Real-time Quantitative Assessment of Surgical Skill

Timothy M. Kowalewski

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Blake Hannaford, Chair

Thomas S. Lendvay

Howard J. Chizeck

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Real-time Quantitative Assessment of Surgical Skill

Timothy M. Kowalewski

Chair of the Supervisory Committee:
Professor Blake Hannaford
Electrical Engineering

Faculty surgeons must accurately evaluate trainee performance in an apprenticeship training model in order to establish the skill level of their trainees and minimize their iatrogenic impact on patients. This is a subjective, resource-intensive process that has historically lacked quantitative rigor, with the exception of simple, summary measures such as task time. Moreover, the rise of minimally invasive laparoscopic surgical techniques has only increased the responsibilities of surgical faculty since it is more technically challenging and takes longer to master than traditional techniques. In the past two decades, technology has augmented surgical education with a number of simulators and robotic platforms. These technologies enable a more objective and automated evaluation of surgical skill since they provide quantitative data of the surgical act. This can, in turn, decrease the time, risk, and resource cost of training for students and faculty alike.

One such platform—a laparoscopic box trainer called the Electronic Data Generation and Evaluation (EDGE) system—is herein used to collect a large, multi-institutional corpus of surgical data using the widely-adopted and heavily-validated Fundamentals of Laparoscopic Surgery (FLS) tasks. This corpus consists of subject demographics, tool motion data and corresponding video. A novel approach is proposed which seeks to find skill measures that apply universally across a variety of surgical procedures and enable skill evaluation in real-time using only tool motion data. Multiple criteria are used to rigorously define a set of "true experts" and additional skill categories. Vector quantization is used to extract rele-

vant features from the surgical tool data and quantitative methods establish which features best discriminate skill. Hidden Markov models employ these features for temporal analysis capable of real-time feedback. Two criteria are adopted to determine whether these models provide value over simple task time. The models are shown to discriminate skill levels and provide significant value over simple task time. In contrast, the FLS scoring system is shown to provide no significant value over task time.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Professor Blake Hannaford for his invaluable mentorship, support and advice over the years and Dr. Thomas S. Lendvay for his constant inspiration, his passion for advancing his field, and his admirable willingness to collaborate so closely. I am also indebted to Professor Chizeck for his advice and willingness to help me complete this work.

I would like to particularly thank Lee White who has been a phenomenal collaborator, colleague, and contributor, as well as my fellow students at the Biorobotics Laboratory, whose comradery makes the lab such a great place to work. I would like to acknowledge Simulab Corporation for their funding of this work and Lloyd Murrey and Aaron Erbeck of Simulab for their hard work related to this project and for being such great people to work with.

Most of all, I would like to thank my wife, Jennifer, for her love, patience, and unwavering support and my parents and sister who have given so much to make my long education possible.

DEDICATION

Dla Czesława Bieńkowskiego

Chapter 1

INTRODUCTION AND BACKGROUND

1.1 Thesis Goal

This investigation hinges on the following question:

What, quantitatively, is the measure of surgical skill in a given segment of time?

The goal of this study is to answer this question by identifying novel dynamic metrics. This approach moves beyond assessing skill level as a general property of a surgeon. When the results of such a measure are averaged over time, this still provides an assessment of the skill level as a general property, but a dynamic metric provides further meaningful analysis. Unlike a summary metric of skill which only computes a total score at the end of a task without regard to the ordering of data samples, a dynamic metric computes skill scores for each instant or segment of time within the task. Namely, given N seconds of time-dependent surgical tool path data, what is the quantitatively-measured skill level of the subject? Moreover, when did the subject perform poorly within the N seconds, and what exactly was the deficiency? This would enable online performance feedback to a training surgeon, for example, to accelerate learning curves for surgical skill. These dynamic metrics, ideally, should be sufficiently tractable computationally to allow near real-time evaluation.

The scope of this thesis will be constrained to technical, psychomotor surgical skill alone. The motivation for this work and its focus, formulation, and constraints are enumerated in the following sections.

1.2 Motivation

Since its introduction into general surgery, laparoscopic surgery, or minimally invasive surgery (MIS), has developed rapidly in both scope of application and complexity [34, 5]. The number and diversity of MIS procedures have mushroomed. Despite benefits of the

laparoscopic approach like reduced postoperative pain and scarring, shorter hospitalization, earlier resumption of normal activity, and significant cost savings for the health care provider, laparoscopic surgery is more demanding and requires greater concentration than open surgery [34, 80]. Operating times are longer and there is more surgical fatigue and stress because of the remote intervention associated with the laparoscopic approach [34, 15]. Furthermore, research has indicated that laparoscopic surgery is associated with a higher rate of complications than open surgery [34, 22, 79], particularly during the early part of the junior surgeon's MIS experience [34, 73]. In this context, authorities such as Satava [34, 96] and Darzi et al. [34, 16] have declared that the need for metrics in laparoscopic surgery is urgent.

Traditional surgical education has suffered from key deficiencies including a lack of objectivity in performance evaluation; a growing training gap due to tightening resource constraints, and increase in number and diversity of skills to acquire; [8] and arduous, prolonged learning curves associated with skill acquisition (5-7 years) [66]. In the past two decades, technology has augmented surgical education with a plethora of simulators which enable a higher level of objectivity for performance evaluation [97]. Simulators offer promise of semi-automated mentoring which can decrease time, risk, and resource costs associated with training. While progress is evident among surgical disciplines, deficiencies remain.

While the need for metrics is indeed urgent and objectivity is now available, little or no work has been done in systematically defining metrics which could dynamically capture the "essence" of surgical skill. That is, the skill rating at any moment or time segment within a surgical task as opposed to summary information like total time or total tool path. What specifically and quantifiably determines this dynamic "essence" of surgical skill or ineptitude remains an open question. The benefit of an answer to this questions is the possibility of automated, real-time (proximate) feedback to a surgical trainee. The ultimate motivation for such work is to accelerate or mitigate the prolonged, arduous learning curves associated with surgical skill acquisition as well as to provide accurate, granular skill measurement for the purpose of skill maintenance and re-accreditation.

1.3 Review of Literature and the State of the Art

The purpose of this section is to identify the state of the art regarding quantitative surgical skill evaluation, the conventions and vocabulary in use, and unexplored areas that would most benefit from additional work. This section is divided into seperate reviews of the surgical literature, motor learning and development literature, and engineering literature. For similar reviews that were published after this review was complete the reader is referred to [14] and especially [84].

1.3.1 The Surgical Literature

Over 70 articles within the surgical literature were reviewed. Relevant surgical literature was identified by faculty recommendation, exposure from previous work and via searches of PubMed, Science Direct, and Google Scholar using various combinations of the keywords: surgery, surgical, laparoscopic, endoscopic, MIS, minimally invasive, skill(s), metrics, assessment, skills pyramid, Miller pyramid, MISTELS, OSATS, FLS, and robotics. Relevant articles were also discovered from referenced work within the reviewed literature. Representative articles are cited in the discussion below.

An essential aspect of simulation is validation, or verifying the effectiveness of assessment or training in a simulation. There are many types of validation [23, 87]. Assessment principles typically applied in surgical simulation appear below [8, 112, 74]:

- Face validity is the extent to which the examination resembles real life situations.

- Construct validity is the extent to which a test measures the trait that it purports to measure. One implication of construct validity is the extent to which a test discriminates between various levels of expertise.

- Content validity is the extent to which the domain of interest is measured by the assessment tool. It is the appropriateness of measures tested in the simulation to the task to be trained, as determined by a task analysis (e.g., while trying to assess technical skills we may actually be testing knowledge).

- Concurrent validity is the extent to which the results of the assessment tool correlate with the gold standard for that domain. For example, if the gold standard for surgical skill evaluation is double-blind video review with an OSATS grading scale by a panel of surgical faculty, strong correlation between OSATS scores and the assessment tool is evidence of concurrent validity.

- Predictive validity measures this (concurrent) correlation as well, but as a prediction of future performance in the real environment. Consequently, this is relevant to pre-assessment, as for example the prediction of a medical student's future performance as a surgeon.

- Reliability is the ability of a test to generate similar results when repeatedly applied. When assessments are performed by more than one observer, another type of reliability test is applicable that is referred to as inter-rater reliability, which measures the extent of agreement between two or more observers.

- Transfer of Training (TOT) is considered the most important type of validity. It is the degree to which training in the simulation leads to improved performance in the real environment. When training in simulation leads to worse performance in the real environment, this is termed negative training transfer.

The reviewed literature unanimously argues the need for objective metrics in simulation and practice and recognizes the potential value of simulation-based training. Barring technology and automation, earlier methods such as the Objective Structures Assessment of Technical Skill (OSATS) employed manual, subjective evaluation of performance via expert review of video-recorded procedures [70, 86]. Objectivity was argued based on a consistent checklist and preset Likert scale evaluations with categories such as "Respect for Tissue," "Time and Motion," "Instrument Handling," etc. Such methods are equally applicable in both simulation and real surgical environments and scale well across the different tasks or modalities (e.g., robotics, laparoscopy, endoscopy, open surgery). However, they require at least one human proctor to manually evaluate each individual's work, which is expensive and

does not scale well to large numbers or concurrent trials. Such approaches also invariably suffer from the subjectivity of the evaluator's judgment.

Virtual Reality (VR) was introduced into surgical simulation in 1993 [95] and continues to be evaluated as a training tool for surgical skill with varying degrees of granularity and success [35, 103, 43, 117, 34, 13]. Proficiency-based evaluation and training arose from within this corpus of VR surgical simulation studies [33]. Proficiency methods are based on the repetition of tasks or procedures until pre-determined criteria are met within a set number of consecutive task repetitions. This approach deals well with the large amount of variability inherent within and among subjects. Application of proficiency-based methods have spread beyond VR since their inception in 2005.

The benefits of VR simulation include the ability to deploy the same environment between subjects and tasks and so offer a consistent training platform for trainees, low cost of long term use, ease in data collection, and ease of tracking the virtual environment. Drawbacks include high initial cost, steep cost increases for better realism in visual representation, internal modeling or haptic rendering, and the inability to extract similar data from real cases.

In terms of metrics, VR natively supports automation and objectivity in recording metrics, more so than in reality-based procedures or simulations. Time to task is automatically computed along with more novel tool path metrics such as path length, economy of motion, and smoothness. Recording complete tool trajectories is trivial. Such information can provide a rich source for dynamic analysis, though this source of data and its subsequent, potential dynamic analysis are basically ignored in the surgical literature. Because VR systems synthesize their environments, tracking of virtual tissue and objects and how the surgeon interacts with them is also trivial. Thus, once the expense of creating the environment is incurred, it is inexpensive to automate the accurate detection of both procedural and cognitive errors in VR. This is a major benefit of VR.

The bulk of the surgical literature in the VR simulation area has focused primarily on validation. That is, in establishing that skills acquired during simulation trials ultimately transfer to operating room (OR) performance. To quantitatively establish validity (e.g. face, content, or concurrent validity) these studies rely almost exclusively on summary metrics

like task time, path length, and economy of motion (path length divided by task time) and provide mixed results about the validity of simulators to train OR-transferable surgical skills [109].

Reality-Based (RB) simulators consist of physical objects that either mimic anatomy with varying degrees of realism or simply provide inexpensive, non-anatomical objects as a means for basic manipulation. These simulators employ real surgical tools used in the OR or slightly modified versions. Perhaps the most notable of these is the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS). The original purpose of MISTELS was to develop a series of structured tasks to objectively measure laparoscopic skills [21, 20]; these tasks were not necessarily developed to systematically accelerate or optimize the learning curves for skill acquisition. It originally consisted of seven laparoscopic tasks (peg transfers, pattern cutting, clip and divide, endolooping, mesh placement and fixation, and suturing with intracorporeal or extracorporeal knots) executed on inexpensive materials like gauze, rubber grommets, latex gloves, tubing and foam. The chief metrics used in MISTELS are task time and an error penalty. These metrics are combined into a single score based on the following formula:

$$Score = PresetConstant - TimePenalty - ErrorPenalty \tag{1.1}$$

The time penalty is simply the task completion time. Both the preset constant (cutoff time, the maximum allowable time for the task) and error penalty are unique to each of the seven tasks. MISTELS was successfully validated with varying degrees of granularity [19, 30, 55, 29, 26, 27]. Eventually, the Fundamentals of Laparoscopic Skills (FLS) committee, mandated in the late 1990s by the Society of American Gastrointestinal Endoscopic Surgery (SAGES), adopted the MISTELS program with the exception of two tasks [79]. (Clipping tubular structure and securing a mesh were found to lack utility.) Since this adoption, a number of studies reinforced the validation of the MISTELS/FLS paradigm [31, 28, 18, 107, 106, 110, 25]. Most notably, given proficiency-based training, translation of skills to the OR was established [88, 57] along with positive evidence for its utility in skill retention and maintenance [101, 11].

FLS and similar RB simulators are cheaper than VR simulators because they require less technology: they need no additional development to accomplish realism in accurate models or visual and haptic rendering. Moreover, some recent work has shown that RB simulators train technical skills more effectively than VR counterparts [10]. RB simulators also obviate additional simulation and validation steps by providing physical interaction with a real environment and objects. They provide accurate force feedback and tactile sensation induced by real physical objects that VR may only approximate at considerable expense. The realism of low cost RB simulation is limited mainly by the mechanical properties of RB tissue surrogates that may not adequately mimic human anatomy without cost increase. RB validation typically focuses on the metrics used for skill scoring and does not need to address the quality of haptic realism in simulation since the subject is already interacting with real-world objects. However, the acquisition of these skill metrics typically requires manual oversight for timing and particularly with evaluating errors for task-specific penalty scores. FLS trainers, like most RB methods, do not utilize tool motion analysis either for summary metrics like path length and economy of motion or for dynamic metrics or force information.

Robotics provides a platform in which dry-lab simulations and OR procedures can both be logged in an identical manner and yield consistent, automatically generated metrics. This would be ideal for validation studies of dry-lab or realistic VR training skills transferred to the operating theater. However, Intuitive Surgical, the company which currently deploys the vast majority of surgical robotic platforms, does not allow open access to the data streams internally collected during operation. Some work is underway for creating VR tasks intended to train or evaluate robotic skills which resemble FLS constructs, but these are not as developed or validated as the FLS program and remain an active area of research at this time [62, 102, 54]. If dynamic metrics are successfully created based on tool trajectories from VR or RB simulation, they would be naturally well-suited to extend into surgical robotics.

The stratification of surgical skills has been accomplished with varying degrees of resolution, sometimes called granularity. Perhaps the most abstract (and most heavily cited) example is Miller's pyramid [72] reproduced in Figure 1.1. Miller's approach stratifies sur-

gical skill from the perspective of an instructor or evaluator.

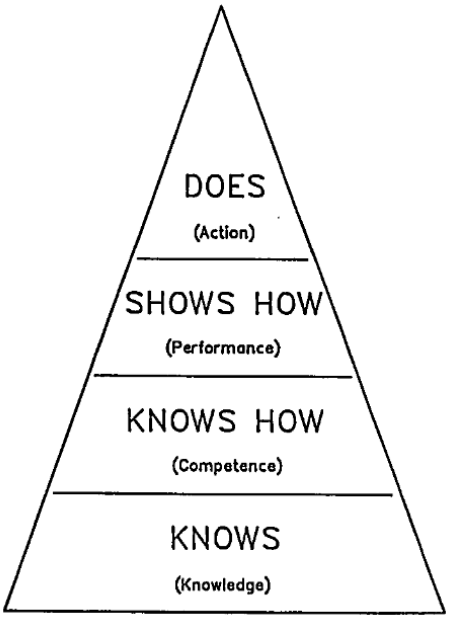


Figure 1.1: Miller's Pyramid: a "framework for clinical assessment" [72].

The implicit premise in Miller's four-layer pyramid is that certain skills are foundational; they must be developed before others can be addressed. Typically, a finer degree of granularity is used in the surgical literature in reference to skill acquisition, particularly in simulation. The literature often distinguishes between cognitive and technical skills [79]. According to Miller's pyramid, this would place cognitive skills at the bottom two levels: "knows" and "knows how." Technical skills would belong to the top two layers: "shows how" and "does," with simulation typically falling into the "shows how" layer.

Technical skills are often further stratified into visuospatial and psychomotor skills [21, 20]. Visuospatial skills consist of being able to accurately reconstruct and navigate a 3D environment based on one's depth perception of 2D video that is typically displayed along a different axis than that of the tool interaction. In his comprehensive decomposition of skill categories, Satava further distinguishes psychomotor, visuospatial, perception, and haptics skills [97]. However, he does not define or discuss the definitions of these notions. In this

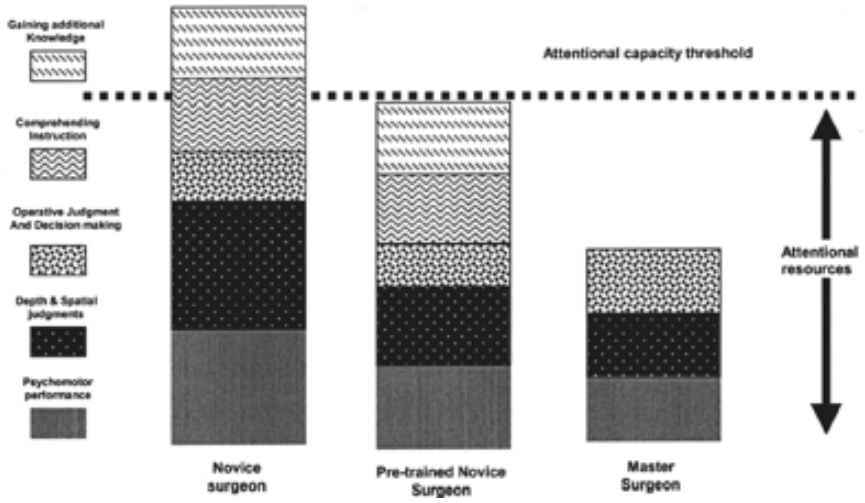


Figure 1.2: Gallagher's hypothetical attentional resource map indicates the benefits of simulation training [33].

work, it is unclear what is meant by perception. Also, presumably, haptics refers to a subject's ability to perceive haptic cues such that resolution of more subtle haptic cues implies stronger haptic abilities. However, Satava does not elucidate this topic.

Many of these finer distinctions of technical skill arise due to a change in focus. Whereas Miller's pyramid was constructed primarily from the point of view of the evaluating clinician, the simulation literature moved towards stratifying skills from the perspective of the trainee and his perception. Perhaps the most significant advance in this regard came from Gallagher's hypothetical map of attentional resources across different training levels, reproduced below in Figure 1.2 [33]. In this map, Gallagher et al. suggest that an individual surgeon has a fixed attentional capacity threshold. A novice surgeon must consciously attend to at least five items: psychomotor performance, depth and spatial judgments, operative judgment and decision making, comprehending instruction, and gaining additional knowledge. For a typical novice surgeon, the simultaneous combination of these demands is beyond their attentional capacity. As a result, their ability to learn in at least some of these categories is significantly diluted. Gallagher et al. suggest that simulation-based pre-training of novice surgeons can refine technical skills like psychomotor performance and

depth and spatial judgments such that most or all of the categories receive sufficient attention. This reasonably supposes that once trained, technical aspects will demand less attention, thus freeing attentional resources for the acquisition of other important skills or knowledge.

This hypothetical map was suggested primarily as a means to motivate pre-training of surgical trainees via simulation. It did not rigorously analyze the process of and neurophysiological elements involved in the relationships between attention, skill categories, and skill acquisition. Nonetheless, it does suggest that implementing validated, objective metrics for technical skills can be used to evaluate whether surgeon trainees are ready for higher-level instruction or learning based on their available attentional resources. In this work, Gallagher et al. do not establish these metrics, at least not in a rigorous, quantitative way.

Some more recent robotics studies from the University of Nebraska have proposed some novel, more sophisticated metrics [75, 52, 53]. Movement time intervals (e.g., time spent reaching for an object, time spent holding) and the coefficient of their variation allowed for finer granularity in temporal analysis. Another metric, curvature k, was computed by

$$k = \frac{|\dot{r} \times \ddot{r}|}{|\dot{r}|^3} \tag{1.2}$$

where r is the position of a 3-dimensional point, $\dot{r}$ is the corresponding velocity, and $\ddot{r}$ is the acceleration of that point.

Phase portraits of position vs. displacement were suggested for bimodal analysis: the extent of using both left and right hands together (see Figure 1.3). From this phase portrait, the suggested Mean Absolute Relative Phase (MARP) value was derived,

$$MARP = \sum_{1}^{N} \frac{|\Phi_{RPi}|}{N} \tag{1.3}$$

which measures the extent to which tools are out of phase and was found to be significant (in-phase registers with lower MARP, out-of-phase induces higher MARP). Moreover, electromyogram (EMG) signals were evaluated and also indicated a correlation to skill level.

The metrics used in the reviewed surgical literature were collected into Table 1.1. The

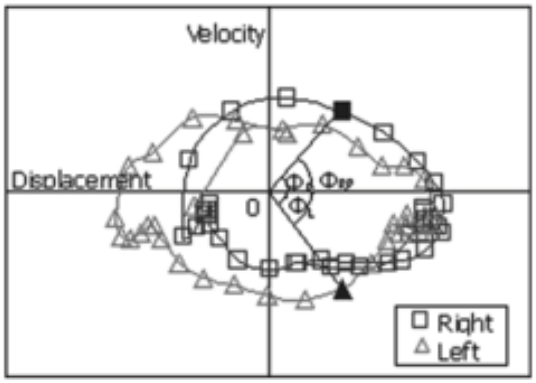


Figure 1.3: A representative phase portrait (velocity vs. displacement), from both left and right instrument tips where the phase angles (Φ) are identified [75].

table clearly establishes the ubiquity of static metrics in the literature, with task time being the most prevalent. Any of the listed platforms which compute economy of motion (EoM) and/or tool path, implicitly acquire and potentially log time-dependent tool path data. Thus, all of these approaches are amenable to automated, dynamic metrics. However, such dynamic metrics are neither explored nor applied in the surgical literature.

Table 1.1: Procedural metrics employed in the surgical literature: "time" implies the total time for completing a task; "EoM" is the economy of motion (*pathlength*/*time*, where path length may or may not include angular path), procedural errors are errors specific to a task or procedure (e.g., erroneous cuts, incorrect anatomy identification, dropped objects).

Type	**Source**	**Static/Summary Metrics**	**Dynamic Metrics**
General, RB	OSATS [74, 70]	manual subjective rating for categories such as: respect for tissue, time and motion, instrument handling, knowledge of instruments, etc.	*
VR	MIST-VR, Mentice [35, 103]	time, EoM, procedural errors	†

*Satava's metrics overview [97] distinguishes psychomotor, visuospatial, perception, and haptic as targets for metrics but does not provide quantitative attributes to measure these targets.

†Expert reviewers manually rate dynamic phenomena such as "fluid moves vs. tentative, stiff, or awkward

Table 1.1: (continued)

Type	Source	Static/Summary Metrics	Dynamic Metrics
VR	VIST, Mentice [116, 9, 12]	time, time per subtask, procedural errors, proximate feedback based on errors (force measurements are possible though not typically used in validation studies)	
VR	VR TURP Simulator [111, 83]	time, error count, mass resected, blood loss, fluid use	
VR	PicSor and Fundamental Abilities [32]	orientation error (degrees)	
VR	LapMentor, Simbionix [105, 67, 76]	time, time per subtask, movement count, path length, EoM, procedural errors, blood loss	
VR	GI Mentor, Simbionix [104, 37]	time, time per subtask, EoM, procedural errors, diagnostic errors	
VR	ES3 (Lockheed Martin, Bethesda MD) [6]	time, EoM, accuracy, navigation, errors	
VR	LapSim (Surgical Science) [108, 60]	time, EoM, path length, errors	
VR	HystSim, VirtaMed[47]	time, EoM, errors, fluid use, movie playback	
RB	MISTELS [19, 30]	time, errors	
RB	FLS [79]	time, errors	
RB	Pro-MIS (CAE, Formerly ProHaptica) [42, 10]	time, errors, EoM	
RB	ADEPT (Advanced Dundee Endoscopic Psychomotor Tester) [69, 98, 3]	time, errors, error time	

moves," "unnecessary moves," "efficient time/motion," "economy of movement," "excessive force on tissue vs. careful handling of tissue." These are very important qualitative descriptions that deal specifically with tool-tissue dynamics. However, they are manually/subjectively assessed, and even though a reviewer evaluates the dynamics throughout the entire task, the final score is cumulative, essentially rendering this a static metric.

Table 1.1: (continued)

Type	**Source**	**Static/Summary Metrics**	**Dynamic Metrics**
General, Open, RB	ICSAD (Imperial College Skill Assessment Device) [17, 112, 74]	time, path length, number of movements	
Robotic	DaVinci Surgical Robot (Intuitive Surgical, Mimic Technologies) [102, 54, 62]	time, EoM, errors, time instruments are out of view	
Robotic	DaVinci Surgical Robot (Nebraska Studies) [75, 52, 53]	time, EoM, curvature, speed, grip force; bimanual analysis via phase portraits and mean absolute relative phase (MARP), EMG analysis	‡

Many of the metrics in Table 1.1, such as procedural errors, error time, accuracy, blood loss, fluid use, etc., are specific not only to a particular task or procedure (e.g., FLS Peg Transfer or Cutting, TURP) but specifically fixed to a certain modality. For example, the amount of unnecessary tissue resected may be cheaply computed in VR, but may be difficult or impossible within RB, robotics or traditional manual MIS. However, universal dynamic metrics may potentially transcend many or all of these borders. Figure 1.4 illustrates this notion with a Venn diagram. Only potential metrics mutually held in common are enumerated.

In summary, the surgical literature is primarily concerned with the validation of skills transfer. Technical skills are primarily stratified into cognitive, psychomotor, and visuospatial/depth perception and do not consider finer granularity of skills based on neurophysiological attributes involved in human motor learning. Simulation tasks are designed with the goal of adequately addressing skills demanded by the surgical discipline, but not

‡A novel study which significantly advanced the type of metrics used; however, only summary metrics were extracted and analyzed.

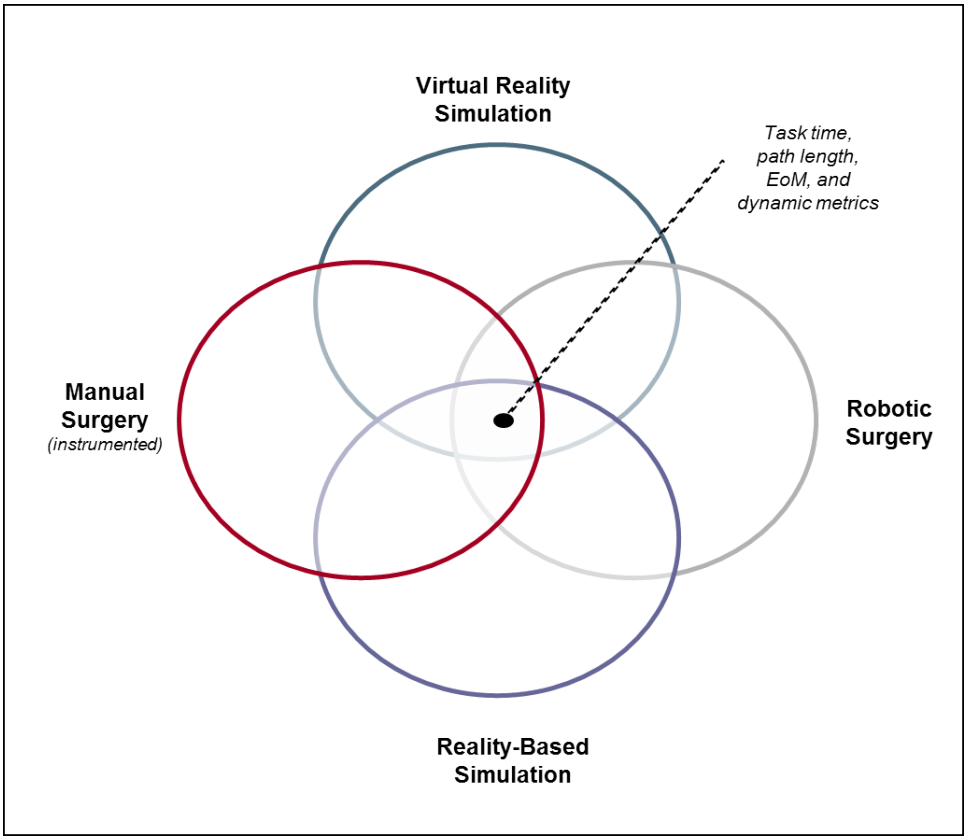


Figure 1.4: Venn diagram of the four modalities of performing or training minimally invasive surgery and the feasible technical metrics they all hold in common: time, path length, EoM and dynamic metrics.

necessarily with the goal of most efficiently training and/or evaluating those skills. Clinicians have approached the skill evaluation problem by employing statistical tools and static, summary metrics common in their field. Typical measures include total time-to-task, total errors, or economy of motion. Statistical methods focus exclusively on cumulative or summary information. They neglect the dynamic, temporal or sequential phenomena inherent in the task; yet such phenomena could, for example, intimate at what point in time, space, or procedural context a subject exhibits a degree of technical skill or ineptitude. Such information could be used to generate proximate (immediate) feedback to accelerate trainee learning curves. It may also "tag" or index certain time segments of a recorded procedure to accelerate review or be used for selective storage (e.g., only skill-deficient or unusual parts of the video recording are selected for review or storage).

There is a gap in the corpus of surgical simulation literature: dynamic metrics have neither been effectively proposed nor implemented, despite available tool path data. Unlike task-specific metrics which are confined to a particular procedure or platform, dynamic metrics may enjoy greater task independence and be consistently deployed across VR, RB, robotic, and manual MIS environments. Thus, if established, dynamic metrics would have a broad impact across the surgical discipline.

1.3.2 Motor Control, Learning, and Development Literature

Gallagher's hypothetical attentional resource map (see Figure 1.2) finds both conceptual and empirical support in the motor learning literature. ("Motor" in this field is synonymous with "muscle.") For example, the single channel theory of attention and its supporting evidence reveal that

> ... attentional demand is usually estimated indirectly by the extent to which the tasks interfere with each other. Processing sensory stimuli (or performing other processes early in the sequence) can apparently be done in parallel, with little interference from other tasks. But processes associated with response selection or with response programming and initiation interfere greatly with other activities [100, p.121].

Since early stages of surgical training deal heavily with response selection and programming, this strongly supports Gallagher's notion of attentional resource strain. Moreover, "some evidence suggests that directing one's attention to movement or environmental cues may differ according to one's skill level" [100, p.121]. Also of interest is that "other evidence, based on secondary task techniques, suggests that attention demands are highest at both the initiation and termination stages of movement" [Ibid.]. Such observations suggest strategies for developing relevant dynamic metrics. However, Schmidt and Lee conclude that "even though attention has had a long history of thinking in psychology, we are still unclear about its nature and the principles of its operation—indeed, even its definition" [Ibid.].

The literature recognizes that "learners appear to pass through various phases when acquiring skill:

1. The cognitive phase, in which emphasis is on discovering what to do, e.g., observing the target motor skill. Trainees are most responsive to verbal instruction or feedback in this stage.

2. The associative phase, in which the concern is with perfecting the movement patterns.

3. The autonomous phase, in which attentional requirements of the movement appear to be reduced or even eliminated" [100, p.429].

A key theoretical distinction in the motor learning literature is between open-loop control—sometimes synonymous with schemas or motor programs—and closed-loop motor control (also termed feedback control).

Closed-loop motor control [100, Ch.5, attr. to [2]] uses perception to consciously, continuously adjust muscle movements, when, for example, threading a needle. A subject looks at the needle then at the thread, then moves the thread towards the needle. He looks at the needle again, then looks at the thread again, then again corrects his movement. The process is repeated dozens of times until the thread is through the needle. Each stimulus-response adjustment takes at least 200 milliseconds [100, Ch.7, p.216]. Ten adjustments combine for a total of two seconds. Closed-loop motor control has two advantages:

1. It enables precise, accurate control.

2. It enables execution of novel movements (e.g., threading a needle on the deck of a rolling ship).

Closed-loop motor control has two disadvantages.

1. It is slow.

2. It requires significant attention.

Closed-loop motor control is well suited for acquiring new skills or for executing skills rarely needed. It is inappropriate and unsuccessful for fast-paced, frequently used skills. However, there is a growing body of literature which indicates that the theory of closed-loop motor control does not adequately describe the process of motor learning or at least is not solely responsible for it [100, p.411-413].

Open-Loop motor control [100, Ch.6, attr. to [99]] does not depend on feedback for its completion. For example, some muscle movements have execution times in tens of milliseconds—well below the 200ms threshold of a typical sensation-to-execution cycle. This approach bypasses the sensation (15ms), perception (45ms), and response selection (75ms) stages and directly invokes the response execution stage (15ms). This is accomplished via the execution of preprogrammed movements, called motor programs or schemas, without perceptual feedback. This phenomenon is sometimes referred to colloquially as "muscle memory" or "kinesthetic learning." Open-loop motor control has two advantages:

1. It is fast (approx. 20 ms reaction time to disturbance).

2. It requires little or no attention, especially during movement execution.

Open-loop motor control has several disadvantages:

1. If a motor program contains errors, performance is degraded.

2. After a motor program is initiated, it typically cannot be adjusted or stopped.

3. If a mistake is made, e.g., a target is missed, it may not be detected or detectable.

4. Developing open-loop control of a motor skill requires long practice, especially for adults.

5. Novel situations cannot be handled well.

In terms of new motor skill acquisition, the literature's major hypothesis is that closed-loop control is used to gradually increase skill until open-loop control can take over. Since open-loop control schemas are the result of significant training, it is possible that novel metrics may be created to identify the degree of schema development or reliance. Schmidt et al. state that "a linear [speed-accuracy] trade-off appears to occur in movement tasks that encourage a preprogrammed, open-loop control process; a logarithmic [e.g., Fitt's Law] trade-off occurs in the performance of tasks that encourage closed-loop, corrective processes"[100, p.217]. He adds, however, that "theories can claim a number of lines of experimental support, but neither is capable of explaining all the evidence on motor learning" [100, p.430]. More likely, hierarchical loops, or combinations of closed and open loops, make up the motor learning and control inherent in humans [Ibid.]. In fact, there is no shortage of theories or hypothetical models for motor learning in the literature. However, two alternatives that may provide insight for dynamic metrics are Bernstein models and a dynamic-systems perspective.

Bernstein models (developed in 1967) assert that "learning involves the process of solving the degrees of freedom problem: figuring out ways in which the independent parts of the moving body can be organized in order to achieve a task goal" [100, p.423]. According to Bernstein, motor learning occurs in three stages, each of which enjoy significant experimental support:

1. Freezing degrees of freedom. Degrees of freedom are frozen or fixed to allow as few body parts as possible to move independently. This allows for crude initial success since fewer things can go wrong.

2. Releasing and reorganizing degrees of freedom. This allows for greater independent motion and subsequently for a higher level of success. Bernstein also suggested that in this stage independent degrees of freedom may combine or "couple" to perform one functional movement.

3. Exploiting the mechanical-inertial properties of limbs. This allows more precise, faster and more efficient motion (or a combination thereof) [Ibid.].

Apart from the empirical support for this hypothesis, Bernstein's approach is compelling for the surgical task since the addition of laparoscopic tool and fulcrum effectively creates an entirely different kinematic chain for the surgical trainee to master. Thus, unlike typical adult skill acquisition like learning to play an instrument or throwing with precision, learning laparoscopic skills may be more akin to motor development in early life. Also, identifying frozen degrees of freedom may be a viable target in tool path analysis for a dynamic metric.

Dynamical systems and dynamic pattern theory provide perhaps the most popular, controversial, and unexplored perspective in motor learning. This approach posits that a system changes state over time in search of stability: stable patterns are formed, become unstable, and new patterns are formed to re-establish stability [100, p.260]. A key element of this approach emphasizes that learning new skills modulates the set of previously acquired skills [100, p.430]. While inherently complex, this approach has yielded some simple, descriptive metrics for bimanual cooperation, especially given in-phase and out-of-phase motions. It is particularly well suited for cyclic or repetitive motions. Moreover, if successful, it is promising as a means to realize dynamic metrics for surgery.

A classic component of dynamical systems analysis is the phase portrait: a plot of a data set vs. its time derivative (e.g., position vs. velocity parameterized by time). Two examples appear in Figures 1.5 and 1.6. Note how the features which discriminate the subjects are revealed by the use of a phase portrait. The differences in the time series data are not as obvious or intuitive.

An alternate metric for surgical tool path data could invoke the optimized-submovement model. This model of motion correction suggests that impulses to correct a motion may occur during the execution of a trajectory (contrary to schema theory) but in discrete

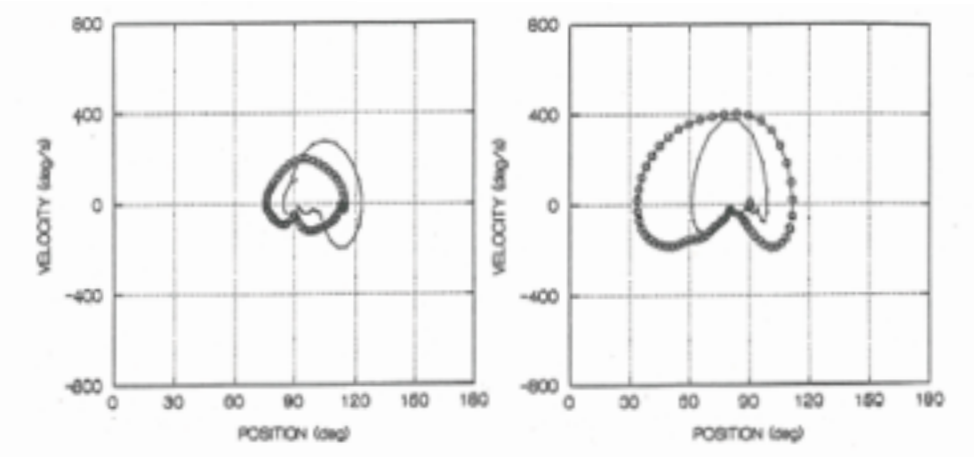


Figure 1.5: Left: the phase portrait of the thigh of an adult ($-o-$) and a newly walking infant ($--$). Right: the phase portrait for shank of an adult ($-o-$) and a newly walking infant ($--$) [113, p.86].

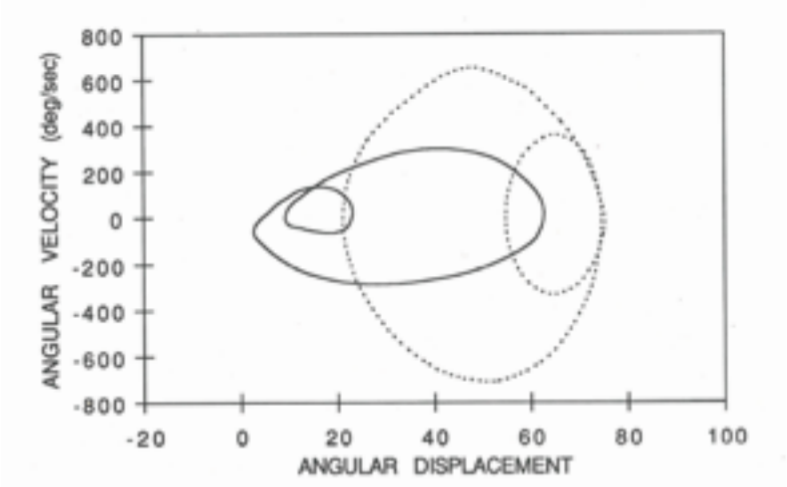


Figure 1.6: Phase portrait for the right knee. Walking: solid line (—). Running: dashed line (- - -) [113, p.215].

installments. In particular, a more recent experiment using this method for analyzing aiming tasks reports that about 80% of all movements were composed of two or three submovements (i.e., one or two corrections; 43% and 37%, respectively) [100, p.240].

1.3.3 The Engineering and Computer Science Literature

Statistical Models

Since the 1970s Hidden Markov Models (HMM's) have enjoyed considerable success in computer speech recognition and voice identification [81]. They also showed promising results when applied to robotics problems such as human task segmentation or task identification [41, 119, 51, 50, 49].

Rosen et al. successfully applied Markov modeling techniques to surgical skill/performance evaluation [92, 94, 91] in part by developing the Blue-DRAGON [89, 93, 68] operating room surgical data capture device and a subsequent, smaller desktop version known as the Red-DRAGON [38, 39]. (See Figure 1.7 below.) The Blue-DRAGON employed a novel instrumented mechanism and was used to record a large database of surgeon-tool interactions for common laparoscopic procedures executed in live pigs. This exposed surgery to modern signal processing and led to validating the Markov modeling approach for surgical skill recognition [90] by correlating Markov-based scores with scores generated by surgical faculty review of task video. (Both the Red-DRAGON and the use of HMM's for surgical skill evaluation are pending patents and were licensed by the UW Office of Technology Transfer to Simulab Corp., Seattle, WA.)

The use of HMM's for surgical performance measurement and processing has gained considerable momentum since its inception at the University of Washington Biorobotics Lab. Primary contributors reside at Johns Hopkins University [59, 63, 64] but development has spread internationally [71, 24]. The strong reception of surgical Markov modeling in academia has spurred research activity in this field and suggests introduced alternative models which could potentially outmode classical HMM's by offering better performance in surgical applications [85].

While HMM's have enjoyed decades of success in signal processing, newer, more sophis-

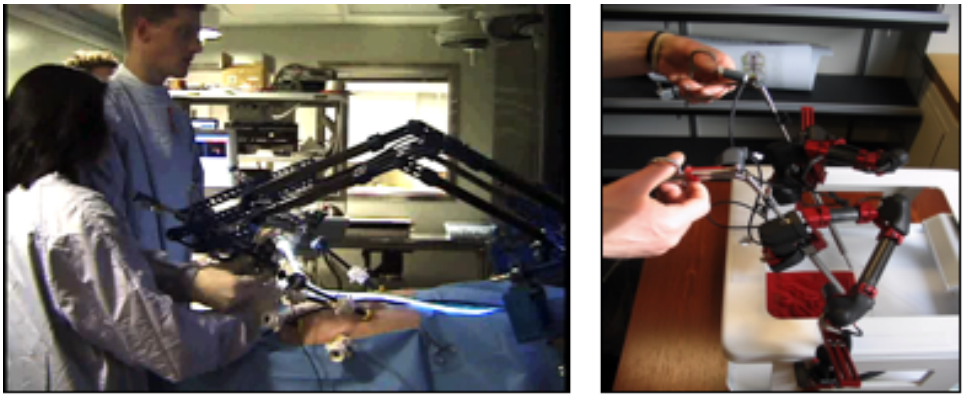


Figure 1.7: The Blue-DRAGON (left) collecting data during surgical training in live pigs and the subsequent, smaller Red-DRAGON (right) in use on an artificial tissue model.

ticated machine learning algorithms can outperform HMM's and extend their application to a larger set of problems. A notable example of applying Bayesian networks to medicine is MIT's Pathfinder, a statistical model trained by a large database of expert clinical knowledge, which has outperformed expert clinicians in diagnosing hematopathologies [45, 44]. A Dynamic Bayesian Network (DBN) is a generalization and superset of a HMM [56]. Recent work in human-computer interaction suggests that DBN's are better-suited for quantitatively processing non-deterministic human behaviors than conventional HMM's and that DBN's can accommodate a diversity of input data types [77, 48]. This suggests that DBN's could outperform HMM's in surgical performance evaluation. Moreover, due to their reliance on Bayesian statistics and Bayesian belief propagation, DBN's exhibit tremendous flexibility and are highly amenable to expert supervision in regards to structure formation.

The framework of Bayesian Filters can further generalize DBN's and have likewise shown versatility in coping with complex dynamical systems under a unifying stochastic framework [114]. However, applying DBN's or Bayesian Filtering to the surgical skill problem is challenging, particularly in the development of universal metrics: it is not intuitively clear how to define the structure of these models in a way favorable to surgical skill recognition.

Eye Movement Tracking

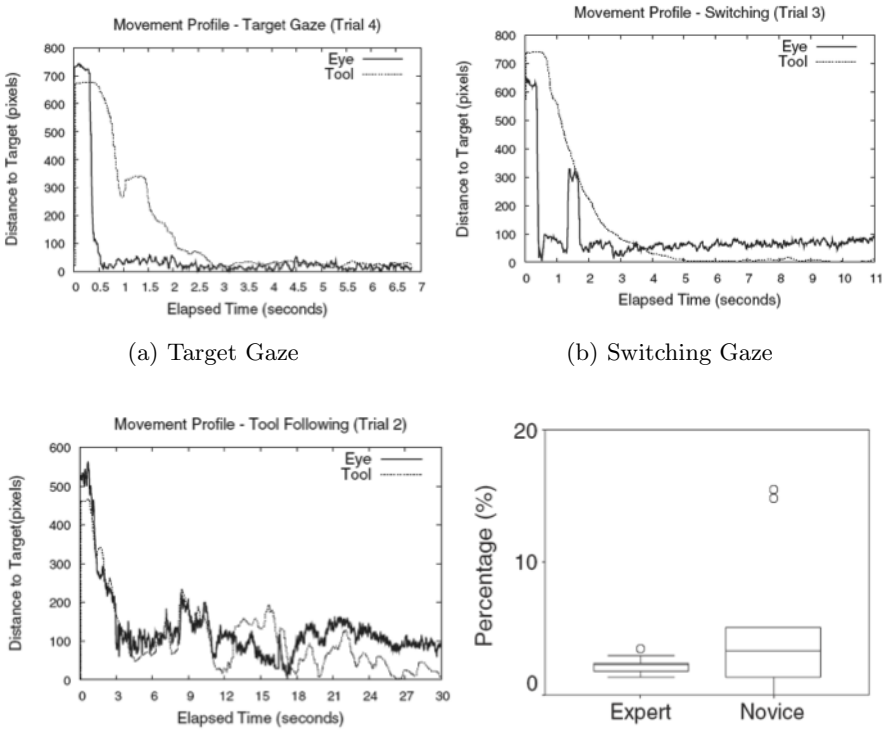
An interesting computer science study of laparoscopic skill tracked eye movement and investigated its relationship to skill level. Five novices and five experts were presented with a VR laparoscopic targeting task where a target appeared in a laparoscopic simulation and they were to touch the target in minimal time with a laparoscopic tool:

> To see if the performance differences between groups were accompanied with eye movement differences, we looked at the amount of eye gaze on the tool and then characterized their eye behaviour through eye and tool movement profiles. In terms of eye gaze behaviour, novices tended to gaze at the tool longer than experts. Several eye gaze behaviours identified in this study, including target gaze, switching, and tool following, are similar to previous findings. The target gaze behaviour was the preferred strategy for experts, and novices tended to tool follow more frequently than experts [61].

