to improve the safety and efficacy of spinal needle insertions by increasing success rates in vertebral level identification.

1.3 Contributions

- A system is developed, Lamina Identification and Tracking system (LIT), for the automatic identification of lumbar vertebral levels, given a known paramedian scan path of an US transducer. This work proposes a deep network architecture, making use of sparse auto-encoders, to detect vertebrae in standard paramedian 2D US images of the lumbar spine. This network is trained on data collected from 15 volunteers, and has results evaluated against an expert sonographer. LIT also adopts an algorithm to label and track the respective vertebrae in 2D US image sequences.

- A second system is developed, the Spine Level IDENTification system (SLIDE), for the automatic identification of lumbar vertebral levels, given a known transverse scan path of the US transducer. This work proposes a framework that uses a deep convolutional neural network (CNN) for automatic classification of transverse spinal US planes. This work also proposes a novel state machine, based on a priori knowledge of the clinical imaging protocol, to identify the vertebral level. A real-time implementation of this framework is shown. A user display is provided graphically as well as via an augmented reality (AR) projector-based display.

- A clinical trial to evaluate SLIDE is performed. This trial included 76 parturient (at-term women about to give birth) patients from the British Columbia Women’s Hospital. An evaluation of vertebral level identification via manual palpation versus SLIDE, compared to a freehand US gold-standard was performed. US data from this trial was collected and analyzed to further evaluate and improve the ML components of SLIDE.
1.4 Outline

This thesis covers the background of spinal analgesia, anesthesia, and the relevant technologies for the proposed systems; the details of the two systems developed; and the evaluation and implication of the clinical trials performed on parturient patients. The outline of the thesis is as follows:

Chapter 2: Background Provides an overview of basic US theory and spinal analgesia and anesthesia practices as well as a review of prior work in vertebral level identification via US and relevant ML techniques used in US.

Chapter 3: LIT: Lamina Identification and Tracking using deep networks with unsupervised feature learning Describes the data collection, development and evaluation of a proposed vertebral level identification and tracking system, LIT.

Chapter 4: SLIDE: automatic Spine Level IDENTification system using a deep convolutional neural network Describes the data collection, development and evaluation of a second proposed system, SLIDE. SLIDE is constructed to solve some of the problems identified in the first system while also identifying vertebral levels.

Chapter 5: Clinical evaluation of SLIDE Describes the protocol and evaluation of a clinical study that includes 76 parturient patients, including the results of updating SLIDE’s internal ML model based on the data from these patients.

Chapter 6: Conclusion and future work Summarizes the objectives of the research and the contributions made, and describes potential applications and directions for this research to be continued.
Chapter 2

Background

*I not only use all the brains that I have, but all I can borrow.*
– Woodrow Wilson

2.1 Spinal anatomy

The spine is a critical structure, giving support to the body. In a regular spine, there are 33 vertebrae, with 7 cervical vertebrae, 12 thoracic vertebrae, 5 lumbar vertebrae, 5 fused sacral vertebrae, which composes the sacrum, and 4 fused coccygeal vertebrae, which composes the coccyx. Vertebrae can be given shorthand names, such as thoracic levels 1 to 12 being referred to as T1 to T12, or lumbar levels 1 to 5 being referred to as L1 to L5. An intervertebral gap can be specified by naming two adjacent vertebrae, such as L3-L4, the intervertebral gap between the third and fourth lumbar vertebrae.

Extending from the brain, the spinal cord travels down the spine and branches sensory and motor nerves to the rest of the body, acting as the neurological pathway between the body and the brain. The anatomical region which communicates with nerves branching from a given vertebral level is shown in a dermatome chart, as in Figure 2.1. The spinal cord initially consists of a tight bundle of fibers in the upper region of the spine. Towards the inferior (low) end of the spine, often assumed to be near the T12-L1 gap, the spinal nerves tapers at the conus medullaris, resulting in the lower cauda equina traveling through the lumbar vertebrae [3]. Cauda equina
Figure 2.1: Dermatome chart. Anatomical mapping of skin sensation related to spinal nerve location. Artist: Mikael Häggström, used with permission.
2.2 Obstetric analgesia and anesthesia

The pain from labour can be expected to be the most severe pain that a woman will face in her lifetime [3]. As such, many techniques have been researched and developed to relieve the woman’s labour pains. Women can opt to have non-neuraxial or neuraxial analgesic techniques, where analgesia is defined as the absence of pain sensation. Many non-neuraxial techniques exist, including lumbosacral massage, hypnobirthing, acupuncture, sedative anxiolytics, and systemic opioids [3]. There is a wide range of efficacy and research within these methods. Alternatively, many
women may opt to have neuraxial analgesia, which is the most effective technique and directly affects the nervous system [3]. In the United States alone, the annual rate of spinal analgesia for labouring women was approximately 2.4 million in 2012 [60]. Among all labouring women, approximately 60% choose to have spinal analgesia [46].

In neuraxial analgesia, pain relief is caused by the use of local anesthetics directly delivered to the central nervous system via a percutaneous injection to the epidural space, illustrated in Figure 2.4 [3]. Local anesthetics are a drug class that temporarily stops the conduction of nerves after regional administration [3]. They work by disrupting the action potential of the nerves, stopping signals from propagating.

The goal of neuraxial analgesia and anesthesia is to safely reduce the patient’s sensation of pain during childbirth, while not adversely affecting the obstetric outcomes. Neuraxial anesthesia differs from analgesia in that it not only affects the patient’s sensation of pain, but can also reduces motor function. Anesthesia, which is stronger, is primarily used during Cesarean sections operations, where the baby is removed from the mother surgically. This is opposed to analgesia primarily given
Figure 2.4: Sagittal view of the lumbar vertebrae with prominent anatomies labeled. In this illustration, the needle is inserted into the epidural space via the L3-L4 intervertebral gap, as is convention in epidural injections. Artist: Vickie Earle, used with permission.

during labour. Drug developments have created both analgesia solutions and anesthetic solutions. Analgesic agents are diluted concentrations of local anesthetics and opioids. Anesthetic agents have higher concentrations that cause motor function blockage and can potentially affect obstetric outcomes [3]. Pain from labour primarily comes from nerves extending between T10 and S4, and thus must be blocked for sufficient analgesia effects [3].

Analgesic agents can be administered by anesthesiologists with three primary techniques: epidural analgesia, continuous spinal analgesia (CSA), and combined spinal epidural (CSE) [3]. These techniques have various pros and cons associated with each. Epidurals involve the threading of a catheter in the epidural space, to provide a passageway for the delivery of the analgesic agents. The epidural needle is commonly inserted at the L3-L4 lumbar vertebrae, with a needle large enough for a catheter to be threaded through [3]. This injection is often done with the loss-of-resistance technique [59] whereby the epidural needle is advanced in midline of the patient’s back, with pressure induced from a syringe full of either air
or saline. Once the needle punctures the ligamentum flavum (LF), it enters the epidural space, losing resistance and injecting the air or saline. Then a catheter can be threaded into this space for delivery of medication throughout labour. With the large needle diameter, puncturing the dura mater becomes a risk-factor in this procedure because cerebrospinal fluid (CSF) can eject through the needle, and continually eject out of the spinal canal, often leading to debilitating post-dural puncture headache (PDPH) [52].

CSA involves the threading of a catheter for analgesic agents directly into the spinal canal. This technique is normally the compensation for an accidental dural puncture in an epidural, where it is decided to thread a catheter through the already punctured dura mater, rather than re-attempt the epidural [3]. The main disadvantage of this technique is the potential fatal danger to the patient if the wrong medication is given directly to the spinal canal.

CSE involves the insertion of both an epidural needle, and a smaller spinal needle, whereby an epidural needle is inserted into the epidural space, and a spinal needle is threaded through the epidural needle to intentionally puncture the dura mater and deliver a spinal anesthetic agent within the spinal canal [3]. By using a smaller needle, less CSF leaks from the spinal canal reducing the chance of PDPH. Following the spinal injection, a catheter is threaded into the epidural space for continuous delivery of medication during the rest of the birth. In some cases, such as when a patient is in too much pain to withstand the larger gauge epidural needle, the two injections are done independently [3]. CSEs are known to provide faster onset of analgesia than a conventional epidural [3], but the risk of meningitis is increased because the dura mater is intentionally punctured [3].

For anesthesia, a fourth technique exists referred to as a single-shot spinal. A single-shot spinal involves the placement of a needle directly into the spinal canal, which delivers an anesthetic agent in a single dose and then is removed. Much like in a CSE, single-shot spinals have a quick onset. However, without the continuous drug delivery from the epidural, they have a short duration, and as such are typically only used for cesarean sections or in facilities that do not have epidural equipment.
2.2.1 Complications

Some complications were discussed in the above section, but this section provides a more thorough overview of the potential complications from obstetric spinal analgesia. It specifically provides insight into the challenges that this thesis aims to address.

Although neuraxial analgesia and anesthesia are considered safe overall, there are known risks from the procedure. Some risks are due to the medication itself, such as hypotension (decreased blood pressure) and pruritus (severely itchy skin) [3]. Other risks are inherent to any percutaneous procedure, such as infection; specifically meningitis, which is at a higher risk when the dura mater is punctured. Other risks specific to spinal analgesia include failed analgesia (inadequately or malpositioned epidural catheter), unintended dural punctures, epidural hematoma and abscess, neurologic deficits, and in extremely rare cases, death [3]. Unintended dural punctures occur in 1-8% of all cases, with 50-80% inducing PDPH due to the lost CSF through the large epidural needle [52] [3]. The incidence of dural punctures is inversely related to the experience of the anesthesiologist [3]. Epidural Hematoma and abscess occur on the order of 1 in 100,000 [3]. Symptoms include neurologic problems, which quickly increase as pressure builds causing spinal compression.

When the spinal needle is inserted at a higher level than targeted, the severity of these complications can be increased. This is specifically the case with dural punctures. As in Figure 2.2, the spinal cord is a tight bundle of nerves above the conus medullaris. Once tapered, the nerves form the cauda equina. In the context of dural punctures, cord damage is more likely above the conus medullaris, where the needle can permanently damage the cord, rather than in the cauda equina, where the needle can push aside nerve fibers [42]. This is why physicians must identify the vertebral level for these procedures.

2.2.2 Manual palpation

The current standard practice for identification of the spinal injection site is via manual palpation. Manual palpation involves the feeling of the patient’s back for bony landmarks to identify both the target intervertebral gap and the midline of
the spine. To find the target lumbar interspace, the physician first finds the peaks of both iliac crests. The line between these two points is known as the Tuffier’s line, sometimes referred to as the intercristal line. In pregnant patients, this line is assumed to cross the L3-L4 intervertebral gap, which is most often the target interspace [3].

In practice, the Tuffier’s line does not always cross the L3-L4. In non-pregnant patients, the line more commonly crosses the L4 body, or the L4-L5 intervertebral gap [3]. Pelvic rotation caused by pregnancy shifts the relative location of the Tuffier’s line. Furthermore, aging affects the relative anatomical location of the Tuffier’s line [12]. This is important because it indicates that the true anatomical location of the Tuffier’s line is not known via manual palpation alone.

Multiple studies have been performed investigating the accuracy of anesthesiologists in identifying lumbar intervertebral gaps. The findings include vertebral level identification accuracies of 29% when looking in a healthy adult population of n=100, compared to an magnetic resonance imaging (MRI) gold standard [7] and 30% with n=50, compared against X-ray [19]. In another study, looking in pregnant adult population of n=121, accuracies were shown to be 55% when compared between the actual injection site and post-procedure freehand US [63]. The use of US as a gold standard should be understood to have potential errors, as shown by a study by Watson et al. showing that an anesthesiologist using freehand US mislabeled the L3-L4 interspace 25% of the time, compared to MRI [62]. In another study, the use of freehand US resulted in mislabeling the L3-L4 interspace 29% of the time, compared to X-ray [19].

2.3 Ultrasound in spinal analgesia

The research of US aiding spinal procedures is rapidly advancing. In obstetric analgesia, it has been shown that US can improve injection success rates by revealing anatomical information not otherwise visible by manual palpation [6, 54]. Adoption of US in clinical practice is limited, with manual palpation still being the clinical standard of care. One contributing factor is the lack of experience in this challenging image interpretation [14].

When US is used for injection site planning, the typical scan procedure is to
first, find the appropriate vertebral level in a paramedian view and mark it on the skin. Then, the midline is identified and marked in a transverse view [58]. The intersection of the two marks provides the skin injection site. The anatomical planes are shown in Figure 2.5. However, this protocol requires the operator to have a strong understanding of 2D cross-sectional spinal anatomy to interpret US images. It also requires them to be able to setup and use an US machine, with the dexterity required to position the transducer correctly.

The rest of this section first gives an introduction to US, then describes the literature on various technologies and systems that have been developed for identifying the US transducer position for puncture site identification.
2.3.1 Introduction of ultrasound

This section briefly introduces the principals of US. US is a relatively new imaging modality, with initial clinical applications in the 50’s [64]. US is unique in medical imaging in that it can non-invasively image anatomies within the body in real time, using non-ionizing radiation. Compared to other non-ionizing modalities such as MRI, US machines are significantly less expensive to purchase and to operate.

