AUTOMATED AND ADAPTIVE SURGICAL ROBOTIC TRAINING USING REAL-TIME STYLE FEEDBACK THROUGH HAPTIC CUES

by

Marzieh Ershad

APPROVED BY SUPERVISORY COMMITTEE:

____
Ann Majewicz Fey, Chair

____
Robert Rege

____
Mark Spong

____
Mehrdad Nourani

To my loyal and beyond supportive family: my dad who has been a role model for me; and my mom and sister, who have always been encouraging and supportive and kept me going to accomplish this commission.

AUTOMATED AND ADAPTIVE SURGICAL ROBOTIC TRAINING USING REAL-TIME STYLE FEEDBACK THROUGH HAPTIC CUES

by

MARZIEH ERSHAD, BSc, MSc

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2019

ACKNOWLEDGMENTS

First and foremost I would like to thank my supervisor Dr. Ann Majewicz Fey for her great support, kindness, and guidance during my studies, for this accomplishment would have not been possible without her unlimited support and consideration. She would kindly devote her time and patiently provide advice throughout the study and assist me in different stages throughout this program. In addition to providing technical guidance for my work, she would patiently and compassionately provide advice and share experiences to help choose my career path. Her enthusiasm, care, and passion for her work and students always inspired me. I cannot express how fortunate and grateful I am to have had her as my supervisor and guide through this path.

I am extremely grateful to Dr. Robert Rege for being generous with his time for valuable discussions, great ideas and feedback throughout my work. I would like to thank my dissertation committee members Dr. Mark Spong and Dr. Mehrdad Nourani for taking the time to review and provide feedback on the content of my dissertation.

Thank you Dr. Schweitzer for chairing my defense.

I would like to thank my labmates at Herolab and all my friends who helped me through this journey.

Thank you to my loyal and beyond supportive family: my dad, mom, and sister, adorable Amir Mojtaba, and Aala and most recently Ali. I cannot express my gratitude and appreciation for all your love and encouragement. Dad, you have always been my role model in every step. Mom, thank you for all your encouragement and prayers. Maryam, thank you for your never ending motivation and support.

November 2019

AUTOMATED AND ADAPTIVE SURGICAL ROBOTIC TRAINING USING REAL-TIME STYLE FEEDBACK THROUGH HAPTIC CUES

Marzieh Ershad, PhD
The University of Texas at Dallas, 2019

Supervising Professor: Ann Majewicz Fey, Chair

Evaluating robotic surgical skills efficiently and effectively while providing surgical residents with intuitive and easy to understand feedback is an important problem. This dissertation focuses on the design, implementation, and experimental validation of a style-based surgical skill assessment haptic guidance and training method that addresses key challenges related to surgical training systems.

The first topic of this dissertation is finding differences in the style of movement of an expert versus a novice while performing a task. We leverage crowd-sourced assessment to define the stylistic attributes of surgical motion and find features within movement and physiological data to classify style. We show through an experimental human subject study that stylistic behaviors can distinguish between different levels of expertise and have more discriminating power than standard scoring systems for surgical robot simulators.

The second topic of this dissertation is to automate the stylistic behavior detection in near-real-time to be able to provide feedback to the user during task performance. For this purpose, a data driven model was developed and tested for each proposed stylistic behavior adjective. These models enable the detection of a deficiency as soon as it occurs.

The final step in a training system is to provide meaningful feedback to the user. Since each individual learns in a different way, it is important to customize training for each individual. The third topic of this dissertation is to provide the trainees with adaptive feedback based on their stylistic performance. For this purpose, the deficiency of a user's stylistic behavior is detected and is used to provide an appropriate force feedback to help correct movement. Three types of force feedback for each stylistic behavioral adjective are evaluated in this study and the best feedback for each style is found through an experimental human subject study.

The work in this dissertation provides a groundwork for adaptive, automatic and real-time surgical skill training. This method can also be extended for coaching in other areas other than surgical applications, such as sports.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Problem definition

Robotic surgery is becoming increasingly popular as the robot provides improved stereo vision, dexterity, and precision for more complex procedures and in confined regions of the body compared to open surgery and minimally invasive laparoscopic surgery (MIS) (Morris, 2005). Telesurgery is another advantage of robotic surgery. The translation of movements from the master side manipulator to the patient side manipulator in surgical robots (Fig. 1.1), enables performing operations and assisting other surgeons without the need to be present at the surgical site.

Due to the unique techniques required in robotic surgery, surgeons must acquire several technical skills different from open surgery or MIS. Thus, adequate training of surgeons is necessary for acquiring mastery in performing surgical procedures using the robot. The lack of tactile sensations and force feedback present in open surgery, and to a lesser extent in MIS, make surgical robotic training a challenging task, due to increased dependence on visual perceptual cues and reliance on proprioception.

Patient safety is the major priority in surgical procedures. As surgical skill is directly associated with surgical complication rates, efficient training is crucial for achieving successful outcomes (Birkmeyer et al., 2013). Surgical training for technical skills has been built based on William Halsted's method of See One, Do One, Teach One (Carter, 1952); however, this has caused controversy in the field due to concerns for patient safety since in this method, learning is done through practicing on real patients; thus, different approaches have been proposed to alleviate this matter.

For robotic surgical training, surgical simulators are being used widely due to moving training outside of the operating room and providing a patient free learning environment. Simulators' automatic assessment and scoring mechanisms also enable objective assessment of trainees' perfor-

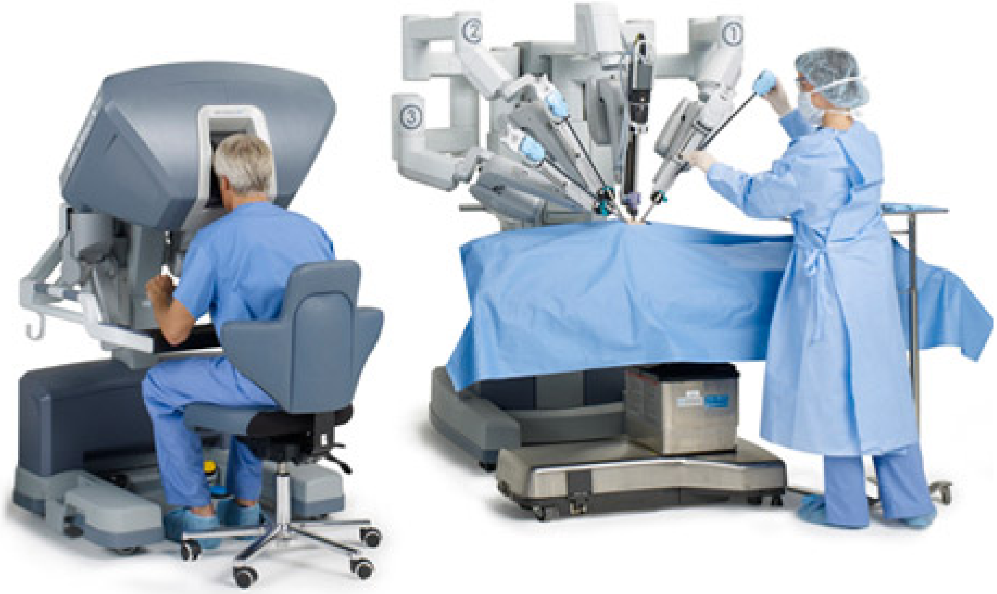
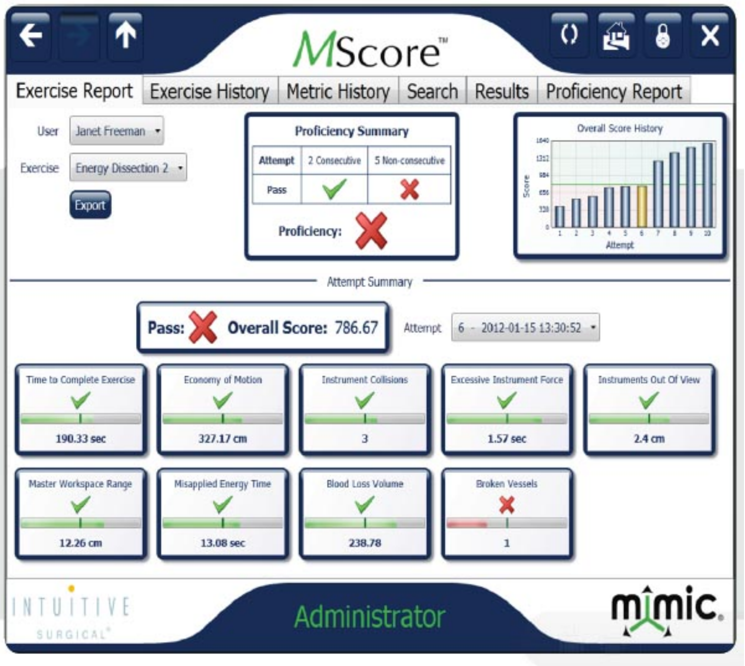


Figure 1.1: The da Vinci surgical robot, Intuitive Surgical Systems

mance; however, some advancements in these simulators can improve training and ensure effective results. Some potential advancements are discussed here.

A crucial step in surgical training is the evaluation of technical skills learned by trainees. Simulators use recorded hand movement data to calculate descriptive metrics and provide feedback to trainees based on their performance. Examples of these metrics from the *da Vinci*® *Skills Simulator*^{TM} include time to complete the task, economy of motion, instrument collision etc. (Fig. 1.2) These quantitative metrics do not imply how one must move differently to behave more like an expert. ***Focusing on differences in the quality and the style of movement of an expert vs. a novice***, could provide rich insight on the differences in performance between the two, and can be used to guide novices to adopt expert-like styles in their movements.

Simulators have automated the evaluation of surgical skill by providing objective assessment and eliminating the need for direct observation from experts which is time and resource intensive. They provide descriptive metrics to the trainee at the end of the task as a feedback for their performance; however, the limited number of metrics calculated by simulators may not represent the



User Janet Freeman

Exercise Energy Dissection 2

Proficiency Summary

Attempt	2 Consecutive	5 Non-consecutive
Pass	✓	✗

Proficiency: ✗

Attempt Summary

Pass: ✗ Overall Score: 786.67

Attempt 6 - 2012-01-15 13:30:52

Time to Complete Exercise	✓	190.33 sec
Economy of Motion	✓	327.17 cm
Instrument Collisions	✓	3
Excessive Instrument Force	✓	1.57 sec
Instruments Out Of View	✓	2.4 cm
Master Workspace Range	✓	12.26 cm
Misapplied Energy Time	✓	13.08 sec
Blood Loss Volume	✓	238.78
Broken Vessels	✗	1

Figure 1.2: The *da Vinci*® *Skills Simulator*^{TM} scoring, Intuitive Surgical Systems

skill deficiencies of the surgeons efficiently. In addition, these metrics are calculated based on the whole duration of the performed task, and scores are not provided to the trainee until the end of the simulation task. ***Automatic surgical skill modeling and evaluation based on data driven models that make use of the whole kinematic data set***, instead of considering limited features selected by expert knowledge, enable extraction of underlying information in the movement data. Furthermore, ***real-time error detection could provide feedback to the user as soon as errors are made***.

Since each individual learns in a different way, customizing training to focus on each individual's technical skills throughout task performance leads to more efficient learning. ***Adaptive training is another potential improvement of simulator-based training***. In addition, descriptive features given to the user at task completion by current simulators do not provide information on how to modify movement to improve performance. Feedback can be provided to the user during task performance through different types of feedback including audio, vision, haptic etc. ***Provid-***

ing trainees with proper feedback relevant to the deficiency in their movement in real-time, can improve trainees' movements, and guide them towards more expert-like behavior.

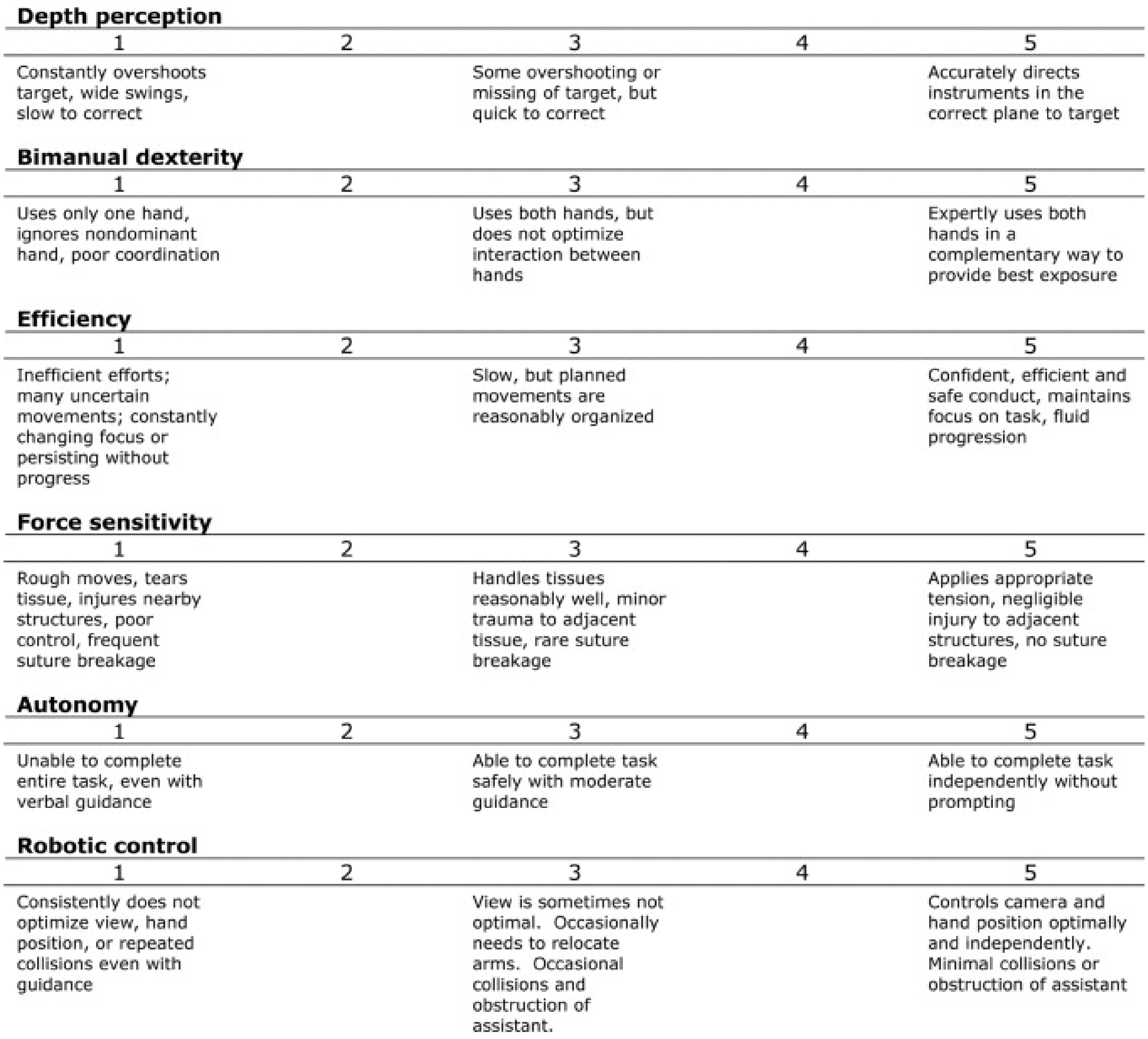
1.2 Prior Work

In this section, first the most widely used tools and methods in surgical skill assessment found in the literature are discussed. Then different methods and models for automating the evaluation process and differentiating technical skill levels are presented. Finally different methods for applying feedback to the user, as well as different types of feedback are included.

1.2.1 Surgical Skill Evaluation Tools and Techniques

Direct Observation

Traditionally, technical skill assessment in surgery was done through directly observing trainees' performance by surgical experts and providing self-reported procedure logs, with no specific criteria. Global ratings and checklists provide a structured method of assessment which improve the traditional direct observation method. Since different types of surgery require different skill sets and techniques, they require different evaluation and assessment metrics associated with them (Peters et al., 2004; Morris, 2005; Vassiliou et al., 2010). Thus, global ratings and check lists have been proposed for all major types of surgery to focus on criteria important to that type of surgery. Examples of checklist for three major types of surgeries are objective structured assessment of technical skills (OSATS) for open sugery (Martin et al., 1997), global operative assessment of laparoscopic skills (GOALS) for minimally invasive laparoscopic surgery (Vassiliou et al., 2005), and global evaluative assessment of robotic skills (GEARS) for robotic surgery (Goh et al., 2012). The GEARS checklist is shown in Fig. 1.3.



Depth perception

1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target

Bimanual dexterity

1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure

Efficiency

1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression

Force sensitivity

1	2	3	4	5
Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage

Autonomy

1	2	3	4	5
Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate guidance		Able to complete task independently without prompting

Robotic control

1	2	3	4	5
Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant

Figure 1.3: Global Evaluative Assessment of Robotic Skills (Goh et al., 2012)

While these methods have been significantly studied and validated, they require considerable resources and time from a team of experts to evaluate trainee performance. To reduce the overall burden of direct observation-based surgical skill assessment methods, various non-direct methods for surgical assessment have been proposed and some of the most popularly used ones will be discussed here.

Dexterity Analysis

One of the most widely used methods in this field is dexterity analysis which focuses on quantitative analysis of surgeon's hand or tool motion for evaluating technical skills. Quantifying the assessment of surgical skill is done through recorded motion data which provide descriptive statistical features such as time, path length, number of movements, speed, and motion smoothness to differentiate expertise levels (Grober et al., 2010; Schijven et al., 2002; Derossis et al., 1998; Liang et al., 2018). In motion analysis these descriptive statistical features are used to evaluate performance. Motion analysis devices, such as the Imperial College Surgical Assessment Device (ICSAD) (Datta et al., 2001), have been developed which autonomously measure the total distance moved by each hand, number of movements, and the total time using electromagnetic trackers. ICSAD has been validated for use in open surgery as well as laparoscopic surgical simulation (Datta et al., 2001; Taffinder et al., 1999). Other dexterity analysis methods for skill assessment in MIS include Advanced Dundee Endoscopic Psychomotor Tester (ADEPT) (Francis et al., 2002), an objective scoring method for evaluating endoscopic task performance, and McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) (Fried et al., 2004; Acosta and Temkin, 2005). For the assessment of robotic surgical skills, most studies have been carried out on the *da Vinci® Skills Simulator™* and the *da Vinci® Surgical System™* (Intuitive Surgical, Inc., CA, USA). Stergiou et al. studied the ability of objective kinematic measurements including task completion time, distance travelled, speed, curvature, and relative phase, in distinguishing expert and novice performance on da vinci robot (Stergiou, 2008).

These metrics provide useful information about trainees' technical skills; however, they are limited, selected by expert opinion and could be missing additional key features.

Crowd Source Assessment

Crowd-worker evaluation has recently gained a lot of attention in surgical skill assessment due to its efficiency and reliability. Crowd-sourcing is a method where workers complete simple online tasks for a small monetary compensation. The cost-effectiveness and time efficiency of this method holds great promise for many labor-intensive applications such as language processing and information retrieval, and its reliability has been the topic of several research publications (Callison-Burch, 2009; Alonso et al., 2008; Brabham, 2010; Blohm et al., 2013). For surgical educational research, several studies have shown the effectiveness of crowd-sourcing in evaluating surgical technical skills. After the developement of CSATS, a crowd-sourcing method for surgical skill evaluation, by Chen et al. (Chen et al., 2014), several studies have shown the effectiveness of crowd-sourced assessment of surgical skills in various procedures including: cricothyrotomy (Aghdasi et al., 2015), robotic radical prostatectomy (Ghani et al., 2016), robotic urologic surgery (Holst et al., 2015), optical biopsy for bladder cancer diagnosis (Chen et al., 2016), general surgery residency training (Deal et al., 2016), and basic laparoscopic urologic skills (Kowalewski et al., 2016a), to name a few. In environments where surgical skill motions can be captured only through video, crowd-sourcing can be an effective evaluation method (Kowalewski et al., 2016b).

However, despite its time-efficiency in evaluating performance compared to expert evaluation, crowd sourcing is still an offline evaluation method and can not provide the trainee with online feedback.

1.2.2 Automating Surgical Skill Assessment using Feature Extraction and Skill Modeling

The tools and methods discussed in the previous section, require human interaction to either directly observe and evaluate performance, or to extract meaningful metrics that are then used for evaluation. Automating the evaluation process has been the focus of many studies in the literature, since it increases the time and cost efficiency. Recently, simulators have found their way in surgical training practices. They can record and store large amount of data that can later be processed. Data science is an emerging field in surgical training and evaluation which is used to extract useful information from these large amount of data, and assists with objective computerized surgical coaching and active learning (Maier-Hein et al., 2017). The most common methods for automatic surgical skill assessment is discussed here.

Machine Learning Models using Descriptive Features

In addition to directly using the global movement features discussed in section 1.2.1, these descriptive movement features are used as input to machine learning models to automatically distinguish between different levels of expertise (Hung et al., 2018; Fard et al., 2018).

As mentioned earlier, these hand-crafted features require expert and domain-based knowledge, lack flexibility, and are time consuming and labor intensive to compute. In addition, they may not be able to capture underlying information in the rich kinematic data as they only focus on a limited set of movement characteristics. Also, due to the stochastic nature of human motion, descriptive features alone may not sufficiently capture a user's skill level in repetitive tasks (Reiley and Hager, 2009b).

Statistical models

To take into consideration the stochastic property of time-series motion data, statistical modeling of time-series data and unsupervised learning methods have been widely used for activity recognition and extracting procedural knowledge for skill assessment. Studies focusing on skill assessment using procedural knowledge, usually rely on surgical activity recognition and skill assessment at the subtask level (Lin et al., 2005, 2006; Reiley et al., 2011) and include different granularity levels such as motion, activities, steps, phases, and procedure. Some studies focus on detecting surgemes which are low-level surgical activities (Lea et al., 2015), while others focus on higher level activity detection at the maneuver level. These activities are built up of a sequence of surgemes and can be more useful for providing feedback to the users at this level (DiPietro et al., 2016). They provide the ability to analyze simple movements for quantitative assessment of surgical skill (Uemura et al., 2016). The goal of skill modeling is to uncover and measure the underlying characteristics of skill hidden in measurable motion data. Statistical models can be used to describe the sequence of motions for automatic task recognition. Hidden Markov Models (HMMs) which determine hidden parameters from observed data, are one of the most widely used statistical models in surgical skill assessment. Reiley et al. performed gesture classification using HMM to model each gesture in a trial after applying linear discriminant analysis (LDA) to the continuous kinematic data for dimensionality reduction (Reiley et al., 2008). In another study Reiley et al. applied an HMM on discrete spectrum features extracted from a short time Fourier transform of the velocity features in the kinematic data with the goal of evaluating surgical skill in tasks and sub-tasks (Reiley and Hager, 2009b). In (Reiley and Hager, 2009a), they studied the relation between subtasks and skill using the same features. Varadarajan et al. accomplished a surgical recognition task (jointly identifying gesture boundaries and assigning gesture labels) by applying an HMM to features derived from modeling the kinematic data (after an optional LDA) using Gaussian mixture models (GMMs) (Varadarajan et al., 2009). Their topology for HMMs was entirely derived form data. These studies show that statistical models such as HMMs are able to recognize skill level and surgemes from motion data. Pattern discovery methods can also be used to identify surgical procedural sequences that may be useful for surgical skill evaluation in the future (Huaulmé et al., 2017).

These gesture classification methods require large amount of labeled data which is obtained through manual annotations thus, are time consuming and labor intensive. Furthermore, these methods are task dependent as they focus on different gestures in a specific task and do not focus on the general style of movement during task performance.

Deep Learning

Deep learning, a popular method in data science that extracts features from raw data and is based on learning data representations, is widely being used in surgical skill assessment for automating the assessment. It has been used with different data modalities including kinematic data, video, imaging, and functional neuroimaging data for the purpose of surgical skill assessment. Wang et al. used kinematic data from robot end effectors to train deep networks for assessing surgical skill levels (Wang and Fey, 2018). Jin et al. used deep learning to track tool movement in laparoscopic surgery from video data and extracted metrics such as tool usage patterns and movement range for surgical skill assessment (Jin et al., 2018). Law et al. used deep networks to train a robot instrument tracker, then classified technical skill level using the movement of the robotic instruments (Law et al., 2017).

All the methods mentioned above are able to differentiate different levels of expertise; however, they do not focus on the structural differences in the quality of movement between the performance of a novice relative to an expert in order to be able to provide them with relevant and more meaningful feedback.

1.2.3 Adaptive and Customized Training Feedback

Adaptive technology is introduced into training devices to develop customized user-specific training which results in more effective learning. An adaptive system can be known as a supervisor that instructs each trainee based on his/her unique performance and provides specific instructions on how to proceed or adjusts the task for each individual to ensure best results for each trainee. An adaptive system consists of a control loop that detects changes in the output from a desired point, this can be done using machine learning approaches. Feedback is then applied to modify the response to move it towards the desired performance (Vaughan et al., 2016). Thus, adaptive systems require constant monitoring and measurement of performance, an adaptive variable, and a methodology to adjust the variable to enhance performance (Kelley, 1969).

In adaptive training, the user's performance is evaluated based on a specific criteria and the next training step is adapted accordingly. Task difficulty level is one element of focus in adaptive training. The difficulty of the task is updated based on user's performance to adjust the level of challenge and enhance learning. This has been studied in digital games (Charles et al., 2005).

The stimulus or the type of feedback provided to the user (such as visual, audio, or haptic feedback) is another element of focus in adaptive training in which the type of feedback is adapted based on the user's performance. In customized learning using haptic feedback, the user's movement is modified to match a desired movement.

To study the effect of haptic feedback on user's performance, two types of haptic feedback are typically used: reflective feedback and guidance feedback. The former provides the user with a feeling of touch and force in interacting with an object in environments where these sense are missing. In virtual environments this is done through haptic rendering. The latter provides users with haptic cues, and assists the user in correcting movements to improve performance.

Reflective Feedback

Haptic feedback in VR simulators improves training (Basdogan et al., 2004; Tholey et al., 2005; Cosman et al., 2002). Many studies show that the loss of haptic feedback in MIS and RMIS can cause errors and thus, safety issues (Joice et al., 1998; Okamura, 2004; Tang et al., 2005; Morino et al., 2006; Gutt et al., 2004; Antoniou et al., 2011). The lack of haptic feedback (both force and tactile) causes an inappropriate level of force applied to the tissue (Enayati et al., 2016). Incorporating haptic feedback can alleviate this (Tavakoli et al., 2005), and result in reduced performance error (Demi et al., 2005; Ortmaier et al., 2007; Wagner et al., 2002).

Tactile feedack decreases the force applied to the tissue and hence reduces tissue damage (Wottawa et al., 2016; King et al., 2009). Force feedback can improve psychomotor skills in the early training phase (Van der Meijden and Schijven, 2009; Ström et al., 2006; Lamata et al., 2006; Kim et al., 2003). Yanping et al. developed a simulator with force feedback for learning bone-sawing skills. The results indicate a higher improvement in novice performance compared to expert's (Lin et al., 2014). An experiment was conducted to determine if the cognitive load in the early stages of learning basic laparoscopic skills is too much to be able to undertake haptic cues by novice surgeons. The results indicate that haptic feedback improves performance accuracy even while cognitively loaded (Cao et al., 2007). Vibrotactile feedback can be used to provide information such as surface texture or contact force. An in vitro study by McMahan et al. shows that this type of feedback is preferred by users yet it does not improve performance in terms of applied force or completion time (McMahan et al., 2011); however, an in vivo study showed that vibration feedback can increase the immersion and situational awareness of surgeons (Bark et al., 2013).

This type of reflective feedback, though proving to be helpful in providing the user with a feeling of touch and force in teleoperated environments where these senses are missing, does not provide any cues to the user on how to modify movement to improve performance. Guidance feedback which addresses this issue is discussed in the following section.

Guidance Feedback

Haptic guidance can enhance learning new motor skills in robotic environments where an instructor is not present to guide the user on how to modify his/her movement. Different studies have shown the effectiveness of haptic feedback in developing motor skills (Boulanger et al., 2006; Solis et al., 2003; Feygin et al., 2002), as well as providing movement guidance (Stanley and Kuchenbecker, 2012; Norman et al., 2014). A common type of training motor skills using haptic feedback, is transferring expert skills in which an expert's movements are recorded and played back to train a novice (Yang et al., 2008). However, (Gibo and Abbink, 2016) showed that haptic feedback can help discover new movement strategy rather than following a specific trajectory or enforcing a specific movement. They provided the subjects with an environment to explore different types of movement using haptic feedback and adopt the best strategy. Haptic disturbances which suggests disturbing the movemnet instead of guding the user can also improve motor skills (Lee and Choi, 2010).

While all these methods prove the effectiveness of haptic feedback in movement guidance, they do not focus on performance feedback. Jantscher et al. designed and implemented a framework that provides vibrotactile feedback method based on movement smoothness. They proved that the smoothness-based feedback improved accuracy compared to trajectory based feedback methods (Jantscher et al., 2018). They provided the subjects with vibrotactile cues with a degree of pleasantness relative to their performance; however, results in the literature show scenarios in which force feedback assists the user in performing a task, yet is perceived negatively by the human user (Gwilliam et al., 2009), and other scenarios in which force feedback does not improve performance, yet is preferred (McMahan et al., 2011). These results indicate that objective performance metrics and subjective user response surveys may not be sufficient for understanding the intuitiveness of a control interface.

1.3 Contribution

In this dissertation, we propose novel methods to address the problems mentioned above. We develop the groundwork for building a surgical robotic training system for autonomous assessment of user's performance, while providing real-time, automatic, understandable, and customized feedback to match each individual's requirements.

1.3.1 Proposing a Novel Method in Surgical Skill Assessment Based on Stylistic Behavior

We believe that there is an underlying behavioral style that distinguishes experts from novices. We leverage the idea that perception of expertise is an instinctual response. This means a casual observer, who may not know anything about the objectives and goals of the observed behavior, can still distinguish between a novice and an expert by observing their movements and their performance. Inspired by this fact and recent work on crowd source assessment of skills, we built a lexicon of surgical expertise. The lexicon was chosen based on traits important in robotic surgery defined by the six skill domains in GEARS, and with consultation with surgical faculty.