Three movement profiles for eye motion appear in Fig. 1.8. These figures indicate the distance between what pixel a subject's eyes focus on at each time instant and the pixel centered at the laparoscopic target. Fig. 1.8a shows target gaze: eyes quickly fix on a target (in 0.5 seconds) and stay fixed there for the duration of the move-tool-to-target task. The tool reaches the target within 3 seconds. Fig. 1.8b shows switching gaze: the eyes switch between the tool and target; the tool reaches the target in more time (4.5 seconds). Fig. 1.8c shows tool following: eye gaze stays fixed on the tool and almost never leaves it. Reaching the target with the tool takes considerably longer (18 seconds for first hit). Finally, Fig. 1.8d indicates that on average, novices spend a higher percentage of their time gazing at the tool.

Table 1.2: Eye movement behavior distributions for expert and novices over all trials [61].

Group	**Target Gaze**	**Switching**	**Tool Following**	**Loss**
Expert	73.3%	33.3%	8.9%	4.4%
Novice	53.3%	17.8%	26.7%	2.2%



(a) Target Gaze

(b) Switching Gaze

(c) Tool Following

(d) Percentage of Time Spent Gazing at Tool

Figure 1.8: Eye vs. tool movement profiles typifying different types of gaze patterns. Eyes gaze at the target (a), eyes switch between target and tool (b), and eyes gaze at tool (c). The total percentage of time spent gazing at the tool (d) is shown for both skill levels [61].

There are several ramifications of this study in light of the surgical and motor learning literature reviewed above. First, the different movement profiles and their associated task times corroborate the notion of Gallagher's attentional resources; the tool following profile of a novice indicates active attentional focus on the tool, while the target gaze of the expert suggests a level of autonomy in the manipulation task. Second, the difference in gaze and targeting patterns across skill levels was previously suggested in the motor learning literature reviewed (see discussion of open-loop vs. closed-loop motor control). It is experimentally reproduced here in a VR laparoscopic setting. Third, this presents strong evidence of open-loop control in the expert (and hence faster performance) vs. closed-loop control in the novice, at least in the sense of a visual feedback loop. This same study also makes the following two important observations [61]:

- Laparoscopic tool movement is unlike direct hand movement because proprioceptive feedback from hand position does not map directly to the tool tips [citing Smith et al. 2000] necessitating additional visuomotor and spatial transformations [citing MacKenzie et al. 1999].

- Tactile feedback from tool movement is minimal because of friction between the tool and the cannula (a tunnel like structure surrounding the tool at the patient entry point), and thus, the surgeon has a greater reliance on the indirect visual information [citing Cuschieri 1995; Ibbotson et al. 1999].

1.4 Novel Contributions of this Dissertation

This work makes the following novel contributions:

1. A universal metrics approach for surgical skill: instead of creating task-specific solutions, it emphasizes skill measures that generalize across different surgical procedures, tasks, and modalities (e.g. robotics, manual laparoscopy, and virtual reality simulators).

2. A real-time metrics approach for surgical skill: instead of merely summary results at

the end of the task, real-time metrics could enable immediate feedback to the trainee as well as highlighting specific segments within a procedure where skill was lacking.

3. A universal way to register surgical task spaces via tracing a calibration task block with tools and the Iterative Closes Point algorithm to register the kinematics.

4. Validation of EDGE hardware that quantifies the accuracy of toolpath and related metrics for Reality-Based surgical simulators. This reveals an advancement in the state of art since higher accuracy is reported over prior art.

5. A large, multi-institutional, high quality data set of surgical data (447 individual runs for three different tasks for 98 subjects).

6. Quantitative feature selection to establish which features of surgical toolpath data best discriminate skill.

7. Two criteria to establish whether a surgical skill metric adds value beyond a stopwatch.

8. Statistical and deterministic skill metrics which add value beyond a stopwatch; specifically, a segmented Hidden Markov Model with an evaluation method that enables real-time analysis of skill for each segment of surgical motion.

9. Evidence that the widely adopted and heavily validated FLS scoring methodology does not add significant value beyond a stopwatch.

Chapter 2

EQUIPMENT DESCRIPTION AND DATA COLLECTION

This chapter details the equipment and methods used to collect a large dataset of surgical motion and video from surgeons of all skill levels. Data was collected at three sites in the United States: the University of Washington, Seattle; the University of Minnesota, Minneapolis-St. Paul; and three locations in New Orleans. The key piece of equipment was the first generation Electronic Data Generation and Evaluation (EDGE) Platform by Simulab Corp. (Seattle, WA) shown in Fig. 2.1. This is a commercialized version of the RedDRAGON prototype previously described in Chapter 1 (see Fig. 1.7). EDGE is a dry lab reality-based box-trainer which can obtain high-accuracy motion (position and orientation) measurements along with grasping force [118]. Details for both the EDGE platform and data collection appear in following sections.

(a) First Generation EDGE (b) Second Generation EDGE

Figure 2.1: The EDGE Platform, by Simulab Corporation, is based on a spherical mechanism developed by the BioRobotics lab. It consists of a pair of interchangeable surgical tools whose motion is constrained to rotating about a remote center in the same way the motion of laparoscopic instruments is constrained by patient access ports. Both the (a) first generation and (b) second generation models are shown. Only the first generation model was used for data collection in this study.

2.1 The EDGE Platform: Calibration and Data Integrity

2.1.1 Overview

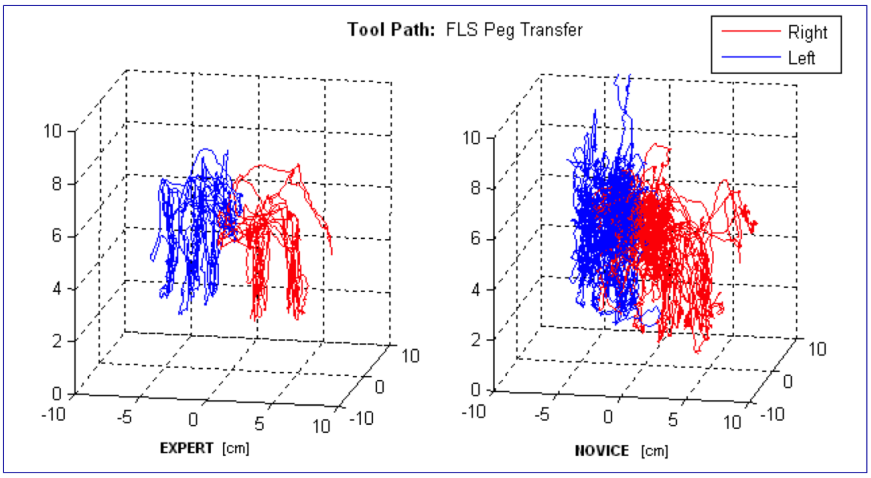


Figure 2.2: 3D Plots of example EDGE tool path data for the left (blue) and right (red) hands. A single iteration of a Peg Transfer task from a novice surgeon (right) and an expert faculty surgeon (left) is shown.

EDGE hardware consists of an instrumented mechanism which holds laparoscopic tools about a fixed pivot point. These were either Stryker Endoscopy (San Jose, CA) tools with interchangeable tool tip inserts (5mm diam, 33cm length with 250-080-282 Maryland Grasper or 250-080-267 Endo Metzenbaum Curved Scissor inserts) or Karl Storz (Tuttlingen, Germany) curved needle drivers (26173 KAL and KAR). Position sensors (potentiometers and optical encoders) measure tool position (x, y, z in cm) in tool roll (r in degrees), and grasper angle (θ in degrees). A calibrated strain-gauge measures grasping force (Fg in Newtons). EDGE includes custom software that time-stamps all sensors measurements for each hand at 30 Hz with synchronized video capture of the task. Videos are compressed on-the-fly with MPEG4 codecs and average approximately 10MB per minute of video. Timing is automatically recorded and begins and ends when tools are moved in and out of a fixed "homebase" position directly behind each task block which, in turn, is rigidly and repeatably held to a fixed coordinate system.

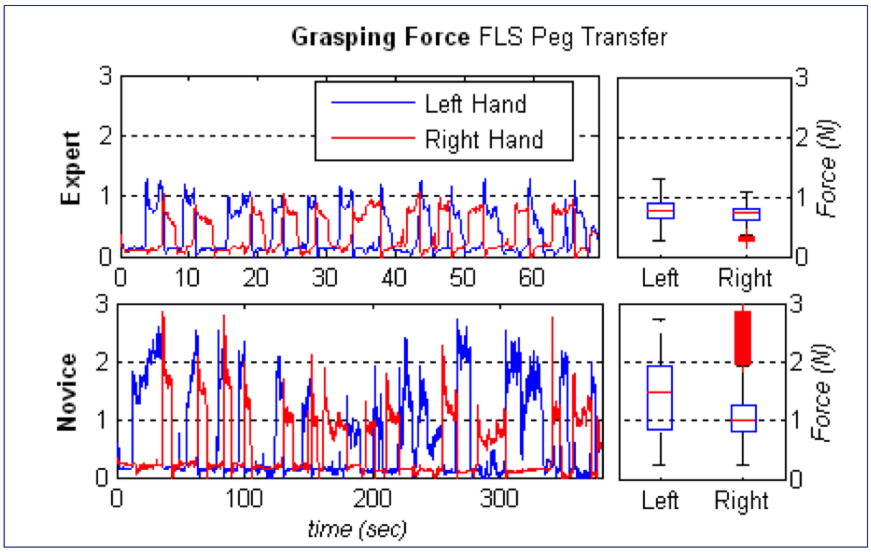


Figure 2.3: The grasping force for both left (blue) and right (red) hands vs. time for Peg Transfer task iterations of a novice (bottom) and a faculty expert (top) surgeon.

EDGE employed a laptop computer running Microsoft (Redmond, WA) Windows XP (UW sites only) or Windows 7 (all other sites) interfaced via USB 2.0 to EDGE during all recordings. EDGE's custom software stored tool data and task video files and labeled them with unique codes to identify subject, site, task, date, and time of each iteration. All tool motion and demographic data for all iterations was compiled into a single database, sorted, and analyzed by custom scripts in MATLAB software (Mathworks Inc., Natick, MA).

Plots of 3D toolpath of a novice (no prior FLS experience) and proficient subject (faculty surgeon) appear in Figure 2.2 along with corresponding scores for a single Peg Transfer iteration. Figure 2.3 shows the left and right hand grasping force plotted in time for the same iteration.

2.1.2 EDGE Sensors

EDGE measures the same seven variables for each hand. These variables and the sensors it uses to measure them are shown in Table 2.1. Some of these variables merit further discussion.

The position of the tool tip is derived from the potentiometer-measured angles of joints

Table 2.1: Details regarding all sensors used in EDGE. The same convention is used for both hands. To indicate which hand the sensor corresponds to, each label is postfixed with "L" for left and "R" for right. (e.g., QqL indicates left grasping angle.)

Target	**Label**	**Units**	**Sensor**	**Description**
Joint 1	Q_1	$^\circ$	Potentiometer	Angle between EDGE base and first link.
Joint 2	Q_2	$^\circ$	Potentiometer	Angle between first and second link.
Tool Rotation	Q_3	$^\circ$	Optical Encoder	Angle of tool rotation about its axis.
Tool Insertion	d	cm	Optical Encoder	Insertion of tool along its axis.
Grasp Angle	Q_g	$^\circ$	Potentiometer	Angle of grasper handle.
Grasp Force	F_g	N	Strain Gauge	Force applied at grasper handle.
Homebase	–	–	Switch	Activated when tool is removed from home position.

one and two (Q_1, Q_2) and the tool insertion (d) measured by an incremental optical encoder. The angles provided by the potentiometers are absolute: they supply a fixed voltage to angle mapping that does not change unless the mechanism is physically altered by loosening set screws and manually re-adjusting the potentiometer in its fixtures. The incremental encoder resets to zero each time the device is powered up and measures distance of tool insertion traveled by the tool relative to that position. While the potentiometers provide absolute position, it is difficult to calibrate their orientations in a pre-specified, repeatable way between various EDGE machines. Thus, these three position sensors are calibrated in software when the tool is at rest in its home position. This is accomplished by sliding the tool tips into "homebase" receptacles which repeatably set the two joint angles and tool insertion in a fixed, repeatable way. The presence or absence of a tool tip in the "homebase" position actuates an electro-mechanical switch. Upon exiting "homebase," software internally sets the values for each sensor position sensor to zero. This value is used for calibrating the home pose, which registers absolute position for the linear optical encoder and reference angles for tool position.

The tool rotation variable Q_3 only measured relative position since there was no home position based on tool geometry that could be used as a calibration reference and since the internal electronics reset to 0° whenever the device was powered up. Thus, the initial position (0°) was not guaranteed to be consistent between trials. This variable also had

an unbounded range: a user could continuously rotate the tool through dozens of complete (360°) revolutions during recording in either direction resulting in values far below 0° or well above 360°. While this can easily be mapped to the range [0 360] with a modulo operation, such a mapping loses information and creates additional issues for subsequent computations like time derivatives. To overcome these issues, only the tool rotation rate, not position, was considered for analysis. The rotational velocity of the tool (in $^\circ/sec$) about its axis is denoted as dQ_3. The d prefix indicates the time derivative d/dt of the variable. Two types of motion occur with this variable: slow deliberate rotation, as when a subject uses their wrist to orient the tool, and a "flicking" of the rotational knob with a finger, which can intermittently spike to very high rates. Thus, dQ_3 was passed into the function $(k\pi/2)\arctan(\pi dQ_3/k)$ where $k = 125^\circ/sec$, the maximum threshold for slow deliberate tool rotation established by recording normal wrist rotation with the tool and the "flicking" of the rotational knob on the tool with the finger. Values below this range experience only a linear scaling; values that exceed this range are compressed in a nonlinear way and map into the range $\pm\,(0.5\ 1)$ in normalized coordinates once they are established (discussed in Chapter 4). This version of tangent function-associated data derived from tool rotation rate dQ_3 is termed $dRta$.

The grasp variables consist of grasper angle Q_g and grasping force F_g. These sensors reported absolute values. However, EDGE allows users to switch out tool inserts for different tasks, such as Maryland graspers for the block transfer task or a curved shear (scissor) for the cutting task. These tool inserts had slightly different lengths, which resulted in a different angle at the grasper handles (as measured by the potentiometer). Thus, a software calibration scheme was used where the end effectors (e.g., graspers or scissors) were set in the closed position with no additional force applied. The corresponding voltage reported by the potentiometer and strain gauge was used to set the zero values of Q_g and F_g respectively.

2.1.3 Potentiometer Range Calibration and Verification

EDGE angular potentiometers require an added step for calibrate and verify their range of measurement. EDGE requires relatively long runs of wiring between the potentiometers

and its internal active electronics. Tight space constraints prohibit the addition of electromagnetic shielding and the electronics hardware only senses voltages in the range of $0-5V$. This results in a poor signal to noise ratio (SNR) in measured data, especially if the potentiometers employ a ratiometric sensing scheme when powered from a 5-Volt reference. To increase SNR, a non-ratiometric sensing scheme is implemented. The potentiometers are powered from EDGE's 12-Volt supply and are mounted so that the joint angle range reachable by the mechanism falls within the $0-5V$ range used by the electronics sensing input. This increases SNR but potentially introduces measurement error. Additionally, the potentiometers have a small dead zone where no voltage change is registered despite physical rotation, which is not accurately described by the manufacturer. Thus, a physical calibration scheme is required to correctly establish and verify the mapping of measured voltage to the real-world angle of each potentiometer in this non-ratiometric setup for each EDGE. A jig was designed and machined to empirically establish this mapping (see Fig. 2.4).

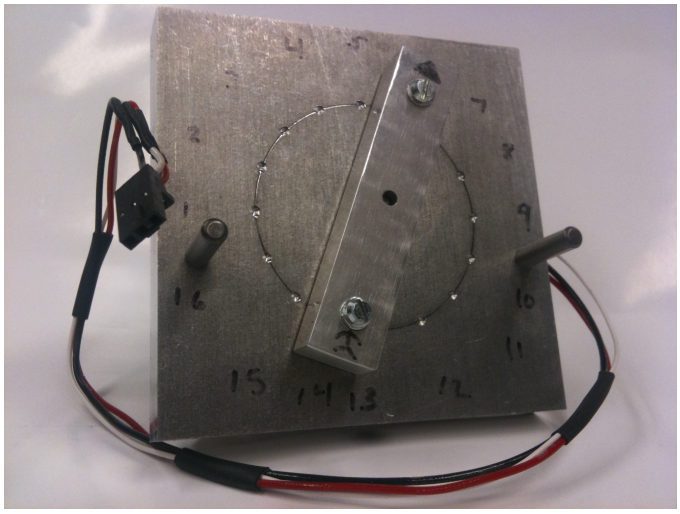


Figure 2.4: The machined potentiometer calibration jig used to verify the accuracy of all potentiometer-based angle measurements in EDGE.

Table 2.2: Angular potentiometer calibration results for the three EDGE machines used in this study. The actual 12-Volt power supply under load voltage level is shown. The resulting calibration values are in units of $inc/^\circ$, where inc refers to units of internal software increments such that $1\,inc = 0.005V$.

Platform ID	**Actual Voltage**	Q_1L	Q_2L	Q_gL	Q_1R	Q_2R	Q_gR
EDGE1	11.97 V	-0.1469	-0.1452	-0.1472	-0.1471	-0.1470	-0.1460
EDGE2	11.96 V	-0.1441	-0.1441	-0.1441	-0.1440	-0.1442	-0.1442
EDGE3	11.98 V	-0.1440	-0.1440	-0.1441	-0.1441	-0.1440	-0.1440

Kinematics Verification and Calibration

Forward kinematics refers to the nonlinear mapping from measured angles of joints one and two (Q_1, Q_2) and the tool insertion (d) to the position of the tool tip in Cartesian space (x, y, z). The forward kinematic equations which govern this mapping for EDGE's spherical mechanism are described in [39]. However, these equations make certain assumptions about the geometry of the assembled mechanism. While CAD models for EDGE were built to closely follow the assumed link parameters in [39], there are small differences. More importantly, the assumed orientation and position of the base frames are substantially different. This results in a more ergonomic system since the dexterous workspace in EDGE is centered on the task but substantially alters the coordinate system. The corrected kinematics are specified by updated Denavit-Hartenberg (DH) parameters shown in Table 2.3 and the corresponding DH homogenous transformation matrix used to construct the forward kinematic equations appears in Eqn. 2.1. The equations which map the joint angles to tip position in the base frame $\{0\}$ are given by ${}^0_4T = {}^0_1T\,{}^1_2T\,{}^2_3T\,{}^3_4T$.

$$ {}^{i-1}_{i}T = \begin{bmatrix} \cos\theta_i & -\sin\theta_i & 0 & 0 \\ \sin\theta_i\cos\alpha_{i-1} & \cos\theta_i\cos\alpha_{i-1} & -\sin\alpha_{i-1} & 0 \\ \sin\theta_i\sin\alpha_{i-1} & \cos\theta_i\sin\alpha_{i-1} & \cos\alpha_{i-1} & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.1} $$

An additional difference in EDGE concerns joint origins and calibration of absolute position.

Table 2.3: The Denavit-Hartenberg parameters extracted from EDGE designs used to compute the forward kinematics. All joint angle entries shown are for the left hand side; the right hand side angles have an opposite sign. Entries that differ by design from [39] are indicated with (‡); entries that only differ by an offset to aid in calibration are indicated with (†).

$i-1$	i	$i+1$	α_{i-1}	d_i †	θ_i †
0	1	2	0	0	$Q_1 - \hat{\theta}_1$
1	2	3	$-75°$	0	$Q_2 - \hat{\theta}_2$
2	3	4	$-59.89°$ ‡	0	$Q_3 - \hat{\theta}_3$
3	4	–	0	$-d - \hat{d}_4$	0

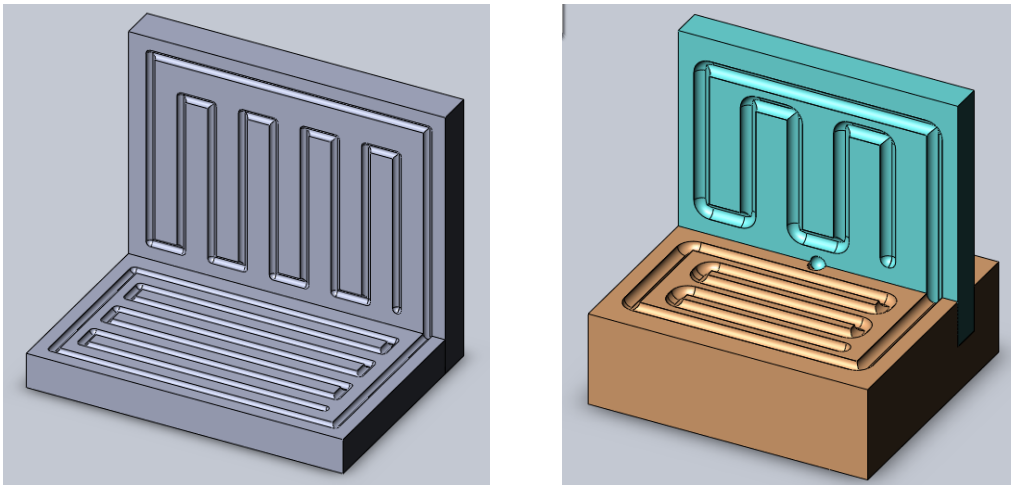
Table 2.4: Joint offsets used to establish home position when the tool tips are at rest in the "home base" receptacles. The first generation EDGE values are absolute offsets. The second generation values show the additive offset relative from the first generation values (Δ). Both were derived from CAD models or manual measurement with the assembled system.

Offset	**1st Gen.**	**2nd Gen.** (Δ)
$\hat{\theta}_1$	$30.7°$	$-29.68°$
$\hat{\theta}_2$	$81.9°$	$-26.37°$
$\hat{\theta}_3$	$2.9°$	$0.0°$
$\hat{d}_4$	23.57 cm	7.5 cm

In [39] all joints are at their origin such that they evaluate to zero (the zero-pose) when all joint axes lie in the same plane. This pose is not physically reachable by EDGE when fully assembled, so it cannot be used to calibrate the system. Instead, the tool tips are placed in the "homebase" position (home pose) and the resulting joint angles between the zero pose and home pose are either extracted from the 3D CAD models or are physically measured in the mechanism. These joint offsets appear in Table 2.4 and are embedded in the EDGE software. When the tool tip is removed from "homebase" position, the four joint angles are set to zero in software and the offsets are applied separately before recording the values. However, both are made available through the application for more intuitive online diagnostics.

Perhaps the most challenging calibration issue lies in establishing the base frame offset. While the new base frame position and orientation can be measured or approximated from the design drawings, manufacturing tolerances, and from the assembled system, additional sources of error exist that render such an approach inaccurate. For example, certain parameters cannot be physically measured in the assembled system such as the location and orientation of the sphere centers (i.e., the intersection of all joint axes of the spherical mechanism). An empirical approach was adopted to address these issues: the tools would trace a known geometry that fills the three dimensional task space and which is set in the correct coordinate system, then discrepancy between measurements and ideal geometry would provide the needed base frame offset. This approach accounts for all discrepancies introduced during manufacture, assembly or system use.

The calibration reference block was designed and manufactured as the target geometry (see Fig. 2.5). It includes a trough that the tool tip follows through a rectilinear grid pattern. This pattern was chosen as it can provide an intuitive, qualitative view of where deviations exist when recorded toolpath data is plotted as 2D projections onto the xy, yz, and xz planes. A more rigorous technique involves extracting a point cloud of the reference rectilinear tool path and computing the error from corresponding points of recorded tool path travel. The reference point cloud was extracted by sampling the rectilinear tool path at 10 points/cm. The coordinate system used by the reference point cloud is the intended base frame coordinate system that defines the EDGE task space. Fig. 2.6 shows both the

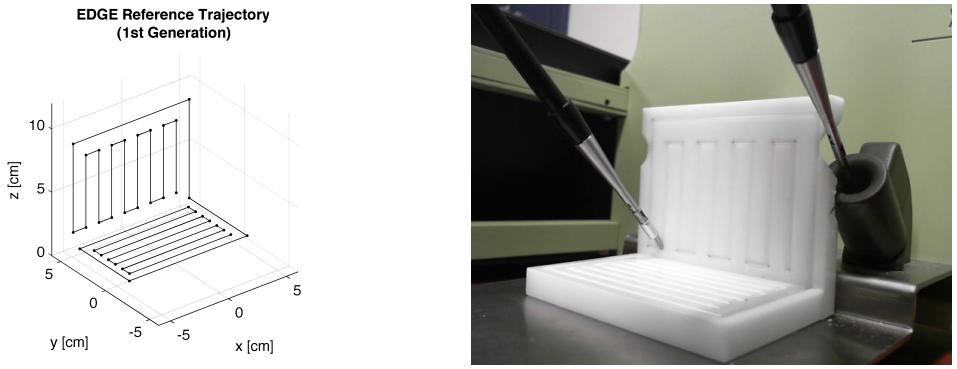


(a) First Generation (b) Second Generation

Figure 2.5: 3D CAD models of the calibration blocks used in both generations of the EDGE machine. The first generation block (a) was machined from Delrin, which provided low friction but also low dimensional stability over time. The second generation version (b) was machined from aluminum, which is more dimensionally stable and allows an electrical contact sensor to indicate when the tool tip is in contact with the block during a recording.

reference tool path and the calibration reference block tracing procedure.

The Iterative Closest Point (ICP) algorithm [120] was adopted to quantitatively compute the correct base frame offset between the desired reference geometry in the correct coordinate system and the actual data collected. This correction provides the translation and rotation which minimize the error between collected data and reference geometry without needing to parametrically define which collected data point belongs to what segment of the reference geometry. An outline of the algorithm as it is used in this work is provided by Algorithm 2.1.



(a) (b)

Figure 2.6: EDGE reference trajectory data capture. The reference trajectory is shown in (a) is the intended coordinate system of the EDGE task space. Control points used to define the piece-wise rectilinear path are highlighted at the corners. Data capture (b) from tracing the reference block with tool tips uses a small aluminum cuff to stiffen the tool tip jaws. One tool remains in "homebase" while the other traces the reference path.

input : Point clouds of reference geometry and captured data Ref and D, initial estimate of transformation matrix T_o, threshold to stop algorithm th.

output: Transformation matrix T providing closest fit between Ref and D.

while *not at threshold th* **do**

Associate points with nearest neighbor criteria;

Estimate T using mean squared error cost function;

Transform points D using estimated parameters;

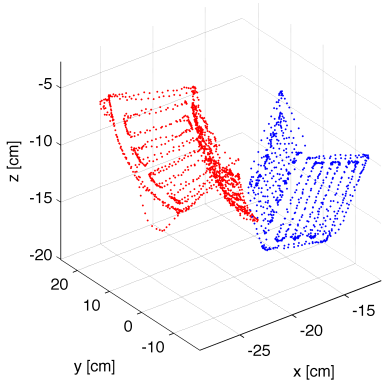
end

Algorithm 2.1: The Iterative Closest Point Algorithm.

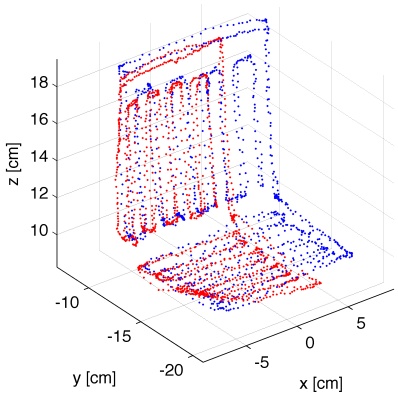
The nearest neighbor criteria it uses is, for example, a k-D tree or Delaney tessellation. The output transformation matrix T contains the position translation vector $P = [P_x, P_y, P_x]^T$ and 3×3 rotation matrix R that minimize the error between the reference trajectory Ref and captured data D. It assumes the form:

$$T = \begin{bmatrix} R & P \\ [0\ 0\ 0] & 1 \end{bmatrix} \tag{2.2}$$

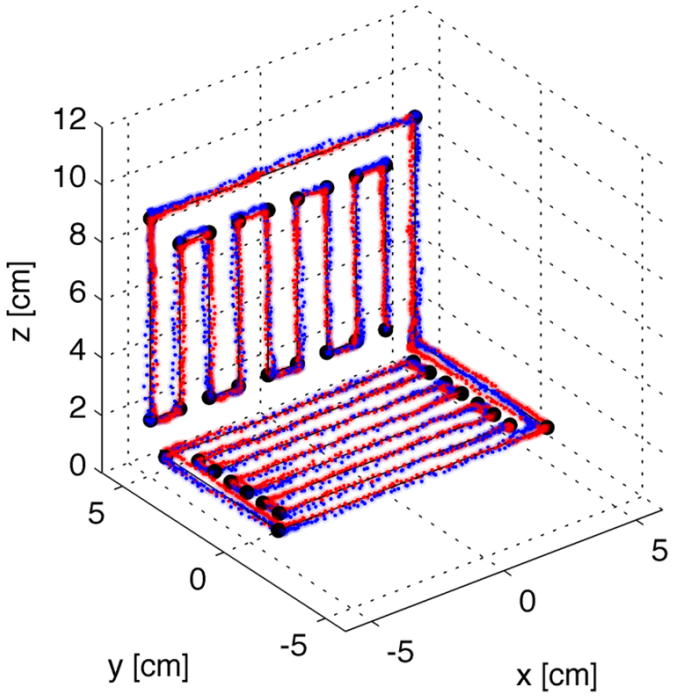
A comparison of the original kinematics implemented in the Red DRAGON, corrected kinematics based on analytically-derived updates to the DH parameters, and the corrected kinematics based on the ICP algorithm appear in Fig. 2.7. The original implementation was flawed (Fig. 2.7a): there was a sign error in the equations, the orientations of the base frame were completely distorted, and the origin position was off by as much as 20 cm. In the corrected DH parameters version (Fig. 2.7b) a significant deviation exists in both left and right tools due to mechanical sagging of the EDGE machines after initial assembly. This was due to a design flaw that allowed the settling of plastic structural members. Such errors are difficult to model quantitatively or eliminate during assembly, illustrating the need for an empirical calibration approach such as the ICP algorithm can provide. This approach clearly provides the best kinematic accuracy in the correct desired coordinate system (Fig. 2.7c).



(a) Original Red DRAGON Kinematics



(b) Analytically Adjusted DH Parameters



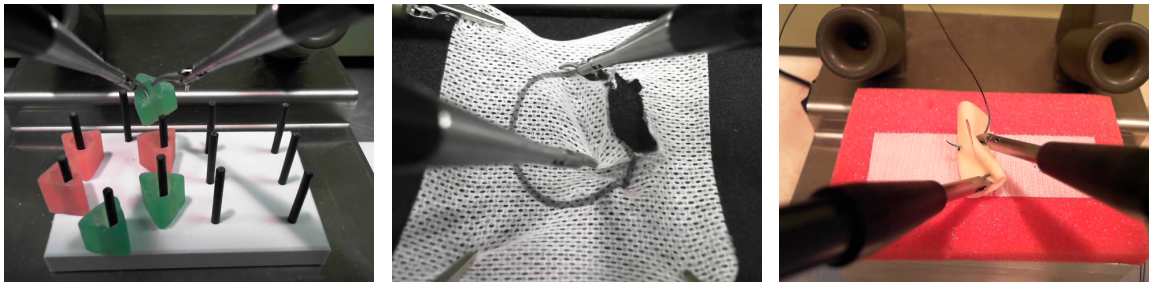
(c) ICP Algorithm Adjustments

Figure 2.7: Forward kinematics verification via calibration block traces left (blue) and right (red) hand tools from EDGE using: (a) the original Red DRAGON kinematic equations—both hands had significant translation and rotation offsets; (b) from kinematics that employed theoretically-derived corrections based on designed deviations from original Red-DRAGON parameters; and (c) from final adjustment via ICP algorithm. The ICP algorithm method shows the best fit to the reference geometry (black), not only between the two traces, but also to the absolute reference frame.

2.2 Multi-Institutional Data Collection

This study employed the EDGE platform to collect tool motion and task video data from faculty and training surgeons at multiple surgical centers in the United States. The subjects were asked to perform the Fundamentals of Laparoscopic skills (FLS) tasks which, were attached to the EDGE platform in a rigid, repeatable way to maintain the same coordinate system. This section details the methods used for the data collection and provides an overview of the number, type, and location of all task iterations successfully recorded with EDGE.

2.2.1 Surgical Task Description and FLS Scoring



(a) Peg Transfer (b) Circle Cutting (c) Suturing

Figure 2.8: EDGE Screenshots of videos recorded during performance of the three FLS tasks: Peg Transfer, Cutting, and Suturing.

Three of the five FLS tasks were used in the study: Peg Transfer, Cutting, and Intracorporeal Suturing. Descriptions of each task appear below along with representative screenshots in Fig. 2.8. An iteration was defined as one complete execution of a single task. Subjects were asked to complete three iterations of the Peg Transfer task, two iterations of the Cutting task, and two iterations of the Suturing task, in that order. Subjects were also invited to perform additional iterations of any task if they were willing to do so. Each subject was introduced to each task via printed instructions. For each iteration, time (t) started upon removing either tool tip from a fixed "home-base" position and stopped when both tools are returned to that position. The published FLS scoring methodology [21, 28]

was adopted to calculate FLS scores for each task (denoted FLS_{task}). Explicit computation is shown in Table 2.5 and each task's error variables E_{Err} are described below.

Table 2.5: Equations used to compute FLS scores [21, 28].

FLS Task	**FLS Score**
Peg Transfer	$FLS_{peg} = (300 - t - 17E_{dr})/237$
Cutting	$FLS_{cut} = (300 - t - 2E_a)/280$
Suturing	$FLS_{sut} = (600 - t - E_{pd} - E_g - E_q)/520$

Peg Transfer (PegTx)

The Peg Transfer task, also called block transfer, employed two curved Maryland graspers. Instructions were to transfer six blocks in minimum time and with minimal errors, from one side to another and back again without regard for order or color of blocks, to transfer each block mid-air between hands, and to avoid dropping blocks. EDGE automatically computed task time t. Each video was later manually reviewed to count the total number of non-recovered drops considered to be errors (E_{dr}).

Circle Cutting

The Circle Cutting task (Cutting), also called pattern cutting, employed a curved Maryland grasper in the non-dominant hand and a curved sheer in the dominant hand for the duration of each task. Instructions were to cut gauze along a marked circular pattern ($diam = 4cm$) in minimum time and with minimal error and to begin by either making a puncture anywhere on the circle or cutting in from the gauze edge. Task time t was automatically computed. The accumulated area (mm^2) cut beyond the marked circle boundary composed the error count E_a. To minimize subjective grader error and exploit automation, cut circles were flattened and electronically scanned, and cutting error was automatically computed via ImageJ (NIH), a public domain image processing suite [82]. The wand tool automatically outlined and measured out-of-bound areas through an edge-finding algorithm that measures

pixel values. Scans included a printed reference line in order to scale the image from pixels to mm.

Suturing

The intracorporeal suturing task (Suturing), also called knot tying with intracorporeal suture, employed two curved small needle drivers. Subjects had a choice of a ratcheting or non-ratcheting mechanism at the beginning of each task. Instructions were to complete the task in minimum time and with minimum errors. The task was to puncture a Penrose drain at marked entry and exit dots with a 2-0 V-20 tapered half circle needle and 12.5 cm of suture and tie a surgeon's knot with two initial throws and one final throw. Errors included distance away from puncture dots in mm (E_{pd}), gap of sutured slit in mm (E_g), and knot quality (E_q) where 0 indicated a secure knot, 10 a slipping knot, and 20 a knot that came apart. Task time t was automatically computed, but errors were manually determined by two FLS-certified graders.

2.2.2 Subject Pool Description

Subject enrollment was approved and registered under Western IRB 19125-A/B. This multicenter effort included surgeons of various skill levels from the University of Washington Medical Center, the University of Minnesota Medical Center, and three sites in the city of New Orleans enumerated below. The subject pool spanned General Surgery, Urology, and Gynecology specialties. It consisted of active surgical faculty, surgical fellows, residents, and experienced practicing surgeons. Medical students pursuing surgical practice or FLS-experienced technicians were also enrolled in the study.

2.2.3 Data Collection Sites

Three EDGE platforms, one dedicated to each task, were deployed at each site to maximize subject throughput and allow up to three simultaneous subjects. An approved study administrator set up the equipment at each site on a daily basis and subjects were invited to voluntarily participate in the study whenever their schedules allowed. Subjects were allowed

to complete the study over multiple sessions.

Site: University of Washington

Data were collected over a three week period at the University of Washington Institute for Simulation and Interprofessional Studies (ISIS) surgical simulation training center or at the Center for Vidoe-Endocscopic Surgery (CVES), a porcine lab laparoscopic training center.

Site: University of Minnesota

Data were collected over a period of two weeks mostly at a surgeons' OR lounge during regular operating hours. Some subjects did their tasks at SimPORTAL, the adjacent surgical simulation training center of the University of Minnesota.

Site: New Orleans, Louisiana

Data were collected over a ten day period. Sites included the Louisiana State University Health Science Center, University Hospital New Orleans, and Interim LSU Public Hospital.

2.2.4 Questionnaire

Subject demographics were recorded after consent was obtained. The de-identified questionnaire included subject's gender, age, handedness, training level, surgical specialty, laparoscopic experience, approximate number of relevant procedures done (where a subject completed more than half of the case) time since last laparoscopic procedure, total number of FLS tasks done in the past, and FLS certification status.

A post-task questionnaire invited feedback regarding the acceptability of EDGE based on categorical Likert scales and written general comments. A subject's oral comments during the study were also noted in the general comments section.

2.2.5 Summary Overview of Collected Data

Not all subjects completed all requested iterations in the study. Some subjects voluntarily completed additional iterations. Incomplete iterations or iterations with corrupted data

such as missing video which prevented post-task scoring were excluded from analysis. This reduced the total number of attempted iterations from 583 to 447. An overview of the collected data appears in Table 2.6. This provides the quantity of successfully recorded iterations according to task and site categories.

Table 2.6: Overview of all collected EDGE data.

	UW	**UMN**	**NOLA**	**TOTAL**
"Expert" Subjects	6	8	3	17
Total Subjects	32	35	31	98
Peg Transfer Iterations	78	88	27	193
Cutting Iterations	61	53	51	165
Suturing Iterations	0	59	30	89
Total Iterations	139	200	108	447
Total Time (hours)				22.7

2.3 Data Pre-Processing, Filtering, and Derivatives

The EDGE hardware and software provide raw tool motion data in the form of time-stamped joint angles, grasp variables, and kinematically-derived tool tip position. In order to accurately extract motion features such as tool tip speed, acceleration, or cumulative tool tip path length, additional data processing steps are required. This pre-processing step is intended to verify the accuracy of extracting time derivatives and tool path length, and to determine optimal filtering parameters.

2.3.1 Signal Characterization and Filtering from Collected Data

Accurate time derivatives of sensor data demand maximal noise suppression. The frequency domain representation of the time derivative operation shows that signal gain increases without bound with increasing frequency. Since human hand motion is a low-frequency, bounded phenomena, this indicates that additive white noise can drastically degrade the results of time differentiation. High frequency noise components can be greatly amplified above the true sensor signal. This problem is exacerbated by repeated differentiation, e.g.

to extract acceleration.

EDGE hardware incorporates a variety of filtering stages in its signal acquisition electronics. However, due to long wiring runs and an unavoidable lack of electromagnetic shielding of some of the components, noise cannot be adequately eliminated by the electronics hardware alone. Moreover, the filter cutoff frequencies and oversampling designed in the electronics are designed primarily to eliminate aliasing in analog to digital conversion and to have a passband of 15Hz. This passband was intended as a conservative threshold. Since surgical tool motion tasks, especially those used in this study, actually exhibit lower frequency spectra, this admits additional noise into the recorded data. Even if this additional noise level is small, it has the potential to degrade time derivative calculations. Thus, additional filtering can help.

To establish the optimal cutoff frequency for implementing additional filtering in software, sensor signals from all recorded data were characterized with power spectral analysis. Additionally, to isolate only noise generated by the system, a "stationary task" was recorded. This consists of placing the tool tips in a stationary position chosen to induce the largest error: when the tools are at their maximal insertion in surgical tasks. This induces the largest change in Cartesian position for a small change in joint position (the magnitude of the determinant of the manipulator Jacobian is highest). In other words, such a position is most sensitive to noise-induced error.

Spectra were estimated using Welch's averaged, modified peridogram spectral estimation method with a Hamming window of 1024 samples at 50 sample increments. For each of the three FLS tasks, all recorded iterations were concatenated before computation and the resulting spectra were all normalized to their individual peak level for each task. The stationary task consisted of a single iteration lasting 30 seconds.

Figure 2.9 shows the plots of these power spectra for the three FLS tasks and the stationary task and highlights the highest and lowest bandwidth sensors: the grasp variables of grasp angle Qg and force Fg for both left and right hands and the tool rotation variable Q_3. This variable was denoted Q_3-rel because the tool rotation sensor reported only relative values since it lacked an absolute reference (see Section 2.1.2). For all FLS tasks, the position-related joint variables Q_1, Q_2, and d showed very similar power spectra and

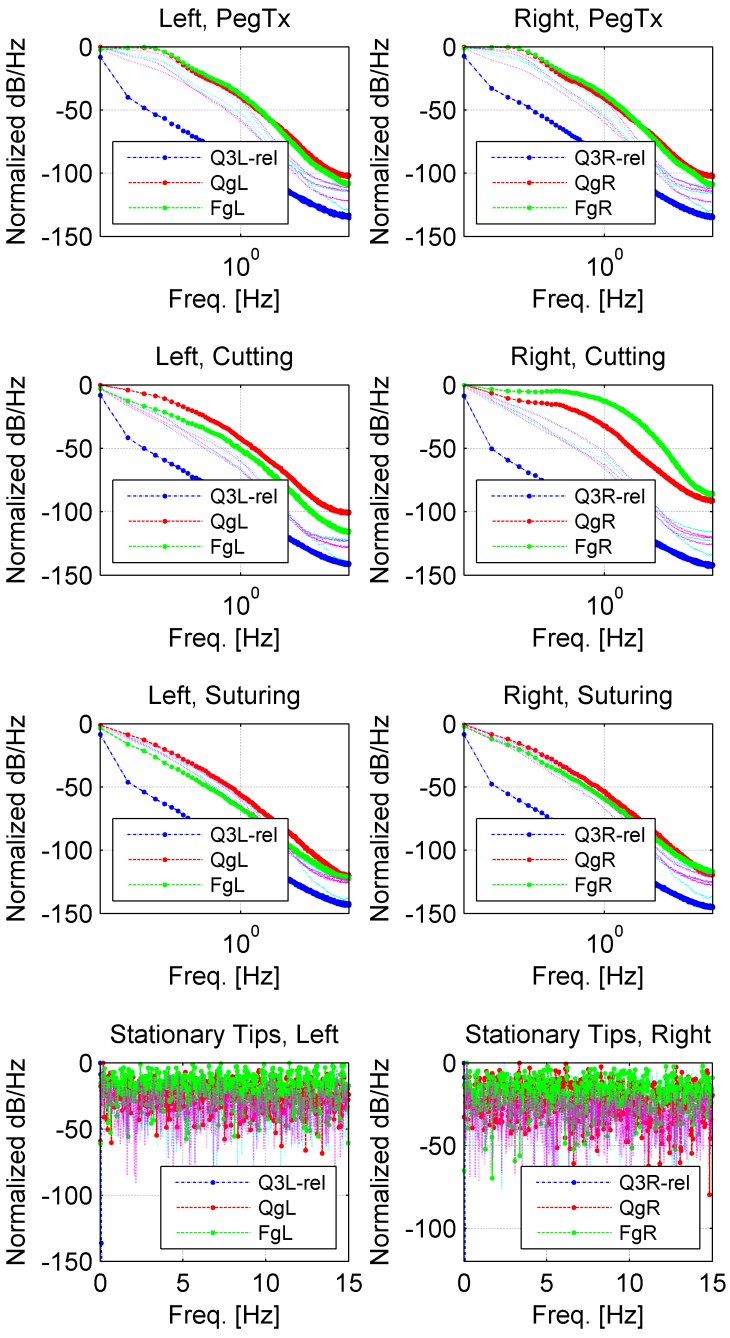


Figure 2.9: The normalized power spectral densities for left and right hands of the Peg Transfer (top), Cutting (upper middle), and Suturing Tasks (lower middle) as well as 30 seconds of the tool tips recorded at a stationary position (bottom). The stationary tool position was chosen such that sensor noise would induce the largest error. The legends indicate the lowest and highest bandwidth variables: tool rotation Q_{3-rel} and grasp variables F_g and Q_g.

typically fell between the maximum and minimum spectra of the other sensors. All spectra were consistent between the different FLS tasks with the exception of the right hand grasp variables in the cutting task. These grasp variables, Q_gR and F_gR, exhibited substantially increased bandwidth. This stems from the "snipping" behavior exhibited in this task: subjects would typically make frequent, repeated cuts as if cutting paper and the scissor tool tip (curved shears) exhibited substantial static friction, resulting in jerky motion that registered higher frequency content. Unlike the FLS task spectra, the stationary task exhibits a nearly constant, flat spectrum. This confirms that the sensor noise is white noise and legitimizes the need to eliminate the higher frequencies clearly attributed to noise and not the relevant signals in the lower part of the spectrum of each sensor.