Modern US machines have a processing computer, with a user interface (UI). US transducers are used to scan a patient and are attached to the computer. An example US machine can be seen in Figure 2.6. Successful advances are being made to make US transducers wireless, and to reduce the size of the processing computers, with some systems now working with mobile phones.

US transducers function by producing ultrasonic sound waves that propagate into tissues, and upon interaction with various tissue types, reflect back to the transducer. The sound waves are in the range of 1-20 MHz for medical US. The waves are produced by the conversion of electrical to mechanical energy via piezoelectric crystals. Likewise, the crystals act as receivers to convert returning mechanical waves to electrical signals that will be processed.
In an idealized model, a single piezoelectric crystal will send a wave into tissue in a single line. As this wave travels through the tissue, when it encounters different tissue types with differing acoustic impedances, it will partially reflect towards the crystal. The signal received at the crystal can be used to calculate the intensity and the depth of the reflection based on assumptions of the speed of the wave in the tissue. This is visualized in Figure 2.7 (a). This single line calculation is referred to as A-mode US, which calculates the amplitude response, and can help identify transitions in boundaries. If several crystals are positioned in a one-dimensional (1D) line, and used to calculate successive amplitude lines, then a 2D image can be constructed. This is visualized in Figure 2.7 (b). This 2D US image is referred to as a B-mode image, which calculates the amplitude of several lines, and visualizes the brightness of the lines. Three-dimensional (3D) US also exists, whereby a transducer either contains a 2D grid of crystals, or a 1D line of crystals that is swept back and forth to produce a stack of 2D images.

One drawback of US is the many artifacts that appear in the final images. The
Figure 2.8: Example 2D US B-mode image. Basic anatomy is labeled and acoustic shadowing beneath the highly reflective bone is identified.

practical aspects of transmitting and receiving ultrasonic waves into tissue result in these artifacts, which sonographers must be trained to understand. The most relevant artifacts applicable to this thesis are speckle, slice-thickness, and acoustic shadowing. An exemplary 2D B-mode US image is shown in Figure 2.8. This image contains anatomies including skin, fat, muscle, ligament, and bone. Speckle can be seen throughout the image, with a granular appearance. Acoustic shadowing occurs when the US beam loses power due to reflections from highly reflective material (such as bone, as seen in the image), or from many reflections as the wave travels deeper. This artifact causes inherent limitations to the use of US in extremely obese patients, or when trying to image anatomies hidden behind highly reflective boundaries, such as at bone (e.g. the brain is hidden behind the skull) or air (e.g. bronchial structures are hidden in the lung). In practice, B-mode images actually contain information from a slice with some thickness, referred to as slice-thickness artifacts. Unlike the ideal model, waves from one crystal do not travel in a single line (in A-mode US), nor a single plane from a 1D line of crystals (in B-mode US). In B-mode images, this results in super-imposed anatomy of multiple parallel planes beneath the transducer. This is further elaborated in Chapter 4.
2.3.2 Ultrasound technologies for spinal anatomy understanding

With the challenges of US image interpretation, and complex spinal anatomy, much work has gone into developing technologies for the understanding of spinal anatomy in US images.

Chen et al. introduced a navigation system utilizing tracked US for spinal needle insertion procedures [10]. In their system, it is assumed that a pre-existing computed tomography (CT) scan has already taken place. Specifically, they used a phantom model of the lumbar spine. Using an electromagnetic (EM) tracked transducer, they reconstructed a 3D US volume of the lumbar model from a series of 2D images. This volume was then registered to the CT volume in a biomechanically constrained manner. Upon registration, they were able to construct digitally reconstructed radiographs of the initial CT volume to match the placement of the transducer. This effectively shows the CT data in the plane of the US transducer through a UI in real time. Although promising, this work relies on multiple key factors including: a preoperative CT; rough manual segmentation of the preoperative CT; initial alignment of the CT to the US volume; as well as EM tracking hardware. With regards to the initial alignment problem, their work was done entirely on a phantom, which implies that an in-vivo solution would first require identifying the vertebral level with respect to the transducer for the possibility of success.

To aid in visualization of the spinal anatomy present in US images, without the need for a pre-operative CT scan, Rasoulian et al. developed an atlas based system, that integrates with EM tracked US [49–51]. Their work proposes the augmentation of US images with a registered statistical shape model of the spine, based on statistical modes of variation from previously acquired, manually segmented CT scans. Through utilizing tracked US, a 3D volume from a series of 2D images could be generated. The statistical model is registered to the volume to find the best fit. The 2D scene could then be augmented with the relevant cross-sections of the fitted model, outlining bony anatomy in the US image to the operator.

Further work includes Rasoulian et al.’s augmentation of 3D US volumes with a statistically derived model. In this work, their prior atlas-based systems were applied to 3D volumes taken at the intervertebral level of interest, acquired with a 3D US transducer. Their goal is to provide an augmented scene of the sagittal
midline in real time such that the target LF, as well as the needle, could be viewed during an injection. Brudfors made the system real-time [8]. This project has been extended to include pre-operative CT and MRI scans of the patient to improve the intra-operative registration of the US images to the statistical model [4, 5]. This is done by performing a concurrent registration between the modalities. Although valuable, this work also relies on the use of external tracking hardware, which takes time to set up, and requires additional workspace. All of these systems require the manual identification of vertebral levels. Therefore, the work in this thesis not only has the possibility of acting as a standalone system, but also has the potential to aid the aforementioned atlas-based systems.

Ungi et al. additionally discusses in detail EM tracking hardware and software for spinal interventions, making use of tracked transducers, needles, and external markers [61]. Other spinal visualization and identification systems have been investigated, which utilize other external tracking hardware. For example, Rafii et al. develop a camera-based system, which constructs an US panorama with reference to patient-mounted fiducial markers for epidural navigation [48].

The primary constraint with the aforementioned systems is their reliance on external tracking hardware. This hardware adds additional costs and additional workflow steps to any solution. Thus, work has also gone into the development of solutions that do not rely on external tracking hardware, but rather work purely off of software-based image analysis. Kerby et al. constructed a panorama image of the lamina in the sagittal plane using an image matching algorithm, based on the sum squared of absolute differences of pairs of images, identifying vertebral levels in the constructed panorama [31]. Tiouririne et al. studied the performance of a commercial computer-aided detection tool for lumbar anatomy, but the device doesn’t specifically identify vertebral level [57]. Yu et al. and Leng et al. have done work to automatically identify the vertebral level in both the sagittal plane as well as the transverse plane of US. They did so with a variety of techniques that involve hand-crafted feature extraction, as well as support vector machine (SVM) ML applied to those features [39, 65–68]. Their work, which has a similar focus as this thesis, has achieved high success rates on their data with their various methodologies. However, their use of hand-crafted feature extraction, with techniques such as template matching, depend on accurate calibration of matching thresholds for a
given dataset and are not robust to substantial anatomical variability in the population. Therefore, further investigation into systems that can accommodate a range of anatomical variability is warranted, which this thesis aims to address.

2.4 Machine learning in ultrasound

The use of ML in medical images has become very prominent. ML can be used in different ways to achieve various goals, with clinical relevance. This section aims to cover the fundamentals of ML with regards to image classification, segmentation, localization, and transfer learning. This section will also discuss the advancements made by other groups in the field. Although this section focuses on the application of ML to US, a current survey of deep learning in medical imaging is given by Litjens et al., which can be consulted for an extensive list of relevant works [40].

2.4.1 Classification

Image classification is a well defined problem. Given an image and a set of possible classes, the goal is to choose which class the image fits into, ideally with a probability or likelihood estimation. One of the earliest examples of ML applied to image classification is the recognition of hand-written zip code digits. In the work initiated by Denker et al. [13] and followed up by Le Cun et al. [37], hand-written digits were classified by neural networks with error rates as low as 1%, along with the ability to quantify the confidence in the classification.

Neural networks are data processing structures, which take input $x_n$ in $\mathbb{R}^N$ and map to the output $\hat{y}_m$ in $\mathbb{R}^M$ through a function, $f$. In image classification, $\hat{y}_m$ corresponds to the likelihood for each of $M$ classes. The function to calculate $\hat{y}_m$ can be simplified as $\hat{y}_m = f(Wx_n + b)$. Here $W$ is an $M \times N$ matrix of network parameters and $b$ is a bias vector. Neural networks can be composed of multiple layers, where the information is processed layer-by-layer from the input, $x_n$, to the output, $\hat{y}_n$. One configuration is a fully-connected neural network and is exemplified in Figure 2.9. As data flows from the input of the network through subsequent layers, salient features are extracted and result in the output, $\hat{y}_n$.

Neural networks are trained by optimizing parameters of $f$, $W$, with respect to a set of training data $X$, with ground truth labels $Y$. A loss function is computed,
often by SVM or softmax, comparing the output $\hat{Y}$, to $Y$. One method for optimizing the parameters in $f$, is performed by minimizing the loss in the training data via iterative back-propagation of gradients in parameters. There are many nuances in this extensive field of research, but one important point is that training networks on a set of data is prone to overfitting. Overfitting occurs when a network becomes optimized for the training data set, but performs significantly worse on unseen test data. One common solution is the use of regularization, which penalizes parameters of $f$ from becoming too specific to the training dataset.

Le Cun et al.’s work was one of the first to start utilizing back-propagation to train network parameters [37], and in 2012, Krizhevsky revolutionized image classification by utilizing the power of modern graphics processing unit (GPU) computing technology along with CNN architectures, outperforming all other image classification systems before that time [34]. Since then, advancements of neural networks have been growing dramatically.

CNNs are a variant of neural networks which were initially inspired by an inves-
Figure 2.10: Example convolutional neural network architecture.

tigation into the visual cortex by Hubel et al. [28], and focus on analyzing images. CNNs extract low, mid, and high-level features through ascending layers of the network [69]. Currently, the CNN is the most accurate ML model in image classification problems [40]. In convolutional neural networks, the number of parameters are reduced from the fully-connected neural networks due to parameter sharing via convolutional kernels. An example CNN architecture is shown in Figure 2.10. Many modern image classification CNN architectures follow similar architectural schemas to this mock network.

This common schema is composed of the following: an input image, followed by a repeating series of convolutional layers, often paired with a non-linear function and pooling function applied to the output. The final layers are conventionally a series of fully-connected layers, with the final output layer being equal to the number of classes in a given data set. Much like a conventional neural network, this structure results in salient features being extracted from the input data, contributing to the final results. However, CNNs are intended for image data, where spatial relations in the input data must be taken into account.

Image classification appears frequently in modern medical image analysis literature. In one work with multilayered perceptron neural networks, Chen et al. diagnosed breast tumors in US [9]. In work on paramedian spinal US, Pesteie et al. develop a real-time US image classification network [47]. This system is used within Chapter 3 to aid in image acquisition. Another example of classification in US involves the plane localization in fetal US via a deep CNN [11].
2.4.2 Segmentation and localization

Segmentation and localization are two other image analysis ML problems, which have many uses in medical image analysis. In segmentation, the pixels (in a 2D image) or voxels (in a 3D image) corresponding to a target object should be identified. The result will include a contour around one or more target objects. Similarly in localization, the general location of a target object in an image should be identified. Conventionally, this involves course segmentation and can result in bounding boxes, or centroid of an object. Certain ML architectures can be used to solve both of these problems, and have similar fundamental structures as in the classification architecture described in Section 2.4.1. In the context of US, image segmentation can be used to solve problems such as anatomy segmentation or anatomy identification via localization.

Milletari et al. use a neural network to segment various anatomies in 3D US data using a hough forest on the learned features for classification [43, 44]. They showcase their technique on segmentation of the midbrain, left ventricle of the heart, and the prostate. In another segmentation work, Qinghua et al. use a CNN architecture to automatically segment breast lesions in 2D US images [27]. The current trend is to use CNNs for most image-based tasks, as they are designed specifically for processing images [40].

2.4.3 Transfer learning

CNNs are rapidly improving in image classification. This is largely due to the improvements in hardware, the availability of larger data sets and bigger models, and new ideas and algorithms [55]. One problem with deep CNNs is that they are composed of a very large number of parameters that must be learned from training images. It takes many images to train a network’s parameters to a useful state, so it is challenging to optimally train deep CNNs with small data sets [45]. Transfer learning can be used to help train CNNs on smaller data sets [45, 69]. Transfer learning takes knowledge learned from one domain and applies it in another domain. Oquab et al. [45] and Zeiler et al. [69] showed that transfer learning can make use of pre-trained CNNs to solve classification tasks in a new domain. A network is first trained on one dataset and the learned parameters are utilized in
training a network for a different problem with a different dataset. As previously mentioned, trained CNNs extract low, mid, and high-level features through ascending layers of the network. These features can be used for various tasks such as classification or segmentation [69]. Networks that have been trained on millions of images have optimized their parameters for visual recognition tasks, and can transfer that knowledge to other related tasks. Transfer learning re-uses the parameters that were optimized in lower layers of a network to extract features in another context. To make use of the transferred knowledge, the final layers of the network can be retrained entirely with the task specific dataset. Optionally, the lower layers of the network can be fine-tuned by allowing their parameters to update, but with a much lower learning rate than the final layers which are being retrained. Work by Chen et al. in plane localization in fetal US uses transfer learning, showcasing its applicability in the context of medical imaging [11]. Kong et al. additionally emphasized the success of transfer learning in medical imaging by utilizing the technique to identify end-diastole and end-systole frames in US images sequences [32].

