12-8-2017 4:30 PM

Skills Assessment in Arthroscopic Surgery by Processing Kinematic, Force, and Bio-signal Data

Behnaz Poursartip
The University of Western Ontario

Supervisor
Trejos, Ana Luisa
The University of Western Ontario Joint Supervisor
Naish, Michael D.
The University of Western Ontario Joint Supervisor
Patel, Rajni V.
The University of Western Ontario

Abstract

Arthroscopic surgery is a type of Minimally Invasive Surgery (MIS) performed in human joints, which can be used for diagnostic or treatment purposes. The nature of this type of surgery makes it such that surgeons require extensive training to become experts at performing surgical tasks in tight environments and with reduced force feedback. MIS increases the possibility of erroneous actions, which could result in injury to the patient. Many of these injuries can be prevented by implementing appropriate training and skills assessment methods.

Various performance methods, including Global Rating Scales and technical measures, have been proposed in the literature. However, there is still a need to further improve the accuracy of surgical skills assessment and improve its ability to distinguish fine variations in surgical proficiency.

The main goal of this thesis is to enhance surgical, and specifically, arthroscopic skills assessment. The optimal assessment method should be objective, distinguish between subjects with different levels of expertise, and be computationally efficient. This thesis proposes a new method of investigating surgical skills by introducing energy expenditure metrics. To this end, two main approaches are pursued: 1) evaluating the kinematics of instrument motion, and 2) exploring the muscle activity of trainees.

Mechanical energy expenditure and work are investigated for a variety of laparoscopic and arthroscopic tasks. The results obtained in this thesis demonstrate that expert surgeons expend less energy than novice trainees. The different forms of mechanical energy expenditure were combined through optimization methods and machine learning algorithms. An effective two-step optimization method for classifying trainees into detailed levels of expertise is proposed that demonstrates an enhanced ability to determine the level of expertise of trainees compared to other published methods. Furthermore, performance metrics are proposed based on electromyography signals of the forearm muscles, which are recorded using a wearable device. These results also demonstrate that the metrics defined based on muscle activity can be used for arthroscopic skills assessment. The energy-based metrics and the muscle activity metrics

demonstrated the ability to identify levels of expertise, with accuracy levels as high as 95% and 100%, respectively.

The primary contribution of this thesis is the development of novel metrics and assessment methods based on energy expenditure and muscle activity. The methods presented advance our knowledge of the characteristics of dexterous performance and add another perspective to quantifying surgical proficiency.

Keywords: Performance metrics, motor skills assessment, energy expenditure, muscle activity, arthroscopy, minimally invasive surgery, surgical training, motion and force measurement, electromyography

Statement of Co-Authorship

The thesis presented here has been written by Behnaz Poursartip under supervision of Drs. Trejos, Naish and Patel. In this project, Drs. Trejos, Naish, and Patel have supervised the research on every aspect including idea development, experiment design, data processing, interpreting the results, and reviewing the manuscripts. This project was also supervised by Dr. LeBel from the medical point of view. Dr. LeBel has contributed in experiment design, interpreting the results, and revising the manuscripts. Two chapters of this thesis have been published in peer-reviewed journals, as described in the following paragraphs.

The study presented in Chapter 3 has been published as:

- B. Poursartip, M. E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Analysis of energy-based metrics for laparoscopic skills assessment," *IEEE Transactions on Biomedical Engineering* (2017). Available online: http://ieeexplore.ieee.org/document/7932145 (accessed on 19 May 2017).

    - B. Poursartip developed the idea, analyzed the data, and wrote the manuscript.

    - The data used was collected as part of Dr. Trejos' doctoral work under the supervision of Dr. Patel.

The study presented in Chapter 4 has been published as:

- B. Poursartip, M. E. LeBel, L.C. McCracken, A. Escoto, R. V. Patel, M. D. Naish, and A. L. Trejos, "Energy-based metrics for arthroscopic skills assessment," *Sensors*, vol. 17, no. 8 (2017): 1808.

    - B. Poursartip developed the idea, performed the data processing and manuscript preparation. She was also involved in experiment design and running the experiment.

    - L. C. McCracken prepared the experimental setup and was involved in running the experiment.

    - A. Escoto cooperated in setup preparation.

Dedicated to my beloved family,

Parvin, Yadollah, Behrang, and Babak,

who have always encouraged me to learn and grow.

Acknowledgments

I would like to offer my sincere thanks to Drs. Ana Luisa Trejos, Michael Naish, and Rajni Patel for their constant support, patient guidance, and motivation. This project would not have been possible without their help and supervision. I would like to thank Dr. Marie-Eve LeBel for her dedication to this project and invaluable advice and feedback on the medical aspects of this work.

Special gratitude is also extended to Drs. Abbas Samani, Ken McIsaac, George Knopf, and Antony Hodgson for serving as my examination committee and for their helpful suggestions.

Many thanks are also due to Abelardo Escoto, Laura McCracken, and Christopher Ward for their effort in preparing the experimental setups and data recording.

I would like to thank my colleagues and officemates who gave me the joy of working in such a friendly and creative environment: Mahta Khoshnam, Mahya Shahbazi, Farokh Atashzar, Karen Siroen, Mahsa Khosravi, Vahid Mehrabi, Saeed Bakhshmand, and Pouya Soltanizarrin. Many thanks to Dr. Gerard Guiraudon for all the kindness, life lessons, and sharing his experiences with me. I am also grateful to my friends I met in London for their support and understanding: Sahar & Alireza, Rayehe, Neda & Hosein, Nahal & Ghazal, Mehrnoosh S., Nooshin & Ehsan, Mehrnoosh & Alireza, Shima & Hadi, and Elmira.