In order to quantitatively determine the suitable, data-derived cutoff frequency for each sensor, a threshold of 99.8% of the cumulative bandwidth was adopted. This was intended to sufficiently capture the relevant data but eliminate irrelevant noise. The cumulative power spectra were computed using trapezoidal numerical integration and appear in Fig. 2.10 along with the 99.8% thresholds. Thresholds for the stationary task were not computed. The stationary tips' recorded values verify the virtually non-existent noise of the optical encoder sensors d and Q_3. Additionally, the nearly linear cumulative spectra of the other sensors suggests that the primary source of these sensors (potentiometers and strain gauges) is indeed thermal white noise.

The stationary task also provides an important threshold for computing tool path length. Since path length is an integration operation, additive noise will accumulate without bound and tool path length will accrue with time even if the tools are physically stationary. To eliminate this source of error, the integration operation can ignore adjacent samples that fall below some threshold of motion: the movement threshold M_{th}. If the threshold is larger than the typical noise-induced level it will eliminate relevant data in the computation, i.e., very slow tool motion will not count towards total tool path. Thus a minimum threshold is desirable. The stationary task provides the worst-case noise-induced level of signal amplitude change in the position sensors. Figure 2.11 plots the left hand Cartesian tip position variables vs. time to demonstrate the noise-induced phantom motion in the stationary task. It shows both the unfiltered signals and the signals filtered with a tenth-order Butter-

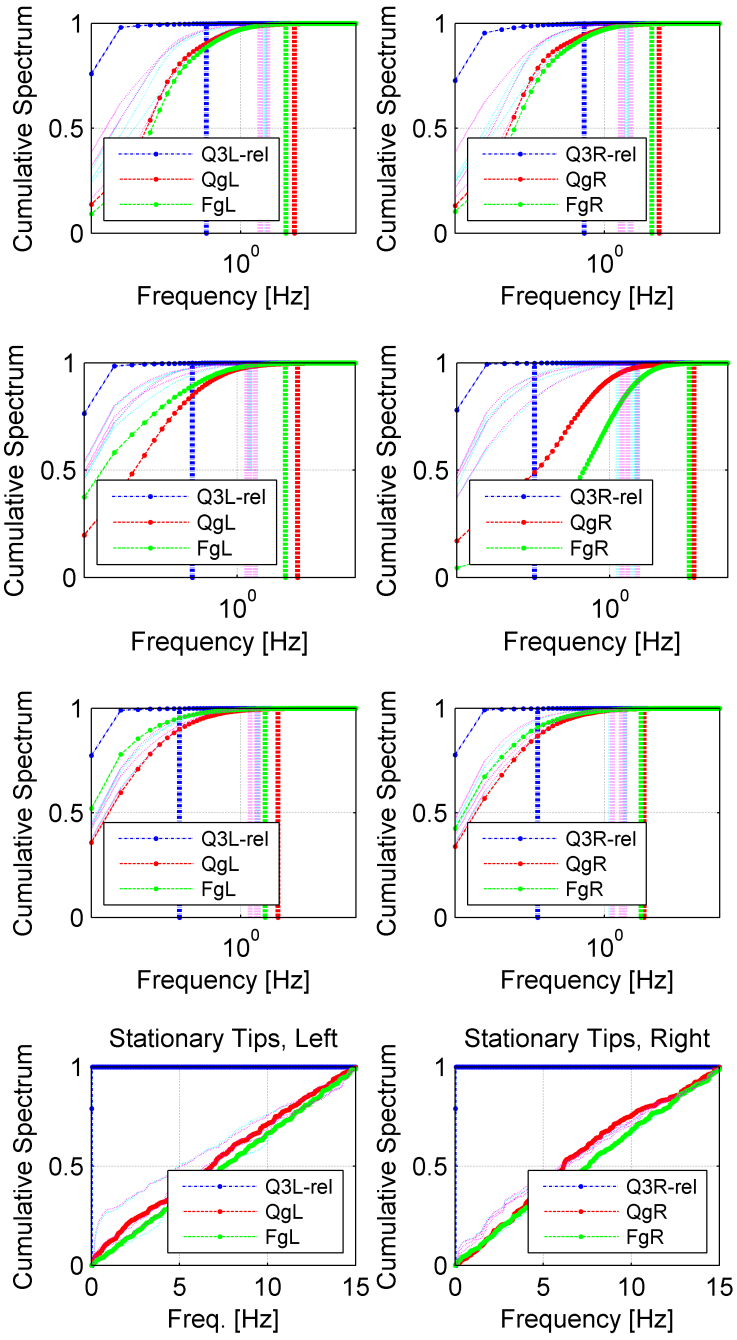
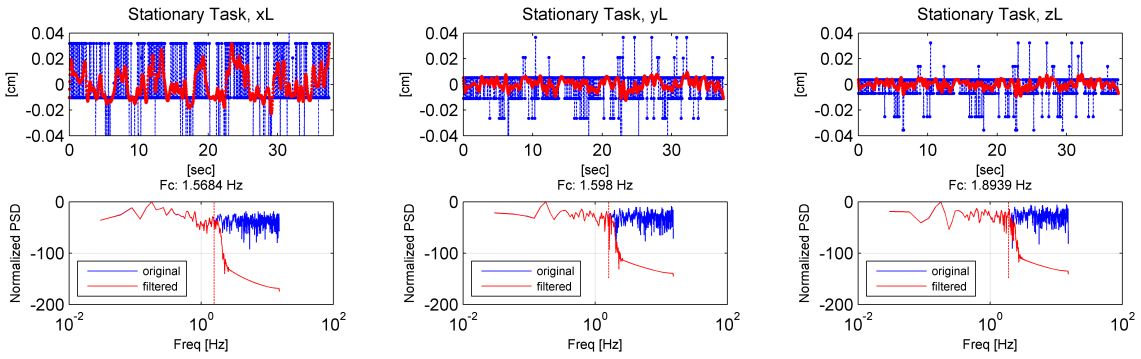


Figure 2.10: The cumulative power spectra for left hand (left side) and right hand (right side) of Peg Transfer (top), Cutting (upper middle), Suturing (lower middle), and Stationary tasks (bottom) were estimated via trapezoidal numerical integration. The 99.8% cumulative bandwidth thresholds are indicated with vertical lines colored to match their source sensors. Note that the left and right hand bandwidths are not consistent, especially for the cutting task. Thresholds for the stationary task were not computed.



(a) (b) (c)

Figure 2.11: The tip position in time (top) and frequency (bottom) for the worst-case Cartesian tip position values during the Stationary Task, showing both filtered (red) and unfiltered data (blue).

worth lowpass filter using a cutoff frequency established by the 99.8% cumulative spectrum thresholds. The Butterworth filter was employed to ensure a flat pass band. A second, time-reversed application with the same tenth order filter ensured no phase delay artifacts. In Fig. 2.11, the filtered signal, though better, still shows a range of motion even though the tools were physically motionless. This maximum per-sample range of motion (difference) for each variable sets minimal movement threshold M_{th} for accurately computing tool tip path length and subsequent derivatives.

The greater of the left or right hand cutoff frequencies for each sensor found in Fig 2.10 was adopted as the cutoff frequency for that sensor. This is because the majority of subjects in the database were right-handed; thus a left-handed subject's higher bandwidth motions may be incorrectly filtered out. Since tool insertion d and tool rotation $Q3$ sensors are optical encoders, they added negligible noise and were not suitable for the bandwidth-based noise estimation. Thus a cutoff frequency of 5Hz was used in these cases. Similarly, the worst-case value (largest magnitude) of either the left or right hand movement threshold was used to compute the movement threshold M_{th}, and any subsequently extracted features like tool path length used this threshold. The same tenth-order lowpass Butterworth double pass (forward and reversed time) scheme described above was applied to all sensor data prior

to any subsequent skill evaluation analysis in this work. Table 2.7 provides the computed numerical values for all filter cutoff frequencies and movement thresholds.

Table 2.7: Filter cutoff frequencies Fc (in Hertz, unless otherwise specified), and the Movement Threshold M_{th} used to compute tool path length and its derivatives.

	Peg Transfer $F_c\ (Hz)$	**Cutting** $F_c\ (Hz)$	**Suturing** $F_c\ (Hz)$	Movement Threshold M_{th}
Q_1	1.78	1.86	1.63	0.0193 (deg)
Q_2	1.60	1.24	1.24	0.0174 (deg)
d	5.00	5.00	5.00	0.0000 (cm)
Q_3	5.00	5.00	5.00	0.0564 (deg)
Q_q	3.61	6.92	2.54	0.0733 (deg)
F_g	3.05	6.21	2.37	0.0363 (N)
x	1.57	1.51	1.24	0.00699 (cm)
y	1.60	1.33	1.51	0.00565 (cm)
z	1.89	1.89	1.60	0.00292 (cm)

To illustrate the need for both filtering and the movement threshold M_{th} for accurately computing tool path length, Figure 2.12 compares all four possible conditions in computing the tool path for the stationary task. The raw data without any threshold results in greatest error. Within 30 seconds, over 15 cm of tool path are erroneously recorded despite the fact that the tools are motionless. Filtering the data substantially reduces this error to approximately 1 cm after 30 seconds. Without filtering, when only a static movement threshold of 0.05 is applied to path length directly, even if it is five times larger than the data-derived values shown in Table 2.7 (i.e., $0.05 > 0.0095 = \| [M_{th,x}, M_{th,y}, M_{th,z}] \|$), it performs quite poorly. After 30 seconds, approximately 5 cm of error accrue. However, when the data-derived motion thresholds M_{th} are applied for each sensor along with filtering, path length error is zero for the entire stationary task.

2.3.2 Time Derivative Extraction and Validation

EDGE does not provide any rate sensors. Thus, to obtain rates like velocity or acceleration, these values must be extracted from position. The evident presence of unwanted white noise in the recorded signals can seriously degrade the accuracy of such derived rates, despite the

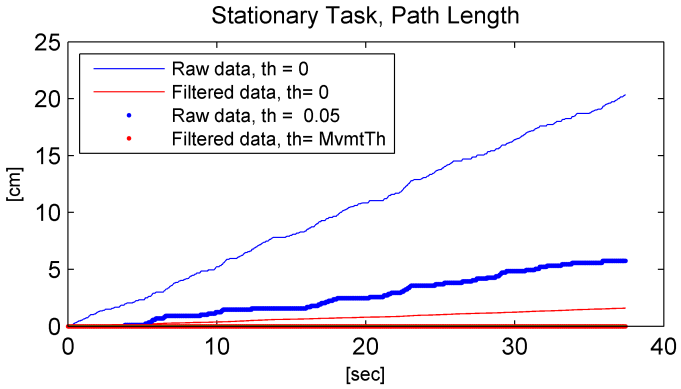


Figure 2.12: Accuracy of tool path length computation in the stationary task (only right hand cumulative tool tip path is shown). The legend indicates different thresholds used in computing tool path length. The most accurate performance occurs on filtered data where the movement threshold M_{th} is used, which consists of a unique value for each sensor derived in the same manner as in Figure 2.11 and presented in Table 2.7.

presence of filtering. For example, a simple first-order difference only approximates the derivative and successive differences exhibit more sampling artifact. More sophisticated techniques exist which provide better approximations of derivatives from data. Three techniques were evaluated in addition to the first-order difference to establish which one provides the most accurate time derivative approximations in the frequency range of interest. These include commonly used five-point stencils such as

$$f'(i) \approx \frac{-f_{+2}+8f_{+1}-8f_{-1}+f_{-2}}{12/T}$$

with T as the sampling period [1] where the notation $f_{+n} = f(i+n)$, a digital-domain implementation of a low pass infinite impulse response continuous time differentiator [4], and a Holoborodko smooth noise-robust central difference method [46]. For the Holoborodko technique, an 11th order differentiator exact up the fourth order was adopted:

$$f'(i) \approx \frac{322(f_{+1}-f_{-1})+256(f_{+2}-f_{-2})+39(f_{+3}-f_{-3})-32(f_{+4}-f_{-4})-11(f_{+5}-f_{-5})}{1536/T}.$$

Test data consisting of a 0.3 Hz sine wave was differentiated with each of the four different

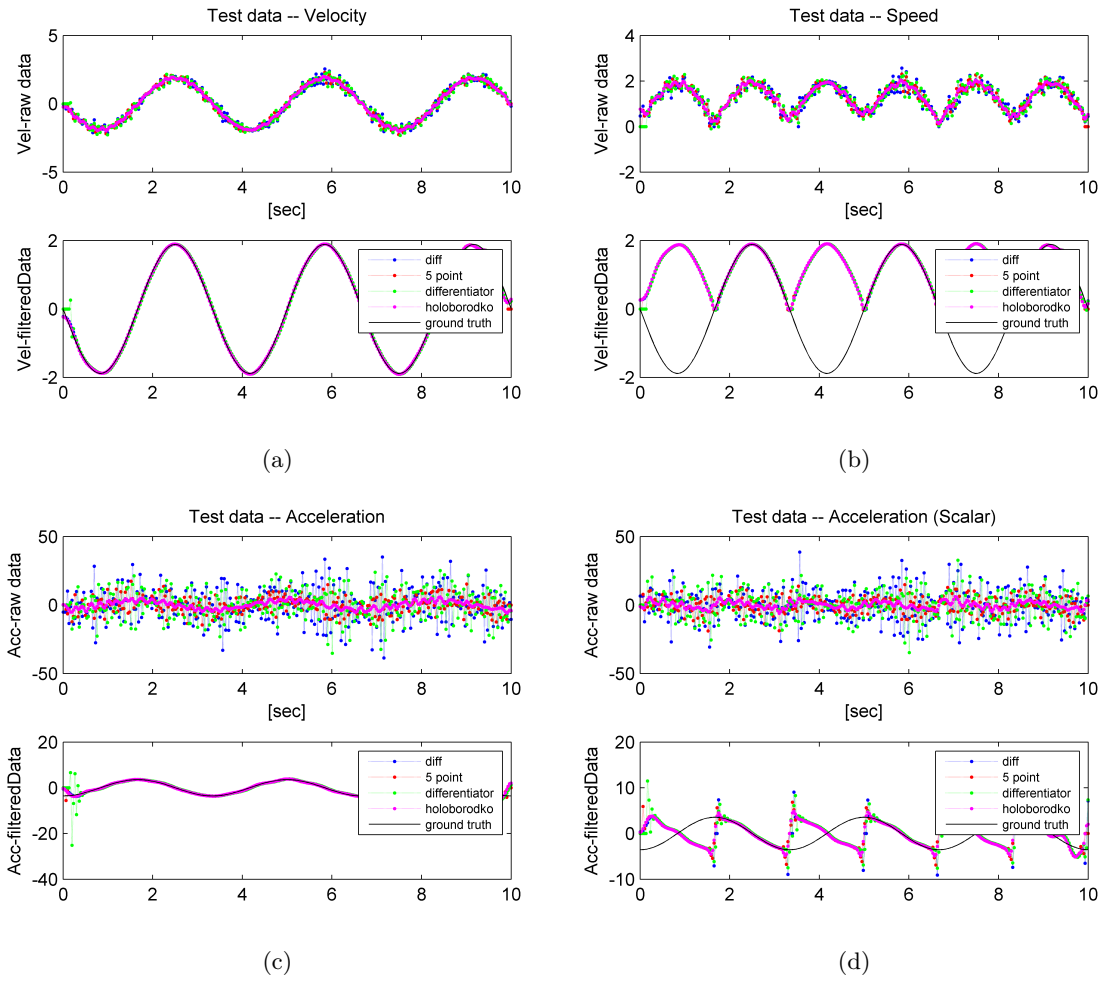
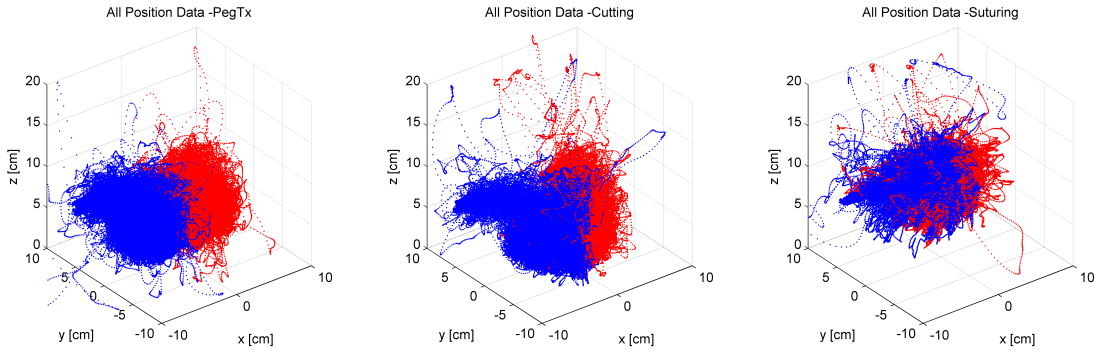


Figure 2.13: Validation of derivative extraction algorithms for (a) vector velocity and (b) its magnitude speed as well as (c) acceleration and (d) scalar acceleration. Of the five candidate algorithms evaluated, the Holodboroko central difference method proves most robust in all cases.



(a) (b) (c)

Figure 2.14: Cartesian task space plots of all concatenated tool tip data for all tasks.

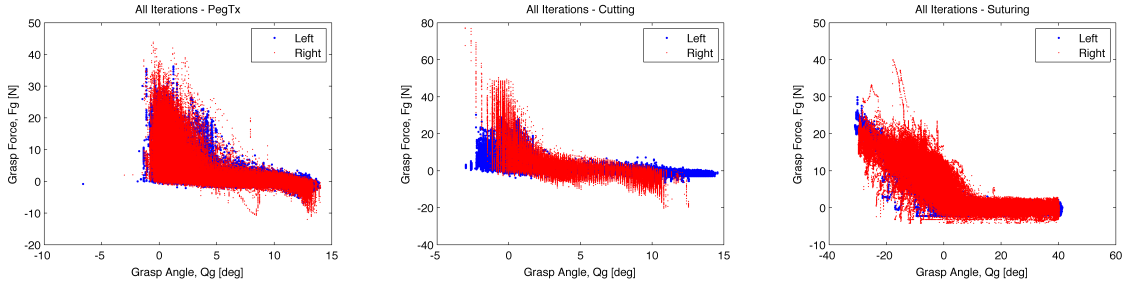
techniques enumerated above to approximate tool tip velocity, acceleration, speed (velocity magnitude) and scalar acceleration (magnitude of speed derivative). Figure 2.13 shows how each technique performed for all approximation tasks with both filtered and unfiltered data. The additional filtering dramatically reduces error, especially for higher derivatives. The Holoborodko central difference method proves most robust in all approximations (velocity, acceleration, etc.) and for both filtered and unfiltered data. It also minimizes initialization artifact. Moreover, its satisfactory performance was verified over the frequency range [0.1Hz 1.5Hz]. Thus, the Holoborodko method was adopted for all data differentiation tasks in this study such as extracting rates like velocity, speed, acceleration, or jerk (the time derivative of acceleration).

2.3.3 Overview of All Data Streams and Outliers

For each task, each sensor's entire data sequence was individually concatenated over all iterations of that task to generate summary plots. The purpose of this was to visualize typical data ranges for the different tasks and to examine the characteristics of outliers.

Figure 2.14 shows 3D plots of each task's tool path data concatenated over all recorded iterations. The left and right hand motion ranges appear to be most distinct in the peg transfer task and least so in the suturing task. This figure also indicates that for all tasks,

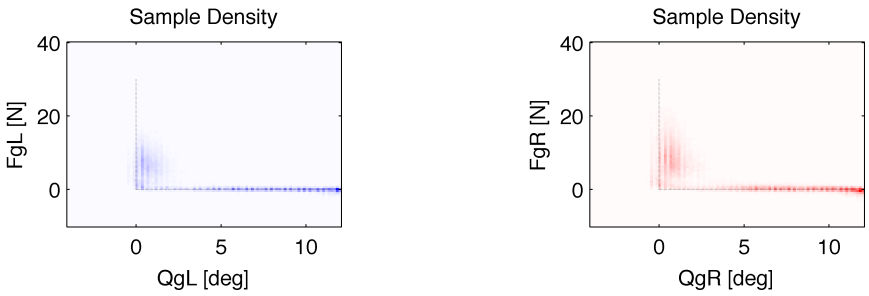
the majority of the time the tool tips appear to lie within a ball-like region with the occasional outlier samples induced by rare, reach-like events far beyond this region. In the peg transfer task, these events seem to point downward most often, suggesting dropped blocks. Such reach-like events occur most frequently for the suturing task which is probably due to the knot tightening motion that must occur at least three times in each correctly executed suturing task. This knot tightening procedure, particularly in inexperienced subjects, sometimes results in moving the tool tips out of the field of view. Since these events are rare, this suggests that threshold levels can easily be established to determine "rare events." These could be either rarely-executed user motions which may indicate suspicious activity or provide information about skill. Additionally, such events could indicate a sensor malfunction if values are beyond the threshold for an unexpectedly long period of time.



(a) (b) (c)

Figure 2.15: All iterations of (a) Peg Transfer, (b) Cutting, and (c) Suturing task grasp variables Q_g and F_g, the angular displacement and applied force at the grasper.

The grasp variables, grasp angle (Qg) and force (Fg) were also plotted against each other (see Fig. 2.15). These plots indicate how the ranges of motion and values such as decision thresholds can depend on the task. Decision thresholds such as regions in grasp space that indicate a grasped object vs. an empty grasper can be as simple as maximum or minimum sensor values. For example, a combined grasp force above 5 N and grasp angle above 1° may indicate that an object is being grasped, whereas grasp angles below 0° with moderate force (e.g., less than 20 N) may indicate that the user is grasping nothing for the block transfer task (Fig. 2.15a). However, such thresholds appear to be sensitive to the task.



(a) Left Hand Grasping (b) Right Hand Grasping

Figure 2.16: The Grasp variables for all iterations of the Peg Transfer task. The cloud plots (a) and (b) show sample density for left and right hand respectively. Thus, the darker the color the more samples exist in that region.

This is in part because the different tasks use different tools. Most notably, the suturing task has entirely different mechanism and handle and the tool metal shaft that actuates the grasper shaft exhibits less stiffness. This is evident in the sloped region of Fig. 2.15c. For suturing tools, the thresholds which worked to indicate the presence or absence of an object with the peg transfer graspers will not work. Thus, such decision thresholds derived from the grasp variables should be specific to the tool type used in a given task.

While Figure 2.15 adequately shows the spread of the data and highlights the outliers, it does not indicate how frequently the data samples occur in a given region. To accurately show this, a sample density plot is required where the relative frequency of samples is indicated by the color intensity. Darker colors indicate higher frequency. Figure 2.16 shows such a plot for the peg transfer task. It indicates that the majority of data fall quite close to the 0 N level for grasp angles above 4°: the dense points clustered about the right of the Qg axes indicate open graspers. The smudge near the inside corners indicates the most frequent grasp events. The size of this smudge indicates that the most frequent types of grasp events actually fall within a substantially smaller area than the wider spread indicated in Figure 2.15.

Chapter 3

SKILL CATEGORIES

Any candidate metric for scoring surgical skill requires a means to evaluate its accuracy. For construct validity—the extent to which a test measures the trait that it purports to measure—the test is the extent to which that metric discriminates between different levels of expertise. However, this requires correctly selected sets of data which accurately represent those levels of expertise. Without such a standard, construct validity cannot be tested, nor can additional validity analysis be undertaken. To this end, it is crucial to establish a "ground truth" that defines "true experts" as accurately as possible to successfully evaluate candidate skill metrics. Three such standards are often used in clinical practice. These include demographics-based groupings into skill categories, performance-based ranking, and manual evaluation by surgical faculty. All three techniques are adopted and ultimately combined in this study and are described in the sections below.

3.1 Overview of Individual Criteria for Skill Categories

The data used to establish skill categories in this study come from the multi-center data collection using the EDGE platform, described in Chapter 2 in Sections 2.1 and 2.2. Since each task iteration done on EDGE records data in addition to a video of the tool motion, additional performance evaluation is possible after data collection. This video review by faculty surgeons is described below, along with other individual criteria that are typically used to establish skill categories.

Performance-Based Skill Identification

Performance measures such as task time or operative errors often serve as a basis for skill identification in surgery. The FLS tasks are standardized and provide their own validated individual scoring mechanism for each task (see Table 2.5). The empirical distributions of

these FLS scores for all recorded iterations appear for each task in Figure 3.1. While this provides a validated and published way of scoring performance for a given task iteration, the specific thresholds used to determine FLS proficiency—whether a subject passes or fails the FLS exam based on a weighted combination of scores from all tasks—are not public. The pass/fail FLS proficiency criteria may suggest a convenient way to establish skill boundaries in the collected EDGE data. While the actual thresholds are not published, the method used to establish the thresholds is published and consists of taking the sum of normalized FLS scores for all five tasks over a pool of pre-determined competent and non-competent laparoscopic subjects and finding the cutoff score that gives best specificity and sensitivity [29]. Fraser et al. based competency in [29] on demographic criteria: medical students and junior surgical residents were classified as non-competent and chief senior residents, surgical fellows and attending laparoscopic surgeons were classified as competent. Taking the sum or scores over all tasks resulted in fewer overall misclassifications, but this potentially allowed an individual to pass the FLS certification even if they performed very poorly in one task but well in the others. Without repeating a similar procedure on the data collected with EDGE, the per-iteration FLS scores would be able to rank relative performances but not establish subject group boundaries. However, this approach hinges on the assumption that demographically-identified skill categories used to define competent/non-competent are accurate in the first place.

Demographics-Based Skill Identification

Levels of expertise are typically employed as a basis for construct validity (e.g., the setting of the FLS pass/fail criteria [29]). However, they may be difficult to objectively establish. Common practice employs professional or academic rank such as faculty surgeons compared to medical students or training surgeons in the post graduate years (residency) of study ranked by year: PGY1, PGY2, etc. This approach typically works well for groups at the extremes, however, in practice there is typically significant variation among faculty at a given institution in the amount of experience they have or the types of cases in which they specialize in. Similar disparities exist between trainees at different institutions or among

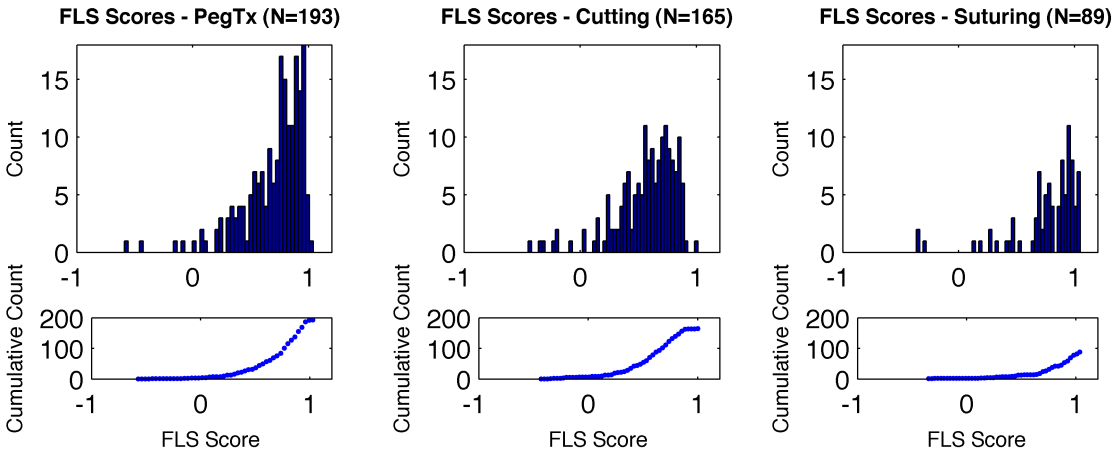


Figure 3.1: The FLS score distributions of all collected data for (a) Peg Transfer (b) Cutting and (c) Suturing tasks (top) as well as the cumulative score count per task (bottom).

trainees within a given institution.

A second approach approximates expertise with the total number of cases done. Since surgical training typically utilizes an apprenticeship model, trainees typically complete only parts of an entire case. The percentage of the total case they complete usually increases with training level. Thus the definition of "doing a case" must be clearly defined when employing this approach. Moreover, the resulting case count itself is approximate as case counts are not usually documented. Finally, even an accurate case count typically does not account for the underlying variety or subspecialty of skills favored by the surgeon: two individuals with identical laparoscopic case count may in practice use different techniques—one may never suture, the other only do high volumes of a specific procedure.

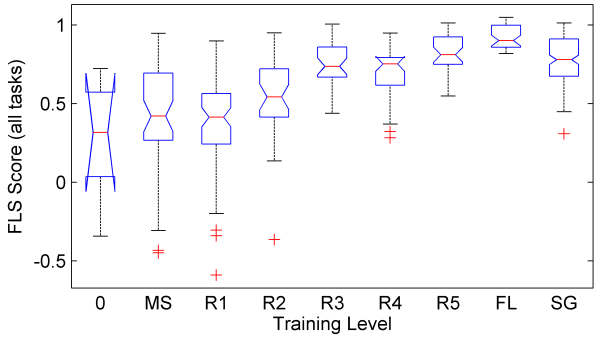
Alternative measures such as age, subspecialty or specific procedural preferences may also be considered. Age is easy to collect from voluntary subjects and universal to all subjects (assuming subjects volunteer such information). It may be used to generally distinguish among faculty, for example, though individuals may vary quite widely in the age they start or complete training. Categories that attempt to distinguish finer details of surgical practice or training often suffer from low resulting data counts: there is such variety

among the subject pool that not enough subjects are available to result in statistically significant numbers in these domains.

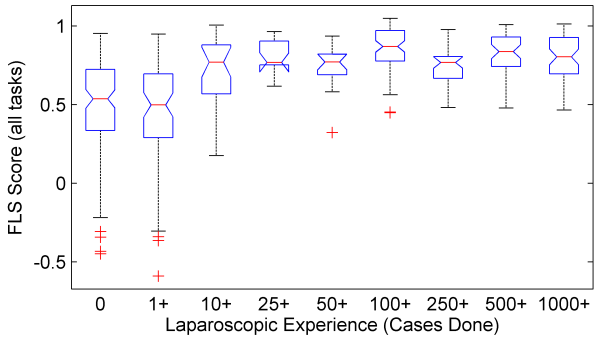
Demographic-based rankings are typically used to establish a small number of distinct skill groups as they do not provide a continuous scoring of skill. Thus they may be adequately suited for construct validity which requires categorical groupings, but not for concurrent validity which requires a continuous, graduated scoring scale for correlation analysis. Their combination with the continuous scoring scale of FLS may better address this issue. Figure 3.2 illustrates the relationships between FLS scores and demographic categories for the entire database. The box plots for 5th-year residents, fellows, and practicing surgeons (Fig 3.2a) exhibit notable differences: the practicing and faculty surgeons' group (SG) scores substantially lower than both 5th-year residents and fellows at FLS. This suggests some disagreement between these two criteria typically used to establish skill level.

Faculty Evaluation–Based Skill Identification

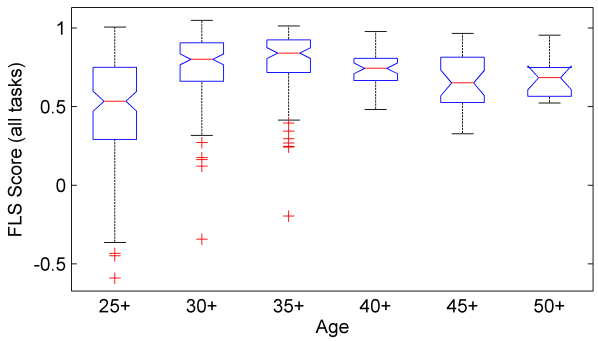
The current "gold standard" for evaluating skill consists of a blind, structured performance review by one or more surgical faculty. Established sets of Likert scales for various performance domains such as OSATS are typically employed as their inter-rater reliability has been established in the literature [70, 86]. This approach is resource-intensive, however, so it is less common or when it is practiced only a subset of the dataset is typically chosen.



(a) Box Plots For Training Level.



(b) Box Plots for Laparoscopic Experience.



(c) Box plots for Age.

Figure 3.2: Box plots of FLS scores vs. demographic categories. Such categories are often used to establish construct validity. (a) Training Level (0=none, MS=medical school, R1-5=post graduate year of surgical residency, FL=fellow, SG=practicing surgeons). (b) Laparoscopic Experience: the self-reported estimate of lifetime total number of laparoscopic cases performed. 1+ indicates subjects who had performed 1 or more cases, where they did 50% or more of the case. (c) Age: 25+ indicates subjects who are 25 years or older, etc.

3.2 Skill Categories and Combined Criteria to Establish "Ground Truth"

The most common criteria to establish skill level typically used in the surgical literature were detailed in the previous section. To some degree, skill level rankings provided by these approaches conflicted with each other among the collected data. Thus, criteria were combined to better estimate the Ground Truth (GT) of skill, i.e., the "absolutely true" level or category of skill. Each skill category was represented by a set of individual iterations. Combinations of skill criteria were used to establish sets of iterations (individual logs) that represented categories of skill for each task. The most rigorously established is the Ground Truth Expert set. It aims to represent "true expert" performances at the expense of a low number of iterations. Two alternative sets increase the number of iterations but trade off potential accuracy. These are primarily constructed to enable cross-validation of statistical results. Additional sets represent the Intermediate and Novice skill levels. All categorical sets are detailed below.

3.2.1 The Ground Truth Expert Set (GT-Exp)

Three methods commonly used in the surgical literature to identify "experts" to establish construct validity were employed. All were then combined to establish a set of individual iterations (not subjects) that exemplifies "true expert" skill. First, a subject's self-reported demographic criteria were considered. These included training level (None, Medical School, Post Graduate Years of residency 1 through 5, Fellowship, and Practicing Surgeon) as well as the self-reported estimate of all-time total number of laparoscopic cases performed. Only practicing laparoscopic surgeons and fellows who had completed more that 100 laparoscopic cases (where they did 50% or more of the case) were considered candidate subjects for the "true expert" candidate subject pool. Second, the highest FLS-scoring iteration for each task of each expert candidate was taken as a set of "true expert" candidate iterations. Finally, the third approach utilized an Objective Structured Assessment of Technical Skill (OSATS) protocol [70, 86] which was modified to focus exclusively on psychomotor skills, denoted p-OSATS and shown in Table 3.1. Only the "true expert" candidate iterations were considered for p-OSATS review. The videos of these iterations were randomly renamed and

ordered before evaluation. Two faculty surgeons (coders A and B) served as p-OSATS reviewers. Reviewers were blind to the identity and demographics of the subjects whose videos they reviewed and to the scores of the other reviewer.

Table 3.1: Psychomotor OSATS (p-OSATS) grading scale used to evaluate and numerically code psychomotor skill [70, 86].

Score	**Bimanuality**	**Motion Quality**
1	One arm paralyzed, offering no help to complete the step	Unnecessary, hesitant or awkward movements of tools
2		
3	Using both arms most of the time, but clear perceivable bias of accomplishing most of the task with dominant hand	Reasonably efficient movements of tools but frequent non-effective moves
4		Mostly Fluid
5	Both arms naturally complementing each other. Optimal use of non-dominant hand	Elegant, fluid, and efficient movements of tools

Reliability of p-OSATS

To address inter-session reliability, coder A reviewed all videos again, but in a different randomized order approximately 10 days after first review. Coder B's scores were compared to Coder A's combined scores for inter-rater reliability. Both coders were asked to limit their coding to psychomotor characteristics of the tool motion in p-OSATS alone, but were repeatedly asked to verbalize their thoughts during review which were subsequently recorded for each reviewed iteration.

Correlation coefficients were adopted to measure the level of ranking agreement between scores. Pearson's r was adopted as a measure of linearity, Spearman's ρ as a measure of monotonicity in the data. If Pearson's $r > 0.5$ is considered a strong fit and $0.4 < r < 0.5$ is considered a moderate fit, then both p-OSATS domains exhibited acceptable reliability ($p < .05$ or less) for the Peg Transfer and Suturing tasks, with strongest reliability for the

Table 3.2: Inter-session and inter-rater reliability of pOSATS scores. Coders only scored the Ground Truth experts candidates subset ($N = 56$). Inter-session compares coder A's scores with himself; Inter-rater compares coder A's mean scores with coder B's. Pearon's r and Spearman's ρ with corresponding (p-value) are shown.

Inter-session:		**Peg Tranfer**	**Cutting**	**Suturing**
Bimanuality	r	0.57 (0.01)	0.49 (0.03)	0.76 (0.00)
	ρ	0.45 (0.04)	0.51 (0.02)	0.69 (0.00)
Motion Quality	r	0.74 (0.00)	0.34 (0.14)	0.85 (0.00)
	ρ	0.70 (0.00)	0.32 (0.17)	0.90 (0.00)
Inter-rater:		**Peg Tranfer**	**Cutting**	**Suturing**
Bimanuality	r	0.55 (0.01)	0.04 (0.86)	0.70 (0.00)
	ρ	0.45 (0.05)	0.04 (0.88)	0.66 (0.01)
Motion Quality	r	0.63 (0.00)	0.27 (0.26)	0.78 (0.00)
	ρ	0.60 (0.00)	0.21 (0.38)	0.80 (0.00)

"Motion Quality" p-OSATS domain. Inter-session and inter-rater reliability were weakest for the Cutting task. See Table 3.2 for details.

Post analysis of recorded reviewer comments made during p-OSATS review of cutting tasks indicated coder B frequently emphasized handling of the gauze—a cognitive skill—and coder A intentionally avoided doing so to limit his review to the strictly psychomotor categories of p-OSATS, despite a strong desire to lower scores based on the same observed cognitive error in gauze handling. This suggests that including additional cognitive categories in scoring or more detailed coder training may improve overall reliability.

From the pool of Ground Truth Expert candidates (demographically identified practicing surgeons and fellows who completed more than 100 laparoscopic cases) the video of each subject's best FLS-scoring log for each task was submitted for blind p-OSATS review. The resulting scores for all evaluations appear in Fig. 3.3. Only logs whose video review scored 3 or above for both p-OSATS domains (Bimanuality and Motion Quality, see Table 3.1) made it into the Ground Truth Expert (GT-Exp) set. The GT-Expert set consists of logs (single iterations) not individuals. The selected logs demonstrate high performance by all three approaches used to determine skill: demographics, performance ranking, and faculty

evaluation.

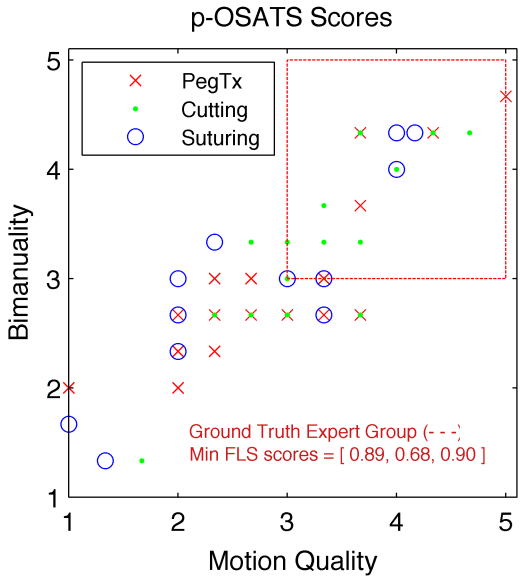


Figure 3.3: Ground Truth expert set determination based on p-OSATS scores. The red box indicates the boundary for inclusion into the Ground Truth Expert (GT-Exp) set. The minimum FLS scores of this set (used to determine FLS-Proficiency) were $FLS_{th} = (.89, .68, .90)$ for Peg Transfer, Cutting and, Suturing, respectively.

3.2.2 The Ground Truth Expert Plus Set (GT-Exp+)

The total number of iterations in the GT-Experts group is low ($N = 19$). In order to enable cross-validation of expert classification, each GT-Expert group subject's remaining logs that scored above FLS_{th} were added to the GT-Expert group to create the GT-Experts Plus (GT-Exp+) group ($N = 41$). These logs, however, did not have corresponding p-OSATS scores for their videos.

3.2.3 The FLS-Experts Set (FLS-Exp)

The FLS-Expert group consists of all logs (from any subject) that scored above the FLS threshold FLS_{th} set by the GT-Experts group. This followed the criteria originally used by FLS to determine their internal (unpublished) threshold for proficiency. This group is the most populous expert group.

3.2.4 The FLS-Intermediate Set (FLS-Int)

The FLS-Intermediate (FLS-Int) group was determined about the midpoint between FLS_{th} and $FLS_{NoviceMax}$. The 15th percentile centered at this midpoint for a given task was used to establish the minimum and maximum FLS thresholds for membership in this group. The FLS-Intermediate group can be used to test construct validity.

3.2.5 The FLS-Novice Set (FLS-Nov)

The FLS-Novice (FLS-Nov) group consisted of the lowest 15percentile of FLS scores. This determined the maximum FLS score for the Novice group: FLS_{NovMax}. Any logs that scored below this threshold composed this group. The FLS-Novice group can be used to test construct validity.

3.2.6 Summary Overview of Skill Categories

An overview of the criteria used to create sets of logs by skill category appears in Table 3.3 along with the specific numerical values employed. The minimum FLS-score of the GT-Expert set for each task determined the FLS threshold $FLS_{th,task}$ for that task (see Fig. 3.3) similar to the method originally employed by FLS in [29] to determine their internal threshold for proficiency. The distribution of skill categories among FLS scores appears in Fig. 3.4. Finally, Figure 3.5 illustrates how the number of iterations is reduced for each skill category by successive application of its criteria.

Table 3.3: Summary of criteria used to select each set of iterations and its intended purpose. N refers to the total count of iterations of each set.

Set	N	**Criteria**	**Thresholds**
GT-Exp	24	Ground Truth expert set: practicing surgeons' and fellows' best FLS-scoring logs whose video's p-OSATS scores both exceeded thresholds (see Fig. 3.3)	Bimanuality ≥ 3 AND Motion Quality ≥ 3
GT-Exp+	62	Ground Truth expert plus set: All logs of GT-Exp group subjects that scored above FLS_{th}	FLS_{th} =(.89, .68, .90)
FLS-Exp	157	FLS Expert set: All logs over threshold	$FLS_{th} = (.89, 68, .90)$
FLS-Int	71	FLS Intermediate set: 15th percentile of FLS scores about midpoint between FLS_{th} and FLS_{NovMax}	$FLS_{Int} =$ ([.59 .73], [.44 .57], [.66 .71])
FLS-Nov	67	FLS Novice set: All logs below 15^{th}-percentile FLS score.	$FLS_{NovMax} =$ (.52, .65, .48)

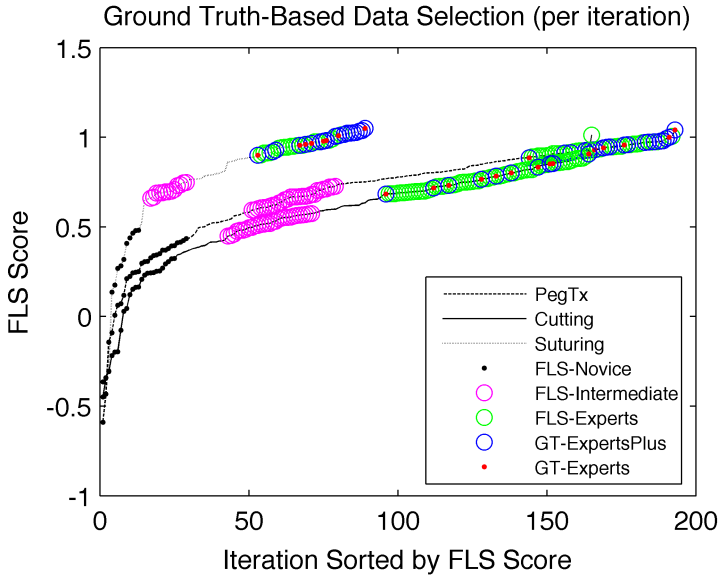


Figure 3.4: Skill Set selection overview and the resulting distribution of FLS scores. Note that Ground-Truth (GT) Expert logs are sparsely spread at the upper range of FLS scores; choosing only the top percentile of FLS scores would exclude logs deemed as expert performances by the combined p-OSATS and Laparascopic Experience criteria of determining expert performance.