1.3.2 Finding Quantitative Metrics Associated with the Styles of Movement

After presenting the idea that the style of movement can define the proficiency of the subject, an important research question was: what quantitative metrics correspond to expert-like style? We sought to identify metrics embedded in the trainee's style of movement to be able to communicate with the trainee, through more understandable language, on how to improve deficiencies in surgical style.

1.3.3 Automatic and Real-time Skill Level and Stylistic Behavior Detection

We automated the identification of users' performance based on their stylistic behavior. We proposed a method that automatically captured underlying structures in raw movement data in near real time, while eliminating the need for complex feature engineering and expert knowledge.

1.3.4 Adaptive Training and Feedback Algorithms

We introduced a framework that provided the user with customized force feedback. Based on the users performance, a deficiency in the movement style was detected with reference to a desired output (expert knowledge), a relevant force feedback was applied to the user to correct the movement. We evaluated the effectiveness of automated force feedback training system through an experimental study.

1.4 Dissertation Overview

In this chapter, we presented the motivation for our research in automatic and customized robotic surgical training. We also presented the relevant prior work for this research.

Chapter 2 describes our novel method in surgical skill assessment using stylistic behavior. A lexicon of surgical expertise is proposed and evaluated through crowd-workers, who rated the performance of the users. Quantitative metrics associated with the contrasting adjectives in the lexicon are found through an extensive search. We extract several metrics from sensor measurements and find the best metric for each word pair based on the best correlation with the crowd source assessment, and also correlation with expert ratings. The metrics are then validated by their ability to distinguish among different levels of expertise.

The metrics extracted in Chapter 2 require a lot of knowledge to process the data and find the correct features. In Chapter 3 we focus on automating the surgical skill level assessment as well as the evaluation of stylistic behavior in near real-time, while eliminating the need for feature engineering and manually extracted features from motion data. A dictionary learning algorithm is developed to detect the good or poor performance for each stylistic behavior (each word in the lexicon presented in Chapter 2) using the raw kinematic data. A separate dictionary is learned and evaluated for each word, and the results of the performance detection is presented in this chapter.

In Chapter 4, we conduct an experiment to examine the effect of three different types of force feedback (spring, damping, and spring plus damping) on the improvement of the user's stylistic behavior. The users perform a simulated needle steering task using a haptic device. A deficiency in the user's stylistic behavior is detected using the algorithm developed in chapter 2. One of three types of feedback is applied to the user once a deficiency in the style is detected. All types of feedback are tested for all movement styles, and the the results for the best type of feedback for the improvement of each style is presented.

Finally in Chapter 5, we summarize our main results and present topics for future work in robotic surgical training that build upon ideas presented in this dissertation.

CHAPTER 2

ASSESSMENT OF SURGICAL SKILL USING STYLES OF MOVEMENT

2.1 Introduction

In this chapter, we propose and evaluate a novel method for surgical skill assessment using styles of movement. Many surgical assessment metrics have been developed to identify and rank surgical expertise; however, some of these metrics (e.g., economy of motion) can be difficult to understand and do not coach the user on how to modify behavior. We aim to standardize assessment language by identifying key semantic labels for expertise.

In recent years, the field of data-driven identification of surgical skill has grown significantly. Methods now exist to accurately classify expert vs. novice users based on motion analysis (Howells et al., 2008), eye tracking (Ahmidi et al., 2010), and theories from motor control literature (Nisky et al., 2014), to name a few. Additionally, it is also possible to rank several users in terms of expertise through pairwise comparisons of surgical videos (Malpani et al., 2014). While all these methods present novel ways for determining and ranking expertise, an open question remains: how can observed skill deficiencies translate into more effective training programs? Leveraging prior work showing the superiority of verbal coaching for training (Porte et al., 2007), we aim to develop and validate a mechanism for translating conceptually difficult, but quantifiable, differences between novice and expert surgeons (e.g. more directed graphs on the known state transition diagrams (Reiley and Hager, 2009b) and superior exploitation of kinematic redundancy (Nisky et al., 2014)) into actionable, connotation-based feedback that a novice can understand and employ.

A great musician, an all-star athlete, and a highly skilled surgeon share one thing in common: the casual observer can easily recognize their expertise, simply by observing their movements. These movements, or rather, the appearance of the expert in action, can often be described by words such as fluid, effortless, swift, and decisive. Given that our understanding of expertise is so innate and engrained in our vocabulary, we seek to develop a lexicon of surgical expertise through combined data analysis (e.g., user movements and physiological response) and crowd-sourced labeling (Chen et al., 2014; Malpani et al., 2014). The central hypothesis to this study is that human perception of surgical expertise is not so much a careful, rational evaluation, but rather an instinctive, impulsive one.

Prior work has proposed that surgical actions, or surgemes (e.g. knot tying, needle grasping, etc.) are ultimately the building blocks of surgery (Lin et al., 2006). While these semantic labels describe the procedural flow of a surgical task, we believe that surgical skill identification is more fundamental than how to push a needle. It is about the quality of movement that one can observe from a short snapshot of data. Are the movements smooth? Do they look fluid? Does the operator seem natural during the task? The hypothesis that expertise is a universal, instinctive assessment is supported by recent work in which crowd-workers from the general population identified surgical expertise with high accuracy (Chen et al., 2014). Thus, the key to developing effective training strategies is to translate movement qualities into universally understandable, intuitive, semantic descriptors.

We believe that there is an underlying behavioral style that distinguishes experts from novices. We leverage the idea that perception of expertise is an instinctual response and propose an approach to assess surgical skill level based the behavioral style of the user.

This chapter describes three studies. The first section 2.2 describes and evaluates our proposed lexicon for evaluating surgical skill. The second section 2.3 focuses on finding metrics associated with the words in the lexicon. The third section 2.3 describes a method for automating the GEARS scoring in robotic surgery using our lexicon.

2.2 Defining and Validating a Lexicon for Surgical Style

Inspired by studies showing the benefit of verbal coaching for surgical training (Porte et al., 2007), we selected a set of semantic labels which could be used to both describe surgical expertise and coach a novice during training. The lexicon of surgical expertise was proposed by consulting robotic surgical experts and is also based on skill components important in robotic surgery, as mentioned by the GEARS assessment tool. The six preliminary word pairs chosen and their corresponding data metrics are listed in Table 2.1. While by no means exhaustive, this study aims to determine if this preliminary set of stylistic adjectives could be used for surgical skill assessment, by associating these adjectives with measurable data metrics. As this was a preliminary study, we selected adjective pairs that were commonly used and also had some logical data metric that could be associated with them. For example, crisp/jittery can be matched with a jerk measurement, and relaxed/tense can be matched to some metric from electromyography (EMG) recordings. Galvanic skin response (GSR) is another physiological measurement anxiety (Critchley et al., 2000), thus serving as a basis for a calm/anxious word pair. The choice of word pairs and the corresponding metric is not unique; however, in this section, we simply aim to determine whether or not these word pairs have some relevance in terms of surgical skill evaluation. This study was published in (Ershad et al., 2016).

Table 2.1: Lexicon of Stylistic Behavior and Proposed Metrics

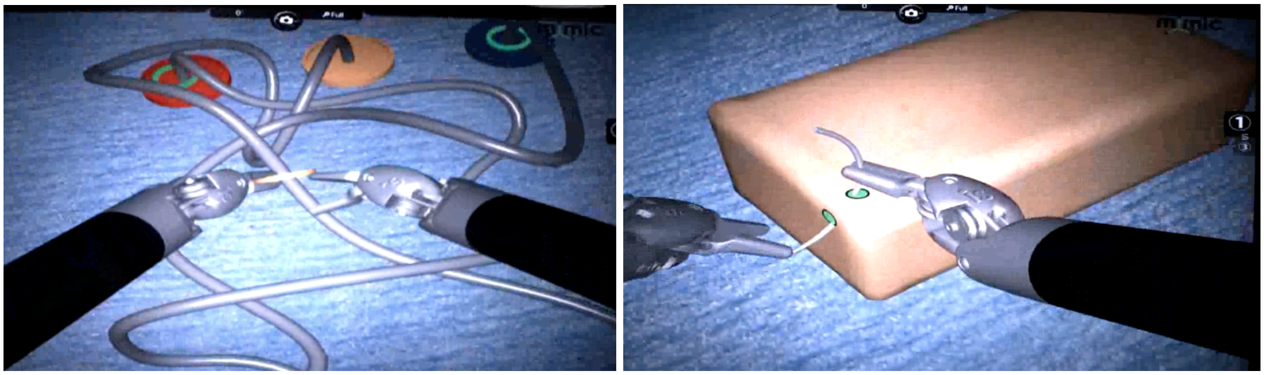
Positive Adjective	Positive Metric	Negative Metric	Negative Adjective
Crisp	High mean jerk	Low mean jerk	Jittery
Fluid	Low ang. velocity var	High ang. velocity var	Viscous
Smooth	Low acceleration var	High acceleration var	Wavering
Swift	Short completion time	long completion time	Sluggish
Relaxed	Low normalized EMG	High normalized EMG	Rough
Calm	Low GSR event count	High GSR event count	Tense

Our hypothesis is that crowd-rated adjectives, which simultaneously correlate to expertise level and measurable data metrics, will be good choices for an automated coaching system. Many studies have recently investigated the effectiveness of crowd-sourcing for the assessment of technical skills and have shown correlations to expert evaluations (Chen et al., 2014; Malpani et al., 2015; White et al., 2015; Kowalewski et al., 2016b). These studies support the hypothesis that the identification of expertise is somewhat instinctive, regardless of whether or not the evaluator is a topic-expert in his or her evaluation area. The goal of this portion of our study is to see if the crowd can identify which of the chosen words for expertise are most important or relevant for surgical skill. Therefore, we conducted an experimental evaluation of our word pairs and metrics through a human subjects study, using the *da Vinci*® *Skills Simulator*^{TM} (on loan from Intuitive Surgical, Sunnyvale, CA). Users were outfitted with a variety of sensors used to collect metric data while performing tasks on the simulator.

2.2.1 Human Subject Study and Simulated Surgical Tasks

Three subjects were recruited to participate in this study, approved by both UTD and UTSW IRB offices (UTD #14-57, UTSW #STU 032015-053). The subjects (right handed, 2545 years old) consisted of: An expert (+6 years clinical robotic cases), an intermediate (PGY-4 surgical resident) and a novice (PGY-1 surgical resident). All subjects had limited to no training using the da Vinci simulator; however, the expert and intermediate had exposure to the da Vinci clinical robots.

The simulated surgical tasks chosen for this study were used to evaluate endowrist manipulation, needle control and driving skills (Fig. 2.1(a,b)). Endowrist instruments provide surgeons with range of motions greater than a human hand. Thus, these simulated tasks evaluate the subject's ability to manipulate these instruments. The needle driving task evaluates the subject's ability to effectively hand off and position needles for different types of suture throws (forehand and backhand) and while using different hands (left and right). All subjects first conducted two non-recorded warm up tasks (i.e., Matchboard 3 for endowrist manipulation warm up and Suture Sponge 2 for needle driving warmup). After training, subjects conducted the recorded experimental tasks for endowrist manipulation (Ring and Rail 2) and needle driving (Suture Sponge 3). For the purposes of data analysis, each task was subdivided into three repeated trials, corresponding to a single pass of a different colored ring (i.e., red, blue or yellow), or two consecutive suture throws.



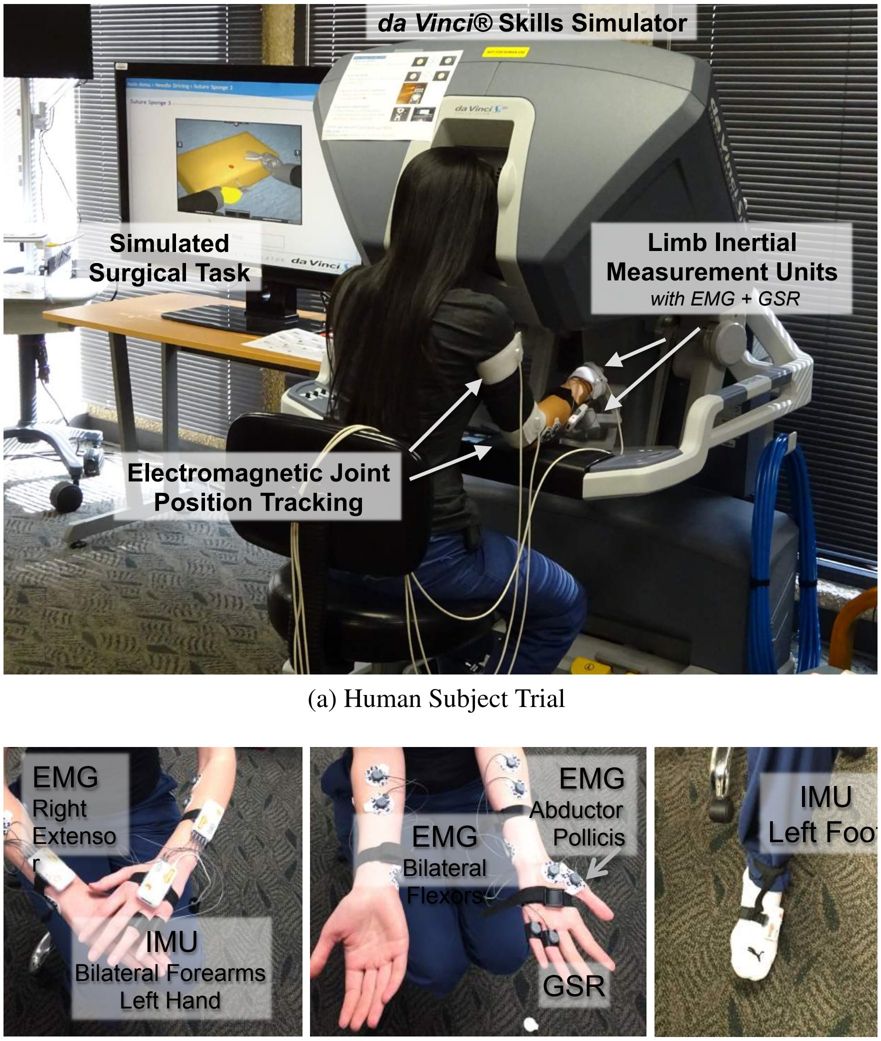
(a) Ring and Rail Task

(b) Suture Sponge Task

Figure 2.1: Two simulated tasks on the *da Vinci*® Skills Simulator were chosen including: (a) Ring and Rail 2 and (b) Suture Sponge 3. These tasks were chosen for their ability to evaluate endowrist manipulation and needle driving and control skills.

2.2.2 Experimental Setup and Data Collection System

To quantify task movements and physiological response for our semantic label metrics, we chose to measure joint positions (elbow, wrist, shoulder), limb accelerations (hand, forearms), forearm muscle activity with EMG, and GSR. Joint positions were recorded using electromagnetic (EM) trackers (trakSTAR, Model 180 sensors, Northern Digital Inc., Ontario, Canada) with an elbow estimation method as described in (Nisky et al., 2014). Limb accelerations, EMG and GSR were measured using sensor units from Shimmer Sensing, Inc. (Dublin, Ireland). Several muscles were selected from EMG measurement including bilateral (left and right) extensors, and a flexor on the left arm, which are important for wrist rotation, as well as the abductor pollicus, which is important for pinch grasping with the thumb (Criswell, 2010). These muscles were recommended by a surgeon educator. For each subject, baseline data was collected including arm measurements for the elobow position estimator, and maximum voluntary isometric muscle contractions (MVIC) for normalization and cross-subject comparison (Halaki and Ginn, 2012). We also recorded videos of the user posture and simulated surgical training task with CCD cameras (USB 3.0, Point Grey, Richmond, Canada). Robot Operating System (ROS) was used to synchronize all data collection. The experimental setup and sensor locations are shown in Fig. 2.2(a,b).



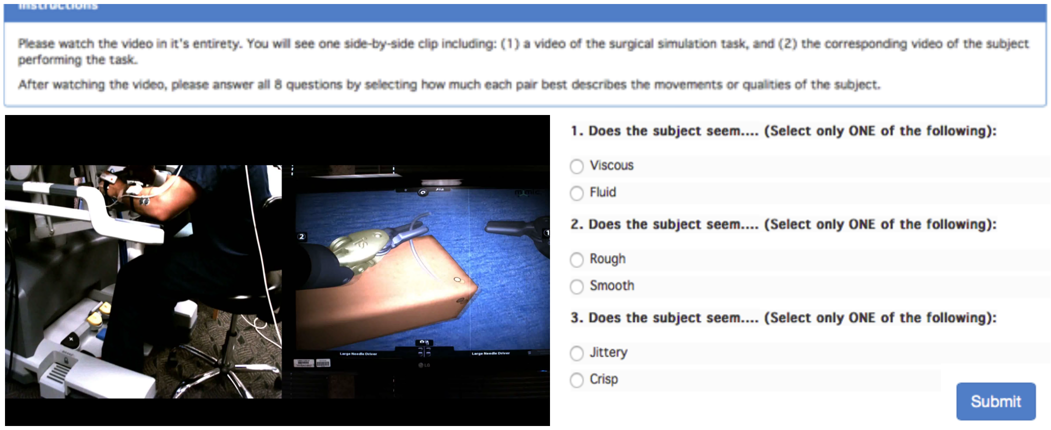
(a) Human Subject Trial

(b) Shimmer Sensor Placement

Figure 2.2: Human subjects conducted simulated tasks using the *da Vinci*® Skills Simulator while wearing various sensors for motion, position, and physiological response measurement (a). Inertial measumrements were obtained from the hands and forearms, as well as the foot, while electromyography signals were recorded from various muscles relevant to robotic surgical tasks. Galvanic skin response was also measured (b).

2.2.3 Crowd-Worker Recruitment and Tasks

For each trial, side-by-side, time-synchronized videos of the simulated surgical task and user posture were posted on Amazon Mechanical Turk. The videos ranged in length from 15 s to 3 min and 40 s. Anonymous crowd-workers (n=547) were recruited to label the videos using one from each



Instructions

Please watch the video in it's entirety. You will see one side-by-side clip including: (1) a video of the surgical simulation task, and (2) the corresponding video of the subject performing the task.

After watching the video, please answer all 8 questions by selecting how much each pair best describes the movements or qualities of the subject.

1. Does the subject seem.... (Select only ONE of the following):

○ Viscous
○ Fluid

2. Does the subject seem.... (Select only ONE of the following):

○ Rough
○ Smooth

3. Does the subject seem.... (Select only ONE of the following):

○ Jittery
○ Crisp

Figure 2.3: Crowd-workers were recruited on Amazon Mechanical Turk to select between contrasting pairs of adjectives to describe videos of the subject posture and task performance, played simultaneously. Crowd-workers were required to select one of each adjectives pair to earn credit for the job.

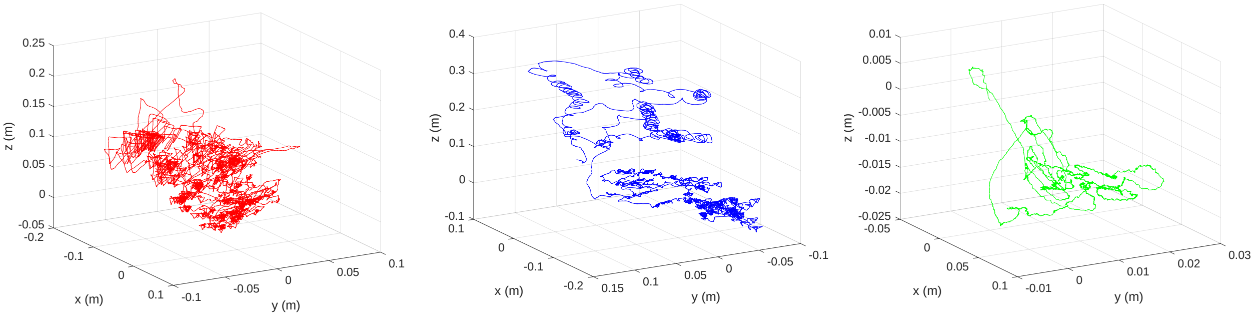
of the six contrasting adjectives pairs (Fig. 2.3). Crowd-workers received $0.10 for each video and were not allowed to evaluate the same video more than once

2.2.4 Data Analysis Methods

For each word pair, many options exist for correlation to a desired metric (e.g., sensor location, muscle type, summary static etc.). In this preliminary study, we selected metrics based on logical reasoning and feedback from surgical collaborators. To measure crisp vs. jittery hand movement, we calculated the standard deviation of the trial average mean value of jerk from the inertial measurement unit (IMU) mounted on the subjects right hand. Similarly, fluid/viscous was measured by variability in angular velocity of the same IMU. Smooth/rough was measured by the variability in acceleration magnitude of an IMU mounted on the right forearm. The acceleration magnitude was calculated for each time-sampled x, y, and z accelerations to eliminate the effects due to gravity [15]. To measure the calmness vs. anxiousness of the user, the GSR signal was processed using the Ledalab EDA data analysis toolbox in Matlab to count stressful events (Benedek and Kaernbach, 2010b). Mean EMG levels were recorded through electrodes placed on the forearm extensor as a measure of relaxedness vs. tenseness of the subject. In order to compare EMG levels between different subjects, these signals were normalized using the maximum of three repeated EMG signals during a maximal voluntary isometric contraction (MVIC) for each muscle. All EMG signals were high pass filtered using a fourth order Butterworth filter with a 20Hz cut-off frequency to remove motion artifacts and were detrended, rectified, and smoothed (Halaki and Ginn, 2012). The EM tracker data was used to visualize user movements and was not correlated to any word pairs. Finally, a Pearsons R correlation was used to compare the crowd-worker results and data metrics for each word pair.

2.2.5 Results and Discussion

The trajectories for each subject's wrist movements as measured by the EM tracker are shown in Fig. 2.4. As expected, the expert movements are tighter and smoother than the intermediate and novice.



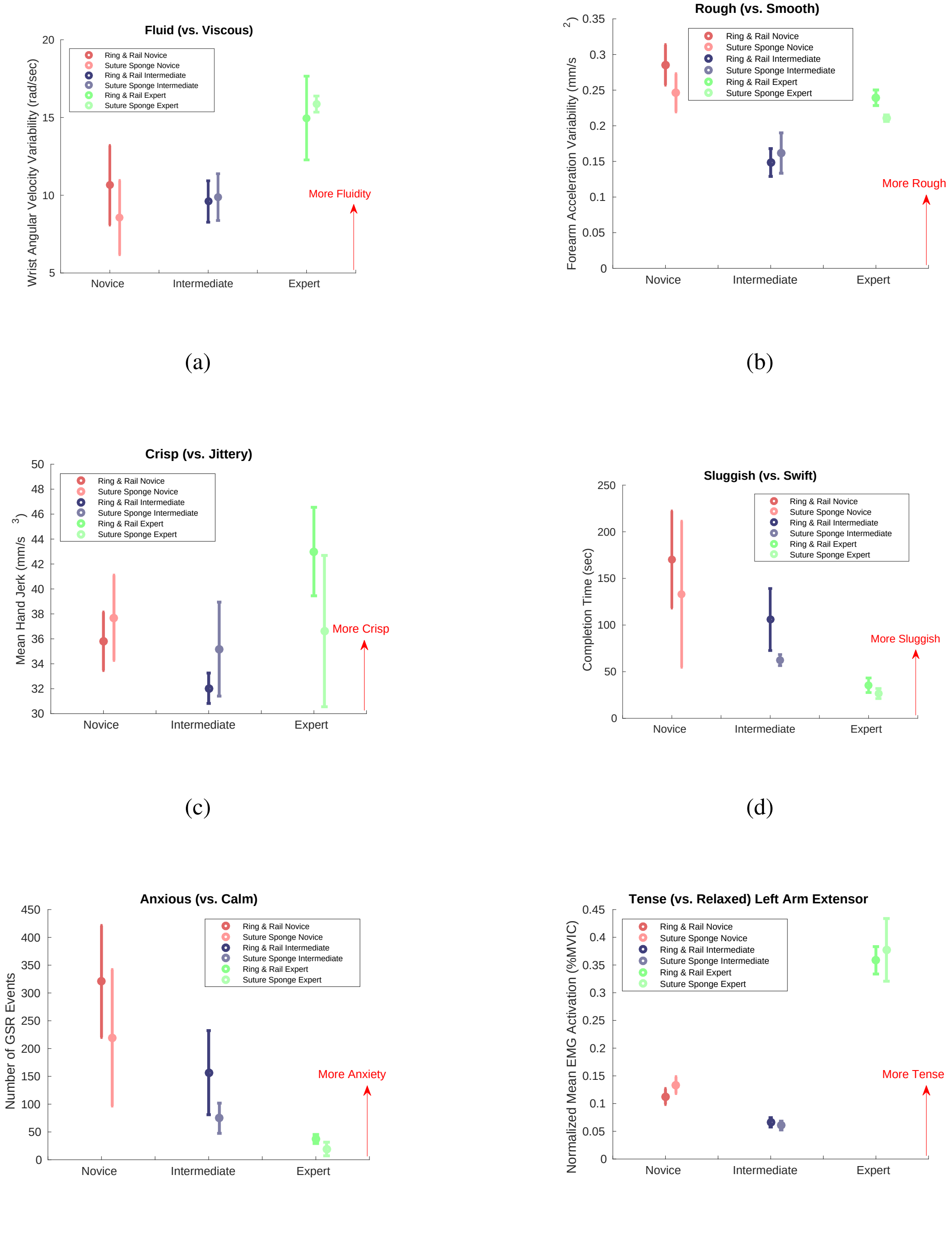
(a) Novice

(b) Intermediate

(c) Expert

Figure 2.4: Wrist trajectory of subjects performing Ring and Rail 2 (red ring)

Figure 2.5 compares the mean and standard deviation of each chosen metric through all trials among three subjects. Of the 547 crowd-workers recruited, 7 jobs were rejected due to incomplete



(a) (b)

(c) (d)

(e) (f)

Figure 2.5: Mean and standard deviation of all metrics for all trials and subjects.

labeling assignments, resulting in 30 complete jobs for each of the 18 videos posted. The results of the analysis can be seen in Fig. 2.6.

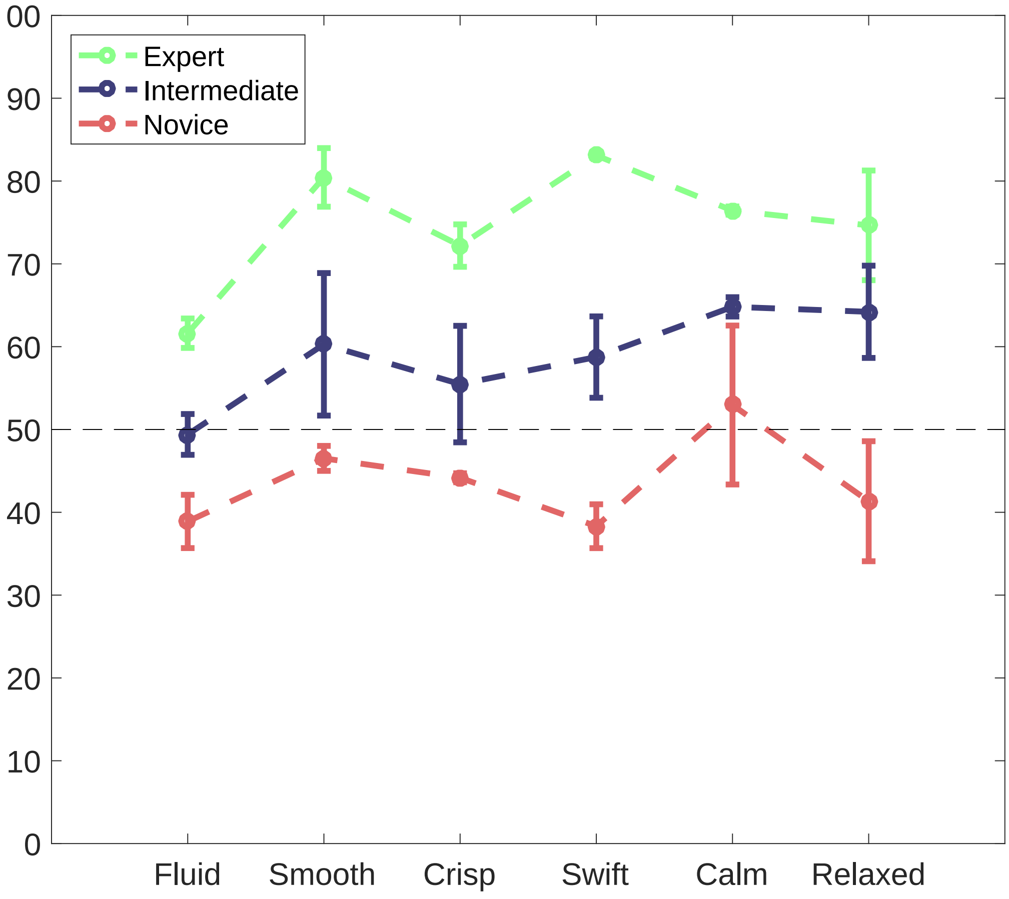


Figure 2.6: Results of the crowd-worker assignments

An ANOVA analysis was conducted to identify significant groups in terms of expertise level, type of task, and repetition for the data metrics, as well as crowd sourced data. Additionally, the crowd sourced data was evaluated for significant differences in terms of word assignment rates. Table 2 summarizes the significant statistical results ($p \leq 0.05$), significant groups (post-hoc Scheffe test), and the correlation between the crowd ratings and data metrics. For nearly all metrics, there was no significant effect due to task or repetition. The expert exhibited better performance on all metrics, with the exception of the smooth/rough and relaxed/tense. This could be due to a poor choice of metrics, or data collection errors with the expert EMG signal or baseline. The crowd assigned significantly better semantic labels to the expert, then the intermediate and novice. Additionally, the crowd rated the ring and rail tasks with significantly lower ratings than the suturing task, and evaluated the second repetition across all subjects as worse than the first and last. The magnitude of the data metric to crowd rating correlation ranged from 0.25 to 0.99. The best correlated metric to word-pair was swift/sluggish and the worst correlated metric was smooth/rough followed by crisp/jittery.

