RICE UNIVERSITY

Using Motion-Based Metrics to Objectively Classify Surgeon Skill and Assess Performance with Augmented Feedback

by

Benjamin Kramer

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE Master of Science

APPROVED, THESIS COMMITTEE:

Marcia K. O'Malley, Chair
Professor of Mechanical Engineering and
of Computer Science

Andrew J. Dick
Assistant Professor of Mechanical Engineering

Erzsébet Merényi
Research Professor in Statistics and in
Electrical and Computer Engineering

Houston, Texas
April, 2016
ABSTRACT

Using Motion-Based Metrics to Objectively Classify Surgeon Skill and Assess Performance with Augmented Feedback

by

Benjamin Kramer

Objective evaluation of surgical skill based on assessment of tool movements made by the surgeon is a rapidly developing area of research. Motion-based performance metrics are computed from data recorded in real-time during surgical procedures. The approach offers advantages over the subjective observation based grading schemes typically used to characterize the expertise of the endovascular surgeon. In this thesis, these motion-based metrics are applied in two novel scenarios. First, the use of artificial neural network machine learning techniques demonstrates that these metrics that quantify tool movement smoothness and quality can be used to classify surgeons as experts, intermediates, or novices. This finding extends prior work that simply showed correlations between such metrics and expertise level. Second, motion-based metrics are used not to directly assess surgical expertise and skill, but to evaluate the benefits of new tool visualization technologies made possible with real-time electromagnetic (EM) sensing and tracking of endovascular tool tip movements. Visualizations based upon EM sensing technology are shown to increase motion smoothness while also decreasing radiation exposure during a procedure.
Acknowledgments

While there are undoubtedly countless individuals to whom I owe thanks from my time at Rice University, there are a few in particular. I would first like to thank Dr. O’Malley for providing an opportunity for me to pursue a graduate degree under her guidance. I learned a lot about robotics and surgery, research in general, and myself. It has truly been a pleasure to be a part of the MAHI Lab. I would like to thank Dylan Losey, Andrew Erwin, Chad Rose, Craig McDonald, Laura Blumenschein, Evan Pezant, Ted Artz, and Kyle Fitle for creating a great atmosphere of collaboration and friendship. I would also like to thank Dr. Merényi for pushing me to present my work involving machine learning. I really enjoy the subject matter and would like to pursue it further some day. I would also like to thank Dr. Dick for participating on my thesis committee.

To my dear friend Clark, once a Wolverine, always a Wolverine.

And finally, I would like to thank my family, my girlfriend Emily, and my cat Kirby. Each of them has found the time and energy in their busy lives to support me, and I greatly appreciate it.
Contents

Abstract ii
Acknowledgments iii
List of Illustrations vi
List of Tables xi

1 Introduction 1
  1.1 Basics of Endovascular Procedures 2
  1.2 Motion Tracking 7
    1.2.1 Methods for Tracking Motion 7
    1.2.2 Applications of Sensors 8
  1.3 Motivation 9

2 Motion-Based Quantitative Metrics 10
  2.1 Smoothness in the Time Domain 11
    2.1.1 Average Tangential Acceleration 11
    2.1.2 Root Mean Dimensionless Jerk 11
    2.1.3 Normalized Speed 12
    2.1.4 Submovement Extraction 12
  2.2 Smoothness in the Frequency Domain 15
    2.2.1 Spectral Arc Length 15
  2.3 Task Kinematics 16
    2.3.1 Path Length 16
    2.3.2 Catheter Turns 16
3 SOM and LVQ Classification of Endovascular Surgeons 18
   3.1 Experimental Protocol ........................................ 20
   3.2 Machine Learning Techniques ................................. 23
   3.3 Endovascular Surgical Skill Classification Methods ........ 23
      3.3.1 Classification with LVQs .................................. 24
      3.3.2 Clustering with SOMs .................................. 26
      3.3.3 Input Data and Class Labels ............................. 29
   3.4 Results and Discussion ..................................... 32
      3.4.1 LVQ Classification Results ............................. 32
      3.4.2 SOM Clustering Results ................................ 35
   3.5 Conclusions .................................................. 41

4 Objective Analysis of Augmented Visual Feedback during Robotic Catheterization 42
   4.1 Platform ....................................................... 43
      4.1.1 Viewing Modes ........................................ 47
   4.2 Pilot Study .................................................. 50
      4.2.1 Methods ................................................ 51
      4.2.2 Results ............................................... 54
   4.3 Follow-Up Study ............................................ 57
      4.3.1 Procedural Modifications ............................... 58
      4.3.2 Results and Discussion ............................... 58
   4.4 Conclusions ................................................ 66

5 Conclusions and Future Work .................................. 67

Bibliography .................................................. 69
Illustrations

1.1 Sample fluoroscopy image with (1) labeled guidewire tip of the catheter [8] ................................................................. 3

1.2 Profile of patient showing (a) where the catheter is inserted and (b) the basic design of the catheter [4] ................................................. 4


3.1 A comparison of the manual and virtual simulators which were navigated during the various tasks. In both platforms the surgeon is operating a catheter—in (a) the tip position is tracked using a magnet at the tool-tip, while (b) is a virtual reality simulator with tip tracking based on image processing .................................................. 21


3.3 Each of the plotted points corresponds to one of the eleven motion metrics derived from the surgeon’s trajectory. Average input vectors associated with a novice, intermediate, and expert surgeon are shown (standard error bars included) ......................... 30
3.4 Sample LVQ results from one fold of the four-fold stratified cross-validation procedure: training classification left; testing classification right. The black pixels represent novice surgeons, the grey pixels represent intermediate surgeons, and the white pixels represent expert surgeons. The top plot is the desired classification, the middle plot is the actual classification, and the bottom plot is the error. ................................................................. 34

3.5 SOM learning history using the same visualizations as in Figure 3.6. Each pair of visualizations from left to right represents 10,000 learning steps, starting at 10,000 on the left. Given that there is little observable change between the visualizations at 90,000 and 100,000 learning steps, and that both appear to be similar to the network at 80,000 learning steps, this illustrates that the training has converged. 36
3.6 SOM final results. Both visualizations provide insight into the final weight values of the 49 prototypes (7 by 7 grid) and how the data space maps onto the SOM grid. The left visualization is a modified U-Matrix [52]; red-scale represents the relative number of data vectors mapped to each prototype (i.e., density), while the gray-scale bars signify the distance between prototypes in the data space. The redder the neuron, the more input vectors are contained within its Voronoi cell; likewise, the darker the bar separating neurons, the greater the difference between their weight vectors. The right visualization shows the known surgeon classifications projected onto the SOM lattice—here red signifies novice, green represents intermediate, and blue indicates expert surgeons, with color intensity representing the number of mappings (more intensity again means increased density). Black neurons indicate that no input vectors are mapped to a particular node. The clusters found in the mU-matrix are identified using white lines in the right visualization. While novices (red) are primarily separated, clustering in the upper and lower left, intermediates (green) and experts (blue) are largely intermingled, clustering along the right side, a result which supports the LVQ findings.
3.7 SOM weight vector plotted in the grid cells. This figure shows both the final results of the SOM grid with the known classifications projected onto the lattice, as well as the weight vector of each PE with respect to the average weight vector across all nodes. The weight vector of a given PE is shown in black, while the average weight vector across all nodes is plotted in a dotted magenta line. The color coding of the prototypes is the same as before, with a slight fading of the colors in order to better visualize the weight vectors. Black boxes were used to mark the SOM cluster boundaries identified in Figure 3.6.

4.1 Magellan Catheter showing the guidewire running through the leading catheter (leader) telescoping within a flexible sheath.

4.2 Magellan Robotic System showing the operating table, C-arm, and remote workstation from which the surgeon operates the robotic catheter. The surgeon is presented with imaging of the procedure on the screen depending upon the chosen viewing mode and is given a control panel, as well as foot pedals, for controlling the robotic catheter and fluoroscopy.

4.3 Magellan Robotic System viewing modes. a) Traditional Single Plane Fluoroscopy (Control/Fluoro) b) Electromagnetic Fluoroscopy Biplane (EM Biplane) c) Electromagnetic 3-D Endovascular (EM Endo) d) Electromagnetic 3-D Biplane (EM 3-D)

4.4 Rigid fluid-filled aortic aneurysmal phantom used during the procedure with a) EM field generator placed under the angiography table b) Robotic Catheter and c) removable simulated contralateral gate.
4.5 Sample trajectories from a subject using the various viewing modes. It should be noted that the tool-tip motion for Mode 1 appears to be significantly greater in length and less intentional than the tool-tip motion using the other viewing modes. .......................... 57

4.6 Comparison between tasks one and two showing the apparent difference in difficulty between the two. The larger number of catheter turns (a) and longer path lengths (b) for task 2 indicate higher level of difficulty for task 2 compared to task 1. ............... 60

4.7 Interaction between Task and Viewing Mode visualized using metrics (a) Number of Submovements, (b) Catheter Turns at 165°, (c) Root Mean Dimensionless Jerk (RMDJ), and (d) 3-D Path Length ........... 63

4.8 Spectral Arc Length interaction plot showing the decreased performance using the EM3E viewing mode on Task 1. ............... 64
Tables

3.1 Classification Summary. The mean and standard deviation for training and testing accuracy. Poor results likely stem from incorrect class labels, particularly between intermediates and experts. 32

3.2 Average confusion matrix over the four folds. Data is given in the form % of hits (number of hits). Diagonal elements represent correctly classified data, while off-diagonal elements show incorrect classifications. 33

4.1 A contrast t-test was run for each metric evaluating the control mode (SPF) against the three EM modes (EMFB, EM3E, and EM3B) using the conventional (−3, 1, 1, 1) contrast to test whether the means were significantly different. These contrast t-tests were evaluated across both tasks as well as for each task individually. Bolded values were significant. 56

4.2 Guidewire Mixed Design ANOVA - all modes. The table shows the p-values for each metric given the within subject and between subject factors. Each factor is labeled with its respective degrees of freedom. Significant results are bolded. 59

4.3 Guidewire Mixed Design ANOVA - SPF, EMFB, and EM3B only. The table shows the p-values for each metric given the within subject and between subject factors. Each factor is labeled with its respective degrees of freedom. Significant results are bolded. 61
4.4 Guidewire Metrics - Contrast T-Test comparing SPF against EMFB,
EMFB+,, and EM3E. Significant results are bolded. . . . . . . . . . . . 62

4.5 Guidewire Metrics - Contrast T-Test comparing SPF against EMFB
and EM3E. Significant results are bolded. . . . . . . . . . . . . . . . 62
Chapter 1

Introduction

A primary challenge in assessing and evaluating surgical skill is the subjective, and hence non-standardized, nature of such evaluations [1]. Presently, observation by an expert surgeon is the primary method of assessment. During the course of these assessments, expert surgeons grade a trainee’s performance using a designated grading scale. While these structured grading scales are intended to universalize evaluation, they still rely heavily on the objectivity of the observing surgeon. The time required for experts to perform these assessments as well as the objectivity of such experts has been largely questioned [2]. As an alternative to observation and grading, a surgeon’s qualification level is often linked to caseload [3]. Within the field of endovascular surgery, a surgeon is considered to be of an intermediate skill level after performing 50 surgeries, and an expert after performing 500 surgeries [4]. This measure of competence assumes that experience breeds expertise; however, the correlation between caseload and surgeon skill is unproven [5]. Despite the subjectivity of these assessments and uncertain correlation between caseload and surgeon skill, there is a lack of sufficiently validated alternatives. An objective and quantitative assessment tool, well suited to the complexity of endovascular surgery, is needed to fill this void.
1.1 Basics of Endovascular Procedures

Endovascular procedures in particular lack objective assessment and are particularly challenging due to limited visual fidelity and tools that are difficult to navigate. An endovascular surgery is a form of minimally invasive surgery used to treat problems affecting the blood vessels. Minimally invasive surgeries are alternatives to "conventional" surgeries involving a single large incision. By introducing tools through a number of small incisions, surgeons are able to decrease the likelihood of infection while also shortening recovery times and discomfort for patients [6].