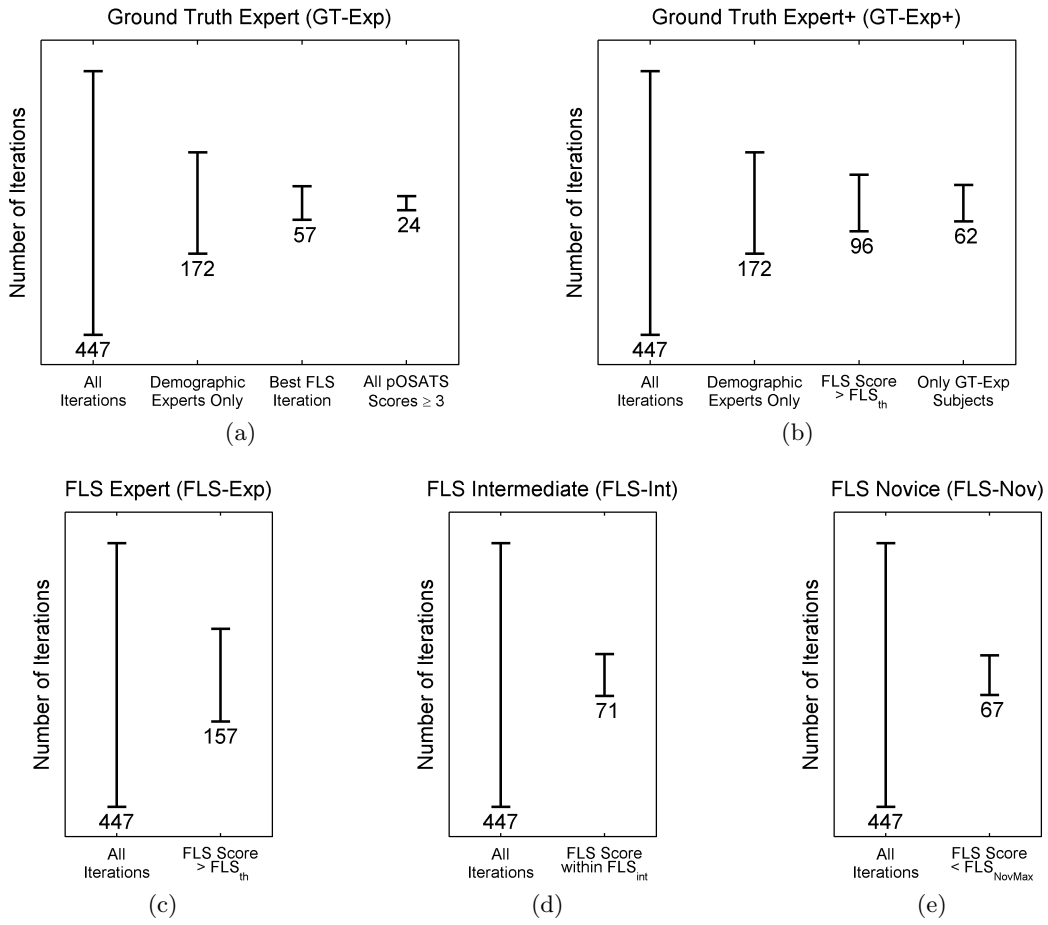


Figure 3.5: The reductions in the total number of iterations based on adding each successive criteria that define each skill category. The total counts for all tasks are shown below each bar. (a) The Ground Truth Expert set iterations were all p-OSATS-reviewed yielding highest accuracy but fewest iterations. (b) The Ground Truth Expert Plus (+) set contains some additional iterations that do not have p-OSATS scores, but come from the same subjects found in Ground Truth Experts and score above the FLS score threshold, FLS_{th}. The remaining skill categories (c,d,e) were obtained based on FLS scores alone.

Chapter 4

FEATURE EXTRACTION AND SELECTION

Algorithms to quantify surgical skill are herein developed by a four-step process (see Fig. 4.1). First, the feature extraction step attempts to reduce continuous surgical tool motion and force data into a more intuitive and computationally tractable data stream of discrete "features." For example, vector quantization can reduce the data dimension and result in more tractable discrete outputs. Segmentation of the data provides a means of establishing sequential context within a task, which can give a more intuitive understanding of the analysis: one can review the short video clip that corresponds to the segment of motion data to help establish what happened during that segment. Countless feature extraction schemes are possible, so several hypotheses derived from observation and prior work are used to implement a pool of feature extraction schemes. Then the feature selection step employs quantitative methods to suggest which feature extraction schemes best discriminate surgical skill, effectively ranking their suitability. This results in a small set of best-candidate extracted features. The sequential modeling step then incorporates these extracted features into models of sequential (temporal) characteristics of surgical skill.

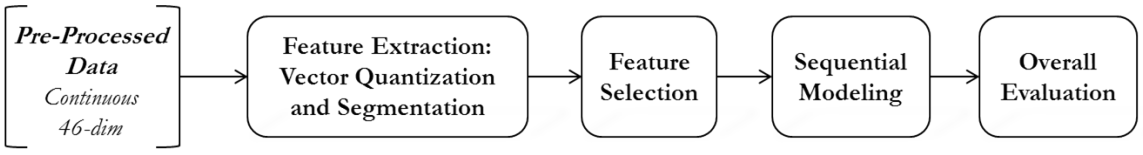


Figure 4.1: Overview of algorithm development for surgical skill evaluation based on time-sampled surgical tool motion data.

Finally, the overall evaluation step evaluates which sequential models best discriminate surgical skill and presents the criteria used to make this evaluation. These criteria are applicable to all surgical skill metrics—not just those developed here—to enable a uniform

comparison among alternative approaches. This chapter details the feature extraction and selection steps. The subsequent chapter describes the sequential modeling and overall evaluation steps.

4.1 Feature Extraction

A number of data sources can potentially be utilized for surgical skill evaluation. This work makes use of 46-dimensional continuous data derived from twelve time-sampled sensors (via a pre-processing step described in Chapter 2, Section 2.3) along with corresponding video from the EDGE platform. Additional data sources can easily be added. This results in high-dimensional data sets. Yet the desired output of quantified surgical skill is typically of low dimension and cardinality: it may even be a binary Pass/Fail output. High dimensional data incur computational challenges like processing time, memory, and algorithmic complexity for tractable solutions. Thus, it is necessary to identify and extract only the most relevant pieces of data from the higher dimensional space and disregard the irrelevant data. This process is called feature extraction. Its purpose is to significantly decrease the dimension and cardinality of the input data into a more tractable and relevant feature output. In this work, only discrete features are considered. Following the approach of Rosen et al., the widely-used signal processing technique of Vector Quantization (VQ) [36] is adopted to perform this extraction for each feature.

A feature is herein defined as a mapping:

$$f_{feature} : \mathbb{R}^N \mapsto \mathbb{Z}_k \tag{4.1}$$

where N is the dimension data (e.g. all dimensions or a subset of pre-processed continuous data) consisting of real numbers $\mathbb{R}$, which maps to a range of zero-based integers $\mathbb{Z}$ with cardinality k: $(0, 1, 2, ..., k)$. In VQ, these integers are called codewords. The integers merely serve as labels; they do not necessarily indicate ranking or order in the mapping. (For example, if three raw data samples map to labels 1, 2, and 100, then data labeled 1 is not necessarily closer or "more similar" to data labeled 2 than data labeled 100 is.) Thus, initially the order of integers in a feature may be arbitrarily assigned, but once established,

it must remain consistent for all subsequent use of that feature.

4.1.1 k-means Vector Quantization

Rosen et al. adopted k-means clustering to establish the VQ mapping $f_{feature}$ [90]. Given the codebook size k, the objective of k-means is to find a partitioning in $\mathbb{R}^N$ such that

$$\underset{S}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{4.2}$$

where x_j is the j-th vector-valued data point in partition S_i and μ_i is the mean of all data points in S_i. The set of all centroids μ for a given partitioning scheme is called the VQ codebook, and each μ_i is called a codeword corresponding to $\mathbb{Z}_k$ in Eqn. 4.1. The overall measure of how well such a codebook represents or compresses the data is called the distortion:

$$distortion = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{4.3}$$

Note that the distortion for a dataset with n points, the k-means codebook of size 1 (i.e., where $k = 1$) is simply the total variance of the data (unnormalized variance) and a codebook of size n is identically equal to the data set.

The work reported here differs from the approach of Rosen et al. in several ways. Rosen's data included seven additional sensors present in the Blue DRAGON but absent in EDGE: binary tissue contact and six degrees of freedom (DoF) of force/torque sensor in the shaft of the tool for a total of 13 sensor dimensions for each hand. In contrast, the present method utilizes only six sensors in EDGE for each hand (x, y, z position, tool rotation, grasp angle, and grasp force). The 6-DoF force/torque sensor was eliminated due to excessive cost and difficulty of integrating it into a surgical tool without severely constraining its range of motion. The binary tissue contact sensor was also eliminated since the capacitive sensing scheme it employed was not compatible with the objects in the FLS tasks or most other dry-lab tasks.

Rosen used a supervised VQ training paradigm in which 10-word codebooks were trained with hand-selected data that represented 15 prototypical surgical gestures with a total code-

book size of 150. In contrast, the present work uses an unsupervised training approach: no manual pre-labeling of data is required as with Rosen's hand-selected training sets to find optimal VQ codebooks. This is primarily due to the missing 6-DoF force/torque and tissue contact sensors which provided data crucial to Rosen's 15 prototypical gesture approach. Instead, a variant of k-means called the Full Search Greedy VQ algorithm specifically developed by Kowalewski et al. to improve on Rosen's original approach provides unsupervised training and is used in this chapter to optimize codebook training and provide lower distortion for surgical applications [58]. The comparison presented in [58] between the traditional k-means algorithm employed by Rosen et al. and other variants is reproduced in Table. 4.1. The present study adopts the Full Search Greedy VQ algorithm.

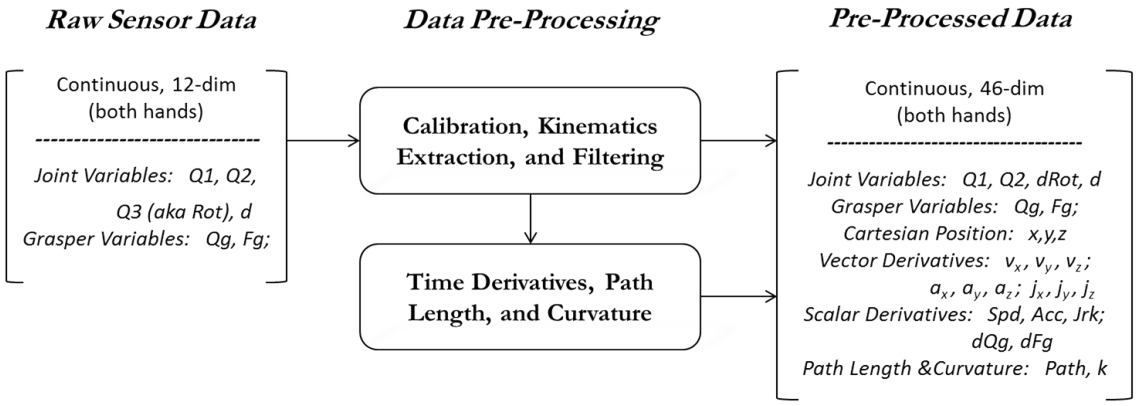
Table 4.1: Overview of different VQ algorithms taken from [58]. The Full Search "Greedy" VQ algorithm was adopted. *Normalized execution time is based on identifying 250 codewords using GLA on Sun Ultra Enterprise 450 machines with dual UltraSPARC II processors in 5 min.

	GLA (Method I)	*Modified GLA (Iterated Method I)*	*Full Search "Greedy" VQ Algorithm*	*Traditional K-means Algorithm*
Step 1	Place first word at mean vector of data	Place first word at mean vector of data	Place first word at mean vector of data	Initially choose code-book size N and randomly distribute N points in the data space
Step 2	Increase the size of the codebook by 2^n, splitting all code-words	Increase the size of the codebook by 1, splitting the most populous codeword(s)	Find the codeword giving largest distortion drop when split (full search through codebook, using step 4 each time) and split it	Iteratively migrate each codeword towards local point clusters to min-imize global distortion
Step 3	Relocate each word until a (local) minimum of distortion is reached	Relocate each word until a (local) minimum of distortion is reached	Relocate each word until a (local) minimum of distortion is reached (using GLA techniques)	Continue step 2 until the (percent) change in distortion is less than threshold
Step 4	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Repeat steps 1-3 with different random initializations to approach global extrema
Normalized Execution Time*	1	72	1440 (degrades greatly w/ larger book sizes)	Forbiddingly Slow (sensitive to initial conditions)
Relative Distortion	High	High	Low	Lowest (theoretically) (highly sensitive to initial conditions)

4.1.2 Feature Input Variables and Normalization

The continuous inputs for feature extraction schemes are provided by the pre-processing steps of Chapter 2, Section 2.3. To summarize how the pre-processed data emerge from the raw sensor signals of EDGE, Figure 4.2 provides a general overview of the data pre-processing step. Each feature extraction scheme may select its own, unique subset of continuous inputs from among the 46 variables of pre-processed data. This selection of input variables from the pre-processed data determines the input dimension of the feature extraction scheme.

Figure 4.2: Overview of the data pre-processing stage. The output, pre-processed data, provides the 46 input variables for feature extraction schemes.



It is not obvious which combination of input variables of a feature extraction scheme provides the most relevant information for skill evaluation. Choosing only a subset instead of all input variables has several benefits. It is computationally more tractable and requires less data collection for statistical significance as it avoids the curse of dimensionality. Also, it may reveal that certain sensors are not necessary and that they can be removed from the system, thus reducing cost and complexity associated with the hardware. Rosen et al. evaluated only one set of input variables (tool tip velocities, grasp variables, forces, torques, and binary contact) as inputs to a single, supervised codebook of size 150. This is effectively a single feature extraction scheme with only one set of input variables, one codebook size,

and one data segmentation approach. In contrast, the present study evaluates multiple candidate feature extraction schemes with variety in all three areas: input variable selection, codebook size, and task segmentation schemes. The input variable selection is described immediately below, while the other two are detailed in subsequent sections.

Input Variable Subset Selection

A total of 11 different subsets of input variables from the 46 variables of pre-processed data were considered as inputs to feature extraction schemes. Each scheme is referred to by an identification tag that abbreviates the input variables selected. These tags serve to identify candidate features in a way that communicates meaningful information about the choice of variables. For example, $Feat_{Vel-dRta-Gr,k}$ refers to a feature extraction scheme that uses vector tool velocity, time-derivative (rate) of tool rotation, and grasp variables—grasp angle and grasp force—as input variables and a codebook size of k to find the optimal codebook. Once such a codebook is established, $Feat_{Vel-dRta-Gr,k}$ maps the continuous input variables into a sequence of discrete integer codewords: any data points falling closest to the j-th codeword, that is, into partition S_j of Eqn. 4.2 get mapped to integer j. This completes the feature extraction for that given scheme and input data.

The first feature extraction scheme, $Feat_{Vel-dRta-Gr,k}$, attempts to be as close to the work of Rosen et al. as possible given the substantial differences introduced by EDGE (e.g., lack of sensors and unsupervised VQ codebook training). This was intended to provide a starting point for development as well as a reference for evaluating whether EDGE can provide similar skill discrimination with fewer sensors. Rosen, Sinanan et al. based their feature extraction scheme on domain-specific knowledge that depended heavily on force/torque and tissue contact data, drawing from extensive surgical faculty teaching experience. Since such data is no longer available for EDGE, Rosen's feature extraction scheme cannot be identically replicated. It uses the same variables of velocity, tool rotation, and grasp variables and provides no surrogates for force or tissue contact. The velocity variable in EDGE is tip position in Cartesian space. Rosen et al. used joint velocities since consistent Cartesian coordinates systems were not available between trials. This makes the resulting model

potentially device-specific as joint rates might change with mechanism even if the surgeon moves the tool tip in an identical way. The tool rotation rate $dRta$ is also modified by a tangent inverse function (see Chapter 2, Section 2.1.2). The grasp variables Q_g and F_g should be identical between EDGE and the work of Rosen et al.

The input variable subset denoted by Vel-dRta-Gr-dQg includes the rate of grasp angle change in addition to three variables described above. It is evident from observing a variety of surgical procedures that there is considerable variety in grasp behavior as displayed by grasp angle alone. Some objects are small and register nearly zero grasper angle, Q_g, while others are so large, the grasper angle Q_g is only marginally different from a fully open grasper. Also, a variety of grasping behavior was observed among the subjects,with some keeping the graspers open while moving and others keeping them closed unless interacting with tissue. However, the rate of grasp angle always changes while subjects are actuating the graspers. Thus, it was hypothesized that grasping rate could provide information that can indicate when grasping was taking place and for how long, which could in turn serve to discriminate skill level.

EDGE provides accurate Cartesian position information within a task and between the different machines due to its design and calibration procedure (see Chapter 2, Section 2.1.2). Observation of recorded task videos suggested that how subjects positioned tools either helped or inhibited their performance in the FLS tasks. Also, experienced surgeons appeared to require less open space than novices to complete task objectives like block transfers or knot tying. The difference in tool trajectories between an expert and novice surgeon shown in Figure. 2.2 illustrates this phenomena. Thus, the vector of Cartesian tip position variables was hypothesized to contain relevant, skill-discriminative information. The vector of three-dimensional position, indicated by Pos, was combined with the same rotation rate and grasp variables. This combination was labeled Pos-dRta-Gr.

Perhaps the most consistent observation made while watching surgeons use EDGE or in reviewing surgical video of tool motion was that viewers can rather quickly discriminate the approximate skill level of the subject. No specific knowledge of the task seems to be required, rather the quality of tool motion seems sufficiently indicative. This is particularly the case when watching two videos side-by-side. Further investigation of such videos showed

that it did not matter what the subjects were doing so much as how they were executing their motions. Direction or location did not seem to make a difference. Thus, it was hypothesized that scalar measures of motion like speed (magnitude of vector velocity), magnitude of acceleration, magnitude of jerk (the time derivative of acceleration), and tool path curvature would be sufficient to discriminate skill based on this observed phenomena. The scalar motion variables of speed, acceleration, jerk, and curvature were thus combined tool rotation rate and grasp variables. This combination was labeled SpdAccJrkCrv-dRta-Gr.

The remaining input combinations were based on excluding variables. The purpose of this was two-fold. First, additional segmentation schemes were developed (described in the subsequent sections) that used the grasp variables to segment motion. This effectively combined the grasp variables and motion variables in a different manner suggesting they can be excluded from the VQ input set as they were not expected to provide much additional relevant information. Second, it was hypothesized that scalar tool motion variables alone could discriminate skill, since observing the quality of tool motion in video reviews appeared to distinguish skill most easily. However, because it was not clear what quantitative aspects of tool motion contributed to this phenomenon, a number of combinations of scalar tool motion variables was considered: SpdAcc-dRta, SpdAcc, SpdAccJrk, and SpdCrv. The combination of Grasp variables Q_g and F_g was labeled Gr. This combination, along with its time derivatives labeled dGr, was also employed as an input combination. It was not expected to discriminate skill on its own, but rather to serve as an alternative way of combining grasp variables and motion variables for subsequent analysis. Finally, the rotation rate dRta was also adopted as a single-variable input simply to see how much information it provides on skill determination. This was due to the observation that novice surgeons often failed to use their wrists to rotate tool tips into optimal positions while more experienced surgeons did this consistently throughout their tasks.

The list of proposed input variable subsets for all feature extraction schemes used in this work appears in Table 4.2 along with descriptions and the tags used to identify them.

Table 4.2: The list of pre-processed data variable subsets used as inputs for feature extraction schemes and their identification (ID) tags. Position is in Cartesian coordinates, speed is the magnitude of velocity, acceleration and jerk are the magnitudes of the first and second time derivatives of position respectively, and curvature $\frac{|\dot{p}\times\ddot{p}|}{|\dot{p}|^3}$ refers to the curvature of the tool tip trajectory.

ID	**Description**	**Dimension**	**Units**
Vel-dRta-Gr	Vector velocity, rotation rate, grasp variables	6	$[cm/s, cm/s, cm/s,$ $^\circ/s, N/s]$
Vel-dRta-Gr-dQg	Vector velocity, rotation rate, grasp variables, derivative of grasping angle	7	$[cm/s, cm/s, cm/s,$ $^\circ/s,\ ^\circ/s, N/s]$
Pos-dRta-Gr	Vector position, rotation rate, grasp variables	6	$[cm, cm, cm,$ $^\circ/s, N/s]$
SpdAccJrkCrv-dRta-Gr	Speed, acceleration, jerk, curvature, rotation rate, grasp variables	7	$[cm/s, cm/s^2, cm/s^3,$ $1/cm,\ ^\circ/s,\ ^\circ/s, N/s]$
SpdAcc-dRta	Scalar speed, acceleration, rate of rotation	3	$[cm, cm/s^2,\ ^\circ/s]$
SpdAcc	Scalar speed, acceleration	2	$[cm, cm/s^2]$
SpdAccJrk	Scalar speed, acceleration, jerk	3	$[cm/s, cm/s^2, cm/s^3]$
SpdCrv	Scalar speed, curvature	2	$[cm, 1/cm]$
Gr	Grasp Variables: grasp angle Qg and force Fg	2	$[\,^\circ, N]$
dGr	Derivative of Grasp Variables	2	$[\,^\circ/s, N/s]$
dRta	Rotation rate	1	$[\,^\circ/s]$

Input Normalization

Another substantial departure from the Rosen et al. approach includes normalizing all input data to the VQ codebook. The algorithm used to train codebooks from data attempts to minimize the pairwise Euclidean distance between each data point and the nearest centroid (Eqn. 4.2). Thus if the range of one input variable is substantially smaller than another, it will effectively be ignored by the algorithm. For example, the grasping angle may only move through a range of 10° while tool rotation has a range of 360°; moreover, some variables employ completely different units. Another issue is that of extreme outliers. The algorithm will erroneously assign entire clusters to extreme outliers even if they are very rare. To circumvent these issues, each input variable is normalized with a linear transformation that maps its 2^{nd} and 98^{th} percentiles to -1 and 1, respectively.

Histograms of each sensor's data were generated along with upper and lower percentile markers to illustrate the normalization of input variables. Representative cases appear in Figure 4.3, but the complete set of histograms for all candidate input variables partitioned by task and left or right hand of each hand appears in the Appendix (Figures A.1 ff.). These histograms illustrate the need to address outliers. They also indicate that some input variables are not normally distributed, particularly the grasp variables. Hard decision boundaries at the upper and lower 0.5 percentile regions were used to discard outliers. Since these events happen so rarely, they are excluded from influencing the codebook training. The 98th percentiles serve as a "soft" boundary that models can use for normalization: most of the data is within the upper and lower 98th percentiles but the distribution tails may extend relatively far beyond the percentile values. Furthermore, this method suggests a simple mechanism for on-line recognition of anomalous behavior: when sensor values exceed these thresholds, this can be tagged as atypical behavior for a given surgical task. For each task, such detection may identify task-specific events. For example, moving the needle beyond the camera's field of view during suturing, a dropped block in Peg Transfer, or detaching the gauze from the task plate in cutting.

The effect that identifying and removing the outliers from the data has on the overall 3D tool path trajectories for each task can be seen in Figure 4.4. The rare, far-reaching

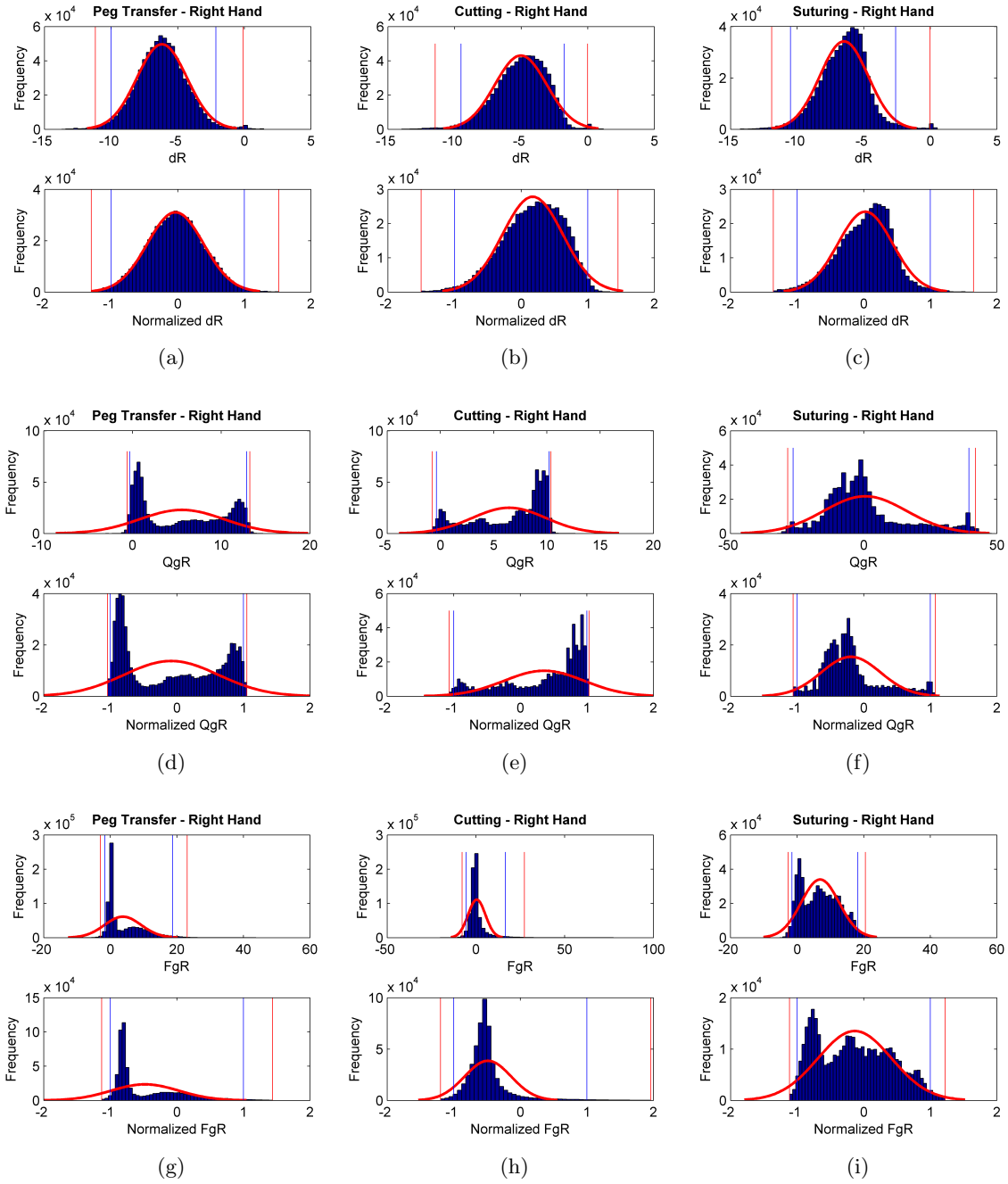


Figure 4.3: Representative histograms and normalization of input variables for all three tasks for the right hand only. dR is the tool insertion distance in cm. Qg is the grasper angle (in °) and Fg the grasp force (in Newtons). 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers. Note that some distributions are clearly non-normal and multi modal.

paths are eliminated. Such data points can cause the vector quantization algorithms to provide incorrect or anomalous results due to their reliance on Euclidean distance: points that are very far away from the majority of the data will tend to accrue their own, very sparsely-populated clusters.

4.1.3 Optimal and Maximum Codebook Sizes

Given that Rosen's method worked successfully with 150 codewords for 26 sensor-based Blue DRAGON data, 150 was taken as a maximum codebook size for the 12 sensor-based EDGE data. It is reasonable to assume that the pruning of dimensions decreases overall variance in the data. Since EDGE provides less than half of the original dimensions of the Blue DRAGON data, a smaller codebook size can yield the same normalized distortion rate (compression ratio) as that of the Blue DRAGON since there is less information to compress. The EDGE codebook size that provides an equivalent compression rate is herein called the optimal codebook size for EDGE. Because the underlying data from Blue DRAGON and EDGE are of substantially different lengths and type, it is not enough to simply equate distortion, but rather distortion rate, that is, the percentage of decrease in normalized distortion due to an increase of codebook size from k to $k+1$ where normalized distortion, d_n, is defined as

$$d_n = \frac{\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2}{\sum_{x_j} \|x_j - \mu_i\|^2}. \tag{4.4}$$

The distortion rate which is taken to be equivalent to that of Rosen et al. occurs when $\Delta d_n / \Delta k \leq 0.1\%$ based on the value convergence threshold used in [90]. An added benefit of the Full Search Greedy VQ algorithm over traditional k-means is that it generates intermediate optimal codebooks of increasing size up to k. Thus, for any given feature, 150 codebooks were trained from size 1 to 150. Then a 4^{th} order monotonic rational curve was fit to the data via non-linear regression to ensure differentiability on the domain [1 150]. The codebook size below the threshold slope of $\Delta d_n / \Delta k \leq 0.1\%$ is taken as the optimal codebook size for a given feature. This process is illustrated in Fig. 4.5. In all cases, individual codebooks were trained separately for each hand. Thus, the encoding of a surgical task uses two unique codebooks of a particular size to encode (vector quantize) the data for

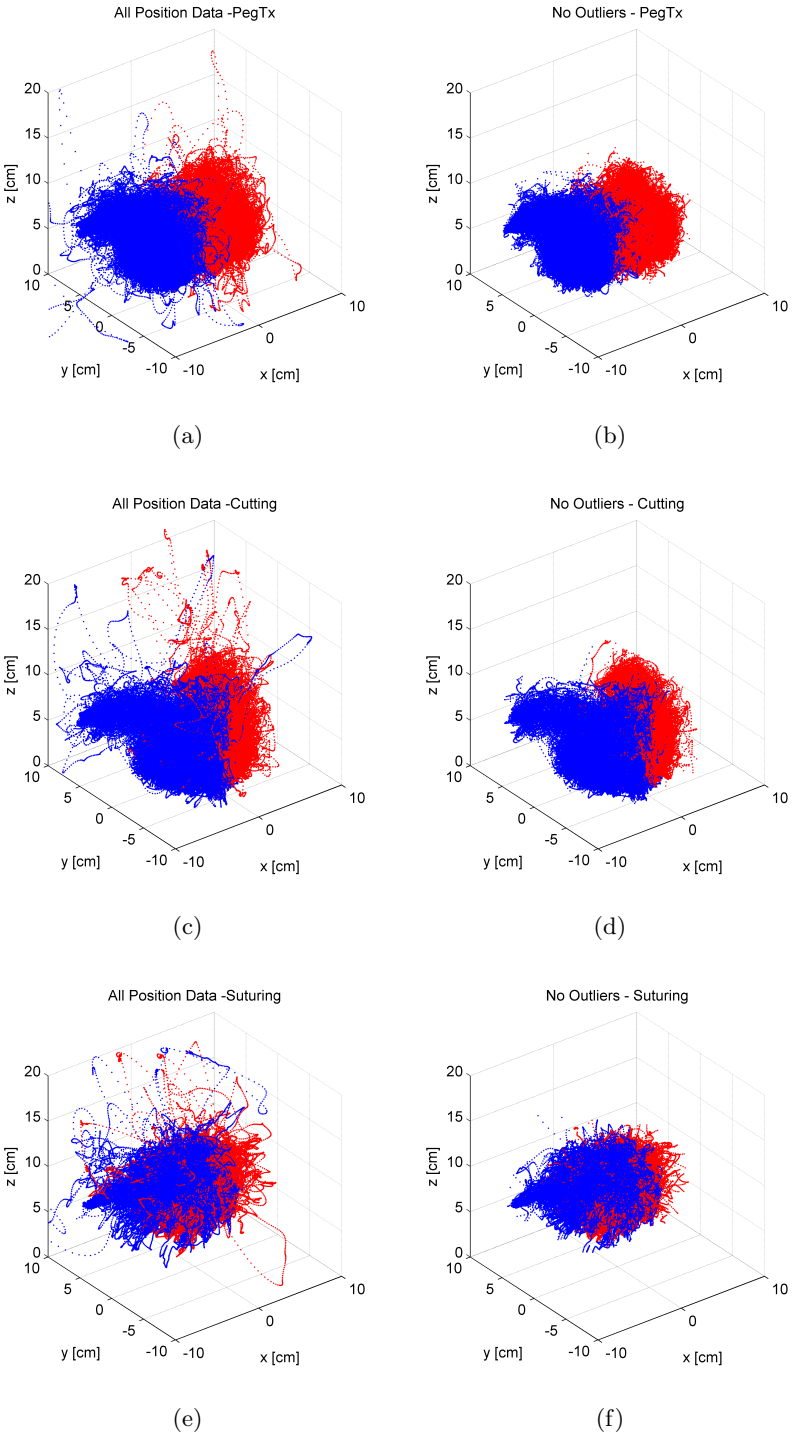


Figure 4.4: Cartesian task space plots of left (blue) and right (red) hand tool tip trajectory of all concatenated task data (left) and all data with outliers removed (right) for all tasks. A sample was removed if any of its vector values exceeded percentile limits.

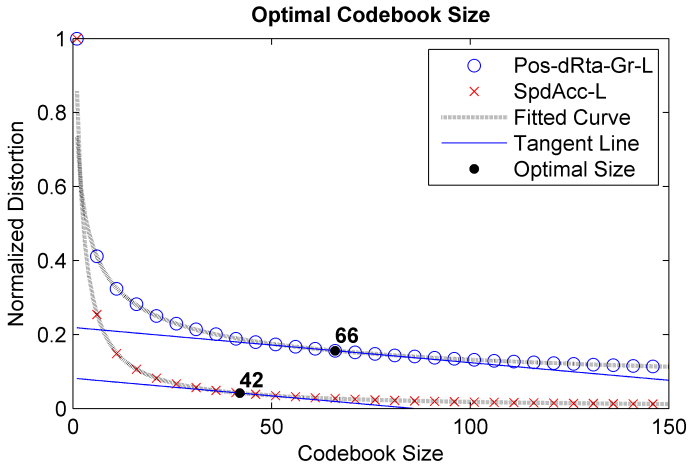


Figure 4.5: Distortion curves from codebook training used to determine optimal codebook size. Two sets of input dimensions: Pos-dRta-Gr-L and SpdAcc-L each have 150 codebooks trained. The nonlinear fitted curve provides a smooth derivative to find the closest codeword where the tangent line drops below slope 0.1%. For clarity only every 5th codebook size is shown.

each hand.

4.1.4 Universal Segmentation Schemes

In addition to the whole-task analysis employed by Rosen et al., three methods for surgical task segmentation are explored. Prior segmentation work proved quite successful but was strictly tied to a single robotic platform and very specific to a single suturing task [65, 85, 115]. In contrast, the present work attempts to find segmentation schemes that are universally applicable across the broad spectrum of surgical procedures, specialties, and modalities (i.e., manual, open, robotic, or laparoscopic). The development of such a universal approach to segmentation schemes opens the possibility of universally applicable skill metrics as well.

Two observations enable such universal segmentation schemes: all surgeons must move a tool or end effector to a target position, and once the target is reached, virtually all tools provide some "actuation" that the surgeon executes. Perhaps the most common example of an actuation is a grasp of tissue, but this can be generalized to cutting, cauterizing,

coagulating, etc. Also, each individual movement can be construed as a "reaching" event which is ubiquitous in all surgical procedures, independent of the modality.

The goal of these segmentation schemes is to add additional structure to the models to improve their skill discriminative power. For example, two whole-task sequences may prove difficult to compare since even if two surgeons start a task in the same way at the same time, they will naturally diverge over time, making direct comparisons difficult: one surgeon may be reaching for a target at time t while the other may have already been holding tissue for several seconds. The resulting tool trajectories would look very different at time t, even if both surgeons have the same skill level. However, a segmentation scheme that divides each task in a consistent way—by grasp events for example—would enable better comparison between similar segments.

The following schemes leverage these observations in an attempt to universally segment surgical motion. These labels are used throughout this work to identify the segmentation scheme. For simplicity, the discussion enumerates only grasping activities such as grasps or grasp events, but the approach is generalizable to "cuts," "coagulations," etc.

- SegGr: "Grasp events" derived from Grasp variables Qg and Fg. These were evaluated identically for all types of tools across tasks (e.g., graspers, shears, needle drivers). When Gr variables crossed a threshold $[3°, 4N]$ for more than 200ms, a grasp event was registered and maintained until the threshold was transgressed again for at least another 200ms. Additionally, a 5% hysteresis trigger level was used to reduce transient behavior. Tool motion segments between grasp events (i.e., while the graspers were open) were labeled *betweenGr* Segments and *withinGr* segments while the graspers remained closed. This method is generalizable to other types of actuation events; for example, the use of electro-cautery can be considered a generalized grasp event analogous to a closed grasper.

- SegFg: "Grasp events" derived solely from grasp force Fg. When Fg crossed a threshold ($[.5N]$ for needle drivers and $[3N]$ for all other tools) for at least 200ms, a grasp event was registered and maintained under 5% hysteresis until the threshold was transgressed again for another 200ms or more. Tool motion segments between grasp events

(i.e., while the graspers were open) were labeled *betweenFg* Segments and *withinFg* segments while the graspers remained closed. As in the case of SegGr, this method can also be generalized to different tool types.

- SegZSpd: When tool speed fell below 1.5 cm/s for at least 200ms, this indicated a "zero speed" event. Segments between such events were labeled *betweenZSpd*. Because all surgical tools move, this approach can be generalized to virtually any surgical tool.

A plot illustrating a segmentation scheme on grasp variable data from a FLS peg transfer task is shown in Fig. 4.6.

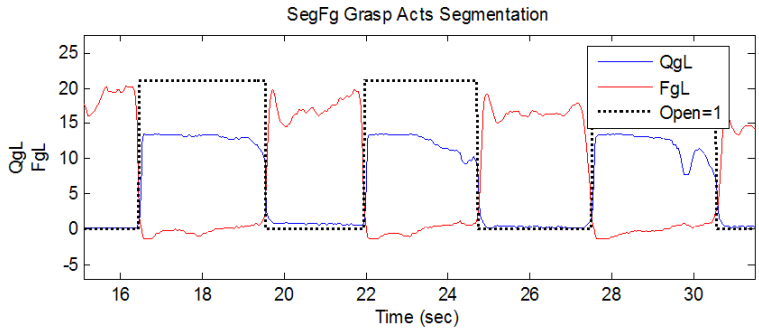


Figure 4.6: Example of the SegFg segmentation scheme in action for a Peg Transfer task performed on EDGE. The dotted black line indicates the binary state of the Grasp: 0 indicates "Closed" and the binary 1 indicating "Open" was scaled by 20 to make it easier to see in the plot. QgL indicates the grasping angle (in °) and FgL the grasping force (in Newtons) for the left hand. A grasp "action" or act was defined whenever SegFg transitioned between Closed and Open states.

4.2 Feature Selection

The number of possible features can grow without bound. For example, each of the 11 input variable combinations, all 150 trained codebooks for both hands, and the whole task approach and the three suggested segmentation schemes considered up to this point already result in over 10,000 possible combinations for candidate feature extraction schemes. It is important to identify and use only the features that best discriminate skill level and to avoid those that do not. Information Gain [40, 78] is the primary criterion herein adopted to score feature performance. In this application, information gain is simply the mutual information between the class variable C which partitions data by skill group and a given feature X.

$$I(X;C) = \sum_{x,c} p(x,c) \log \frac{p(x,c)}{p(x)p(c)} \tag{4.5}$$

Information Gain is applied in three different ways to establish the best possible set of candidate features for subsequent statistical models to use. These are described in the following sections. In all cases, the class variable C consists of the Ground Truth Expert (GT-Expert), FLS-Intermediate, and FLS-Novice sets.

4.2.1 Static Features

The computed overall information gain for all features appears in Table 4.3. This table effectively sums the information gain over both hands for all tasks. It suggests that position-based features provide the most discriminative information for skill level. This reflects favorably on position-based features, providing some evidence in support of the hypothesis that they contain skill-discriminating information. However, this success has a drawback. These position-based features tend to be more task-specific: they better discriminate skill at the expense of not generalizing as well to different tasks. For example, the tool tips spend the majority of task time in a different place or volume in the Peg Transfer task than in the Suturing task. An expanded version of this table appears in Appendix C as Table C.1. It details the minor differences in information gain across tasks and hands.

Table 4.3: The list of features and their scores, sorted by best total information gain (IG) cumulative over all tasks and both hands. The KL-divergence normalized by total entropy appears in parentheses. This is effectively a distance measure of average distance between distributions of the different classes. A value of 0 indicates almost no difference, suggesting poor discrimination.

Feature	**Overall IG**
Pos-dRta-Gr	0.1964 (0.04)
Vel-dRta-Gr-dQg	0.1029 (0.02)
Vel-dRta-Gr	0.0908 (0.02)
SpdAcc-dRta	0.0450 (0.01)
SpdAcc	0.0403 (0.01)
Gr	0.0388 (0.02)
SpdAccJrk	0.0325 (0.01)
dGr	0.0247 (0.01)
SpdAccJrkCrv-dRta-Gr	0.0247 (0.01)
dRta	0.0082 (0.00)
SpdCrv	0.0068 (0.00)

4.2.2 Sequential Features

The order of time samples is critically important in the surgical task. If the samples of a trajectory were randomly permuted and replayed, this would result in chaotic motion. Unfortunately, traditional IG will provide identical results under any permutation of the order of data samples. Moreover, traditional IG may completely overlook relevant sequential information. For example, suppose a hypothetical feature of cardinality 2 encodes expert and novice tool motion into the following sequences:

$$\begin{aligned} Seq_{Expert} &= AAAAAAAABBBBBBBB \\ Seq_{Novice} &= ABABABABABABABAB \end{aligned} \tag{4.6}$$

According to traditional IG, this feature provides no discrimination; the relative frequency of codewords is identical for both Expert and Novice sequence. However, the sequential information clearly discriminates between the two. Thus, the Information Gain (IG) criterion is modified to the conditional mutual information between the class and next sample, given

a present sample:

$$I(X_{t+1}; C|X_t) = \sum_{x_t} p(z) \sum_{x,c} p(x_{t+1}, c|x_t) \log \frac{p(x_{t+1}, c|x_t)}{p(x_{t+1}|x_t)p(c|x_t)} \tag{4.7}$$

and is herein called the sequential information gain (SIG). Unlike IG, SIG would indicate that the binary hypothetical feature suggested above strongly discriminates between the example skill classes.

The computed SIG for all features appears in table 4.4. As before, these are summed over all tasks and both hands. SIG recommends a different set of features as the best candidates for sequential analysis than the features recommended for static analysis by IG. This suggests that sequential models, like Hidden Markov models, should employ the former—specifically, SpdAcc-dRta—as input variables. This appears to support the hypothesis that scalar motion data can discriminate skill without need for vector direction.

Table 4.4: The list of sequentially grouped features and their scores, sorted by best total sequential information gain (SIG).

Feature	**Overall SIG**
SpdAcc-dRta	0.1756
SpdAcc	0.1302
SpdAccJrk	0.1276
dGr	0.0986
Vel-dRta-Gr-dQg	0.0492
SpdAccJrkCrv-dRta-Gr	0.0445
Vel-dRta-Gr	0.0442
SpdCrv	0.0346
dRta	0.0263
Gr	0.0026
Pos-dRta-Gr	0.0136

An expanded version of this table appears in Appendix C as Table C.2. It details the minor differences in SIG across tasks and hands.

4.2.3 Segmented Sequential Features

The goal of the segmentation schemes developed in Section 4.1.4 was to aid in discriminating skill by comparing tool motion in similar contexts. This should amplify the sequential information gain (SIG) for skill discrimination. SIG was calculated for each segmentation scheme. The resulting SIG values for all segmentation schemes appear in Table 4.5, where again, each entry is summed over all tasks for both hands. There are two criteria used to suggest the ideal candidate: highest SIG and lowest standard deviation. Since each SIG value in Table 4.5 is a sum of SIG values over all tasks and both hands, lower standard deviation will imply consistent discrimination across tasks and between hands. This is desirable for the universal applicability of skill metrics between different surgical tasks, different handedness, and possibly different modalities of surgery (e.g., robotic, manual, etc.).

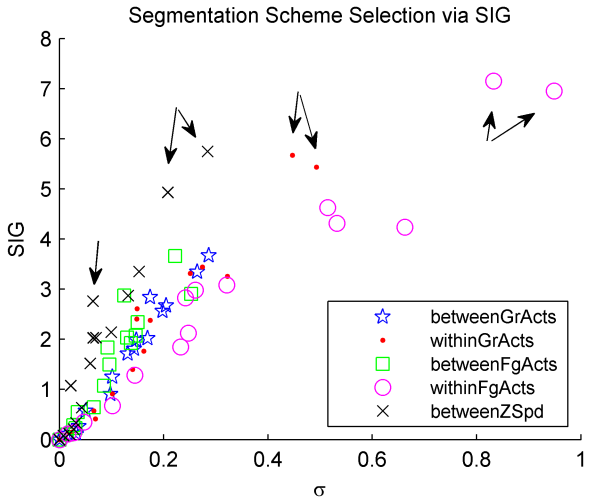


Figure 4.7: Segmentation scheme feature selection based on the sequential information gain (SIG) and the standard deviation σ among tasks and both hands for a given combination of segmentation scheme and extracted feature (VQ codebook).

The value of overall Sequential Information Gain vs. the standard deviation σ is plotted in Figure 4.7 to clarify how these two criteria interact. No single segmentation scheme exhibited highest overall SIG and lowest standard deviation simultaneously. Rather, Figure

Table 4.5: The list of segmented features and their sequential information gain (SIG) scores vs. segmentation scheme chosen. The standard deviation of the overall SIG across both hands for all tasks indicates how well the SIG value generalizes for all tasks and appears in parenthesis (*std*). The entries elected as based on high sequential information gain $I(Feat_{t+1}; Class|Feat_t)$ or combined low standard deviation are marked in bold. The lower portion of the table groups features in which the maximum codebook size of 150 was used (as opposed to the optimal codebook size discussed earlier in this chapter).