Table 2.2: Crowd Data Statistical Analysis Summary

Metric/Crowd Correlation	Subject (E, I, N)		Task (RR, SS)		Repetition (1-3)		
	p	Significance	*p*	Significance	*p*	Significance	
Fluid/Viscous	0.82	0.0005	E > I, N	0.0374	RR > SS	0.1134	n/a
Smooth/Rough	0.25	0.0001	I > E, N	0.1240	n/a	0.3366	n/a
Crisp/Jittery	0.63	0.073	n/a	00.7521	n/a	0.9128	n/a
Calm/Anxious	0.98	0.035	E < I< N	0.2286	n/a	0.9504	n/a
Relaxed/Tense	0.76	<0.0001	E > N> I	0.6834	RR > SS	0.6291	n/a
Swift/Sluggish	0.99	0.0028	E > N	0.1659	n/a	0.8541	n/a
Crowd worker Rating		<0.0001	E > I> N	<0.0001	SS > RR	0.0005	2< 1,3

E – Expert, I – Intermediate, N – Novice, RR – Ring and Rail 2, SS – Suture Sponge 3

2.2.6 Summary of Defining and Validating a Lexicon for Surgical Style

In this section, we presented a lexicon of surgical expertise composed of contrasting adjective pairs, associated with quantitative movement or user-based metrics. We showed that crowd-labeled training videos correlate strongly to expertise level. The data metrics typically also corresponded to expertise level; however, there were some discrepancies in terms of the smooth/rough metric and relaxed/tense metric. Finally, the crowd-workers identified differences based on task and repetition not seen in the data metrics, and not all data metrics correlated to trends in the crowd-worker ratings. In the next section we will determine which metrics are best correlated, and who/what is more correct in evaluating expertise - the crowd or the metrics. We also expand the lexicon to include additional semantic labels and will identify, specifically, which better predict expertise.

2.3 Finding Metrics associated with Surgical Styles

Stylistic coaching is often present in mentored training with an expert. This type of coaching is lacking from current surgical simulators which focus on quantitative metrics (e.g., workspace range, instrument collisions, economy of motion, etc.). Verbal feedback has been shown to efficiently improve performance in technical skills (Porte et al., 2007), and thus supports our hypothesis of stylistic-based coaching. Therefore, an important research question arises: what quantitative metrics are perceived when observing an expert, and how are these metrics translated to verbal descriptors of movement? In the previous section, we proposed a method for surgical skill assessment using a semantic lexicon of stylistic behavior to describe surgical expertise. In this section, we expand the lexicon proposed previously and aim to find the data metrics which most highly correlate to the lexicon of surgical expertise. The extended lexicon is shown in Table 2.3. Furthermore, to evaluate the effectiveness of these metrics for differentiating different levels of expertise, the metrics are used to classify subjects to different expertise levels. This study was published in (Ershad et al., 2018b)

Table 2.3: Extended Lexicon of Stylistic Behavior

Positive Adjective	Negative Adjective
Crisp	Jittery
Coordinated	Uncoordinated
Fluid	Viscous
Deliberate	Wavering
Swift	Sluggish
Smooth	Rough
Relaxed	Tense
Calm	Anxious

2.3.1 Experimental setup

A block diagram describing the procedures in this study is shown in Fig. 2.7. Each object in the diagram is discussed in detail in the following. The experimental setup is similar to the study in the previous section 2.2; however, data collection is discussed in more detail in this section.

Subjects and Simulated Surgical Tasks

We extended the previous study by recruiting 14 subjects including four experts (practicing robotic surgical faculty), three fellows (surgical fellows post residency), three intermediates (PGY-4 surgical residents), and four novices (medical students). All subjects were asked to perform two tasks on a *da Vinci*® *Skills Simulator*™ (same as the tasks discussed in section 2.2.1). The study protocol was approved by both UTD and UTSW IRB offices (UTD #14-57, UTSW #STU 032015-053).

Data Acquisition

To find quantitative metrics associated with the adjectives in our lexicon, task movements and physiological response, as well as videos of the user and of the training tasks, were recorded. In order to record data from all sensors and cameras simultaneously, Robot Operating System (ROS) (Quigley et al., 2009) was used.

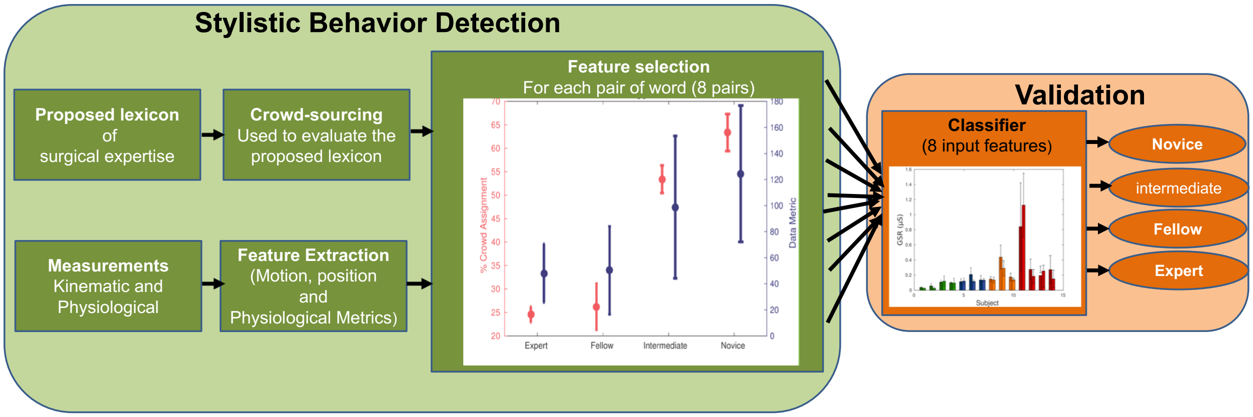

Stylistic Behavior Detection
Proposed lexicon of surgical expertise
Crowd-sourcing Used to evaluate the proposed lexicon
Measurements Kinematic and Physiological
Feature Extraction (Motion, position and Physiological Metrics)
Feature selection
For each pair of word (8 pairs)
% Crowd Assignment
Data Metric
70
65
60
55
50
45
40
35
30
25
20
180
160
140
120
100
80
60
40
20
0
Expert
Fellow
Intermediate
Novice
Validation
Classifier
(8 input features)
GSR (µS)
Subject
Novice
intermediate
Fellow
Expert

Figure 2.7: The proposed method combines crowd-selected adjectives of expertise, and the metrics associated with crowd-assignment rates, through the use of a machine learning classifier to determine levels of expert-like style.

Position data Position data were acquired using electromagnetic trackers (trakSTAR, Model 180 sensors, Northern Digital Inc., Ontario, Canada) at 256 Hz, placed on the wrist, elbow and shoulder joints (Fig. 2.2a). Position derived metrics are commonly used in surgical skill assessment (Moorthy and Munz, 2003; Datta et al., 2001). Wrist, elbow, and shoulder joints are shown to be effective in arm posture analysis in Robotic surgery (Nisky et al., 2013).

Motion data Motion data were gathered through kinematic measurements with a sampling frequency of 512 Hz using inertial measurement units (IMU), embedded in the Shimmer sensors (Shimmer Sensing, Inc., Dublin, Ireland) located on the hand and forearm, which control clutching and tool manipulation, and the foot, which controls the camera view of the robot (Fig. 2.2 (b)). The sensors include integrated 6 DoF inertial sensing via accelerometer, gyroscope, magnetometer and altimeter sensors.

Physiological data The physiological responses used in this study included electromyography (EMG) and galvanic skin response (GSR), which were recorded using Shimmer sensors with a sampling frequency of 1024 Hz and 512 Hz respectively. The EMG and GSR sensor positioning is shown in Fig. 2.2(b). EMG signals were recorded from forearm extensors, flexors, and policis, which are the main muscles activated in robotic surgery (indicated by an expert surgeon), while performing a robotic surgical task on *da Vinci*® simulator. For EMG signal acquisition, two sensors were placed 2cm apart along the longitudinal midline of the desired muscle, parallel to the muscle fibers. Maximum voluntary isometric muscle contractions (MVIC) were collected for normalization and cross-subject comparison. For this purpose, each subject was asked to contract the desired muscle with maximum strength for three repetitions of 10 seconds each. The maximum value among these three repetitions for each muscle was selected as the baseline MVIC.

GSR was recorded to gain information of the stress level of the user. Two electrodes of the shimmer GSR sensor were placed on two neighboring fingers for signal acquisition.

Videos Two CCD cameras (USB 3.0, Point Grey, Richmond,Canada) were used to record videos of the subject performing the task, as well as the task being performed on the simulator.

Crowd Sourced Assessment of Stylistic Adjectives

The paired videos were randomized and rated by the crowd using the adjectives proposed in the lexicon. For each subject trial, a side-by-side, time-synchronized video of the simulated surgical task and user posture were posted on Amazon Mechanical Turk. A total number of 84 videos (14 subjects and 2 tasks, 3 repetition each) were posted. Anonymous crowd workers were asked to label each video based on the 8 adjective pairs in the proposed lexicon. A screen shot of the videos and questionnaire posted on Amazon Mechanical Turk is shown in Fig 2.3. They received $0.10 for each video and were not allowed to evaluate the same video more than once. A total of 1680 surgical assessment tasks were requested (20 unique crowd-workers assessed each of the 84 videos), out of which 27 tasks were rejected due to incompletion.

Faculty Expert Assessment of Stylistic Adjectives and GEARS Scoring

Seven faculty experts were also recruited to assess the videos (each expert evaluated 12 of the 84 videos). The crowd ratings were comparable to expert ratings. However, the crowd assessment took 3 days to complete, while it took more than a month for seven experts to complete rating 84 videos. The paired videos were also rated based on the GEARS domain by 7 faculty experts through Qualtrics survey.

2.3.2 Candidate Metrics for Stylistic Adjectives

A total of 83 metrics were calculated from the motion, position, and physiological data measurements as potential metrics to associate with each stylistic adjective pair. The chosen metrics for measuring fluidity, smoothness, crispness, and deliberateness of movement, and anxiety levels of the user, were proposed based on inspiration from the literature (Balasubramanian et al., 2015; Anderson et al., 2012; Grober et al., 2010; Postacchini et al., 2015), as well as metrics that could easily be computed from our sensor data. The calculated metrics are discussed below. Table 2.4 shows all the metrics used in this study.

Position-Based Data Metrics

The position of the shoulder and the wrist were measured directly using EM trackers. However, because of the metal structure of the armrest and its interference with the elbow measurements from the EM tracker, we estimated the elbow position using the shoulder and wrist positions measured by the EM trackers, as discussed in (Nisky et al., 2014). The elbow position was measured by the average of two estimates below:

$$X_e^s(t) = X_s(t) + L_{se} * r_{xs}(t) \tag{2.1}$$

$$X_e^w(t) = X_w(t) + L_{ew} * r_{xw}(t) \tag{2.2}$$

where X_e^s and X_e^w are the estimate of the elbow position using the shoulder and wrist sensor respectively, L_{se} and L_{ew} are the length of the upper arm and the forearm respectively, which were measured by placing EM trackers on the shoulder, elbow, and wrist joints prior to the experimental trials. The direction of the longitudinal axis of the shoulder and wrist sensors, are r_{xs} and r_{xw}, respectively.

The cumulative path length of the mentioned limbs was calculated as the sum of Euclidean distances between sample points (Anderson et al., 2012), from Eq.2.3.

$$\sum_{k=0}^{n} ||p(k) - p(k-1)|| \tag{2.3}$$

where $p(k)$ is the position of the sensor in the kth sample for a total number of n samples. The difference in distance traveled by both hands was also calculated. The movement speed for both hands was calculated from the derivative of the position measurements and the mean and standard deviation of the speed were calculated.

Motion-Based Data Metrics

The IMU sensors were placed on the hands, forearms and the foot (which controls the camera view with the pedal). The position of the motion sensors is shown in Fig. 2.2(b). Thus all IMU measurements discussed in the following, are measured for these limbs. The X, Y, and Z directions for the measurements associated with the IMU sensor is shown in Fig. 2.8. Limb acceleration and angular velocity were measured using the Shimmer sensors. For each sensor we calculated:

- The mean and standard deviation of the angular velocity in X, Y, and Z directions.
- The mean, standard deviation, and maximum of jerk in Z direction.
- The mean and standard deviation of the magnitude of the jerk.
- The mean and standard deviation of the magnitude of the linear acceleration.
- The number of movements using the magnitude of the resultant acceleration calculated as follows, which was used in each instance of the acceleration measurement to eliminate the effects due to gravity (Grober et al., 2010).

$$a_{net} = sqrt(a_x^2 + a_y^2 + a_z^2) \tag{2.4}$$

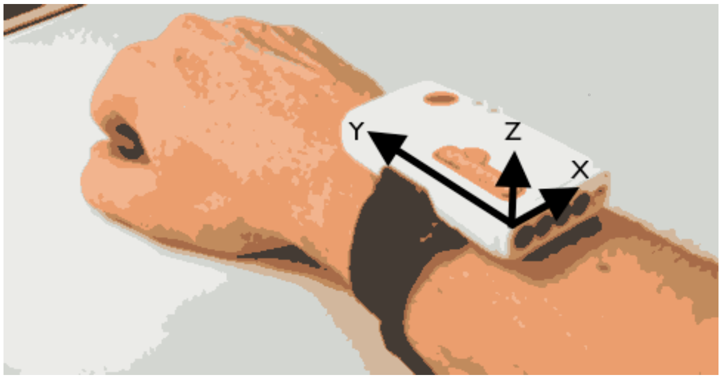


Figure 2.8: A standard orientation was selected for all Shimmer sensors to ensure appropriate computation of various metrics. The sensors were mounted on each subject with the +Y axis placed proximal on the limb with the base of the sensor mounted medially on the dorsal side of the limb.

The net acceleration was then smoothed using Gaussian filter with a 0.5 standard deviation (Shenoi, 2005). The peaks in the smoothed net acceleration signal, which indicate an increase and decrease in the acceleration, count the number of movements (Grober et al., 2010).

- A jerk cost function (jerk trajectory) was calculated as follows

$$\sum_{k=0}^{n}(jerk_x(k)^2 + jerk_y(k)^2 + jerk_z(k)^2) \tag{2.5}$$

  where n is the number of samples in the signal.

- Since surgical movements are inherently dynamic, we developed a new heuristic metric, the maximum of average jerk within 1/8 time intervals of the overall task completion time T (i.e. $T/8$). This metric captures the average jerk within a short segment of the overall task.

Physiological Data Metrics

The amplitude of the EMG signal was used to compare subject performance. In order to compare EMG signals, the signal was normalized using the maximum voluntary isometric contraction

(MVIC) signals (Halaki and Ginn, 2012). After the EMG signal was acquired, DC offset was removed and the signal was rectified and filtered by a non-zero phase lag, third order low-pass Butterworth filter with a cut-off frequency of 20 Hz to eliminate noise due to motion artifacts (De Luca et al., 2010).

The average and variability in the EMG signal gathered from the left and right extensor, right left flexor and the left pollicis was used to assess whether subjects were relaxed or tense.

The number of events in the GSR signal was calculated from the GSR Shimmer sensors using Ledalab (Benedek and Kaernbach, 2010a), and it was time averaged for each subject. The standard deviation of the GSR signal was also calculated.

Table 2.4: Calculated Metrics

Sensors	Metrics
Position Based Metrics	Distance traveled (4 metrics)
4 EM trackers	Difference in distance traveled by both hand
	Mean speed in right and left hand (4 metrics)
	Standard deviation of speed in right and left hand
Motion Based Metrics	Number of movement (4 metrics)
4 IMU sensors	Mean angular velocity in x, y, and z direction (12 metrics)
	Standard deviation of angular velocity in x, y, and z direction(12 metrics)
	Maximum jerk in Z direction (4 metrics)
	Mean jerk in Z direction (4 metrics)
	Standard deviation of jerk in Z direction (4 metrics)
	Mean of jerk magnitude (4 metrics)
	Standard deviation of jerk magnitude (4 metrics)
	Mean of maximum jerk in 1/8 time intervals (4 metrics)
	Mean of acceleration magnitude (4 metrics)
	Standard deviation of acceleration magnitude (4 metrics)
	Jerk trajectory (4 metrics)
Physiologocal Based Metrics	Mean magnitude of 4 EMG signal (4 metrics)
4 EMG sensors	Standard deviation magnitude of 4 sensors (4 metrics)
	Standard deviation of GSR
GSR sensor	Number of GSR events

2.3.3 Metric Selection

Among all the proposed metrics found for each adjective pair, the best metric associated with each pair was found using a cross correlation coefficient. The metrics as well as the crowd ratings were arranged based on the expertise levels, and Pearson correlation was calculated for each proposed metric associated with each adjective pair, and the relevant crowd ratings. The metrics with the highest correlations were chosen. The correlation between the crowd ratings and the faculty ratings for each adjective pair were also calculated.

2.3.4 Metric Normalization

Since the units of the data metrics are different, the range of value that they take are also significantly different. This difference in dynamic range can negatively affect the performance of a classifier. To overcome this problem, for each trial the data metrics were normalized using the following method

$$X = (D - min(D))/(max(D) - min(D)) \tag{2.6}$$

Where D is the array containing data metrics associated to each trial. To minimize the effect of outliers, Winsorizing was used such that the bottom 5% of values are set equal to the minimum value in the 5th percentile, and the upper 5% of values are set equal to the maximum value in the 95th percentile.

2.3.5 Metric Evaluation in Distinguishing Expertise Levels

To evaluate the metrics' ability in distinguishing different levels of expertise, two classifiers were trained using the chosen metrics. The trained classifiers were used to differentiate expert levels (expert, fellow, intermediate, or novice) based on the metrics associated with the stylistic behavior of each subject.

A Gaussian Naive Bayes classifier was used with the prior probabilities set as class relative frequencies distributions. The distribution parameters of the model are estimated through cross validation on the training set.

A Naive Bayes classifier is a probabilistic classifier based on conditional probability and Bayes' theorem. For the space of decision with M classes, the probability that a set of feature vectors $X = X_1, X_2, ...X_n$, belong to the correct class m_i is determined as:

$$P(m_i|X) = P(m_i \cap X)/P(X) \tag{2.7}$$

Classification is done according to the following rule:

$$X \in m_i \quad \text{if} \quad P(m_i|X) > P(m_j|X) \quad \text{for all} \quad i \neq j$$

Based on Bayes theorem, equation 2.7 can be written as:

$$P(m_i|X) = P[P(X|m_i)P(m_i)]/P(X) \tag{2.8}$$

with the naive assumption of each feature X_p being conditionally independent of every other feature X_q for all $p \neq q$, this equation can be simplified as discussed in (De Moraes and dos Santos Machado, 2007). However this classifier has proven to perform well in spite of the independence assumption on the features (Zhang, 2004). In this study, a multi-class Naive Bayes classifier was trained using the selected features to assess its ability to classify the subjects based on expertise level.

A SVM classifier with radial basis function (RBF) kernel and parameters (C, γ) was also used.

$$K(u, v) = exp(-\gamma|u - v|^2), \gamma > 0 \tag{2.9}$$

Gamma is the free parameter of the Gaussian radial basis function. C controls the cost of misclassification on the training data. The parameters of the RBF kernel were tuned through a grid-search on C and γ using cross-validation. Various pairs of (C, γ) values were tried and the one with the best cross-validation accuracy was chosen.

2.3.6 Cross Validation

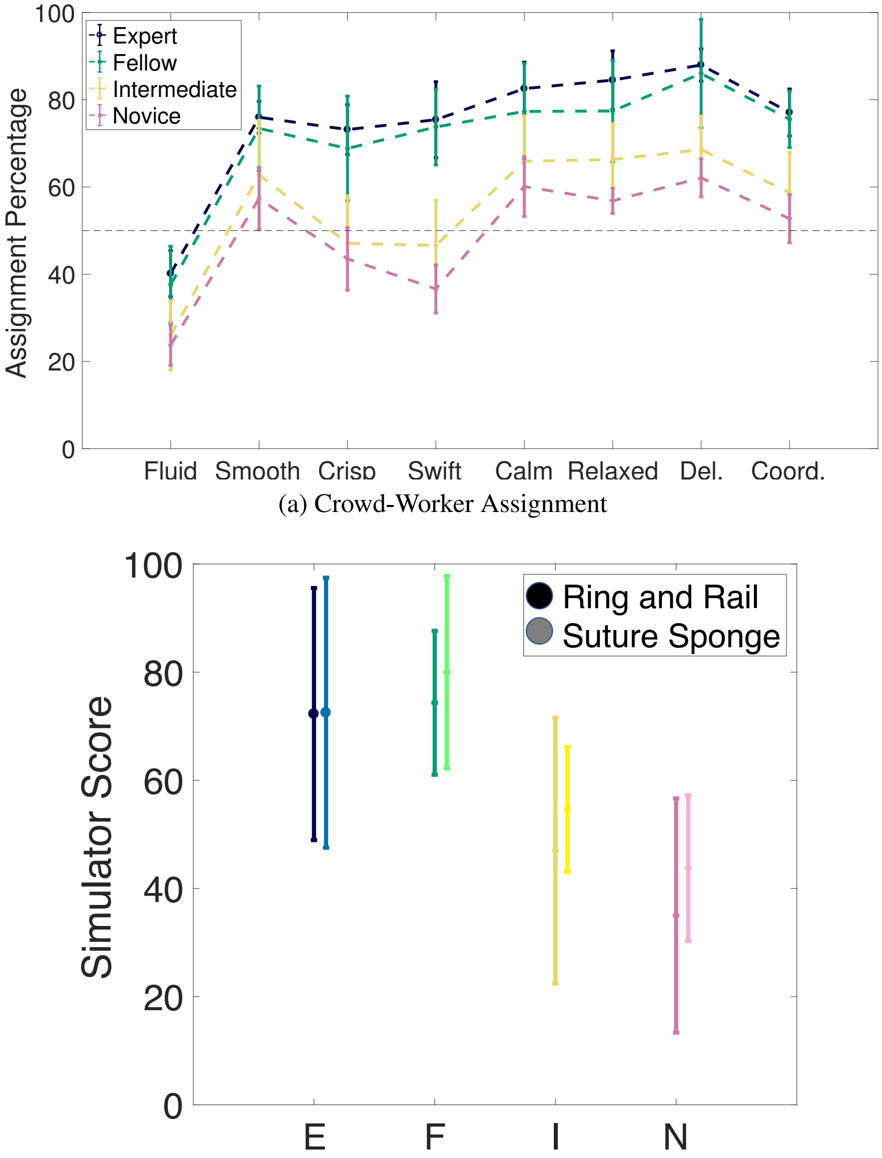
Two types of cross validation were performed. The first one was a simple k-fold cross validation which partitions the data set into k partitions, trains the classifier on k-1 partitions, and tests it on the remaining partition. This is repeated for every partition. The second method was leave one user out (LOUO) cross validation in which all data from one user is left out while the classifier is trained on the remaining users' data, the trained classifier is then tested on the left out user's data. All analysis were done in Matlab.

2.3.7 Results and Discussion

Results of Crowd Assignments

The results of the crowd source rating can be seen in Fig. 2.9(a). This figure shows the mean and standard deviation of the percentage of the crowd assignments, relating the positive adjective to the expertise level, for each adjective pair in our lexicon. The percentage was calculated by counting the total number of times a positive adjective was chosen for a single video and the average was calculated over all the videos for the particular level of expertise. As indicated by these results, the crowd assigned the best performance to experts, followed by fellows, then intermediates, and then novices for all eight pair of adjectives. The mean and standard deviation of the simulator scores for the subjects, based on their expertise level, is shown in Fig. 2.9(b). As indicated in this figure, the simulator scores show a similar performance for the expert (mean = 72.37%, standard deviation = 22.35%) and fellow (mean = 78%, standard deviation = 13.29%), with a slightly better performance for fellow. The scores associated with the intermediate and novice are in agreement with the crowd assignments for expertise levels.

Statistical analysis was done on the crowd ratings for each pair of adjectives to identify significant difference in terms of expertise level, type of task, repetition, and their interactions. This was also done on the overall crowd rating, which is the average assignment percentage for each subject



(a) Crowd-Worker Assignment

(b) Simulator Scores

Figure 2.9: Crowd-worker assignment percentages were computed for each adjective pair, shown for each positive adjective (a). The scores for each task, Ring and Rail 2 (3 repetitions) displayed in dark colors and Suture Sponge 3 (3 repetitions) displayed in light colors, as computed by the *da Vinci*® Standalone Simulator, were also averaged for experts (N=4), fellows(N=3), intermediates(N=3), and novices(N=4) (b).)

for all adjective pairs. These results can be seen in Table 2.5. As indicated by the results, although ratings progressively increased with levels of expertise, differences between experts and fellows, and between novices and intermediates were not significant. However, significant differences were noticed when the expert and fellow levels were combined and compared to the combined intermediate and novice levels for almost all of the adjectives. This demonstrates that, to some extent, the levels of expertise can be distinguished between subjects regardless of the task type or the number of repetitions. The simulator scores indicated significant difference between the combined expert and fellow levels compared to the novice level but no significant difference was seen on the intermediate level. This shows that, as opposed to the simulator, the crowd was able to distinguish the intermediate level from the expert and fellow level.

Correlation Between Crowd Assignments, Data Metrics, and Faculty Ratings

The correlation coefficients between the crowd rating and the proposed data metric for each pair of adjective, the crowd and faculty ratings, and the data metrics and faculty ratings are indicated in Table 2.6. In the crowd/metric correlation, a positive correlation shows that a higher value of the metrics relates to the positive adjective. A negative correlation shows that a lower value of the metric relates to the positive adjective.

Table 2.5: Crowd Data Statistical Analysis Summary

	Subject (E,F, I, N)		Task (RR, SS)		Repetition (1-3)	
Source	*p*	Significance	*p*	Significance	*p*	Significance
Fluid/Viscous	0.0004	{E, F} > {I, N}	0.0156	RR > SS	0.9465	n/a
Smooth/Rough	0.0057	{E, F} > N	0.1175	n/a	0.7777	n/a
Crisp/Jittery	0.0001	{E, F} > {I, N}	0.8746	n/a	0.4534	n/a
Swift/Sluggish	<0.0001	{E, F} > {I, N}	0.5216	n/a	0.2163	n/a
Calm/Anxious	0.0013	E > {N, I} & F > N	0.1390	n/a	0.6349	n/a
Relaxed/Tense	0.0001	E > {N, I} & F > N	0.1708	n/a	0.6567	n/a
Deliberate/Wavering	0.0001	{E, F} > {I, N}	0.5101	n/a	0.6726	n/a
Coordinated/Uncoordinated	<0.0001	{E, F} > {I, N}	0.9324	n/a	0.7095	n/a
Overall Crowd worker Rating	<0.0001	{E, F} > {I, N}	0.3307	n/a	0.5042	n/a
Simulator Scores	0.0015	{E, F} > {N}	0.3887	n/a	n/a	n/a

E – Expert, I – Intermediate, N – Novice, RR – Ring and Rail 2, SS – Suture Sponge 3
For the word pair ”calm-anxious” and ”deliberate-wavering” there was a significant difference in the task-repetition interaction (p = 0.0412 and 0.0484 respectively)

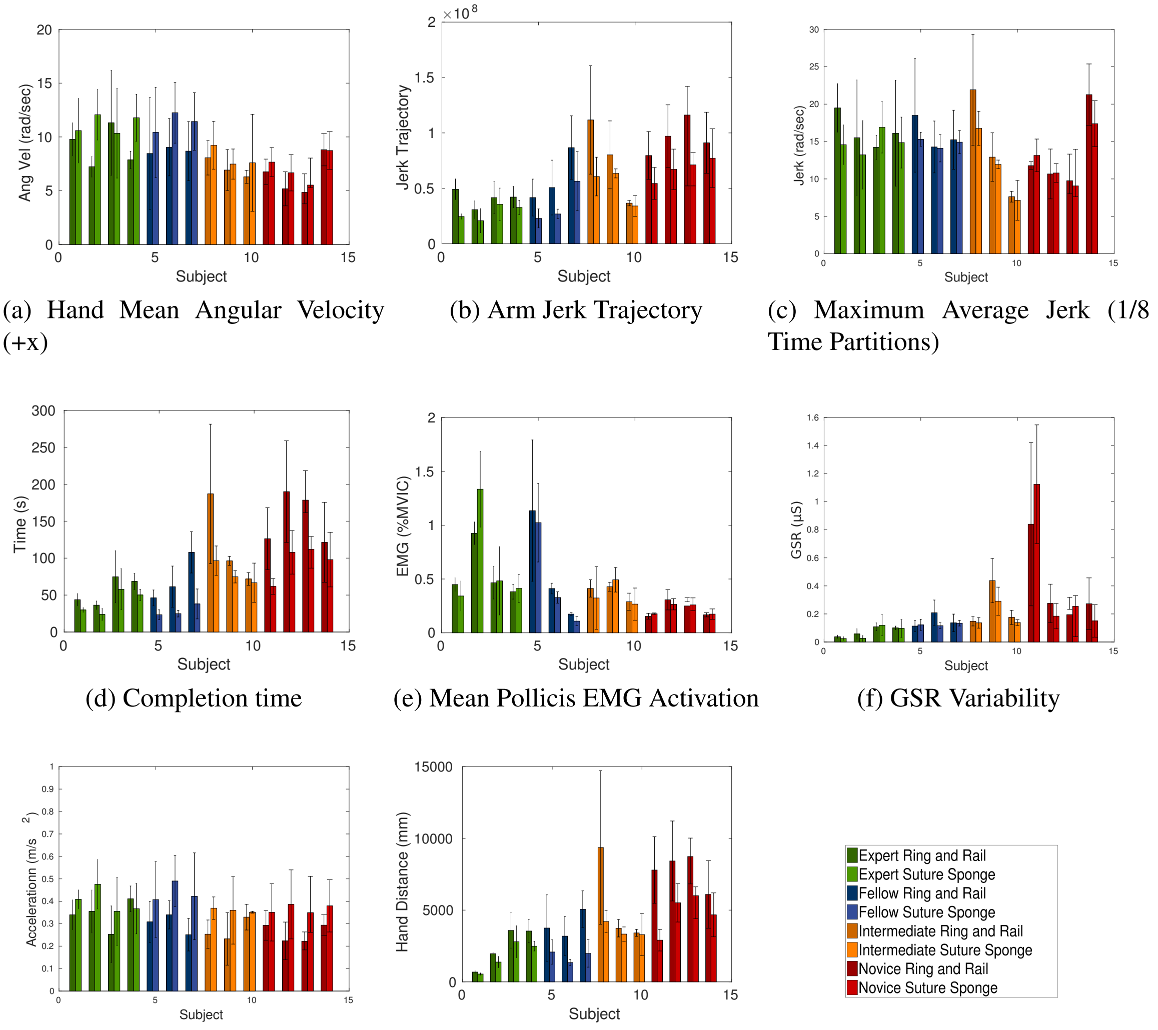
Table 2.6: Crowd/Metric/Faculty Rating Association and Correlation

Adjectives	Data Metrics	Metric/Crowd Correlation	Faculty/Crowd Correlation	Faculty/Metric Correlation
Fluid/Viscous	Mean angular velocity of the hand	0.99	0.95	0.97
	Number of movements of the hand[a]	0.99		0.98
Smooth/Rough	Jerk trajectory of the arm	-0.98	0.96	-0.92
Crisp/Jittery	Maximum of average jerk intervals	0.99	0.96	0.97
	Jerk standard deviation of the hand[b]	0.99		0.96
Swift/Sluggish	Completion time	-0.99	0.98	-0.97
Calm/Anxious	Variation in the GSR	-0.94	0.96	-0.97
Relaxed/Tense	Mean EMG in the pollicis	0.99	0.99	0.99
Deliberate/Wavering	Mean of peak amplitude in hand acceleration	0.98	0.97	0.95
Coordinated/Uncoordinated	Difference in right and left hand distance	-0.97	0.95	-0.90

[a] Alternative metric for fluid/viscous.