ML can be used as a tool to build models based on available datasets, giving their parameters statistical meaning, and making them capable of identifying and handling non-hand-crafted generic features. In this thesis, neural networks, CNNs, classification, localization, and transfer learning are utilized in the development of LIT and SLIDE.

2.5 Summary

Spinal analgesia, specifically in obstetrics, is a challenging procedure. Approximately 60% of all labouring women in the United States opt to receive epidurals to deal with the immense pain of birth [46], resulting in millions of annual procedures. The purpose is to reduce pain via analgesic agents, which requires the successful delivery of medication at the correct location. Among other considerations, it is important that physicians choose the correct vertebral level for needle insertion. This mitigates possible complications with the procedure.

The current clinical practice for identifying the vertebral level in obstetrics is predominantly manual palpation, without the aid of medical imaging (and so is
considered “blind”). The use of US has been shown to improve the accuracy of vertebral level identification \[19\]. However, it is not commonly used for a number of reasons, including the challenge of interpreting images filled with speckle and artifacts.

Many groups have developed solutions to address the problem of US not being utilized in aiding spinal needle procedures. There are still many reasons that these systems have not yet become the clinical standard, including trade-offs between technological improvements and overhead monetary and workflow costs of necessary setup, such as external tracking hardware.

Some groups have investigated entirely software based solutions to US analysis in this problem space. Such systems do not rely on external tracking hardware, but have other shortcomings. The work in this space relies on techniques that involve hand-crafted feature extraction. The work in this thesis aims to develop and clinically test software based solutions for automated vertebral level identification, via the application of ML techniques for US image analysis. The goal of this work is to aid pre-puncture planning for percutaneous spinal needle procedures, especially in obstetric spinal analgesia.
Chapter 3

LIT: Lamina Identification and Tracking using deep networks with unsupervised feature learning

3.1 Introduction

As described in Chapter 2, anesthesiologists need to identify the lumbar vertebral levels of their patients. The clinical practice of manual palpation alone has shown the accuracy of a physician in identifying the L3-L4 intervertebral gap, which is the injection site for epidural analgesia, is only 30% [19]. The use of medical imaging has been shown to improve pre-puncture target level identification. As fluoroscopy is prohibited due to ionizing radiation, US is an adequate alternative because it provides anatomical information to the operator, not obtainable by manual palpation alone[6, 54], without ionizing radiation.

Much work has already gone into the development of solutions for vertebral visualization and level identification via US technologies. But, the current proposed solutions either make use of external tracking hardware, or rely heavily on hand-crafted computer vision techniques. Current non-tracked systems that rely on
hand-crafted features, such as with template matching [39, 65–68] are not robust to substantial anatomical variability in the population. Template matching also depends on accurate calibration of matching thresholds for a given dataset. Therefore, there is a need for a purely 2D US solution for vertebral level identification with no change to the hardware, working directly on the US imaging sequence, which can accommodate a range of anatomical variability. The work in this chapter aims to automate the process of identifying and labeling vertebrae, while tracking them throughout a sequence of US images. This system works on paramedian US images, with the transducer starting from the sacrum and moving cephalad.

In this chapter, an overview is provided for the development of a system, LIT, which makes the following contributions: 1) a deep network architecture is proposed to detect the vertebrae in standard paramedian 2D US images of the lumbar spine. The proposed network has been trained on a data set acquired from 15 volunteers and the experimental results have been evaluated against an expert sonographer. 2) An algorithm is implemented to label and track the respective vertebrae in the 2D US image sequences. By using statistical modeling based on deep learning, compared to conventional approaches such as template matching [68], our approach can inherently accommodate a range of anatomical variations in patients.

3.2 Methods and materials

3.2.1 Data acquisition

LIT’s data flow is shown in Figure 3.1. US images were acquired from 15 participants using informed written consent from all subjects, after approval of the study by the Clinical Review Ethics Board (Certificate H15-01310). Patients were in the prone position during data collection to provide consistent probe contact. The US data was generated with a C5-2/60 curvilinear transducer, at 7 cm depth and 3.3 MHz, with a SonixTouch machine (Analogic Corp., Richmond BC, Canada). This produced images with 616 × 820 pixels with geometrically determined axial and lateral scales of 0.14 mm and 0.18 mm per pixel respectively. For all of the participants, 7 cm depth was sufficient to ensure visibility of important landmarks including the anterior lamina reflections as well as the deeper anterior com-
Figure 3.1: Data flow of LIT. Original US images are augmented with the bone enhanced images by a weighted sum, and downsized to $83 \times 110$. A $27 \times 27$ sliding window generates image patches for pixel-level classification. Unsupervised features are extracted from the image patch by two sparse auto-encoders which were previously trained, and classified by a fully-connected layer, to identify the vertebrae in the image. Each vertebra is then labeled with its respective level and is tracked in the sequential images.

plex (AC) reflections. These landmarks are expected to be present in good quality paramedian US images, and the lamina is necessary to be present as it is what will be identified and tracked for level identification.

The path of US data collection starts with the transducer in the paramedian plane with the L5-S1 intervertebral gap approximately centered in the image, as shown in Figure 3.2. The transducer is then moved cephalad along the paramedian plane until the L1 is visible. US of the spine in the paramedian plane show many anatomies, including the laminae shown in Figure 3.3, which have a characteristic wave-like pattern that is used for identification. Each of the US recordings contain between 500 and 1500 frames. US data was recorded with this known path on the left and right side of 15 participant’s spines, resulting in 30 data sets. The data was acquired by a novice sonographer, but was aided by a real-time lamina-detection tool to help ensure the transducer was kept in a paramedian plane, with lamina constantly visible [47]. The use of this tool aided consistent paramedian data collection, when acquired by a novice.
3.2.2 Ultrasound bone enhancement

Low contrast, speckle, and shadowing have introduced problems in related US research [68]. To counteract these issues, multiple techniques have been proposed. One such technique is to enhance bone surfaces in the US images. This technique enhances features near the targets of interest, which in this work are the reflections of the laminae. A bone enhancement algorithm, as proposed by Anas et al., was used to augment the input to the neural network with higher emphasis on visible laminae [2]. The enhancement is done by estimating the location of bone through a combination of phase symmetry and shadow information. Phase symmetry is expected to have a high magnitude at tissue boundaries [2]. Building on top of Foroughi’s bone enhancement model, which searched for shadows in images, the phase symmetry map was multiplied by an estimated shadow map, to output a bone response estimation [2] [16].

As indicated by Figure 3.1 in LIT, the first step of data processing is to augment the raw US images by the bone estimation image, because the lamina bone is the target anatomy in the image. A weighted sum of the raw US image and the
Figure 3.3: (a) Collection path on the left and right paramedian planes of all participants, moving cephalid from the L5-S1 intervertebral gap up to the L1 vertebrae. (b) The paramedian imaging plane intersects the vertebrae as indicated by the purple plane. Artist: Vickie Earle, used with permission. (c) Exemplary US image in the paramedian plane showing the wave-like pattern of laminae. In this image, the reflection of four lamina are visible.

Bone estimation image is calculated, as shown in Figure 3.4. An empirically determined ratio of 70% to 30% between the enhanced image and the raw image are combined. Data from the raw image is used to account for potential errors in the bone enhancement algorithm, and to provide non-sparse images into subsequent subsections of LIT. This new bone enhanced image is used for further processing.

3.2.3 Vertebrae identification

Vertebral level identification involves accurate detection of vertebrae in US images via discriminative anatomical features, which characterize typical vertebra. As shown in Figure 3.2 and Figure 3.3 in the paramedian plane, the posterior position of the lamina creates US echoes that appear as a subtle curve in the original and bone enhanced images. A learning algorithm can be used to identify the peak of the curve and accordingly distinguish anterior peaks of vertebral lamina from other anatomy. Since the peaks of the laminae are distinct anatomical features, they consistently appear in paramedian US images of the subjects. Therefore, obtaining
Figure 3.4: LIT bone enhanced image. A weighted sum of the initial raw B-mode image and a bone response estimation.

A feature set in an unsupervised fashion that can yield a generic representation of the target is beneficial. In order to automatically extract the features of the posterior peaks of the laminae, a deep network architecture is utilized for pixel-level classification. The input image is downsized from $616 \times 820$ to $83 \times 110$, to ease computational cost. This downsizing was determined empirically to ensure images could be processed on the order of minutes.

Sparse auto-encoder

Two layers of the network, as shown in Figure 3.1, are based on sparse auto-encoder (SAE). SAE is a specific configuration of a neural network that results in unsupervised learning. The algorithm aims to minimize the reconstruction error from inputs $x_n$ to outputs $\hat{x}_n$ [33]. Given an input sample $x_n$ in $\mathbb{R}^N$, the encoder would map it to a hidden activation, called encoding $h_m$ in $\mathbb{R}^M$. It can be visualized in Figure 3.5 and mathematically defined by the following function

$$h_m = f_E (W_E x_n + b_E), \quad (3.1)$$

where $f_E(z)$ is the encoder’s nonlinear activation function, and $W_E$ is an $M \times N$ encoder weight matrix and $b_E$ is the encoder’s bias vector. At the decoding stage, the SAE reconstructs the input data from the latter hidden activations using a linear
Figure 3.5: Example SAE configuration. $x$ and $\hat{x}$ both have the same dimensionality, which is higher than the encoded dimensionality of the hidden layer, $h$.

or nonlinear activation function:

$$\hat{x}_n = f_D(W_Dh + b_D), \quad (3.2)$$

where similarly, $f_D(z)$ is decoder’s activation function, $W_D$ is the $N \times M$ decoder weight matrix and $b_D$ is the bias vector. Finally, $\hat{x}_n$ is the reconstructed input. During the training stage, the SAE applies back-propagation to train the reconstructed values to equal the input. With $M < N$, the SAE is forced to obtain a nonlinear compressed representation of the input, which can be considered as the feature vector.

The SAE model enables one to specify that only a certain percent of the hidden neurons have a high probability to be activated by adding a penalization term to the training loss function [21]. Therefore, the reconstruction loss function can be
defined as:

\[
E = \frac{1}{K} \sum_{i=1}^{K} (x_n^i - \hat{x}_n^i)^2 + \lambda \|W\|_2^2 + \beta \sum_{j=1}^{M} \text{KL}(\rho \| \hat{\rho}_j ) ,
\]

(3.3)

where \(K\) is the number of training samples, \(W\) is the encoder and decoder weight matrix, and \(\lambda\) and \(\beta\) are the \(\ell_2\) norm and sparsity regularization parameters, respectively. \(\text{KL}(\rho \| \hat{\rho}_j )\) denotes the Kullback-Leibler divergence, where \(\hat{\rho}_j\) is the average activation of hidden unit \(j\), and \(\rho\) is its desired value.

**Hierarchical feature learning**

Multiple layers of encoders can be hierarchically stacked to obtain a representation of the data at multiple levels of abstraction, resulting in a deep network. This can be achieved by a greedy layer-wise training strategy [1], which first involves training an SAE with 512 hidden neurons on the original images to transform the input \(x_n\) to \(h_m^{(1)}\) that consists of the activations of the hidden layer. Then, a second SAE with 256 hidden neurons is trained on \(h_m^{(1)}\) to produce a new activation vector \(h_l^{(2)}\) from \(h_m^{(1)}\). Note that the dimension of the feature space gradually decreases with the increase in the abstraction level. The number of hidden neurons were chosen to be powers of 2, and empirically selected as to greatly reduce the dimensionality (from 9130 to 512 to 256), while still converging upon training. Since SAEs are trained on the augmented US images, which contain the enhanced bone surfaces, the extracted top level representation yields distinctive feature sets that separate the target lamina images from other regions. This is required for uniquely identifying and separating vertebrae in paramedian US images, to then be tracked.

Finally, after training the SAEs, a fully-connected layer is trained on the top-level representations, i.e. \(h_l^{(2)}\). The fully-connected layer is trained against the binary labels of the patches indicating whether they contain lamina peaks or not. Patches were sized to either contain 0 or 1 lamina peaks. In the proposed architecture, the fully-connected layer contains 100 neurons.

The \(\ell_2\) and sparsity regularization parameters for both of the SAEs are \(\lambda = 0.001\) and \(\beta = 1\), respectively. Also, the desired average activation value, \(\rho\), is set to 0.05 for both SAEs. These are default values when generating an SAE in Matlab.
which were kept as empirically determined values that resulted in convergence.

**Generating data set and training**

The proposed deep network identifies the peaks of the laminae of vertebrae through pixel-level classification of the augmented US image. Therefore, the training data should include labeled pixels from all regions in the image. For each pixel of the image, \( p \), the training sample consists of a patch, where \( p \) is located at the center of the patch. Patch size should be chosen to contain the vertebrae. The patch size is set to span the length of the average posterior vertebral height and spacing, 34 mm, in order to captures the anatomic information of the lamina in [20]. Based on the resolution of the down sized image, the patch size is set to \( 27 \times 27 \) pixels.