I greatly appreciate those who supported me in any respect during the completion of this dissertation. I cannot find words to express my heartfelt gratitude to my immediate family, who supported and encouraged me in the study. They have been an important and indispensable source of love and spiritual strength.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Arthroscopic surgery is a type of Minimally Invasive Surgery (MIS) that can be used for diagnostic or treatment purposes on human joints. Similar to other MIS procedures, this type of surgery reduces post-operative pain, the risk of infection, and bleeding compared to open surgery [1]. Consequently, patients can recover, start rehabilitation, and go back to normal activities faster. Recently, arthroscopy has become the most performed orthopaedic procedure [2].

Apart from theoretical knowledge, motor skills substantially affect the outcome of MIS [3]. In particular, arthroscopy requires the learning of fine motor skills and enhanced visual-spatial abilities [4, 5]. It has been reported in [6, 7] that orthopaedic residents do not feel prepared and confident when they start participating in the Operating Room (OR), even when their participation is supervised. In addition, the limited practice time available in the OR, the high cost, and the possibility of tissue injury necessitate appropriate training and skills assessment outside of the OR. It was reported in [8] that a high percentage of the mortality rate in hospitals is related to medical errors; and from that large percentage, 50% is preventable during surgery. An appropriate assessment method can assist in providing high quality health care to patients by ensuring that surgeons are truly proficient with their instruments.

1.1   Motivation

One of the problems in the area of surgical training is that there is no clear definition of proficiency. Previously, there was a significant reliance on the outcome of the task, or the time required to complete the task. However, it is now argued by many researchers that these factors are not appropriate for a thorough analysis of surgical proficiency [9, 10]. Trainees can demonstrate a similar task completion time as experts and minimize the number of errors in their performance. However, their limited capability in transferring these skills to the operating room is an indication of their lack of proficiency. In addition, by exploring other performance metrics, significant differences can be found between two subjects with the same task completion time. A thorough assessment method is required to prevent trainees from being misguided and instead give them a realistic report on their performance [9]. Surgical assessment methods are required that can detect small variations in the performance of subjects with various levels of expertise.

For arthroscopic training in particular, there is a lack of standardized assessment methods for arthroscopic skills. Most of the studies in this area have focused on developing simulators without as much consideration about how the simulators can be used to objectively assess performance.

The lack of an appropriate training system in the area of arthroscopy motivated researchers at Canadian Surgical Technologies and Advanced Robotics (CSTAR) to develop a physical knee simulator [11, 12] and a physical shoulder simulator [13]. In addition to the development of simulators, objective assessment methods are required to realize a complete training platform. In 2015, a study was published and discussed by the Association for Surgical Education (ASE) Simulation Committee about the challenges in the use of surgical simulators and included possible solutions to these challenges. A lack of optimal assessment methods was mentioned as one of the major gaps in training with a surgical simulator [14]. Although performance metrics have been explored previously for arthroscopy, the assessment has been limited to a few metrics. This problem limits the criteria of expertise that can be explored and also our knowledge of the skills required for arthroscopy. The knowledge of arthroscopic proficiency and standardized assessment methods are inadequate for realizing an appropriate training system.

1.2 General Problem Statement

One of the factors that influences the quality of health care is fluid, purposeful instrument motion, which can be limited for some surgeons. Currently, a major part of surgical competence assessment is performed by expert evaluators observing the performance of trainees, whose performance is rated based on Global Rating Scales. However, this assessment method is subjective and the ratings vary for different evaluators [15]. This inconsistency of evaluation was also noticed in a previous study in our group. In this study, two different expert surgeons blindly evaluated the performance of trainees by watching video of their performance. A significant percentage of subjects were rated differently by the two expert surgeons. Differing opinions are inevitable because there is no clear definition of proficiency, and also because there might be features of performance that are not observable in a video.

Developing an objective and comprehensive method of surgical skills assessment prevents medical errors by preventing less-prepared surgeons from operating on human patients, or at least ensuring that developing surgeons are supervised by proficient surgeons. A better understanding of what constitutes surgical proficiency will also inform the development of more appropriate training simulators that can focus on enhancing the required skills at different stages of learning. In addition, this knowledge is beneficial for the development of training curricula.

This thesis, proposes a new method of investigating surgical skills by introducing energy expenditure metrics. Since energy expenditure is mentioned as one of the features of general motor skills [16], investigating this parameter might assist in developing a more comprehensive surgical skills assessment method. Including energy expenditure, in addition to the previously developed metrics that have shown high capability in skills assessment, can enhance the accuracy of the current assessment methods, enable us to differentiate between detailed levels of expertise, and improve the guidance provided to trainees.

1.3 Objectives

The main goal of this thesis is to enhance surgical, and specifically, arthroscopic skills assessment. The optimal assessment method should include the following characteristics:

- be objective, and not require the presence of an assessor,

- distinguish between subjects with different levels of expertise, including those with in-termediate skill levels,

- indicate the trainee's dexterity as well as the patient's safety, and not rely solely on the task outcome,

- be computationally efficient in order to inform trainees about their performance in a reasonable amount of time.