Feature ID	betweenGrActs	withinGrActs	betweenFgActs	withinFgActs	betweenZSpd
Gr	0.12 (0.03)	0.42 (0.07)	0.28 (0.03)	0.35 (0.05)	0.21 (0.03)
dGr	0.91 (0.10)	0.91 (0.10)	0.64 (0.07)	1.29 (0.14)	0.63 (0.04)
Pos-dRta-Gr	1.71 (0.13)	2.61 (0.15)	1.84 (0.09)	2.83 (0.24)	2.14 (0.10)
Vel-dRta-Gr	2.02 (0.17)	3.26 (0.32)	1.92 (0.14)	**4.24 (0.66)**	2.87 (0.13)
SpdAccJrk	1.81 (0.14)	1.76 (0.16)	1.50 (0.10)	2.12 (0.25)	1.52 (0.06)
Vel-dRta-Gr-dQg	1.99 (0.15)	**3.44 (0.27)**	2.34 (0.15)	4.32 (0.53)	3.35 (0.15)
SpdAccJrkCrv-dRta-Gr	0.57 (0.05)	0.57 (0.07)	0.55 (0.04)	0.67 (0.10)	0.34 (0.03)
SpdCrv	0.26 (0.04)	0.15 (0.02)	0.23 (0.03)	0.11 (0.02)	0.11 (0.02)
SpdAcc	1.25 (0.10)	1.40 (0.14)	1.07 (0.09)	1.85 (0.23)	1.07 (0.02)
SpdAcc-dRta	2.67 (0.20)	2.40 (0.15)	2.07 (0.15)	2.98 (0.26)	2.03 (0.06)
dRta	0.11 (0.01)	0.15 (0.02)	0.10 (0.01)	0.14 (0.03)	0.08 (0.01)
Vel-dRta-Gr-dQg150	3.35 (0.26)	5.67 (0.45)	3.66 (0.22)	**7.15 (0.83)**	**5.75 (0.28)**
Vel-dRta-Gr150	2.84 (0.17)	5.44 (0.49)	2.87 (0.12)	**6.95 (0.95)**	**4.93 (0.21)**
SpdAcc-dRta150	3.68 (0.29)	3.32 (0.25)	2.90 (0.25)	4.63 (0.51)	**2.76 (0.06)**
SpdAccJrk150	2.56 (0.20)	2.38 (0.17)	2.03 (0.13)	3.09 (0.32)	2.03 (0.07)

4.7 indicates a trade-off relationship between the two: SIG increases at the cost of greater standard deviation, suggesting that more discriminative segmentation schemes may rely on task-specific information to improve their performance.

The segmentation schemes were first grouped into those generated by optimal codebook size (see Section 4.1.3) and those generated by maximal codebook size. The latter are indicated by the lower portion of Table 4.5 and by the 150 mark in their labels. This was to enable a comparison of both approaches in modeling skill. The five columns in the table resolve into schemes derived from three categories: segmentation based on grasp variables (grasp angle and grasp force) as indicated by *GrActs*, those based on grasp force alone indicated by *FgActs*, and those derived solely from zero speed segments as indicated by *ZSpd*. Only two out of the three categories were selected for both the optimal and maximal codebook size groups. This was to minimize the risk of expending resources to pursue analysis that was less likely to yield skill-discriminant models but still enable the possibility of comparison within each group. By this criteria, for the optimal codebook size the *withinGrActs* segmentation scheme coupled with the *Vel-dRta-Gr-dQg* codebook and *withinFgActs* coupled with the *Vel-dRta-Gr* codebook proved most successful and *betweenZSpd* was rejected. For the maximal codebook size group, SIG values were substantially higher. The two highest-ranking categories were *withinFgActs* and *betweenZSpd* for segmentation schemes. This was true for both the *Vel-dRta-Gr-dQg150* and *Vel-dRta-Gr150* codebooks. To test the hypothesis that scalar motion variables alone like speed and acceleration magnitude can discriminate skill in a task-invariant way, the highest SIG-scoring combination of segmentation scheme and codebook was selected to analyze this hypothesis: *SpdAcc-dRta150* and *betweenZSpd*. In summary, the top two SIG-scoring features for *withinFgActs*, the top two for *withinGrActs*, the top two for *betweenZSpd*, and the highest-scoring feature with lowest standard deviation for *betweenZSpd* were elected for subsequent analysis. These entries, identified as ideal candidates for subsequent segmented model training, are indicated in bold in Table 4.5. Figure 4.7 indicates these points with arrows.

4.3 Conclusion

This chapter detailed the feature extraction and feature selection steps—the first two steps in the four step process undertaken to analyze and quantify surgical skill in this work. The feature extraction step indicated the space of input variables and consisted of selecting subsets of these input variables for analysis. This was based on a combination of hypotheses and an attempt to establish a point of reference with prior work to enable comparison of this work with prior art. The subsets of input variables were then normalized and fed into a vector quantization algorithm which generated up to 150 codebooks for each hand and task. The smallest codebook size which achieved a 0.1% drop in normalized distortion per codeword increase was adopted as the optimal codebook, and codebooks of the maximal size were also preserved for analysis. Finally, a framework for universal segmentation schemes was developed and three categories of particular segmentations were adopted: grasp variables (grasp angle and grasp force), grasp force, and zero speed based segmentation.

The feature selection section developed quantitative criteria based on information gain to determine which features and segmentation schemes best discriminate skill. The static information gain did not consider sequential information and indicated that position-based features discriminate skill the best under this criterion. However, the static information gain yielded relatively lower discrimination compared to similar analysis that leveraged sequential information in the data. Thus, it was not adopted for subsequent analysis. The sequential information gain (SIG) showed better performance and suggested that position-based features rank poorly, and that scalar motion-derived features such as those based on speed and acceleration appear to perform the best under the SIG criterion. Finally, SIG was combined with segmentation schemes to yield some of the highest information gains, suggesting that segmentation may indeed boost skill discrimination. However, this was shown to come at the cost of task independence, as higher SIG values for segmentation schemes showed greater deviation (less consistent performance) when applied across all tasks and both hands.

The features and their segmentation schemes were selected as follows. A whole-task

approach that did not segment data and was based on *Vel-dRta-Gr* and optimal codebook size was chosen as the closest feature to prior work by Rosen et al [90]. The remaining combinations are indicated in Table 4.5 and Figure 4.7. These selections were used in the following chapter for further sequential modeling and analysis of skill.

Chapter 5

SEQUENTIAL MODELS AND OVERALL EVALUATION

Surgery is foremost a sequential phenomenon. This work seeks to utilize this sequential information via tractable and scalable models that can potentially run in near real-time, concurrent with a surgical task. The goal of this chapter is to propose and evaluate sequential models for surgical skill—dynamic skill metrics—that meet these criteria. It completes the last two steps of the four-step algorithm development process reproduced in Fig. 5.1: Sequential Modeling and Overall Evaluation. Only the features identified as optimally discriminative of skill in the previous chapter's Feature Selection Section 4.2 are utilized for sequential models and subsequent evaluation.

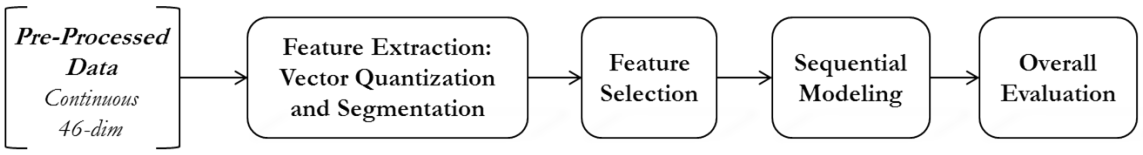


Figure 5.1: Overview of algorithm development for surgical skill evaluation based on time-sampled surgical tool motion data. (The same diagram is shown in Fig. 4.1, repeated here for convenience.)

This chapter first defines criteria used to determine whether or not a candidate sequential model succeeds. Then the Sequential Modeling section describes Hidden Markov Models (HMMs), variants of HMMs which improve skill evaluation, and the evaluation methods adopted to classify surgical data with these models. Finally, the Overall Evaluation section presents the results regarding how well the sequential models performed relative to the evaluation criteria.

5.1 Evaluation Criteria: Value Beyond a Stopwatch

Task time is perhaps the most widely used performance metric in surgery. It is inexpensive: a cheap stopwatch can suffice. It is universal: it applies to any surgical specialty, procedure, or modality (e.g., surgical robotics, laparoscopy, open surgery). However, it remains a cumulative summary metric and not a dynamic metric, providing little information or instructional feedback as to why or when a subject exhibited poor performance. Any new measure of skill which merely provides another summary result that correlates very highly with task time, then, is not necessary. Moreover, tool-motion derived metrics such as those explored in this work require equipment much more expensive than a simple stopwatch. If they provide no value over a stopwatch, then this added cost is unjustified. A novel metric can provide value over a stopwatch when it either discriminates skill in situations where a stopwatch cannot or if it is a dynamic metric capable of indicating when or why a subject performed poorly within a task. Ideally, it would meet both criteria.

Two methods were adopted to assess the extent that candidate skill metrics add information beyond simple task time.

- Normalization by $(1/T)$: For cumulative measures like total number of events or the sum of individual scores for each time-sample, the effective average value can be extracted to a constant by the total time. The two methods are:

- Constant-Time Window $[1 : T_w]$: Given only the first N-seconds of a task, (e.g. 45 seconds), what score does the skill metric report for that time window? A skill metric which provides value over task time should effectively discriminate skill between two constant-time windows from different attempts of the same procedure.

In both cases, task time would provide no discrimination of skill between any two attempts of a given procedure. These criteria are applicable to all surgical skill metrics, not just the sequential models developed here. Thus, they can also be used to evaluate and compare a number of alternative skill metrics: those presented in this work, future models, or skill metrics widely used in practice, for example, by commercially available surgical simulators.

5.2 Sequential Modeling

5.2.1 Hidden Markov Models

A hidden Markov Model (HMM) is a widely used statistical framework that models sequential phenomena [81]. It assumes the sequential phenomena can be sufficiently modeled by a Markov process (given a current sample, past and future are statistically independent) with a finite set of discrete hidden states that are not directly observable, but are inferable through a conditionally dependent observation variable. Fig. 5.2 provides a graphical model that illustrates these dependencies. The discrete hidden state at time t is Q_t and it is conditionally dependent on state Q_{t-1} in the immediate past. The observation made at time t denoted by X_t depends on the state Q_t. These relationships are repeated for all sample times. The prior success of HMM's in surgical skill recognition suggests that these modeling assumptions are acceptable for surgical applications [90, 85].

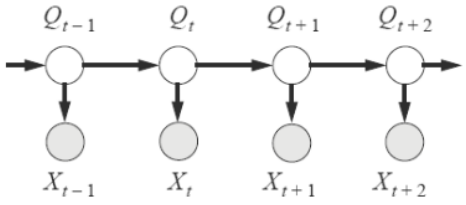


Figure 5.2: A graphical model view of a Hidden Marlov Model illustrating the dependencies between hidden states Q and visible observations X at different points in time t.

This work adopts Rabiner's naming convention in which a HMM, λ, is defined by three parameters: the state transition matrix A, the observation matrix B, and the initial state distribution π [81]. The number of states N, is a constant parameter of the model. It must be determined initially. The entries a_{ij} of A indicate the probability of transitioning from state i to state j, implying A is an $N \times N$ matrix. The number of possible observations, M, is also a pre-set constant. The entries b_{ij} of B indicate the probability of witnessing the observation labeled j while in state i, implying that B is an $N \times M$ matrix. The individual rows of A and of B must satisfy the standard stochastic constraints that their elements can only take on real-valued probabilities in the range [0 1] and each row of A and B must

sum to 1. A number of decisions must be made before a HMM can be applied. These include choosing the number of states N, deciding whether the observations are continuous or discrete, selecting the cardinality of observations M, and determining how real-world data are to be represented by observations via the random variable X_t. The next section details these choices for applying HMMs to skill evaluation via tool motion data.

5.2.2 HMM Variants Used in this Study

Alternative parameters and additional variations in applying HMMs to surgery beyond the technique of Rosen et al. are proposed with the goal of better skill discrimination. These variants were developed based on hypotheses derived from observations of the surgical tasks and attempts to improve on either the tractability or discriminative power of Rosen's methods. The first hypothesis was that models based on segmented data would provide more discriminative power: if segmentation schemes would create similar start and stop points for trajectories, this would allow a more favorable comparison of two trajectories by placing them in the same temporal context of each segment, rather than over the entire task. The shorter sequence lengths that result provide more tractable computation as well as a meaningful decomposition of the task which could provide feedback synchronized to the corresponding short segments of video in the task. The second hypothesis was that tool tip motion alone could discriminate skill based on scalar variables like speed and magnitude of acceleration. This was derived from observations that the quality of surgeons' motions during video review of their tool motion appeared to quickly indicate their level of technical proficiency. This quality of motion appeared to depend mostly on how surgeons accelerated and decelerated towards their targets and less so on the directional information provided by vector-valued counterparts like velocity. The benefit of such an approach, if proven successful, was reduced dimensional complexity and hypothesized task invariance. Both of these hypotheses gained support in the feature selection portion of Chapter 4. As multiple candidate features and segmentation schemes emerged from that work, this Chapter adopts a number of different models and features to establish which hypothesized approach provides best skill discrimination.

The model implementation of Rosen et al. [90] serves as a starting point to construct HMMs in this work. Rosen et al. used a Markov model of 15 states for each hand with discrete observations of 150 possible codewords from a vector quantization step. The observations were grouped into 15 sets of 10 codewords where each set mapped to a surgical "utterance" drawing on an analogy to speech recognition and segmentation. These sets were deterministically mapped to the states: there was zero probability that an observation in the i-th set mapped to a state other than i. The present work adopts the convention of discrete HMMs (as opposed to continuous), with one model for each hand as used by Rosen et al. In all models, vector quantization maps raw data into a sequence of 1-dimensional discrete integers (codewords) via codebooks suggested by the feature selection step. The cardinality of observations, M, is determined by the codebook size. Fifteen states for each hand serve as a starting point for analysis. All models analyzed in this work fall under the three categories described below.

Whole-Task vs. Segmented Models

The approach of Rosen et al. trained each model with entire iterations of a given task. We adopted the Whole-Task approach as a starting point: each model (one for each task and one for each hand) was trained with multiple iterations in their entirety. Thus, the initial state distribution, π, corresponds to the most likely state at the beginning of each task. An alternative method used the surgical task segmentation schemes suggested by Chapter 4, Section 4.1.4: *withinFgActs* and *betweenZSpd*. These selected segmentation schemes generated grasp-based motion segments and individual submovements respectively.

In the Segmented Models approach, each model (one for each task and each hand) was trained on all individual segments from multiple iterations of a given task. Thus, the initial state distribution, π, corresponds to the most likely state at the beginning of each individual segment. HMMs trained with the Whole-Task method are identified by a tag "WholeTask." Those trained with Segmented Models are identified by the segmentation type: "SegFg" refers to the *withinFgActs* segmentaton scheme and "SegSpd" implies the *betweenZSpd* segmentation scheme.

Optimal Features vs. 150-Codeword Features

All features used in this study consist of vector quantization codebooks which provide the discrete sequence of observations for HMMs. The choice of input variable subsets which serve as inputs to these codebooks were detailed in Section 4.1.2 and their normalization in Section 4.1.2. The process of codebook training, described in 4.1.3, resulted in hundreds of trained codebooks; Once the training process was complete for each choice of input variables, 150 codebooks were trained for each hand and each task. The number 150 was selected as the maximum codebook size based on the work of Rosen et al. to serve as a common point of comparison between that study and the current work. Since the number of input dimensions is decreased, it was expected that codebook sizes less than or equal to 150 should offer sufficient descriptive power relative to the work of Rosen et al.

Two criteria were used to select from among the hundreds of trained codebooks. The first was termed the optimal codebook size, also described in 4.1.3, and it was established by selecting the codebook size that generated the same distortion rate as the Rosen et al. work. Since both hands generated similar but often unequal optimal codebook sizes, the larger of the two was adopted for that set of input dimensions for a given task. The other criteria consisted of using the maximum codebook size, which provides less distortion at the cost of greater computational intensity and fewer average samples per codeword. Models which employ this maximum codebook size of 150 are denoted by the "-150" tag appearing after the model identifier. Models with the "-Opt" tag employed the optimal codebook size instead.

Dense vs. Sparse Left-to-Right Bakis Models

Two fundamental types of discrete Hidden Markov Models (HMMs) were considered: Dense HMMs and Bakis-type Left-to-Right HMMs. Dense HMMs are characterized by lack of sparsity (0 probabilities) in their state transition matrix A, observation matrix B, and initial state distribution probability π. This allows transitions from any one state to all other states and non-zero probabilities for observing any codeword in any state. Left-to-Right models differ in that their upper-triangular state transition matrix A disallows

transitions from state i to j if $j < i$. Such models are intended to characterize the time-evolution of individual, similar motion segments in a more intuitive way: all segments start at state 1 and all segments end at the last state (e.g., 15 for a 15-state model). The increasing progression through time samples is mirrored by the increasing progression of state numbers since—like time index—state number cannot decrease in Bakis models. As this differs substantially from the Rosen et al. method, the Bakis model state size was not preset to 15. In these models, the number of states is proportional to the average state duration as provided in [81, Section D]. Thus the mean length of observing the same symbol was computed. This was used to suggest the mean duration in a given state, and hence the expected number of states in a given model. The sparsity induced by Bakis models also proves less computationally intensive since computations with zero-probability factors need not be computed. Models are either denoted by "Dense." or "Bakis." tags to denote the different allowed state transition schemes.

5.2.3 Model Training

The different HMMs (model type and feature combination) explored in this study are summarized in Table 5.1. All models were trained on the ground truth Expert and Novice data sets separately to generate an expert and novice model of surgical skill, each consisting of one model for each hand. This was repeated for all three tasks. Thirty random initializations for each model were trained with the Baum-Welch/Expectation Maximization algorithm (tolerance: .0001, 250 iterations max). The resulting model which best fit the training data (according to highest log likelihood of the training data given the trained model) was adopted for subsequent analysis in the overall evaluation step.

5.2.4 HMM Skill Evaluation Methods

Once a Hidden Markov Model is trained, an evaluation method is required to quantify how well a sequence of observations from a given task iteration fits the model. The standard HMM computation of the probability of an observation given the model, $P(O|\lambda)$, decreases geometrically with task time (i.e., the number of samples) and typically results in a strong

Table 5.1: Overview and naming convention for all HMMs types and features employed in model training.

Model ID	States	Description	Codebook Size (Peg Transfer, Cutting, Suturing)
Dense.WholeTask(Vel-Opt)	15	Most similar to Rosen et al. approach, Vel-dRta-Gr codebook.	(56, 65, 70)
Dense.SegFg(VeldQg-Opt)	15	"Within Grasp" segmented HMM, Vel-dRta-Gr-dQg codebook.	(58, 67, 70)
Dense.SegFg(VeldQg-150)	15	"Within Grasp" segmented HMM, Vel-dRta-Gr-dQg codebook.	(150, 150, 150)
Dense.SegSpd(VeldQg-150)	15	Submovement segmented HMM, Vel-dRta-Gr-dQg codebook.	(150, 150, 150)
Dense.SegSpd(Spd-150)	15	Submovement segmented HMM, SpdAcc-dRta codebook.	(150, 150, 150)
Bakis.SegFg(VeldQg-Opt)	12	Left-to-Right "Within Grasp" segmented HMM, Vel-dRta-Gr-dQg codebook.	(58, 67, 70)
Bakis.SegFg(VeldQg-150)	25	Left-to-Right "Within Grasp" segmented HMM, Vel-dRta-Gr-dQg codebook.	(150, 150, 150)
Bakis.SegSpd(Spd-150)	49	Left-to-Right Submovement segmented HMM, SpdAcc-dRta codebook.	(150, 150, 150)

correlation with task time. This may not be amenable to surgical skill recognition; for example, ten minutes of an expert's observation sequence will yield a lower score than one minute of a poorly performing novice simply because there are more samples, even though the sequential data is consistently different between both observations. In this study, two HMM evaluation methods were adopted to cope with this issue. They are detailed below, along with a description of with how each method specifically controls for task time.

Symmetric Dissimilarity (Statistical Distance)

Rosen et al. adopted the approximation for the Kullbeck-Leibler (KL) divergence between HMMs provided by Rabiner [81]:

$$\begin{aligned} D(\lambda_1, \lambda_2) &= \frac{1}{T}[\log P(O^{(2)}|\lambda_1) - P(O^{(2)}|\lambda_2)] \\ &\neq D(\lambda_2, \lambda_1) \end{aligned} \tag{5.1}$$

Where $O^{(i)}$ is an observation of length T generated by the i-th model. For symmetric distance, this becomes:

$$\begin{aligned} D_s(\lambda_1, \lambda_2) &= \frac{1}{T} \log \frac{P(O^{(2)}|\lambda_1)P(O^{(1)}|\lambda_2)}{P(O^{(1)}|\lambda_1)P(O^{(2)}|\lambda_2)} \\ &= \frac{D(\lambda_1, \lambda_2) + D(\lambda_2, \lambda_1)}{2} \\ &= D_s(\lambda_2, \lambda_1) \end{aligned} \tag{5.2}$$

In both cases, the sequence length T must be chosen beforehand and should be long enough to sufficiently represent the data, e.g., the length of the longest iteration or of the training set. Rosen trained a HMM on expert data to represent expert skill and any new iteration to be scored was used to train a second HMM. The symmetric distance, D_s, between these two was adopted as the score of how "close to expert" a given iteration was. This successfully circumvents the strong correlation with task time inherent in a simple $P(O|\lambda)$ computation. However, this method cannot be computed in real time: the complete sequence is required to train the model the time required for model training is prohibitively long. Moreover,

this method only approximates KL-Divergence by generating random observations from the models. If the models are only slightly dissimilar, this approach may produce conflicting results based on pseudo-random number generation. Thus, it may not be well-suited to robustly measure small differences between models.

Log Likelihood Ratio (LLR)

This work proposes an evaluation approach that differs substantially from that of Rosen et al. In this approach, a HMM (containing individual models for left and right hands) is trained on expert data and another on novice data. The likelihood of how well an observation fits the expert model, $P(O|\lambda_{Exp})$ versus the novice model $P(O|\lambda_{Nov})$ can be established by a ratio of the two likelihoods. The logarithm (log) of the likelihood ratio, or Log Likelihood Ratio (LLR) then can be established for each task iteration as:

$$\begin{aligned} LLR(t, t_0) &= \log \prod_{k=t_0}^{k=t} \frac{P(O_k|\lambda_{Exp,L}, \lambda_{Exp,R})}{P(O_k|\lambda_{Nov,L}, \lambda_{Nov,R})} \\ &= \sum_{k=t_0}^{k=t} \log[P(O_k|\lambda_{Exp,L}, \lambda_{Exp,R})] - \log[P(O_k|\lambda_{Nov,L}, \lambda_{Nov,R})] \end{aligned} \tag{5.3}$$

Where,

$$\log[P(O|\lambda_L, \lambda_R)] = \log P(O|\lambda_L) + \log P(O|\lambda_R) \tag{5.4}$$

and t_0 indicates an initial time sample and t a sample at a later time. These two values select the corresponding range of samples from the observation sequence, $O_{t_0:k}$. When the Log Likelihood Ratio abbreviation is written without arguments as LLR, the entire observation sequence is assumed. The value of LLR indicates the following:

$$\begin{cases} LLR > 0 : \text{Expert more likely} \\ LLR < 0 : \text{Novice more likely} \end{cases}$$

The two evaluation criteria established in Section 5.1, Normalization and Constant Time

Window, are readily applied to the Log Likelihood Ratio. The Time-Normalized version of LLR, indicated as $LLR \cdot (1/T)$, is simply LLR divided by the total number of samples. The constant time-window version is indicated as $LLR[1 : T_w]$ where $t_0 = 1$, the first time sample, and $t = T$, the length of the constant time window, for equation 5.3. For this work, the value of the time window was adopted as 45 sec, the shortest recorded task iteration.

The LLR approach has several advantages over model dissimilarity and it also successfully circumvents the strong correlation with task time inherent in a simple $P(O|\lambda)$ computation. To evaluate (score) an iteration, it does not require additional model training. In fact, it only requires the linear-time computation of $P(O|\lambda)$ likelihoods which enables real-time computation on data as it is being generated. Its results are exact and repeatable, allowing more robust comparison for iterations statistically similar to an expert model. Finally, it can be used as a "dynamic metric," one that reports on the skill level at each instant or segment within a surgical task, as opposed to a summary metric which only provides cumulative information at the end. This is demonstrated in Figure 5.3, which shows the individual likelihoods from a range of observation samples extracted from an expert iteration relative to both expert and novice models.

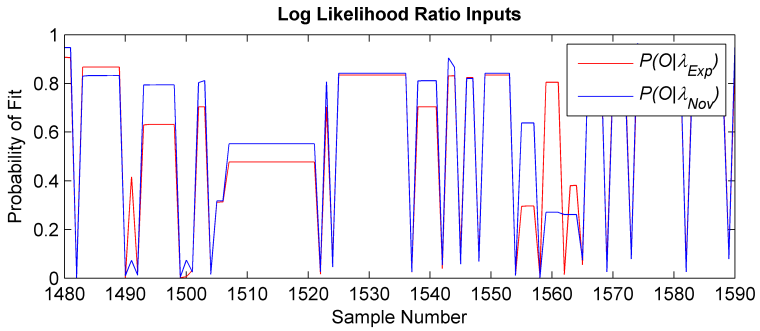


Figure 5.3: The time-varying likelihoods of a range of observation samples taken from an expert iteration to an expert model (red) and a novice model (blue) for a time range of approximately 3.5 sec. These are the two inputs to the Log Likelihood Ratio (LLR). When the probability of expert fit is better for a given sample than the novice fit, $LLR > 0$ and vice versa. This illustrates a truly dynamic skill metric: one that measures skill at each instant within a task.

5.3 Overall Evaluation and Results

The overall evaluation step attempts to establish whether any candidate skill metric successfully discriminates surgical skill and simultaneously provides value over a simple stopwatch. Satisfying this step establishes the construct validity of any time-controlled skill metric: the extent to which it discriminates between carefully established levels of surgical expertise when task time cannot. For the HMM candidates, our aim is to establish what choice of features, model type, and model parameters best discriminate skill as well as what evaluation method works best. The following sections report the overall evaluation of each candidate skill metric.

5.3.1 Rosen et al. Model and Symmetric Dissimilarity Evaluation

To analyze how well the modeling, features, and evaluation method of Rosen et al. work with this data set, a Symmetric Dissimilarity analysis was carried out with the DenseWholeTask (Vel-Opt) HMM. In addition to the expert HMM trained from GT-Expert data, an additional HMM was trained for each task iteration with the same procedure as training all other models. Symmetric distance Ds was computed with $T = 2000$ samples for all models in the ground truth data set. This value generates observations approximately 67 seconds in length and was found to provide sufficient data to generate consistent results. Longer sequence lengths would generate slightly more robust (repeatable) comparisons but take longer to process.

The results appear in Figure 5.4. Expert iterations should gravitate to the top left of the plot, near the expert group but distant from the novice group. Novices should appear near the bottom right. Instead, there is considerable overlap and some novice iterations appear closer to the desired expert area than actual expert iterations. It appears the novice model is better at discriminating skill than the expert model. Overall, the approach of Rosen et al. does not work well in this data set for any task.

The asymmetry between the classification power of the novice vs. the expert models suggests that the training methodology, selected model topology, or selected features are not ideal. For example, if there is some similarity between novices and experts, training a model

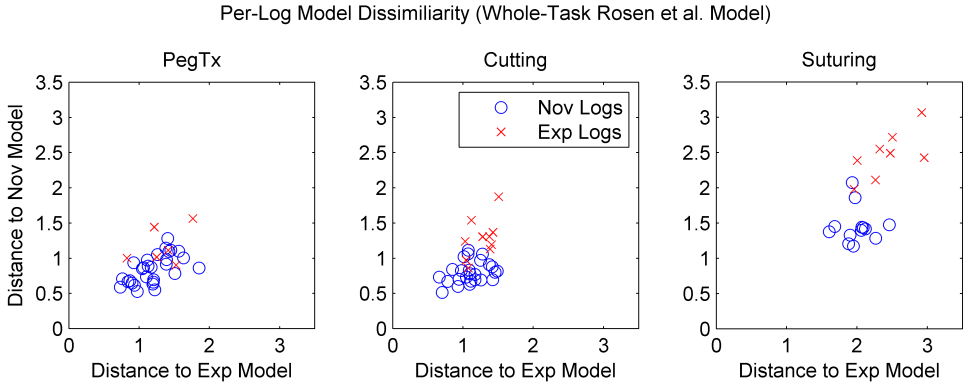


Figure 5.4: Statistical distance between each per-log trained model (one model per one iteration) and models trained with data from expert (GT-Exp) and novice (FLS-Nov) group data.

on only the expert data and another on only the novice data provides no guarantee that these two models will exhibit strong discriminating power between the two groups. Rather, if a discriminative training criterion like Maximum Mutual Information was optimized at each EM step, this would result in more discriminative models, provided that the features and sequential model provided enough information gain in the first place.

5.3.2 Log Likelihood Ratio Evaluation For All HMM Variants

The method of time-controlled Log Likelihood Ratio (LLR) evaluation was adopted for all candidate HMMs appearing in table 5.1. The results of skill discrimination between the Ground Truth Expert set (GT-Exp) and FLS Novice set (FLS-Nov) appear in Figures 5.5-5.10. The Time-Normalized version of LLR is indicated as "$LLR(1/T)$" whereas the constant time-window version is indicated as "$LLR[1 : T_w]$" and used the first 45 seconds of all iterations. All results are summarized in Table 5.2.

Only the Dense.SegFg(VeldQg-Opt) models provide zero misclassification for both methods of controlling for time (normalization and constant time window), but only for the Peg Transfer and Cutting Tasks (see Figure 5.5). The Dense.WholeTask(Vel-Opt) also has a zero misclassification rate for the Cutting task using Time Normalization ($1/T$) but not when using the constant time window $[1 : T_w]$. This model also appears to provide the

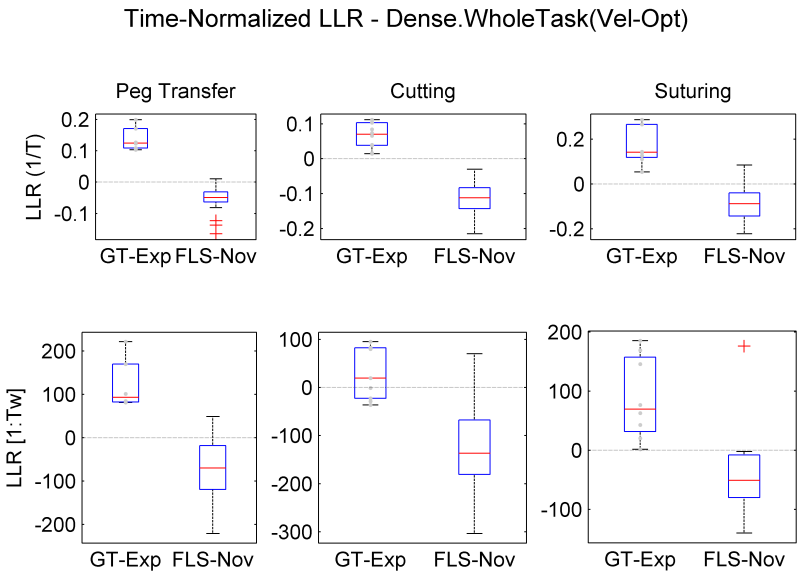


Figure 5.5: Time-Controlled Log Likelihood Ratio evaluation for Dense.WholeTask(Vel-Opt) HMM, the model most similar to that of Rosen et al.

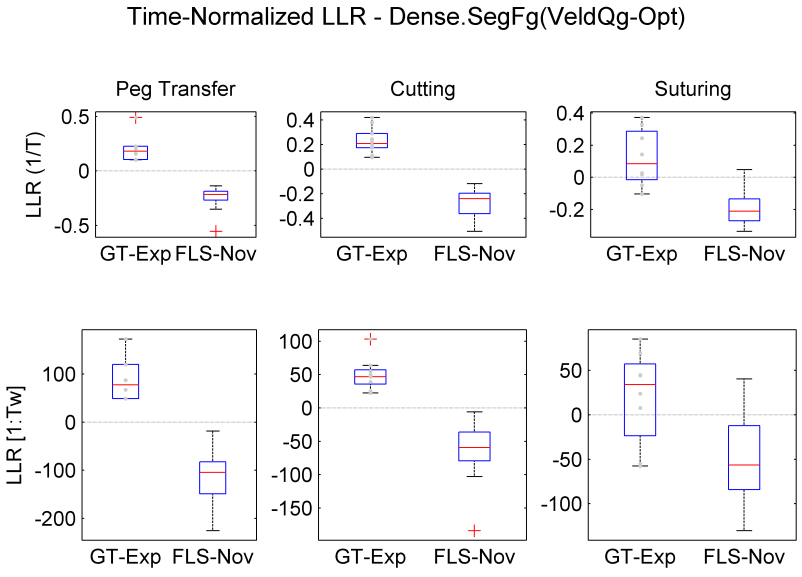


Figure 5.6: Time-Controlled Log Likelihood Ratio evaluation for Dense.SegFg(VeldQg-Opt).

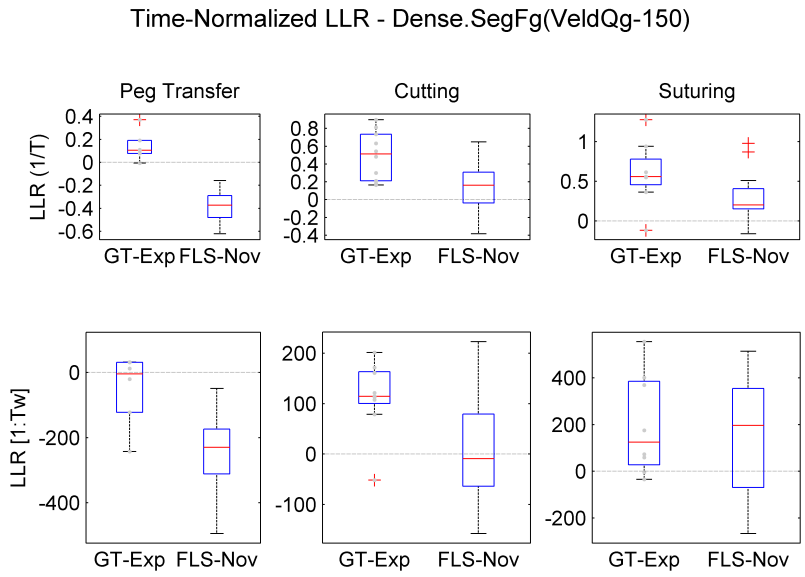


Figure 5.7: Time-Controlled Log Likelihood Ratio evaluation for Dense.SegFg(VeldQg-150).

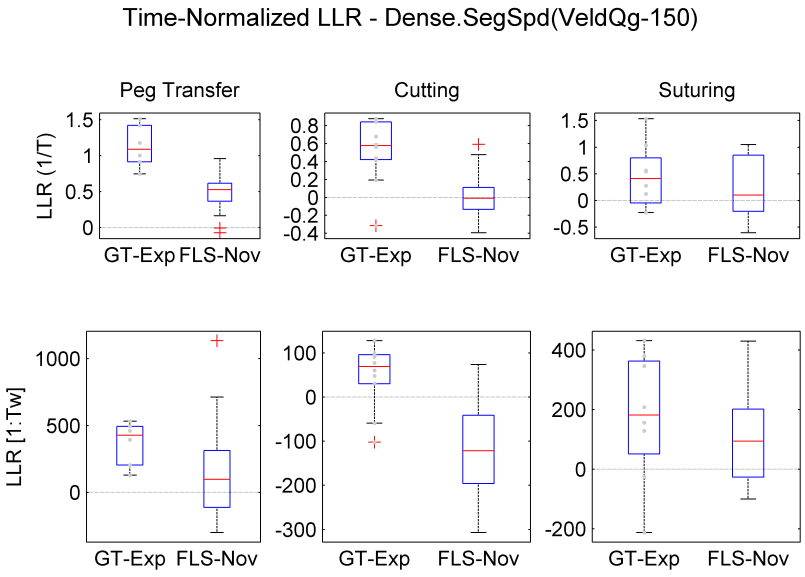


Figure 5.8: Time-Controlled Log Likelihood Ratio evaluation for Dense.SegSpd(VeldQg-150).

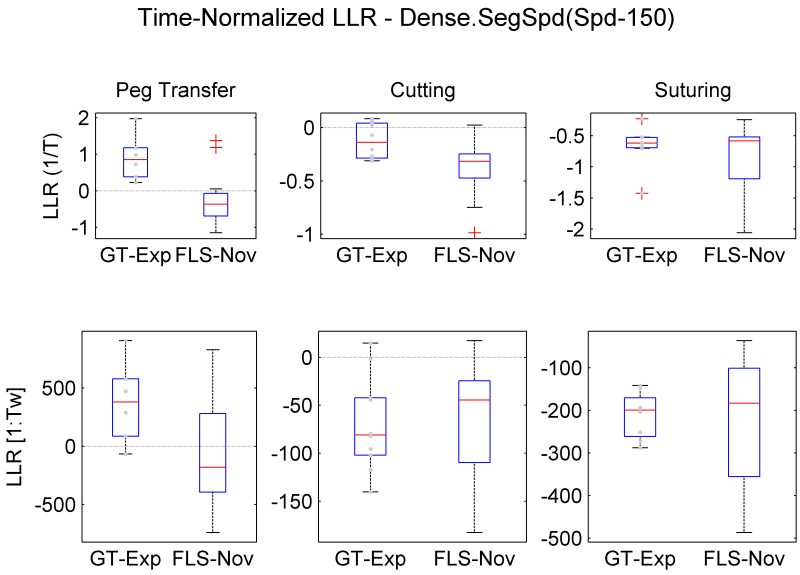


Figure 5.9: Time-Controlled Log Likelihood Ratio evaluation for Dense.SegSpd(Spd-150).

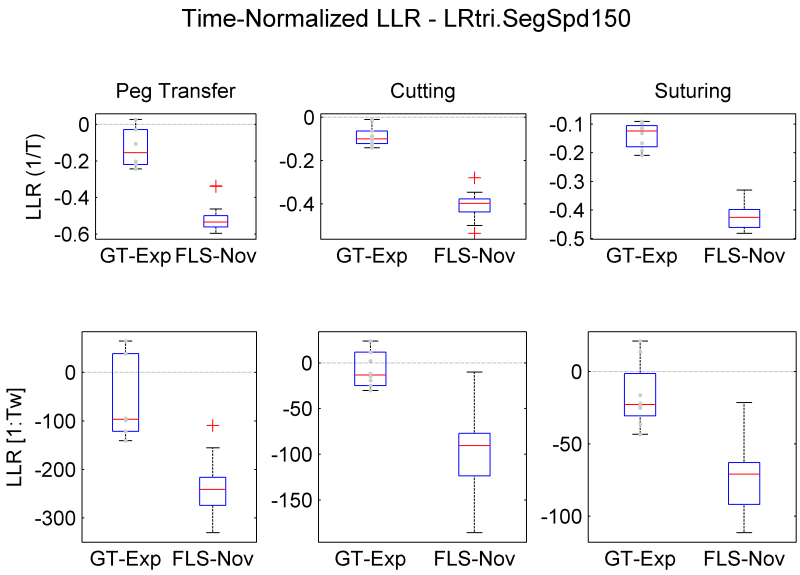


Figure 5.10: Time-Controlled Log Likelihood Ratio evaluation for Bakis.SegSpd(Spd-150).

Table 5.2: Summary of time-normalized Log Likelihood Ratio HMM results for discrimination of the Ground Truth Expert and FLS Novice groups. Entries where any misclassification occurs are marked with "−", once for each group that exhibits any misclassification. Entries marked with $*$ exhibit statistically significant ($p < 0.05$) differences between means and the difference is provided in parentheses. Larger differences imply better separation.

	PegTx	Cutting	Suturing
Dense.WholeTask(Vel-Opt)	−	(0.18)*	−
Dense.SegFg(VeldQg-Opt)	(0.45)*	(0.50)*	−−
Dense.SegFg(VedQgl-150)	−	−	−−
Dense.SegSpd(VedQg-150)	−	−−	−−
Dense.SegSpd(Spd-150)	−	−−	−
Bakis.SegFg(VeldQg-Opt)	−	−	−
Bakis.SegFg(VeldQg-150)	−	−	−
Bakis.SegSpd(Spd-150)	−	−	−

best discrimination for the Suturing task (Figure 5.5), but there is a single iteration from the FLS-Novice set that is incorrectly classified as an expert. Ideally, no misclassification should occur, especially since the evaluated iterations were used to train the models. Only these two model types—Dense.SegFg(VeldQg-Opt) and Dense.WholeTask(Vel-Opt)—provided zero misclassification on the training set. Thus only these two models were used in subsequent analysis.

Table 5.3: The layout, description, and computed values of a confusion matrix. All derived values are indicated in green.

		Ground Truth: "Is expert"		
		True	False	
Classifier Results: "Is expert"	True	True Positive (TP)	False Positive (FP)	Positive Predictive Value (PPV) = TP / (TP + FP)
	False	False Negative (FN)	True Negative (TN)	Negative Predictive Value (NPV) = TN / (TN + FN)
		Sensitivity = TP / (TP+FN)	Specificity = TN / (FP+TN)	Accuracy = (TP+TN) / (TP+FP+FN+TN)

A confusion matrix quantifies the performance of a binary classifier. Table 5.3 demonstrates the components, terminology, and derived values of a confusion matrix. The confusion matrices for the two models yielding fewest misclassifications appear in Figures 5.11 and 5.12. This scores the performance of these models when using the Log Likelihood Ratio evaluation approach with both methods of time normalization for each of the three tasks. Ideally, there would be zero false positives and zero false negatives. In this work, false positives can be more deleterious than false negatives. A false positive would imply an incompetent surgeon may be misclassified as an expert, potentially exposing patients to higher risk of iatrogenesis if allowed to operate based on this test. A false negative would misclassify an expert surgeon as non-competent. While it may place an added burden on proficient faculty or result in unnecessary additional training for them, this avoids putting patients at risk. Thus, if two competing models have the same total of false positives and false negatives, the model with fewer or zero false positives is elected as the better model.

The confusion matrix provides multiple numbers that emphasize different attributes of binary classification, but a single number is preferred for overall comparison between candidate models. Measures like positive predictive value, negative predictive value, and accuracy may prove misleading as overall measures when the underlying class sample sizes substantially differ. For example, there are relatively few "true expert" iterations available in this work for each task compared to the number of FLS novices. According to the definitions provided in the last column of Table 5.3, these measures are skewed by the prevalence of novices over experts in the input data. Matthews correlation coefficient (MCC) is adopted to overcome such issues and provide an acceptable single number-measure of performance for this work [7]. The MCC value is computed directly from the confusion matrix and is closely related to Pearson's χ statistic. Both the equation for MCC and its relationship to χ are provided in Equation 5.5 below.

$$\begin{aligned} MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ &= sign(MCC) \times \sqrt{\frac{\chi^2}{N}} \end{aligned} \tag{5.5}$$

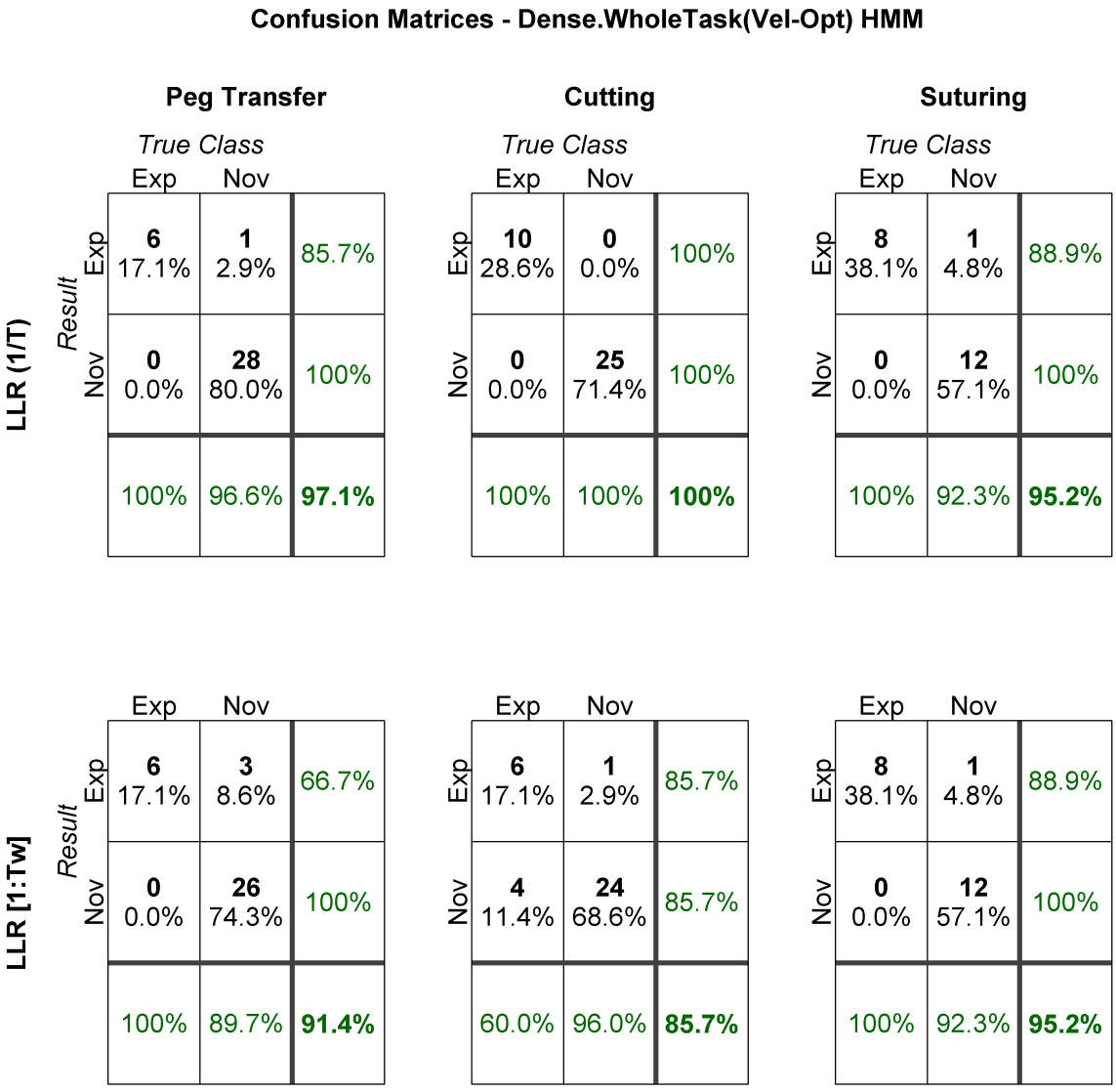


Figure 5.11: The confusion matrices of the Dense.WholeTask(Vel-Opt) HMM, the model most similar to that of Rosen et al. It uses no segmentation (whole task) and optimal VQ codebook size. The top row controls for task time via the $(1/T)$ time normalization and the bottom row via the $[1 : T_w]$ method, the first 45 seconds of each iteration.