Surgeons are presently able to visualize the inside of a patient during a procedure using fluoroscopy, shown in Figure 1.1. A fluoroscopic feed of the patient’s vasculature (the arrangement of blood vessels in an organ or part of an organ) is achieved by passing a continuous X-ray beam through the body part being examined [7]—typically the aorta for endovascular procedures. A fluoroscopic view may be difficult to interpret as a result of its two-dimensional nature and monochromatic coloring. Often times it can be especially difficult to differentiate the vasculature of the patient from the bones and muscles in the surrounding area. This issue can be mitigated by the use of contrast dye. As one would expect, contrast dye provides contrast to the image, as seen in the darker sections of Figure 1.1. When needed, contrast dye is flushed through the catheter into the vessel branches of the vasculature in order to provide better information about the vessels. The contrast dye is quickly diluted within the bloodstream and must be injected each time it is needed.
Presently, endovascular procedures are limited by the required utilization of fluoroscopy and contrast dye. Because fluoroscopy requires a continuous X-ray beam, patients may experience prolonged exposure to radiation, especially for more complex procedures [9]. Increased exposure to radiation can lead to radiation-induced burns and cancer [9]. Contrast dye, while not as dangerous, poses a few complications as well. A small percentage of individuals are allergic to contrast dye [7], making it unavailable during a procedure and thus more difficult for a surgeon to visualize the vasculature. For everyone else, contrast dye is difficult for the kidneys to process [7], therefore limiting the amount which can be used. Because a patient’s exposure to fluoroscopy and contrast dye must be limited during a procedure, there is perceived to be a finite amount of visualization which a surgeon can acquire during a given procedure.
Figure 1.2: Profile of patient showing (a) where the catheter is inserted and (b) the basic design of the catheter [4]

In order to perform an endovascular procedure, a surgeon uses a combination of fluoroscopy and contrast dye to steer a catheter (Fig. 1.2 (b)), inserted into the vasculature of the patient (typically through the femoral artery near the hip), to the location of concern. This process is known as catheter cannulation. A full catheter is composed of an inner guidewire, a middle catheter (also known as the leader), and an outer sheath. Steering of the entire catheter involves patiently fishing out in front with the guidewire and then sliding the leader and sheath over the guidewire. Because the guidewire is the smallest and most pliable of the tools, a surgeon will use the guidewire to progress along a vessel in order to avoid damaging vessel walls [10]. Guidewires often feature a j-tip in order to further prevent vessel wall damage and to allow for partial hooking with the guidewire. Once a path has been established by the guidewire, the leader, and then the sheath, will be brought along the guidewire.
For instances in which the pliable nature of a guidewire makes it difficult to cannulate specific branches of the vasculature, the leader and sheath can be brought closer to the guidewire in order to create support and rigidity. Skilled surgeons are able to effectively balance these techniques, using limited visibility, in order to successfully cannulate the catheter to the procedure location.

![Diagram of Aortic Stent Graft Placement](image)

Figure 1.3: Placement of an Aortic Stent Graft at the aneurysm [11]

Once positioned, a catheter can be used to deploy medical devices, such as balloons or stents, in order to treat the patient. One of the more common endovascular procedures involves the treatment of an aneurysm. An aneurysm, pictured in Figure 1.3, is a balloon-like bulge in the wall of a blood vessel. Installing an aortic stent graft at the location of the aneurysm provides an artificial passageway for blood to flow through without placing pressure on the aneurysm sac and risking rupture.
In order to better facilitate cannulation and deployment, robotic catheters have been developed offering greater tool-tip manipulation. Among these is the Hansen Magellan Robotic System (MRS) (Fig. 1.4). The functionality of the system, resembling that of manual catheterization, relies upon the concept of a guidewire running through a leading catheter which telescopes within a flexible sheath. The MRS enables enough stability to reliably use the catheters and wires to navigate through rather than interact with the vessel wall, reducing wall injuries [12], [13]. Its capability to be remotely steered with six degrees of freedom offers potential benefits: better catheter orientation and maneuverability facilitating vessel cannulation [14], as well as reduction in radiation exposure [15]. Despite these advancements, surgeons are still provided limited visibility of the procedure and forced to expose patients to large amounts of radiation and contrast dye during cannulation.
1.2 Motion Tracking

1.2.1 Methods for Tracking Motion

New technologies allow for tracking of movements during surgical procedures, offering both improved visualization opportunities and opportunities for objective assessment. The tracking of surgical instruments supports innovative research, decision making, and skill assessment [16], [17], [18].

One method of motion tracking is optical trackers. Using a sufficient number of image sensors and triangulation, it is possible to record the 3-D position of an object [19]. These systems produce data with three degrees of freedom for each marker. Because of the line of sight requirement associated with cameras, optical markers struggle with occlusion and require a well-established operating space. An insufficient number of cameras on a marker makes triangulation, and therefore position data, impossible [20].

As an alternative to optical trackers, computer vision techniques can be used for motion tracking. Using thresholding and machine learning algorithms, it is possible to analyze images and properly identify structures within the images [21]. While these techniques are often limited to two-dimensional images, orthogonal images would allow for three-dimensional positioning. The fine-tuning required to accurately track an object is a possible disadvantage of this technique.

A final method is Electromagnetic (EM) tracking. It involves the use of an electromagnetic field generated in space and detected at a remote location [22]. Knowing the signals from the source and sensor, it is possible to calculate the position and orientation of the sensor with six degrees of freedom in real-time. While the accuracy of these sensors can be limited by neighboring metallic objects [23], the ability to
locate them close to the point of interest and track motion within an object, such as
the body, makes EM sensors ideal for surgical applications.

1.2.2 Applications of Sensors

Using motion data, it is possible to objectively evaluate a subject’s performance
during task completion. Motion-based metrics can be calculated from the data in
order to provide quantitative assessment tools. Recent efforts across multiple domains
have looked at motion-based metrics as an objective method of evaluation. These
metrics are sometimes directly derived from the position data of the tracked object. In
other scenarios, metrics are computed from the position data in an attempt to quantify
movement quality as defined by fundamental theories of human motor control. Target
hitting tasks, among other simple motion tasks, have been analyzed using motion
metrics in order to objectively quantify motion smoothness [24]. Fittle et al. used
motion-based metrics to evaluate the effects of rehabilitative robotics on impaired
subjects following incomplete spinal cord injury [25]. Movement characteristics in
individuals with motor impairment compared to whose without are comparable to
the relationship between novice and expert movement characteristics.

These objective measurements were similarly translated to the laparoscopic surgery
domain [26], [27]. Hofstad et al. used motion-based metrics to study the differences
between expert and nonexpert performance in minimally invasive surgeries [27] while
Hernandez et al. examined the learning curve on the da vinci system [26]. Estrada
et al. further translated these metrics to the endovascular surgery domain [28]. The
findings showed correlations between surgeon skill levels and motion-based metrics [4].
1.3 Motivation

In this thesis, applications of motion-based metrics for the classification of endovascular surgeons according to skill level and for the assessment of performance augmentation are presented. By combining motion-based metrics with machine learning techniques, it is possible to explore the metrics that can best be used to classify surgical skill. Similarly, artificial neural network clustering helps to illustrate the shortcomings of current caseload-based qualification criteria. Using the motion-based metrics also allows for the evaluation of augmented visual feedback on the performance of surgeons testing on a robotic platform. Chapter 2 will discuss the specifics of the motion-based metrics use for assessment of skill. Chapter 3 will explore attempts to classify surgeons using these metrics and machine learning techniques. Chapter 4 will show how recent advancements in the technology provided to surgeons can improve performance. Chapter 5 provides a summary of the conclusions of this work, and offers some potential future directions.
Chapter 2

Motion-Based Quantitative Metrics

In this chapter, motion-based quantitative metrics are discussed. While some of these metrics are derived directly from the instrument kinematic data and others are computed from the kinematic data according to theoretical motor control, both methods attempt to characterize the quality of motion during a task. Movement smoothness, in particular, has been shown to be an effective measure of motor performance during task completion. Reaching tasks have been analyzed using motion metrics in order to quantify the coordination and smoothness of subject movement [24]. Based upon the successful assessment of subject motion in simple motor-related tasks such as reaching, as well as more complex motor-related tasks such as laparoscopic procedures [26], Estrada et al. extended these motion-based metrics to endovascular surgery [4], thereby providing a means by which one can objectively measure surgeon performance. All motion-based metrics used in this thesis are calculated from the three-dimensional position data of the tool tip over the duration of a procedure. Because of the limitations of optical motion capture discussed in Chapter 1, data were collected using EM sensing and computer vision techniques for Chapter 3, and EM sensing exclusively for Chapter 4. Velocity and acceleration data is calculated using linear approximations of the data given the known time between recorded data points.
2.1 Smoothness in the Time Domain

This section presents metrics which analyze the smoothness of motion with respect to time. By examining how a tool-tip position changes over time, it is possible to quantify the smoothness of motion over the course of such movement.

2.1.1 Average Tangential Acceleration

Average tangential acceleration, shown to characterize arm movement during hitting and drawing tasks, is the derivative of tangential speed. A smoother movement will involve gradual changes in velocity and therefore have a lower average tangential acceleration [29], [30]. This allows for differentiation between subjects whose speed profiles are similar in magnitude but one subject’s motion involves less smooth changes in direction and speed. Average tangential acceleration is computed using Equation 2.1 where \(N\) is the vector length and \(x, y, z,\) and \(t\) are the recorded position and time data.

\[
TA(t) = \frac{1}{N} \sum_{j=1}^{N} \frac{d}{dt} \left( \sqrt{\left( \frac{dx_j}{dt} \right)^2 + \left( \frac{dy_j}{dt} \right)^2 + \left( \frac{dz_j}{dt} \right)^2} \right)
\]  

(2.1)

2.1.2 Root Mean Dimensionless Jerk

Jerk, the derivative of acceleration, is also considered to be an effective way to quantify the smoothness of a movement. From the theory that a smoother acceleration profile is indicative of movement smoothness, jerk allows us to test whether or not the acceleration is smooth (lower jerkiness) or sharp (higher jerkiness) [24]. In order to make the measurement more universally applicable, the jerk of a movement is non-dimensionalized and transformed using the mean and square root to account for
the complexity and duration of a task [31]. This allows us to compare the jerkiness of subjects with different task completion times. Root Mean Dimensionless Jerk (RMDJ) is computed using Equation 2.2 where \( T_1 \) and \( T_2 \) are the start and end time, respectively, \( T \) is the duration and \( PL \) is the 3-dimensional path length, as defined in Equation 2.7.

\[
RMDJ = \sqrt{\left(0.5 \int_{T_1}^{T_2} (x'''(t))^2 + (y'''(t))^2 + (z'''(t))^2 \, dt\right) \cdot \frac{T^4}{PL^2}}, \tag{2.2}
\]

### 2.1.3 Normalized Speed

The normalized speed metric, similar to the other metrics, again attempts to quantify the consistency of a user’s motion. In order to do this, the mean speed of the motion is divided by the peak speed [32]. Thus, an individual user’s speed variance throughout the task is quantified as it is a relationship between the mean speed and peak speed.