After carefully studying the previously developed metrics for arthroscopy and other MIS procedures, it was noticed that surgical proficiency has not been investigated from the view point of energy expenditure. The focus of this thesis is to investigate energy expenditure during arthroscopic performance.

1.4   Overview of the Thesis

In considering energy expenditure to assess surgical skill, two main approaches were pursued: 1) evaluating the kinematics of instrument motion, and 2) exploring muscle activity features. This thesis is organized into the following chapters to address its objectives:

Chapter 2   Literature review: Various metrics that have been used for arthroscopy and other areas of MIS that can inform arthroscopic skills assessment are reviewed in this chapter. In addition, various types of surgical simulators and methods of implementing the assessment are reviewed.

Chapter 3   Analysis of Energy-based Metrics for Laparoscopic Skills Assessment: This chapter considers energy metrics based on instrument motion and contact force. Various types of energy are combined, and the optimized combination is found. A novel method is proposed and evaluated for improving the efficiency of detecting the change in energy expenditure according to different levels of experience. This chapter provides an initial analysis of energy

metrics using a laparoscopy data set to evaluate the concept of energy-based performance metrics.

Chapter 4   Energy-Based Metrics for Arthroscopic Skills Assessment: In this chapter, energy-based metrics are investigated for arthroscopic skills through a different approach. Machine learning algorithms are explored to classify trainees based on the energy content of their performance.

Chapter 5   Muscle Activity Analysis for Surgical Skills Assessment: This chapter investigates the use of muscle activity for developing skills assessment metrics. These metrics are cross validated to detecting the level of expertise of trainees. In addition, forearm motions are inspected and found to be informative for differentiating between various states of training.

Chapter 6   Conclusions and Future Work: In this chapter, the methods and findings of this thesis are summarized. Suggestions and guidelines for future research in this area are also provided.

Appendix A   Sensitivity Analysis of Energy-based Metrics to Additional Mass: This appendix provides a complimentary analysis on the effect of removing mass from the energy equations.

Appendix B   Approvals: The ethics approval is provided in this chapter.

Bibliography

[1] S. Bayonat, M. García, C. Mendoza, and J. Ferniindez, "Shoulder arthroscopy training system with force feedback," in *International Conference on Medical Information Visualisation-BioMedical Visualisation (MediVis)*, London, England, Jul 5–7, 2006, pp. 71–76.

[2] R. Treuting, "Minimally invasive orthopedic surgery: arthroscopy," *The Ochsner Journal*, vol. 2, no. 3, pp. 158–163, 2000.

[3] R. J. Koehler, S. Amsdell, E. A. Arendt, L. J. Bisson, J. P. Bramen, A. Butler, A. J. Cosgarea, C. D. Harner, W. E. Garrett, T. Olson *et al.*, "The arthroscopic surgical skill evaluation tool (asset)," *The American Journal of Sports Medicine*, vol. 41, no. 6, pp. 1229–1237, 2013.

[4] J. Espregueira-Mendes and M. Karahan, "Motor learning principles for arthroscopic motor skill teaching," in *Motor Skills Training in Orthopedic Sports Medicine*. Springer, 2017, pp. 7–15.

[5] P. C. Unalan, K. Akan, H. Orhun, U. Akgun, O. Poyanli, A. Baykan, Y. Yavuz, T. Beyzadeoglu, R. Nuran, B. Kocaoglu *et al.*, "A basic arthroscopy course based on motor skill training," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 18, no. 10, pp. 1395–1399, 2010.

[6] M. Karahan, G. M. Kerkhoffs, P. Randelli, and G. J. Tuijthof, *Effective Training of Arthroscopic Skills*. Springer, 2015.

[7] M. P. Hall, K. M. Kaplan, C. T. Gorczynski, J. D. Zuckerman, and J. E. Rosen, "Assessment of arthroscopic training in us orthopaedic surgery residency programs: a resident

self-assessment," *Bulletin of the NYU Hospital for Joint Diseases*, vol. 68, no. 1, pp. 5–5, 2010.

[8] B. McCrory, C. A. LaGrange, and M. Hallbeck, "Quality and safety of minimally invasive surgery: past, present, and future," *Biomedical Engineering and Computational Biology*, vol. 6, p. 1, 2014.

[9] D. Stefanidis, M. W. Scerbo, J. R. Korndorffer, and D. J. Scott, "Redefining simulator proficiency using automaticity theory," *The American Journal of Surgery*, vol. 193, no. 4, pp. 502–506, 2007.

[10] A. L. Trejos, "A sensorized instrument for minimally invasive surgery for the measurement of forces during training and surgery: development and applications," PhD Thesis, Dept. of Electrical and Computer Engineering, University of Western Ontario, 2012.

[11] A. Escoto, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "Force sensing-based simulator for arthroscopic skills assessment in orthopaedic knee surgery." in *Medicine Meets Virtual Reality 19, NextMed*, 2012, pp. 129–135.

[12] A. Escoto, F. Le Ber, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "A knee arthroscopy simulator: design and validation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, July 3-7, 2013, pp. 5715–5718.

[13] L. C. McCracken, "Development of a physical shoulder simulator for the training of basic arthroscopy skills," Master's Thesis, Biomedical Engineering Program, University of Western Ontario, 2015.