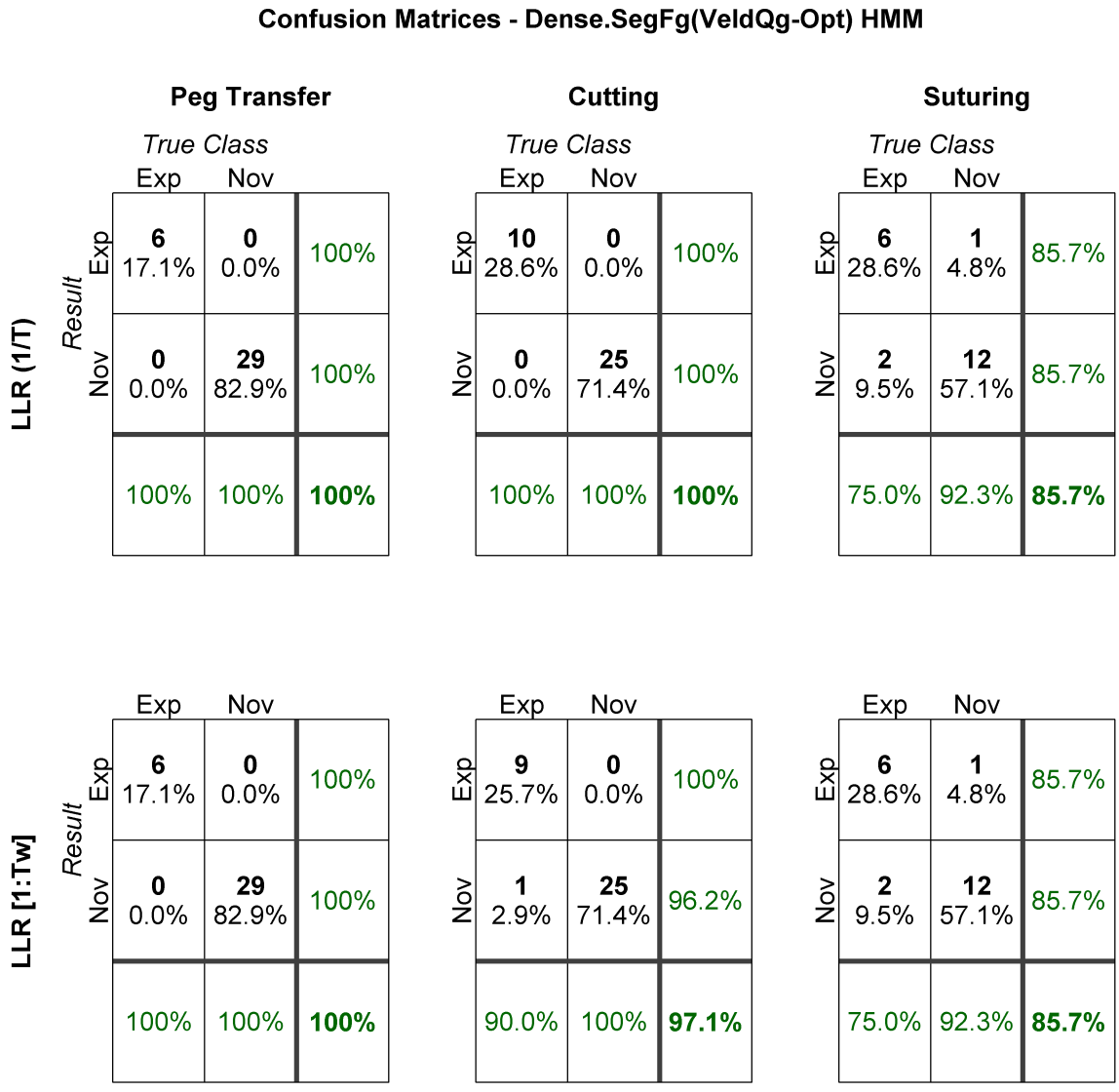


Figure 5.12: The confusion matrices of the Dense.SegFg(VeldQg-Opt) HMM. This model uses grasp force-based segmentation and optimal VQ codebook size. The top row controls for task time via the $(1/T)$ time normalization and the bottom row via the $[1 : T_w]$ method, the first 45 seconds of each iteration.

where N is the total sample size. The overall comparison of MCC values is presented in Table 5.4. This table sums the confusion matrix values of both time normalization methods for each task. The MCC for the overall sum of confusion matrices over all tasks appears in the last column. All models that exhibited zero MCC values in one or more columns were excluded. According to this table, the Dense.SegFg(VeldQg-Opt) model has best overall performance as well as showing the highest MCC values for the Peg Transfer and Cutting Tasks. The Dense.WholeTask(Vel-Opt) ranks second in overall performance but provides the best results for the Suturing task.

Table 5.4: Overall Mathews correlation coefficient results of HMM training set classification. Only HMMs with no zero entries are shown.

	Peg Transfer	Cutting	Suturing	All Tasks
Dense.WholeTask(Vel-Opt)	0.84	0.82	0.91	0.85
Dense.SegFg(VeldQg-Opt)	1.00	0.97	0.69	0.90
Dense.SegFg(VeldQg-150)	0.79	0.29	0.09	0.39
Dense.SegSpd(VeldQg-150)	0.21	0.48	0.10	0.25
LRtri.SegSpd150	0.47	0.39	0.29	0.38

Cross Validation

A simple cross-validation step was performed in order to determine how well the results of the previous section can generalize to new data that is not part of the training set. An evaluation set of iterations representing each skill level containing data not used in the training set was classified by the same LLR evaluation method. Specifically, the Ground Truth Experts Plus (GT-Exp+) and FLS-Experts (FLS-Exp) sets were used to see how well evaluation could be generalized beyond the GT-Expert set used in training for the case of proficient subjects. FLS-Intermediate (FLS-Int) iterations were never used during training, but their LLR evaluation scores were expected to fall between the Expert and Novice Sets. Only the models which provided total separation within the training set (between the GT-Experts set and the training set remainder) for at least one task were considered for cross validation. The results appear in Fig. 5.13-5.14.

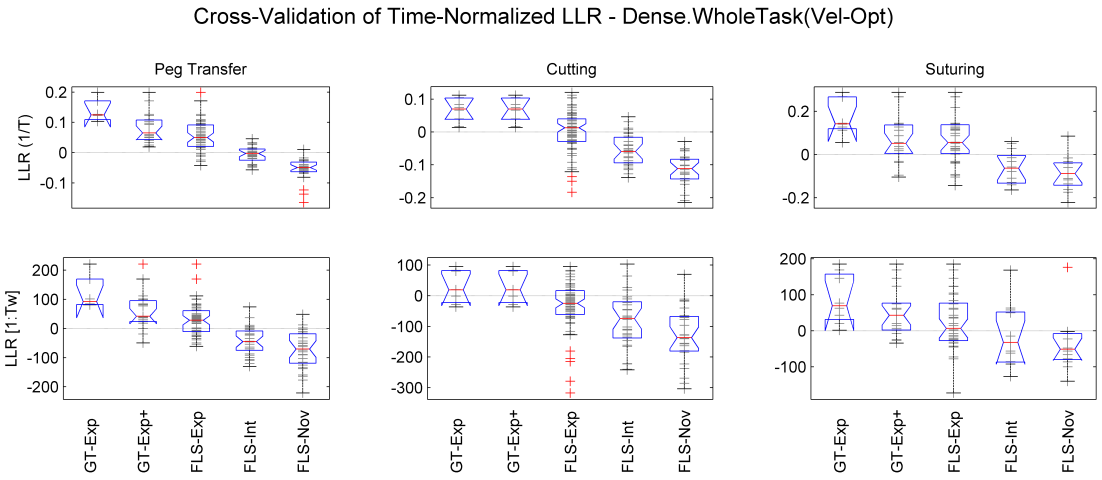


Figure 5.13: Cross-Validation for the Dense.WholeTask(Vel-Opt) HMM, the model most similar to that of Rosen et al. The middle three categories in each plot resulted from iterations not present in the training set.

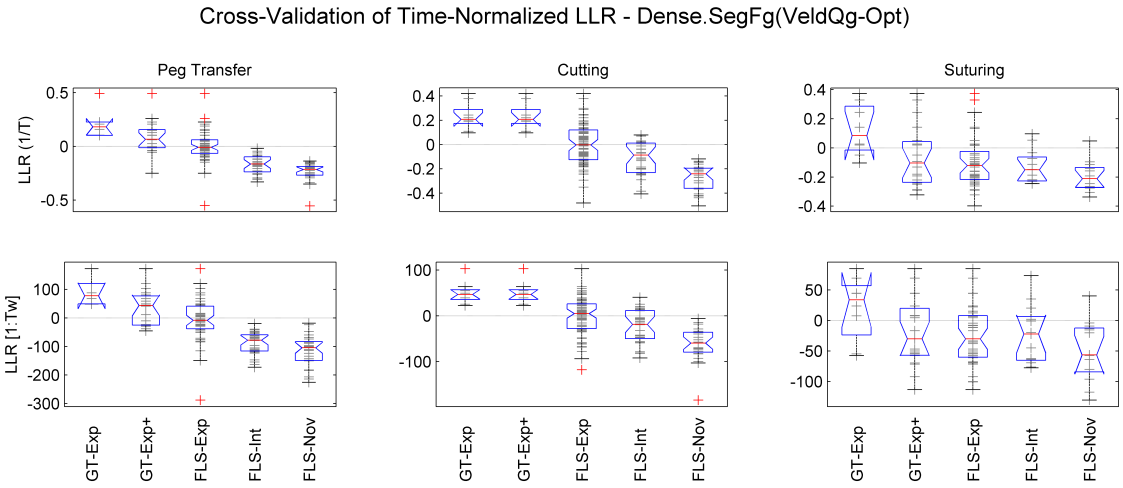


Figure 5.14: Cross-Validation data for the Dense.SegFg(VeldQg-Opt) HMM. The middle three categories in each plot resulted from iterations not present in the training set.

A one-way analysis of variance (ANOVA) was performed to determine whether the means of classified skill groups significantly differed at the $\alpha < 0.01$ significance level. Table 5.5 shows the results of the one-way ANOVA. All F-tests for both models indicate significant difference among means for both time normalization methods across all three tasks.

Table 5.5: One-way ANOVA results for cross validation of the two best HMM classifiers of the previous section. The value of the F-statistic is shown with corresponding p-value in parenthesis. All p-values are well below the 0.01 significance threshold.

	Normalization	Peg Transfer	Cutting	Suturing
Dense.WholeTask(Vel-Opt)	LLR (1/T)	35.8 (4.6e-020)	35.4 (2.3e-020)	4.89 (1.3e-003)
	LLR [1:Tw]	35.1 (8.9e-020)	25.3 (1.2e-015)	2.32 (6.3e-002)
Dense.SegFg(VeldQg-Opt)	LLR (1/T)	35.8 (4.6e-020)	35.4 (2.3e-020)	4.89 (1.3e-003)
	LLR [1:Tw]	35.1 (8.9e-020)	25.3 (1.2e-015)	2.32 (6.3e-002)

Following the positive results of Table 5.5, a multiple comparison test was used to determine which means were significantly different. The Tukey-Kramer method was used to test for pairwise differences while addressing the familywise error rate to avoid increasing the likelihood of false discoveries due to multiple comparisons. The Ground Truth Experts set served as the basis for all pairwise comparisons. Figures 5.15 and 5.16 show the multiple comparison results for cross validation of the two best-performing models on the training set from the previous section. Each figure shows the two methods of controlling for task time for all three tasks. The Ground Truth Experts group is marked in blue and the dotted lines around its whiskers indicate the range which determines whether the mean of an adjacent group is significantly different or not at the ($p < 0.05$) level. Red indicates a significant difference between means, gray does not. The whisker segments about each datum indicate the allowable range for which no overlap implies a significant pairwise difference. All evaluations indicate a significant difference from the Ground Truth Expert set to the FLS Intermediate and FLS Novice sets with the exception of the Suturing task. There is no significant difference between the Ground Truth Experts set (training set) and the Ground Truth Experts Plus set (evaluation set). However, there was a significant difference between the Ground Truth Experts and the FLS Experts in all cases except in the Suturing task

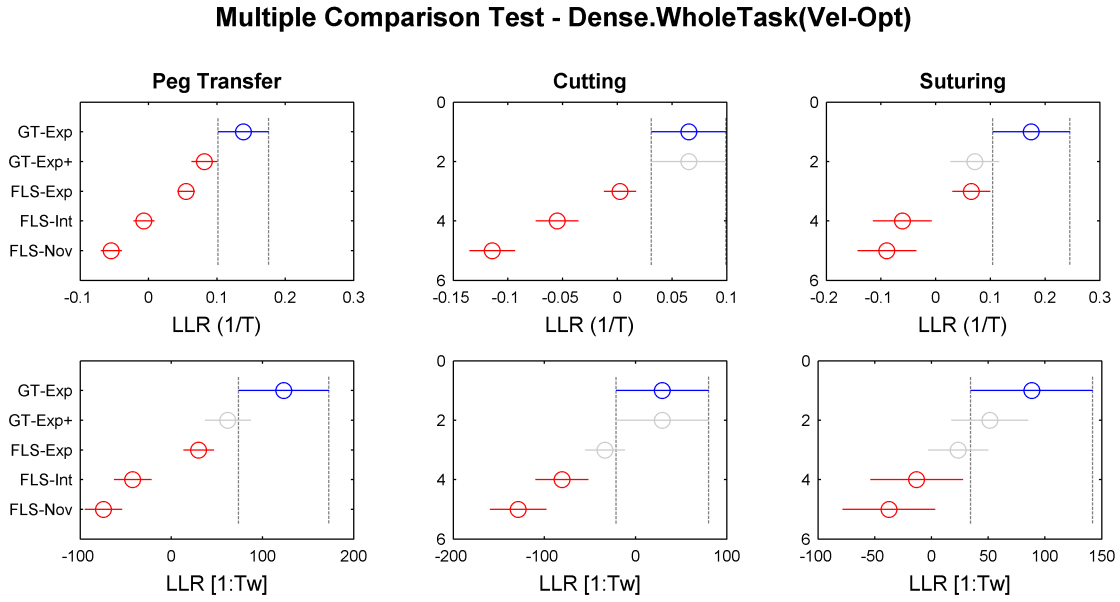


Figure 5.15: Multiple comparison test of Dense.WholeTask(Vel-Opt) for cross validation. The Ground Truth Expert group (blue) was used as the basis for pairwise comparisons. Red indicates a significant difference between means, gray indicates the lack of a significant difference.

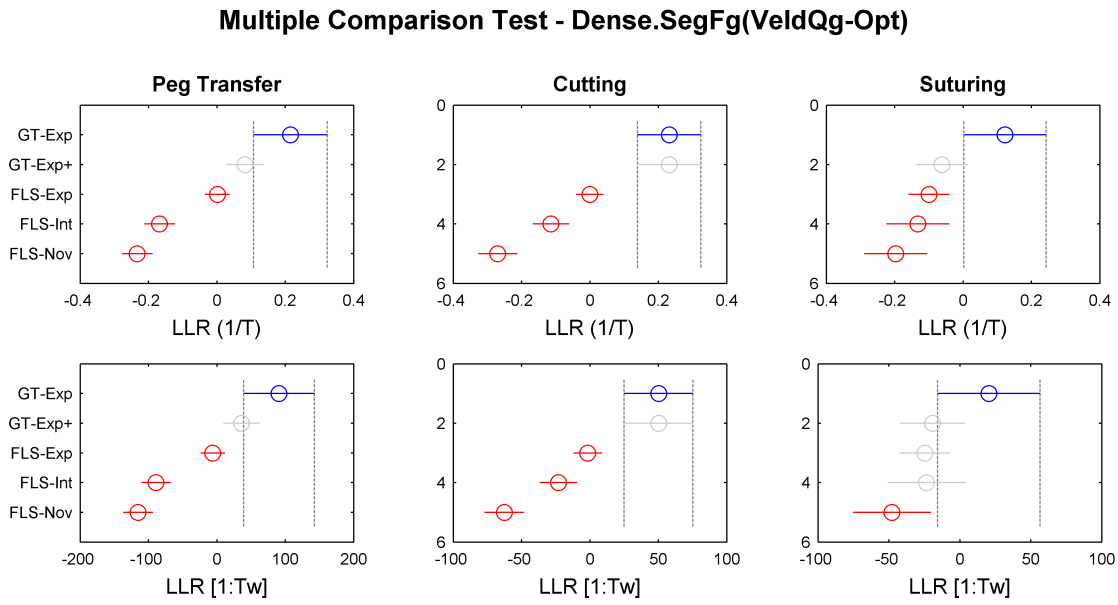


Figure 5.16: Multiple comparison test of Dense.SegFg(VeldQg-Opt) for cross validation. The Ground Truth Expert group (blue) was used as the basis for pairwise comparisons. Red indicates a significant difference between means, gray indicates the lack of a significant difference.

with the constant time window $T[1 : T_w]$ time normalization method.

5.3.3 Alternative Metrics of Skill

The Sequential Modeling step of the previous section used resource-intensive statistical modeling to compute skill level based on sequential data. However, some novel deterministic metrics of skill can be more easily computed based on the same input data. Specifically, counts of movements or events based on the segmentation schemes provide an easily computable metric that has some precedent in both the surgical and motor learning reviewed literature (see Chapter 1). We used the same criteria of establishing value over task time (see Section 5.1) to evaluate these segment counts and FLS scoring below.

Event Counts

The Segmentation Schemes described in Section 4.1.4 and presented in Table 4.5 offer two simple ways to count the number of events: generalized "grasp events" with segmentation based on grasp force (SegFg), or submovements with segmentation based on zero-speed events (SegZSpd). The combination of these two counts may provide a more robust means of discriminating skill than either count alone. Fig. 5.17 shows a time-normalized grasp count which effectively discriminates skill for Peg Transfer: a task that should have only 12 total grasps per hand when no drop errors are made. It does not perform so well when applied to less artificial, more surgically relevant tasks like cutting or Suturing. Grasp and Submovment counts were combined by direct multiplication and two different time normalization schemes were adopted: the standard $(1/T)$ scaling and $(1/T^2)$ to ensure time-independence via the product of two time-normalized quantities. These results appear in Fig. 5.18 and 5.19.

A one-way analysis of variance (ANOVA) was performed to determine whether the means of classified skill groups significantly differed at the $\alpha < 0.01$ significance level. Table 5.6 shows the results of the one-way ANOVA. All F-tests for the right hand and both hands indicate significant difference among means. The right hand shows the strongest difference in means. The left hand indicates either less significant differences or no differences among

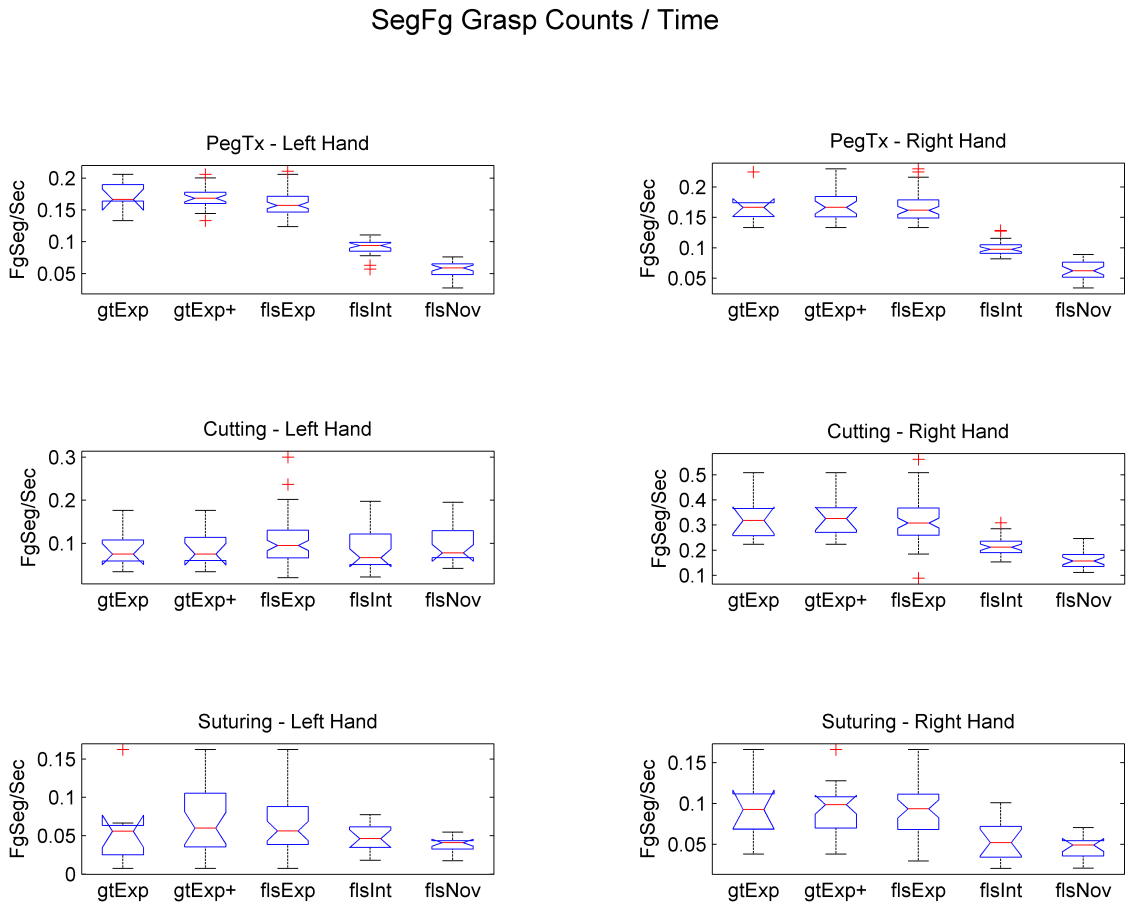


Figure 5.17: Time-controlled segment count for number of grasps per iteration for both left and right hands derived from SegFg segmentation scheme. Time normalization provided by $(1/T)$.

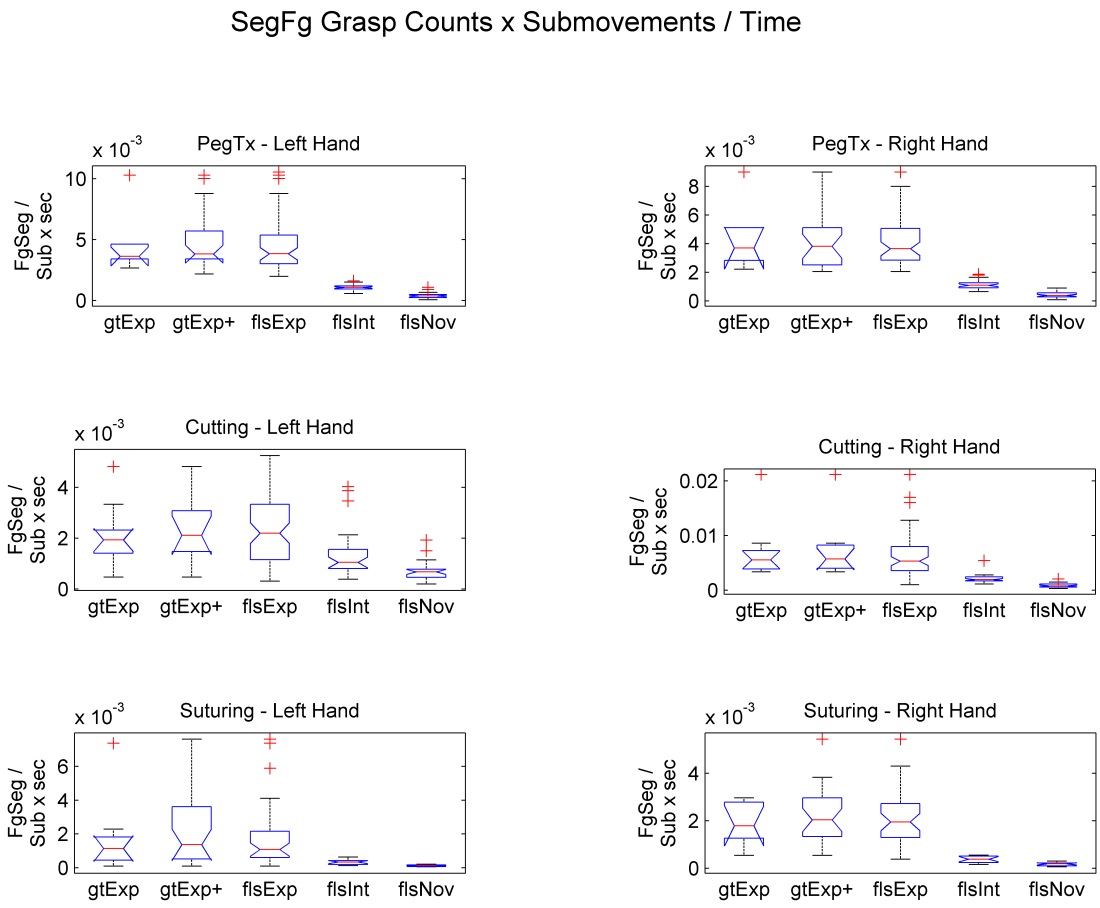


Figure 5.18: Time-controlled combined movement count for number of grasps times number of submovements per iteration for both left and right hands derived from SegFg and SegZSpd segmentation schemes. Time normalization provided by $(1/T)$.

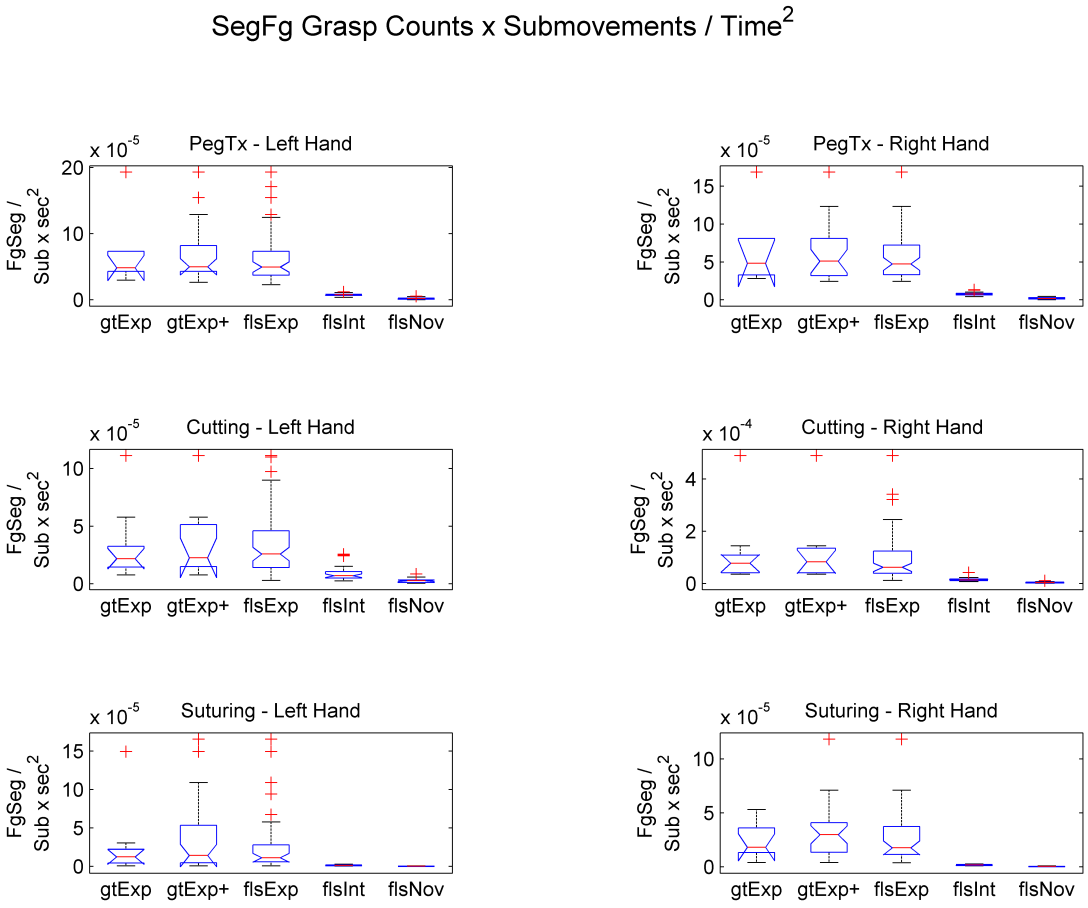


Figure 5.19: Time-controlled combined movement count for number of grasps times number of submovements per iteration for both left and right hands derived from SegFg and SegZSpd segmentation schemes. Time normalization provided by $(1/T^2)$.

Table 5.6: One-way ANOVA results for cross validation of event count-based skill metrics. The value of the F-statistic is shown with corresponding p-value in parentheses.

Metric	Task	Left Hand	Right Hand	Both Hands
FgSeg / Time	Peg Transfer	270 (9e-063)	149 (3.8e-048)	383 (3.1e-110)
	Cutting	0.635 (0.64)	31.8 (7.6e-019)	6.73 (3.5e-005)
	Suturing	2.62 (0.04)	9.94 (9.7e-007)	9.35 (6.4e-007)
FgSeg x SubMv / Time	Peg Transfer	10.7 (1.5e-007)	14 (1.4e-009)	23.3 (1.5e-016)
	Cutting	14.6 (5.1e-010)	21.9 (4.6e-014)	18.2 (2.5e-013)
	Suturing	40.2 (1.7e-019)	20.3 (4.7e-012)	51.1 (3.3e-029)
FgSeg x SubMv / Time^2	Peg Transfer	48.2 (5.5e-025)	57.3 (5.1e-028)	99 (7.9e-052)
	Cutting	0.486 (0.75)	19.9 (5.7e-013)	5.68 (0.00021)
	Suturing	1.1 (0.36)	5.25 (0.00075)	4.54 (0.0016)

means, particularly for the Cutting and Suturing tasks.

A multiple comparison test was used to determine which means were significantly different. The Tukey-Kramer method was again used to test for pairwise differences The Ground Truth Experts set served as the basis for all pairwise comparisons. Figures 5.20 – 5.22 show the multiple comparison results for cross validation for the event count-based metrics. The Ground Truth Experts group is marked in blue and the dotted lines around its whiskers indicate the range which determines whether the mean of an adjacent group is significantly different or not at the ($p < 0.05$) level. Red indicates a significant difference between means, gray does not. The whiskers of each point indicate the allowable range for which no overlap implies a significant pairwise difference. All right hand evaluations indicate a significant difference from the Ground Truth Expert set to the FLS Novice sets, even for the Suturing task. There is no significant difference between the Ground Truth Experts set (training set) and the Ground Truth Experts Plus set or FLS-Experts set (evaluation sets). Finally, only the right hand evaluation of the FgSeg/Time metric provides significant separation of means across all tasks for all three skill categories: Expert, Intermediate, and Novice. It also shows no significant variation within the Expert category (comprised of the GT-Expert, GT-Expert+, and FLS-Expert sets). Similarly, the two event count metrics based

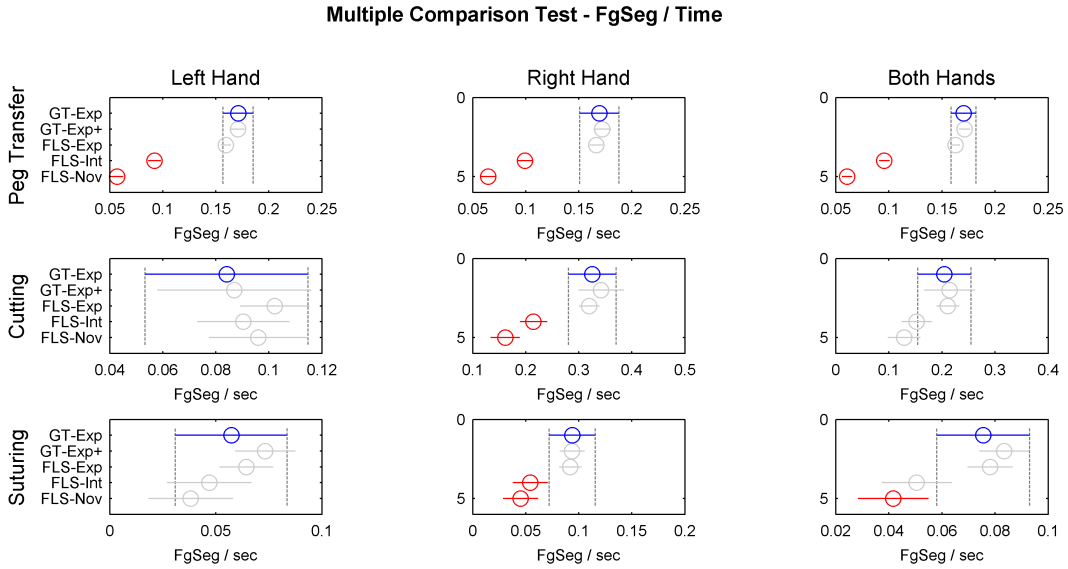


Figure 5.20: Multiple comparison test for cross validation of the grasp force-based segmentation rate (FgSeg/Time). The Ground Truth Expert group (blue) was used as the basis for pairwise comparisons. Red indicates a significant difference between means, gray indicates the lack of a significant difference.

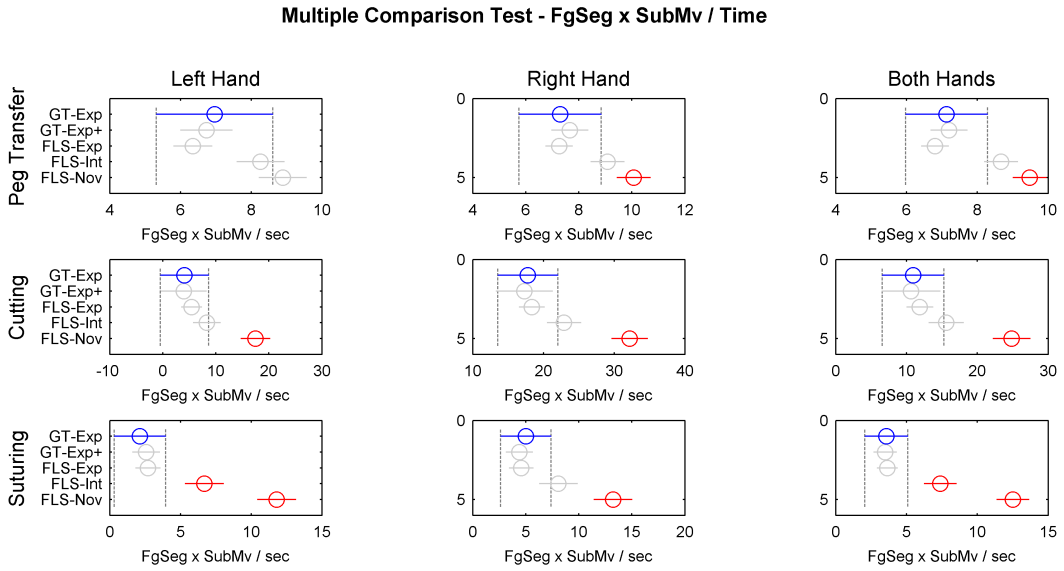


Figure 5.21: Multiple comparison test for cross validation of the product of grasp force-based segmentation (FgSeg) and submovement count (SubMv) normalized for time by (1/T). The Ground Truth Expert group (blue) was used as the basis for pairwise comparisons. Red indicates a significant difference between means, gray indicates the lack of a significant difference.

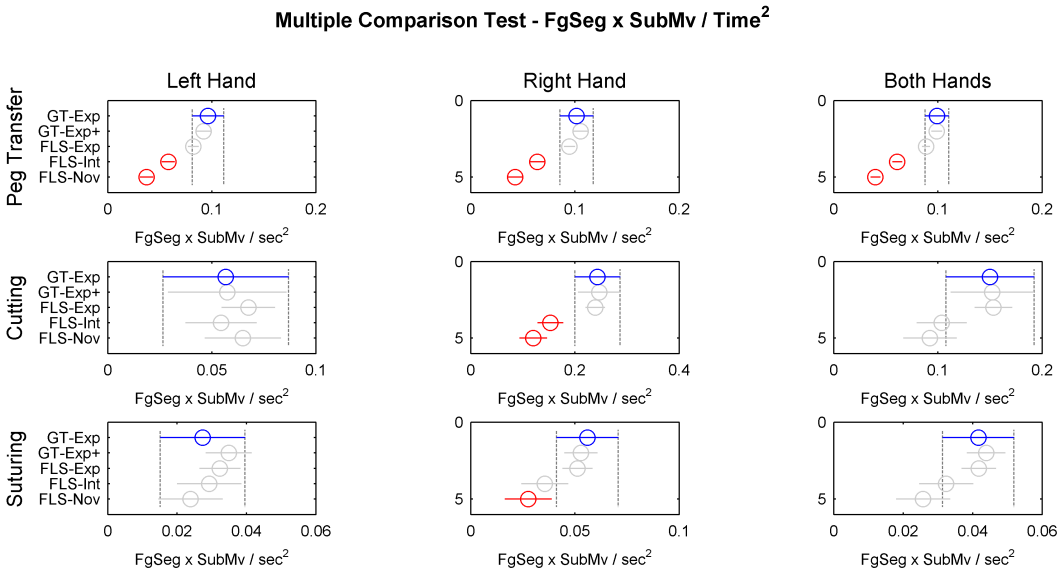


Figure 5.22: Multiple comparison test for cross validation of the product of grasp force-based segmentation (FgSeg) and submovement count (SubMv) normalized for time by ($1/T^2$). The Ground Truth Expert group (blue) was used as the basis for pairwise comparisons. Red indicates a significant difference between means, gray indicates the lack of a significant difference.

on $FgSeg/Time$ and $FgSegxSubMovements/Time^2$ discriminate skill levels equally well for the Peg Transfer task across all hand combinations. Unlike the HMMs used in sequential modeling, these metrics succeed in discriminating skill for the most clinically relevant task of Suturing.

FLS Scoring

This work has made an effort to establish whether the proposed skill metrics add value beyond task time. The same question can be asked about the widely adopted FLS-Scoring protocol whose scoring equations are repeated in Table 5.7. According to these FLS scoring equations, strict correlation between FLS and task time is required if and only if there are no errors. As error count increases, the correlation between FLS and time should decrease: individuals who commit errors should incur a measurable penalty independent of how quickly they complete a task iteration. In this manner, FLS scoring adds value beyond task time.

To experimentally determine the degree of value over task time, the correlation between FLS score and task time was computed based on the entire corpus of collected data. Figure 5.23 shows a scatter plot of FLS score vs. task time for all iterations of all tasks recorded in this work. Table 5.8 shows the correlation between these two variables for each task. Pearson's r indicates the degree of a linear correlation and Spearman's ρ the degree of monotonicity.

Table 5.7: Equations used to compute FLS scores [21, 28].

FLS Task	**FLS Score**
Peg Transfer	$FLS_{peg} = (300 - t - 17E_{dr})/237$
Cutting	$FLS_{cut} = (300 - t - 2E_a)/280$
Suturing	$FLS_{sut} = (600 - t - E_{pd} - E_g - E_q)/520$

According to the results in Table 5.8 the correlation between FLS and task time is never substantially reduced by errors for any task for the entire database. This suggests that the error scores in our data set—which must be manually collected and computed for each FLS

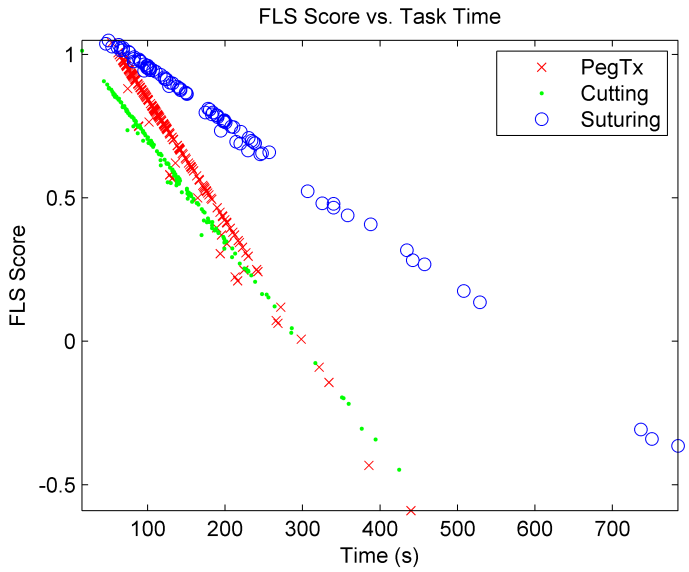


Figure 5.23: Scatter plot

Table 5.8: Correlation of task time with FLS scoring.

		PegTx	**Cutting**	**Suturing**
Time (s)	r	-0.99 (0.00)	-1.00 (0.00)	-1.00 (0.00)
	ρ	-0.99 (0.00)	-1.00 (0.00)	-1.00 (0.00)

task iteration—have a negligible effect on the final score given these equations and may not warrant the added resource cost they incur. If our data set is representative of the true distribution of FLS scores in practice and if our FLS scoring was properly applied, then the apparent correlation of FLS and task time is so strong that FLS scoring provides virtually no benefit over a simple stopwatch.

5.4 Conclusion

The goal of this work is to determine quantitative measures of surgical skill over a given segment of time with an emphasis on technical surgical skill. Chapter 2 detailed the equipment used to collect surgical data and provided quantitative validation of the accuracy of sensor measurements, kinematics, and noise suppression as well as derived measures such as tool path and time derivatives. Notably, the EDGE platform was shown to discriminate cumulative tool tip path length change with a threshold of 0.007cm between samples. Also, the frequency content of all signals except grasp variables was shown to be below 2 Hz. Chapter 2 also provided a universal way to register surgical task spaces by tracing a reference calibration block trajectory and using the Iterative Closest Point algorithm to register coordinate systems. This enables consistent measurements between different surgical data collection systems such as the three EDGE machines used in this study as well as alternative data sources that can benefit from the algorithms developed here.

A multi-institutional data collection using this validated data collection equipment resulted in 447 iterations of three FLS tasks (22.7 hours of surgical toolpath data, subject demographic information, and synchronized video) from a variety of surgeons at three geographically diverse institutions in the United States. Chapter 3 investigated the definition of skill categories as typically established in the surgical literature via subject demographics, performance measures, and double-blind review of task videos by surgical faculty. It demonstrated disagreement between these three criteria, then motivated and defined the development of a rigorous approach to defining a "ground truth" of surgical skill by combining all three criteria. These criteria were applied to determine a set of Ground Truth Expert task iterations (not subjects) that characterized expert skill. These were used for training and evaluation of surgical skill metrics in this work.

The feature extraction and feature selection steps, detailed in Chapter 4, provided substantial dimensionality reduction and mapped input sensor variables into a small, discrete set of values for sequential processing via vector quantization (VQ). This generated multiple codebooks through a variant of the k-means algorithm and resulted in a codebook of optimal size and of the maximum size (150 codewords) for each choice of input variables and each task. The choice of input variables, intended to boost relevant information for skill discrimination, was based on a combination of hypotheses and an attempt to establish a point of reference with the prior work of Rosen et al. to enable comparison of this work with prior art. Notable examples included the hypothesis that scalar motion variables may be sufficient to discriminate skill and that segmenting data in a meaningful way may aid in comparing similar trajectories. A framework for universal segmentation schemes was developed and three types of particular segmentation schemes were adopted: grasp variables (grasp angle and grasp force), grasp force alone (SegFg), and speed based segmentation (SegSpd).

The feature selection step in Chapter 4 used quantitative criteria based on information gain to determine which features and segmentation schemes best discriminated skill. Static information gain indicated that position-based features best discriminate skill when sequential information is completely ignored. Sequential information gain (SIG) showed better performance and suggested that position-based features rank poorly once sequential information is considered, and that scalar motion-derived features such as those based on speed and acceleration appear to perform the best under the SIG criterion, thus lending empirical support for this hypothesis. Finally, SIG was combined with segmentation schemes to yield some of the highest information gains, suggesting that segmentation may indeed boost skill discrimination. However, this was shown to come at the cost of task independence as higher SIG values for segmentation schemes showed greater deviation (less consistent performance) when applied across all tasks and both hands, indicating it worked well for some tasks but not others. Chapter 4 identified the optimal candidate features and segmentation schemes for the sequential analysis of Chapter 5. These included a whole-task approach (no segmentation) based on *Vel-dRta-Gr* (vector tip velocity, rotation rate, and grasp angle and force variables) of optimal codebook size, which was the closest feature to prior work by

Rosen et al. [90]; features based on *Vel-dRta-Gr-dQg* (which added grasp rate) and scalar variables *SpdAcc-dRta* (tip speed, acceleration magnitude, and normalized rotation rate) of either optimal (-Opt) or maximum codebook size (-150); and segmentation based on grasp force (SegFg) or speed (SegSpd).

Chapter 5 proposed two methods to control for task time to indicate whether a skill metric provides value over a stopwatch: $(1/T)$ which normalized a score by time and $[1 : T_w]$ which took the first 45 seconds of each iteration. These two methods can be applied to any surgical skill metric, not just the metrics developed and analyzed in this work. Three types of sequential statistical models were adopted for analysis, all based on HMMs. The first, which was derived to be most similar to that of Rosen et al., employed no segmentation (Whole Task), no assumptions about model structure (Dense) other than a state size of 15, and had feature inputs most similar to the Rosen's work. The second employed the same dense structure but also incorporated segmentation based either on SegFg or SegSpd since these provided the highest sequential information values in the feature selection analysis. The third model also used segmentation but used Bakis (left-to-right) model structure and larger state size. All HMMs analyzed in this work appear in Table 5.1. Chapter 5 also provided two methods of evaluating surgical skill via HMMs: symmetric dissimilarity used by Rosen et al. and log likelihood ratio (LLR). The models were trained on the Ground Truth Expert and FLS Novice sets provided in Chapter 3.