### 2.1.4 Submovement Extraction

It has been hypothesized that the movement required to complete a task can be broken down into a number of submovements, each resembling a unimodal, bell-shaped function [33]. Using segmentation techniques such as the scattershot optimization algorithm proposed by Rohrer et al. [33], the tool tip movement over the course of a task can be broken down in order to extract the individual submovements which together compose the entire task motion. A sample submovement decomposition is illustrated in Figure 2.1.
Algorithms have been developed to extract submovement functions from both minimum-jerk profile curves (MinJ) and support-bounded lognormal curves (LGNB).

Minimum jerk submovements are characterized using the amplitude of the peak, $A$, the time at which the peak occurs, $t$, and the duration of the movement, $\omega$, outlined in Equations 2.3 and 2.4.

$$v(\tau) = \frac{A}{1.875} \left( 30 \left( \frac{\tau - t + \frac{\omega}{2}}{\omega} \right)^2 - 60 \left( \frac{\tau - t + \frac{\omega}{2}}{\omega} \right)^3 + 30 \left( \frac{\tau - t + \frac{\omega}{2}}{\omega} \right)^4 \right),$$

$$t - \frac{\omega}{2} \leq \tau \leq t + \frac{\omega}{2} \quad (2.3)$$

$$v(\tau) = 0 \ , \ \text{otherwise} \quad (2.4)$$

Support-bounded lognormal curves are defined by five independent parameters as outlined in Equation 2.5 [34], therefore allowing them to represent a wide range of submovement-like shapes.
\[ B(t) = \frac{D (T_1 - T_0)}{\sigma \sqrt{2\pi} (t - T_0) (T_1 - t)} e^{\left(\frac{1}{2\sigma^2} \left[ \mu \left( \frac{t - T_0}{T_1 - T_0} - \mu \right)^2 \right] \right)} \]  

for \( T_0 \leq t \leq T_1 \), where \( D \) is the displacement resulting from the movement, \( T_0 \) is the movement start time, \( T_1 \) is the end time, \( \mu \) controls the skewness (asymmetry), and \( \sigma \) determines the kurtosis (“fatness”) of the curve [34].

Regardless of the method used for extraction, identifying these individual submovements allows for investigation into the complexity and smoothness of task completion. Due to the computational intensity of submovement extraction, LGNB curves were exclusively used in Chapter 4 given the greater flexibility of submovement shape. Chapter 3, however, uses both types of curves.

**Number of Submovements**

As suggested, the number of submovements is simply the number of bell-shaped sub-profiles that are identified to comprise the composite velocity profile. Disjointed motion will result in a larger number of submovements being extracted from the data. More difficult tasks have been associated with movements that contain a larger number of submovements [4]. Thus, smoother and more adept task completion will exhibit fewer submovements.

**Submovement Duration**

Submovement duration can then be calculated using the number of submovements in a motion profile. Defined as the total time of task completion divided by the number of submovements, submovement duration is a measure of average time spent for each submovement. Smoother motion profiles will involve longer, more methodical
submovements whereas uncoordinated motion will be erratic and have more submovements and therefore shorter submovement duration.

2.2 Smoothness in the Frequency Domain

While most of the metrics are computed in the time domain, it is also important to explore the frequency domain. It stands to reason that a smoother, more coordinated movement will consist of primarily low-frequency components, whereas an unsmooth movement will be composed of higher-frequency components.

2.2.1 Spectral Arc Length

The frequency spectrum of a movement can be evaluated using the Fourier magnitude spectrum of the path length. Therefore, spectral arc length is defined as the negative arc length of the amplitude and frequency-normalized Fourier magnitude spectrum of the speed profile (Equation 2.6). Spectral arc length is considered to be extremely robust and easily calculated [35]. A smoother movement will consist of primarily low-frequency components.

\[
\eta_{sal} = -\int_0^{\omega_c} \sqrt{\frac{1}{\omega_c} + \frac{d\hat{V}(\omega)}{d\omega}} \, dt
\]

\[
\hat{V}(\omega) = \frac{V(\omega)}{V(0)}
\]  

(2.6)

where \( V(\omega) \) is the Fourier magnitude spectrum of \( v(t) \), and \([0, \omega_c]\) is the frequency band occupied by the given movement. \( \omega_c = 40\pi \text{ rad/s} \) (which corresponds to 20 Hz) covers the normal and abnormal aspects of human movements such as tremor [36].
2.3 Task Kinematics

Kinematic metrics come directly from the measured position, velocity, and acceleration of tool tip motion. These measurements can be easily calculated in real-time and do not require knowledge of the entire motion profile.

2.3.1 Path Length

Path length serves as a simple and easily calculated performance measure. By looking at the total three-dimensional path length which the tool tip traverses over the course of a procedure, it is possible to draw basic conclusions about user performance [37], [38]. Using the three-dimensional position data of a task, path length is calculated according to Equation 2.7.

\[
PL = \int_0^T \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \, dt
\]  

(2.7)

Given a task involving movement from point A to point B, more efficient and skilled users will generally require a shorter path length to reach the objective. Longer path lengths can be indicative of difficulty with a specific aspect of the task, a lack of familiarity with the equipment being used, or an overall lower level of skill or confidence. There are numerous reasons why the path length of a user may vary in magnitude; however, this is certainly a good starting metric for differentiating user performance.

2.3.2 Catheter Turns

In order to explore the complexity of various tasks, the catheter turns metric was developed for this research. In many cannulation tasks, surgeons will attempt to extend the catheter tip to a specific location and if unsuccessful, will retract the
catheter slightly before reattempting. It stands to reason that on a per-person basis, a more difficult task will involve more failed attempts. Similarly, on a per-task basis, less skilled subjects will have more failed attempts. Therefore, the three-dimensional velocity data, derived from the position data, was used to calculate the angle between subsequent velocity vectors. The number of times over the course of a task where the angle value between subsequent vectors exceeded $135^\circ$, $150^\circ$, and $165^\circ$ was recorded. Per the input of individuals familiar with endovascular procedures, $150^\circ$ was settled on to be the ideal angle value. It is also important to note that the catheter turns were not recorded if the tool tip speed did not exceed $0.05 \text{mm/s}$, thus avoiding the incorporation of unintentional direction reversals. As shown in Chapter 4, the results of this metric closely aligned with the results of the verified metrics, suggesting that it may be useful as a quickly calculated method of evaluation.
Chapter 3

SOM and LVQ Classification of Endovascular Surgeons*

Medical advancements in recent years have increased the popularity of endovascular surgery as an alternative to more traditional surgical methods [40]. As discussed in Chapter 1, endovascular surgery is a form of minimally invasive surgery (MIS) which allows access to various parts of the body through blood vessels and the endovascular system. The surgeon introduces a catheter into the vasculature of the patient, typically via the femoral artery, and from there navigates the catheter to the desired location so as to perform some type of procedure. During these procedures, surgeons must rely on fluoroscopy and other forms of medical imaging in order to determine tool position. This imaging is often limited, and complications may go unnoticed until they become too serious; therefore, it is imperative that surgeons be proficient at endovascular techniques. Aside from the risk of possible complications, surgeon skill level significantly affects clinical outcomes after successful surgeries [41].

As a result, there is medical interest in understanding an effective means to determine a surgeon’s skill [2]. There are presently two preeminent methods for assessing a surgeon. The most common involves an expert observing task completion by a novice, which is entirely subjective and vulnerable to significant amounts of variability [42]. The second method is simply a measurement of the number of cases performed by the surgeon; although it stands to reason that an individual with more practice will

*Large portions of this chapter appear in [39]
likely be better, it is also likely that individual surgeons will improve at different rates. Either method is insufficient, and therefore a primary goal of the endovascular community is the development of an objective assessment technique [37], [43].

In an effort to more objectively study surgeons, sensors have been used to record the tool tip trajectory [44]. The results are then processed to calculate a variety of motion-based metrics; the most indicative of these metrics are correlated to user smoothness, such as minimum jerk [45] and spectral arc length [35]. An alternative, yet similarly-minded, method is the extraction of submovement number and duration from a larger task [34]. To date, researchers have attempted to show that there exist correlations between these movement metrics and the standard methods of skill evaluation. Surgeon force and motion signatures have been leveraged to objectively assess performance; hidden Markov models were then used to learn the nonlinear mapping between performance data and skill [46]. Lin et al. demonstrated the ability to decompose a surgical procedure into a series of sub-tasks by parsing raw motion data in order to provide on-line training feedback [47]. Estrada et al. specifically quantified the correlation between various metrics and the standard methods of surgeon evaluation on both manual and robotic platforms [28].

While the statistically significant correlation between various objective metrics and current subjective assessments is an important initial finding by Estrada et al. [28], it fails to provide a holistic approach to skill classification. Hence, the motivation for my work is to understand the mapping between movement metrics and surgeon proficiency. In this chapter, I present a novel surgeon skill classification method that is based on artificial neural network machine learning techniques. Successfully mapping motion-based performance metrics to skill may improve training procedures, reduce the amount of oversight required, and ultimately automate the
task of training surgeons. I will first describe the data set and framework for the analysis. Next, I will discuss some fundamentals of machine learning, as well as the advantages and disadvantages of various techniques. Then, the results of my analysis will be discussed. I acknowledge the contributions of my collaborators on this work.

3.1 Experimental Protocol

Tool-tip trajectories were recorded for fifteen surgeons completing four separate endovascular tasks (navigating to various targets in an anatomical model) over three experimental sessions using two different experimental platforms, a virtual reality simulator and a physical model (see Figure 3.1). All subjects were familiar with the endovascular domain. Five of the subjects were deemed “novices,” six were labeled “intermediates,” and the remaining four were regarded as “experts,” with these classifications based on their previous caseload. This experiment is described in detail in [28]. For the purpose of this thesis, post-hoc analysis of the already collected data was carried out.
Figure 3.1: A comparison of the manual and virtual simulators which were navigated during the various tasks. In both platforms the surgeon is operating a catheter—in (a) the tip position is tracked using a magnet at the tool-tip, while (b) is a virtual reality simulator with tip tracking based on image processing.