To generate the training data set, the locations of the peaks of laminae, as the anatomical landmarks, were annotated. Pixels within a \( 15 \times 15 \) bounding box were also given positive labels to generate multiple positive pixel samples per vertebra peak, set to approximately half the width of the patch size. Outside of this bounding box, equal numbers of negative samples were selected from uniformly random locations. The balanced image patches corresponding to the target and non-target labels were used for unsupervised training of the SAEs. Without balancing, the negative samples would heavily outweigh the positive samples and likely lead to training a network that always predicts negative results. The respective annotations were used to train the fully-connected neural network with 100 neurons in the hidden layer. In total, 43,586 image patches from three subjects were used for training the network. The proposed network was trained, validated and tested on 70%, 15% and 15% of the data set, respectively.

### 3.2.4 Vertebrae level labeling and tracking

The output of the network is a binary pixel-level classification map, which predicts the location of lamina peaks. To generate final coordinate predictions, referred to as responses, the classification map is post-processed by dilation to fill holes, and erosion to remove small false-positives and help separate predictions. The centroids of remaining structures are obtained.

A tracking algorithm is used to assign vertebral levels to the responses of the
vertebrae identification subsystem, and track them in sequential US frames. The algorithm matches responses based on horizontal distances between points (the distance in the x-coordinate in the US frame). This is a bipartite matching problem between points in each subsequent frame [17]. Pseudo code in Algorithm 1 describes how updates are made. Each series of connected points are referred to as a track. $T_n$ represents all tracks in frame $n$. $T_n$ is composed of individual tracks, $t^n_i \forall i$ in $\mathbb{R}^I$, where $I$ is the number of tracks at frame $n$. $R_n$ is the set of all responses, $r^n_j$, from the vertebrae identification subsystem. At time $n$, all $r^n_j$ are used to update $t^n_{i-1}$, with a Euclidean distance function, $d$, that includes a forward-motion bias factor, $\gamma \geq 1$, since the US procedure is intended to move uni-directionally. The bias factor reduces incorrect matches if the transducer jitters and to reduce the effect of incorrectly identified responses. This is done iteratively, where matches from tracks, $T_{n-1}$, to $R_n$, are done $\min(|T_{n-1}|, |R_n|)$ times, as per Algorithm 1.

**Algorithm 1:** Sequential frame matching between tracks in the previous frame, $T_{n-1}$, and new responses, $R_n$. 

\[
(i,j)_{\text{match}} = \arg\min_{i,j} (d(t^n_{i-1}, r^n_j)), \forall t^n_{i-1} \in \mathbb{R}^I, \forall r^n_j \in \mathbb{R}^J \quad (3.4)
\]

\[
d(x_1, x_2) = \begin{cases} 
 x_2 - x_1, & \text{if } x_2 \geq x_1, \\
 \gamma^* (x_1 - x_2), & \text{otherwise.} 
\end{cases} \quad (3.5)
\]

On each frame, the algorithm has three steps: 1) link tracks to responses; 2) spawn new tracks for unmatched responses; and 3) prune any tracks that have no
responses for a given time frame [17].

Some restrictions are placed on the scanning protocol to provide guidelines for parameter selection in LIT. The anatomical knowledge and clinical protocol that aid this algorithm are:

- The initial frames contain L5 in the superior half of the image due to the initial transducer placement constraint. When LIT starts processing data, it labels the first identified lamina superior to the center of the image as the L5.

- Posterior vertebral height and inter-vertebral spacing, together, has a minimum of 27 mm, as shown by previous research [20]. This value guided the pixel-patch size for pixel-level classification.

- Inter-vertebral distances remain unchanged since the patient is still, and therefore vertebra should not move with respect to one another. This indicates that all movement of lamina in subsequent images is caused by the transducer motion.

- The sonographer moves the transducer below 6 cm/s, while remaining in the paramedian plane to view the lamina. This value was chosen to restrict the distance lamina could travel between subsequent frames. This empirical value restricts motion to approximately half the span of the average posterior vertebral height and spacing (15 mm) per 10 frames [20].

Finally, the tracks are assigned to vertebral levels, starting at L5 and being enumerated in the superior direction as vertebrae are detected. Tracks can spawn and be pruned multiple times, giving the tracking subsystem the ability to recover even if it loses track of a given vertebra for a period of time. Examples of the output of LIT are shown in Figure 3.6.

**Labeling**

Cartesian coordinate location and vertebral level labels of sequential US frames were manually identified for all 30 data sets, in order to compare against system outputs. Additionally, 10 images from each data set were labeled by an expert sonographer, to characterize the performance of the vertebra identification subsystem when run on all 30 participants.
Figure 3.6: Horizontal position of lamina peaks vs frame number, with labels and LIT outputs compared. Examples of lamina tracking successes in (a) vs. certain level mismatches shown in (b) and (c). (a) Coloured points indicate the outputs of LIT, and black “X”s are the manually identified peak position labels. The first label in time, that has the highest position (y-axis) corresponds to L5, and the next label corresponds to L4 and so on. Label levels always correspond to the closest label in the next frame (the closest “X”). (b) An off-by-one tracking level mismatch, where all system outputs track labels, but indicate the wrong level because the L5 was missed (indicated by the grouping of “X”s on the left without matching LIT outputs). Low contrast in many initial frames caused L5 to be missed. (c) An example of two level mismatches (red points jump from L4 labels to L3 labels, but converge back to L4 labels).
Figure 3.7: Confusion matrix for the classification of test image patches from paramedian US as either lamina (0) or not lamina (1).

3.3 Results

The supervised training of the lamina peak identifier is performed on six of the available 30 data sets. It produced a sensitivity of 96%, a precision of 96%, a specificity of 96%, and a negative predictive value of 96% when run on all training, validation, and testing data. Identification testing was performed against all 30 data sets in another experiment. An expert sonographer labeled a random selection of frames from each data set, skipping frames that do not depict the lamina. The same testing procedure for the supervised training was repeated, where an equal number of pixels surrounding positive labels, and pixels randomly selected elsewhere in the US image were tested by the identification network. The network produced a sensitivity of 95%, a precision of 95%, a specificity of 95%, and a negative predictive value of 95%, as shown in Figure 3.7. This shows generality in the trained network, since the data has not been seen during training.

It is desirable to have the output predicted tracks nearest to the labels with the
same vertebral level for each frame. The error of level labeling is evaluated by the percentage of level mismatches in all data sets.

Labels include a vertebral level and are matched to a track within a frame, matching if the distance is within half the minimum posterior inter-vertebral distance, 14 mm [20]. Level mismatches are defined as the divergence from one matching of labels and tracks to another, for a pair of subsequent frames. For example, an analyzed data set with two mismatches is shown in Figure 3.6 (c). Initially, L5 and L4 predictions are correctly matched to their respective levels (as specified by the labels). However, the L4 track clearly deviates away from the L4 label, resulting in the predicted L5 and L4 locations actually corresponding to the L5 and L3 locations. This is one mismatch. LIT later recovers, when L5, L4, and L3 predictions correctly match to their respective levels again. However, by definition, this is considered to be a second mismatch.

Accordingly, the average number of level mismatches is 11, in an average 164 frames analyzed in each data set. Between pairs of subsequently analyzed frames, the matching of vertebral level labels was correct in 94% of cases, when compared to matches of manually selected labels. Figure 3.6 shows an example of LIT’s outputs with no mismatches and two different cases of mismatches.

### 3.4 Discussion and conclusion

A vertebra identification, labeling, and tracking system for US images of the lumbar vertebrae is proposed, called LIT. LIT includes a learning algorithm for identification, followed by a detection-based tracking subsystem. Compared to other similar systems that analyze data after an entire sweep of the lumbar spine is done, this system tracks frame by frame and is capable of recovering from errors.

This work exemplifies the ability of an unsupervised training algorithm to be applied to US, to address a clinical need. Features were able to be learned from the available data and applied to unseen data. Pixel-level segmentation via patch based lamina classification performed with 95% sensitivity and 95% precision when evaluated on test data annotated by an expert sonographer. This indicates that a deep network with features learned from unsupervised-training can be useful in spinal US. The vertebral level dynamic matching success rate is 94%, which indicates
that the tracking algorithm is suitable for this application. However, because the algorithm allows for labels to be lost and later recovered and LIT is not real-time, no direct metric of vertebral level identification can be drawn from this work on the available off-line data. In similar work, Yu et al. use template matching and frame stitching to identify vertebral levels in paramedian images, reporting 97% success in their analysis [68]. A real-time implementation of LIT, compared to a gold standard, would be necessary to produce a directly comparable metric. The approximate speed of each section of LIT is on the order of minutes for processing each frame. The bone enhancement and lamina localization, implemented in Matlab, take approximately 550 ms and 280 sec respectively. The inter-frame tracking, implemented in C++, takes approximately 15 ms. The lamina localization was able to be sped up by adding a stride to the pixel-level classifier. A stride of 4, empirically chosen, caused a speed up of $16 \times$ in these analyses.

There exist other shortcomings of this work, which motivate potential future work. The current implementation does not perform at real time. Optimizations via GPU processing would be beneficial for this architecture. Exploration into related deep learning architectures for lamina vertebral response identification could potentially result in more optimized methods. For example, end-to-end deep networks, which segment an entire image in one forward pass of a network, may show faster performance metrics. Additionally, this work is based on an US scanning protocol that involves two steps: first identifying the vertebral level in the paramedian plane, and then rotating the transducer into the transverse plane to find the patient’s midline. This scanning protocol can be simplified to keep the transducer entirely in the transverse plane. These limitations are addressed in Chapters 4 and 5 via other forms of ML applied to US images to automatically identify vertebral levels. SLIDE is developed to achieve real-time performance, and be able to analyze data off-line and in a clinical setting to evaluate true level identification accuracy.
Chapter 4

SLIDE: automatic Spine Level IDEntification system using a deep convolutional neural network

4.1 Introduction
In this chapter, an overview is provided for the development of a second vertebral level identification and labeling system. The work from Chapter 3 exemplified the potential success of deep networks for feature extraction in spinal US. But, LIT is not currently capable of processing images in real time, with high processing demands from its pre-processing steps and pixel-level classification schema. A real-time solution is necessary for clinical evaluation. When further comparing to LIT, the clinical imaging protocol for pre-puncture target identification could be improved if the transducer could remain in the transverse plane throughout the entire scan. Conventionally if US is used, it is first used in the paramedian plane for level identification and then later used in the transverse plane for midline identification. This work simplifies the scanning protocol to remove the need for scanning in the paramedian plane. Like LIT, this system identifies vertebral levels with-
out using external tracking hardware, providing a solution to physicians with low overhead.

In this work, deep CNNs are used to automatically identify features in US images, as they are well suited for image classification tasks. The proposed system is referred to as SLIDE: the Spine Level IDEntification system. Development is focused on the transverse view of the spine because it enables both level identification and puncture site localization without the need for a rotation of the imaging plane. Hence this scan path simplifies the clinical workflow. Automatic level identification of vertebral level in both the transverse and paramedian views has been explored by Yu et al. [65–68] using either a template matching approach or by classification of hand-crafted features from the US data using a SVM. The feature extraction process in this work is proposed to be automated by taking advantage of the statistical modeling power of CNNs. Using CNNs also has the potential to accommodate a range of anatomical variations in patients, which cannot be easily captured by using hand-crafted features. Kwitt et al. also proposed generic solutions for target localization in US, using dynamic texture models [35]. However, SLIDE’s use of a priori knowledge, combined with statistical classification allows for multiple spine level localizations.

This work makes the following contributions: 1) a framework is proposed, using a deep CNN, for automatic classification of transverse spinal US planes. The framework uses a state-machine, based on a priori knowledge of the clinical imaging protocol, to identify the vertebral level. 2) A real-time implementation of the framework is demonstrated. To facilitate workflow, an AR interface is also proposed. A small projector is mounted to the transducer to project the detected vertebral level onto the patient’s skin adjacent to the transducer.

4.2 Methods and materials

4.2.1 System overview

SLIDE is a real-time system for identifying the lumbar level of patients. The anatomy and representative US images are shown in Figure 4.1. SLIDE is implemented as an extension of 3D Slicer [15] and provides both a desktop display as
Figure 4.1: Anatomy and representative US images of the lower vertebrae. Overlaid on the lumbar spine is the desired path of a operator. The operator begins at the sacrum, with the US transducer oriented in the transverse plane, and moves cephalad up to the L1 vertebra. The associated US images on the right show typical images from three distinct planes. Modified work from artist: Vickie Earle, used with permission.

well as an AR display for an operator as shown in Figure 4.2. The system architecture is shown in Figure 4.3, which shows the data flow within SLIDE as well as the hardware interactions. A secondary computer along with a projector are used in the current architecture. It is possible to integrate SLIDE into an US machine that has a GPU, thereby requiring no additional computing hardware.

An operator must start scanning a patient’s midline in the transverse plane until the sacrum is automatically detected. SLIDE can discriminate between three characteristic transverse planes of the lower spine; the sacrum, intervertebral gaps (“gap”), and vertebral bones (“bone”), which are shown in Figure 4.1. Once SLIDE automatically detects the sacrum, the operator follows a cephalad path up to the L1 vertebra, while keeping the transducer near the patient’s midline. In doing so, the transducer will go from the patient’s sacrum to their L5-S1 gap, to their L5 vertebra, and so on. By automatically classifying and processing the sequential US
Figure 4.2: SLIDE displays. The left figure shows the desktop display integrated into 3D Slicer, indicating the current predicted level and current predicted plane, along with the associated confidence from the deep network output (showing the gap classification in green with >99% confidence). The right figure shows the AR display projected onto the patient’s skin (enlarged for visibility in the figure). In this diagram, the predicted level, which is being projected, matches the gold standard markings on the participant’s skin.

data, SLIDE is able to detect transitions between sacrum, intervertebral-gap, and vertebral bone planes thereby identifying the vertebral level. By using a transverse midline imaging plane, the puncture site at the midline (seen as left-right image symmetry) can be accurately determined with a single scan.