The symmetric dissimilarity method failed to discriminate skill despite its substantial computational cost relative to LLR-based evaluation. The results of Figure 5.4 indicate that with this method, the expert model proved less discriminative than the novice model. The tool motion of experts and novices may exhibit more similarities than differences or the skill-relevant differences are perhaps too subtle to be picked up by the modeling, training method, or feature extraction schemes employed, despite the observation-driven hypotheses used to develop them. This suggests future work should focus on features, models, and model training that employ discriminative techniques. These techniques should identify and focus on the set of subtle but important differences in the data while ignoring the majority of common attributes irrelevant to skill discrimination. Since these differences are subtle and difficult to articulate, such an a posteriori approach my provide better results

than an a priori hypothesis-driven approach.

The method of log likelihood ratio (LLR) established in Eqn. 5.3 and 5.4 performed better than the Symmetric Dissimilarity approach of Rosen et al. in a number of ways. Unlike the dissimilarity method, the LLR approach provided several instances of zero misclassification in the training set, as depicted by the zero entries in Figures 5.11 and 5.12. LLR evaluation resulted in much faster computation, which is amenable to real-time feedback. It also provided a means of establishing skill at each point in time for a given task as shown in Fig. 5.3. Thus, it enables a "dynamic metric" analysis, which was the original goal of this work in contrast to the cumulative score of summary metrics like task time or path length. However, since there is no supervised labeling of skill at each individual point in the collected data, there is no way to verify the accuracy of this "dynamic metric" approach on a point-by-point basis. Instead, only the cumulative classification result is used to determine whether this point-by-point analysis works on average since the LLR method effectively integrates all point-by-point scores for each iteration. Finally, this evaluation method, in conjunction with the Dense.SegFg(VeldQg-Opt) HMM provides value beyond task time: a number of training set classification results with zero misclassifications (Figures 5.11 and 5.12) were obtained by controlling for task time with both time normalization methods: scaling by $(1/T)$ or considering only the first 45 seconds of all tasks $[1 : Tw]$. In both cases, task time was effectively eliminated.

The Dense.SegFg(VeldQg-Opt) HMM establishes construct validity for the Peg Transfer and Cutting tasks by completely differentiating ground truth expert and novice skill sets in the training set with zero false positive misclassifications, indicating no novices were misclassified as experts (see Fig 5.6 and 5.12). This was true for both time normalization schemes. This model employs a dense 15 state HMM, grasping-force based segmentation, and features based on optimal codebook size for vector velocity, rotation rate, grasp variables (Qg, Fg) and grasp rate. All other candidate HMM models failed to provide such results. No HMM model established such a degree of skill separation in the training set for FLS Suturing, the most clinically relevant task.

The Dense.WholeTask(Vel-Opt) HMM was the most similar to the approach of Rosen et al. It provided fewest misclassifications in the FLS Suturing task of all HMMs. This model

also used a dense 15-state HMM but without the segmentation or grasp rate (dQg) present in the Dense.SegFg(VeldQg-Opt) HMM. The Dense.WholeTask(Vel-Opt) model presented a Mathews correlation coefficient (see Eqn. 5.5 and Table 5.4) of 0.91 for the Suturing task but 0.85 overall MCC for all tasks. The Dense.SegFg(VeldQg-Opt) model yielded a perfect 1.00 MCC value for the Peg Transfer task, 0.97 for the Cutting task, and 0.90 overall MCC for all tasks.

These two models, Dense.SegFg(VeldQg-Opt) and Dense.WholeTask(Vel-Opt), outperformed all other HMMs: they received MCC scores above 0.90 for each task. If MCC values above 0.90 are adopted as a minimal per-task classification performance threshold for the training set, then only these two models are acceptable (See Table 5.4). All other HMMs are disqualified since they received lower MCC scores for each task as well as substantially lower overall MCC scores. This suggests that a number of hypotheses regarding skill modeling are either incorrect or not verifiable with the extracted features, model types, or training methods used. Specifically, HMMs employing scalar tool motion alone failed to discriminate skill; the maximum codebook size of 150 decreased classification performance compared to the smaller optimal codebook size. This suggests that sequential information gain does not adequately model HMM-based sequential analysis and that the sparsity introduced in the Bakis left-to-right HMMs does not, in fact, serve to better discriminate skill.

Cross validation for the Dense.SegFg(VeldQg-Opt) and Dense.WholeTask(Vel-Opt) HMMs indicated a significant difference ($p < 0.01$) among the mean LLR scores of both training and evaluation sets via a one-way ANOVA analysis for both methods of time normalization (see Figures 5.13, 5.14, and 5.5). Notably, the Ground Truth Expert Plus set (GT-Exp+) and FLS-Intermediate set (FLS-Int) included iterations not used in model training. The mean score from the GT-Exp set to the FLS-Intermediate and FLS-Novice iterations was statistically different ($p < 0.05$, using the Tukey-Kramer method for multiple comparison) for both models across all tasks and time normalizations with the exception of the constant time window normalization $[1 : T_w]$ for the Suturing task in the Dense.SegFg(VeldQg-Opt) HMM (see Figures 5.15 and 5.16). This verifies that models can correctly classify non-experts in a statistically significant way, assuming the FLS-Intermediate set contains mostly non-proficient iterations. Similarly, there was no significant difference between the

GT-Exp and GT-Exp+ sets, suggesting the correct classification of expert iterations that were not previously encountered by the models since they were not part of the training set. However, the FLS-Expert set showed more divergence in its mean value from the GT-Exp set (see Figure 5.15 and 5.16). In fact, the mean of FLS-Expert set is shown to statistically differ from the GT-Exp set ($p < 0.05$, using the Tukey-Kramer method for multiple comparison) for at least one time normalization method in both models for all tasks. The FLS-Expert set was the most populous but least rigorously-established of all expert sets (see Chapter 3). Thus, it is unclear how its results reflect on the cross validation of the two HMMs.

We also evaluated three alternative skill metrics based on a deterministic computation of combined movement event counts. The algorithm—the either individual force-based grasp counts or the product of grasp and submovement counts—was controlled for time in either of two ways: normalization by $(1/T)$ or by $(1/T^2)$. It provided some of the best time-controlled skill evaluation when only the right hand was considered (Fig. 5.18 and 5.19). Cross validation was carried out in the same way as for the two HMM models above. A one-way ANOVA determined there was significant difference ($p < 0.01$) between the mean values of the three metrics for both hands and for the right hand alone, but not for the left hand alone (see Table 5.6). The FgSeg/Time metric provided the most consistent evaluation results when only considering the right hand. The distance between means from the GT-Exp set to the FLS-Intermediate and FLS-Novice sets was statistically different ($p < 0.05$, using the Tukey-Kramer method for multiple comparison) for all tasks for this FgSeg/Time metric for the right hand (see Figures 5.20 – 5.22). This established that non-proficient iterations in the evaluation set were, on average, correctly classified by this metric when only the right hand side was used. The same test also showed lack of significant difference from the mean of the GT-Exp set to those of the GT-Exp+ and FLS-Expert groups. This established, on average, correct classification of proficient subjects. Unlike the two HMM methods described immediately above, when considering the right hand alone the FgSeg/Time metric succeeded in discriminating skill for the Suturing task. Also, this metric is much less complex and computationally expensive than the HMM methods. However, once the left hand was considered, it did not perform as well.

Finally, FLS Scoring was shown to correlate almost perfectly with task time (Fig. 5.23 and Table 5.8). If our implementation of the published FLS scoring method is correct and our iterations are representative of FLS in practice, this indicates that FLS scoring provides little or no additional information beyond task time. This suggests that the added resource cost of manually computing errors and FLS scores may not be justified.

This work establishes two HMMs which quantitatively measure surgical skill: one model for the Peg Transfer and Cutting tasks and another for the Suturing task. Both models are equipped with a means of real-time evaluation that can enable "dynamic metrics," which report skill level for each time instant. An alternative, deterministic metric based on the rate of force-derived grasp counts is also shown to successfully discriminate surgical skill when considering the right hand alone. Most importantly, all metrics are evaluated to determine if they provide benefit over task time. While FLS scoring fails to provide this benefit, the combination of novel skill metrics found in this work does.

BIBLIOGRAPHY

[1] M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. Dover publications, 1964.

[2] J.A. Adams. A closed-loop theory of motor learning. *Journal of motor behavior*, 1971.

[3] R. Aggarwal, K. Moorthy, and A. Darzi. Laparoscopic skills training and assessment. *British journal of surgery*, 91(12):1549–1558, 2004.

[4] M.A. Al-Alaoui. Novel digital integrator and differentiator. *Electronics Letters*, 29(4):376–378, 1993.

[5] Anonymous. Cholecystectomy, practice transformed. *Lancet*, 338:789–790, 1990.

[6] H. Arora, J. Uribe, W. Ralph, M. Zeltsan, H. Cuellar, A. Gallagher, and MP Fried. Assessment of construct validity of the endoscopic sinus surgery simulator. *Archives of otolaryngology–head & neck surgery*, 131(3):217, 2005.

[7] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[8] P. Batalden, D. Leach, S. Swing, H. Dreyfus, and S. Dreyfus. General competencies and accreditation in graduate medical education. *Health Affairs*, 21(5):103, 2002.

[9] M. Berry, T. Lystig, R. Reznick, and L. Lönn. Assessment of a virtual interventional simulator trainer. *Journal of Endovascular Therapy*, 13(2):237–243, 2006.

[10] Sanne Botden, Sonja Buzink, Marlies Schijven, and Jack Jakimowicz. Augmented versus virtual reality laparoscopic simulation: What is the difference? *World Journal of Surgery*, 31:764–772, 2007.

[11] A.O. Castellvi, L.A. Hollett, A. Minhajuddin, D.C. Hogg, S.T. Tesfay, and D.J. Scott. Maintaining proficiency after fundamentals of laparoscopic surgery training: A 1-year analysis of skill retention for surgery residents. *Surgery*, 146(2):387–393, 2009.

[12] R.A. Chaer, B.G. DeRubertis, S.C. Lin, H.L. Bush, J.K. Karwowski, D. Birk, N.J. Morrissey, P.L. Faries, J.F. McKinsey, and K.C. Kent. Simulation improves resident performance in catheter-based intervention: results of a randomized, controlled study. *Annals of surgery*, 244(3):343, 2006.

[13] HR Champion and AG Gallagher. Surgical simulation—a 'good idea whose time has come'. *British journal of surgery*, 90(7):767–768, 2003.

[14] MK Chmarra, CA Grimbergen, and J. Dankelman. Systems for tracking minimally invasive surgical instruments. *Minimally Invasive Therapy & Allied Technologies*, 16(6):328–340, 2007.

[15] A. Cuschieri et al. Whither minimal access surgery: tribulations and expectations. *American Journal of surgery*, 169(1):9, 1995.

[16] A. Darzi, S. Smith, and N. Taffinder. Assessing operative skill. *Bmj*, 318(7188):887–888, 1999.

[17] V. Datta, S. Mackay, M. Mandalia, and A. Darzi. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model1. *Journal of the American College of Surgeons*, 193(5):479–485, 2001.

[18] B. Dauster, A.P. Steinberg, M.C. Vassiliou, S. Bergman, D.D. Stanbridge, L.S. Feldman, and G.M. Fried. Validity of the mistels simulator for laparoscopy training in urology. *Journal of endourology*, 19(5):541–545, 2005.

[19] AM Derossis, M. Antoniuk, and GM Fried. Evaluation of laparoscopic skills: a 2-year follow-up during residency training. *Canadian journal of surgery. Journal canadien de chirurgie*, 42(4):293, 1999.

[20] AM Derossis, J. Bothwell, HH Sigman, and GM Fried. The effect of practice on performance in a laparoscopic simulator. *Surgical endoscopy*, 12(9):1117–1120, 1998.

[21] MD Derossis, M. Anna, MD Fried, M. Gerald, M. Abrahamowicz PhD, MD Sigman, H. Harvey, MD Barkun, S. Jeffrey, MD Meakins, et al. Development of a model for training and evaluation of laparoscopic skills. *The American journal of surgery*, 175(6):482–487, 1998.

[22] D.J. Deziel, K.W. Millikan, S.G. Economou, A. Doolas, S.T. Ko, and M.C. Airan. Complications of laparoscopic cholecystectomy: A national survey of 4,292 hospitals and an analysis of 77,604 cases. *The American journal of surgery*, 165(1):9–14, 1993.

[23] W. Dick and N. Hagerty. *Topics in Measurement: Reliability and Validity.* McGraw-Hill Book Company, 330 West 42nd Street, New York, NY, 10036, 1971.

[24] A. Dosis, F. Bello, D. Gillies, S. Undre, R. Aggarwal, and A. Darzi. Laparoscopic task recognition using hidden markov models. *Studies in Health Technology and Informatics*, 111:115–122, 2005.

[25] L.S. Feldman, J. Cao, A. Andalib, S. Fraser, and G.M. Fried. A method to characterize the learning curve for performance of a fundamental laparoscopic simulator task: Defining. *Surgery*, 146(2):381–386, 2009.

[26] L.S. Feldman, S.E. Hagarty, G. Ghitulescu, D. Stanbridge, and G.M. Fried. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *Journal of the American College of Surgeons*, 198(1):105–110, 2004.

[27] L.S. Feldman, V. Sherman, and G.M. Fried. Using simulators to assess laparoscopic competence: ready for widespread use? *Surgery*, 135(1):28, 2004.

[28] SA Fraser, LS Feldman, D. Stanbridge, and GM Fried. Characterizing the learning curve for a basic laparoscopic drill. *Surgical endoscopy*, 19(12):1572–1578, 2005.

[29] SA Fraser, DR Klassen, LS Feldman, GA Ghitulescu, D. Stanbridge, and GM Fried. Evaluating laparoscopic skills: Setting the pass/fail score for the mistels system. *Surgical endoscopy*, 17(6):964–967, 2003.

[30] GM Fried, AM Derossis, J. Bothwell, and HH Sigman. Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. *Surgical endoscopy*, 13(11):1077–1081, 1999.

[31] G.M. Fried, L.S. Feldman, M.C. Vassiliou, S.A. Fraser, D. Stanbridge, G. Ghitulescu, and C.G. Andrew. Proving the value of simulation in laparoscopic surgery. *Annals of surgery*, 240(3):518, 2004.

[32] AG Gallagher, R. Cowie, I. Crothers, J.A. Jordan-Black, and RM Satava. Picsor: an objective test of perceptual skill that predicts laparoscopic technical skill in three initial studies of laparoscopopic performance. *Surgical endoscopy*, 17(9):1468–1471, 2003.

[33] A.G. Gallagher, E.M. Ritter, H. Champion, G. Higgins, M.P. Fried, G. Moses, C.D. Smith, and R.M. Satava. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, 241(2):364, 2005.

[34] AG Gallagher and RM Satava. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. *Surgical Endoscopy*, 16(12):1746–1752, 2002.

[35] A.G. Gallagher, Ph. D, K. Richie, B. Sc, N. McClure, MD, and J. McGuigan. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World journal of surgery*, 25(11):1478–1483, 2001.

[36] A. Gersho and R.M. Gray. *Vector quantization and signal compression*, volume 159. Springer, 1992.

[37] T.P. Grantcharov, L. Carstensen, and S. Schulze. Objective assessment of gastrointestinal endoscopy skills using a virtual reality simulator. *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, 9(2):130, 2005.

[38] S. Gunther, J. Rosen, B. Hannaford, and M. Sinanan. The Red DRAGON: a multimodality system for simulation and training in minimally invasive surgery. *Studies in health technology and informatics*, 125:149, 2007.

[39] Scott Gunther. Red dragon: A multi-modality system for simulation and training in minimally invasive surgery. Master's thesis, University of Washington, May 2006.

[40] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[41] B. Hannaford and P. Lee. Hidden markov model of force torque information in telemanipulation. *International Journal of Robotics Research*, 10(5):528–539, 1991.

[42] Haptica. Promis validation. Online, 2010. http://www.cae.com/en/healthcare/promis.simulator.asp.

[43] G.B. Healy. The college should be instrumental in adapting simulators to education. *Bulletin of the American College of Surgeons*, 87(11):10, 2002.

[44] D. Heckerman, A. Mamdani, and M.P. Wellman. Real-world applications of bayesian networks. *Communications of the ACM*, 38(3):24–26, 1995.

[45] D.E. Heckerman and Stanford University. Knowledge Systems Laboratory. *Probabilistic similarity networks*. MIT press, 1991.

[46] Pavel Holoborodko. Smooth noise-robust differentiators. *online: http://www.holoborodko.com/pavel/numerical-methods/numerical-derivative/smooth-low-noise-differentiators/*, 2012.

[47] HystSim. Hystsim. Online, 2011. http://simbionix.com/simulators/clinical-validations/hystsim/.

[48] T. Inamura, M. Inaba, and H. Inoue. A dialogue control model based on ambiguity evaluation of users' instructions and stochastic representation of experiences. *Journal of Robotics and Mechatronics*, 17(6):697, 2005.

[49] T. Inamura, H. Tanie, and Y. Nakamura. From stochastic motion generation and recognition to geometric symbol development and manipulation. In *International Conference on Humanoid Robots*, 2003.

[50] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura. Embodied symbol emergence based on mimesis theory. *The International Journal of Robotics Research*, 23(4-5):363–377, 2004.

[51] K. Itabashi, K. Hirana, T. Suzuki, S. Okuma, and F. Fujiwara. Modelling and realization of the peg-in-hole task based on hidden markov model. In *Robotics and Automation, 1998. Proceedings. 1998 IEEE International Conference on*, volume 2, pages 1142–1147. IEEE, 1998.

[52] Oleynikov D. Judkins, T.N. and N. Stergiou. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surgical endoscopy*, 23(3):590–597, 2009.

[53] Timothy Judkins, Dmitry Oleynikov, and Nick Stergiou. Objective evaluation of expert performance during human robotic surgical procedures. *Journal of Robotic Surgery*, 1:307–312, 2008.

[54] P.A. Kenney, M.F. Wszolek, J.J. Gould, J.A. Libertino, and A. Moinzadeh. Face, content, and construct validity of dv-trainer, a novel virtual reality simulator for robotic surgery. *Urology*, 73(6):1288–1292, 2009.

[55] E.J. Keyser, A.M. Derossis, M. Antoniuk, H.H. Sigman, and G.M. Fried. A simplified simulator for the training and evaluation of laparoscopic skills. *Surgical endoscopy*, 14(2):149–153, 2000.

[56] D. Koller and N. Friedman. *Probabilistic graphical models*. MIT Press, 2009.

[57] J.R. Korndorffer et al. Simulator training for laparoscopic suturing using performance goals translates to the operating room. *Journal of the American College of Surgeons*, 201(1):23–29, 2005.

[58] T.M. Kowalewski, J. Rosen, L. Chang, M. Sinanan, and B. Hannaford. Optimization of a vector quantization codebook for objective evaluation of surgical skill. In *Proc. Medicine Meets Virtual Reality 12*, pages 174–179, January 2004.

[59] D. Kragic, P. Marayong, M. Li, A.M. Okamura, and G.D. Hager. Human-machine collaborative systems for microsurgical applications. *The International Journal of Robotics Research*, 24(9):731–741, 2005.

[60] P.S. Kundhal and T.P. Grantcharov. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surgical endoscopy*, 23(3):645–649, 2009.

[61] B. Law, M.S. Atkins, AE Kirkpatrick, and A.J. Lomax. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the Eye tracking research applications symposium on Eye tracking research applications ETRA2004*, volume 1, pages 41–48. ACM Press, 2004.

[62] T.S. Lendvay, P. Casale, R. Sweet, and C. Peters. Initial validation of a virtual-reality robotic simulator. *Journal of Robotic Surgery*, 2(3):145–149, 2008.

[63] M. Li and A.M. Okamura. Recognition of operator motions for real-time assistance using virtual fixtures. In *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2003. HAPTICS 2003. Proceedings. 11th Symposium on*, pages 125–131. IEEE, 2003.

[64] H.C. Lin, I. Shafran, D. Yuh, and G.D. Hager. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery*, 11(5):220–230, 2006.

[65] Henry Lin, Izhak Shafran, Todd Murphy, Allison Okamura, David Yuh, and Gregory Hager. Automatic detection and segmentation of robot-assisted surgical motions. 3749:802–810, 2005.

[66] A. Liu, F. Tendick, K. Cleary, and C. Kaufmann. A survey of surgical simulation: applications, technology, and education. *Presence: Teleoperators & Virtual Environments*, 12(6):599–614, 2003.

[67] S.M. Lucas, I.S. Zeltser, K. Bensalah, A. Tuncel, A. Jenkins, M.S. Pearle, and J.A. Cadeddu. Training on a virtual reality laparoscopic simulator improves performance of an unfamiliar live laparoscopic procedure. *The Journal of urology*, 180(6):2588–2591, 2008.

[68] M. Lum. Kinematic optimization of a 2-DOF spherical mechanism for a minimally invasive surgical robot. Master's thesis, University of Washington, December 2004.

[69] A.I.M. Macmillan and A. Cuschieri. Assessment of innate ability and skills for endoscopic manipulations by the advanced dundee endoscopic psychomotor tester: predictive and concurrent validity. *The American journal of surgery*, 177(3):274–277, 1999.

[70] JA Martin, G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison, and M. Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

[71] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and evaluation of surgical performance using hidden markov models. *Biomedical Engineering, IEEE Transactions on*, 53(10):1911–1919, 2006.

[72] GE Miller. The assessment of clinical skills/competence/performance. *Academic medicine: journal of the Association of American Medical Colleges*, 65(9 Suppl):S63, 1990.

[73] M.J. Moore, C.L. Bennett, et al. The learning curve for laparoscopic cholecystectomy. *The American journal of surgery*, 170(1):55–59, 1995.

[74] K. Moorthy, Y. Munz, S.K. Sarker, and A. Darzi. Objective assessment of technical skills in surgery. *Bmj*, 327(7422):1032, 2003.

[75] K. Narazaki, D. Oleynikov, and N. Stergiou. Robotic surgery training and performance: identifying objective variables for quantifying the extent of proficiency. *Surgical endoscopy*, 20(1):96–103, 2006.

[76] Arts MA Grantcharov T Okrainec A, Greco E. Does assessment during virtual surgery predict actual intraoperative performance? Presentation, Association for Surgical Education Annual Meeting in April 15-19, 2008 Toronto, Canad, 2008.

[77] N. Oliver and E. Horvitz. A comparison of hmms and dynamic bayesian networks for recognizing office activities. *User Modeling 2005*, pages 119–209, 2005.

[78] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[79] J. Peters, G.M. Fried, L.L. Swanstrom, N.J. Soper, L.F. Sillin, B. Schirmer, K. Hoffman, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*, 135(1):21–27, 2004.

[80] A. Prasad, RJ Foley, et al. Day case laparoscopic cholecystectomy: a safe and cost effective procedure. *The European journal of surgery= Acta chirurgica*, 162(1):43, 1996.

[81] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[82] W.S. Rasband. Imagej. U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997-2011.

[83] H.H. Rashid, T. Kowalewski, P. Oppenheimer, A. Ooms, J.N. Krieger, and R.M. Sweet. The virtual reality transurethral prostatic resection trainer: evaluation of discriminate validity. *The Journal of urology*, 177(6):2283–2286, 2007.

[84] Carol E. Reiley, Henry C. Lin, David D. Yuh, and Gregory D. Hager. A review of methods for objective surgical skill evaluation. In *Surgical Endoscopy*, July 2010.

[85] C.E. Reiley, H.C. Lin, B. Varadarajan, B. Vagvolgyi, S. Khudanpur, DD Yuh, and GD Hager. Automatic recognition of surgical motions using statistical modeling for capturing variability. *Studies in health technology and informatics*, 132:396, 2008.

[86] R. Reznick, G. Regehr, H. MacRae, J. Martin, and W. McCulloch. Testing technical skill via an innovative "bench station" examination. *The American journal of surgery*, 173(3):226–230, 1997.

[87] R.K. Reznick. Teaching and testing technical skills. *The American journal of surgery*, 165(3):358–361, 1993.

[88] E.M. Ritter and D.J. Scott. Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. *Surgical innovation*, 14(2):107, 2007.

[89] J. Rosen, J.D. Brown, L. Chang, M. Barreca, M. Sinanan, and B. Hannaford. The BlueDRAGON—a system for measuring the kinematics and dynamics of minimally invasive surgical tools in-vivo. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 2, pages 1876–1881. IEEE, 2002. blue dragon.

[90] J. Rosen, J.D. Brown, L. Chang, M.N. Sinanan, and B. Hannaford. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. *Biomedical Engineering, IEEE Transactions on*, 53(3):399–413, March 2006.

[91] J. Rosen, L. Chang, J. D. Brown, B. Hannaford, M. Sinanan, and R. Satava. Minimally invasive surgery task decomposition - etymology of endoscopicsuturing. *Studies in Health Technology and Informatics - Medicine Meets Virtual Reality*, 94:295–301, January 2003.

[92] J. Rosen, B. Hannaford, C.G. Richards, and M.N. Sinanan. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *Biomedical Engineering, IEEE Transactions on*, 48(5):579–591, 2001.

[93] J. Rosen, M. Lum, D. Trimble, B. Hannaford, and M. Sinanan. Spherical mechanism analysis of a surgical robot for minimally invasive surgery - analytical and experimental

approaches. *Studies in Health Technology and Informatics - Medicine Meets Virtual Reality, (MMVR05)*, 111:422–428, January 2005.

[94] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. *Computer Aided Surgery*, 7(1):49–61, 2002.

[95] R.M. Satava. Virtual reality surgical simulator. the first steps. *Surgical endoscopy*, 7(3):203, 1993.

[96] RM Satava. The need for metrics in surgical education. *Surgical endoscopy*, 13(11):1082, 1999.

[97] R.M. Satava, A. Cuschieri, and J. Hamdorf. Metrics for objective assessment. *Surgical endoscopy*, 17(2):220–226, 2003.

[98] MP Schijven, J. Jakimowicz, and C. Schot. The advanced dundee endoscopic psychomotor tester (adept) objectifying subjective psychomotor test performance. *Surgical endoscopy*, 16(6):943–948, 2002.

[99] R.A. Schmidt. A schema theory of discrete motor skill learning. *Psychological review*, 82(4):225, 1975.

[100] R.A. Schmidt and T.D. Lee. *Motor control and learning: A behavioral emphasis*. Human Kinetics Publishers, 2005.

[101] D.J. Scott, E.M. Ritter, S.T. Tesfay, E.A. Pimentel, A. Nagji, and G.M. Fried. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. *Surgical endoscopy*, 22(8):1887–1893, 2008.

[102] A.S. Sethi, W.J. Peine, Y. Mohammadi, and C.P. Sundaram. Validation of a novel virtual reality robotic simulator. *Journal of Endourology*, 23(3):503–508, 2009.

[103] N.E. Seymour, A.G. Gallagher, S.A. Roman, M.K. O'Brien, V.K. Bansal, D.K. Andersen, and R.M. Satava. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery*, 236(4):458, 2002.

[104] Simbionix. Gimentor validation & publication. Online, 2011. http://simbionix.com/simulators/clinical-validations/gi-mentor/.

[105] Simbionix. Lapmentor validation & publication. Online, 2011. http://simbionix.com/simulators/clinical-validations/lap-mentor/.

[106] D. Stefanidis, J.R. Korndorffer, et al. Proficiency maintenance: impact of ongoing simulator training on laparoscopic skill retention. *Journal of the American College of Surgeons*, 202(4):599–603, 2006.

[107] D. Stefanidis, R. Sierra, J.R. Korndorffer, et al. Intensive continuing medical education course training on simulators results in proficiency for laparoscopic suturing. *The American journal of surgery*, 191(1):23–27, 2006.

[108] SurgicalScience. Lap sim validation & publication. Online, 2011. http://www.surgicalscience.com/productsmain/others/validation-studies/.

[109] L.M. Sutherland, P.F. Middleton, A. Anthony, J. Hamdorf, P. Cregan, D. Scott, and G.J. Maddern. Surgical simulation: a systematic review. *Annals of Surgery*, 243(3):291, 2006.

[110] L.L. Swanstrom, G.M. Fried, K.I. Hoffman, and N.J. Soper. Beta test results of a new system assessing competence in laparoscopic surgery. *Journal of the American College of Surgeons*, 202(1):62–69, 2006.

[111] R. Sweet, T. Kowalewski, P. Oppenheimer, S. Weghorst, and R. Satava. Face, content and construct validity of the university of washington virtual reality transurethral prostate resection trainer. *The Journal of urology*, 172(5):1953–1957, 2004.

[112] M. Tavakol, M.A. Mohagheghi, and R. Dennick. Assessing the skills of surgical residents using simulation. *J Surg Educ*, 65(2):77–83, 2008.

[113] E. Thelen, L.B. Smith, D.J. Lewkowicz, and R. Lickliter. *A dynamic systems approach to the development of cognition and action*. Number 2. MIT Press, 1994.

[114] S. Thrun, W. Burgard, and D. Fox. Probabilistic robotics. 2005.

[115] Balakrishnan Varadarajan, Carol Reiley, Henry Lin, Sanjeev Khudanpur, and Gregory Hager. Data-derived models for segmentation with application to surgical assessment and training. 5761:426–434, 2009.

[116] Mentice VIST. Mentice vist references. Online, 2008. http://mentice46.kaigan.se/archive/pdf_products/VIST_references.pdf.

[117] J.D. Watterson, D.T. Beiko, J.K. Kuan, and J.D. Denstedt. A randomized prospective blinded study validating acquistion of ureteroscopy skills using a computer based virtual reality endourological simulator. *The Journal of urology*, 168(5):1928–1932, 2002.

[118] Andrew S. Wright, Timothy M. Kowalewski, and Blake Hannaford. Novel laparoscopic box trainer with integrated force and positioning sensors. In *12th World Congress of Endoscopic Surgery, Emerging Technology Session, National Harbor, MD*, April 2010.

[119] J. Yang, Y. Xu, and C.S. Chen. Human action learning via hidden markov model. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 27(1):34–44, 1997.

[120] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1992.

Appendix A

HISTOGRAMS AND NORMALIZATION OF ALL VQ INPUT DATA

The following pages provide the histograms of all data dimensions used as inputs for training VQ codebooks. They are separated by task name and left or right hand data. Raw histograms demonstrate the percentiles (2 and 98) used to determine the linear normalization that maps the percentile points to normalized coordinates -1 and 1 respectively. The percentiles used to exclude outliers are also presented. The histograms of normalized data demonstrate the effect of this normalization and exclusion of outliers.

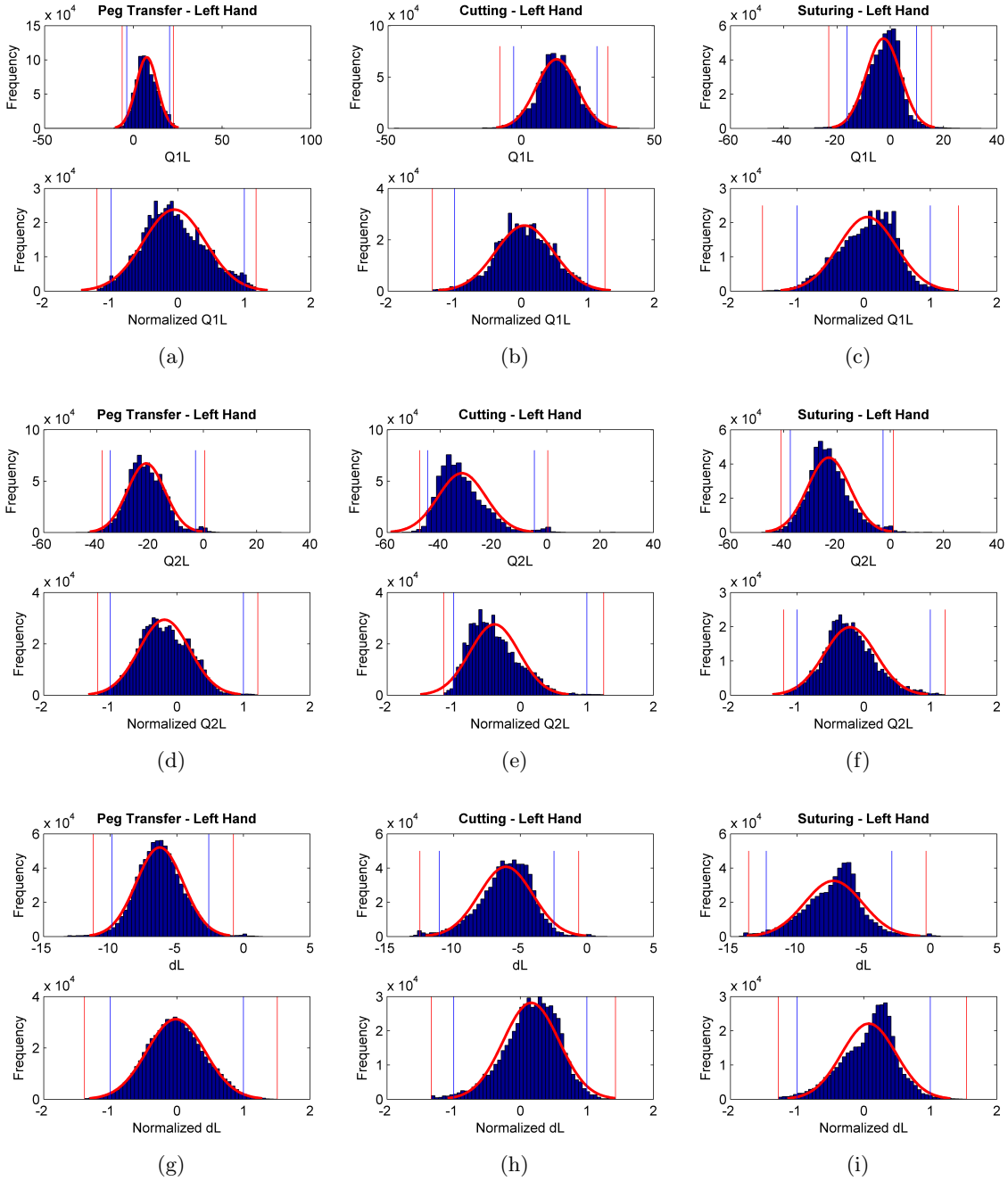


Figure A.1: Histograms of tool position sensor variables for the left hand. Q1 and Q2 refer to angles in degrees of EDGE Joints 1 and 2 respectively; dL is the tool insertion distance in cm. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

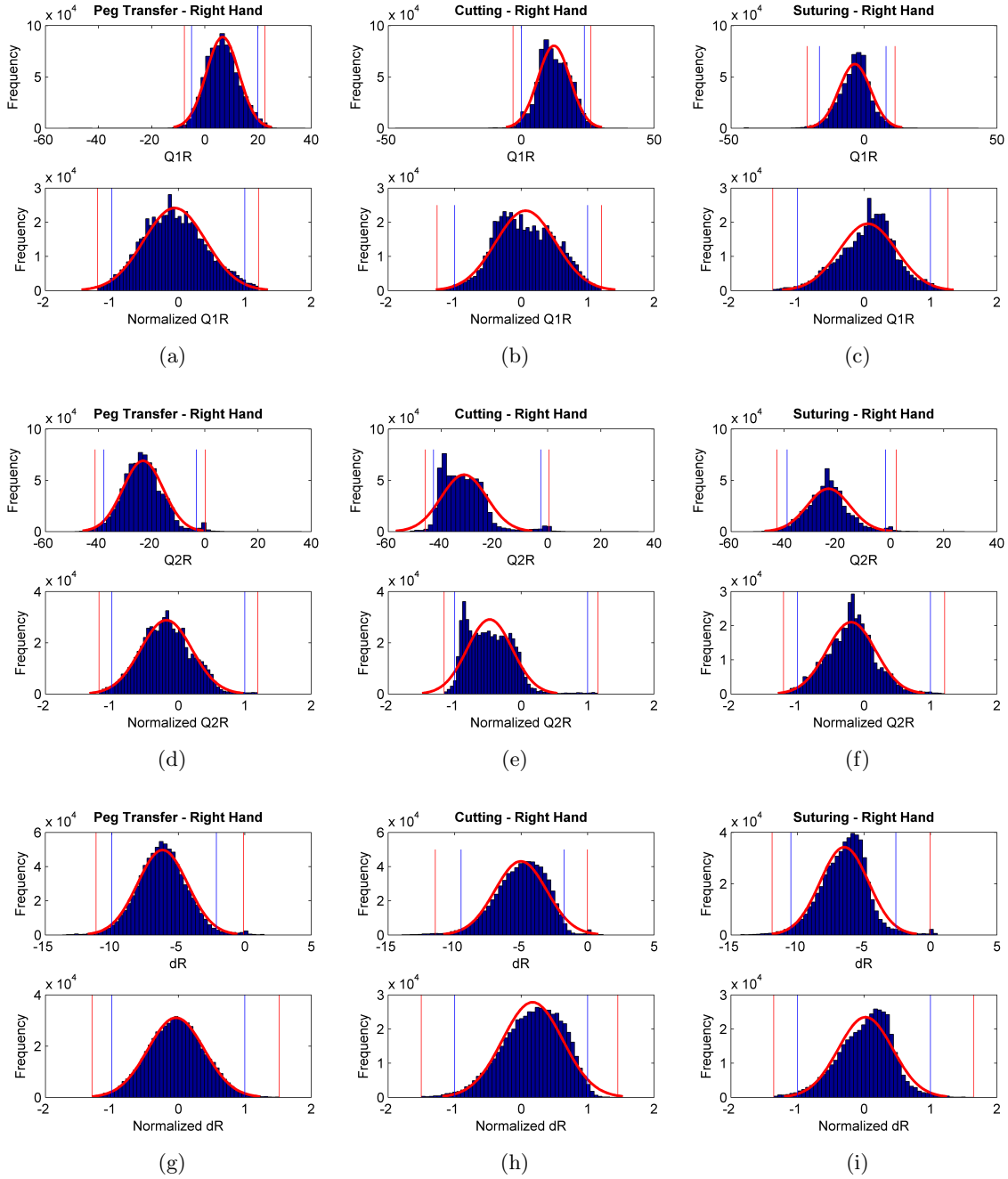


Figure A.2: Histograms of tool position sensor variables for right hand. Q1 and Q2 refer to angles in degrees of EDGE Joints 1 and 2 respectively; d is the tool insertion distance in cm. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

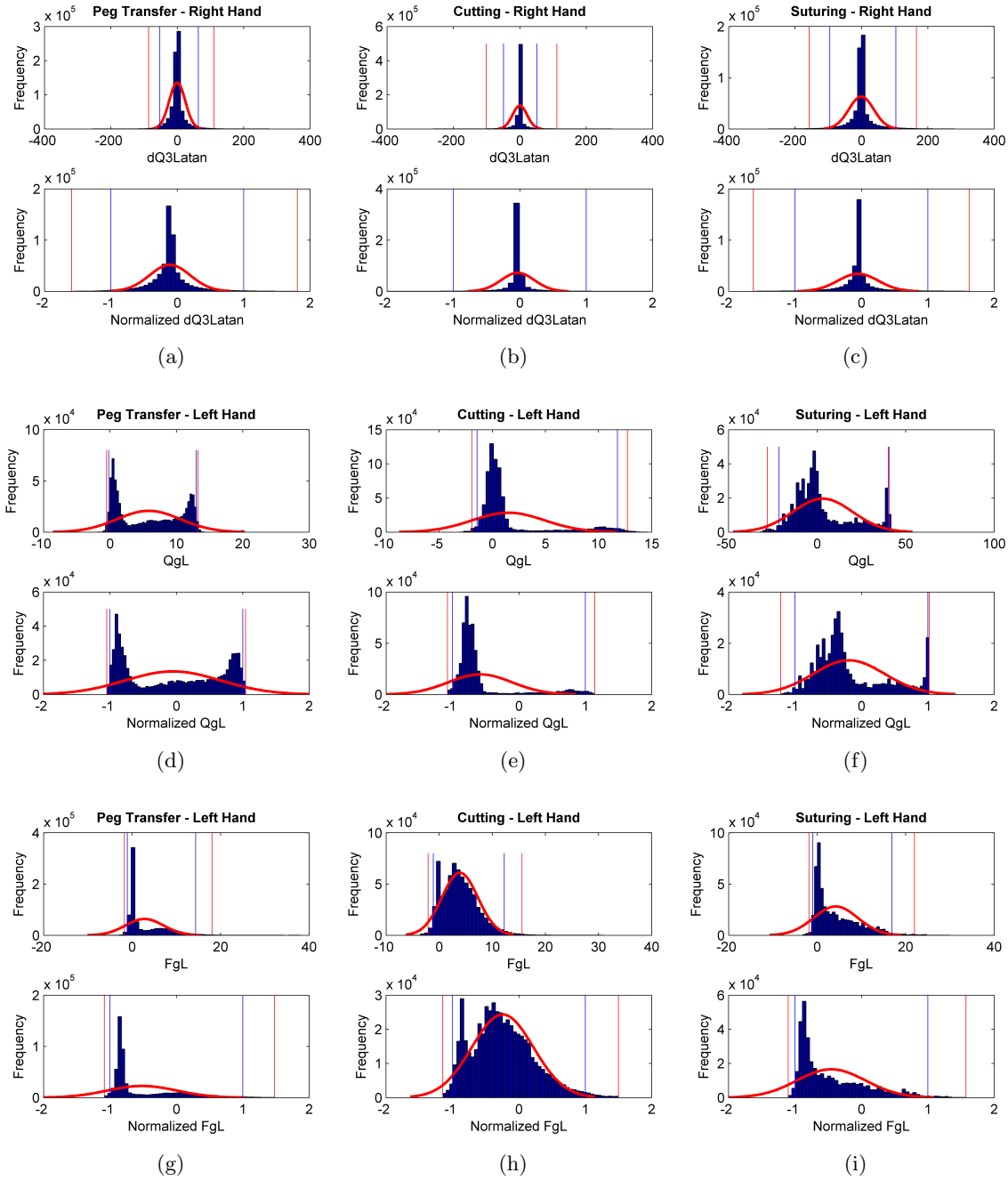


Figure A.3: Histograms of rotation rate and grasp sensor variables for the left hand. The rotational velocity of the tool (in $°/sec$) about its axis is dQ3atan; Qg is the grasper angle (in °) and Fg the grasp force (in Newtons). 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

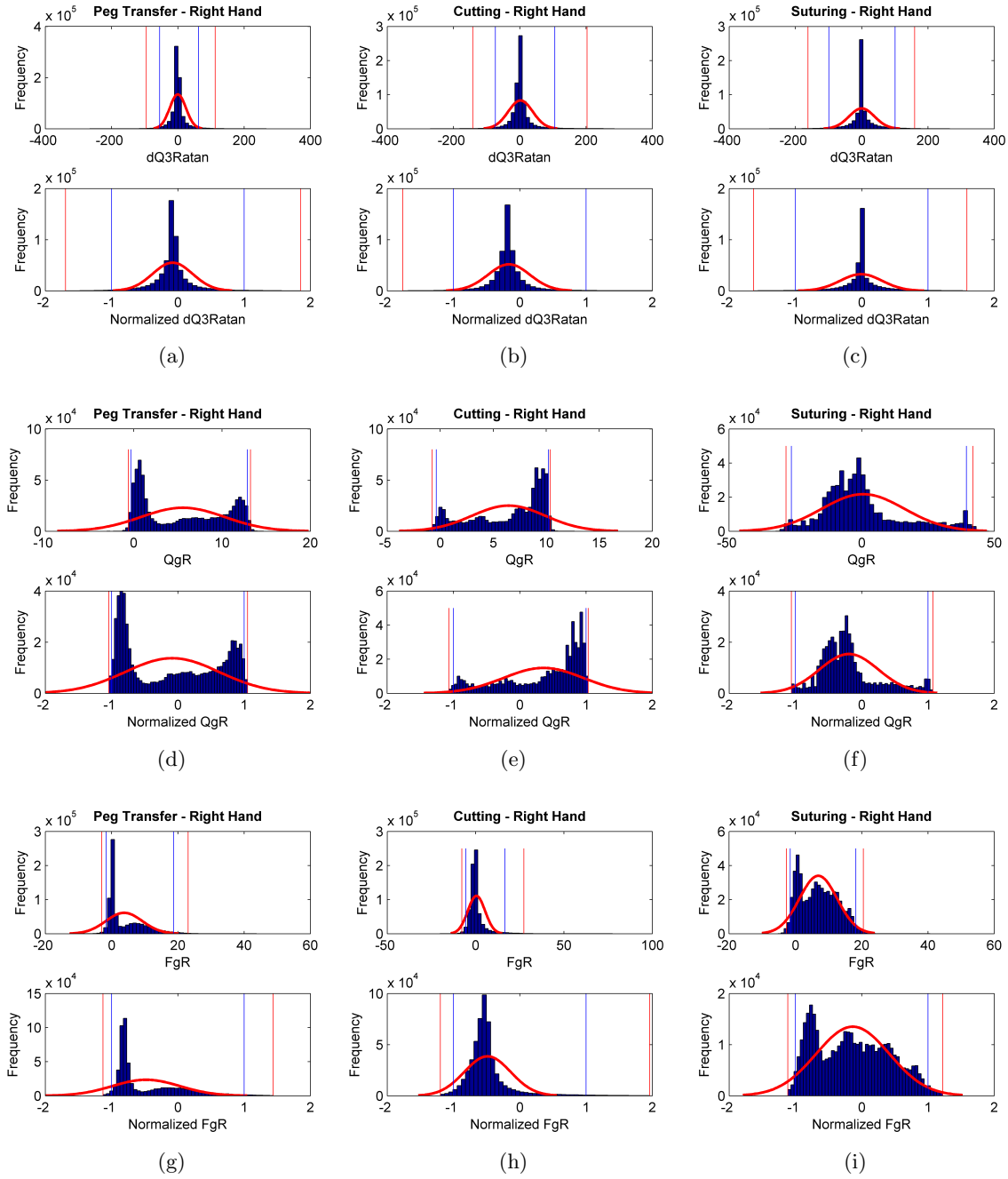


Figure A.4: Histograms of tool position sensor variables for right hand. The rotational velocity of the tool (in $^\circ/sec$) about its axis is dQ3atan; Qg is the grasper angle (in $^\circ$) and Fg the grasp force (in Newtons). 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