**Fundamentals of Endovascular Skills Model**

The Fundamentals of Endovascular Skills (FEVS) Model was developed to serve as a standardized, inanimate model. It was designed to feature anatomically inspired geometry and therefore provide a realistic representation of the vasculature of a human
for surgeons to practice on. The model allows for testing of various skills, including but not limited to: access via the femoral artery, basic catheter and guidewire navigation, and catheterization of selected blood vessels. As detailed by Estrada et al. [48], the model is 28” in length and incorporates the labeled vasculature features (Fig. 3.2).

Figure 3.2: Fundamentals of Endovascular Skills (FEVS) Model. Serves as a standardized, inanimate model allowing for practice of endovascular navigation [11].

**Simbionix Angio Mentor Endovascular Simulator**

The Simbionix Angio Mentor Endovascular Simulator is a module of the larger Simbionix Angio Mentor Ultimate training simulator. The Angio Mentor is a computer-based training platform featuring simulated anatomies and tools. Users are able to interact and practice with the simulator in much the same way as the physical model,
even experiencing haptic feedback during the procedure. On top of these capabilities, the Angio Mentor provides C-Arm and patient table manipulation, as well as vital signs monitoring. Essentially, the Angio Mentor attempts to incorporate all of the direct and indirect aspects of a normal procedure in order to create the most realistic training environment possible.

3.2 Machine Learning Techniques

Artificial neural networks (ANNs) are biologically-inspired networks composed of large numbers of interconnected neurons. By implementing simple learning algorithms within these networks, ANNs can be utilized to approximate non-linear functions between large sets of inputs and outputs. The neuron connections are given weighted numeric values which are adjusted over the training phase of a network, thus allowing it to adapt to the dataset and simulate learning. While there are lots of possible learning algorithms to employ, the primary delineation is between supervised and unsupervised learning. For this research, both a supervised and unsupervised learning algorithm were applied to classify surgical skill and the results evaluated. Supervised learning relies upon a known, or well understood, mapping between the input and output space; however, unsupervised learning attempts to project the shape of the input data space onto a lower dimensional space, without the use of imposed classifications.

3.3 Endovascular Surgical Skill Classification Methods

The supervised learning was accomplished by first training a Learning Vector Quantization (LVQ) to classify surgeons using standardized novice, intermediate, or expert labels, and then study the LVQ’s accuracy using testing data. As discussed, super-
vised learning requires that there be a known mapping between inputs and outputs. In the case of this research, correlation had previously been shown between some of the input space and the designated surgeon classifications of novice, intermediate, and expert. Thus, these classifications were employed in the training of the LVQs. I hypothesize that the traditional “novice, intermediate, and expert” labeling—while commonly assumed to be correct—does not actually reflect the motion data, and, as such, a more sophisticated classification model is recommended.

A more sophisticated classification model requires greater knowledge of the input data, thus suggesting the employment of unsupervised learning. The unsupervised learning was accomplished by training a Self-Organizing Map (SOM). SOMs were used to examine the underlying clusters; by comparing these SOM clusters with pre-labeled classes, I can evaluate the veracity of the medically imposed class labels used in LVQ training. The secondary goal was to identify which motion patterns contribute most to the surgeon’s classification; this knowledge may improve the feedback which can be provided during and after the surgeon’s training.

3.3.1 Classification with LVQs

In order to determine the mapping from input data to desired classification, I used an LVQ with supervised learning [49]. More specifically, an LVQ2 with stratified four-fold cross-validation was employed. An LVQ learns by prototype-based supervised classification. Using a winner-take-all Hebbian learning algorithm, the randomly initialized prototypes in the network space are adjusted in order to represent the data. For each input vector, the closest prototype in the network is selected. If the selected prototype is of the same classification as the input vector, the winning prototype is moved closer towards the input vector according to the previously defined
Algorithm 1 LVQ Pseudocode

Input:

\[ X : \text{input vectors (size } N) \]
\[ \alpha : \text{learning rate} \]
\[ n : \text{max iterations} \]
\[ p : \text{number of prototypes} \]
\[ e : \text{min error} \]

Output:

\[ MAT_{final} : \text{prototype vectors} \]

1: \( MAT \leftarrow \text{Initialize prototypes over input vector data space} \)
2: \textbf{while} \( \text{learningstep} < n \ \text{AND} \ error > e \ \textbf{do} \)
3: \textbf{for} \( i = 1 \text{toN} \ \textbf{do} \)
4: \( BMU_i \leftarrow \text{nearest prototype to } X_i \)
5: \( dist \leftarrow \text{euclidian distance between } BMU_i \ \text{and } X_i \)
6: \textbf{if} \( X_i \ \text{Class} = BMU_i \ \text{Class} \ \textbf{then} \)
7: \( BMU_i \leftarrow \text{move towards } X_i \ \text{according to } dist \ \text{and } \alpha \)
8: \textbf{else} 
9: \( BMU_i \leftarrow \text{move away from } X_i \ \text{according to } dist \ \text{and } \alpha \)
10: \( error \leftarrow \text{percentage of incorrectly classified } X \ \text{using BMU} \)
11: \textbf{if} \( error < e \ \textbf{then return} \)
12: \( error_{final} \leftarrow \text{percentage of incorrectly classified } X \ \text{using BMU} \)
13: \( MAT_{final} \leftarrow MAT \ \text{after updating each learning step} \)
learning rate. Conversely, if the selected prototype is of an incorrect classification, it is moved further away from the input vector. By repeating this simple task over a large number of learning steps, the prototypes within the network will ideally relocate to better align with the input vectors.

In my case, the LVQ was initialized with 120 prototypes, as this was found to provide the best classification accuracy, where forty prototypes were allocated to novices, forty-eight were allocated to intermediates, and the remaining thirty-two were allocated for experts. Thus, this prototype allocation was done in proportion to class size. LVQ neurons were randomly initialized and scaled to the range of the input data. Ideally, the trained LVQ would adapt to the externally imposed classification structure, and, as such, would serve as an autonomous means towards identifying the class of the surgeon’s skill—novice, intermediate, or expert—based solely on motion metrics. On the other hand, the reliability of the traditionally imposed class labels may be questionable [50]. These labels are based on the number of cases performed; however, it is conceivable that a surgeon could perform a large number of cases with improper technique, and therefore be labeled an “expert” by this traditional evaluation while actually maintaining a “novice” level of ability. In order to examine the performance of the classification with LVQs, I will show confusion matrix data and statistics across all four folds, as well as a visualization of the best results.

3.3.2 Clustering with SOMs

If classification accuracy using an LVQ is found to be unsatisfactory, further analysis may be warranted. Further analysis of the input data—and, in particular, clusters present in the input data—was performed and visualized through the use of SOMs [51]. SOMs, the successor to LVQs, differ in the fact that they are unsupervised.
Algorithm 2 SOM Pseudocode

Input:

- \(X\) : input vectors (size \(N\))
- \(l\) : grid dimension
- \(\alpha\) : learning rate
- \(\sigma\) : neighborhood
- \(n\) : max iterations
- \(e\) : min error

Output:

- \(MAT_{final}\) : SOM grid

1: \(MAT \leftarrow \) Initialize \(l \times l\) grid over input vector data space

2: while learningstep < \(n\) AND error > \(e\) do

3: for \(i = 1\) to \(N\) do

4: \(BMU_i \leftarrow\) nearest prototype in grid to \(X_i\)

5: \(BMU_{\sigma} \leftarrow\) set of prototypes within the defined neighborhood of \(BMU_i\)

6: \(dist \leftarrow\) euclidian distance between \(BMU_i\) and \(X_i\)

7: \(BMU_i \leftarrow\) move towards \(X_i\) according to \(dist\) and \(\alpha\)

8: \(BMU_{\sigma} \leftarrow\) move towards \(X_i\) according to distance from \(BMU_i\) and \(\alpha\)

9: \(error_{total} \leftarrow\) total \(dist\) of prototypes from input vectors \(X\) using BMU

10: if \(error_{total} < e\) then return

11: \(error_{final} \leftarrow\) total \(dist\) of prototypes from input vectors \(X\) using BMU

12: \(MAT_{final} \leftarrow MAT\) after updating each learning step
The lack of supervision allows the network to capture internal characteristics of the data space which are not readily apparent. Usually these internal characteristics are represented by an SOM in the form of a two-dimensional map. Similar to an LVQ, the weights of the two-dimensional network space are randomly initialized. Over the course of the learning steps, the closest neuron (prototype) is selected as the winner and is adjusted according to a learning rate to move closer towards the input vector. Rather than the winner-take-all approach of the LVQ, an SOM features some type of neighborhood function to move the neighboring neurons closer to the input vector as well. This creates clusters within the data and ensures learning and generalization of the network instead of simple memorization.

In my case, I leveraged forty-nine prototypes for the SOM, which were arranged into a seven-by-seven rectangular grid in the lattice space. A Gaussian neighborhood function was used while updating the prototypes, and mU-matrix visualization was employed to visualize clusters. The rationale for using an SOM was to capitalize upon the strengths of unsupervised learning; I sought to obtain an objective view of the data structure without needing potentially erroneous labels. Therefore, there were two primary goals behind this SOM application. First, I wanted to validate or disprove the classification labels (novice, intermediate, and expert) previously used for the LVQ training. By superimposing these labels over the SOM lattice while visualizing SOM clusters, I could test label veracity and hopefully understand why the LVQ machine learning underperformed. Second, I wanted to identify clusters within the data in order to determine the relative importance of surgeon attributes and motion methods when distinguishing between skilled and unskilled surgeons. Comparison of the input vectors associated with different clusters will clarify which motion metrics were consistent and which varied amongst clusters. These insights may enable more
efficient evaluation of surgeons and more directed training strategies. SOM clustering will be revealed through plots of the lattice space.

3.3.3 Input Data and Class Labels

The data for this research were recorded for fifteen surgeons over three sessions while completing four separate tasks. For the purposes of this research, I did not differentiate between the platforms (manual or virtual), sessions, or tasks, yielding a total of 120 separate trials. The motion metrics associated with each trial—described previously—were then utilized as a unique input vector; hence, results were obtained using 120 input vectors. Metrics which were not previously described are explained in the original work by Estrada et al. [28].
Figure 3.3: Each of the plotted points corresponds to one of the eleven motion metrics derived from the surgeon’s trajectory. Average input vectors associated with a novice, intermediate, and expert surgeon are shown (standard error bars included).

The input vectors for the LVQ and SOM neural networks were constructed from previously calculated motion metrics (Fig. 3.3). These metrics were all computed from the three-dimensional catheter position data, which was collected at 30 Hz frequency. Based upon the findings of Estrada et al. [28], I selected motion metrics which were shown to individually correlate with traditional skill labels. Eleven metrics (listed below) were chosen, and each comprised an element of the eleven-dimensional input vectors. Although the units for the various metrics are not detailed here, it should be noted that they were kept consistent throughout the work. I found that the best
results occurred with the inclusion of:

1. Spectral Arc Length

2. Average Submovement Duration (LGNB profile)

3. Average Submovement Duration (MinJerk profile)

4. Number of Submovements (LGNB profile)

5. Number of Submovements (MinJerk profile)

6. Normalized Velocity

7. Mean Arrest Period Ratio (10% threshold)

8. Completion Time

9. Submovement Overlap (LGNB profile)

10. Submovement Overlap (MinJerk profile)

11. Average Frequency

Example input vectors can be seen in Fig. 3.3. Note that these values are all well defined over a continuous range, and that the chosen metrics mitigated statistical outliers which may skew results of the input-space neural networks.