4.2.2 Data acquisition

US images were acquired from 20 participants with informed written consent from all subjects, after approval of the study by the Clinical Research Ethics Board (Certificate H15-01310). The scans were done by an expert sonographer. The US data sets were obtained with a C5-2/60 curvilinear transducer and a SonixTouch machine (Analogic Corp., Richmond BC, Canada). The machine settings were adjusted for each subject to ensure the depth included the echo of the AC near the bottom of the frame. This is done because the AC is a prevalent feature in intervertebral gap images that only appears if the US echo travels between spinous processes. Data were acquired from subjects by scanning their lumbosacral vertebrae in a number of ways in order to collect data for both classification training
as well as full-system level identification testing. All data were recorded in the transverse plane, maintaining the anatomical midline as close to the center of the image as possible, with the transducer approximately perpendicular to the skin. For plane classification training and validation, recordings of greater than 100 sequential US frames were obtained in each of the subject’s vertebral planes of interest: their sacrum, and all intervertebral gaps and vertebral bones of their lower spine (sacrum, L5-S1, L5, L4-L5, up to L1). Exemplary images of these planes can be found in Figure 4.1.

Recordings of a continuous sweep from the sacrum up to the L1 vertebra of each participant were also obtained. These recordings were used for testing SLIDE’s full-system vertebral level identification, rather than plane classification. The test recordings contain on average 850 frames, dependent on the scan speed.

4.2.3 Plane classification

Transfer learning

As described in Chapter 2, CNNs are powerful tools for image classification tasks, and are an active topic of research in the computer vision community. SLIDE applies transfer learning from pre-trained CNNs to the task of transverse spinal US classification, due to its limited number of training images. Transfer learning has shown success in taking knowledge learned from one domain and applying it to another, both in natural computer vision tasks, as well as with medical im-

Figure 4.3: System architecture, with data flow and hardware configuration.
Figure 4.4: Inter-frame NCC comparison between all pairs of frames in a given data set. The diagonal shows the NCC with a given frame compared to itself, having value of 1. A given row or column (mirrored along the diagonal) shows the NCC comparison between one frame and all others. A square wave in a row or column would be expected if such a strategy could identify planar transitions (which is not present).

ages [11, 32]. This is because networks that are trained on millions of images optimize their parameters for visual recognition tasks, and can transfer that knowledge to other tasks.

The success of transfer learning applied to transverse spinal US image recognition is analyzed in this work by fine-tuning four pre-trained CNNs: AlexNet [34], GoogLeNet [55], ResNet [24], and a parameter-reduced variant of AlexNet, SqueezeNet [29]. Microsoft’s deep residual network, ResNet, has multiple variants with various depths. Due to hardware memory limitations, this work implemented and analyzed the ResNet-50 variant. These networks were selected due to their success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Each of
the four networks were trained on 1.2 million ILSVRC training images containing 1000 classes of objects. As shown by Zeiler et al., transfer learning for deep CNNs results in superior training when applied to smaller data sets than fully training a network with that data set [69].

By re-initializing the parameters in the last layers of the network, specifically the fully-connected layers of AlexNet, GoogLeNet, and ResNet-50, and the final convolution layer of SqueezeNet, these models are trained to classify transverse spinal US images. Fully training the final layers and allowing the pre-trained lower layers to update with reduced learning rates is commonly referred to as fine-tuning a model. By fine-tuning a model, the ability of these deep networks to extract rich features with their pre-trained lower layers are utilized, but the domain of the classification task is changed.

Overall, using a statistical model for plane classification, such as a CNN, was additionally motivated by a preliminary rigidly registered inter-frame normalized cross-correlation (NCC) comparison done on each data set. This provided a qualitative assessment of the applicability of simple inter-frame correlation as a potential choice for plane classification. NCC is calculated between two US images, \( f \) and \( t \), with image means, \( \bar{f} \) and \( \bar{t} \), as follows:

\[
NCC(f, t) = \frac{\sum_{x,y}(f(x,y) - \bar{f})(t(x,y) - \bar{t})}{\sqrt{\sum_{x,y}(f(x,y) - \bar{f})^2 \sum_{x,y}(t(x,y) - \bar{t})^2}} \quad (4.1)
\]

Shown in Figure 4.4, no clear discriminative patterns were observed, which could be used to identify planar transitions. The NCC in all pairs of frames stays high immediately around any given frame, but drops off beyond that. This means that a simple method such as an NCC comparison of images in a scan is unsuitable for identifying planar transitions.

**Training**

Caffe is a deep learning framework that can be used for training CNNs [30]. Caffe makes use of the GPU to parallelize the training and deployment of a model, reducing computation time compared to central processing unit (CPU) implementations. Using Caffe, the pre-trained AlexNet, GoogLeNet, ResNet-50, and SqueezeNet
models were fine-tuned to classify transverse spinal US images. Training CNNs is based on the assumption that input data are independently and identically distributed, so that it is correct to assume that future data will behave like previous data. The training and validation data follow this guideline by balancing the number of frames in each class and by shuffling the data. As described in Section 4.2.2, labeled US data were obtained from 20 participants. For each participant, there is one recording of their sacrum, five for each of their intervertebral gaps (L5-S1 through L1-L2), and five for each of their vertebral bones (L5 through L1). To avoid training biases, the number of images from each participant of their sacrum, intervertebral gaps, and vertebral bones were balanced.

The data set is then set up for 20-fold cross-validation by a participant-wise leave-one-out strategy. The black boundary in the US image is removed by cropping the image to the maximum square that would fit inside the reconstructed B-mode US image. This removes black border pixels that do not contribute anatomical information. Images are additionally down sampled, and fit to the network dimensional requirements; 256×256, 227×227, 224×224, and 224×224 for AlexNet, SqueezeNet, GoogLeNet, and ResNet-50, respectively.

There are 8850 class-balanced and labeled US images from the 20 participants. Transverse images, centered on the mid-line of the lower spine, show anatomical symmetry. Due to the symmetry, the number of training data is doubled by augmenting with mirrored images, resulting in a net number of 17,700 images used in the 20-fold cross validation. No other augmentation was performed. Approximately 95% are used in each fold for training, and the remaining images for validation. With 20 datasets, leave-one-out cross-validation is appropriate for characterizing a network and its generality.

### 4.2.4 State machine for improved vertebral level identification

The root problem in identifying vertebral levels from a continuous cephalad scan of the lower back is accurately identifying transitions between planes. One challenge is that the true label of images taken at the boundary of an intervertebral gap and a vertebral bone can be ambiguous. A contributing factor to the ambiguity is US slice-thickness artifacts, which superimpose anatomical information from a slice
Figure 4.5: Transition frame classification ambiguity due to US slice-thickness artifacts. The AC and AP are expected in intervertebral gap images. The lamina is expected in vertebral bone images. In the transition image between gap and bone, the AC, AP, and lamina are present, which make its true class ambiguous.

with finite thickness onto the resulting 2D US image [22]. An example is shown in Figure 4.5, which shows echoes from the anterior complex and the articular processes, typical of intervertebral gap images, as well as bone echos from the lamina, typical of vertebral bone images, superimposed on the transition image. Additionally, SLIDE’s internal CNN can have stochastic classification errors. As such, there must be a method for handling the ambiguity at plane transitions, and simultaneously handling stochastic errors from SLIDE’s internal CNN. A vertebral plane identification state machine is developed to handle transitions and is shown in Figure 4.6.

Images from sequences are given probabilistic estimations of their class, $P$, via a Caffe softmax layer at the end of each network. The probability of the spinal classes are referred to as $P_{Sacrum}$, $P_{Bone}$, and $P_{Gap}$. As seen in Figure 4.7(a), $P$ is noisy, especially around transitions. The noise is caused by stochastic CNN errors, and ambiguous input images near plane transitions. Each component of $P$ is averaged over $N$ frames in order to reduce noise, represented as $P^*$ and shown in Figure 4.7(b). Using $P^*$ as input to the vertebral plane identification state machine results in the output shown in Figure 4.7(c), which showcases a successful example of identifying all vertebral levels using SLIDE. Figure 4.7(c) also shows a failing example when SLIDE’s state machine is not utilized, further described in
The state machine works by applying thresholds to $P^*$. At the start of a scan, $t_s$ is used as a threshold to ensure the sacrum is automatically found. Then, SLIDE must identify undulating transitions between the intervertebral gap and vertebral bones. The state machine must account for stability and instability within its input, $P^*$. It does so by using the threshold values, $t_{\text{stable}}$ and $t_{\text{unstable}}$.

$t_{\text{stable}}$ and $t_{\text{unstable}}$ identify transitions between undulating gap and bone planes, where $t_{\text{stable}} > t_{\text{unstable}}$. When $\text{abs}(P^*_\text{Gap} - P^*_\text{Bone}) > t_{\text{stable}}$, the system can enter a stable state, transitioning to whichever of the two planes has the higher probability. When $\text{abs}(P^*_\text{Gap} - P^*_\text{Bone}) < t_{\text{unstable}}$, the system enters an unstable state, transitioning away from the previously stable plane. When the system goes from a stable state to an unstable state, and back, there are two possible outcomes. In one outcome, the plane identified before and after the instability is the same, which indicates that two level transitions have actually occurred. In the second outcome, the planes identified before and after the instability are different, which indicates that only one transition actually occurred. All transitions in Figure 4.7 exemplify the second scenario.

SLIDE has 4 primary parameters; the averaging length, $N$, and 3 threshold values, $t_s$, $t_{\text{stable}}$, and $t_{\text{unstable}}$. $N$ should be chosen to be small, as it will add latency to SLIDE’s output, relative to the CNN output. In this implementation, $N = 30$, which averages over 0.75 seconds. $t_s$ should be set high to enforce high confidence in the initial identification of the sacrum. The imposition on the operator would be moving the transducer slowly when first identifying the sacrum. In SLIDE, $t_s = 0.8$ is used. A higher delta between $t_{\text{stable}}$ and $t_{\text{unstable}}$ give SLIDE resilience to noise, especially around gap and bone transitions. In SLIDE, $t_{\text{stable}} = 0.3$ and $t_{\text{unstable}} = 0.15$ are used, determined empirically. A discussion of the sensitivity of these values is given in Section 4.4.

If argmax $(P) = P_{\text{sacrum}}$ when SLIDE is already beyond the sacrum state, it resets into the unknown state. Such a condition is anatomically impossible with the required transducer path shown in figure 4.1. Resetting automatically allows an operator to repeatedly scan a patient while SLIDE is continually running.
Figure 4.6: Vertebral plane identification state machine. Averaged classification probabilities for the sacrum, gap, and bone planes are input into the state machine, which transitions between identified planes, thereby counting up a patient’s spine. Certain cases will revert the system into the Unknown state from any other state.

4.2.5 Real-time system

SLIDE is implemented with the following components: a SonixTouch machine with a C5-2/60 curvilinear transducer (Analogic Corp., Richmond BC, Canada); a Public software Library for UltraSound imaging research (PLUS) Server [36] on the US machine to send US images to a client over ethernet; a connected computer, acting as a client and running the real-time analysis software; and an optional PicoPro projector (Celluon Inc., Seoul, Korea) to enable an AR display. The client computer has a 3.4 GHz Intel i7-4770 CPU, 16 GB RAM, and an NVidia GeForce GTX 980 Ti graphics card.

On the client side, the display is implemented in two ways. The first display is as part of a module to the desktop application, 3D Slicer. The second display is as an AR projection of the current level directly onto the participant’s skin. The goal of using a projector in SLIDE is to allow the anesthesiologist to maintain their focus on the patient’s spine, keeping the transducer perpendicular to the skin, and directly seeing where landmarks are located. This could allow them to mark their target location easily. The PicoPro projector is mounted to the US transducer, projecting information onto the skin above the scan location, with 1920 × 720 pixel resolution. The mounting hardware to attach the projector to the US transducer was designed and 3D printed to ensure a rigid and repeatable mounting position, as shown in Figure 4.8.
Figure 4.7: (a) Exemplary output from the softmax layer of the CNN softmax layer, $P$. (b) Averaged values of $P$, $P^*$. (c) Various outputs, compared to the labeled level transitions. The SLIDE output shows a successful result, as there is overlap at all vertebral levels from the sacrum to L1. Latency between plane transitions in SLIDEs output versus the labels is due to a difference in protocol for identifying transition locations. With $t_{stable} = 0$ and $t_{unstable} = 0$, removing the utility of the thresholds in the state machine, a transition-error occurs at L5-Sacrum.
4.3 Results

4.3.1 Transfer learning training

Fine-tuning the pre-trained Caffe models on a three-plane classification task resulted in quick convergence, approaching zero loss within the first 100 iterations. All models were trained for 2000 iterations. The accuracy of training the models is shown in Table 4.1 with average accuracy from 88% to 91%. The 20-fold accuracy of each network was compared by an analysis of variance (ANOVA) test with significance level $p > 0.05$, showing comparable classification accuracy from all of the four models on the US data set.