[14] D. Stefanidis, N. Sevdalis, J. Paige, B. Zevin, R. Aggarwal, T. Grantcharov, D. B. Jones, Association for Surgical Education Simulation Committee, *et al.*, "Simulation in surgery: what's needed next?" *Annals of Surgery*, vol. 261, no. 5, pp. 846–853, 2015.

[15] L. S. Feldman, S. E. Hagarty, G. Ghitulescu, D. Stanbridge, and G. M. Fried, "Relationship between objective assessment of technical skills and subjective in-training evalua-

tions in surgical residents," *Journal of the American College of Surgeons*, vol. 198, no. 1, pp. 105–110, 2004.

[16] E. R. Guthrie, *The Psychology of Learning (Rev. ed.)*.   New York: Harper and Row, 1952.

Chapter 2

Introduction

2.1 Introduction

The number of procedures that are performed through minimally invasive surgery (MIS) is increasing due to the advantages that this type of surgery promises patients [1]. However, the reduced access conditions imply that surgeons require extensive training to become experts at performing surgical tasks in tight environments and with reduced force feedback. Surgeons are required to develop the necessary skills to triangulate within confined spaces, visualize 3D objects from magnified 2D images, and improve their hand–eye coordination ability [2,3]. Surgical simulators provide trainees with the opportunity to practice these skills before proceeding to the operating room [4–6]. Practicing with simulators offers certain advantages such as allowing practice to occur frequently and conveniently, with high accessibility, increased safety for patients, and providing more relaxed conditions for trainees. Surgical simulators would be even more efficient if they were able to provide proper objective feedback to trainees [7]. Such feedback eliminates the need for an expert evaluator and allows trainees to practice independently. However, providing constructive feedback to trainees and assessing their skill level is only possible if proper objective assessment methods are available. Different researchers have proposed a variety of metrics and assessment methods for evaluating the progress of trainees for different type of MIS procedures. Although the basic principles of all MIS procedures are similar, in the sense that no direct sense of touch is obtained and the surgery is performed with long instruments, each type of MIS has certain characteristics. Different sizes of instruments,

in terms of diameter and length, affect the relationship between movement of the handle and the tip of the instrument. In addition, the available space in the human body around the instruments affects the motion profile. As there is constrained space in human joints, the range of motion in arthroscopy is more limited. Vulnerability/hardness of the target tissue also requires practicing the appropriate interaction with tissue, i.e., being gentle enough and effective at the same time. For instance, surgeons need to learn to apply sufficient force when working on bones or on arteries.

Robotic Assisted Minimally Invasive Surgery (RAMIS) is also developed based on similar concepts. Robotic systems usually allow hand motion scaling, 3D visualization, and provide intuitive movements and tremor compensation. Similar to MIS procedures, special training is required for RAMIS for learning the interaction with the robot and its effect on tissue. Altogether, the running principles for different types of MIS are similar, but specific training and assessment methods are required for various surgical conditions. In this chapter, various performance metrics that have been explored for arthroscopy are reviewed. In addition, the studies related to other MIS procedures that can be applicable to arthroscopy are discussed.

In Section 2.2, MIS and its challenges are briefly explained. In Section 2.3, the learning of surgical skills is discussed. Sections 2.4 and 2.5 provide a review of various metrics for technical skills assessment and previous classification methods used for MIS. Finally, in Section 2.6, the shortcomings of the current state of surgical skills assessment are discussed.

2.2 Minimally Invasive Surgery, Benefits and Complications

MIS provides certain advantages for patients, such as shorter hospital stay, less blood loss, and decreased pain and infection risk. Consequently, attention towards this type of surgery has increased [1]. Accordingly, the number of arthroscopic surgeries has increased in recent years. As reported by the American Board of Orthopaedic Surgery (ABOS), knee and shoulder arthroscopy are two of the most commonly performed orthopaedic procedures [8]. However, a high-quality procedure is essential for obtaining the same outcomes as with open surgery. Low-quality performances might negate the benefits of MIS, result in reoperation, or even irreversible damage to patients. The quality of health care can also affect the overal costs of

the health care system [9].

MIS complications increase the possibility of erroneous actions, which result in medical injuries. According to McCrory *et al.* [10] medical errors cause a larger number of deaths than breast cancer and car accidents in the United States. Many of these medical injuries can be prevented by implementing appropriate training and skills assessment methods [10]. In recent years, improvements have been achieved by utilizing engineering knowledge to improve the quality of health care [10]. The development of surgical simulators that are equipped with automated objective assessment methods allows trainees to practice surgical tasks and enhance their technical skills, which ultimately prevents possible injuries to patients [2, 3].

2.3   Learning Surgical Skills

Surgical training has conventionally been performed through the apprenticeship model. This method relies on observing the performance of an expert and gradually increasing involvement of the residents in the operation [11]. The apprenticeship model has disadvantages, such as inconsistency in training, inefficiency in time and cost, and lack of objective assessment and feedback [11, 12]. Surgical simulators are an alternative to the apprenticeship model. Simulators allow trainees to practice surgical tasks multiple times, without time limitations or concerns about patient safety. Various types of surgical simulators and tasks that can be practiced on simulators are described in the following subsections.

An important question in surgical training is to determine what indicates competency in MIS and how much experience is required to obtain it. The criterion of being an active member of the Arthroscopy Association of North America (AANA) is performing 50 arthroscopic operations annually [13]. However, the amount of practice needed to reach proficiency is not clear.