Each of these 120 input vectors was associated with a class label corresponding to the surgeon’s proficiency; the three classes consisted of either “novice,” “intermediate,” or “expert.” Forty input vectors were labeled novice, forty-eight input vectors were denoted intermediate, and thirty-two input vectors were termed expert. When performing supervised learning, stratified four-fold cross-validation was leveraged to
select exclusive sets of ninety input vectors for training and thirty input vectors for testing.

3.4 Results and Discussion

3.4.1 LVQ Classification Results

Using an LVQ, I consistently found that I was able to differentiate the skill groups and correctly classify surgeons within the novice, intermediate, and expert labels 80% of the time for training data and 50% of the time for testing data. Table 3.1 shows the mean classification percentage and standard deviation for a four-fold cross validation.

Table 3.1: Classification Summary. The mean and standard deviation for training and testing accuracy. Poor results likely stem from incorrect class labels, particularly between intermediates and experts.

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>81% (291)</td>
<td>51% (61)</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>4.2% (15)</td>
<td>9.5% (11)</td>
</tr>
</tbody>
</table>

By inspecting the confusion matrices, summarized in Table 3.2, novices were reasonably distinguished from intermediates and experts, but intermediates and experts were largely lumped together.
Table 3.2: Average confusion matrix over the four folds. Data is given in the form % of hits (number of hits). Diagonal elements represent correctly classified data, while off-diagonal elements show incorrect classifications.

| Labels     | | Training Results | | Testing Results |
|------------|| Novice | Intermediate | Expert | Novice | Intermediate | Expert |
| Novice     | | 87.5% (26.25) | 10% (3) | 2.5% (0.75) | 70% (7) | 15% (1.5) | 15% (1.5) |
| Intermediate| | 6.25% (2.25) | 81.9% (29.5) | 11.8% (4.25) | 20.8% (2.5) | 52.1% (6.25) | 27.1% (3.25) |
| Expert     | | 2.1% (0.5) | 27.1% (6.5) | 70.8% (17) | 9.4% (0.75) | 65.6% (5.25) | 25% (2) |

This result is further verified by Figure 3.4(b) wherein there is little distinction between the intermediate (colored grey) and expert (colored white) surgeons in the testing data.

While the definition of acceptable classification is largely application and subject matter dependent, at least 90% would typically be desired given the importance of surgeon qualification in an endovascular procedure. The results obtained by implementing an LVQ were reasonable, but did not provide sufficiently accurate classification given this criteria for the purposes of automated evaluation (Table 3.1). The best results were obtained with an LVQ2 using a learning rate of 0.001 and 10,000 on-line learning steps, although other learning rates and learning step counts were tested. Both the training and testing accuracy were plotted as a function of learning steps to ensure that overtraining did not occur. It is important to also note that while LVQ1, LVQ2, and LVQ3 were tested, there was not significant variation among the performance of these algorithms.
Figure 3.4: Sample LVQ results from one fold of the four-fold stratified cross-validation procedure: training classification left; testing classification right. The black pixels represent novice surgeons, the grey pixels represent intermediate surgeons, and the white pixels represent expert surgeons. The top plot is the desired classification, the middle plot is the actual classification, and the bottom plot is the error.
In particular, the LVQ struggled to distinguish “intermediate” from “expert” surgeons, logically suggesting a larger skill gap from novice to intermediate than from intermediate to expert – depicted in Figure 3.4. I hypothesize that this stems from at least partially inaccurate training labels; the imposed classifications may not truly identify the skill level of each surgeon, since intermediate surgeons, despite having performed fewer cases than experts, may be more proficient than their caseload suggests. Moreover, the use of only three classes is likely insufficient to accurately capture the gradient in surgeon skill, and perhaps more nuanced labels would better reflect the motion data. The overall statistics show that the LVQ procedure netted consistent and accurate training classification, but the testing accuracy and hence machine learning was unacceptable. I conclude that the LVQ was unable to generalize for the given data, and suggest that this inability stems from the lack of labeling precision and correctness for intermediate and expert surgeons. To verify this claim, SOM clusters are explored in the data space.

3.4.2 SOM Clustering Results

Following the failure of LVQs to successfully identify this mapping, SOMs were applied to both test the concerns with the imposed classification labels and help to further explore nuances within the data. Using an SOM, I was able to consistently identify specific clusters within the data space despite using varying initialization conditions.

The best results presented in this chapter were obtained using a seven-by-seven rectangular SOM grid in lattice space, where the forty-nine prototypes were initialized randomly over the input space. The learning rate $\alpha$ started at 0.005 and reached 0.001 following a linear decrease across 100,000 learning steps; similarly, the Gaussian neighborhood width $\sigma$ started at 4 and linearly decreased to 2 over the same number of
learning steps. I experimentally observed the SOM training to converge after around 80,000 to 90,000 on-line learning steps, at which point no changes occurred in the mapping (Fig. 3.5). The results shown in Figure 3.6 were found to be repeatable and superior to those identified using different parameters, which provides confidence in the subsequent conclusions.

Figure 3.5: SOM learning history using the same visualizations as in Figure 3.6. Each pair of visualizations from left to right represents 10,000 learning steps, starting at 10,000 on the left. Given that there is little observable change between the visualizations at 90,000 and 100,000 learning steps, and that both appear to be similar to the network at 80,000 learning steps, this illustrates that the training has converged.
Figure 3.6: SOM final results. Both visualizations provide insight into the final weight values of the 49 prototypes (7 by 7 grid) and how the data space maps onto the SOM grid. The left visualization is a modified U-Matrix [52]; red-scale represents the relative number of data vectors mapped to each prototype (i.e., density), while the gray-scale bars signify the distance between prototypes in the data space. The redder the neuron, the more input vectors are contained within its Voronoi cell; likewise, the darker the bar separating neurons, the greater the difference between their weight vectors. The right visualization shows the known surgeon classifications projected onto the SOM lattice—here red signifies novice, green represents intermediate, and blue indicates expert surgeons, with color intensity representing the number of mappings (more intensity again means increased density). Black neurons indicate that no input vectors are mapped to a particular node. The clusters found in the mU-matrix are identified using white lines in the right visualization. While novices (red) are primarily separated, clustering in the upper and lower left, intermediates (green) and experts (blue) are largely intermingled, clustering along the right side, a result which supports the LVQ findings.
Selecting the learning parameters as described above while observing the system visualizations depicted in Fig. 3.6, the lattice repeatedly converged to a similar, if not the same, solution each time I trained the SOM. Instances in which the lattice did not converge to the results outlined in Fig 3.6 involved some type of rotation of the lattice—however, this did not alter the SOM clustering. Using U-Matrix techniques, I readily discerned some distinct clusters which were identified by the SOM; I then checked these locations with superimposed novice, intermediate, and expert labels in the lattice space, and determined whether there existed agreement between medically defined clusters and clusters identified by the SOM.

From the modified U-Matrix density map and the projection of classifications into lattice space, I can deduce (a) that there exist some SOM clusters which roughly correspond with traditional groups, but (b) other SOM clusters disagree with the medical consensus. These SOM identified clusters are marked in Fig. 3.7. For instance, the bottom left section of the SOM lattice clearly clusters several surgeons who performed poorly, and are correctly labeled as novices. Likewise, the top left SOM cluster corresponds to another group of novice surgeons, which again matches the medical labeling. Moving to the right side of the SOM lattice, however, there are two regions: in the upper right, there exists a mixed cluster—some experts, intermediates, and novices are included here, suggesting labeling inaccuracy. Finally, in the bottom right of the SOM lattice there is a cluster of increasing ability, with intermediates and experts grouped together; perhaps these surgeons are closer in ability than their classification would suggest. By applying SOMs to the input space of motion metrics, I was therefore able to demonstrate that a surgeon’s experience is not sufficient when attempting to classify that surgeon’s skill. Although there are some similarities between the medical labels and SOM clusters, there is also sufficient
disparity to suggest that perhaps more precise skill assessment is required. These findings also explain the inability of the LVQs to distinguish “intermediate” and “expert” surgeons, as SOM clusters revealed overlaps between these classifications.

Figure 3.7: SOM weight vector plotted in the grid cells. This figure shows both the final results of the SOM grid with the known classifications projected onto the lattice, as well as the weight vector of each PE with respect to the average weight vector across all nodes. The weight vector of a given PE is shown in black, while the average weight vector across all nodes is plotted in a dotted magenta line. The color coding of the prototypes is the same as before, with a slight fading of the colors in order to better visualize the weight vectors. Black boxes were used to mark the SOM cluster boundaries identified in Figure 3.6.

In order to further investigate clustering and the distinctions between various groups, it was instructive to look at the weight vector within these individual clus-
ters, as illustrated in Fig. 3.7. There are a few hypotheses which can be formed from visualizing these prototypes and clusters. First, completion time is not necessarily an accurate measure of skill. In fact, completion time appears to be somewhat counter-intuitive; experts often take longer than less successful intermediates and novices, perhaps because they are utilizing slower and more deliberate movements. A quick procedure is ideal, but not if it comes at the cost of deliberate, precise movements. Second, some metrics may provide redundant differentiation, therefore requiring the use of fewer metrics—and other metrics may be entirely irrelevant for classification purposes. Finally, the number of submovements appears to be particularly useful when distinguishing surgeons; the most proficient cluster employed substantially smoother motions than did novice or mixed clusters.

By comparing the differences in weight vectors between members of different clusters, one can visualize which metrics most impact distinctions in surgeon skill. With respect to the average weight vector, novices appear to complete the task in less time but require an increased number of motions; on the other hand, proficient surgeons move slowly but smoothly, reducing submovement duration and number. The combination of SOM clustering and neuron weight vectors reveals errors within traditional labeling and provides insight into important motion attributes. The existing labeling of novice, intermediate, and expert does not agree with knowledge gained through motion metrics (as shown by differences in clustering), and the contribution of various metrics can be analyzed to yield better categorization (as shown by comparing weight vectors).

With these ideas in mind, one can describe five classes of surgeons identified from SOM clustering (Fig.3.7). Class one (lower left) will perform the task slowly and with very little smoothness; likely true beginners. Class two (upper left) will
perform the surgery quickly with little smoothness; likely novice surgeons. Class three (upper right) will perform the surgery quickly at the expense of some smoothness metrics; likely competent surgeons primarily concerned with completion time. Class four (middle right) will perform the surgery above average in terms of time and smoothness; likely experienced surgeons. Class five (lower right) will perform the surgery at an average pace with exceptional dexterity; likely skilled, precise surgeons.