[b] Alternative metric for crisp/jittery .

The metrics that were used for potentially relating to the adjective pair **“fluid/viscous”** were the distance traveled calculated from the position sensors, the number of movements and the mean and standard deviation of angular velocity in the X, Y, and Z direction of the IMU sensors. Among all these metrics, the mean angular velocity and the number of movements of the right hand showed best correlation with the crowd (99% and 99%) and faculty ratings (97% and 98%). A higher angular velocity represents a more fluid movement. As indicated by Fig. 2.10(a), the expert has a better performance in both “ring and rail” and “suture sponge” tasks, compared to those of the novice and the intermediate, however fellows’ performances were very similar to experts. In the case of fellow 2, performance was even better than the experts. For the adjective pair **“smooth/rough”**, the potential metrics were the jerk trajectory, and the mean and standard deviation of the normalized linear acceleration of the IMU sensors. The jerk trajectory of the right forearm showed the highest correlation to the crowd and faculty ratings (-98% and -92% respectively). The minimum jerk trajectory is associated with smoothest movement. This result is consistent with the literature which commonly associate a smooth movement with minimum jerk trajectory (Flash and Hogan, 1985; Balasubramanian et al., 2015). As indicated in Fig. 2.10(b), novices have the highest jerk trajectory in both tasks and the jerk trajectory decreases as the level of expertise increases. Subject number 10 which is an intermediate showed a performance similar to that of an expert for this metric. For the adjective pair **“crisp/jittery”**, the potential calculated metrics were the mean, maximum and standard deviation of the normalized jerk also the jerk in Z direction, and the average of the maximum jerks in 1/8 completion time of the IMU sensors. The average of maximum jerks in 1/8 time intervals and the Jerk standard deviation of the right hand equally showed the highest correlation to the crowd (99% and 99%) and faculty ratings (97% and 96%). The average of the maximum jerks in 1/8 completion time was greater for crisper movements and associated to higher levels of expertise (Fig. 2.10(c)). Subjects number 8 and 14 , which were an intermediate and a novice respectively, showed similar performance to that of an expert in this metric. For the adjective pair **“swift/sluggish”**, the task completion time was calculated which showed -99% correlation



(a) Hand Mean Angular Velocity (+x)

(b) Arm Jerk Trajectory

(c) Maximum Average Jerk (1/8 Time Partitions)

(d) Completion time

(e) Mean Pollicis EMG Activation

(f) GSR Variability

(g) Mean Peak Hand Acceleration

(h) Difference in Right and Left Hand Travel

Figure 2.10: Computed data metrics for each task and subject level are shown in (a-h). These metrics showed the highest correlation with crowd-sourced assignment ratings of the chosen adjectives.

with the crowd and -96% with the faculty ratings. Fig. 2.10(d) shows that the higher the level of expertise, the less time was taken to complete the task. Subject number 8 which is an intermediate, had a performance similar to that of a novice. The potential metrics used for the adjective pair **"relaxed/tense"** included the mean and standard deviation of the normalized EMG signals. The mean normalized EMG from the policis showed the highest correlation with the crowd and faculty rating, both 99%. Fig. 2.10(e) shows that the average normalized EMG is higher for the higher levels of expertise almost for all subjects. Subjects number 6 and 7 have similar performance to that of a novice for this metric. For the adjective pair **“calm/anxious”**, the standard deviation in the GSR signal as well as the number of events were calculated and the GSR standard deviation showed higher correlation, -94% with the crowd and -97% with the faculty ratings. Fig. 2.10(f) shows that the GSR variability decreases as the level of expertise increase showing a calm appearance of the user for higher levels of expertise. For the adjective pair **“deliberate/wavering”**, the mean and standard deviation of the speed from the position sensors were calculated. The average speed of the right hand showed the highest correlation, 98% with the crowd, and 95 % with the faculty ratings. Fig. 2.10(g) shows that experts and fellows had similar performance and novices and intermediates also had similar performance in this metric. For the adjective pair **“coordinated/uncoordinated”** the differences between the distance traveled by each hand was calculated and showed a -97% correlation with the crowd and -90% with the faculty ratings (Fig. 2.10(h)). This difference decreases as the level of expertise increase showing more coordinated movements for higher expertise levels. Subject number 8, an intermediate, showed similar performance to a novice for this metric.

The normalized mean and standard deviation of the metrics for the expertise levels, associated with each adjective in the lexicon is shown in Fig. 2.11.

The signatures associated with three adjectives are represented in Fig. 2.12 for better a understanding of the relation between the proposed metrics and the expertise levels.

Fig. 2.13 compares the average and standard deviation of the crowd-source ratings and the chosen physiological metrics respectively through all trials for all four levels of expertise. Note that the crowd ratings for each pair of word and its relevant metric is shown on the same plot and the two ratings track each other. This demonstrated the correlation between the metrics and the crowd assignments.

Three-way ANOVA tests were performed on the chosen metrics to identify significance in terms of expertise level, type of task, repetition, and their interactions. Similarly, the ANOVA

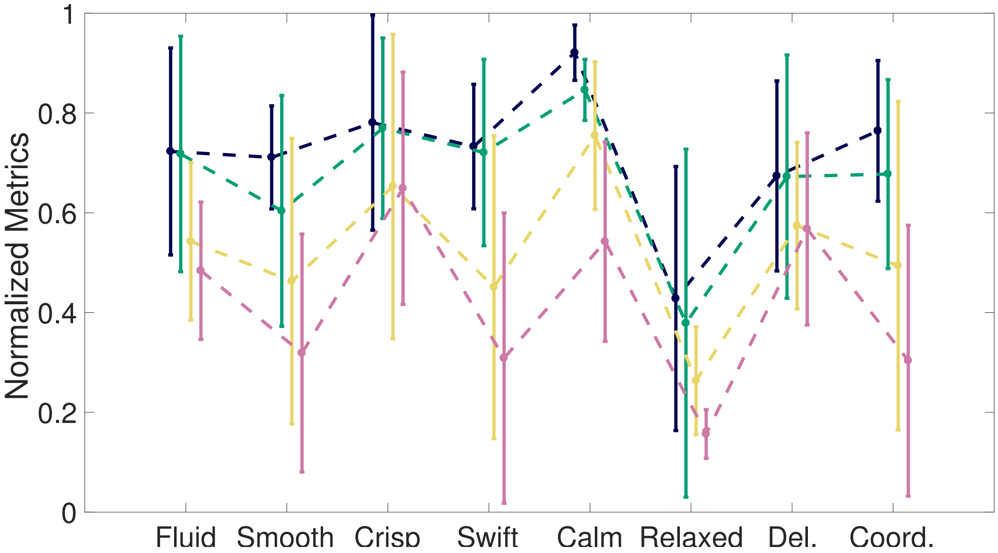
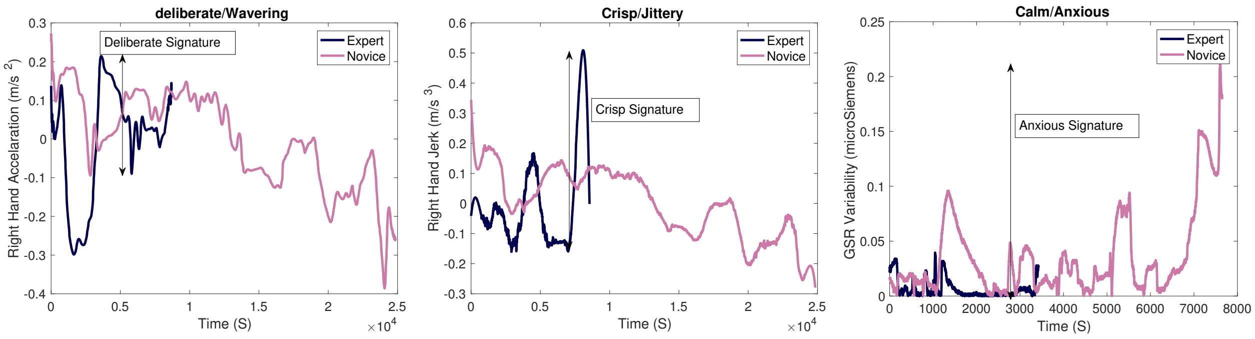


Figure 2.11: Data Metrics mean and standard deviation for each adjective pair.



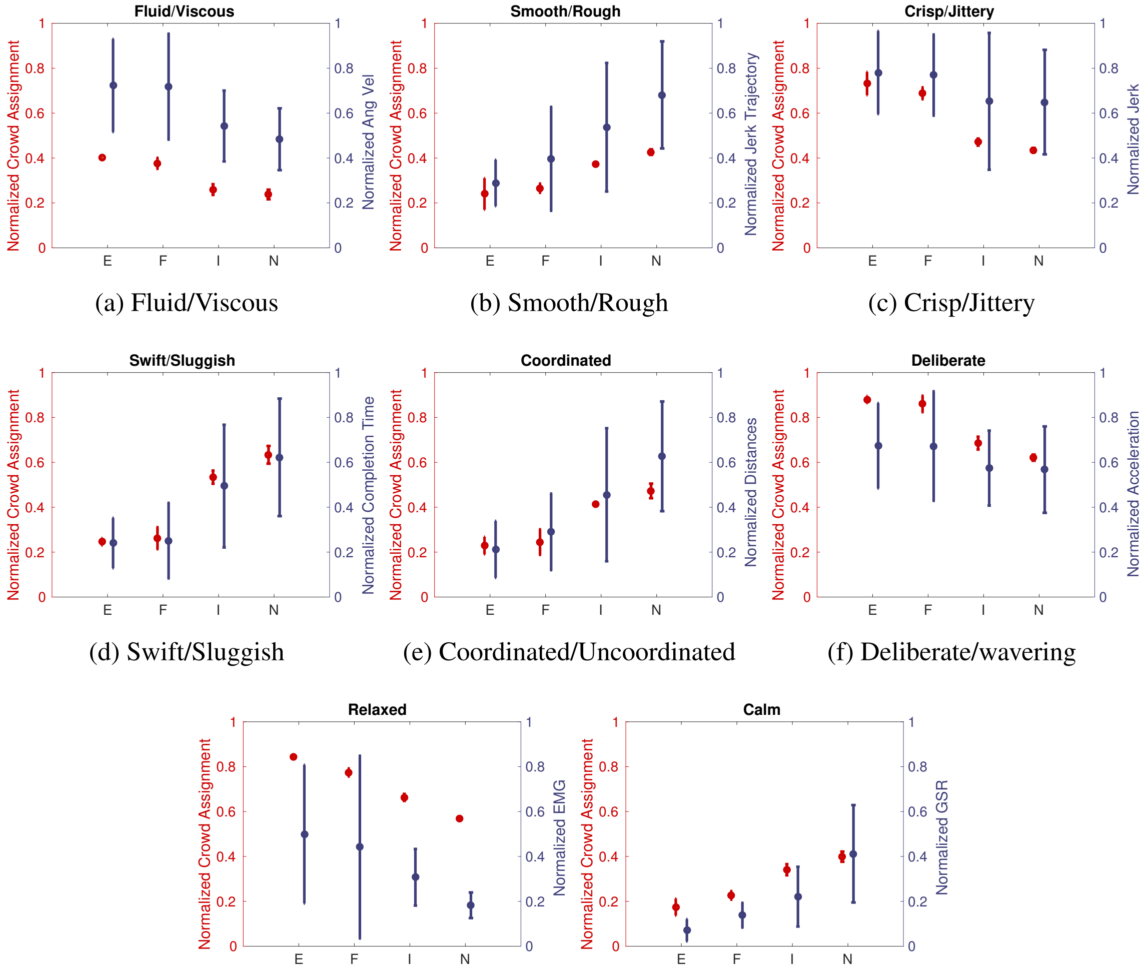
(a) Hand Acceleration

(b) Jerk

(c) Galvanic Skin Response

Figure 2.12: Comparison of expert and novice signatures associated with adjectives (a)“Deliberate”, (b)“Crisp”, and (c)“Calm”. (a)The experts show few large peaks while the novices show more smaller peaks. (b)The experts show greater peaks in the jerk signal while the novices show lower jerk (c) The experts show lower variability in the GSR signal while the novices show higher variability.

test was done on the overall, or total, average of each data metric for each subject. The results of the ANOVA can be seen in Table 2.7. The levels of expertise of expert and fellow were significantly different from novice and intermediate for almost all of the adjectives. No significant difference was seen among the repetitions in all trials. However for some trials the type of task was significantly different.



(a) Fluid/Viscous

(b) Smooth/Rough

(c) Crisp/Jittery

(d) Swift/Sluggish

(e) Coordinated/Uncoordinated

(f) Deliberate/wavering

(g) Relaxed/Tense

(h) Calm/Anxious

Figure 2.13: Crowd-sourced assignment ratings are shown for each word-pair, with the highest correlated data metrics overlaid. Ratings and metrics are broken down between the four levels of expertise: “E:Expert”, “F:Fellow”, “I:Intermediate”, and “N:Novice”.

Table 2.7: Data Metrics Statistical Analysis Summary

	Subject (E,F,I, N)		Task (RR, SS)		Repetition (1-3)	
Source	*p*	Significance	*p*	Significance	*p*	Significance
Fluid/Viscous	0.0451	E > N	0.0071	SS > RR	0.7452	n/a
Smooth/Rough	0.0001	N > {E, F} & I > E	0.0016	RR > SS	0.0464	n/a
Crisp/Jittery	0.2387	E > N	0.4088	n/a	0.1305	n/a
Swift/Sluggish	<0.0001	{I, N} > {E, F}	0.0001	RR > SS	0.0598	n/a
Calm/Anxious	<0.0001	N > {E, F}	0.6861	n/a	0.7965	n/a
Relaxed/Tense	0.0002	{E, F} > N & E > I	0.8261	n/a	0.4730	n/a
Deliberate/Wavering	0.0001	N > {E, F}	<0.5101	RR > SS	0.6726	n/a
Coordinated/Uncoordinated	<0.0035	N > {E, F}	0.0243	RR > SS	0.7730	n/a
Overall Data Metrics	<0.0066	E > N	0.0375	RR > SS	0.3156	n/a

E – Expert, I – Intermediate, N – Novice, RR – Ring and Rail 2, SS – Suture Sponge 3 For the word pair ”calm-anxious” there was a significant difference in the level-task and level-repetition interaction (p = 0.0106 and 0.0001 respectively). For the word pairs ”smooth-rough”, ”crisp-jittery”, ”deliberate-wavering” there was a significant difference in the task-repetition interaction (p = 0.0136, 0.0101, and 0.0416 respectively).

Classification Results in K-Fold and LOUO Cross Validation

The performance of the Naive Bayes and SVM classifiers were compared. The validation was done using 10-fold and LOUO cross validation. Results from the classifications can be seen in Table 2.8. The classifiers' performances were evaluated for four different cases: four groups of expertise; three groups including experts plus fellows, novices, and intermediates; two groups including experts plus fellows and intermediate plus novices; and finally two groups including only experts and novices. In both classifiers the success rate was the best when distinguishing between only experts and novices (for Naive Bayes K-fold: 100± 0% and LOUO: 97± 9%, for SVM K-fold: 100 ± 0 % and LOUO: 94 ± 10 %) and worse for distinguishing among four levels of expertise (for Naive Bayes K-fold: 89± 12 % and LOUO: 83± 15 %, for SVM K-fold: 84 ± 11 % and LOUO: 75 ± 13 %).

Table 2.8: Expertise level classification success rate using metric data

	Naive Bayes		SVM	
	10-Fold	LOUO	10-Fold	LOUO
{E,F,I,N}	89 ± 12 %	83 ± 15 %	84 ± 11 %	75 ± 13 %
{E,I,N}	94 ± 8 %	93 ± 12 %	86 ± 12 %	79 ± 15 %
{EF,IN}	95 ± 7 %	94 ± 14 %	90 ± 7 %	90 ± 9 %
{E,N}	100 ± 0%	97 ± 9 %	100 ± 0 %	94± 10 %

2.3.8 Summary of Finding Metrics associated with Surgical Styles

In this section, we proposed a framework for objective skill level assessment based on metrics associated with stylistic behavior. For this purpose, we developed a stylistic behavioral lexicon based on the six domains described in the Global Evaluative Assessment of Robotic Skills (Goh et al., 2012). The lexicon was validated through crowd sourced analysis. The crowd ratings were highly correlated to faculty ratings for the adjective choices. To find quantitative metrics associated with each stylistic adjective pair, various motion and physiological measurements were gathered, and the best metrics were chosen through correlation with the crowd ratings. A classifier was trained based on the data metrics chosen from the previous step and proved to be able to distinguish the expertise levels.

2.4 Automatic Surgical Skill Rating Using Stylistic Behavior Components

A gold standard in surgical skill rating and evaluation is direct observation, which a group of experts rate trainees based on a checklist with a likert scale, by observing their performance during a surgical task. This method is time and resource intensive. Technical advances have enabled large amount of data acquisition for evaluation purposes. A vast amount of studies have focused on surgical skill assessment to replace direct observation which mainly focus on the identification of novices and experts through quantitative methods; however, very few studies have focused on automatic ratings of surgical skill based on the global rating skills. The global evaluative assessment of robotic skills (GEARS) consists of six domains which demonstrate the fundamental skills in robotic surgery (Fig. 1.3). These domains include depth perception, bimanual dexterity, efficiency, autonomy, force sensitivity and robotic control, and can be associated with a scale between 1 and 5, with 1 indicating a poor performance and 5 indicating a high performance. Brown et al. were among the first to study automatic scoring of a surgeoans skill based on GEARS scores. They used regression based and classification based methods to train a model using force, acceleration, and time measurements. These models were used to predict GERAS scores (Brown et al., 2017). In this section we evaluate our proposed lexicon's ability to automatically rate robotic surgical skill, based on the six domains in GEARS. We compare the score predictions using the stylistic behavior lexicon with the prediction results using features extracted from user kinematic data. This study was published in (Ershad et al., 2018a)

2.4.1 Methods and Experimental Setup

The Experimental setup for this study is the same as the setup discussed in section 2.3.1.

Feature Extraction

For each trial, the following features were extracted.

- Mean and standard deviation of linear velocity, angular velocity, linear acceleration and Jerk in x,y, and z direction from the IMU sesnsors.
- Mean and standard deviation of the norm of linear velocity, linear acceleration, and Jerk from the IMU sesnsors.
- Jerk trajectory from the IMU sesnsors from Eq. 2.5 as discussed in section 2.3.2.
- The maximum of average jerk norm within 1/8 time intervals of the overall task from the IMU sesnsors.
- The distance traveled by the right and left hand, right elbow and shoulder calculated as the sum of Euclidean distances between sample points (Anderson et al., 2012), from Eq. 2.3.
- The difference in distance traveled by both hands.
- The number of movements in right and left hand obtained from Eq. 2.4 as discussed in section 2.3.2

2.4.2 Predicting GEARS Scores

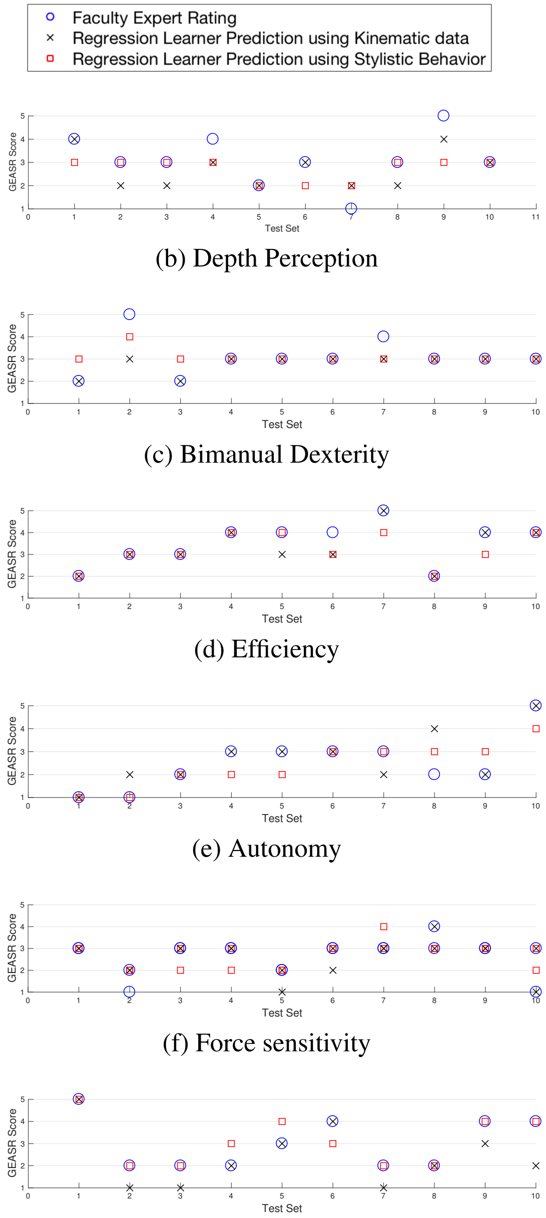
Two methods were compared to predict the GEARS scores. The first method tests our proposed lexicon in their ability to predict GEARS scores. For each domain in GEARS, a regression learner was trained using the crowd ratings for the words in the lexicon. Prior to training the regression learner, adjective selection was performed using stepwise regression. This was performed separately for each of the five GEARS domains. Faculty ratings of the videos based on GEARS provided a gold standard for the regression algorithm. We used leave one user out (LOUO) cross validation for model tuning and testing. In this method, all data from one user is left out while the classifier is trained on the remaining users' data, the trained classifier is then tested on the left out-user's data.

The second model uses the features extracted from the kinematic data gathered from sensors for the prediction. A similar approach was used here. Stepwise regression was used to select kinematic features for each domain in GEARS separately. A regression learner model was trained using the selected features for each domain separately and LOUO was used for model tuning and testing. The results from the two methods were compared.

2.4.3 Results

The average score for each subject was calculated through all trials from the same subject. The developed regression models were used to predict the GEARS scores. Since GEARS scores are integer numbers, the values were rounded. Fig. 2.14 shows the gold standard and predicted scores (between 1-5) for both regression models for 10 different test sets for each domain in GEARS. The developed regression models were used to predict the GEARS scores. The mean and standard deviation of prediction accuracy for the regression learner model using the kinematic data as well as the stylistic behavior lexicon is indicated in Table 2.9. The results show that both regression models are in agreement with the faculty ratings for all domains, with a slightly better performance for the regression learner model trained using the kinematic data, for all of the domains in GEARS except for ”Robot control”. This might be due to the fact that the extracted kinematic features do not fully represent this skill and the lexicon is more successful in capturing the information that the crowd raters grasp from observing the use's movements.

The kinematic model average precision was above 70 % for all all domains, and the stylistic behavior model average precision was above 69 % for all domains. The highest prediction accuracy from the kinematic data model was obtained for the ”Efficiency” domain at 89 ± 11 % and the lowest was obtained for the ”Robot control” domain at 70 ± 16%. For the stylistic behavior model, the highest accuracy was obtained for the domain ”Efficiency” at 86 ± 16%, and the lowest



(b) Depth Perception

(c) Bimanual Dexterity

(d) Efficiency

(e) Autonomy

(f) Force sensitivity

(g) Robot Control

Figure 2.14: Regression learner GEARS predictions (a score between 1-5), for 10 different test sets. Blue circles indicate the ratings from the faculty experts, black x's are the predicted scores from the regression learner model trained using the features extracted from the kinematic data, and red squares are predicted scores from the regression learner model trained using the stylistic behavior lexicon.

was obtained for the ”Autonomy” domain at 69 ± 13. The performance of both models were good;

Table 2.9: LOUO Corss-Validation Accuracy in predicting GEARS Scores

GEARS Domain	GEARS scores Prediction Accuracy Using Kinematic Data	GEARS scores Prediction Accuracy Using Stylistic behavior lexicon
Depth Perception	84 ± 5 %	78 ± 14%
Bimanual dexterity	88 ± 23%	77 ± 18 %
Efficiency	89 ± 11%	86 ± 16%
Force Sensitivity	80 ± 14%	72 ± 10%
Autonomy	81 ± 8%	69 ± 13%
Robot Contorl	70 ± 16%	74 ± 21%

however, the models performed better in some domains compared to the others. This suggests that each individual skill is associated with unique quantitative and qualitative metrics.

2.4.4 Summary of Automatic Surgical Skill Rating Using Stylistic Behavior Components

In this section we evaluated our stylistic behavior lexicon in their ability to predict the GEARS scores. For each domain in GEARS, we trained one regression learner model using the stylistic behavior lexicon and one using the features extracted from user's kinematic data. The results show that the prediction scores in both cases are in agreement with the gold standard faculty ratings. This study contains preliminary results indicating that the stylistic behavior lexicon and features extracted form user motion can be used for automatic rating of surgical skill. Thus, alleviating the need for direct observation by a team of experts to rate a trainee's performance.

2.5 Conclusion

We have proposed a novel method for surgical skill assessment using semantic labels. These labels will be the basis of a future automated coaching system. We hope to extend these methods to other aspects of surgical training, such as: open and laparoscopic skills, team dynamics, patient interactions, and professionalism.

Metrics associated with the semantic labels were calculated from kinematic and physiological measurements; however, the metrics were selected using a time-consuming process of estimating and extracting features. The kinematic and physiological metrics selected might not completely capture the rich information available in these data channels. To overcome these limitations, the next chapter will focus on automatic feature extraction using data driven models which are fed with the higher dimensional raw data to learn the effective features.

This study was carried out on a robotic simulator; however, the simulated environment may not fully represent actual surgical environment and subjects might not be as engaged with the simulation exercise when compared to interaction with a physical robot. In future work, we will study subjects by implementing our proposed method on a *da Vinci*® robot and will evaluate different forms of stylistic feedback to the user including visual, haptic, and verbal cues. Finally, the methods proposed in this chapter might be relevant to other training domains, such as rehabilitation or sports coaching.

CHAPTER 3

AUTOMATIC REAL-TIME SKILL LEVEL AND STYLISTIC BEHAVIOR DETECTION

3.1 Introduction

Automatic skill evaluation is of great importance in surgical robotic training. Machine learning and data driven features play an important role in automating the assessment. A key step in automation using machine learning is to be able to extract meaningful information from the recorded data. Simulators record large amount of movement data which provide rich source of information; however, kinematic data is highly redundant thus, extracting useful information for a better representation is essential. Several dimensionality reduction and feature extraction methods have been proposed to reduce the redundancy in kinematic data and provide an efficient representation. One widely used method for this purpose is dictionary learning methods. Dictionaries are a set of basis vectors that can be used for a new representation of a data set which can bring out features that are not identifiable in the raw representation of the data. Principle component analysis (PCA) is a transformation technique that allows learning a complete dictionary (set of basis vectors). PCA is widely used for feature extraction in human motion analysis and activity recognition (Jeppsson, 2017; Yong et al., 2013; Jarque-Bou et al., 2019; Mollazadeh et al., 2014; Ali and Shah, 2008; Milovanović and Popović, 2012). It finds the direction of maximum variance in the data set and projects the data onto the new subspace of maximum variance components with orthogonal axes. Human motion is constrained by body dynamics and lies in low dimensional manifolds (Scholz and Schöner, 1999; Latash et al., 2002); thus, it can be encoded by sparsity through an over-complete dictionary. One advantage of using an over-complete set of basis vectors is its ability to better extract underlying information from the data set. Dictionary learning and sparse coding has been used for human activity recognition (Bhattacharya et al., 2014; Johnson and Ballard, 2014). In this chapter, we use sparse coding for detecting the quality of movement based on styles of movement.

Extensive research has been done to automatically evaluate surgical skill as discussed in section 1.2.2. These methods, while providing great results in differentiating expertise levels, do not provide any information to the user on how to modify movement to improve performance. Being able to detect the differences in the way a novice moves vs. the way an expert moves, can help provide meaningful feedback to the novice on how to move differently and thus improve results.

In the previous chapter, we leveraged the fact that skill assessment can be done by simple intuition, and introduced a novel concept of distinguishing expertise levels based on user's stylistic behavior which is visible in the movement while performing a task. The stylistic behavior of a subject can represent the ease, efficacy, and expertise with which an individual executes some complex manipulation task. To capture some of these behaviors in robotic surgery, we previously developed a lexicon of positive and negative stylistic descriptors (Table 2.3). We showed that the style of movement defines the user's skill level and is easily distinguishable by non-expert observers. We also proposed some metrics associated with each stylistic behavior, and studied the effectiveness of these metrics in skill level assessment. The metrics were extracted from multiple kinematic and physiological measurements including position, linear velocity, angular velocity, linear acceleration, electromyography (EMG), and galvanic skin response (GSR). This required extensive expert knowledge to select a limited number of metrics. Feature learning from the raw kinematic data, as opposed to knowledge-based feature extraction, can infer important latent information and eliminate limitations of human selected features.