The transferability of surgical skills learned on a simulator to the operating room has not been established yet. Demonstrating the effect of simulator practice on dexterity in the operating room is not easy to accomplish. However, in a follow-up research, it was observed that, after practicing for three years in the operating room, subjects improved their performance on the simulator [13]. For aviation simulators, it has been reported that one hour of training

on the simulator corresponds to half an hour of training on an airplane [14]. In a study by Stefanidis *et al.* [15], it was demonstrated that the transferability of surgical skills to the operating room is superior for trainees who continued their practice until reaching the level of automaticity, compared to participants who stopped practicing after the level of proficiency. In this study [15], automaticity was tested by a secondary task and the NASA-TLX questionnaire. Learning psychomotor skills consumes a huge amount of cognitive energy. As a result, novice trainees usually do not possess the ability to multitask in the early stages of their training. However, through practice, the cognitive requirements for performing surgical tasks decrease [16]. According to [15], robust learning can be achieved after reaching automaticity.

2.3.1 Surgical Simulators

Various types of surgical simulators have been used for training surgical skills including training boxes, physical models, virtual reality (VR) simulators, and cadaver models [17]. Sutherland, *et al.* [18] have reviewed the studies that compared the usefulness of these simulators. A consistent finding of all of these studies was that training with simulators improves performance. This review article, [18], reports controversial findings on the superiority of various types of simulators. Based on this article, it is not clear if training with computer simulation provides a better training than standard surgical drills. In addition, the amount of data on comparison between VR and physical models were not enough to deduce a certain conclusion. It should be noted that these findings are not affected only by the type of simulator but are affected also by the tasks that were studied and the number of participants.

VR simulators have the advantage of realistic modeling of detailed anatomical structures. LapSim is a VR laparoscopic model that is capable of simulating bleeding and tissue deformation [14, 19]. Although some VR models have been equipped with haptic feedback [20, 21], these models still lack the ability to produce a realistic sense of contact with tissue. Munz, *et al.* [14] found higher levels of performance improvement for subjects who practiced on a box trainer than those who practiced on the LapSim model.

Based on a survey of the European Society of Sports Traumatology, Knee Surgery and Arthroscopy (ESSKA) [22], high fidelity simulators are ranked as the most useful training platforms after cadaver models. The highest fidelity can be found in cadaver models. How-

ever, standardization of training cannot be achieved for these models due to pathological and anatomical variations between different cadaver samples [13]. Physical models have been developed for various MIS procedures such as arthroscopy [5, 6, 12], laparoscopy [23], thoracoscopy [24], etc. Focusing on arthroscopy, the Alex Shoulder Professor models and Fully Encased Knee Joint with Patella and Ligaments by Sawbones$^®$ (Vashon Island, Washington) are among the commercially available training platforms for arthroscopy. The Arthroscopy Shoulder with Replaceble Skin by the same company allows for the practice of portal placement. Box trainers are another group of physical models that allow surgical tasks to be practiced on animal tissues or synthesized models with low cost. Additional details about various surgical simulators are provided in [22].

2.3.2 Previously Studied Surgical Tasks

Suturing, which consists of needle driving and knot-tying, is one of the most studied laparoscopic tasks in the literature [25, 26]. In the study by Rodrigues *et al.* [25] a double knot and a three-loop knot were performed on the left needle driver, and one single knot was performed on the right needle driver. In the single, double, and three-loop knots the thread was spiraled once, twice, and three times around the needle, respectively; and the other end of the thread was pulled inside the loops to tighten the knot. In this study, the level of difficulty of this task was scored higher than 5 in a 7-point Likert scale by participants.

Peg transfer and shape cutting [27] are other commonly practiced tasks [26,28,29]. The peg transfer task consists of picking up a number of pegs with one hand (usually the non-dominant hand), transfer the pegs to the other hand, and place them on another board or another side of the same board. The shape cutting task is performed using laparoscopic scissors and requires cutting the edges of a pre-drawn shape on a piece of paper/gauze [30].

Pedowitz *et al.* [31] have studied arthroscopic knots on a Sawbones arthroscopy training station. Following that, the completed knots were tested using the Sawbones knot tester. It was found that, for experienced surgeons, postgraduate year (PGY) 4-5, and residents, the rates of knot failure were 22%, 26%, and 11%, respectively. The inferior performance of experts can be due to unrealistic simulation conditions or inappropriate design of the task. In a study by Srivastava *et al.* [32] variety of navigation tasks were investigated on a virtual

reality shoulder simulator (Mentice Corp. Procedicus shoulder arthroscopy simulator). These tasks included anatomic identification, which required finding 10 anatomic structures in the glenohumeral joint. Trainees were asked to find the structures and then press a foot pedal when the structure was in the middle of the screen. Another task consisted of touching 15 randomly generated colored balls with a probe. A third task addressed scoping and required subjects to find and place a tack in the appropriate view of the scope and press a foot pedal upon a change in the color of that tack. In probe manipulation and anatomical identification, the only significant difference between the three groups of subjects in this study—novice, intermediate, and expert— were found between experts and novices for the third repetition of the tasks and in task completion time. This could have been due to lack of realism in the virtual reality model, inappropriate task design, and the learning curve for the expert group. For the navigation task, however, the difference between experts and novices–intermediate groups was significant. A considerable improvement was also found for the expert groups in the second and third task repetitions. This improvement indicates the learning curve of the experts due to the difference between the model and the human shoulder.