To select a network suitable for deployment in a real-time system, the deployed image classification speeds were analyzed for each of the four CNN architectures. The timing tests were done with the Python implementation of Caffe. Results from classification of 250 images are shown in Table 4.1. An ANOVA test followed by a multiple-comparison test, both with $p < 0.01$, showed that SqueezeNet’s mean speed is significantly faster than the other three networks. Without impacting on accuracy and with significantly faster processing speed, the SqueezeNet CNN was selected for implementation in the real-time system, and system-level analysis. The
<table>
<thead>
<tr>
<th></th>
<th>AlexNet</th>
<th>GoogLeNet</th>
<th>ResNet − 50</th>
<th>SqueezeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>88.7 ± 9.1</td>
<td>88.4 ± 8.3</td>
<td>90.8 ± 6.8</td>
<td>87.8 ± 8.4</td>
</tr>
<tr>
<td>Timing (ms)</td>
<td>343 ± 10</td>
<td>488 ± 14</td>
<td>416 ± 6</td>
<td>44 ± 2</td>
</tr>
</tbody>
</table>

**Table 4.1:** CNN 20-fold cross validation accuracy and CNN classification speed.

details of the SqueezeNet architecture can be read from it’s initial publication [29], but is shown at a high level in Figure 4.9.

### 4.3.2 Vertebral level identification

As stated in Section 4.2.2, each of the 20 participants were scanned to collect data for training and validation purposes, as well as for full system testing purposes. The system-level test data is a continuous sweep from the sacrum up to the L1 lumbar vertebra, moving cephalad in the transverse plane. Again, using a leave-one-participant-out strategy, the `SLIDE` system using SqueezeNet model was used to test the full system vertebral level-identification success rate.

Level-identification success in a recording was defined as the occurrence of overlap in identified levels between manually selected level-labels and the predicted level-labels for all planes of interest in the sequence of images. 17 of the 20 test recordings have their entire lumbar vertebrae identified successfully. Of the three error cases, one has an off-by-one error (missing L5), and another has an off-by-two error (missing L4 and L3). The third error case failed to identify beyond the L5-S1 gap successfully. `SLIDE` was also run without the effect of thresholds, setting $t_{unstable}$ and $t_{stable}$ to 0, thereby simply determining state transitions based on the maximum CNN output. This was successful in only 13 of the 20 recordings, demonstrating the need for the state machine. To further evaluate `SLIDE`, the success at each plane within the data sets was evaluated. Errors were defined as missing a plane, or falsely transitioning away and back to a plane. For example, in Figure 4.7(c), the `SLIDE` Output has 0 transition-errors, but without the effect of thresholds, there is one transition-error at L5-S1. Within the 20 data sets, `SLIDE` has a 95%, 93%, and 93% success rate for sacrum, bone, and gap planes, respectively.
Figure 4.9: SqueezeNet CNN architecture. The left figure shows a high level 10 layer CNN. This architecture includes simple (direct) and complex (convolutional) bypass connections between layers. The right figure shows an expanded view of SqueezeNet’s Fire module. Numbers beneath a layer indicate that layer’s depth.
4.3.3 Real-time implementation

The proposed work was ported from Python to a C++ application and successfully implemented in a real-time setting. US data is processed at the maximum observed speed of the US machine; 40 frames per second (FPS). The implementation provides real-time feedback via two displays: 3D Slicer and a projected AR on the patient’s back. The projection is offset from the transducer to an unobstructed region of the back, cephalid of the transducer. The operator should be trained to understand that projected information corresponds to the plane of the transducer. The projector is laser-based, so it is always in focus. The projected text was clearly visible in normal office ambient lighting conditions, reflecting clearly even with the presence of US coupling gel. The predicted vertebral level of the transducer is projected onto the patient, but with the PicoPro, any other desired information could simultaneously be projected.

4.4 Discussion and conclusion

A real-time ML system is presented that successfully identifies lumbar vertebral levels in US images. SLIDE provides a display in both a desktop application and an AR display. It includes a deep learning framework for transverse plane identification, combined with an additional anatomical state machine, which is used for identifying vertebral level transitions of the US transducer during a scan.

Four deep networks were optimized on transverse lumbar US data, with transfer learning, and analyzed for accuracy and speed. The mean accuracy of each network after fine-tuning to the US data set had no statistical difference. SqueezeNet was significantly faster than the other three networks, and thus was used in the real-time implementation. The classification accuracy of SqueezeNet was 88% on 20-fold cross-validation.

 Compared to similar systems that incorporate tracking hardware, the proposed solution is less expensive. SLIDE can be used without any extra hardware, as the AR display is optional, and a GPU is now commonly embedded in standard US machines. Additionally, compared to other similar systems that extract pre-defined features, this proposed solution classifies images through entirely learned methods.

The need for such a system is shown by the high vertebral-level identification
error rates of “blind” palpation, which is only 30% accurate [19]. The comparable metric of this system on 20 test recordings show 85% accuracy. In one of the three error cases, SLIDE’s CNN failed to distinguish between gap and bone images with high enough probability to identify level transitions above L5-S1. In the other two cases, the errors seem to have been caused by vertebral transitions occurring too quickly. The errors could have been avoided by enforcing a slower scan speed during data collection, without changing SLIDE. To investigate this hypothesis, the two failing test recordings were augmented with repeated frames of the bones that SLIDE missed. With this augmentation, SLIDE successfully identified all lumbar vertebral levels in those two cases, bringing success up to 19 of 20 test recordings (95% accuracy). Requiring 21 augmented frames for correction gives insight into a potential maximum speed of the transducer. Estimating vertebral bone length to be > 1.5 cm [20], a maximum scan speed could be 3 cm/s.

Yu et al. reports near 100% accuracy in similar analysis using methods that include the extraction of pre-defined features, which are then trained by ML models such as SVM [65–67]. Their analysis focuses on the ability of their system to classify each individual vertebral plane in a recording, rather than localizing a given vertebral level from a recording. At the system level, Yu et al. identifies vertebral planes by counting consecutive plane classifications and comparing to a constant threshold [65–67]. This comparison makes the assumption of consistently correct classifications, and only characterizes the ability of the system in terms of local plane classification. Comparably, SLIDE identifies planar transitions from a recording by classifying transverse spinal US images and also handling errors in that classification. SLIDE does this with its internal state machine, thereby localizing any given lumbar level, while reducing the assumption that consecutive images must be classified correctly. Lastly, when distinguishing from Yu’s methods, the use of hand-crafted feature extraction suggests that Yu’s methods may not solve the classification problem when presented with anatomically varied images that do not include the pre-defined features. In contrast, CNNs are known to learn rich discriminative features and have out-performed hand-crafted features before [45].

The parameters in SLIDE’s state machine were determined via manual tuning. A sensitivity analysis was performed over the $t_{stable}$ and $t_{unstable}$ parameters, and it was found that a 25% change of the parameters caused a reduction of success.
from 85% to 70%. Another sensitivity analysis of the averaging length, $N$ showed a 20% change caused no reduction of the 85% success. In future work, the use of a recurrent neural network (RNN) will be explored as an alternative method for determining state transitions, to remove the need for manually identified parameters.

For SLIDE, training its CNN on more data may improve the results from 85%. Thus, it is desirable to incorporate data from more participants to improve the accuracy of the trained model. Data from pregnant patients is preferable. With the training methodology in this work, there is high correlation of images taken at each unique plane (11 planes from sacrum to L1), which indicates redundant training data. Being able to increase the breadth of the training data set could yield higher classification accuracy. Furthermore, integration of SLIDE into a standard anesthesiology workflow is desirable to be able to quantify results relative to the true levels in a larger clinical study. Both of these motivations inspire a clinical study with parturient patients, described in detail in Chapter 5.
Chapter 5

Clinical evaluation of SLIDE

5.1 Introduction

Multiple clinical studies have investigated the accuracy of manual palpation, compared to US, in intervertebral level identification. In a 50 person study of patients about to undergo spinal X-ray, Furness et al. showed that the traditional level identification using palpation is only 30% accurate [19]. In a study on obstetric patients, Whitty et al. investigated the agreement between visible puncture sites on 121 postpartum period patients against US operated by an anesthesiologist [63]. This study showed a 55% agreement between the identified level of the neuraxial needle puncture site and the US scan. In another study Lee et al. investigated the agreement between levels determined by manual palpation, and those found with the use of US in 51 at term obstetric patients, with respect to the Tuffier’s line [38]. They found only a 14% agreement between manual palpation and US in the identification of the vertebral level at the Tuffier’s line. The variations within the findings of these studies are unknown, but may have dependence on the expert performing the palpation, or an uneven distribution of patients amplified by relatively small study sizes.

In related research, Broadbent et al. show that the spinal cord can terminate below the L1 [7]. Thus, the risk of accidental selection of a higher interspace than intended for injection has an implication for increased likelihood of spinal cord damage. These findings motivate further investigation into the agreement between
manual palpation and US in the obstetric setting, and further motivate technologies that could improve the outcome of clinical level identification.

The success of SLIDE in its initial study with 20 participants showcased the potential of its clinical applicability. However, no participants in the study were pregnant, and all participants had a healthy body mass index (BMI). Thus, the results do not necessarily reflect obstetric clinical applicability. In the context of obstetrics anesthesia, women are approximately 37 weeks into their pregnancy and are at the end of their gestational period. Pregnancy incurs many physiological changes, affecting the weight of the patient and musculoskeletal properties. For example, lumbar lardosis is common among parturients, which is a curvature of the spine that narrows interspaces [3]. Therefore, there is a need to test the performance of SLIDE in an obstetric clinical setting.

Testing SLIDE in the clinical setting provides two benefits. First, it can validate the generality of SLIDE’s plane classification and level identification algorithms in a real-time context, without technical changes from the preliminary study. This can be directly compared against manual palpation in identifying vertebral levels, which can quantitatively assess the clinical value of SLIDE. Second, data from the pregnant patients can be recorded and used to update the classification algorithm used within SLIDE. As described in section 4.2.3 SLIDE classifies US images via a statistical ML model. As such, the model can be updated, or completely retrained using more data. Using more representative data of the expected application-specific patient population could help build a more robust model.

In this chapter, the design, execution, and analysis of a clinical test is described. The test took place at the British Columbia Women’s Hospital. The test provides initial clinical performance metrics for SLIDE, and the construction of a database of lumbar US with a parturient population. This data is further used to evaluate and retrain SLIDE’s internal CNN. As far as the author knows, this is the first clinical test with parturient patients to investigate the automatic analysis of US for vertebral level identification.
5.2 Experimental design

The clinical study is designed to compare the performance of SLIDE versus the performance of manual palpation, both compared against a gold-standard freehand US scan. In this work, SLIDE was used without the AR display, using only the 3D Slicer display. The primary outcome of the study is the identification of the L3-L4 intervertebral gap. The secondary outcome is the identification of the remaining gaps (the L1-L2, L2-L3, L4-L5, and L5-S1 interspaces). The use of US as a gold-standard is necessary, given the patient population is entirely obstetric, prohibiting the possible use of imaging modalities that use ionizing radiation, and the time and patient pose constraints on performing an MRI. Given that, it is possible that freehand US could provide an incorrect gold-standard. In an investigation by Furness et al., it was found that the vertebral level identification of freehand US only correlated with the true X-ray gold-standard in 71% of the cases [19]. To compensate for this, two members of the study team, of which all are familiar with lumbar US, had to agree on the freehand US results. The two members were allowed to discuss what they saw in the US images until they both felt confident with the analysis. In the statistical analysis of this study, the consensus freehand US is considered true.

Because no needles were inserted, results in terms of distance units to puncture sites were not suitable. Instead, level identification of each interspace is analyzed as a binary outcome. The location of vertebral levels from the various methods are recorded via a removable transparency sheet, which is taped to the patient’s back, as shown in figure 5.1. Recording results on the removable transparency provides blinding of the results between methods. The full protocol description is given below in section 5.2.2.