In this chapter we aim to eliminate metric extraction, and detect the quality of performance based on each style automatically. We first test some automatic feature extraction methods in their effectiveness in distinguishing different expertise levels. Then we aim to automatically identify user's performance based on their stylistic behavior. We propose a method that automatically captures underlying structures in raw movement data in near real time, while eliminating the need for complex feature engineering and expert knowledge for extracting metrics as discussed in the previous chapter. We propose a dictionary learning method to find basis vectors which can uncover hidden underlying structures in raw kinematic data to model stylistic behavior. For each stylistic behavior adjective, first a codebook is learned. Then, feature extraction is done using the learned codebook. These features are used to train a classifier which is then evaluated and tested for performance.

3.2 Surgical Skill Level Assessment using Automatic Feature Extraction Methods

Automatic feature extraction methods can make use of the whole kinematic data set instead of considering limited features selected by expert knowledge. Feature learning from sequential kinematic data for the purpose of human activity recognition has been studied in various applications and has proven to be efficient in activity recognition from motion sensors (Bhattacharya et al., 2014; Mazilu et al., 2013). These methods reveal meaningful information in the data without requiring explicit domain-based knowledge, and can be used to extract a reduced subset of new features from the original high-dimensional data and discover unknown patterns form low-level sensor data.

Inspired by studies in human activity recognition and to deal with feature redundancy in kinematic data, we aim to study the effectiveness of three automatic feature extraction methods, which provide high quality discriminant features, in detecting underlying information from time-series kinematic data in distinguishing different surgical expertise levels. These methods include principle component analysis (PCA), independent component analysis (ICA), and linear discriminant analysis (LDA). All three methods are linear transformation techniques which reduce the dimension of the original input data. PCA aims to eliminate the information redundancy. ICA estimates components as statistically independent as possible. LDA improves the separability of samples in the subspace.

These automatic feature extraction methods are used to extract a reduced subset of new features from the original high dimensional raw data and discover unknown pattern form low-level sensor data.

The resulting features are used as input to four different classifiers to evaluate the ability of classification based on these features. In addition, for the purpose of real-time user proficiency level detection, the minimal time interval for accurate classification of skill levels is studied. This is an important parameter for the development of future feedback methods in surgical training. This study was published in (Ershad et al., 2018)

3.2.1 Experiment Setup

In this study, the position data recorded form EM trackers (trakSTAR, Model 180 sensors, Northern Digital Inc., Ontario, Canada) at a sampling rate of 256 Hz. was used from the experiment setup discussed in section 2.3.1. The movement of the shoulder, elbow, and wrist from the dominant hand and positions from both hands in x, y, and z directions (15 measurements total) were recorded. 14 subjects including four experts (practicing robotic surgical faculty), three fellows (surgical fellows post residency), three intermediates (PGY-4 surgical residents), and four novices (medical students) were recruited to perform two simulated tasks (suture sponge and ring and rail) on the *da Vinci*® *Skills Simulator*™.

3.2.2 Kinematic Data Windowing and Real-Time Evaluation

For the purpose of detecting and correcting trainees' stylistic behavior during the performance of a task, as opposed to providing feedback about their performance at the end of the task, we considered feature detection on intervals of the acquired kinematic signals while performing a task. For each trial, the kinematic data were normalized and windowed. To avoid significant ringing from the windowing process we applied a Hanning envelope along the time dimension of each window so that movements could be decomposed into overlapping windows. Different sizes of time intervals were considered including: 0.04 s (10 samples), 0.25 s (64 samples), 0.4 s (100 samples), 1 s (256 samples), 2 s (512 samples), for the purpose of finding the minimal time interval for real-time detection and control. 15 position measurements for each sample was considered and arranged in n-dimensional vectors where $n = 15 \times num.samples = 150, 960, 1500, 3840, 7680$, respectively, for each chosen time frame. The minimal window size which could accurately predict user stylistic behavior was suggested. Specifically, we considered measurements of the form $x_i \in$

R^n, where x_i is a vector containing all measurements within the ith frame and n is the length of the frame, i.e., the unlabeled measurements are represented as the set $X = \{x_1, x_2, ..., x_k\}, \quad x_i \in R^n$

3.2.3 Feature Extraction Methods

To deal with feature redundancy in kinematic data, feature extraction methods were used which provide high quality discriminant features. The data was separated into two groups of train and test data sets. The training data includes data from each of the four expertise levels. Each of the feature reduction methods (PCA, ICA, LDA) were applied on the whole training data matrix, obtained by arranging all the n-dimensional vectors in the training set. This yields a transformation matrix, which was then applied on the test data to get the test data on the same feature space as the training data.

3.2.4 Classification and Cross-Validation

Four classifiers including Naive Bayes, Support Vector Machine (SVM), Nearest Neighbor (NN), and Decision Tree were used to evaluate the performance of the extracted features, in distinguishing among different levels of expertise. Performance of the classification methods was examined using a 10-fold cross validation.

3.2.5 Results and Discussion

Feature Extraction and Classification Results

Figure 3.1 shows the cumulative variance explained results of PCA. As indicated by the figure, the first 15 principle. components account for more than 95% of the variance. Thus, resulting in reduction of feature size from 900 to 15 features. The results of a 10-fold cross validation for four classification methods for each of the feature extraction methods are shown in Table 3.1. The results indicate that PCA features, in combination with the SVM classifier, are most effective in classifying among expertise levels from arm joints and hand position data.

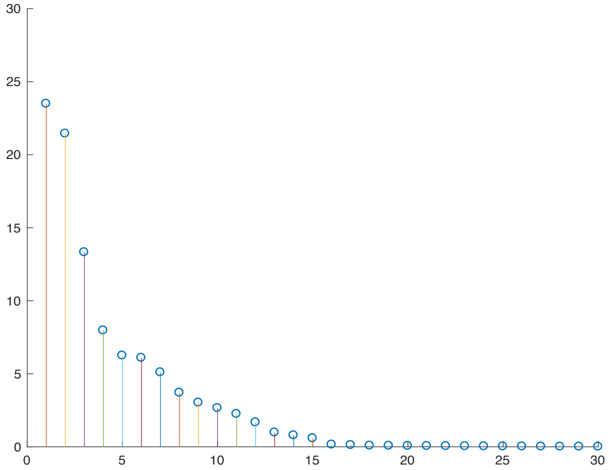


Figure 3.1: Variance explained of all principal components.

Table 3.1: Classification Success Rate

Classifier / Feature Extraction	SVM	Nearest Neighbor	Decision Tree	Nave Bayes
PCA	98 ± 0.21%	97 ± 0.40%	88 ± 0.71 %	65 ± 0.67 %
ICA	88 ± 0.38%	74 ± 1.12%	65 ± 0.80%	54 ± 1.94%
LDA	43 ± 0.33%	63 ± 1.49%	58 ± 1.74%	64 ± 1.04%

3.2.6 Classification Results for Different Time Intervals

Figure 3.2 compares the SVM classification accuracy for PCA features in different time intervals. This shows that the highest accuracy is achieved in 0.25 sec with 64 samples.

3.2.7 Summary of Surgical Skill Level Assessment using Automatic Feature Extraction Methods

We proposed a method for detecting user's surgical proficiency level, within short time intervals. We studied the effectiveness of three feature extraction methods from hand position movements during two robotic surgical training tasks, for the purpose of finding underlying patterns in kine-

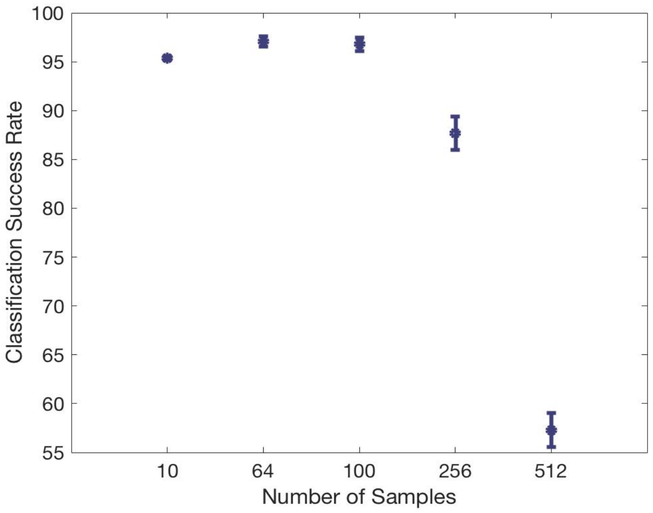


Figure 3.2: SVM classification for PCA for different number of samples.

matic movement data to distinguish among different expertise levels. The results indicate that among the three feature extraction methods studied, PCA is the most effective. This implies that the redundancy in kinematic data can be removed by PCA features which extract linearly uncorrelated features from the original low-level data set. Also, PCA features are able extract useful underlying information from kinematic data for distinguishing trainee's proficiency level. The minimal time frame size for accurately classifying different expertise levels using PCA features has also been studied. Automatic feature extraction using windowed kinematic data can be useful in real-time assessment of skill level for the purpose of real-time feedback and customized training in simulator based training.

3.3 Automatic and Near Real-time Stylistic Behavior Assessment in Robotic Surgery

In this section, we aim to automatically identify users' performance based on their stylistic behavior in near-real-time without the requiring expert knowledge feature extraction.

For this purpose, we first leverage the time and resource efficiency of crowd-sourcing in evaluating surgical videos based on our proposed lexicon(Table 2.3). Then, we use this data to train multiple models, each representing one of the stylistic behavior pairs. The models are learned from the raw kinematic data through sparse dictionary learning, a method which represents the input data as a linear combination of a set of basis vectors. This new representation of the data makes the problem of classification easier. Using short time intervals enables near-real-time error detection, which could identify deficiencies in the movement as soon as they appear. This also leads to feedback methods during training, rather than the more traditional methods of post-task feedback. Assessing surgical skill based on the style of movement can also help provide customized feedback aimed at training to mastery of a task, not merely proficiency. This study was published in (Ershad et al., 2018b).

3.3.1 Method Overview

An overview of our stylistic behavior detection method is shown in Fig. 3.3. After collecting the data, we reduce the dimensionality of the kinematic data using PCA. Then, a set of basis vectors that capture characteristic patterns of user's movement are found. These vectors enable transforming sensor data to the feature space spanned by the basis vectors in the codebook. Once the basis vectors are learned, the training data is used to find the sparse codes (coefficients of the basis vectors) which are then used to train a classifier. After training the classifier, the stylistic behavior in new sensor movement recordings can be recognized by transforming the new data into the same feature space and applying the learned classifier. Detailed discussion for each step is provided here.

3.3.2 Dimensionality Reduction

Time series sensor data, which in our experiment corresponds to 3D position data, is used in this study. To remove redundancy in kinematic data, and thus reduce the computation time and

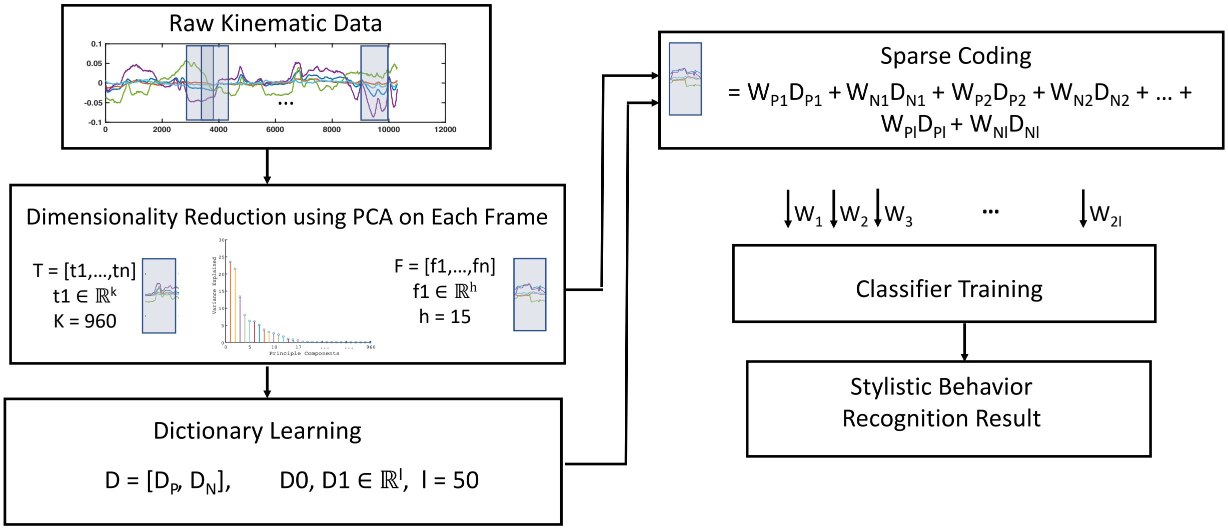


Figure 3.3: Proposed algorithm for one stylistic behavior adjective: Feature reduction using PCA is applied on each frame of the raw kinematic data. A codebook is obtained using this feature reduced kinematic dataset. This codebook is used to extract sparse features, which are then used to train a classifier. The raw test data goes through the same feature reduction and basis vector transformation process.

complexity, PCA is used. PCA has been widely used for dimensionality reduction in kinematic data, which provides a linearly uncorrelated set of basis vectors that reduce correlation in the dataset, while preserving variability (El Moudden et al., 2016; Karg et al., 2009; Ershad et al., 2018b; Kirkwood et al., 2011; Milovanović and Popović, 2012). We apply PCA on the training data to find a reduced dimension feature space. The test data is also transformed to the same feature space.

3.3.3 Codebook Learning

Predefined dictionaries are widely used in signal decomposition (Mallat, 1989; Dhamala et al., 2008); however, designing and adapting a dictionary based on a desired dataset better captures complexities and latent structures in the data (Hoyer, 2002). In this study, a set of over-complete basis vectors are learned from the sensor data. The new representation of the data, based on the learned basis vector, is sparse which means that there are few non-zero components in the representation.

To represent the good or poor performance of each stylistic behavior adjective, a codebook is learned from the kinematic data. Two types of basis vector sets are learned for the negative (D_N) and positive (D_P) performance of each adjective pair. The total codebook $D_{total} = \{D_P, D_N\}$ is then the concatenation of these two separate basis vector sets, which is used to find the sparse codes in each time frame (Wright et al., 2009). The positive and negative labels used for learning the codebook are obtained through crowd-sourced ratings for each trial. This method is further discussed in Section 3.4. Once the over-complete codebook is constructed, any input sensor recording can be represented as a linear combination of the basis vectors in the codebook:

$$T = \sum_{i=i}^{n} w_i d_i \tag{3.1}$$

where T is the input signal, n is the number of basis vectors in the codebook, w_is are the sparse code coefficients, and d_is are the basis vectors. The codebook learning problem is to optimize:

$$min \sum_{i=1}^{k} ||T_i - Dw_i||_2^2 + \lambda ||w_i||_1 \tag{3.2}$$

which contains two problems: 1) finding the codebook $D = d_1, ..., d_n$ such that the reconstructed signal Dw_i is as similar as possible to the original signal T_i (i.e., $||T_i - Dw_i||$ is minimum), 2) finding the activation coefficients $W = w_1, ..., w_k$, such that they are sparse. The parameter λ controls the trade off between these two constraints. To learn the codebook, we used the K-SVD algorithm which is an iterative method that alternates between the two processes mentioned above (Aharon et al., 2006). The over-complete codebook enables better capturing of the underlying structures in the data compared to complete codebooks (i.e., PCA). The sparsity results in a small subset of the over-complete basis vectors to be activated; thus, resulting in a reduced computational complexity.

3.3.4 Feature Representation and Classification

Once the codebook has been learned ($D_{total} = \{D_1, D_2\}$), sparse codes for the reduced feature training dataset are calculated by optimizing the following equation using the LARS-Lasso algorithm (Efron et al., 2004).

$$\hat{w_i} = argmin||T_i - D_{total} w_i||_2^2 + \lambda ||w_i||_1 \tag{3.3}$$

To find the sparse codes for a new set of measurements, the reduced feature dataset is mapped to the space represented by the basis vectors in the codebook, and the set of sparse coefficients, $\hat{w}_i$, is calculated. The sparse codes from the training set are used to train a classifier for each of the stylistic adjectives.

3.4 Experimental Study

3.4.1 Data Collection

The subject recruitment and simulated tasks used in this experiment are similar to those described in the previous chapter discussed in detail in Section 2.3.1. The raw kinematic movement data were obtained from positions of the shoulder, elbow, and wrist from the dominant arm, and positions from both hands in x, y, and z directions (15 measurements total). Wrist, elbow, and shoulder joints are shown to be effective in arm posture analysis in robotic surgery (Nisky et al., 2013). Hand motion data has also been the primary data set for surgical skill evaluation (Datta et al., 2002, 2001; Smith et al., 2002; Hayter et al., 2009; Darzi and Mackay, 2001). The positions of the shoulder, wrist, and hands were measured directly using EM trackers at a sampling rate of 256 Hz; however; due to metallic interference from the armrest on elbow measurements, the elbow position was estimated using measured shoulder and wrist positions (Nisky et al., 2014).

The sensor positioning is shown in Fig. 3.4. Videos were also recorded for crowd-source assessment. The crowd assessment of the videos is described in detail in in the previous chapter in section 2.3.1. Robotic Operating System (ROS) was used for synchronizing the data collected from the cameras and the EM trackers.

Two datasets are important in this study: (1) ground-truth stylistic behavior labels and (2) the raw kinematic data for learning. The ground-truth labels were obtained through crowd-sourced assessments. A positive label was assigned to a trial if it was rated as positive by more than or

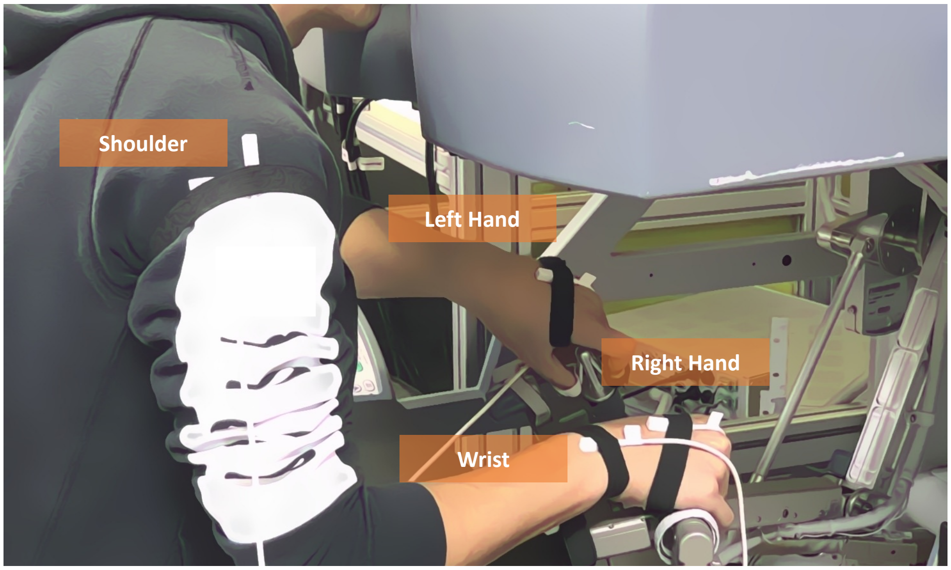


Figure 3.4: Sensor Positioning: Human subject conducted simulated tasks using the da Vinci Standalone Simulator while wearing an electromagnet tracker

equal to 50% of the crowd-workers and a negative label was assigned if it was rated as positive by less than 50% of the crowd-workers.

3.4.2 Data Pre-Processing

Frame extraction

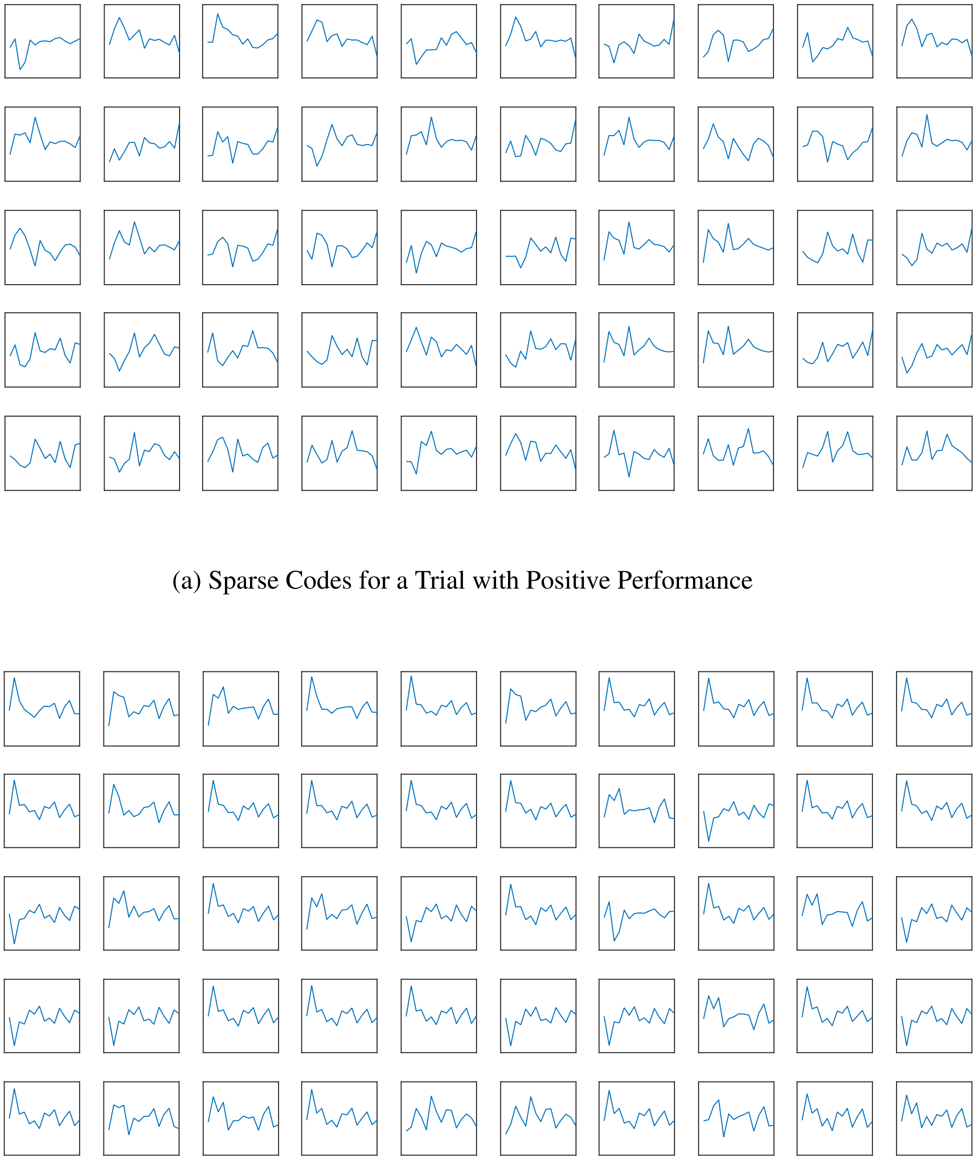
For each trial, the sensor data were normalized and windowed into small time intervals of 0.25 s. Each data frame containing 64 samples for 15 signal measurements (X, Y, and Z positions for each of the five electromagnet trackers) was arranged into a 960 dimension vector. To avoid significant ringing from the windowing process, we applied a Hanning envelope along the time dimension of each window so that movements could be decomposed into overlapping windows. Using short time intervals enables near real-time stylistic behavior recognition.

Dimensionality Reduction

For the training data, PCA was applied on the whole training data matrix that was obtained by arranging all the 960-dimensional vectors in the training set into the matrix. This provides a transformation matrix which is used to project the test data frames onto the same space as the training dataset. Input frames of position readings are projected onto the linear PCA subspace, that retains at least 95% variance of the data resulting in significant reduction in dimension size form 960 to 15-dimensional data vectors. Removing the correlations among the position data within each window improves the coding process in sparse coding.

3.4.3 Codebook Learning and Sparse-Code Feature Extraction

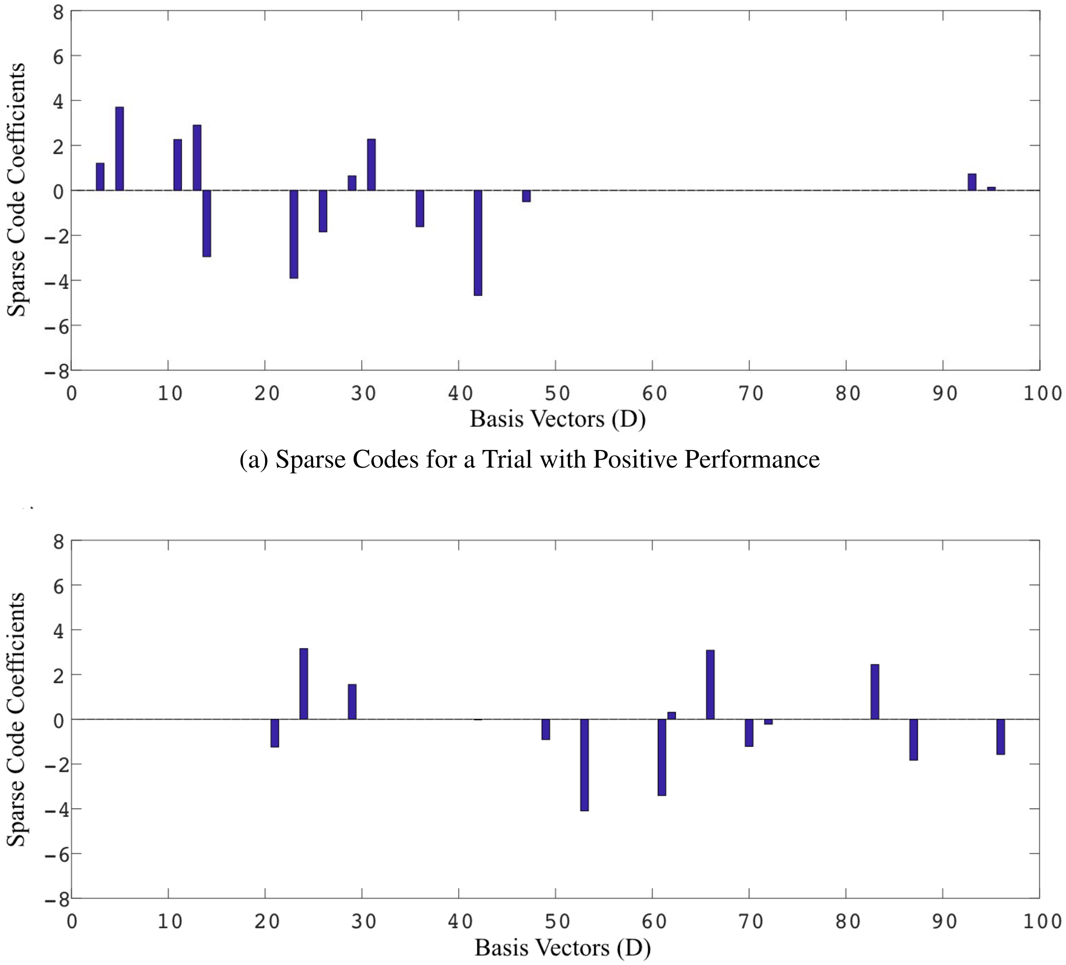
Based on the dictionary learning method for recognizing stylistic behavior, as described in section 3.3.3, we derived a codebook of basis vectors from the training dataset. We used the crowd-sourced ratings of positive or negative performance of each adjective as labels for each trial to capture the stylistic behavior characteristics in the sensor measurements. Eight different codebooks were learned, one for each adjective pair. The positive and negative codebook learned for one stylistic behavior (Fluid/Viscous) is shown in Fig. 3.5. Once the codebooks were learned using the training dataset, new frames of input data were projected onto these set of basis vectors and the sparse codes for each codebook was obtained using the sparse coding method explained in section 3.3.4. The new representation of the input signal using the learned over-complete codebook is sparse. For a trial labeled with positive performance for a specific adjective, the sparse code representation of that signal indicates most non-zero coefficients for the positive basis vector (D_P), and for the trial labeled with negative performance, the negative basis vector (D_N) contains most non-zero coefficients. The coefficients of the codebook for two randomly chosen positive and negative trials of one adjective pair (Fluid/Viscous) are shown in Fig. 3.6.



(a) Sparse Codes for a Trial with Positive Performance

(b) Sparse Codes for a Trial with Negative Performance

Figure 3.5: Examples of dictionaries learned from two trials. Plot (a) shows a the basis vectors learned from a positive performance. Plot (b) shows a the basis vectors learned from a negative performance. The dictionary used in this study includes 50 basis vectors for each of the two types of positive and negative performances. The total code book is the concatenation of these two dictionaries with 100 basis vectors.



(a) Sparse Codes for a Trial with Positive Performance

(b) Sparse Codes for a Trial with Negative Performance

Figure 3.6: Examples of sparse codes from two trials. D_{1-50} are the basis vectors related to the positive performance, and D_{51-100} are the basis vectors related to a negative performance. Plot (a) shows a trial with positive performance and has most non-zero coefficients for the first 50 basis vectors. Plot (b) shows a trial with negative performance and has most non-zero coefficients for the second 50 basis vectors.

3.4.4 Classification

A Support Vector Machine (SVM) classifier was trained for each of the stylistic behavior adjectives using the training data. We considered radial basis functions (RBF) as the Kernel function of the SVM classifiers. The optimal parameters of RBF kernal (C and γ) were tuned using a grid search through 10-fold cross validation. The C parameter controls the trade off between a smooth decision boundary and the classification accuracy, and the γ parameter is the inverse of the standard deviation of the RBF kernel (Gaussian function). Various values of (C and γ) were tested and the one with the highest cross validation accuracy was selected. The trained classifier was then used to evaluate stylistic behavior performance for the test dataset.