3.5 Conclusions

Based on the results of the LVQ and SOM, there does appear to be some consistent mapping between motion metrics and desired classification; using the LVQ, around 50% testing accuracy was achieved. I hypothesized that this poor LVQ machine learning, particularly when discerning between intermediate and expert surgeons, stemmed from inaccurate class labeling. Using the SOM approach, I was able to identify some clusters which roughly corresponded to the known classification groups; however, I also discovered that several clusters disagreed with the given labels. Indeed, from Fig. 3.6 I was able to conclude that the traditional labeling based on surgeon experience disagreed with SOM clustering in the motion metrics. I was further able to suggest which metrics may best be able to indicate ability, as can be seen in Fig. 3.7. By replacing the subjective medical grouping with the actual measured features, one may be able to improve on skill assessment for endovascular surgeons. Similar to the work by Cotin et al. [37], this suggests that it may be better to first identify statistics which are significant to expert clusters, and then create a scoring system which classifies users based on their accordance with those statistics. Summarily, SOM clustering, as seen in Figure 3.7, helps accomplish the goals of both disproving classical labels and suggesting improved alternatives.
Chapter 4

Objective Analysis of Augmented Visual Feedback during Robotic Catheterization*

While it is important to have the capability to objectively differentiate between surgical skill levels, an equally important goal is to evaluate the extent to which technological interventions can augment surgical skill. This chapter explores visualization tools that provide the surgeon with greater information about the anatomy of the patient and the present location of the tool-tip using electromagnetic (EM) tracking and the utility of the previously discussed objective metrics to evaluate the surgeon’s performance relative to the control condition, traditional navigation using single plane fluoroscopy.

EM tracking has been used to enhance visualization in several clinical applications, especially in neurosurgery [54], cardiac electrophysiology [55], [56], bronchial [57], vascular [58] and spinal interventions [59]. Valderrabano et al. showed significant reduction in median fluoroscopy time when using an EM-assisted navigation system for left ventricular lead implants in canine models: 6 seconds versus 96 seconds for conventional navigation ($p < 0.001$) [55]. In a pilot study including 17 patients undergoing Endovascular Aneurysm Repair (EVAR), Manstad-Hulaas et al. compared the catheterization of the contralateral leg of the main body using standard 2D fluoroscopy versus EM guidance [60]. They found that the number of attempts to gain access to the contralateral gate was significantly lower when using EM guidance. How-

*Large portions of this chapter appear in [53]
ever, there were no differences regarding contrast volume or total procedure time, and the radiation dose was significantly higher during procedures involving the EM system [61], primarily because the variables included the extra radiation and time spent acquiring the intra-procedural CT scan and testing the novel navigation technology system.

Other imaging technologies aim to significantly reduce fluoroscopy time and radiation exposure. Very recently it was demonstrated that intraoperative image registration between preoperative Computed Tomography Angiography (CTA) and intraoperative cone beam Computed Tomography (CT) when performing Fenestrated Endovascular Aneurysm Repair (FEVAR) did significantly reduce radiation exposure and fluoroscopy time [62], but 40 to 60 minutes of fluoroscopy were still required for each procedure. In addition, combined usage of pre-operative image fusion and endovascular robotics has been recently shown to reduce contrast agent usage and facilitate cannulation [10].

In this chapter, the applicability of motion-based metrics to evaluate the value of EM-based visualization during robotic catheterization is studied. Guidewire movements were tracked and motion-based metrics were computed for traditional navigation using single plane fluoroscopy and three enhanced viewing modes. These enhanced modes are expected to augment surgical performance by reducing procedure time, radiation exposure, and tool-vessel interactions through 3-D visualization of target anatomical structures. I acknowledge the contributions of my collaborators.

4.1 Platform

This study was conducted using the Hansen Medical Magellan robotic catheterization system. The system was designed to facilitate vessel navigation, guidewire manipula-
tion, and endovascular device delivery by means of a steerable catheter tip, controlled through the use of a remote workstation control interface. The functionality of the system relies upon the concept of a guidewire running through a leading catheter which telescopes within a flexible sheath (Fig. 4.1).

![Figure 4.1: Magellan Catheter showing the guidewire running through the leading catheter (leader) telescoping within a flexible sheath.](image)

Since its approval by both the FDA and CE (an abbreviation for the French phrase “Conformité Européenne,” meaning “European Conformity”), the Magellan Robotic System (MRS) (Magellan, Hansen Medical Inc., Mountain View, CA) is increasingly used to support endovascular navigation, especially in complex and challenging vascular procedures. The MRS enables enough stability to reliably use the catheters and wires to navigate through rather than interact with the vessel wall, reducing wall injuries [12], [13]. Its capability to be remotely steered with six degrees of freedom offers potential benefits: better catheter orientation and maneuverability facilitating
vessel cannulation [14], as well as reduction in radiation exposure [15].

One main limitation of the current robotic catheter technology is the lack of real-time 3-D localization of the catheter tip. Even though the catheter has 3-D manipulation capabilities, it is still navigated based on 2D X-ray fluoroscopic projection images. Most of the time, multiple orthogonal fluoroscopic views are needed to deduce the exact 3-D location of the catheter in the vascular anatomy. The need to make a number of radiographic adjustments during a procedure can potentially pose a number of logistical issues depending upon the situation, and at a minimum exposes the patient and interventionalist to increased radiation. Although the robotic catheter provides data on the degrees of flexion and rotation of the catheter tip, it is only able to display the commanded input and not the real-time position and orientation of the tool tips, thus making it difficult to establish a precise 3-D orientation or to trust that a commanded input is achieved. This void of knowledge is where EM sensing technology has the potential to benefit such procedures.

Electromagnetic (EM) tracking systems have been used as a tracking technology in medical applications since the mid-2000s [55], [54], [56], [59]. These systems localize small EM field sensors in an EM field of known geometry, similar to the fields used in magnetic resonance imaging. The EM field is generated at some point in space and detected at a remote location. The whole system consists of (1) an electro-magnetic-field generator, (2) a compatible 3-D sensor, which is fixed at a remote location in the body and measures the fields generated by the source, and (3) a processor whose function is to relate the signals from the source and the sensor. So far, the Aurora tracking system (Aurora Window Field Generator, Northern Digital Incorporated, Waterloo, Ontario, Canada) is the only EM tracking system that has been studied in ex-vivo and pre-clinical models [63]. The accuracy of the tracking system was eval-
uated using target registration error (TRE), which measures the difference between the location of the tracked device calculated by the tracking system and its actual position. Recent EM field generators with open rectangular frames (window field generators) that have been properly integrated into patient tables, without causing any fluoroscopic imaging interference, have been shown to demonstrate a TRE of $1.28 \pm 0.79$ mm in a rigid model study [64]. In the most recent animal studies, the accuracy reached a TRE of $4.18$ mm [60].

As shown in Figure 4.2, the Magellan system is mounted to the operating table and the surgeon is located at a remote workstation where all of the necessary data is presented to him or her. At this station, the surgeon is able to operate the C-arm, activate the fluoroscopy, and drive the robotic catheter. The robotic catheter is controlled using a keypad with controls for rotation, extension, and flexion.
Figure 4.2 : Magellan Robotic System showing the operating table, C-arm, and remote workstation from which the surgeon operates the robotic catheter. The surgeon is presented with imaging of the procedure on the screen depending upon the chosen viewing mode and is given a control panel, as well as foot pedals, for controlling the robotic catheter and fluoroscopy.

4.1.1 Viewing Modes

The studies performed on the Hansen Magellan system featured four types of viewing modes. Each of these viewing modes offers different advantages and disadvantages to the surgeons.
Figure 4.3: Magellan Robotic System viewing modes. a) Traditional Single Plane Fluoroscopy (Control/Fluoro) b) Electromagnetic Fluoroscopy Biplane (EM Biplane) c) Electromagnetic 3-D Endovascular (EM Endo) d) Electromagnetic 3-D Biplane (EM 3-D)
The first viewing mode is the traditional single-plane fluoroscopy image (Fig. 4.3 (a)). In this viewing mode, the operator must control the C-arm to the desired orientation and when the fluoroscopy is active, the operator sees a black and white two-dimensional representation of the model. The advantage of this viewing mode is its familiarity. This mode is used in both robotic and manual catheterization, and experienced surgeons are accustomed to successfully performing surgeries with this viewing mode. One disadvantage of the viewing mode is the limited information presented about the catheter position, as it is a flattened two-dimensional representation of the anatomy and tool-tip. The limited visual fidelity often results in longer procedure times and radiation exposure.

The second viewing mode is an electromagnetic (EM) fluoroscopy biplane. In this mode, the surgeon is presented with a biplane view of the model. As shown in Fig. 4.3 (b), the views used to create the biplane are stored fluoroscopy images with EM information overlaid to provide real-time information about the position of the guidewire, catheter, and sheath without the use of active radiation. In addition to the advantageous decrease in radiation exposure since the tool tips can be visualized without need for contrast injection fluoroscopy, this mode also offers the ability for the surgeon to better conceptualize the exact position of the tool-tip as he or she is able to simultaneously view the position in two planes. Furthermore, this mode also incorporates the traditional fluoroscopy view, giving it a sense of familiarity. The disadvantage of this view is that the secondary plane cannot be updated without having to maneuver the C-arm and store the image again. This can create delays in the procedure if the operator needs to switch back and forth between C-arm positions.

The third viewing mode is an EM 3-D Endovascular mode (Fig. 4.3 (c)). In this mode, the operator views the catheter from the angle of a driver in a car and attempts
to maneuver the catheter from a first-person perspective through the vasculature of the model. This mode likely has the steepest learning curve because it is based off an entirely different perspective than the others. That being said, its videogame-esque design could make it a more natural mode with younger surgeons [4].

The fourth viewing mode is an EM 3-D biplane. Much like the EM fluoroscopy biplane, the surgeon is presented with a biplane view of the model. In this case, however, the model is entirely rendered onto the screen. As an additional benefit, the biplane views can be rotated on the screen at the discretion of the operator with the secondary view updating itself to remain orthogonal to the primary view (Fig 4.3 (d)). While this view lacks some familiarity with experienced surgeons, it can operate independent of the C-arm, thus saving time and effort in this regard.

4.2 Pilot Study

The first step in evaluating the usefulness of these viewing modes was to conduct a pilot study. In order to justify subsequent research into the matter, this pilot study needed to show that there is a quantifiable difference in subject performance when using the alternative viewing modes relative to the traditional method. Leveraging the work of Estrada et al. [28], and the previously presented work in this thesis, quantitative motion metrics were used in order to objectively determine surgeon performance. Using a pre-selected variety of motion metrics, simple statistical analysis was completed to quantify the effect of these alternative modes. Given the exploratory nature of the study, strong statistically significant findings were not expected. Rather, the aim was to obtain justification for a larger study.
4.2.1 Methods

Subjects  Six users with various endovascular robotic expertise drove the MRS: two expert vascular surgeons with a minimum of 20 hours of training on the system and experience in performing clinical cases using the robotic platform, and four beginners with only limited training on the robotic system: one vascular surgical fellow and three vascular surgical junior residents.

System  A standard 9Fr† Magellan robotic catheter was specifically modified for the purpose of the study: one EM sensor, composed of two sensing coils, was integrated at the base of the leader articulation section and a tip EM sensor was embedded at the distal extremity of the leader, within its lumen. There was no EM tracking of the robotic sheath. Because the center lumen of the catheter was occupied by the tip sensor, a guide wire was not used during the procedure.