An example filled in transparency sheet, with annotations, is shown in Figure 5.2. It shows the vertical location of gaps, identified via palpation and SLIDE, as well as bones, via freehand US. The purpose of marking the location of vertebral bones with freehand US (the gold standard) is to discretize the intervertebral gap markings into bins. On the resulting transparency sheet, intervertebral gap markings are considered correct if they fall between their expected bone markings, and incorrect otherwise (e.g. the L3-L4 marking from the anesthesiologist is considered correct if it is between the L3 and L4 gold standard markings, and incorrect...
Figure 5.1: Transferring markings from the patient’s back to a fixed transparency sheet. Reference lines are first matched from the corners and sides of the transparency sheet to markings on the patient’s back, and then results are transferred onto the transparency sheet. Marks on the skin are erased to blind results between methods.
Figure 5.2: Example populated transparency sheet. The left figure shows the coloured markings, without annotations. The right figure shows the full sheet, with annotations describing the markings. Horizontal dashed lines are extended from the identified locations of L1 through L5. This example shows success in SLIDE and palpation for all 5 identified gaps. L3-L4 markings are pointed to, and both reside in the “bin” indicated by the L3 and L4 freehand US markings. The column of the palpation markings is further to the right, because it was marked close to the patient’s midline, whereas the SLIDE and freehand US markings are more distal because the transducer and gel cover the midline.

otherwise). The choice of marking bones as the gold standard is in contrast with marking the location of intervertebral gaps as the gold standard. This choice could have produced results in units of distance, but would not make clinical sense. This is because an intervertebral gap is not a point, but rather an opening that spans a distance.
Table 5.1: Demographic data of 76 included patients in SLIDE study

<table>
<thead>
<tr>
<th></th>
<th>Mean (std dev.) or N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>35.4 (4.5)</td>
</tr>
<tr>
<td>BMI</td>
<td>28.9 (3.9)</td>
</tr>
<tr>
<td>Gestational Age (weeks)</td>
<td>38.8 (0.9)</td>
</tr>
<tr>
<td>Parity</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>22 (28.9%)</td>
</tr>
<tr>
<td>1</td>
<td>37 (48.7%)</td>
</tr>
<tr>
<td>&gt;= 2</td>
<td>17 (22.4%)</td>
</tr>
</tbody>
</table>

5.2.1 Statistical analysis methods

Analysis of the clinical study were carried out in R version 3.3.2 [56]. The primary objective is to compare the accuracy of the SLIDE to palpation for determining the location of L3-L4, which is identified using freehand US as a gold standard. The secondary objective of the study is to compare the accuracy of identifying all other intervertebral gaps, which are also measured. There are several ways to think about accuracy and inter-rater agreement, but not all are applicable here. There is assumed to be no error in the estimation of the reference location. This results in a highly skewed distribution of ratings, which makes conventional measures of inter-rater agreement (e.g. Cohen’s kappa) very low or negative, with fairly high raw percent agreement. For highly skewed data such as these, an alternative measure of chance-corrected inter-rater agreement is Gwet’s AC1 statistic [23].

Data are presented as both raw percent agreement between palpation and freehand US, and between SLIDE and freehand US, as well as Gwet’s AC1 for chance-corrected agreement. A mixed-effects logistic regression is also used to compare the raw accuracy rate between the two methods. Mixed-effects regression was used to account for the paired nature of the data with two estimations made on the same individual. Relationships between accuracy and demographic variables, specifically age, BMI, gestational age, and parity (0, 1, 2 or more) are also assessed. Any of these that were significant ($p < 0.1$) were then added to a multi-variable adjusted model with method type. Demographic variables are summarized in Table 5.1.
5.2.2 Data acquisition protocol

The study obtained ethical approval through UBC Clinical Review Ethics Board. The principal investigator was Dr. Anton Chau, and the study was entitled Automated UltraSound Software for Identification of lumbar vertebral levels (AUSSI) with ID H16-02858. For simplicity, AUSSI may be referred to as the “SLIDE study” or “the study” in this section. The study had the following inclusion criteria:

- Normal healthy patient or mild systemic disease (American Society of Anesthesiologists physical status classification I or II),
- ≥ 19 years old,
- Term pregnancy (≥ 37 weeks gestation age),
- Scheduled for elective cesarean section,
- Ability to read English in order to understand the consent form

The study had the following exclusion criteria:

- Patient refusal to participate,
- BMI ≥ 40,
- Scoliosis,
- Previous lower back surgery or known spinal abnormalities,
- Inability to palpate bony landmarks in lumbar region,
- Active labour,
- Allergy to surgical paper tape or felt pen

The study had the following withdrawal criteria (however, no patients were excluded):

- Inability to complete scanning prior to surgery,
- Equipment malfunction during data acquisition
The summary of the approved protocol is as follows:

- Women awaiting elective cesarean sections are approached. They are informed about the study, and asked to participate.

- Consenting women are brought to the study area and guided to a seated position, to emulate how they would be in preparation for an epidural insertion. Specifically, they are asked to sit hugging a pillow, with their chin to their chest, their shoulders down, and pushing out their lower back, thereby pushing out their lower vertebrae.

- A transparent sheet is attached to the participant’s back with medical tape. This sheet is used for transcribing multiple sets of temporary markings from the participant’s back, which can later be analyzed. The location of the sheet is marked on the back to ensure consistent transcription of markings.

- An anesthesiologist palpates the back, identifying all lumbar gaps (from L5-S1 to L1-L2). Note that she does so by first marking L3-L4 with respect to the Tuffier’s line, as identified via standard manual palpation protocol, and then feels higher and lower on the back, with that location as a point as reference. Markings are made on the participants back, transcribed onto the transparency sheet, and then erased form the back to blind the results.

- The US machine is manually set up, selecting a frequency, depth, and gain with the qualitative criteria of having acceptable image quality to visually see common spinal anatomies. It must have an image depth setting to be able to see the AC, which is a deep feature in intervertebral gap images.

- SLIDE is used, identifying all lumbar gaps (from L5-S1 to L1-L2). If there is low confidence in results, SLIDE is allowed to be repeated in up to 5 attempts. Low confidence can occur if two levels are identified to be abnormally close to one another, or far apart from one another. It is assumed that this incorrect behavior could be easily identified by a novice user of the system. Markings are made on the participants back, transcribed onto the transparency sheet, and then erased form the back to blind the results.
• Freehand US is used, with unrestricted movement, such that two investigators must agree on all lumbar vertebral bone locations (from L5 to L1). Markings are made on the participants back, transcribed onto the transparency sheet, and then erased from the back to blind the results.

• The participant’s back is fully cleaned of markings and US transmission gel, and they are brought back to their initial room.

Data collection took place from January, 2017, until June, 2017. In total, 84 women awaiting cesarean section were approached and invited to participate, with 77 women consenting and 7 women declining participation. Of the 77 women that consented and participated, 1 was later excluded due to discovery of exclusion criteria. The woman had a BMI of 42, whereas the criteria states a maximum BMI of 40.

The following data was collected from each participant:

• A transparency sheet with level markings from manual palpation, SLIDE, and freehand US.

• B-mode US frames throughout the duration of the SLIDE scan and from the freehand US scan along with an associated log file, with SLIDE’s CNN and state output for each frame.

• Patient demographics (age, weight, height, BMI, gestation duration, parity).

5.2.3 US data storage

US data from the SLIDE study is stored on a password protected server owned by UBC with AES-256 encryption. Data are stored in individual scans, such that one dataset contains the images from the transducer in the transverse plane starting at the sacrum and moving cephalad towards the L1 vertebra. Between one and three sets of images per patient are manually labeled. Labels are given to frame ranges in the captured sequences, indicating when a sacrum, intervertebral gap, or vertebral bone are visible in the images. Images of the transition between planes are considered too ambiguous for labeling and are not given labels.
5.2.4 SLIDE CNN analysis methods

The performance of SLIDE’s internal plane-classification CNN is further analyzed in this work. Data obtained from the first 60 obstetric patients are labeled (not chosen based on any other criteria). Once labeled, SLIDE’s CNN is evaluated on this data, without any modifications. Then, SLIDE’s CNN is retrained and evaluated using the clinical data.

As described in chapter 4, SLIDE is constructed with a SqueezeNet architecture CNN [29] for plane classification. The model used transfer-learning, and fine-tuned the pre-trained model with data from 20 volunteers.

Retraining of the CNN is performed similarly in this work. The \( n = 60 \) obstetric dataset, \( X_{obs} \), is constructed from an evenly weighted sampling of data from all patients, for each of the sacrum, lumbar intervertebral gap, and vertebral bone planes. \( |S_i|, |G_i|, \) and \( |B_i| \) is the number of labeled images from each plane in patient data set \( i \). Random frames are selected from each labeled set and balanced to \( \min(200, \min(|S_i|, |G_i|, |B_i|)) \). For each patient, there are between 1 and 3 labeled sets. Therefore, a maximum of 1800 images from each patient could be used, limited due to memory constraints. In total, there are 22551 extracted images from \( X_{obs} \), with an average of 1128 per patient. As in Chapter 4, transverse images centered on the mid-line of the lower spine show anatomical symmetry. Due to the symmetry, the number of training data is doubled by augmenting with mirrored images, resulting in a net number of 45102 images used for training. A 20-fold cross validation (equivalent to leave-three-out) training is performed, running for 10 epochs.

5.3 Results

Manual palpation and SLIDE are compared against a gold standard US scan in the 76 patients that are included in the study. The results from these 76 patients are evaluated in a number of ways, which are further discussed in this section. The primary and secondary outcomes of the clinical study are discussed, followed by the analysis of SLIDE’s CNN performance on the acquired data.
Figure 5.3: Identified level by palpation or SLIDE relative to freehand US for the primary outcome: L3-L4 identification. The size of the circles is proportional to the number of cases in each category as indicated by the numbers in the circles.

5.3.1 Primary clinical outcome: accuracy of L3-L4

During manual palpation, the physician first identifies the L3-L4 using standard practice, identifying the Tuffier’s line and labeling the nearest palpable gap as L3-L4. Raw agreement for palpation compared to freehand US is 69.7% with Gwet’s AC1 = 0.59 (95% confidence interval (CI) = 0.41 to 0.77). Raw agreement for SLIDE = 84.2% with Gwet’s AC1 = 0.82 (95% CI = 0.70 to 0.93). In general, where palpation disagrees with freehand US, it tends to find L2-L3 more often than L4-L5. However, when SLIDE disagrees it is equally above or below L3-L4. Figure 5.3 shows the spread of identified locations relative to L3-L4 as determined by the freehand US.

From the mixed-effects logistic regressions there is no significant relationship
between accuracy of locating L3-L4 and age, gestational age, or parity (all \( p > 0.1 \)). However, there is a significant relationship with the BMI, shown in the odds ratio (OR) (OR = 0.83, 95% CI = 0.71 to 0.93, \( p = 0.001 \)). The OR suggests that for every increase in 1 unit of BMI, the odds of accurately identifying L3-L4 go down by about 17% for SLIDE and palpation considered simultaneously. There is also a significant relationship between accuracy and method type (OR = 2.96, 95% CI = 1.20 to 8.71, \( p = 0.02 \)). This remains significant and more or less unchanged after adjusting for BMI (OR = 2.99, 95% CI = 1.21 to 8.65, \( p = 0.02 \)). The OR suggests that the odds of correctly locating L3-L4 are about 3 times higher using SLIDE versus manual palpation after controlling for BMI.

5.3.2 Secondary clinical outcomes: other levels

All other levels have similar relationships between accuracy and method. Table 5.2 shows the results for all levels including the raw agreement, and AC1.

5.3.3 Other clinical outcomes

When SLIDE is used to identify vertebral levels, the transducer must always start at the sacrum and then be moved cephalad towards the L1-L2 intervertebral gap. SLIDE identifies levels based on transitions between identified planes. Therefore, although both the primary and secondary outcomes of the study can be directly compared to manual palpation, there are further metrics of analysis for SLIDE. The transition-accuracy, defined as the percentage of correct transitions between levels is important to consider. For example, if SLIDE has a transition error (it misses a plane, or falsely transitions away and back to a plane), subsequent levels will be incorrect with respect to the gold-standard. Without further transition errors, the implication is that SLIDE is actually performing correctly. It is important to note that transition errors that falsely transitions away and back to a plane are more clinically acceptable than missing a plane. This is because the first error type will result in levels being identified lower than their true location, whereas the second type will result in levels being identified higher than their true location. Inserting a needle lower on the spine incurs less risk, as outlined in Chapter 2. The transition-accuracy of SLIDE is 91.6%, with a more detailed breakdown shown in Table 5.3.
Figure 5.4: Identified level by palpation or SLIDE relative to freehand US for all intervertebral gaps. The size of the circles is proportional to the number of cases in each category. Larger size along the diagonal shows higher performance. As shown along the diagonal, SLIDE outperforms manual palpation at identifying all lumbar interspaces correctly. Also shown is that manual palpation has a higher tendency to incorrectly identify levels higher than the true location, compared to SLIDE. This adds higher risk to the patient than incorrectly identifying levels lower than the true location.
Table 5.2: Summary of accuracy and regression analyses for all levels, between manual palpation and automatic US analysis with SLIDE.