In another study by Howells *et al.* [4] on an Alex Shoulder Professor (Sawbones, Malmö, Sweden), participants completed a probing task that included probing 9 points inside the model. Appropriate probing of these points was monitored by one of the authors of the study. Another task studied in [4, 20] was grasping and removing a ball bearing or other foreign bodies from a joint. Resection with a shaver or punch of parts of a menisci that were specified to trainees by colour is another arthroscopic task studied in the literature [20].

2.4   Surgical Skills Assessment

In previous studies, different assessment methods have been developed and investigated. These methods range from global rating scales (GRS), such as the Arthroscopic Surgical Skill Evaluation Tool (ASSET) used for the knee [33] and the Global Rating Scale for Shoulder Arthroscopy (GRSSA) [34], to assessment methods based on technical parameters. The goal of all of these methods is to provide objective and consistent assessments. However, some of them, such as GRSs require expert evaluators to rate the performance of subjects, which reduces the objec-

tiveness of the proposed metrics. Many metrics aimed at providing objectiveness have been developed, as described in the following subsections.

2.4.1   Pre-processing Techniques

Segmentation methods have been used in previous studies to divide surgical tasks into smaller sections to increase understanding of the difference between the gestures of novices and experts. Ahmidi *et al.* [35] processed tool motions of a septoplasty tissue dissection task. The segmentation was performed based on deviations from the septal plane, moving from the local minimum distance to the next local maximum distance. Additional filtering and constraints including time, length, and position were applied to the segmented data. This group has investigated various automatic segmentation algorithms based on machine learning algorithms, such as skip-chain conditional random chain, to decompose surgical tasks into smaller motions.

Despinoy *et al.* [29] introduced a new technique for automatically recognizing sub-gestures, called *dexemes*, and then classifying these dexemes into their related surgical gestures, called *surgemes*. The position and orientation of the instrument were analyzed to calculate curvature and torsion in several peg transfer tasks and a task consisted of drawing the letter R. Manual segmentation methods based on video recordings has also been employed to decompose different stages of surgical task performance [36]. For instance, in a suturing task, the needle driving and knot-tying were separated in [37] and each one was analyzed individually.

2.4.2   Technical Metrics

In this section, assessment methods developed based on technical metrics that do not need an external evaluator are reviewed.

2.4.2.1   Temporal Metrics

One of the most used metrics for assessing trainee performance is task completion time. It has been shown to best discriminate between novice and expert subjects. Task completion time has been used in [4, 5, 38–44] for skills assessment. It is not clear, however, whether or not performing a task quickly implies better performance. Consequently, in most studies, time

has been combined with other performance metrics to provide a more objective assessment. The number of repetitions required to complete a task successfully is also considered as an assessment method, although secondary criteria need to be defined to determine whether a task has been completed successfully [22]. In addition, hesitation can be determined by time intervals between subtasks [37] or by short intervals in which the instrument does not move [22].

In the study described in [32] that explored the performance of arthroscopic tasks on a virtual reality simulator, the difference between experts and novices in terms of task completion time was statistically significant only in the third trial. Nevertheless, experts always demonstrated shorter task completion times.

2.4.2.2 Outcome Metrics

Assessing the outcome of a task, regardless of how it is performed, is used for skills assessment in some studies [45]. These metrics are defined based on the specifics of the tasks. Examples of outcome metrics are the number of probe and/or scope collisions [38–42, 46], the number of successful identifications of landmarks or structures, and the score assigned to a performance assessment [39, 47]. The number of probe and/or scope collisions can also be an indication of safety and can be determined by identifying the number of times a certain force threshold has been exceeded.

2.4.2.3 Motion-based Metrics

Motion analysis is an objective method of skills assessment. Instrument and hand position have been used to extract biomechanical parameters to determine the proficiency of trainees and surgeons [48]. Several metrics are extracted from motion information, as discussed in the following subsections.

Path Length

The distance that the probe or scope travels has been used in [4, 40–44, 46]. It can be calculated using the following formula in Cartesian space [48].

$$P = \sum_{i=\text{start}}^{\text{end}} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}, \tag{2.1}$$

where x_i, y_i, and z_i represent position in three Cartesian directions for each time step.

In the same way, economy of movement is defined by dividing the ideal path length by the actual path length [22, 40, 49]:

$$E = \frac{P_{\text{ideal}}}{P}. \tag{2.2}$$

The ideal path length can be defined by measuring the minimum distance between the starting point and the target, or based on performance of expert surgeons.

Depth Perception

Depth perception is defined based on movement of the instrument in its longitudinal direction. This metric reflects the perception that subjects have when they convert from the 3D world to a 2D view on a screen, and the ability of subjects to control the instrument. To calculate depth perception, the coordinate system must be converted to that of the instrument to determine movements in the longitudinal direction of the instrument [22].

$$D = \sum_{i=\text{start}}^{\text{end}} \sqrt{(l_{i+1} - l_i)^2}, \tag{2.3}$$

where l_i is displacement in the longitudinal direction of the instrument for each time step.