3.5 Results and Discussion

3.5.1 Dimensionality Reduction

The cumulative variance results of PCA is shown in Fig. 3.7. The first 15 principal components account for more than 95% of the variance. Thus, resulting in reduction of feature size from 960 to 15 features. Finding sparse features for the reduced dimension dataset is more time efficient compared to obtaining features for the original dataset. This enables feature extraction in real-time.

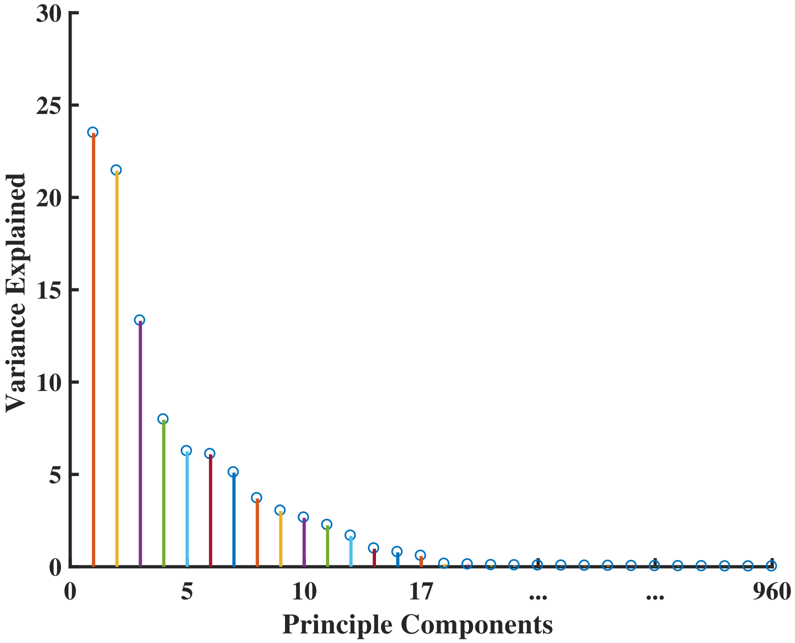


Figure 3.7: Variance explained of all principal components

3.5.2 CodeBook Size

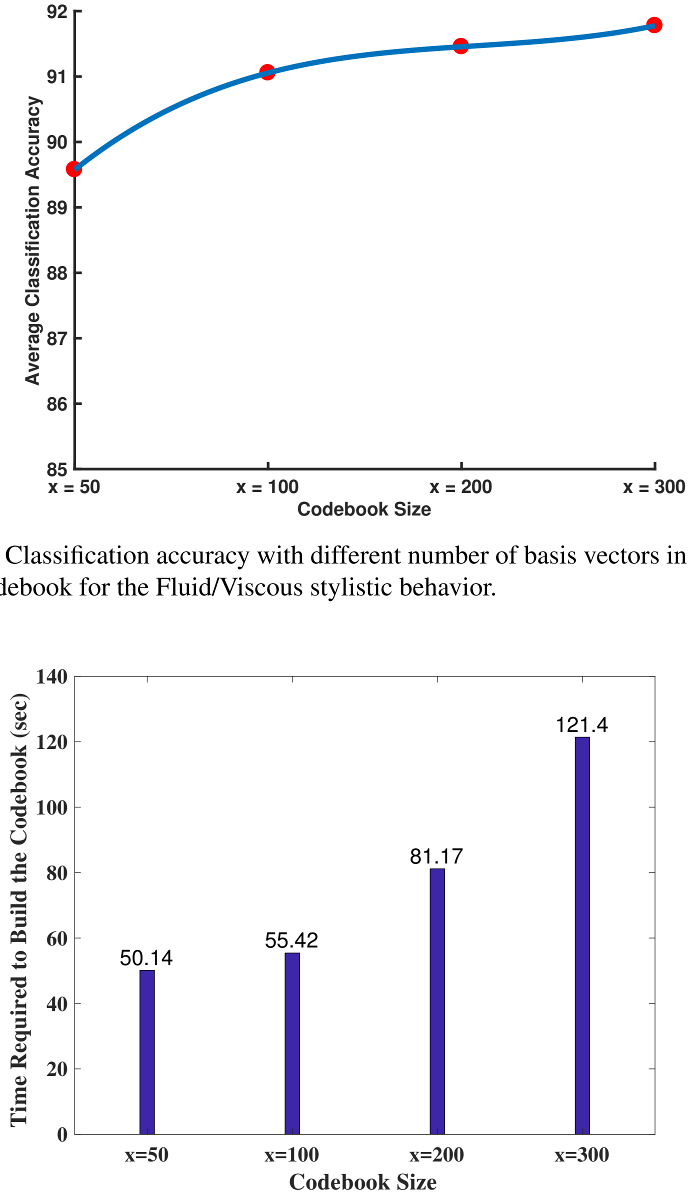
For our sparse coding approach, we studied the effectiveness of different codebook sizes on the classification results. An example of the changes in classification accuracy by changing the number of basis vectors in the codebook for one stylistic behavior is shown in Fig. 3.8(a). Codebook sizes (i.e., number of basis vectors) of 50, 100, 200, 300 were considered.

By increasing the codebook size form 50 to 100 the classification accuracy changes from 89.51 ± 3.82 % to 91.05 ± 4.02 %, and the computational time changes from 50.14 s to 55.42 s. However, further increasing the size of the codebook does not show significant improvement in the performance of the classification (91.05 ± 4.02 % to 91.45 ± 3.55 %, and 91.77 ± 3.33 %), but significantly increases the computational time (50.14 s to 81.17 s, and 121.40 s) as shown in Fig. 3.8(b). Thus, we chose n = 100 for the number of basis vectors used in the codebook.

3.5.3 Classification Performance

Table 3.2 shows the classification accuracy on the test data for each stylistic behavior adjective.

Three classifier models using three different datasets (raw data, PCA features, PCA + sparse coding features) were trained and tested for each stylistic behavior adjective. The results indicate improved performance when using sparse coding with PCA features compared to only using PCA features, or raw data. Our proposed method shows a mean improvement ranging from 18.7% - 68.5% increase in classification accuracy compared to using raw data, and a mean improvement ranging from 1.35 % - 20.59 % increase compared to using PCA feature.



(a) Classification accuracy with different number of basis vectors in the codebook for the Fluid/Viscous stylistic behavior.

(b) Time taken to learn codebooks with different numeber of basis vectors for the Fluid/Viscous stylistic behavior.

Figure 3.8: Tradeoff between classification accuracy and time taken to learn codebooks for different number of basis vectors in the codesbook. The classification accuracy increases with the number of codebooks; however, by increasing the number of basis vectors above n=100, very little increase in accuracy (0.4% (a)) is acquired at a significant increase in computational time (46% (b)). Thus, n=100 is chosen as the optimum number of basis vectors.

Table 3.2: Classification Success Rate

Stlystic Behavior	Raw Data	PCA Features	PCA + Sparse-coding Features
Fluid/Viscous	66.06 ± 2.03%	78.56 ± 2.39%	91.05 ± 4.02 %
Smooth/Rough	47.79 ± 1.78%	70.38 ± 2.14%	78.15 ± 4.00%
Crisp/Jittery	54.53 ± 2.34%	72.53 ± 1.90%	75.05 ± 2.56%
Swift/Sluggish	59.84 ± 3.54 %	58.90 ± 3.83 %	71.03 ± 3.19%
Calm/Anxious	74.50 ± 3.06%	91.85 ± 2.16%	94.75 ± 2.12%
Relaxed/Tense	58.37 ± 2.82%	91.23 ± 2.36%	98.36± 1.02%
Deliberate/Wavering	81.84 ± 2.89%	97.19 ± 1.02%	98.50 ± 0.72 %
Coordinated/Uncoordinated	51.42 ± 1.34%	68.83 ± 2.89%	72.03 ± 3.28%

3.6 Conclusion

In this study, we proposed an automatic and near real-time frame-work for recognizing the quality of stylistic behavior performance in surgical skill. Our approach eliminates the time consuming feature engineering process which requires domain-specific knowledge. For this purpose, we used machine learning and data driven feature extraction methods. For each stylistic adjective, a codebook consisting of basis vectors for the positive and negative performance in short-time intervals was built using the kinematic data and crowd-source assessment. This codebook provides a sparse representation of the data as a linear combination of the basis vectors. The advantage of this method is that the dictionary was inferred from the raw input data as opposed to predefined dictionaries (i.e., wavelet or Fourier transforms) which are not adapted to a specific dataset. In addition, the sparsity of the new data representation reduces the computational complexity. The sparse codes (coefficient of the basis vectors) were then used to train a classifier to recognize positive or negative stylistic behavior performance. We used short-time intervals (0.25 s) in this study thus enabling the recognition of quality of performance and as a result, the detection of deficiency in the style of movement in near real-time.

The most time-consuming aspect of this method is building the codebook; however, this does not need to be done in real-time, and once the codebok is learned the sparse coding can be done in real-time. The time needed for constructing the codebook depends on the number of basis vectors in the codebook, the window size, and number of samples in the window. For reducing the time and complexity of building the codebook, we used PCA to reduce the dimensionality of the data by removing redundancy in the data.

This study used position measurements from users' limbs while performing a task on the *da Vinci® Skills Simulator™*. Future studies will focus on applying the proposed method to data from robot joints and end effectors and integrating it with a da Vinci Research Kit (dVRK) (Kazanzides et al., 2014) to be able to detect the poor performance without body-worn sensors and eventually provide trainees with relevant feedback to improve performance.

In the next chapter we will use the near-real-time stylistic behavior detection method proposed in this chapter to detect the deficiencies in movement as soon as they appear and apply feedback associated with the descriptors to help correct movement styles

CHAPTER 4

ADAPTIVE SURGICAL ROBOTIC TRAINING USING REAL-TIME STYLISTIC BEHAVIOR FEEDBACK THROUGH HAPTIC CUES

4.1 Introduction

Surgical outcomes are highly dependent on surgeon skill levels. Efficient training that provides trainees with appropriate feedback and assists them to achieve expert-like performance is critical for mastering technical skills in surgery, as well as achieving successful outcomes for the procedure and the patient (Curry et al., 2012). Traditional methods in surgical training typically involve an expert observing and evaluating a trainee's performance in simulated surgical tasks or in the operating room, and providing feedback to the trainee on how to improve performance (Cameron, 1997). Automating skill assessment can alleviate the time intensiveness and subjectiveness of these methods; however, finding an effective and efficient feedback method, which is intuitive and easy for the user to understand, is crucial (Hoffman et al., 2015).

For patient-free and more objective training environments, virtual reality (VR) simulators have begun to find their way into surgical training (Gallagher et al., 2005; Badash et al., 2016). Simulators provide factual and quantitative data to the the human user after the completion of each simulated task, such as number of instrument collisions, time to complete the task, and the number of missed targets. These data indicate the success rate of the trainee but do not necessarily provide them with more meaningful feedback on how to modify their movements to improve performance (Sewell et al., 2008).

An ongoing advancement in surgical simulators is to provide the trainee with real-time feedback based on a deficiency in their performance. Errors are computed from the users interaction with the virtual environment and feedback is provided to the user to improve results. Haptic feedback has been widely used for training purposes in simulators (Pezzementi et al., 2008; Halabi et al., 2013; Ko et al., 2017; Luo et al., 2016); however, the feedback is very task-dependent and, to our knowledge, does not provide any feedback related to the style of movements.

Adaptive training is beginning to be explored in virtual reality surgical simulators, developed with the idea that providing relevant and customized training tasks and and feedback to trainees, based on individual strengths and weaknesses, could enhance learning outcomes. The large amount of data recorded and stored by VR simulators enables data-driven analysis and automatic performance evaluation. This enables adaptive training based on each individual's needs (Vaughan et al., 2016). An example of an adaptive robotic surgical training is presented in (Mariani et al., 2018). This study compares training with adaptive curriculum to self-managed training and shows significant improvement in performance and learning skill using an adaptive framework. However, these performance assessment and adaptive feedback methods are largely task-dependent, which limit the generalizability of these approaches.

In this section, we propose a framework to provide task-independent stylistic feedback to the human user during movement-based or surgical training tasks to provide the user with a more intuitive and global understanding of their stylistic deficiencies. We designed, implemented, and evaluated an adaptive training method composed of the following elements: (1) Our proposed framework first evaluates the user's stylistic behavior performance in near-real-time and detects deficiencies in some movement styles as discussed in chapter 3. (2) Next, it provides the user with haptic cues to modify movement to improve performance. A secondary aim of this study is to evaluate the effectiveness of three common types of haptic feedback, namely, spring, damping, and spring-damping feedback that is computed from prior user positions and velocities.

This chapter is structured as follows. In section 4.2 our proposed stylistic assessment and feedback method is discussed in detail. It includes a deficiency detection phase and a feedback applying phase. A deficiency in style is detected from user hand position and velocity data, by comparing to expert style for a variety of stylistic adjectives. Then, the user is provided with either spring, damping, or spring-damping force feedback, depending on their randomized assignment of our feedback groups. We evaluate the effectiveness of our adaptive stylistic force feedback using both performance metrics as well as stylistic changes over the duration of the experimental study. Section 4.3 describes the experiment design and tools used to conduct the experiment. In section 4.4, we present the results of the proposed training method, and discuss the effect of the different types force feedback on styles of movement. Section 4.5 concludes this chapter and suggests the future research in this field. This study was submitted to and is under review in IEEE Transactions on Haptics.

4.2 Methods

Our goal is to improve robot assisted training to help achieve mastery in surgical robotics. For this purpose, we aim to (1) introduce a customized framework in which each individual is provided training based on his/her stylistic deficiencies, (2) provide the trainee with feedback in a timely manner and in near-real-time, (3) introduce a generalizable and task-independent framework which evaluates performance based on the user's style of movement, and (4) develop a more understandable and intuitive way to communicate with the user on how to modify movement to improve performance.

A systematic framework for recognizing the quality of movement through stylistic behavior and applying appropriate feedback for correcting the style was developed using a human machine interface (i.e., a haptic device) and a simulated task. Fig 4.1 shows the block diagram of the proposed method.

4.2.1 Surgical Skill Assessment Using Stylistic Behavior

In Chapter 2, we presented a novel surgical skill assessment method based on the surgeon's stylistic behavior. We proposed a lexicon of surgical styles through collaboration with expert surgeons (Table 2.3), and evaluated the ability of stylistic descriptors to differentiate different expertise levels using metrics correlated to crowd-sourced assessments of surgical style in training videos. In Chapter 3, we proposed an automatic method for detecting the quality of performance, based on these behavioral styles and within 0.25 seconds of movement. In this chapter, we design and

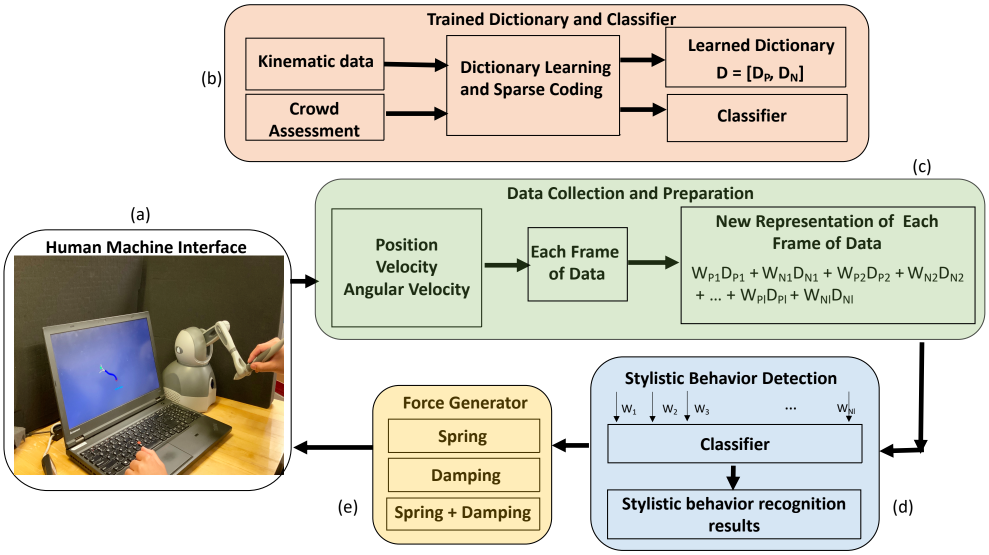

Trained Dictionary and Classifier
(b)
Kinematic data
Crowd Assessment
Dictionary Learning and Sparse Coding
Learned Dictionary $D = [D_P, D_N]$
Classifier
(c)
Data Collection and Preparation
(a)
Human Machine Interface
Position
Velocity
Angular Velocity
Each Frame of Data
New Representation of Each Frame of Data
$W_{P1}D_{P1} + W_{N1}D_{N1} + W_{P2}D_{P2} + W_{N2}D_{N2} + \ldots + W_{PI}D_{PI} + W_{NI}D_{NI}$
Stylistic Behavior Detection
W_1 W_2 W_3 ... W_{NI}
Classifier
Stylistic behavior recognition results
(d)
Force Generator
Spring
Damping
Spring + Damping
(e)

Figure 4.1: System Block Diagram: The human user interacts with a haptic device and the simulation environment (a). In advance of the experiment, training movement data is used to learn a dictionary of stylistic feature codes and a classifier is generated to predict stylistic deficiencies in near-real-time (Ershad et al., 2019) (b). During the experiment, each frame of data from haptic device kinematic measurements is represented into stylistic behaviors by projecting it on the learned dictionary (c). The quality of the users style is detected using a classifier which takes the coefficients of the new representation of the data as an input (d). Finally, force feedback is provided to the user if negative performance is detected. Three different types of feedback forces were evaluated in this study for their effectiveness in improving user style, including spring, damping, and spring damping feedback, computed from prior user positions and velocities (e).

test a framework for automatically detecting the deficiency in movement styles in near-real-time, and examine the effect of three different types of force feedback (spring, damping, spring and damping) on six different styles of movement (Table 4.1).

Table 4.1: Lexicon of Stylistic Behavior

Positive Adjective	Negative Adjective
Fluid	Viscous
Smooth	Rough
Crisp	Jittery
Relaxed	Tense
Deliberate	Wavering
Coordinated	Uncoordinated

4.2.2 Data Acquisition

The Geomagic Touch haptic device (3D Systems, Rock Hill, SC) was used in this study. This device allows for 3-degree-of-freedom force feedback and 6-degree-of-freedom sensing. It is used to both provide the user with the desired movement tasks, as well as force feedback guidance cues based on stylistic deficiencies. Position, and linear and angular velocity measurements were recorded from the stylus of the haptic device at a frequency of 256 Hz. To enable near-real-time performance, stylistic detection was performed on every frame of 30 samples of incoming data (representing 0.12 seconds).

4.2.3 Detecting Deficiencies in Stylistic Behavior

The stylistic behavior performance was detected using the method described in Chapter 2. For clarity, we briefly summarize the method here.

Dictionary Training and Classifier Model Training

For each stylistic behavior, a dictionary containing the basis vectors for the good and bad performance is created using the kinematic data from the right hand manipulator of da Vinici skill simulator from the JIGSAWS data set (Gao et al., 2014). This dataset includes position, velocity, and angular velocity from the robot end effectors (which is similar to the type of data obtained form the Geomagic Touch end effector). The positive and negative labels regarding each stylistic behavior adjective used to train the model were obtained from crowd-sourced assessment on the JIGSAWS video data set (discussed in Section 4.2.3). The input data is then represented as a linear combination of the basis vectors in the dictionary. The dictionary and the coefficients are calculated using an optimization algorithm that iterates between two problems of finding the basis vectors such that the reconstructed signal is as similar as possible to the input signal, and finding the coefficients such that they are sparse. These sparse codes are then used to train a SVM classifier. Six separate codebooks are learned for each of the six stylistic behavior adjectives, leading to six trained classifiers.

Crowd-Sourced Assessment for Positive and Negative Performance for Each Style

The JIGSAWS videos were uploaded to Amazon Mechanical Turk and crowd workers rated the videos based on the quality of performance in the six styles of movement mentioned in Table 4.1. The crowd-workers were asked to rate the video based on either a positive or negative adjective for a given stylistic descriptor (e.g., smooth v.s. rough movement, crisp vs. jittery movement). Each video was rated by 20 crowd-workers. The trial was eventually assigned a positive label if it was rated positive by more than or equal to 50 % of the crowd-workers and was otherwise assigned as negative.

Coefficient Calculation

For a new set of input data (i.e. a new frame of 30 samples), dimensionality reduction is done using principle component analysis (PCA) to remove correlations among the data set, then this reduced dimension data set is projected onto the learned codebook. The new representation of the input signal is sparse. The sparse codes from the new data frame at each point of time is then fed into the trained classifier for performance evaluation. Algorithm 1 shows the pseudo code for this method.

4.2.4 Providing Feedback for Correcting Stylistic Behavior

To avoid confusing the operator with multiple, potentially competing feedback cause, the experiment was divided into 6 blocks and only one stylistic deficiency was detected within this set of movement trials. Based on which style detection algorithm was activated for a given block in the experiment protocol, when a poor performance was detected using the proposed near real-time algorithm, one of the three type of force feedback was turned on. The three types of force feedback that were compared in this study (Fig. 4.2) are discussed below.

Algorithm 1 Style Performance Detection Algorithm

Input: new data

Output: stylistic behavior performance S_i

```
while trial not finished do
    Get every data frame of 30 samples ($df$)
    perform PCA on the new data frame ($df_{PCA}$)
    Project the reduced dimension data set onto the dictionary
        $D = [D_p, D_N]$,
        $D_p \in R^l$ and $D_N \in R^l$,
        l: Number of basis vectors in the dictionary,
        $df_{PCA} = W_{p1}D_{p1}, W_{N1}D_{N1} + W_{p2}D_{p2}, W_{N2}D_{N2} +$
       $\ldots + W_{pl}D_{pl}, W_{Nl}D_{Nl}$
    use the sparse codes ($W_{p1}, W_{N1}, \ldots, W_{pl}, W_{Nl}$) as input to the pre-trained classifier
    $S_i$ = classifier output
        $S_i$ = 1 if poor performance is detected,
        $S_i$ = 0 if good performance is detected)
end while
```

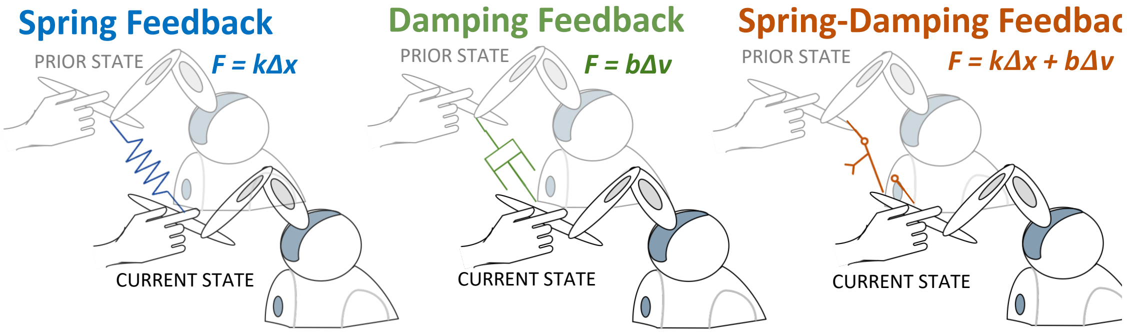


Figure 4.2: Three types of haptic feedback including spring feedback, damping feedback, and spring + damping feedback were studied here for their ability to provide stylistic cues to the human operator. A force feedback was generated based on the users prior position in time.

- Spring Feedback: This was calculated using the difference between the the position of the hand at time t (D_1), and the position at time t-1 (D_2).

$$F_s = K_s(D_1 - D_2) \tag{4.1}$$

The gain K_s was obtained through trial and error and chosen to be 30. The gain was chosen to be high enough so that the user would be able to feel the feedback, but also maintain the stability of the system. This gain was fixed throughout the experiment.

- Damping Feedback: Was calculated using the difference between the the velocity of the hand at time t (V_1), and the velocity at time t-1 (V_2).

  $$F_d = B_d(V_1 - V_2) \tag{4.2}$$

  The gain, B_d was chosen to be 15. A lowpass filter with a cutoff frequency of 100 HZ was used to remove noise and smooth the velocity signal and prevent the system from becoming unstable.

- Spring + Damping Feedback: Was calculated using the difference between the the position and velocity of the hand at time t (D_1, V_1), and the position and velocity at time t-1 (D_2, V_2).

  $$F_{sd} = K_{sd}(D_1 - D_2) + B_{sd}(V_1 - V_2) \tag{4.3}$$

  The gains, K_{sd} and B_{sd} were chosen to be 10 and 5. Similar to the damping feedback, a lowpass filter with a cutoff frequency of 100 HZ was used to remove noise and smooth the velocity signal and prevent the system from becoming unstable.

Algorithm 2 shows the pseudo code for applying the haptic feedback.

4.3 Experimental Setup

Figure 4.3 shows the experimental setup and the simulated needle steering task.

Algorithm 2 Feedback Generation Algorithm

Input: S_i: style performance , 0 otherwise
$S_i = 1$ if good performance is detected,
$S_i = 0$ if bad performance is detected,
Output: f_{out}: force feedback to be applied

```
define feedback type (F)
    $F = F_s$: Spring Feedback,
    $F = F_d$: Damping Feedback,
    $F = F_sd$: Spring+Damping Feedback
while trial not finished do
    if $S_i = 0$ then
        $f_{out} = F$
    end if
end while
```

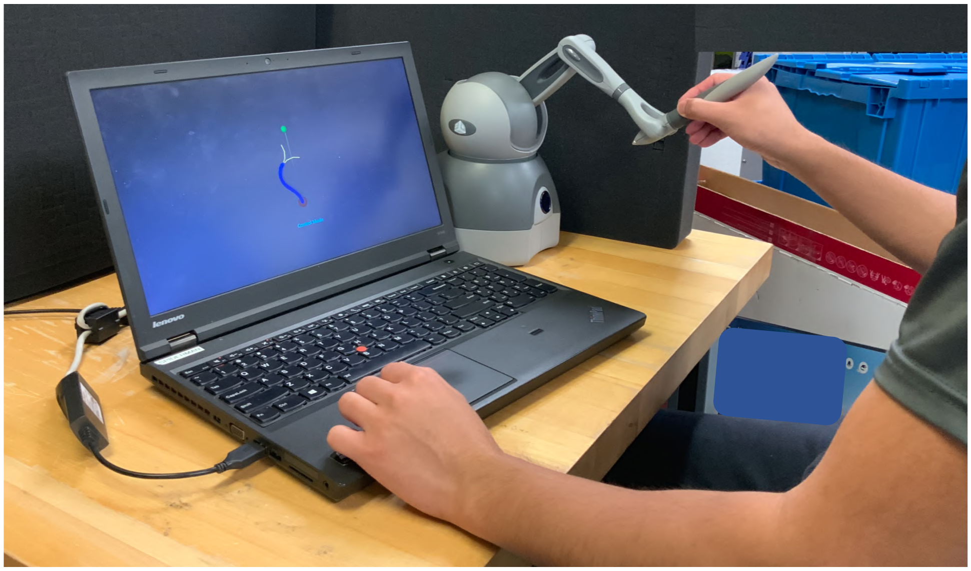


Figure 4.3: Experimental setup showing a subject interacting with the simulated environment using the Geomagic Touch haptic device and the simulated needle steering target reaching task.

4.3.1 Participants

A total of 21 subjects participated in this study. The study protocol was approved by UTD IRB office (UTD # 14-57). Participants had no previously reported muscular-skeletal injuries or diseases, or neurological disorders. The subjects were divided into 3 groups of 7 subjects each. Each group was assigned the same randomized movement task, but only received either spring, damping, or

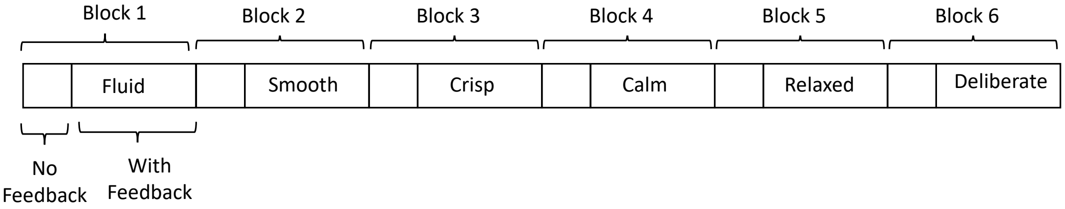


Figure 4.4: An example of an experiment protocol for one subject. The protocol consists of six blocks, each related to one stylistic behavior detection algorithm that was activated for that block. For each block, the user first performed a set of reaching movements with no feedback to enable a baseline computation of style, followed by a set of trials with feedback that was provided, based on measured stylistic deficiencies. For each subject, a single feedback method was provided throughout the whole experiment, but at different points of time, depending on the style detection algorithm for that subject. Hence, a unique feedback relevant to style was provided to each subject.

spring-damping feedback for each of the stylistic adjective blocks. This parallel study design was chosen to allow us to evaluate the effect of the type of haptic feedback on corresponding changes in stylistic behavior.

4.3.2 Simulated Task

The simulated task consisted of reaching a set of targets under a kinematically constrained environment, simulating the control of a steerable needle using Cartesian Space teleoperation (Majewicz and Okamura, 2013). This task was chosen due to its complexity as a single-handed movement and one that naturally hunders movement is a straight-line path, which we felt would not be difficult enough to illicit stylistic changes in the user movements. The movement tasks were developed using C++ and the CHAI 3D haptic rendering library. Users were asked to reach four 5 mm targets, mirrored vertically, at predefined locations which were presented to the user at random. The user was instructed to initialize each trial by moving the virtual stylus to the starting point. After reaching the target the user would end the trail by pressing a button on the stylus. Data was collected from the time the user initialized the haptic device until they defined the end of the trial.

4.3.3 Experimental Protocol

The experiment was divided into six blocks of kinematically constrained movement trials (e.g., controlling a steerable needle under cartesian space teleoperation), with each block corresponding to one of the six stylistic adjectives. Each block includes a baseline segment consisting of two repetitions for each target (a total of 16 reaching trails) with no force feedback applied, and a segment that contains an applied force feedback for five repetitions of movements for each target (a total of 40 reaching trials). In each block, force feedback was provided when a stylistic deficiency was detect for the given adjective corresponding to the block. A 20 sec break was provided to the user between each block. Both target location and ordering of stylistic blocks was randomized for all subjects. An example of the experiment protocol is shown in the Figure 4.4.