<table>
<thead>
<tr>
<th></th>
<th>Palp.</th>
<th></th>
<th>Auto</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% agree</td>
<td>AC1 (95% CI)</td>
<td>% agree</td>
<td>AC1 (95% CI)</td>
<td>raw OR (95% CI)</td>
<td>p value</td>
<td>adj OR (95% CI)</td>
<td>adj p value</td>
<td></td>
</tr>
<tr>
<td>L1-L2</td>
<td>60.0</td>
<td>0.41 (0.19 to 0.64)</td>
<td>0.74 (0.60 to 0.88)</td>
<td>3.10 (1.37 to 7.98)</td>
<td>0.01 (1.40 to 8.20)</td>
<td>3.17</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2-L3</td>
<td>68.4</td>
<td>0.57 (0.38 to 0.76)</td>
<td>0.78 (0.65 to 0.90)</td>
<td>2.79 (1.12 to 8.36)</td>
<td>0.03 (1.13 to 8.74)</td>
<td>2.85</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3-L4</td>
<td>69.7</td>
<td>0.59 (0.41 to 0.77)</td>
<td>0.82 (0.70 to 0.93)</td>
<td>2.96 (1.20 to 8.71)</td>
<td>0.02 (1.21 to 8.65)</td>
<td>2.99</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4-L5</td>
<td>67.1</td>
<td>0.55 (0.36 to 0.74)</td>
<td>0.86 (0.77 to 0.96)</td>
<td>3.59 (1.59 to 8.77)</td>
<td>0.002 (1.67 to 9.65)</td>
<td>3.85</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L5-S1</td>
<td>76.3</td>
<td>0.70 (0.55 to 0.85)</td>
<td>0.93 (0.87 to 0.99)</td>
<td>4.41 (1.65 to 14.56)</td>
<td>0.003 (1.65 to 14.58)</td>
<td>4.42</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All LvlS</td>
<td>46.1</td>
<td>0.11 (-0.15 to 0.37)</td>
<td>0.59 (0.41 to 0.77)</td>
<td>3.26 (1.54 to 7.85)</td>
<td>0.002 (1.55 to 7.87)</td>
<td>3.28</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3.4 SLIDE CNN analysis

Unmodified SLIDE CNN performance

The initial CNN trained purely on 20 non-obstetric volunteers, with results reported in Section 4.3.2 was used in the study. The output of the CNN is computed for each of the labeled frames and compared to the manually assigned label, in order to characterize the classification capabilities of the unmodified network on the new test dataset. The network was re-run on all 22551 labeled frames from 60 parturient patients. The classification accuracy on this set of images is 88.6 ± 6%.
Table 5.3: SLIDE transition-accuracy. The number of correct transitions are compared to the number of incorrect transitions that occurred due to falsely transitioning away and back to a plane (bouncing) or missing a plane. The number of transitions should ideally sum to 76 (the total number of patients) in each column. However, in a few cases, SLIDE failed to transition above a certain level (e.g., L1-L2 column sums to 72, indicating in 4 patients, this metric was not recorded).

<table>
<thead>
<tr>
<th></th>
<th>L1-L2</th>
<th>L2-L3</th>
<th>L3-L4</th>
<th>L4-L5</th>
<th>L5-S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>69 (96%)</td>
<td>69 (96%)</td>
<td>67 (91%)</td>
<td>66 (85%)</td>
<td>68 (90%)</td>
</tr>
<tr>
<td>Bounce</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>1 (1%)</td>
<td>1 (1%)</td>
<td>5 (7%)</td>
<td>5 (7%)</td>
<td>3 (4%)</td>
</tr>
<tr>
<td>Miss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>2 (3%)</td>
<td>2 (3%)</td>
<td>2 (3%)</td>
<td>6 (8%)</td>
<td>5 (7%)</td>
</tr>
</tbody>
</table>

SLIDE CNN retraining on parturient data

The CNN is retrained on the data from the study, $X_{obs}$, consisting of data from 60 parturient women. A 20-fold cross validation training set is used, obtaining an accuracy of $93.8 \pm 5.5\%$. The classification accuracy shows a statistically significant improvement from the unmodified SLIDE CNN classification performance (multiple comparisons test with $p < 0.01$).

5.4 Discussion and conclusions

A study is designed and performed to evaluate SLIDE versus manual palpation, when compared against a gold standard in identifying lumbar intervertebral gaps. This study involves the data collection of 76 parturient patients. As part of the protocol, transverse US scans of the patients lower back are recorded. This data is utilized to analyze the accuracy of SLIDE’s internal CNN without modification, and with the retraining of the network.

The primary outcome of the study, accuracy of identifying the L3-L4 intervertebral gap, show positive results for using SLIDE over manual palpation. The raw agreement for SLIDE is 84.2%, and with manual palpation is 69.7%. The OR between methods is 2.96, or 2.99 when adjusted for BMI, indicating that the odds of
identifying the L3-L4 are about 3 times higher when using SLIDE. The secondary outcome of this study, accuracy of identifying the other four lumbar intervertebral gaps, have similar results to that of the L3-L4. The analysis shows that BMI contributes significantly to both SLIDE and manual palpation, with each increase in 1 unit BMI resulting in a 17% decrease in odds of correctly identifying L3-L4. To the author’s knowledge, no other study has tested the automatic identification of vertebral level from US on parturient patients. These positive results suggest that SLIDE would be beneficial for pre-insertion planning if adopted.

At the core of SLIDE is a plane classification CNN. This CNN must discriminate between transverse planes of the sacrum, as well as lumbar intervertebral gaps and bones. In Chapter 4 it was shown that SqueezeNet is able to classify images in real time, with $88 \pm 9.1\%$ accuracy. This network, trained on 20 non-parturient subjects, is tested on 60 labeled parturient patient’s data and achieves $88.6 \pm 6\%$ accuracy. This indicates that the network is able to classify unseen test data without much loss in mean accuracy. This is especially interesting because the training images are all from non-pregnant patients, whereas this new test data is from an entirely parturient population.

The effect of training on parturient data is also explored, retraining on 60 patient’s data and achieving a 20-fold cross validation accuracy of $93.8 \pm 5.5\%$. This statistically significant increase in accuracy supports the hypothesis that training the network on a more broad, and representative, dataset improve the classification capabilities of the network. Going forward, obtaining more data to aid in training may further improve SLIDE’s CNN classification capabilities further.

It was qualitatively observed that certain patients, even some with lower BMI values, had poor image quality. It may be beneficial to explore an automatic image quality metric, which may be able to estimate if SLIDE will be able to accurately identify vertebral levels. This type of confidence metric could aid anesthesiologists in understanding the limitations of such an automated system.

This study shows positive results for the use of SLIDE, compared to manual palpation, in lumbar vertebral level identification. This study also results in the creation of a valuable database of parturient patient lumbar US scans, which can be utilized to develop or improve various models, including the CNN within SLIDE.
Chapter 6

Conclusion and future work

*I like the dreams of the future better than the history of the past*
— Thomas Jefferson (1816)

6.1 Summary of contributions

In summary, the work of this thesis has three main components: two systems that analyze US images to identify lumbar vertebral levels, via two scanning protocols, and a clinical study of the second system (SLIDE) with 76 parturient patients.

The work from the first system, LIT, entails the development of an image processing pipeline to identify lamina in paramedian spinal US images, and associate them with vertebral levels throughout time. As stated in Chapter 3, this work makes the following contributions: 1) A deep network architecture is proposed to identify peaks of lamina in standard paramedian 2D US, composed of 3 layers; two auto-encoders and a fully connected layer. This network performs pixel-level classification to coarsely segment peaks of lamina in each image. This network is trained on data from 15 volunteers, and produces pixel-level classification accuracy of 95%. 2) An algorithm is implemented that uses the identified location of lamina along with *a priori* knowledge of the paramedian scan path, from the sacrum towards the L1, to associate vertebral levels with the lamina. Between pairs of subsequently analyzed frames, the matching of vertebral level labels is correct in 94% of cases, when compared to matches of manually selected labels.
The work from the second system, SLIDE, is motivated by limitations of the first system. SLIDE aims to function in real time, such that it can be directly compared against other modalities and solutions to vertebral level identification. Further, SLIDE provides a simplified imaging protocol for the clinical workflow. US is often first used in the paramedian plane to identify level, and then used in the transverse plane to identify a patient’s midline. SLIDE functions in the transverse plane, allowing the transducer to remain in the transverse plane for identifying both the vertebral level and the midline. As stated in Chapter 4, this work makes the following contributions: 1) a framework is proposed, using a deep CNN, for automatic classification of transverse spinal US planes. Image classification between the sacrum, intervertebral gaps, and vertebral bones, achieves mean accuracy of 88-91%. SLIDE’s framework uses a state-machine, based on a priori knowledge of the clinical imaging protocol, to identify the vertebral level. In a study of 20 volunteers, all vertebral levels are able to be identified in 17 recorded data sets. 2) A real-time implementation of the framework is demonstrated, with processing speeds faster than the US frame acquisition rate; 40 FPS. An AR interface is also proposed to facilitate clinical workflow, projecting the detected vertebral level onto the patient’s skin adjacent to the transducer.

The initial success of SLIDE motivated investigation into its generality, with representative clinical data. A clinical test was designed and executed to compare the accuracy of level identification via manual palpation and SLIDE, when compared to freehand US as a reference. 76 parturient patients were included in the study, with SLIDE outperforming manual palpation in the identification of all lumbar intervertebral gaps, achieving 84% accuracy in L3-L4 identification. Performing this study allowed for the collection of data from 76 pregnant patients. Transverse data from 60 patients were labeled and used to retrain the classification CNN, to investigate the accuracy of the network, given more data. Classification accuracy on the 60 patient data set went up from 88.6 ± 6% when using the initial network (trained on 20 non-parturient volunteers) to 93.8 ± 5.5% 20-fold cross validation accuracy, showing a statistically significant improvement. This result supports the hypothesis that training the network on a more broad and representative dataset will improve its classification capabilities. Additionally, the data collected in this study can be further utilized in the development of related models.
6.2 Limitations and future work

The work presented in this thesis has limitations and can be improved via further investigation and refinement. Some suggestions are given in the conclusions of prior chapters, but are further explained here. The following is a list of possible improvements and extensions of the work proposed in this thesis:

For LIT:

- One limitation of the current implementation of LIT is the processing speed. Without real-time processing capabilities, LIT cannot be used in a clinical setting, implying that its level identification capabilities currently cannot be quantitatively assessed clinically. LIT has a number of steps which are computationally demanding and could be optimized. Speedups can be realized by parallelizing code where possible and processing on a GPU.

- Beyond processing time, another limitation is the number of steps in the processing pipeline, and the number of parameters in these steps. A number of advancements in end-to-end segmentation neural networks have been proposed for fast image segmentation [41]. An investigation into options for segmenting the location of lamina peaks would be beneficial, such as with end-to-end neural networks. Trade-offs in speed and accuracy should be compared, and an option should be identified that allows for real-time processing. Further, the accuracy of segmentation methods without pre-processing US images should be investigated. The need for manually-chosen parameters may be reduced by using either one or both of these options.

Upon further work, LIT could benefit clinicians because they are already familiar with scanning in the paramedian plane, and such a tool could help increase their confidence in understanding US when identifying vertebral levels.

For SLIDE:

- SLIDE’s internal CNN classification accuracy could be improved, which may result in better overall system performance. Much more data augmentation could be performed via minor translations, rotations, and rigid or non-rigid transformations. Augmenting the data allows for further network training, without requiring the collection of more data, which can improve validation...
accuracy. With enough augmentation, the training database may grow large enough to train a deep CNN from scratch, rather than using transfer learning, which may also improve classification accuracies.

- The state machine within SLIDE is intended to identify transitions in planes based on the temporal output of its internal classification CNN. However, this state machine relies on manually tuned parameters. Success in utilizing temporal information within statistical models has been shown via RNNs. An investigation into using RNNs for identifying planar transitions should be performed, which could eliminate the need for manually tuned parameters (not including network hyper-parameters). Hidden Markov models may also be appropriate for identifying planar transitions.

Throughout the development and analysis of SLIDE, physicians expressed their interest and excitement in such a technology. SLIDE requires a less common transducer scan path for physicians, but could improve their work flow and make US simpler to use. Upon refinement, SLIDE could be used to make US part of the standard of care in spinal needle insertions.

More uses for the clinical database created in Chapter 5:

- In addition to the required transverse imaging protocol of SLIDE, paramedian US data was also collected from most of the parturient patients, following the protocol described in Chapter 3. If this data is labeled, it could be used for further validation of LIT. It could be used as additional training data, which could improve the generality of the localization capabilities.

- With transverse and sagittal images from 76 parturient patients, with various BMIs, using various US parameters, such as frequency, depth, and gain, the database could be valuable for other automated system development and refinement. Exemplary use cases could include the automatic classification, localization, segmentation, and/or depth calculations for commonly visible anatomies such as the LF, lamina, articular processes, or ACs in paramedian images, and spinous processes, facet joints, transverse processes, lamina, articular processes, or ACs in transverse images.

Other future work:
• The work in this thesis is limited to the lumbar spine. Many percutaneous spinal procedures, such as facet joint injections, can occur above the lumbar vertebrae. Therefore, it would be beneficial to extend either or both systems to the entire spine, or at least the thoracic vertebrae. It is feasible that both the methods for segmentation of lamina in paramedian images, and the classification of planes in transverse images may work without modification on thoracic or even cervical vertebrae. But, due to the anatomical differences in regions of the spine, it is likely that databases of these regions will need to be obtained, and networks will need to be developed and trained specifically for success in these regions.

This thesis presents work that shows the potential of aiding physicians in lumbar vertebral level identification via automatic analysis of US. LIT and SLIDE present methods for ML based analysis of US, which utilize unsupervised feature learning, as well as transfer learning. The methods presented, and potential future extensions to these methods, can feasibly be used by physicians in many fields of medicine, when vertebral level identification is required. The techniques shown in this thesis can also aid other related US analysis problems, involving segmentation and classification. Positive results in a clinical study highlight the potential impact of SLIDE, and also resulted in the construction of a database that can be used by other researchers with similar research ambitions.
Bibliography


[3] C. L. Baysinger, B. A. Bucklin, and D. R. Gambling. *A practical approach to obstetric anesthesia*. → pages 2, 6, 8, 9, 10, 11, 12, 13, 60


computing platform for the quantitative imaging network. Magnetic Resonance Imaging, 30(9):1323–1341, 2012. → pages 42


82


83


learning for segmentation of deep brain regions in mri and ultrasound. Computer Vision and Image Understanding, 2017. → pages 23