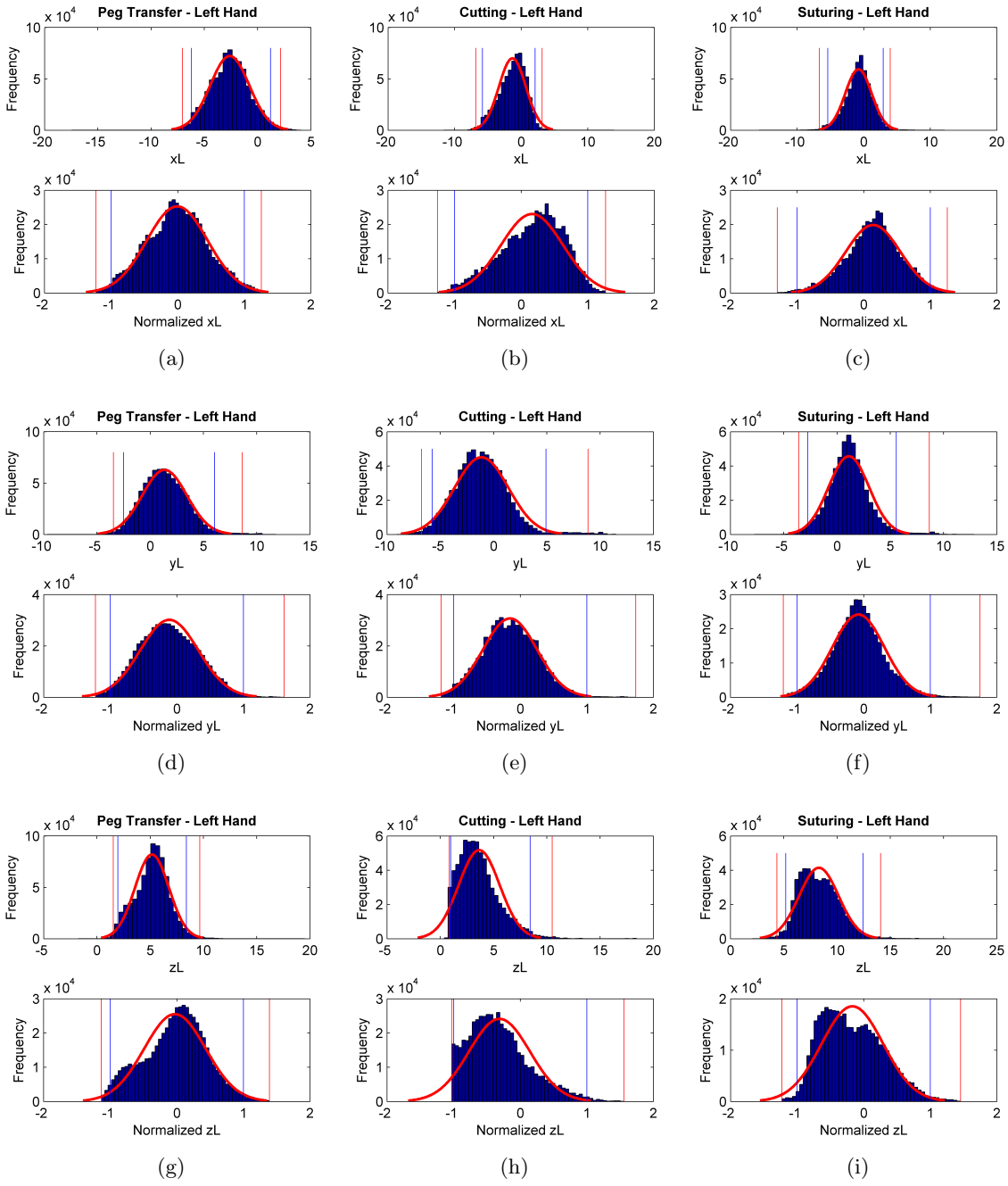


Figure A.5: Histograms of Cartesian position (x, y, z) in cm for the left hand. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

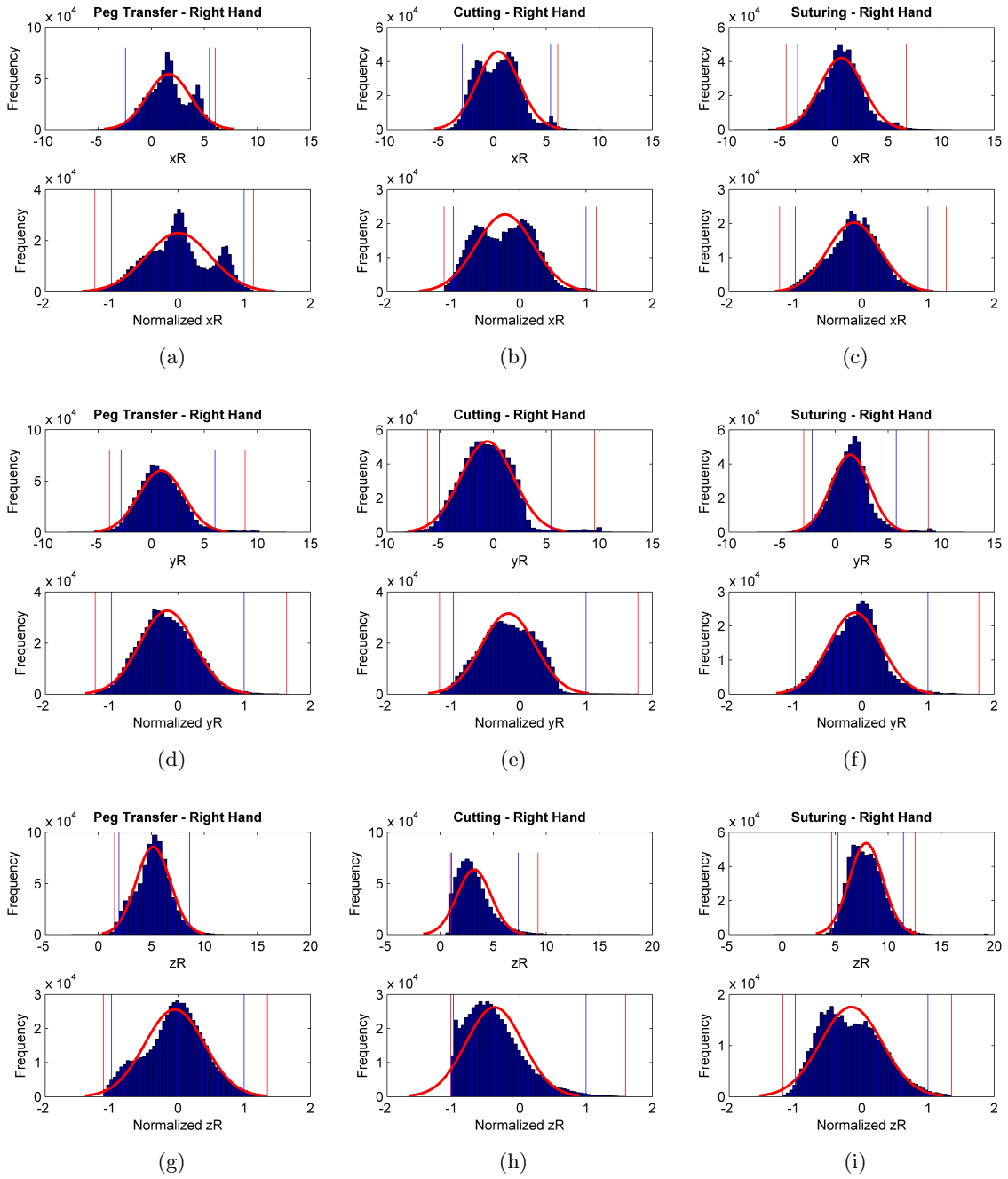


Figure A.6: Histograms of Cartesian position (x, y, z) in cm for the right hand. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

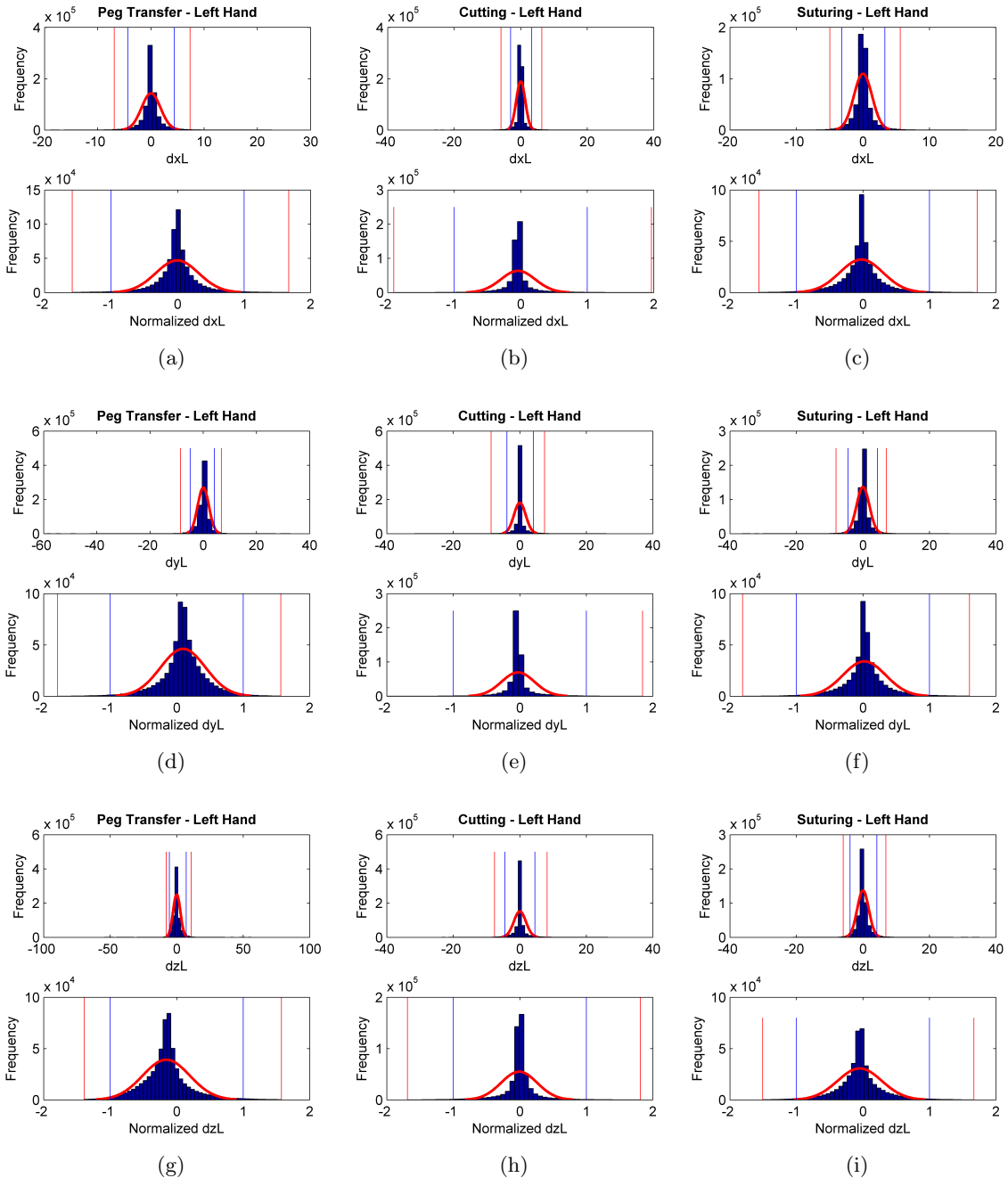


Figure A.7: Histograms of Cartesian velocity (v_x, v_y, v_z) in cm/s for the left hand. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

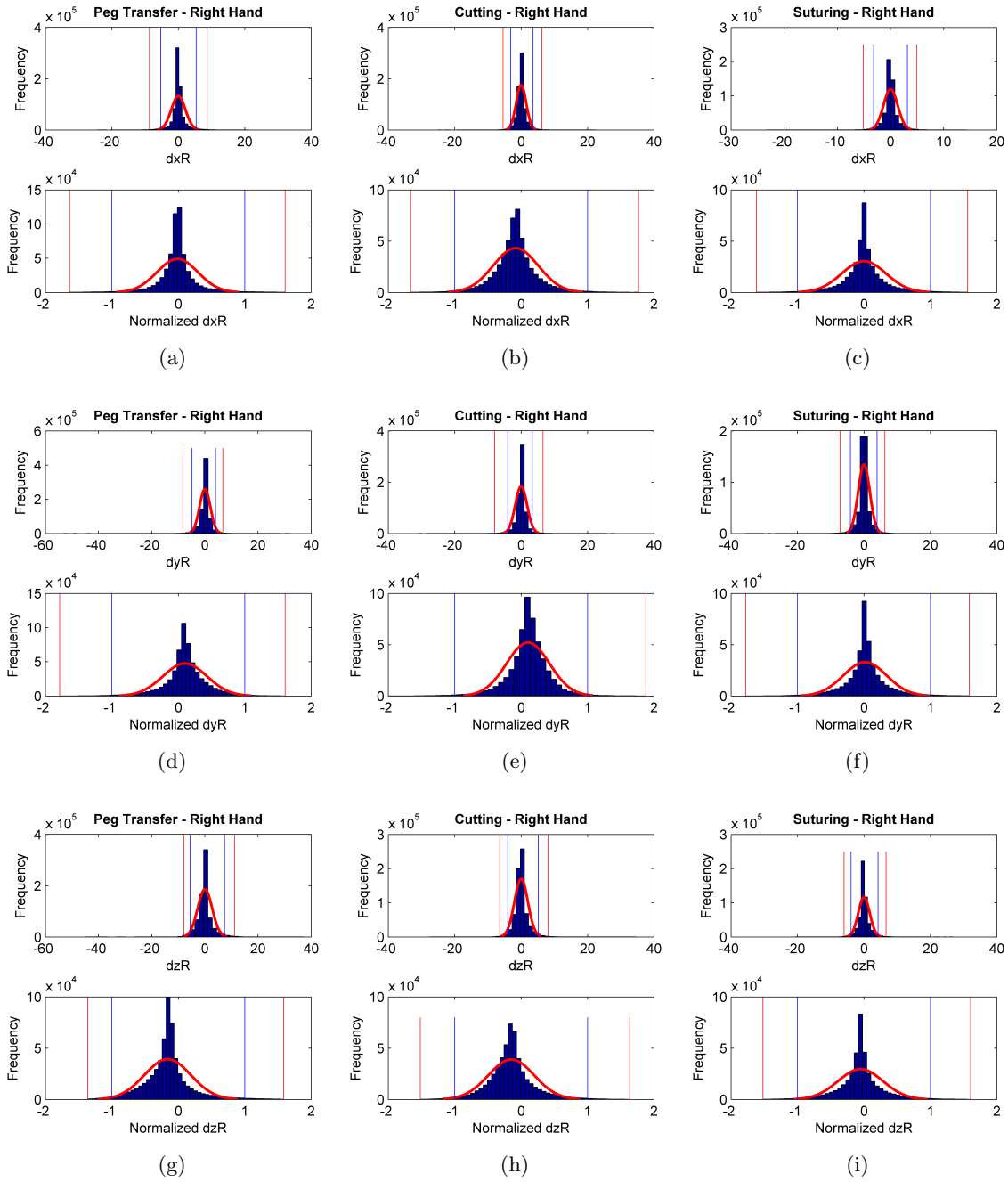


Figure A.8: Histograms of Cartesian velocity (v_x, v_y, v_z) in cm/s for the right hand. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

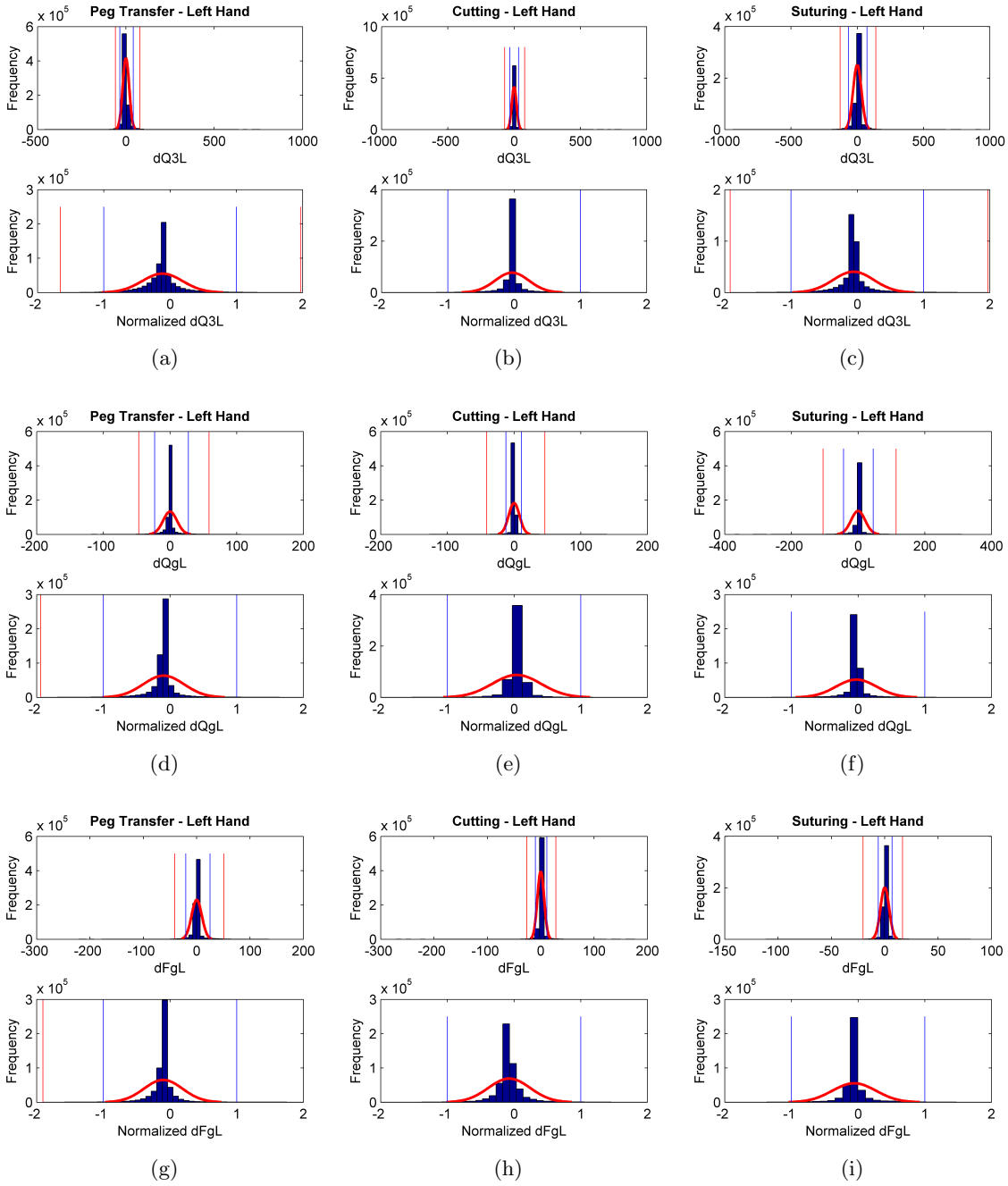


Figure A.9: Histograms of rotation and grasp rates for the left hand. The rotational velocity of the tool (in $^\circ/s$) about its axis is dQ3; dQg is the grasper angle (in $^\circ$/s) and Fg the grasp force (in Newtons/s). 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

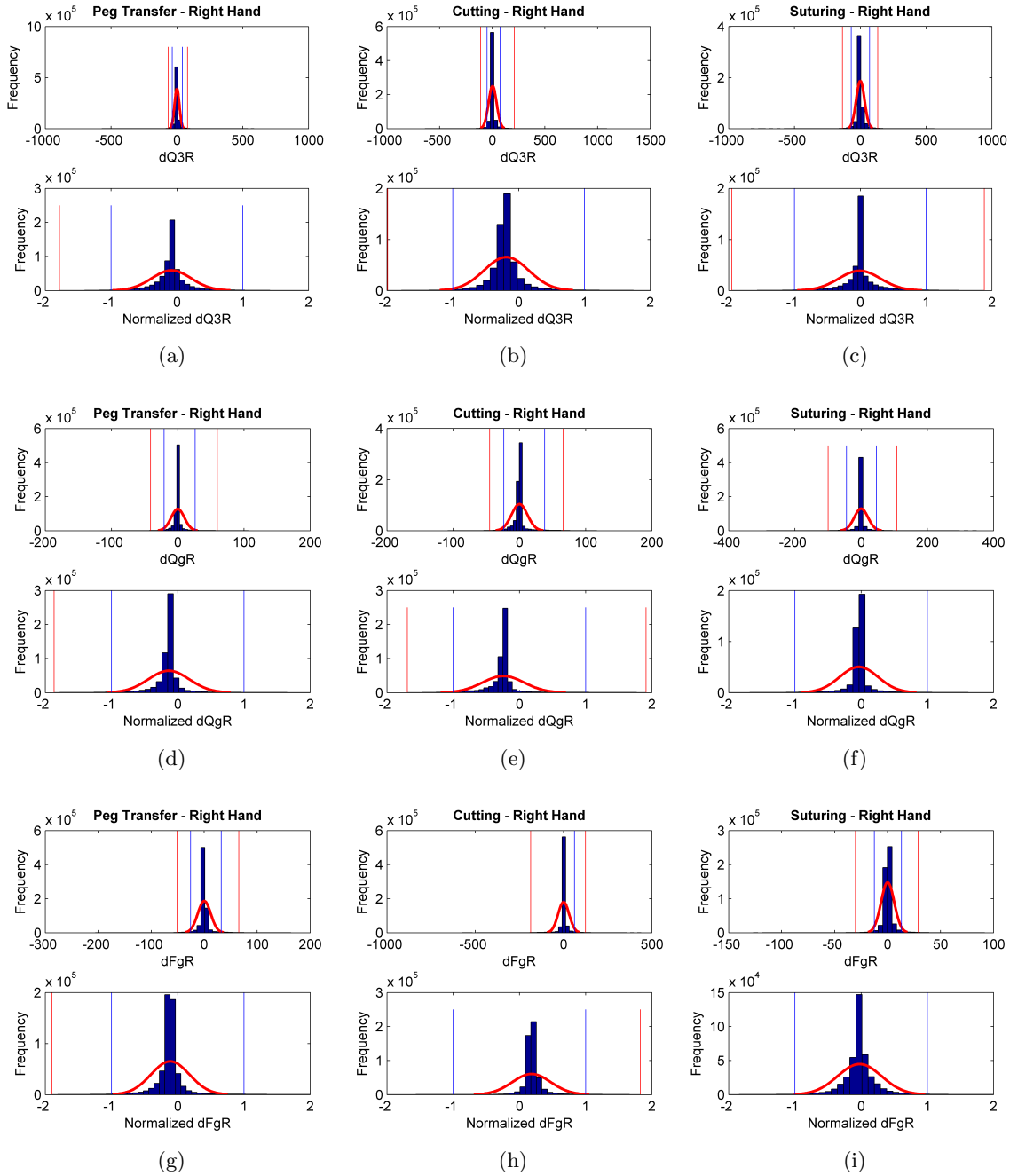


Figure A.10: Histograms of rotation and grasp rates for the right hand. The rotational velocity of the tool (in $^\circ/s$) about its axis is dQ3; dQg is the grasper angle (in $^\circ$/s) and Fg the grasp force (in Newtons/s). 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

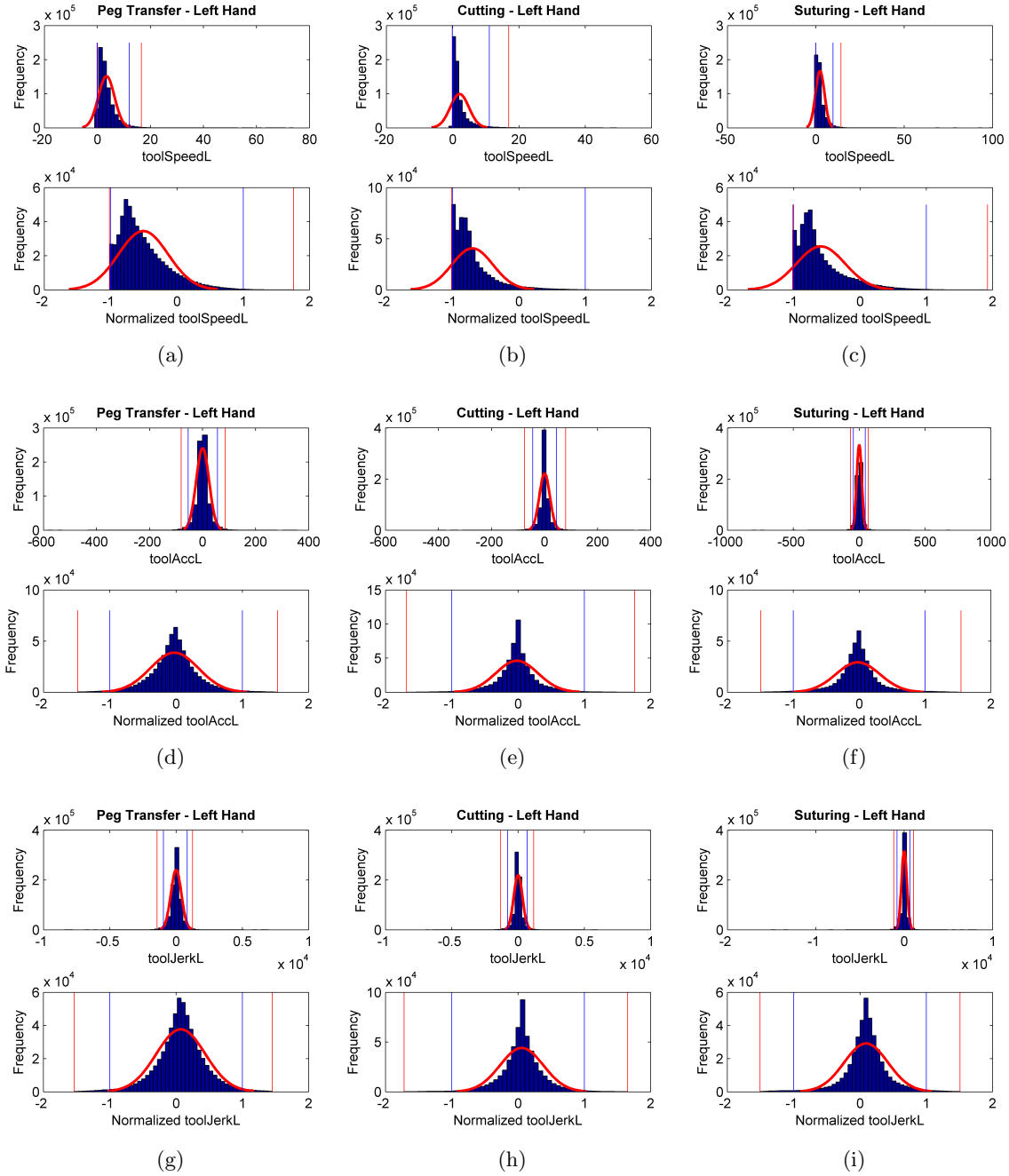


Figure A.11: Histograms of scalar motion rates for the left hand. The toolSpeed is the magnitude of velocity, toolAcc and toolJerk are the 1st and 2nd derivatives of toolSpeed respectively. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

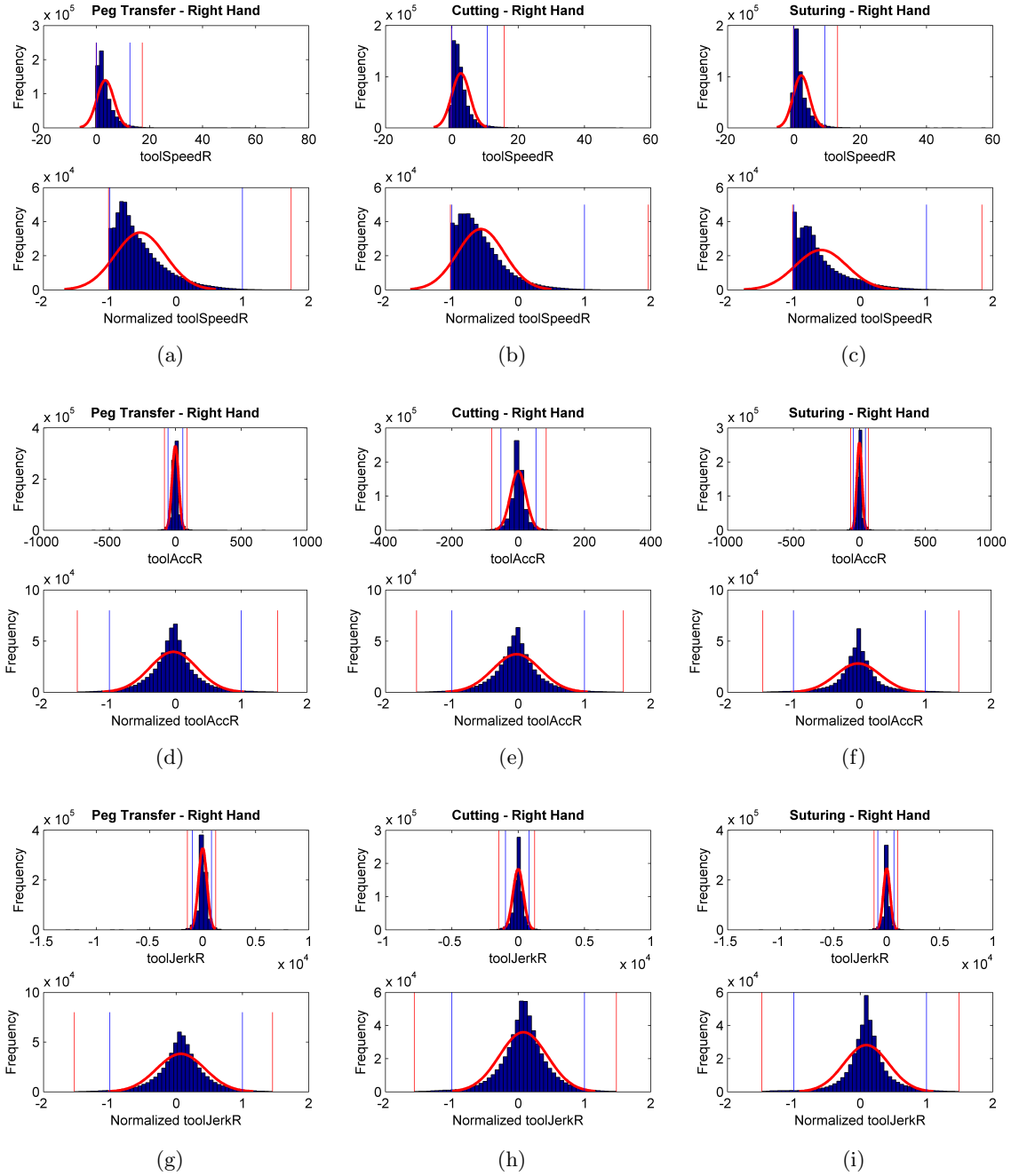


Figure A.12: Histograms of scalar motion rates for the rightt hand. The toolSpeed is the magnitude of velocity, toolAcc and toolJerk are the 1st and 2nd derivatives of toolSpeed respectively. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

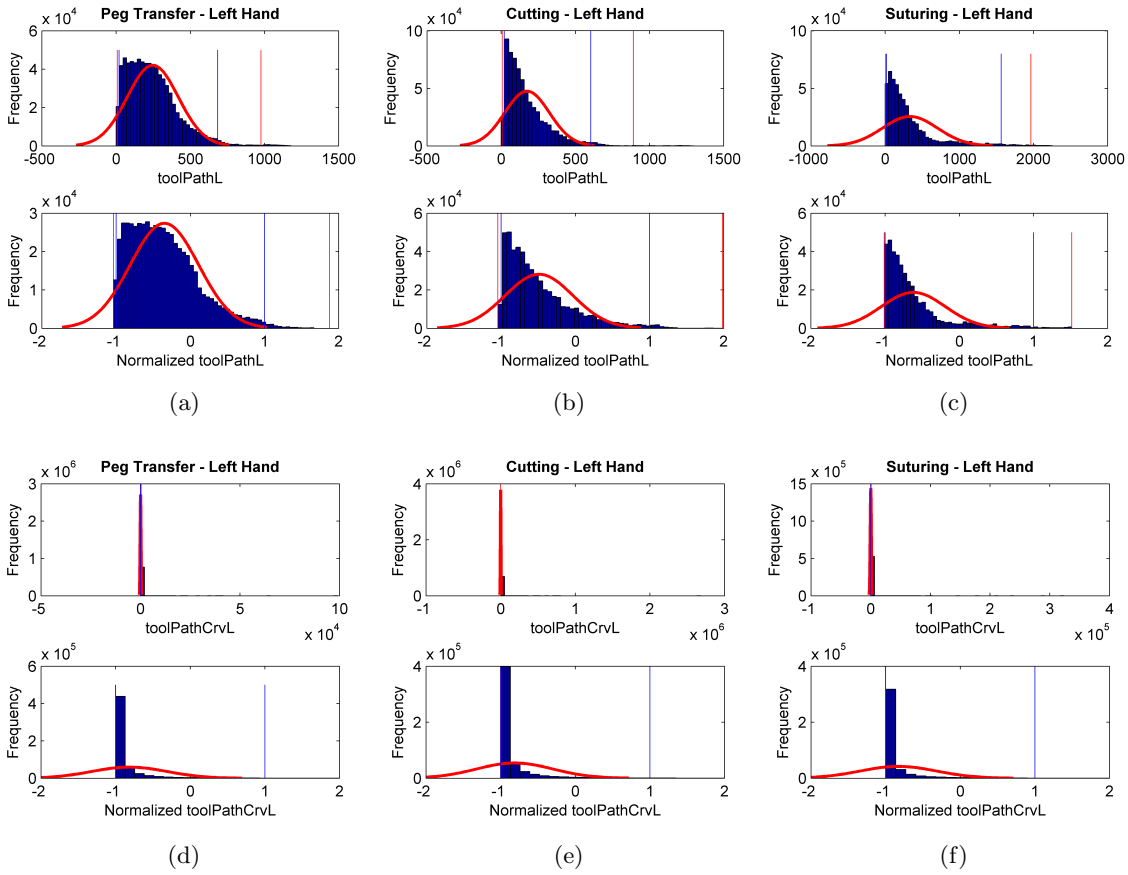


Figure A.13: Histograms of tool path length in (cm) and curvature in ($1/cm$) for the left hand. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

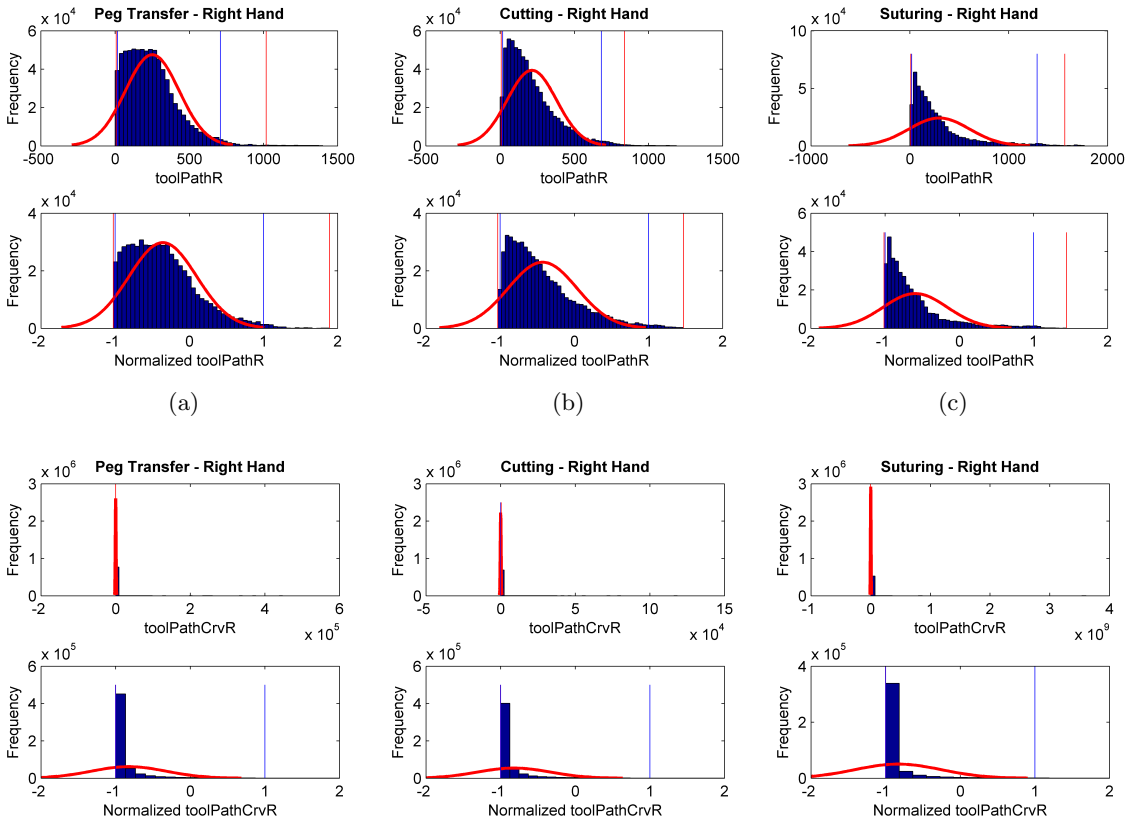


Figure A.14: Histograms of tool path length in (cm) and curvature in $(1/cm)$ for the right hand. 50 bins are used to create each histogram. The inner blue vertical lines indicate the 2nd and 98th percentiles which map to -1 and 1 in normalized sensor space. The outer red vertical lines indicate the 0.5 percentile threshold used to exclude outliers.

Appendix B

GETTING THE FILES: SURGENOME AND THE SVN REPOSITORY

In order to make this work reproducible and enable rapid implementation of future analysis, a content versioning system was adopted. All raw data, processed data, database construction, analysis scripts, generated plots and tables, and all documentation relevant to this project has been placed in a Subversion repository. This repository includes all work carried out in this thesis but also extends to a broader scope. It includes additional data and algorithms not analyzed or presented in this thesis, but highly relevant to the problem of surgical skill evaluation. This broader project, stored in the repository, is called Surgenome and totals over 100 gigabytes. It is available for checkout with Biorobotics lab credentials and any suitable Subversion client at:

`svn+ssh://surgenome.ee.washington.edu/srv/svn/surgenome`

Access to this repository may be available for research purposes upon request from the University of Washington Biorobotics lab or by emailing `tmk@uw.edu`.

Appendix C

ADDITIONAL FEATURE SELECTION TABLES

Table C.1: The list of features and their information gain for each task and each hand. The KL-divergence normalized by total entropy appears in parenthesis. This is effectively a distance measure average distance between distributions of the different classes. A value of 0 indicates almost no difference, suggesting poor discrimination.

	Peg Tranfer	**Cutting**	**Suturing**	**Overall**
Gr-L	0.0023 (0.00)	0.0178 (0.01)	0.0039 (0.00)	0.0240 (0.01)
Gr-R	0.0036 (0.00)	0.0053 (0.00)	0.0058 (0.00)	0.0148 (0.01)
dGr-L	0.0076 (0.00)	0.0012 (0.00)	0.0005 (0.00)	0.0093 (0.00)
dGr-R	0.0105 (0.00)	0.0013 (0.00)	0.0035 (0.00)	0.0153 (0.00)
Pos-dRta-Gr-L	0.0193 (0.00)	0.0524 (0.01)	0.0521 (0.01)	0.1238 (0.03)
Pos-dRta-Gr-R	0.0307 (0.01)	0.0083 (0.00)	0.0337 (0.01)	0.0726 (0.01)
Vel-dRta-Gr-L	0.0137 (0.00)	0.0296 (0.01)	0.0116 (0.00)	0.0549 (0.01)
Vel-dRta-Gr-R	0.0119 (0.00)	0.0087 (0.00)	0.0153 (0.00)	0.0359 (0.01)
SpdAccJrk-L	0.0103 (0.00)	0.0019 (0.00)	0.0034 (0.00)	0.0156 (0.00)
SpdAccJrk-R	0.0095 (0.00)	0.0024 (0.00)	0.0051 (0.00)	0.0169 (0.00)
Vel-dRta-Gr-dQg-L	0.0198 (0.00)	0.0320 (0.01)	0.0114 (0.00)	0.0631 (0.01)
Vel-dRta-Gr-dQg-R	0.0139 (0.00)	0.0089 (0.00)	0.0170 (0.00)	0.0398 (0.01)
SpdAccJrkCrv-dRta-Gr-L	0.0152 (0.01)	0.0027 (0.00)	0.0005 (0.00)	0.0184 (0.01)
SpdAccJrkCrv-dRta-Gr-R	0.0031 (0.00)	0.0012 (0.00)	0.0021 (0.00)	0.0063 (0.00)
SpdCrv-L	0.0035 (0.00)	0.0001 (0.00)	0.0003 (0.00)	0.0039 (0.00)
SpdCrv-R	0.0016 (0.00)	0.0004 (0.00)	0.0009 (0.00)	0.0029 (0.00)
SpdAcc-L	0.0154 (0.00)	0.0027 (0.00)	0.0031 (0.00)	0.0212 (0.01)
SpdAcc-R	0.0115 (0.00)	0.0028 (0.00)	0.0048 (0.00)	0.0191 (0.00)
SpdAcc-dRta-L	0.0169 (0.00)	0.0030 (0.00)	0.0038 (0.00)	0.0238 (0.01)
SpdAcc-dRta-R	0.0113 (0.00)	0.0035 (0.00)	0.0064 (0.00)	0.0212 (0.00)
dRta-L	0.0019 (0.00)	0.0000 (0.00)	0.0006 (0.00)	0.0025 (0.00)
dRta-R	0.0024 (0.00)	0.0013 (0.00)	0.0020 (0.00)	0.0057 (0.00)

Table C.2: The list of sequential features and their scores, sorted by best total information gain $I(Feat_{t+1}; Class|Feat_t)$

	Peg Tranfer	**Cutting**	**Suturing**	**Overall SUM**
Gr-L	0.0000 (0.00)	0.0003 (0.01)	0.0000 (0.00)	0.0004 (0.01)
Gr-R	0.0007 (0.00)	0.0013 (0.00)	0.0002 (0.00)	0.0022 (0.01)
dGr-L	0.0141 (0.00)	0.0499 (0.00)	0.0078 (0.00)	0.0718 (0.00)
dGr-R	0.0071 (0.00)	0.0052 (0.00)	0.0146 (0.00)	0.0268 (0.00)
Pos-dRta-Gr-L	0.0017 (0.00)	0.0016 (0.01)	0.0006 (0.01)	0.0039 (0.03)
Pos-dRta-Gr-R	0.0018 (0.01)	0.0055 (0.00)	0.0024 (0.01)	0.0096 (0.01)
Vel-dRta-Gr-L	0.0055 (0.00)	0.0167 (0.01)	0.0012 (0.00)	0.0234 (0.01)
Vel-dRta-Gr-R	0.0048 (0.00)	0.0082 (0.00)	0.0077 (0.00)	0.0207 (0.01)
SpdAccJrk-L	0.0173 (0.00)	0.0369 (0.00)	0.0168 (0.00)	0.0710 (0.00)
SpdAccJrk-R	0.0125 (0.00)	0.0317 (0.00)	0.0123 (0.00)	0.0566 (0.00)
Vel-dRta-Gr-dQg-L	0.0066 (0.00)	0.0185 (0.01)	0.0022 (0.00)	0.0274 (0.01)
Vel-dRta-Gr-dQg-R	0.0049 (0.00)	0.0103 (0.00)	0.0066 (0.00)	0.0218 (0.01)
SpdAccJrkCrv-dRta-Gr-L	0.0047 (0.01)	0.0140 (0.00)	0.0058 (0.00)	0.0245 (0.01)
SpdAccJrkCrv-dRta-Gr-R	0.0049 (0.00)	0.0107 (0.00)	0.0043 (0.00)	0.0200 (0.00)
SpdCrv-L	0.0035 (0.00)	0.0101 (0.00)	0.0048 (0.00)	0.0184 (0.00)
SpdCrv-R	0.0044 (0.00)	0.0080 (0.00)	0.0038 (0.00)	0.0162 (0.00)
SpdAcc-L	0.0158 (0.00)	0.0439 (0.00)	0.0070 (0.00)	0.0666 (0.01)
SpdAcc-R	0.0186 (0.00)	0.0350 (0.00)	0.0099 (0.00)	0.0635 (0.00)
SpdAcc-dRta-L	0.0159 (0.00)	0.0532 (0.00)	0.0216 (0.00)	0.0906 (0.01)
SpdAcc-dRta-R	0.0158 (0.00)	0.0543 (0.00)	0.0148 (0.00)	0.0850 (0.00)
dRta-L	0.0031 (0.00)	0.0074 (0.00)	0.0024 (0.00)	0.0130 (0.00)
dRta-R	0.0026 (0.00)	0.0055 (0.00)	0.0052 (0.00)	0.0133 (0.00)

VITA

Timothy Kowalewski was born in Wrocław, Poland. He completed his Bachelor of Science in Electrical Engineering at the University of Washington in 2004. He earned a Master of Science degree in Electrical Engineering from the University of Washington in 2009. In 2012 he earned a Doctor of Philosophy degree in Electrical Engineering, also from the University of Washington.