An EM field was generated by a window field generator (placed under the angiography table) and a specific processor related the signals from the source and the sensors to track the sensors in three-dimensional space.

†The French Scale is commonly used to measure the size of a catheter. The French size is simply three times the diameter in millimeters, thus a 9 French catheter is a 3 mm diameter catheter.
Figure 4.4: Rigid fluid-filled aortic aneurysmal phantom used during the procedure with a) EM field generator placed under the angiography table b) Robotic Catheter and c) removable simulated contralateral gate

A rigid fluid-filled aortic aneurysmal phantom was used for this study (Fig 4.4). It consisted of the aorto-iliac bifurcation and left and right renal arteries. It also included a removable simulated gate oriented towards the posterior wall of the aneurysm model.

A 3-D image of the phantom was generated from the C-arm cone-beam CT images acquired using a robotic angiography system. A virtual model of the phantom was created and sent to the EM system for co-registration: this was achieved by placing an EM sensor on a few known points in the phantom and manually aligning them to the virtual model during the setup process.

**Procedure** Subjects were asked to navigate the MRS and to cannulate two targets in the aortic aneurysm phantom: the left renal artery, qualified as a simple cannulation target, and the simulated posterior gate, qualified as a complex cannulation target. Four different visualization modes were used for each cannulation (Fig 4.3):
• Mode 1: Standard 2-D Fluoroscopy (control).

• Mode 2: EM Fluoroscopy Biplane

• Mode 3: EM 3-D Endovascular

• Mode 4: EM 3-D Biplane

To avoid bias resulting from learning process, the order of visualization mode for each cannulation task and each user was randomly assigned. Standard X-ray fluoroscopic imaging was always available during navigation, regardless of visualization mode.

**Performance measure**  Success was defined as positioning the distal tip of the robotic sheath beyond the origin of the target and was verified by fluoroscopic imaging. To quantify radiation exposure, the duration of X-ray usage required during a procedure was recorded as fluoroscopy time. The overall time needed to perform the task was recorded as cannulation time.

Data Analysis was completed using kinematic metrics evaluating the efficiency, consistency, and smoothness of the catheter tip motion [11], [28]. The metrics were calculated from the position data of the tip sensor, reporting at a 4Hz frequency, using the EM tracking system. The selected kinematic metrics were:

1. 3-D Path Length (3-D)

2. Spectral Arc Length (SAL)

3. Root Mean Dimensionless Jerk (RMDJ)

4. Number of Submovements (SubMoveNum)
5. Submovement Duration (SubMoveDur)

6. Catheter Turns (150° or 165°)

**Statistical Analysis**  For each measurement of performance, a mixed design ANOVA was used to analyze the within-subject effects of task and platform, as well as the between subject effect of skill level. For each cannulation task, the results of standard fluoroscopy (mode 1) were compared to the average of the three modes using EM tracking (modes 2,3, and 4) using a contrast t-test.

### 4.2.2 Results

In this single center prospective in vitro study, all six subjects successfully cannulated the two targets. The catheter with an incorporated EM sensor navigated as expected according to all users: inclusion of the sensors inside the catheter, near the catheter tip did not modify or limit catheter articulation or rotation motion.

**Cannulation and Fluoroscopy Times**  For surgeons, the primary concerns are cannulation and fluoroscopy times. These are easily measured and discussed results which allow them to evaluate performance using a common language. Because there is an element of luck involved in how quickly each cannulation trial can be completed, a surgeon’s tool manipulation smoothness and dexterity is a more objective indication of general performance. Similarly, smoothness and dexterity are likely more consistent from trial to trial, thus making them more robust to outlier trials. That being said, it is still important to discuss the cannulation and fluoroscopy times.

For the left renal artery target, mean cannulation times (mm:ss) were 3:55, 3:38, 2:02 and 2:12 for modes single-plane fluoroscopy (SPF), EM fluoroscopy biplane
(EMFB), EM 3-D endovascular (EM3E), and EM 3-D biplane (EM3B), respectively. Mean fluoroscopy times were 129, 16, 1 and 2 seconds respectively. There were no significant main effects of viewing mode or task, likely due to the small number of subjects for the study, therefore contrast t-tests were required to detect any effects within the study. For a simple cannulation target (e.g. renal target), there was no statistically significant difference between standard fluoroscopy mode and the use of EM tracking. For the posterior gate target, mean cannulation times (mm:ss) were 8:12, 4:19, 4:29 and 3:09 respectively for the modes SPF, EMFB, EM3E, and EM3B. Mean fluoroscopy times were 274, 20, 29 and 2 seconds respectively. For a complex cannulation target (e.g. contralateral gate equivalent target), the use of EM tracking modes for the posterior gate cannulation significantly reduced both cannulation and fluoroscopy times when compared to standard fluoroscopy mode ($p = 0.013$ and $p = 0.001$ respectively). There were no statistical differences in fluoroscopy or cannulation times for both tasks when comparing the three visualization modes involving EM tracking (modes EMFB, EM3E, and EM3B). Contrast $p$ values for these metrics and others are listed in Table 4.1.
Table 4.1: A contrast t-test was run for each metric evaluating the control mode (SPF) against the three EM modes (EMFB, EM3E, and EM3B) using the conventional \((-3, 1, 1, 1)\) contrast to test whether the means were significantly different. These contrast t-tests were evaluated across both tasks as well as for each task individually. Bolded values were significant.

<table>
<thead>
<tr>
<th>Metric</th>
<th>3-D</th>
<th>150°</th>
<th>165°</th>
<th>RMDJ</th>
<th>SAL</th>
<th>S(D)</th>
<th>S(N)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Tasks</td>
<td>.017</td>
<td>.021</td>
<td>.082</td>
<td>.091</td>
<td>.054</td>
<td>.0933</td>
<td>.030</td>
<td>.069</td>
</tr>
<tr>
<td>Task 1</td>
<td>.269</td>
<td>.634</td>
<td>.764</td>
<td>.414</td>
<td>.839</td>
<td>.935</td>
<td>.480</td>
<td>.388</td>
</tr>
<tr>
<td>Task 2</td>
<td>.009</td>
<td>.043</td>
<td>.032</td>
<td>.009</td>
<td>.017</td>
<td>1.00</td>
<td>.002</td>
<td>.013</td>
</tr>
</tbody>
</table>

**Motion Metrics** The motion metrics tell a similar story to that of the cannulation and fluoroscopy times; the use of EM tracking and visualization during the left renal artery cannulation task failed to show a significant effect on performance as measured by each of the specific metrics. On the other hand, 3-D path length, spectral arc length, root mean dimensionless jerk and the number of submovements and catheter turns were all significantly improved with EM tracking and visualization of the tip of the robotic catheter during complex cannulation of the posterior gate (Table 4.1). Figure 4.5 illustrates the 3-D path length trajectory of the tool-tip while cannulating the simulated posterior gate for each four different visualization modes.
Figure 4.5: Sample trajectories from a subject using the various viewing modes. It should be noted that the tool-tip motion for Mode 1 appears to be significantly greater in length and less intentional than the tool-tip motion using the other viewing modes.

4.3 Follow-Up Study

After concluding that the initial findings warranted further evaluation, a follow-up study was organized and completed. This study attempted to rectify some of the limitations and shortcomings of the pilot study, while simultaneously increasing the number of participants and thus providing a greater power to detect statistical significance.
4.3.1 Procedural Modifications

The method of data collection for the follow-up study closely resembled the initial study, but slight modifications were made. Namely, the 3-D Endo viewing mode was removed in favor of a newly-developed assisted driving feature. The assisted driving feature incorporates coding which can assist the operator during cannulation and is implemented as an overlay for the EM Fluoro Biplane viewing mode from the previous study. An additional change was an increase in data collection frequency. It was determined that 4 Hz may have limited the ability to properly extract the motion profile of the subjects; therefore, the position data of the EM sensors were recorded at a frequency of 30 Hz. This corresponds with the data rate used by Estrada et al. in determining metric correlations [28].

The other modification of the experiment was the inclusion of a guide wire. The group at Hansen was able to incorporate EM sensors into the guide wire, the leader, and the sheath in order for the system to operate properly in the absence of fluoroscopy. Because of the required articulation within the leader and sheath, the EM sensors in these pieces of the apparatus were placed approximately 2.5 cm back of the respective tips of the leader and sheath. Within the guidewire, the EM sensor was placed approximately 1 cm back of the tip. Additionally, this was the first implementation of a guidewire with the system; a significant step in product development given the importance of a guidewire in cannulation.

4.3.2 Results and Discussion

Significant effects of EM tracking and visualization on guidewire navigation were evaluated using a mixed design ANOVA and computing statistics for each of the motion-based metrics discussed in Chapter 2. Motion-based metrics were calculated
for the guidewire cannulation using the sensor in the tip of the guidewire. The mixed
design ANOVA accounted for the within subject factor of viewing modes and tasks,
and the between subjects factor of skill level. Results from this statistical analysis
are presented in Table 4.2.

Table 4.2 : Guidewire Mixed Design ANOVA - all modes. The table shows the p-
values for each metric given the within subject and between subject factors. Each
factor is labeled with its respective degrees of freedom. Significant results are bolded.

<table>
<thead>
<tr>
<th>Metric</th>
<th>3-D</th>
<th>150°</th>
<th>165°</th>
<th>RMDJ</th>
<th>SAL</th>
<th>S(D)</th>
<th>S(N)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task (1,15)</td>
<td>.009</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.001</td>
<td>.001</td>
<td>.837</td>
<td>.001</td>
</tr>
<tr>
<td>View Mode (2,30)</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td>.007</td>
<td>.154</td>
<td>.014</td>
<td>&lt; .001</td>
<td>.001</td>
</tr>
<tr>
<td>Interaction (2,30)</td>
<td>&lt; .001</td>
<td>.005</td>
<td>.001</td>
<td>.001</td>
<td>.045</td>
<td>.652</td>
<td>.013</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Skill (2,15)</td>
<td>.282</td>
<td>.438</td>
<td>.347</td>
<td>.471</td>
<td>.246</td>
<td>.363</td>
<td>.145</td>
<td>.351</td>
</tr>
</tbody>
</table>

Results of the Mixed Design ANOVA, presented in Table 4.2, show that there
is a significant effect of task and viewing mode, as well as a significant interaction
between the two. For each of these, only one metric is not significant, and the
metric which is not significant is different for each of them. The non-significant
metrics for task, viewing mode, and the interaction between the two are number of
submovements, spectral arc length, and submovement duration, respectively. Based
upon the detected significance of task, this is sufficient to conclude that one task
is completed more easily and smoother than the other. Looking at the results in
Figure 4.6, it is apparent that task one is significantly lower difficulty than task
two, as intended by study design and corroborated by feedback from the subjects.
Using all of the metrics lends more confidence to conclusions drawn from the data, as the findings were substantiated by significant results over an array of metrics, each detecting different movement qualities within a subject’s task completion. While the pilot study failed to find any significant main effects due to its small sample size, these limitations appeared to be mitigated in this study.

![Graphs showing catheter turns and 3-D length](a) and (b)

Figure 4.6: Comparison between tasks one and two showing the apparent difference in difficulty between the two. The larger number of catheter turns (a) and longer path lengths (b) for task 2 indicate higher level of difficulty for task 2 compared to task 1.