4.3.4 Stylistic Behavior Performance Detection

The kinematic data recorded from the haptic device includes user hand position, velocity, and angular velocity all in X, Y, and Z directions, resulting in 9 signal channels. Based on the style detection algorithm that was activated in each block, the new frame of data was projected onto the set of over complete dictionary that was calculated as discussed in 4.2.3. The sparse codes for each incoming frame of data was calculated and used as input to a classifier to detect the performance quality based on the activated detection algorithm for the specific style. The classifier returns 0 if a poor performance is detected, returns 1 otherwise. The detection algorithms were implemented in MATLAB.

4.3.5 Providing Feedback to the User

For each frame of incoming data if a poor performance was detected (output of the classifier was 0), one type of force feedback (spring force feedback, damping force feedback, or spring-damping force feedback), was activated and applied to the user's hand. A custom C++ code was developed to apply the force through the Geomagic Touch device. Robot Operating System (ROS) (Quigley et al., 2009) was used to build the connection between detection algorithm in MATLAB and applying the force to the user through the Geomagic Touch haptic device in C++. Three types of forces, as discussed in section 4.2.4, were studied in this experiment. Each group of subjects was provided with one type of force feedback throughout the whole experiment for all different blocks of style detection.

4.3.6 User Performance Evaluation Metrics

To quantify the quality of performance in each trial in which a feedback was applied, the performance quantity *P* was calculated, as discussed in the following, and assigned to that trial. For each style (i.e., each block in the protocol (Fig 4.4)), the first section of the block where no feedback is applied is used as a baseline for that style. The performance of the user was evaluated using the metric calculated as follows: For each trial, the sum of number of times a one was detected (good performance), was divided by the total number of detection in that trial. This was also done for the baseline trials for each style and averaged over all baseline trials. Then, to compare the performance when a force feedback was applied, to the no feedback condition, the following metric was used:

$$P = (num\_positive\_WF/num\_total\_WF) - mean(num\_positive\_NF/num\_total\_NF) \qquad (4.4)$$

Where:
num_positive_WF: is the number of good performance detected in a trial with feedback,
num_total_WF: is the number of total detections in a trial with feedback,
num_positive_NF: is the number of good performance detected in the baseline trial (no feedback),
num_total_NF: is the number of total detections in the baseline trial (no feedback).

4.3.7 Task Performance Evaluation Metrics

To compare the effect of the three types of feedback on the task performance, three metrics were calculated. The calculated metrics were: (1) time taken to reach the target, (2) needle trajectory straightness (the distance traveled by the needle divided by a straight line to the target), and (3) the needle position error (the distance between the needle and the target at the end of the trial).

4.4 Results and Discussion

We collected a total of 7056 trials (21 subjects, 336 each). Data analysis was carried out for all trials. The results include the evaluation of stylistic behavior improvement, as well as an evaluation of task performance as a function of the different types of haptic force feedback. A NASA Task Load index survey was also conducted to show how users percieved the feedback provided to them in terms of workload.

4.4.1 Effect of Force Feedback on Styles

The effect of each type of force feedback on each style of movement is shown in Figure 4.5. The mean and standard deviation of the quantity associated with good performances (*P*) for the three different types of force feedback (spring, damping, spring-damping) are plotted. This is the average of the metric *P* over all trials related to a style detection in a block. The values above the horizontal line crossing at 0 show the improvement of the movement style when applying feedback with respect to the no feedback condition and the values below this line indicate that receiving feedback did not improve the movement style compared to not receiving any feedback. This plot indicates that the spring force feedback was able to improve the average performance of the styles "crisp", "deliberate", and "relaxed". The damping force feedback improved the "crisp" and "deliberate" styles on average, and the spring+damping force feedback was able to improve the "smooth", "calm", "deliberate" styles on average.

Overall, all styles except for "fluid", showed an average improvement by applying one or more types of force feedback. The "fluid" style however showed the best performance in the absence of the forces studied here. This can be due to the fact that other kinematic metrics, rather than the position and velocity, contribute to the fluidness of movement. In this study only force feedback

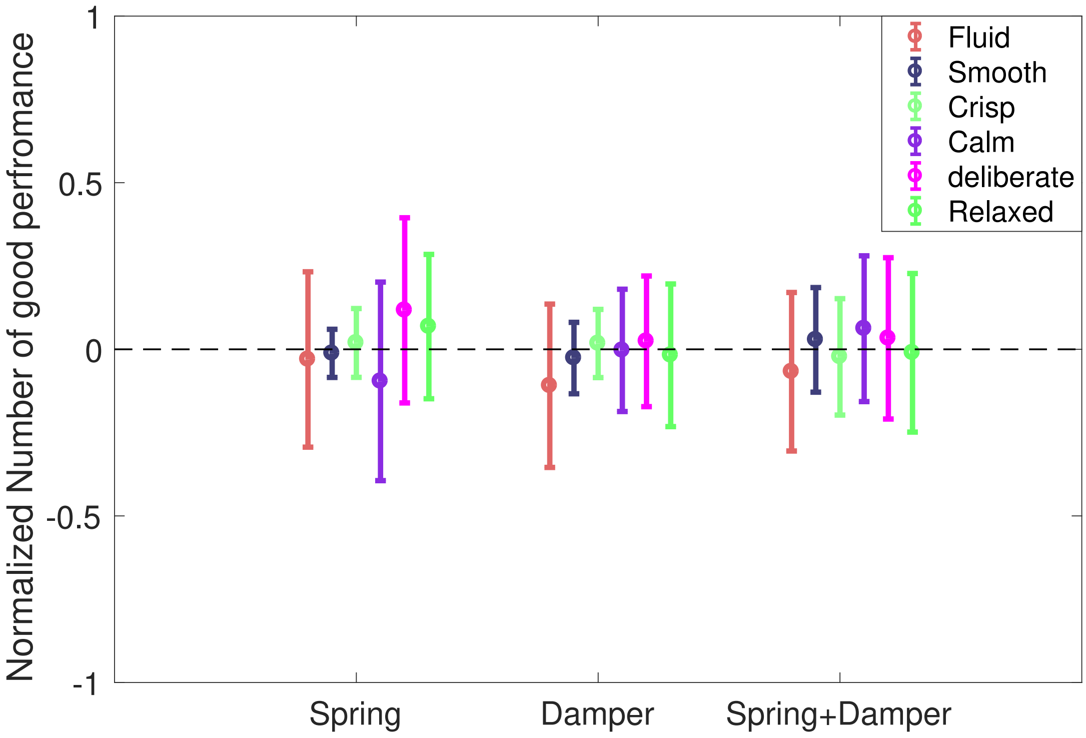


Figure 4.5: Comparing the effects of three different types of haptic feedback on each style. For each group of subjects receiving the same type of feedback, the mean and standard deviation is shown for the number of positive performance normalized to the total number of detections and divided by the baseline stylistic positive performance, for each style. The values above 1 show an improvement in the performance when applying feedback compared to the no feedback condition in the baseline trials.

associated with position and velocity were studied. According to our previous study (Ershad et al., 2018b), the angular velocity of the hand movement is related to the fluidity of the movement. Thus, in future work, applying other types of force feedback which incorporate the effect of angular velocity might help to improve the fluidity of movement. This study was limited by the fact that the haptic device used was not able to provide rotational feedback cues. The style “deliberate” was improved by all types of forces when compared to the no feedback condition; however the most improvement occurred when applying spring force feedback. A post hoc statistical analysis was done to determine significant differences in different types of force feedback, different targets, and task repetitions. For each stylistic behavior adjective (Table 4.2) the Kruskal Wallis test was used to identify significantly different groups. Effect significance is identified for p-values less than 0.05.

4.4.2 Effect of Force Feedback on Task Performance

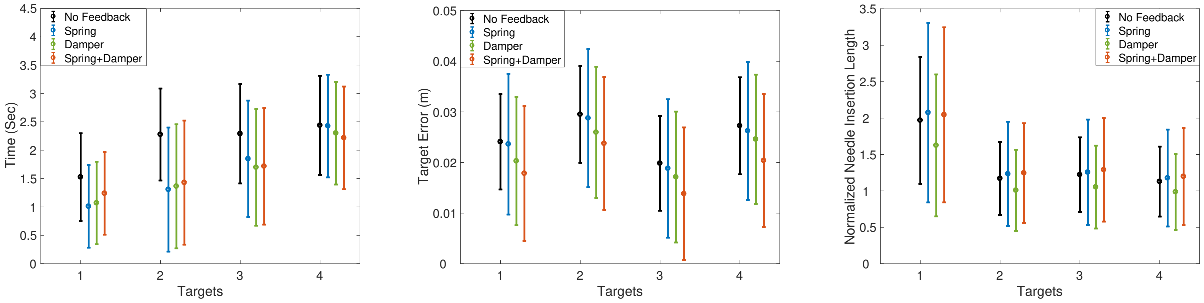
For each trial, task-specific metrics including the time taken to complete the task, needle trajectory straightness, and target error, were calculated to evaluate the effects of different types of feedback, as well as the no feedback condition, on task performance. The mean and standard deviation of each of these metrics were calculated. Figure 4.6 (a) shows that the time taken to reach the target is improved using all three types of force feedback compared to the no feedback condition. For targets 1 and 2, the group with spring force feedback showed the least time taken to reach the target, and for targets 3 and 4, the group with the spring-damping force feedback showed the least time taken to complete the task.

The target positioning error in all four targets was improved using all three groups of receiving force feedback compared to the no feedback condition; however, the group which received the spring+damping force feedback showed the least error (Fig. 4.6 (b)). This indicates that applying force feedback increases the accuracy reaching task regardless of the target.

The straightness of the trajectory traveled by the needle is compared in Fig. 4.6 (c). This figure indicates that for all four targets, the group which received the damping feedback showed a straighter needle path compared to the other two feedback groups and the no feedback condition; however, spring feedback and spring+damping force feedback caused less straightness in the needle trajectory compared to the no feedback condition.

In general, the time and the target error are improved by applying at least one of the forces as feedback improves the task performance compared to absence of any feedback. The straightness of the trajectory however is only improved compared to absence of feedback only when the damping feedback

Target configuration and needle trajectory for all trials for each adjective are shown in Figure 4.7, grouped by the force feedback and color-coded by target error. This figure shows the traces from all trials receiving spring force feedback regardless of the style. The plot visually demonstrates that in general, spring force feedback improves task performance.



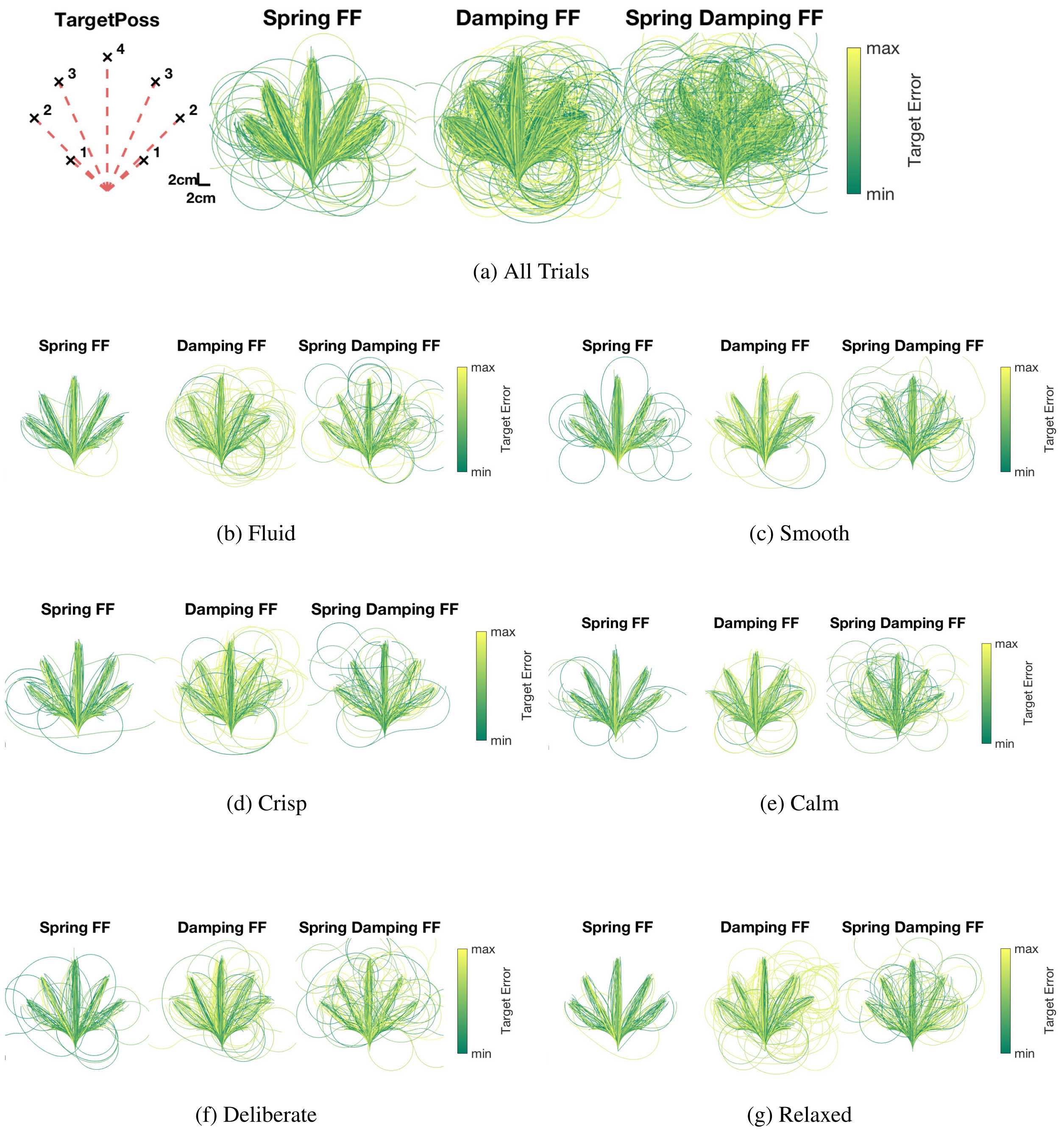
(a) Time to Complete the Task (b) Target Positioning Error (c) Needle Trajectory Straightness

Figure 4.6: Three metrics including (a) time to complete the task, (b) target positioning error, (c) and needle trajectory straightness were used to evaluate the effect of the haptic cues on task performance. For each group of subject receiving the same type of force feedback, the mean and standard deviation of each task performance metric is calculated and compared for 4 target locations.

4.4.3 Statistical Analysis

For each stylistic behavior adjective, as well as each task performance metric, a post hoc statistical analysis was done to determine significant differences in the three types of force feedback, different targets, and task repetitions. The Kruskal Wallis test was used to identify significantly different groups. Effect significance is identified for p-values less than 0.05.

The results from the statistical analysis on different styles of movement (Table 4.2) indicate that for the style *Fluid/Viscous*, the **spring** force feedback showed significant difference in improving the user performance compared to **spring** force feedback but showed no significant difference compared to the **spring+damping** force feedback. For the *Relaxed/Tense, Deliberate/Wavering* styles, the **spring** force feedback showed significant difference in improving the user performance



(a) All Trials

(b) Fluid

(c) Smooth

(d) Crisp

(e) Calm

(f) Deliberate

(g) Relaxed

Figure 4.7: Target layout and resulting needle paths for all subjects under the spring, damping, or spring damping feedback conditions. Needle paths are color-coded by final target error, with green paths indicating smallest error.

compared to the other two types of feedback. For the *Crisp/Jittery* style both the **spring** feedback and **damping** feedback showed significant difference in performance improvement compared to the **spring+damping** feedback. For the *Smooth/Rough* style, the **spring+damping** force feedback showed significant difference in performance improvement compared to the **spring** feedback but no significant difference was found compared to the **damping** feedback. For the *Calm/Anxious* style, the **spring+damping** force feedback showed significant difference in performance improvement compared to the other two feedback types and the **damping** force feedback showed significant difference in improvement compared to the **damping** feedback.

For the task performance metrics, the statistical analysis indicates that for the target positioning error, **spring+damping** feedback shows significant difference in reducing the error compared to **damping** feedback but is not significant compared to **spring** feedback; however, it results in a statistically significant less straighter path traveled by the needle compared to damping feedback. The **spring** feedback results in statistically significant less task completion time compared to **damping** feedback and **spring+damping** feedback (Table 4.3).

The statistical analysis indicate that task repetition shows no statistically significant effect in the different types of stylistic behavior nor the task performance metrics, as opposed to the target location which shows significant importance in both styles and task-specific metrics. This is shown in the third and fourth column of Tables 4.2 and 4.3

Table 4.2: Statistical analysis summary of the effect of different force feedback types, targets, and task repetitions on the stylistic behavior

	Force Feedback		Target		Repetition	
Style	*p*	Significance	*p*	Significance	*p*	Significance
Fluid/Viscous	<0.0035	S>D	0.64259	N/A	0.9916	N/A
Smooth/Rough	0.0568	SD>S	0.0045	1>2	0.3854	N/A
Crisp/Jittery	<0.0035	S, D > SD	0.0045	1>2,3,4	0.6953	N/A
Calm/Anxious	<0.0035	SD>D>S	0.2987	N/A	0.7711	N/A
Deliberate/Wavering	<0.0035	S>D, SD	<0.0035	1>2>3>4	0.4111	N/A
Relaxed/Tense	<0.0035	S>D, SD	<0.0035	1>2,3,4, 2>3	0.6892	N/A

S – Spring, D – Damping, SP – Spring+Damping

Table 4.3: Statistical analysis summary of the effects of force feedback types, targets, task repetition on performance metrics

	Force Feedback		Target		Repetition	
Performance Metric	*p*	Significance	*p*	Significance	*p*	Significance
Target Error	0.0389	D>SD	<0.0035	2>3>1>4	.02033	N/A
Needle Trajectory Straightness	0.0398	SD>D	<0.0035	1 > 2,3,4, 3 > 4	0.914	N/A
Time taken to complete the task	<0.0035	SD, D>S	<0.0035	2,3,4 > 1	0.0498	N/A

S – Spring, D – Damping, SP – Spring+Damping

4.4.4 Subjective User Report

Fig. 4.8 shows the NASA-TLX user report on performing the experiment for the three group of subjects receiving three different types of feedback, based on 6 criteria of mentally, physically, and temporally demanding, subject's' perceived performance, effort and frustration.

The results from user survey (NASA Task Load Index) indicate that subjects who received the spring force feedback found the feedback unpleasant and the tasks more demanding compared to subjects who received the other two types of feedback. This indicates that a haptic feedback can improve task and user performance but still be unpleasant to the user.

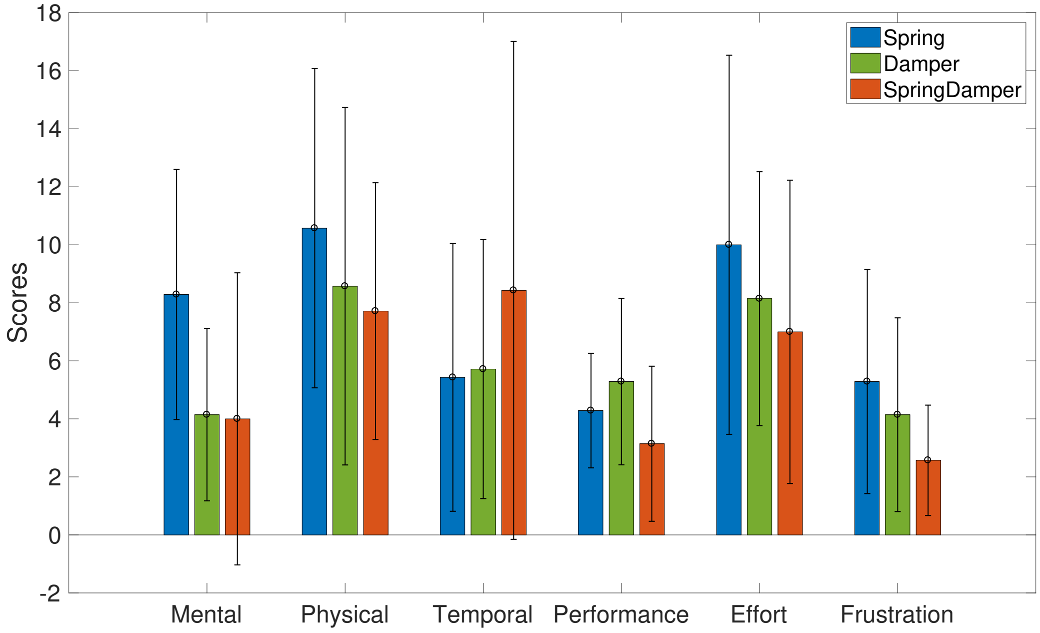


Figure 4.8: NASA Task Load Index

4.5 Summary of Adaptive Surgical Robotic Training Using Real-Time Stylistic Behavior Feedback Through Haptic Cues

In this study we proposed an automatic training framework in a simulated environment which detects a poor performance in the user's style of movement in near-real-time and applies force feedback using a haptic device to help correct the hand movement. We conducted a user study to evaluate three different types of force feedback: (1) spring force feedback, (2) damping force feedback, and (3) spring+damping force feedback for six different behavioral styles. In general, this study demonstrates that haptic feedback used as guidance during a task can help improve task performance, as well as human subject stylistic behavior.

"Spring" force feedback resulted in less time to complete the task, hence faster performance speed. It also helped demonstrate a more fluid, crisp, calm, and deliberate behavior in the user's movement while performing the task.

"Spring+Damping" feedback reduced the target error resulting in a more accurate performance; however, it resulted in a less straight needle path and more time to complete the task. It helped demonstrate a more relaxed performance.

"Damping" force feedback resulted in a straighter line traveled by the needle in the simulated task towards the target; however, it lead to an increased target error and a slower speed resulting in more time taken to complete the task. It helped the user to demonstrate a more crisp performance.

In this study we considered only three type of force feedback related to position and velocity; however, this might not be sufficient for all styles, since other kinematic metrics can also be associated with some styles. In future studies considering other types of force feedback can help improve the performance of the styles that were not improved using only a position or velocity force feedback. Implementing the proposed method in the dvrk through an in vitro study with standard surgical training exercises instead of a simulated environment, can contribute to better evaluating this method and will also be the focus of future studies.

This study provides the groundwork for continued research on user performance based feedback for adaptive training.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This dissertation presents the design, implementation, and experimental validation of an adaptive surgical robotic training system that provides several key contributions related to surgical skill assessment and training: (1) proposing an objective and intuitive surgical skill assessment method based on the style of movement, (2) suggesting quantitative metrics associated with the styles of movement, (3) performing automatic and near-real-time skill assessment based on stylistic behavior, (4) developing a framework for near-real-time, intuitive and customized training. The main objective of this dissertation was to propose a framework for a more intuitive way to provide feedback to users in robotic surgery training. An important feature in training is being able to provide trainees with intuitive and easy to understand feedback that can tell them how to modify movement to improve performance. The current skill evaluation methods, e.g. surgical simulators, provide the trainee with some quantitative metrics at the completion of a task which are sometimes hard to understand. To be able to provide trainees with more intuitive feedback for more effective training, we focused on the style of movement and aimed at improving subjects' stylistic behavior while performing a task by proving real-time feedback.

We introduced a novel method for evaluating technical skills in robotic surgery based on subjects' style of movement. As opposed to many studies in the field of surgical skill assessment that break down a surgical task to building blocks and evaluate the surgeon's performance within those building blocks, we believe that there is a more general characteristics in the movement that differentiates an expert form a novice. This is visible to an observer who has no familiarity with the objectives and technical aspects of a specific task and the human brain can easily distinguish between a good and a poor performance. The quality of performance is expressed through vocabulary using adjectives like "fluid vs. viscous", " smooth vs. rough", etc. which describe the style of movement. Thus, the first aim of this dissertation was to evaluate surgical skill based on different levels of expertise using multiple adjectives describing the quality of movement. This method was tested and evaluated through an experimental study. An experiment was carried out using 4 different groups of expertise including novice, intermediate, fellow, and expert. These groups were identified based on their years of experience in robotic surgery. Paired videos of a subject performing a simulated task on the da Vinci skill simulator, along with the simulated videos were presented to the crowd. The crowd was able to differentiate among the four levels of expertise based on their style of movement. This provides a groundwork for a novel and global method in skill evaluation which is task independent. In the next step, we wanted to be able to quantitatively evaluate a subject's performance based on their style of movement. The subjects in our experiment were wearing multiple inertial measurement unit (IMU) and physiological sensors on their hands and arms to record movement and muscle activities from these limbs while performing the simulated task. We then extracted multiple metrics from the sensors and found the best metrics correlated to each adjective pair through correlation of the metrics with the ratings from the crowd for each adjective describing the styles of movement. The metrics were then evaluated in their ability to distinguish among different expertise levels. This was done using a classifier with the metrics used as input to the classifier. These metrics can provide some insight on how to quantitatively describe certain styles of movement; however, the metrics presented in this study while being effective in distinguishing different expertise levels are by no means exhaustive and other metrics including frequency domain metrics need to be evaluated for their effect in proving meaningful results. This will be the focus of future studies. These labels can be the basis of future automated coaching systems. We hope to extend these methods to other aspects of surgical training, such as: open and laparoscopic skills, team dynamics, patient interactions, and professionalism.

Deciding which metrics to extract and finding a relevant metric that can effectively represent a movement style is very time consuming and requires expert knowledge. In addition, the metrics were calculated upon the completion of a task and could not be used to evaluate performance and provide meaningful feedback to the user during task performance. Real-time evaluation of technical skill is important for providing feedback to the user in a timely manner and assisting with modifying movement by providing meaningful feedback to the user during the task, to improve performance. Another contribution of this dissertation is automating the detection of stylistic behavior in real-time during task performance. To automatically recognize stylistic behavioral performance, a dictionary was learned for each stylistic adjective. The dictionary was derived from the raw data, meaning that as opposed to predefined dictionaries with specific basis vectors, it was learned from the data set. The benefit of this method is that the dictionary can be learned from any type of data which enables capturing the underlying features and characteristics of the specific data set and thus, a more accurate representation of the data. The total dictionary contains two set of basis vectors one, pertaining to the positive performance of a stylistic adjective, and the other one pertaining to the negative performance. These two sets of basis vectors are learned from the trials with positive or negative label assigned through crowd source ratings. Once the dictionary is learned form the training data and the basis vectors are defined, the data is projected onto the dictionary. For a trial with a positive label, the coefficients of the basis vectors belonging to the positive performance contain the most non-zero elements and likewise for a trial with a negative label. A classifier was trained using these coefficients to detect the good or poor performance in each trial. The dictionary and the classifier models are saved and used for new incoming data of the same type. This is done for short frames of data hence, enabling a near real-time assessment.

Once the assessment is done in near-real-time, it is important to provide a relevant feedback to the user as soon as the a poor performance is detected to help improve movement during performance. Real-time and adaptive training is an ongoing topic of research in training simulators. Detecting errors when they appear and giving hints to the user on how to act differently to correct movement can significantly improve performance and out-coming results. Since trainees have different set of skill sets, each individual requires a unique set of training routines focusing on their deficient technical skills, to guarantee efficient training. The other contribution of this dissertation is providing haptic feedback upon the detection of a negative performance in each style of movement. Three different types of haptic feedback including “spring”, “damping”, and “spring +

damping" forces were studied and evaluated for each style to find the best feedback that improves each stylistic behavior. A human subjects study was carried out for this purpose. Each subject was presented with one type of haptic feedback for each style in near-real-time. The experiment protocol allowed the detection of one of the stylistic behaviors over a span of time and thus applying a haptic feedback as soon as a negative performance in that style was detected. The subject's performance was monitored in real-time to see whether it was improved over multiple repetitions of the task when applying the force feedback. The subjects were asked to complete a NASA Task Load index survey at the end of the experiment. The results indicate that what a the human subject might perceive as a pleasant feedback might not always be proper for improving the outcome results.

In conclusion, significant advances were made in this dissertation toward a novel and more intuitive method for assessment and training of technical skills in surgery to help obtain mastery in robotic surgery. This dissertation presents the ground work for an adaptive training method in surgical skill training using real-time haptic feedback based on intuitive stylistic cues. A key difference in the performance of a novice and an expert is in the way that they move their hands during task performance, which is visible to a casual observer who does not know the objectives and technical details of the task being performed but rather instinctively picks up on these differences in hand movement and tool manipulation. This opens up a whole new area for research in skill assessment and training not only in the surgical field but also other fields, such as sports. The idea is that if we can figure out what is it that a human brain can distinguish in the performance of an expert, we can apply the same method to obtain an objective and automatic skill assessment method. In addition, this knowledge can help us achieve a training method with easy to understand feedback which can significantly improve learning technical skills in a task. The effectiveness of skill assessment based on stylistic behavior, proposed in this dissertation, is currently being evaluated for another study in basketball coaching. Furthermore, many interesting problems remain for improvement in adaptive training such as adaptation in task difficulty levels, as well as visual or other types haptic feedback not studied in this dissertation.

REFERENCES

Acosta, E. and B. Temkin (2005). Dynamic generation of surgery specific simulators-a feasibility study. *Studies in health technology and informatics 111*, 1–7.

Aghdasi, N., R. Bly, L. W. White, B. Hannaford, K. Moe, and T. S. Lendvay (2015). Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *Journal of Surgical Research 196*(2), 302–306.

Aharon, M., M. Elad, and A. Bruckstein (2006). *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing 54*(11), 4311–4322.

Ahmidi, N., G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii (2010). Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. *Lecture Notes in Computer Science 6363*(3), 295–302.

Ali, S. and M. Shah (2008). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence 32*(2), 288–303.

Alonso, O., D. E. Rose, and B. Stewart (2008). Crowdsourcing for relevance evaluation. In *Association for Computer Machinery*, Volume 42, pp. 9–15.

Anderson, F., D. W. Birch, P. Boulanger, and W. F. Bischof (2012). Sensor fusion for laparoscopic surgery skill acquisition. *Journal of International Society for Computer Aided Surgery 17*(6), 269–83.