Volume of Motion

Volume of motion corresponds to the volume of an ellipsoid constructed from the standard deviation of movement around 3 main directions. It reflects the required 3D space for the subject to complete the task. This metric can be easily influenced by the direction of movement— if the instrument does not move in one direction, the total volume will be zero [22].

Out-of-View Time

Out-of-view time demonstrates the time span in which the instrument is not in the view of the scope. This condition compromises the safety of the patient as the instrument may damage tissues when it cannot be seen. Out-of-view time can either be calculated from the stream of video images from the scope, or from tracking position of the instrument and scope and defining the view cone, knowing the scope's imaging characteristics [22, 50].

Tip-to-Tip Distance

The tip-to-tip distance, $d_{\text{t–t}}$, indicates the sum of the distance between the instrument tip and the scope tip during the task. This metric is similar to "out-of-view time" in the sense that it can also reflect whether the instrument is out of the scope view, when the distance is more than a certain value. This metric is highly dependent on the task as the distance varies due to task requirements.

$$d_{\text{t–t}} = \sum_{i=\text{start}}^{\text{end}} \sqrt{(x_{i \text{ scope}} - x_{i \text{ probe}})^2 + (y_{i \text{ scope}} - y_{i \text{ probe}})^2 + (z_{i \text{ scope}} - z_{i \text{ probe}})^2} \qquad (2.4)$$

Speed

Speed, the first derivative of position, is another characteristic of trainee motion that is used in the literature as an assessment metric [41, 42]. Different forms of speed such as the normalized speed, the mean speed, the peak speed, the magnitude of the velocity vector, and the number of changes in velocity have been proposed as assessment metrics [48]. Speed can be interpreted as the subject's ability to control position and instrument movements.

Acceleration

Acceleration, the second derivative of position, is another common metric used for skills assessment for arthroscopic simulators. The mean acceleration, the maximum acceleration, and the number of accelerations and decelerations are some of the metrics defined based on acceleration [48]. The integral of the magnitude of the total acceleration vector, IAV, which represents the energy expenditure, is another metric calculated as [40, 51]:

$$IAV = \sum_{i=\text{start}}^{\text{end}} \sqrt{\left(\frac{\Delta^2 x_i}{\Delta t^2}\right)^2 + \left(\frac{\Delta^2 y_i}{\Delta t^2}\right)^2 + \left(\frac{\Delta^2 z_i}{\Delta t^2}\right)^2}, \tag{2.5}$$

where $\frac{\Delta^2(\cdot)}{\Delta t^2}$ is the second derivative in discrete time.

Jerk

Jerk, which is defined as the third derivative of position, represents the smoothness of motion. A lower jerk value indicates a smoother motion [48, 51]. The normalized jerk was defined in [52] as:

$$J_{\text{norm}} = \sqrt{\frac{T^5}{2A^2} \sum_{i=\text{start}}^{\text{end}} \left[\left(\frac{\Delta^3 x_i}{\Delta t^3}\right)^2 + \left(\frac{\Delta^3 y_i}{\Delta t^3}\right)^2 + \left(\frac{\Delta^3 z_i}{\Delta t^3}\right)^2\right]}, \tag{2.6}$$

where T and A are task completion time and the amplitude of the motion, respectively; and $\frac{\Delta^3(\cdot)}{\Delta t^3}$ is the third derivative in discrete time.

2.4.2.4 Force-based Metrics

Force information can be used in the development of objective metrics [5, 44, 53]. However, due to complications in recording the applied force, this data has not been commonly used in performance assessments [9]. It is essential for trainees to learn to apply sufficient force when needed but to be gentle enough with the tissues. Applying too much force may result in tissue damage; however, applying less force than required may lead to ineffective performance. Some of the developed force-based metrics are the average force, the maximum force, the integral of the force, the force range, the force direction, the interquartile range, derivatives of the force, and smoothness of the applied force [22, 37, 48]. These metrics can also be an indication of performance safety and efficiency.

2.4.2.5 Optimized Combination of Metrics

Optimized combined metrics are developed by combining various metrics through an optimization method such as linear and non-linear least square methods. These metrics consider multiple characteristics of performance in one unified metric. Different combined metrics are defined and studied in [37, 54–56].

2.4.2.6 Muscle Activity-based Metrics

MIS demands different muscle activity compared to open surgery due to its difficult postures, use of long instruments, which mostly are not ergonomically comfortable, limited degrees of freedom in manipulating the instruments, etc. According to [57], the muscle activity and total effort required to perform tasks using a laparoscopic grasper are significantly higher than that required for doing the same tasks with open surgery instruments, e.g., a Crile hemostat.

The level of fatigue and stress in MIS is also considerably higher than in open surgery. Fatigue affects the ability of surgeons to focus on the task, decreases their ability to make appropriate surgical decisions, and reduces their ability to use their surgical skills and dexterity [58, 59]. According to Uhrich *et al.* [26] muscle fatigue is lower among surgeons with higher levels of expertise; however, it is still high enough to produce chronic injuries for MIS surgeons. The thresholds for low, medium, and high contraction muscle activities are defined in [60] as 2–5% of Maximum Voluntary Contraction (MVC), 10–14% of MVC, and 50–70% of MVC, respectively.

Analyzing EMG signals allows investigation into patterns of muscle activation that are beneficial to increasing our knowledge about surgical skill development and also identifying the hazards that might be associated with certain surgical instruments and postures for surgeons [58]. This information can be used to adjust the current surgical postures, video display systems [26, 61], and table heights [58], or to develop new instruments that decrease surgeon discomfort and chance of chronic injuries. Different instrument handles have been extensively investigated in [62–64].