T-tests were calculated to test whether or not subjects performed differently on the EM viewing modes EMFB, EM Fluoroscopy Biplane with assisted driving (EMFB+), and EM3B relative to the control viewing mode, SPF. In addition to these t-tests, another t-test was introduced for this study. There was concern that the newly incorporated assisted driving mode may skew the motion metrics which were designed to evaluate human motor control. Therefore, a contrast between viewing mode SPF and viewing modes EMFB and EM3B was tested.
Table 4.3: Guidewire Mixed Design ANOVA - SPF, EMFB, and EM3B only. The table shows the p-values for each metric given the within subject and between subject factors. Each factor is labeled with its respective degrees of freedom. Significant results are bolded.

<table>
<thead>
<tr>
<th>Metric</th>
<th>3-D</th>
<th>150°</th>
<th>165°</th>
<th>RMDJ</th>
<th>SAL</th>
<th>S(D)</th>
<th>S(N)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task (1,15)</td>
<td>.002</td>
<td>.002</td>
<td>.001</td>
<td>&lt; .001</td>
<td>.002</td>
<td>.316</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td>Viewing Mode (2,30)</td>
<td>&lt; .001</td>
<td>.001</td>
<td>.004</td>
<td>.006</td>
<td>.733</td>
<td>.030</td>
<td>&lt; .001</td>
<td>.001</td>
</tr>
<tr>
<td>Interaction (2,30)</td>
<td>&lt; .001</td>
<td>.013</td>
<td>.001</td>
<td>.004</td>
<td>.127</td>
<td>.703</td>
<td>.024</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Skill (2,15)</td>
<td>.356</td>
<td>.474</td>
<td>.417</td>
<td>.699</td>
<td>.486</td>
<td>.419</td>
<td>.169</td>
<td>.562</td>
</tr>
</tbody>
</table>

Contrasts for all modes are presented in Table 4.4, and the contrasts for human motion only modes are presented in Table 4.5. As illustrated by both tables, the t-test is significant for almost all metrics during Task 2, in contrast to the case of Task 1. Spectral arc length appears to be the only metric to not have a significant effect on Task 2 or both tasks. This suggests that the significant main effect and significant interaction detected using spectral arc length as the motion-based metric may differ from the story told by the other metrics. Regardless, it is again apparent that the viewing modes significantly affect performance when the task is more difficult, and appear to add little utility when the task is relatively simple in nature.
Table 4.4: Guidewire Metrics - Contrast T-Test comparing SPF against EMFB, EMFB+, and EM3E. Significant results are bolded.

<table>
<thead>
<tr>
<th>Metric</th>
<th>3-D</th>
<th>150°</th>
<th>165°</th>
<th>RMDJ</th>
<th>SAL</th>
<th>S(D)</th>
<th>S(N)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Tasks</td>
<td>.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.004</td>
<td>.449</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.002</td>
</tr>
<tr>
<td>Task 1</td>
<td>.182</td>
<td>.354</td>
<td>.591</td>
<td>.120</td>
<td>.108</td>
<td>&lt;.001</td>
<td>.033</td>
<td>.096</td>
</tr>
<tr>
<td>Task 2</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.001</td>
<td>.057</td>
<td>.002</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 4.5: Guidewire Metrics - Contrast T-Test comparing SPF against EMFB and EM3E. Significant results are bolded.

<table>
<thead>
<tr>
<th>Metric</th>
<th>3-D</th>
<th>150°</th>
<th>165°</th>
<th>RMDJ</th>
<th>SAL</th>
<th>S(D)</th>
<th>S(N)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Tasks</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.003</td>
<td>.977</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.001</td>
</tr>
<tr>
<td>Task 1</td>
<td>.296</td>
<td>.570</td>
<td>.847</td>
<td>.565</td>
<td>.074</td>
<td>.001</td>
<td>.090</td>
<td>.693</td>
</tr>
<tr>
<td>Task 2</td>
<td>&lt;.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.279</td>
<td>.013</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

This effect can be further visualized in Figure 4.7 where (a) Number of Submovements, (b) Catheter Turns at 165°, (c) Root Mean Dimensionless Jerk (RMDJ), and (d) 3-D Path Length are all plotted to examine the interaction between the two variables. The plotted metrics are representative of trends observed across the complete set of metrics. Improved visualization via EM tracking technology results in smoother task execution, a reduction in catheter turns, and shorter 3-D path lengths when the
task is complex as in task 2. Gains are not statistically significant when the task is simple as in task 1.

Figure 4.7: Interaction between Task and Viewing Mode visualized using metrics (a) Number of Submovements, (b) Catheter Turns at 165°, (c) Root Mean Dimensionless Jerk (RMDJ), and (d) 3-D Path Length

One exception to this trend is the metric Spectral Arc Length (Fig 4.8). With a larger value being equivalent to smoother movement, Spectral Arc Length agrees with the other metrics in identifying Task 1 as easier on the whole, but the large
drop off for the EM3E viewing mode is peculiar. Despite most subjects expressing that Viewing Mode 4 was the most difficult non-traditional viewing mode to operate the robot with, only Spectral Arc Length appears to have detected a decrease in performance when using this mode. Because subjects were still able to more quickly complete the task using Viewing Mode 4 than would have been possible with standard fluoroscopy and because the other metrics are closely correlated to Completion Time, the independence of Spectral Arc Length may explain this anomaly.

![Diagram](image)

**Figure 4.8**: Spectral Arc Length interaction plot showing the decreased performance using the EM3E viewing mode on Task 1.

In evaluating the performance of the subjects, there were a few aspects of the trials to explore. First, there is the process of wire cannulation. This is the time spent from the beginning of the task until the wire is successfully cannulated into the vessel. After this is the leader and sheath cannulation. This involves tracking the leader and sheath over the cannulated wire until all three are successfully cannulated
into the vessel. Therefore, it is possible to inspect three aspects of the trials; the cannulation of the guidewire, the cannulation of the leader and sheath, and the entire process of cannulating all three from start to finish. Evaluation of the guidewire cannulation could be accomplished using the guidewire or leader sensor, discussed below; whereas, evaluating the cannulation of the leader/sheath is accomplished using the leader sensor and evaluation of the entire process would likely require both sensors. Because the pilot study only involved a simulated “wire cannulation,” it was best to begin with just wire cannulation.

The practical benefits of an incorporated guidewire created a dilemma for data collection because it introduced a previously unavailable wire cannulation technique. While experts used the traditional technique of guidewire extension followed by a hooking motion, less experienced subjects used the sheath and leader to aim the catheter before inserting the guidewire into the desired vessel. Thus, depending upon the subject, the tool tip – and therefore the sensor required for extracting movement data – could be either the guidewire or the leader. It is, however, likely that the guidewire sensor is sufficient for movement data regardless of subject technique simply because of the system design. The system initializes a cannulation with the guidewire slightly extended from the leader and subjects using the leader and sheath approach to cannulation tended to keep the guidewire in this position relative to the leader. Therefore in essence, the guidewire sensor location would be approximately equivalent to the leader sensor location throughout most of the trial. Only once the subject completes wire cannulation and moves on to catheter cannulation would the two no longer be in agreement. To simplify the initial evaluation, motion-based metrics for only the wire cannulation using the guidewire sensor were calculated.
4.4 Conclusions

These studies served as demonstration of two very important ideas. First, the combination of EM tracking and flexible robotics is feasible and avoids some of the limitations of current endovascular navigation, namely visualization and tool-tip manipulation. Second, EM-assisted navigation improves surgical performance by presenting the operator with a more detailed, and real-time, mapping of the tool-tip location within the body without the continuous use of fluoroscopy. Endovascular procedures are minimally invasive surgeries but they expose the patient to nephrotoxicity, due to contrast agent injection, and to radiation, at a time when radiation-induced burns and neoplasia are becoming more common. Similarly, this work affirms the previously shown idea that using robotic assistance allows better stability and maneuverability of the catheters and decreases fluoroscopy time and contrast injection [12], [65], [66]. Although these advancements offer alternatives to traditional navigation, fluoroscopy is still needed in every endovascular case, despite the risk of several diseases affecting both the patient and the surgical team. In the case of this research, fluoroscopy was almost exclusively used during EM viewing modes to verify cannulation. The greatest reductions in fluoroscopy time were seen during complex cannulation with EM-assisted navigation.
Chapter 5

Conclusions and Future Work

In this thesis, two applications of quantitative assessment of surgical skill are explored. First, the application of motion-based metrics for the classification of endovascular surgeons according to skill level is presented. While it has previously been demonstrated that motion-based metrics that aim to quantify surgical tool movement quality correlate to skill as determined by structured grading assessments, this thesis advances the field by offering a way to classify surgeons’ expertise level using these metrics combined with machine learning techniques. Artificial neural network clustering helps to illustrate the shortcomings of current caseload-based qualification criteria. Second, this thesis demonstrates the potential of motion-based metrics to quantify the value added by new tool tracking and visualization technologies integrated into a robotic endovascular surgical platform.

Chapter 2 outlined the quantitative motion metrics used throughout the thesis to quantify surgical skill. These metrics have been established and verified across multiple domains, providing confidence in their utility.

Chapter 3 presented initial attempts to utilize motion metrics in order to classify surgeons using artificial neural networks. Novices were easily identified according to conventional surgeon ratings while the neural network often struggled to classify intermediates and experts. Clustering within the data suggests that there is likely less of a distinction between these two groups than often times considered. Therefore the poor classification accuracy obtained by training learning algorithms according to
surgeon caseload is not necessarily indicative of a problem with machine learning. The Self-Organizing Map results suggest that there is some type of hierarchy of proficiency which can distinguish and delineate surgeons according to skill.

Chapter 4 explored the applicability of objective metrics of surgical skill in the evaluation of the benefits of Electromagnetic-assisted navigation. By presenting the surgeons with improved localization of the tool tip during a procedure, surgeon dexterity is greatly increased. Task completion and fluoroscopy times decreased and motion-based metrics indicated improved motion smoothness using EM-assisted navigation. EM-assisted navigation will continue to require further testing, but it offers an alternative to continuous radiation exposure while simultaneously allowing for increased visualization. The results in this thesis suggest that EM-assisted navigation is more beneficial during complex and difficult procedures.

Future work along this trajectory could include a large variety of topics. In terms of metrics themselves, endovascular-specific metrics could be developed and verified against existing measurements of performance. Along the same path, metrics could be more quickly calculated and presented in real-time to surgeons possibly encouraging greater smoothness. Future work could also involve the creation of an improved classification system. The combination of metrics and classification could be applied to vascular training programs in order to improve the efficiency of training. Similarly, objective metrics could allow for classification across all skill levels such that certification could be universally standardized. Future work may also utilize motion-based metrics in order to evaluate new equipment for surgeons in the same way that this thesis did. Whether it be visualization and data representation, or tool-tip capabilities, objective measurements are excellent tools for detecting the effects of new technology on surgical skill.
Bibliography


www.hopkinsmedicine.org/healthlibrary/test_procedures/orthopaedic/
fluoroscopy_procedure_92,p07662.