Antoniou, G. A., C. V. Riga, E. K. Mayer, N. J. Cheshire, and C. D. Bicknell (2011). Clinical applications of robotic technology in vascular and endovascular surgery. *Journal of vascular surgery 53*(2), 493–499.

Badash, I., K. Burtt, C. A. Solorzano, and J. N. Carey (2016). Innovations in surgery simulation: a review of past, current and future techniques. *Annals of translational medicine 4*(23).

Balasubramanian, S., A. Melendez-Calderon, A. Roby-Brami, and E. Burdet (2015). On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabilitation 12*(1), 112.

Bark, K., W. McMahan, A. Remington, J. Gewirtz, A. Wedmid, D. I. Lee, and K. J. Kuchenbecker (2013, Feb). In vivo validation of a system for haptic feedback of tool vibrations in robotic surgery. *Surgical Endoscopy 27*(2), 656–664.

Basdogan, C., S. De, J. Kim, M. Muniyandi, H. Kim, and M. A. Srinivasan (2004). Haptics in minimally invasive surgical simulation and training. *IEEE computer graphics and applications 24*(2), 56–64.

Benedek, M. and C. Kaernbach (2010a). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods 190*(1), 80 – 91.

Benedek, M. and C. Kaernbach (2010b). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology 47*(4), 647–658.

Bhattacharya, S., P. Nurmi, N. Hammerla, and T. Plötz (2014). Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing 15*, 242–262.

Birkmeyer, J. D., J. F. Finks, A. O'reilly, M. Oerline, A. M. Carlin, A. R. Nunn, J. Dimick, M. Banerjee, and N. J. Birkmeyer (2013). Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine 369*(15), 1434–1442.

Blohm, I., J. M. Leimeister, and H. Krcmar (2013). Crowdsourcing: how to benefit from (too) many great ideas. *MIS Quarterly Executive 12*(4), 199–211.

Boulanger, P., G. Wu, W. F. Bischof, and X. D. Yang (2006, Nov). Hapto-audio-visual environments for collaborative training of ophthalmic surgery over optical network. In *2006 IEEE International Workshop on Haptic Audio Visual Environments and their Applications (HAVE 2006)*, pp. 21–26.

Brabham, D. C. (2010). The effectiveness of crowdsourcing public participation in a planning context.

Brown, J. D., C. E. OBrien, S. C. Leung, K. R. Dumon, D. I. Lee, and K. J. Kuchenbecker (2017). Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Transactions on Biomedical Engineering 64*(9), 2263–2275.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Volume 1, pp. 286–295.

Cameron, J. L. (1997). William stewart halsted. our surgical heritage. *Annals of surgery 225*(5), 445.

Cao, C. G., M. Zhou, D. B. Jones, and S. D. Schwaitzberg (2007). Can surgeons think and operate with haptics at the same time? *Journal of Gastrointestinal Surgery 11*(11), 1564–1569.

Carter, B. N. (1952). The fruition of halsted's concept of surgical training. *Surgery 32*(3), 518–527.

Charles, D., A. Kerr, M. McNeill, M. McAlister, M. Black, J. Kcklich, A. Moore, and K. Stringer (2005). Player-centred game design: Player modelling and adaptive digital games. In *Proceedings of the digital games research conference*, Volume 285, pp. 00100.

Chen, C., L. White, T. Kowalewski, R. Aggarwal, C. Lintott, B. Comstock, K. Kuksenok, C. Aragon, D. Holst, and T. Lendvay (2014). Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research 187*(1), 65–71.

Chen, S. P., S. Kirsch, D. V. Zlatev, T. Chang, B. Comstock, T. S. Lendvay, and J. C. Liao (2016). Optical biopsy of bladder cancer using crowd-sourced assessment. *Journal of the American Medical Association (JAMA) Surgery 151*(1), 90–93.

Cosman, P. H., P. C. Cregan, C. J. Martin, and J. A. Cartmill (2002). Virtual reality simulators: current status in acquisition and assessment of surgical skills. *ANZ Journal of Surgery 72*(1), 30–34.

Criswell, E. (2010). *Cram's introduction to surface electromyography*. Jones & Bartlett Publishers.

Critchley, H. D., R. Elliott, C. J. Mathias, and R. J. Dolan (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study. *Journal of Neuroscience 20*(8), 3033–3040.

Curry, M., A. Malpani, R. Li, T. Tantillo, A. Jog, R. Blanco, P. K. Ha, J. Califano, R. Kumar, and J. Richmon (2012). Objective assessment in residency-based training for transoral robotic surgery. *The Laryngoscope 122*(10), 2184–2192.

Darzi, A. and S. Mackay (2001). Assessment of surgical competence Assessment of surgical competence. (June 2009).

Datta, V., A. Chang, S. Mackay, and A. Darzi (2002). The relationship between motion analysis and surgical technical assessments. *American Journal of Surgery 184*(1), 70–73.

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

De Luca, C. J., L. D. Gilmore, M. Kuznetsov, and S. H. Roy (2010). Filtering the surface emg signal: Movement artifact and baseline noise contamination. *Journal of Biomechanics 43*(8), 1573–1579.

De Moraes, R. M. and L. dos Santos Machado (2007). Assessment based on naive bayes for training based on virtual reality. *International Conference on Engineering and Computer Education (ICECE)*, 269–273.

Deal, S. B., T. S. Lendvay, M. I. Haque, T. Brand, B. Comstock, J. Warren, and A. Alseidi (2016). Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. *The American Journal of Surgery 211*(2), 398–404.

Demi, B., T. Ortmaier, and U. Seibold (2005). The touch and feel in minimally invasive surgery. In *IEEE international workshop on haptic audio visual environments and their applications*, pp. 6–pp. IEEE.

Derossis, A. M., G. M. Fried, M. Abrahamowicz, H. H. Sigman, J. S. Barkun, and J. L. Meakins (1998). Development of a model for training and evaluation of laparoscopic skills. *The American Journal of Surgery 175*(6), 482–487.

Dhamala, M., G. Rangarajan, and M. Ding (2008). Estimating granger causality from fourier and wavelet transforms of time series data. *Physical Review Letters 100*(1), 018701.

DiPietro, R., C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager (2016). Recognizing surgical activities with recurrent neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 551–558. Springer.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

El Moudden, I., M. Ouzir, B. Benyacoub, and S. ElBernoussi (2016). Mining human activity using dimensionality reduction and pattern recognition. *Contemporary Engineering Sciences (CES) 9*, 21.

Enayati, N., E. De Momi, and G. Ferrigno (2016). Haptics in robot-assisted surgery: Challenges and benefits. *IEEE reviews in biomedical engineering 9*, 49–65.

Ershad, M., Z. Koesters, R. Rege, and A. Majewicz (2016). Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 508–515. Springer.

Ershad, M., R. Rege, and A. M. Fey (2018a). Automatic surgical skill rating using stylistic behavior components. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1829–1832. IEEE.

Ershad, M., R. Rege, and A. M. Fey (2018b). Meaningful assessment of robotic surgical style using the wisdom of crowds. *International journal of computer assisted radiology and surgery 13*(7), 1037–1048.

Ershad, M., R. Rege, and A. M. Fey (2019). Automatic and near real-time stylistic behavior assessment in robotic surgery. *International Journal of Computer Assisted Radiology and Surgery 14*(4), 635–643.

Ershad, M., R. Rege, and A. Majewicz (2018). Surgical skill level assessment using automatic feature extraction methods. In *Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*. International Society for Optics and Photonics.

Fard, M. J., S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein (2018). Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery 14*(1), e1850.

Feygin, D., M. Keehner, and R. Tendick (2002, March). Haptic guidance: experimental evaluation of a haptic training method for a perceptual motor skill. In *Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS 2002*, pp. 40–47.

Flash, T. and N. Hogan (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of Neuroscience 5*(7), 1688–1703.

Francis, N. K., G. B. Hanna, and A. Cuschieri (2002). The performance of master surgeons on the advanced dundee endoscopic psychomotor tester: contrast validity study. *Archives of Surgery 137*(7), 841–844.

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

Gallagher, A. G., E. M. Ritter, H. Champion, G. Higgins, M. P. Fried, G. Moses, C. D. Smith, and R. M. Satava (2005). Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of Surgery 241*(2), 364–372.

Gao, Y., S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, et al. (2014). Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, Volume 3, pp. 3.

Ghani, K. R., D. C. Miller, S. Linsell, A. Brachulis, B. Lane, R. Sarle, D. Dalela, M. Menon, B. Comstock, and T. S. Lendvay (2016). Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. *European Urology 69*(4), 547–550.

Gibo, T. L. and D. A. Abbink (2016). Movement strategy discovery during training via haptic guidance. *IEEE transactions on haptics 9*(2), 243–254.

Goh, A. C., D. W. Goldfarb, J. C. Sander, B. J. Miles, and B. J. Dunkin (2012). Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. *Journal of Urology 187*(1), 247–252.

Grober, E. D., M. Roberts, E. J. Shin, M. Mahdi, and V. Bacal (2010). Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *American Journal of Surgery 199*(1), 81–85.

Gutt, C. N., T. Oniu, A. Mehrabi, A. Kashfi, P. Schemmer, and M. W. Büchler (2004). Robot-assisted abdominal surgery. *British journal of surgery 91*(11), 1390–1397.

Gwilliam, J. C., M. Mahvash, B. Vagvolgyi, A. Vacharat, D. D. Yuh, and A. M. Okamura (2009). Effects of haptic and graphical force feedback on teleoperated palpation. In *2009 IEEE International Conference on Robotics and Automation*, pp. 677–682. IEEE.

Halabi, O., F. Al-Mesaifri, M. Al-Ansari, R. Al-Shaabi, H. Barki, and S. Foufou (2013). Incorporating haptic and olfactory into surgical simulation. In *2013 International Conference on Cyberworlds*, pp. 52–55. IEEE.

Halaki, M. and K. a. Ginn (2012). Normalization of EMG signals: to normalize or not to normalize and what to normalize to? *Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges*, 175–194.

Hayter, M. A., Z. Friedman, M. D. Bould, J. G. Hanlon, R. Katznelson, B. Borges, and V. N. Naik (2009). Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. *Canadian Journal of Anesthesia 56*(6), 419–426.

Hoffman, R. L., J. Petrosky, M. Eskander, L. Selby, and A. Kulaylat (2015). Feedback fundamentals in surgical education: Tips for success. *Bull Am Coll Surg 100*(8), 35–39.

Holst, D., T. M. Kowalewski, L. W. White, T. C. Brand, J. D. Harper, M. D. Sorenson, S. Kirsch, and T. S. Lendvay (2015). Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. *Journal of Endourology 29*(5), 604–609.

Howells, N. R., M. D. Brinsden, R. S. Gill, A. J. Carr, and J. L. Rees (2008). Motion analysis: a validated method for showing skill levels in arthroscopy. *Arthroscopy: The Journal of Arthroscopic & Related Surgery 24*(3), 335–342.

Hoyer, P. O. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565. IEEE.

Huaulmé, A., S. Voros, L. Riffaud, G. Forestier, A. Moreau-Gaudry, and P. Jannin (2017). Distinguishing surgical behavior by sequential pattern discovery. *Journal of Biomedical Informatics 67*, 34–41.

Hung, A. J., J. Chen, Z. Che, T. Nilanon, A. Jarc, M. Titus, P. J. Oh, I. S. Gill, and Y. Liu (2018). Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *Journal of endourology 32*(5), 438–444.

Jantscher, W. H., S. Pandey, P. Agarwal, S. H. Richardson, B. R. Lin, M. D. Byrne, and M. K. O'Malley (2018, March). Toward improved surgical training: Delivering smoothness feedback using haptic cues. In *2018 IEEE Haptics Symposium (HAPTICS)*, pp. 241–246.

Jarque-Bou, N. J., A. Scano, M. Atzori, and H. Müller (2019). Kinematic synergies of hand grasps: a comprehensive study on a large publicly available dataset. *Journal of neuroengineering and rehabilitation 16*(1), 63.

Jeppsson, J. (2017). Modeling natural human hand motion for grasp animation.

Jin, A., S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei (2018). Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. *arXiv preprint arXiv:1802.08774*.

Johnson, L. and D. Ballard (2014). Classifying movements using efficient kinematic codes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 36.

Joice, P., G. Hanna, and A. Cuschieri (1998). Errors enacted during endoscopic surgerya human reliability analysis. *Applied ergonomics 29*(6), 409–414.

Karg, M., R. Jenke, W. Seiberl, K. Kühnlenz, A. Schwirtz, and M. Buss (2009). A comparison of pca, kpca and lda for feature extraction to recognize affect in gait kinematics. In *3rd IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, pp. 1–6. IEEE.

Kazanzides, P., Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio (2014). An open-source research kit for the da vinci® surgical system. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 6434–6439. IEEE.

Kelley, C. R. (1969). What is adaptive training? *Human Factors 11*(6), 547–556.

Kim, H. K., D. W. Rattner, and M. A. Srinivasan (2003). The role of simulation fidelity in laparoscopic surgical training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 1–8. Springer.

King, C.-H., M. O. Culjat, M. L. Franco, C. E. Lewis, E. P. Dutson, W. S. Grundfest, and J. W. Bisley (2009). Tactile feedback induces reduced grasping force in robot-assisted surgery. *IEEE transactions on haptics 2*(2), 103–110.

Kirkwood, R. N., R. A. Resende, C. Magalhães, H. A. Gomes, S. A. Mingoti, and R. F. Sampaio (2011). Application of principal component analysis on gait kinematics in elderly women with knee osteoarthritis. *Brazilian Journal of Physical Therapy 15*(1), 52–58.

Ko, J., S.-w. Jang, and Y. S. Kim (2017). Development of epiduroscopy training simulator using haptic master device. In *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 542–543. IEEE.

Kowalewski, T. M., B. Comstock, R. Sweet, C. Schaffhausen, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, and J. Kaouk (2016a). Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. *The Journal of Urology 195*(6), 1859–1865.

Kowalewski, T. M., B. Comstock, R. Sweet, C. Schaffhausen, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, and J. Kaouk (2016b). Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. *The Journal of Urology 195*(6), 1859–1865.

Lamata, P., E. Gómez, F. Sánchez-Margallo, F. Lamata, F. Del Pozo, and J. Usón (2006). Tissue consistency perception in laparoscopy to define the level of fidelity in virtual reality simulation. *Surgical Endoscopy and Other Interventional Techniques 20*(9), 1368–1375.

Latash, M. L., J. P. Scholz, and G. Schöner (2002). Motor control strategies revealed in the structure of motor variability. *Exercise and sport sciences reviews 30*(1), 26–31.

Law, H., K. Ghani, and J. Deng (2017). Surgeon technical skill assessment using computer vision based analysis. In *Machine Learning for Healthcare Conference*, pp. 88–99.

Lea, C., G. D. Hager, and R. Vidal (2015). An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 1123–1129. IEEE.

Lee, J. and S. Choi (2010). Effects of haptic guidance and disturbance on motor learning: Potential advantage of haptic disturbance. In *2010 IEEE Haptics Symposium*, pp. 335–342. IEEE.

Liang, K., Y. Xing, J. Li, S. Wang, A. Li, and J. Li (2018). Motion control skill assessment based on kinematic analysis of robotic end-effector movements. *The International Journal of Medical Robotics and Computer Assisted Surgery 14*(1), e1845.

Lin, H. C., I. Shafran, T. E. Murphy, A. M. Okamura, D. D. Yuh, and G. D. Hager (2005). Automatic detection and segmentation of robot-assisted surgical motions. pp. 802–810.

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

Lin, Y., X. Wang, F. Wu, X. Chen, C. Wang, and G. Shen (2014). Development and validation of a surgical training simulator with haptic feedback for learning bone-sawing skill. *Journal of biomedical informatics 48*, 122–129.

Luo, J., P. Kania, P. P. Banerjee, S. Sikder, C. J. Luciano, and W. G. Myers (2016). A part-task haptic simulator for ophthalmic surgical training. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 259–260. IEEE.

Maier-Hein, L., S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. M. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin (2017). Surgical data science for next-generation interventions. *Nature Biomedical Engineering 1*, 691–696.

Majewicz, A. and A. M. Okamura (2013). Cartesian and joint space teleoperation for nonholonomic steerable needles. In *2013 World Haptics Conference (WHC)*, pp. 395–400. IEEE.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11*(7), 674–693.

Malpani, A., S. S. Vedula, C. C. G. Chen, and G. D. Hager (2014). Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8498 LNCS*, 138–147.

Malpani, A., S. S. Vedula, C. C. G. Chen, and G. D. Hager (2015). A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International Journal of Computer Assisted Radiology and Surgery (IJCARS) 10*(9), 1435–1447.

Mariani, A., E. Pellegrini, N. Enayati, P. Kazanzides, M. Vidotto, and E. De Momi (2018). Design and evaluation of a performance-based adaptive curriculum for robotic surgical training: a pilot study. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2162–2165. IEEE.

Martin, J. A., G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown (1997). Objective tructured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery 84*(2), 273–278.

Mazilu, S., A. Calatroni, E. Gazit, D. Roggen, J. M. Hausdorff, and G. Tröster (2013). Feature learning for detection and prediction of freezing of gait in parkinsons disease. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 144–158. Springer.

McMahan, W., J. Gewirtz, D. Standish, P. Martin, J. A. Kunkel, M. Lilavois, A. Wedmid, D. I. Lee, and K. J. Kuchenbecker (2011). Tool contact acceleration feedback for telerobotic surgery. *IEEE Transactions on Haptics 4*(3), 210–220.

Milovanović, I. and D. B. Popović (2012). Principal component analysis of gait kinematics data in acute and chronic stroke patients. *Computational and Mathematical Methods in Medicine,*, 8.

Mollazadeh, M., V. Aggarwal, N. V. Thakor, and M. H. Schieber (2014). Principal components of hand kinematics and neurophysiological signals in motor cortex during reach to grasp movements. *Journal of neurophysiology 112*(8), 1857–1870.

Moorthy, K. and Y. Munz (2003). Objective assessment of technical skills in surgery. *British Medical Journal (BMJ) 327*(7422), 1032–1037.

Morino, M., L. Pellegrino, C. Giaccone, C. Garrone, and F. Rebecchi (2006). Randomized clinical trial of robot-assisted versus laparoscopic nissen fundoplication. *British Journal of Surgery: Incorporating European Journal of Surgery and Swiss Surgery 93*(5), 553–558.

Morris, B. (2005). Robotic surgery: applications, limitations, and impact on surgical education. *Medscape General Medicine 7*(3), 72.

Nisky, I., M. H. Hsieh, and A. M. Okamura (2013). A framework for analysis of surgeon arm posture variability in robot-assisted surgery. *IEEE International Conference on Robotics and Automation*, 245–251.

Nisky, I., M. H. Hsieh, and A. M. Okamura (2014). Uncontrolled manifold analysis of arm joint angle variability during robotic teleoperation and freehand movement of surgeons and novices. *IEEE Transactions on Biomedical Engineering 61*(12), 2869–2881.

Norman, S. L., A. J. Doxon, B. T. Gleeson, and W. R. Provancher (2014, April). Planar hand motion guidance using fingertip skin-stretch feedback. *IEEE Transactions on Haptics 7*(2), 121–130.

Okamura, A. M. (2004). Methods for haptic feedback in teleoperated robot-assisted surgery. *Industrial Robot: An International Journal 31*(6), 499–508.

Ortmaier, T., B. Deml, B. Kübler, G. Passig, D. Reintsema, and U. Seibold (2007). Robot assisted force feedback surgery. In *Advances in Telerobotics*, pp. 361–379. Springer.

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

Pezzementi, Z., D. Ursu, S. Misra, and A. M. Okamura (2008). Modeling realistic tool-tissue interactions with haptic feedback: A learning-based method. In *2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 209–215. IEEE.

Porte, M. C., G. Xeroulis, R. K. Reznick, and A. Dubrowski (2007). Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. *The American journal of surgery 193*(1), 105–110.

Postacchini, R., M. Paoloni, S. Carbone, M. Fini, V. Santilli, F. Postacchini, and M. Mangone (2015). Kinematic analysis of reaching movements of the upper limb after total or reverse shoulder arthroplasty. *Journal of Biomechanics 48*(12), 3192 – 3198.

Quigley, M., K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng (2009). Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

Reiley, C. E. and G. D. Hager (2009a). Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In *M2CAI workshop, Medical Image Computing and Computer-Assisted Intervention (MICCAI), London*.

Reiley, C. E. and G. D. Hager (2009b). Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 435–442. Springer.

Reiley, C. E., H. C. Lin, B. Varadarajan, B. Vagvolgyi, S. Khudanpur, D. D. Yuh, and G. D. Hager (2008). Automatic recognition of surgical motions using statistical modeling for capturing variability. *Studies in health technology and informatics 132*(1), 396–401.

Reiley, C. E., H. C. Lin, D. D. Yuh, and G. D. Hager (2011). Review of methods for objective surgical skill evaluation. *Surgical endoscopy 25*(2), 356–366.

Schijven, M. P., J. Jakimowicz, and C. Schot (2002). The advanced dundee endoscopic psychomotor tester (adept) objectifying subjective psychomotor test performance. *Surgical Endoscopy And Other Interventional Techniques 16*(6), 943–948.

Scholz, J. P. and G. Schöner (1999). The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental brain research 126*(3), 289–306.

Sewell, C., D. Morris, N. H. Blevins, S. Dutta, S. Agrawal, F. Barbagli, and K. Salisbury (2008). Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery 13*(2), 63–81.

Shenoi, B. A. (2005). *Introduction to digital signal processing and filter design*. John Wiley & Sons.

Smith, S., J. Torkington, T. Brown, N. Taffinder, and A. Darzi (2002). Motion analysis. *Surgical Endoscopy 16*(4), 640–645.

Solis, J., C. A. Avizzano, and M. Bergamasco (2003, Nov). Validating a skill transfer system based on reactive robots technology. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pp. 175–180.

Stanley, A. A. and K. J. Kuchenbecker (2012, Third). Evaluation of tactile feedback methods for wrist rotation guidance. *IEEE Transactions on Haptics 5*(3), 240–251.

Stergiou, N. (2008). Objective evaluation of expert performance during human robotic surgical procedures. *Annals of Surgery*, 307–312.

Ström, P., L. Hedman, L. Särnå, A. Kjellin, T. Wredmark, and L. Felländer-Tsai (2006). Early exposure to haptic feedback enhances performance in surgical simulator training: a prospective randomized crossover study in surgical residents. *Surgical endoscopy and other interventional techniques 20*(9), 1383–1388.

Taffinder, N., S. Smith, J. Mair, R. Russell, and A. Darzi (1999). Can a computer measure surgical precision? reliability, validity and feasibility of the icsad. *Surgery Endoscopy 13*(suppl 1), 81.

Tang, B., G. Hanna, and A. Cuschieri (2005). Analysis of errors enacted by surgical trainees during skills training courses. *Surgery 138*(1), 14–20.

Tavakoli, M., R. Patel, and M. Moallem (2005). Robotic suturing forces in the presence of haptic feedback and sensory substitution. In *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.*, pp. 1–6. IEEE.

Tholey, G., J. P. Desai, and A. E. Castellanos (2005). Force feedback plays a significant role in minimally invasive surgery: results and analysis. *Annals of surgery 241 1*, 102–9.

Uemura, M., P. Jannin, M. Yamashita, M. Tomikawa, T. Akahoshi, S. Obata, R. Souzaki, S. Ieiri, and M. Hashizume (2016). Procedural surgical skill assessment in laparoscopic training environments. *International Journal of Computer Assisted Radiology and Surgery 11*(4), 543–552.

Van der Meijden, O. A. and M. P. Schijven (2009). The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: a current review. *Surgical endoscopy 23*(6), 1180–1190.

Varadarajan, B., C. Reiley, H. Lin, S. Khudanpur, and G. Hager (2009). Data-derived models for segmentation with application to surgical assessment and training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 426–434. Springer.

Vassiliou, M. C., B. J. Dunkin, J. M. Marks, and G. M. Fried (2010). Fls and fes: Comprehensive models of training and assessment. *Surgical Clinics of North America 90*(3), 535–558.

Vassiliou, M. C., L. S. Feldman, C. G. Andrew, S. Bergman, K. Leffondré, D. Stanbridge, and G. M. Fried (2005). A global assessment tool for evaluation of intraoperative laparoscopic skills. *American Journal of Surgery 190*(1), 107–113.

Vaughan, N., B. Gabrys, and V. N. Dubey (2016). An overview of self-adaptive technologies within virtual reality training. *Computer Science Review 22*, 65–87.

Wagner, C. R., N. Stylopoulos, and R. D. Howe (2002). The role of force feedback in surgery: analysis of blunt dissection. In *Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS 2002*, pp. 68–74. Citeseer.

Wang, Z. and A. M. Fey (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*.

White, L. W., T. M. Kowalewski, R. L. Dockter, B. Comstock, B. Hannaford, and T. S. Lendvay (2015). Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. *Journal of Endourology 29*(11), 1295–1301.

Wottawa, C. R., B. Genovese, B. N. Nowroozi, S. D. Hart, J. W. Bisley, W. S. Grundfest, and E. P. Dutson (2016). Evaluating tactile feedback in robotic surgery for potential clinical application using an animal model. *Surgical endoscopy 30*(8), 3198–3209.

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(2), 210–227.

Yang, X., W. F. Bischof, and P. Boulanger (2008, March). Validating the performance of haptic motor skill training. In *2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 129–135.

Yong, C. Y., R. Sudirman, N. H. Mahmood, and K. M. Chew (2013). Human hand movement analysis using principle component analysis classifier. In *Applied Mechanics and Materials*, Volume 284, pp. 3126–3130. Trans Tech Publ.

Zhang, H. (2004). The optimality of naive bayes. *Association for the Advancement of Artificial Intelligence 1*(2).

BIOGRAPHICAL SKETCH

Marzieh Ershad received a Bachelor of Science in Electrical Engineering from Shiraz University in 2008, and a Master of Science in Biomedical Engineering from Tehran University of Medical Sciences in 2013. She then joined The University of Texas at Dallas in 2015 to pursue a PhD in Electrical Engineering. Based on her academic record, she was awarded the Mary and Richard Templeton Graduate Fellowship for the 2015-2016 academic year. Her research interests include technological advancements to facilitate surgical procedures for surgeons as well as patients. Her doctoral research focused on the the development of an automatic and adaptive training framework for robotic surgery. Her master's thesis focused on methods to increase the precision of image guided surgical (IGS) procedures in clinical settings.

CURRICULUM VITAE

Marzieh Ershad

Nov 1, 2019

Contact Information:

Department of Electrical Engineering
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.

Voice: (469) 602-2294
Email: `Marzieh.ErshadLangroodi@utdallas.edu`

Educational History:

B.Sc., Electrical Engineering, Shiraz University, 2008
M.Sc., Biomedical Engineering, Tehran University of Medical Sciences, 2013
Thesis: *Pre-operative CT to Intra-operative Patient Registration for Image Guided Spine Surgery*
Ph.D., Electrical Engineering, The University of Texas at Dallas, 2019
Dissertation: *Automated and Adaptive Surgical Robotic Training Using Real-Time Style Feedback Through Haptic Cues*

Experiences:

Research Assistant, The University of Texas at Dallas, Jan 2015 – Jan 2019
Teaching Assistant, The University of Texas at Dallas, Jan 2016 – Dec 2016

Publications:

Journal Articles

- Ershad M., Rege R., Majewicz A., DAdaptive Haptic Movement Training Using Real-Time Surgical Stylistic Behavior Detection, Under review in IEEE Transaction on Haptics

- Ershad, M., Rege, R. and Fey, A.M., Automatic and near real-time stylistic behavior assessment in robotic surgery. International journal of computer assisted radiology and surgery. 2019, 14(4), pp.635-643.

- Ershad M, Rege R, Fey AM. Meaningful assessment of robotic surgical style using the wisdom of crowds. International journal of computer assisted radiology and surgery. 2018, 13(7), pp.1037-1048.

- Ershad, M., Ahmadian, A., Dadashi Serej, N., Saberi, H., and Amini Khoiy, K.: Minimization of target registration error for vertebra in image-guided spine surgery, International journal of computer assisted radiology and surgery., 2014, 9, (1), pp. 29-38

- Ershad, M., Ahmadian, A., and Saberi, H.: Preoperative CT and Intra-Operative Physical Space Registration of the Spine Using an Articulated Model: a Phantom Study, Frontiers in Biomedical Technologies, 2014, 1, (3), pp. 193-199
- Ershad M., Ahmadian A., Dadashi N., Saberi H., Intra-operative 3D navigation and TRE reduction in image guided spine surgery Iranian Journal of Biomedical Engineering 2013.

Conference Proceedings and Presentations

- Ershad M., Rege R., Majewicz A., Surgical Skill Assessment Using Convolutional Neural Networks on Time Series Kinematic data, BMES, GA, Atlanta, Oct 2018
- Ershad M., Rege R., Majewicz A., Automatic Surgical Skill Rating Using Stylistic Behavior Components, EMBC, Honolulu, HI, July 2018
- Ershad M., Rege R., Majewicz A., Automatic and Real-time Stylistic Behavior Assessment in Robotic Surgery, CARS, Berlin, Germany, June 2018
- Ershad M., Rege R., Majewicz A., Surgical Skill Level Assessment Using Automatic Feature Extraction Methods, SPIE, Houston, USA, presented in Feb 2018
- Ershad M., Koesters Z., Majewicz A., Rege R., Meaningful Assessment of Surgical Expertise: Semantic Labeling with Data and Crowds International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2016
- Ershad M., Zahid I., Majewicz A., Meaningfulness in Surgical Skill Assessment: Mapping Semantic Descriptors to Quality of Movement. Late breaking results accepted in IEEE International Conference on Robotics and Automation, 2015
- Ershad M., Ahmadian A., Dadashi N., Saberi H., Amini K., Effect of Landmark Configuration on Target Registration Error for Vertebra: a phantom study, Accepted in SPIE, Florida, USA, Feb. 2013
- Ershad M., Ahmadian A., Saberi H., Anatomical landmark selection for optimum Target registration error in image guided pedicle screw insertion National Spinal congress, Tehran,Iran, 2013
- Ershad M., Ahmadian A., Dadashi N., Saberi H., Automatic landmark detection in spine surfce CT images for registration of pre to intra-operative data International Conference on Electronic Health, Tehran, Iran 2012