The areas of the surgeon's body that are actively involved during MIS procedures are usually the neck, arms, and shoulders [58, 65]. It is hypothesized that increased efficiency in performing surgical tasks can be quantified through analyses of physiological parameters such

as electromyography [66]. This hypothesis has been investigated in a few studies for robotic surgery on the da Vinci^{TM} system. It was shown in [66–68] that practice reduces muscle activity.

Most of the studies that used EMG metrics for evaluating posture or newly developed instruments performed the analysis on participants with approximately the same level of expertise. However, Shafti *et al.* [69] compared the STIFF-FLOP arm with conventional laparoscopic instruments, while evaluating its effect on expert and novice subjects separately. In this study, expert subjects demonstrated considerably lower Root Mean Square (RMS) of EMG magnitude and slightly higher median frequency than novice subjects.

In the following paragraphs, EMG-based metrics that have been used in surgical skills assessment, posture analysis, and other areas related to MIS are described.

Muscle Activation Volume

Muscle activation volume (EMGV) is calculated by integrating the ratio of the EMG signal, which is recorded during the task, to the MVC over time [66, 67]. In [66], a significant reduction in EMGV was found for the extensor digitorum, when comparing pre-practice tests with post-practice and retention tests. In this study, [66], a series of surgical tasks including bimanual carrying, needle passing, suture-tying, and grasping were performed on a da Vinci Surgical System. As was mention in Section 2.1, RAMIS shares common features with manual MIS, which compels investigating this metric for other types of MIS as well. However, it is mentioned in this study [66] that the reduction in EMGV might be due to the reduction in task completion time.

Muscle Activation Rate

Muscle activation rate (EMGR) is calculated by dividing EMGV by the total time that was spent to complete the task. EMGR was investigated in [66, 67] to study surgical robotic performance using the da Vinci System. This metric was less successful than EMGV in demonstrating the difference between pre-practice and post-practice tests in [66].

Relative Activation Time

Relative activation time (RAT) was proposed by Quick *et al.* [70] to compare different lapro-

scopic tasks and instruments. This metric is defined as the time duration that muscle activity exceeds 10% of MVC. As this metric is defined based on the on and off times of muscle activity, it is suitable for tasks that follow a specific temporal order. In addition, in [70], all of the participants were expert surgeons, which reduced variability in task performance. Overall, the more time that a muscle activates, the higher the level of fatigue expected for that muscle [58]. This metric demonstrated significant differences between two different graspers and between two of the three studied tasks.

Median Frequency

The median frequency (MF) of the EMG power spectrum is another metric that is a representation of muscle fatigue [68].

$$\text{MF:} \int_0^{f_{\text{med}}} P(f)df = \int_{f_{\text{med}}}^{f_{\text{max}}} P(f)df, \tag{2.7}$$

where P is power spectrum and f_{med} and f_{max} are median and maximum frequencies of the power spectrum. The higher median frequency is associated with faster performance and less muscle fatigue. It was shown in [68] that training for robotic surgery increases median frequency. In [69] the median frequency was calculated over time and the variation in its value was explored by calculating the Coefficient of Variation (CV). CV equals the standard deviation of MF over its mean value. However, CV of MF did not show a significant difference between novices and experts. Suh *et al.* [71] used median frequency to investigate the effect of distraction on performance. According to the results of their study, distraction has a significant effect on MF. The EMG envelope was also calculated. Overall, it was shown that muscle work increases in the presence of distraction.

Frequency Bandwidth

Frequency bandwidth is the difference between the maximum and the minimum frequency at which the power spectrum is half its maximum level. Bandwidth indicates the range of muscles that are involved in the performance [68].

Dynamic EMG Metrics

Recurrence Quantification Analysis (RQA) and recurrence plots have been incorporated into [72] to study muscle fatigue during MIS. The determinism percentage is a novel metric proposed in this study, which is defined based on the predictability of the method and the diagonal lines in the recurrence plot. According to Panahi *et al.* [72] determinism percentage is associated with muscle fatigue. muscle activity of the upper arm and shoulder was investigated during a real surgical operation. The trapezius and deltoid were activated continuously during MIS to allow the surgeon to maintain posture and watch the video display. Biceps and triceps were recruited to manipulate instruments in higher frequencies than the deltoids and trapezius. Consequently, higher levels of fatigue were noticed in the deltoids and trapezius. Since the EMG sensors could not be attached to sterilizable parts of the arms and hands, forearm and hand muscles were not examined in this study.

The results of the above-mentioned studies show that the proposed metrics are highly task dependent and more studies should be performed to define EMG-based metrics with the capability of discriminating between subjects with different levels of proficiency for each task. Future EMG signal analysis for the development of objective performance metrics should incorporate different parameters of the EMG signal and study the pattern of muscle activation during different MIS tasks.

2.5 Classifying Trainees based on their Performance

The criteria of expertise can be defined based on various performance features. Some studies have investigated the applied forces required to operate on delicate tissues, such as menisci, the facial nerve, and sigmoid sinus [53, 73]. Since any damage to these tissues can be irreversible, special care must be taken when working around vulnerable tissues. In a study by Tuijthof *et al.* [53] the force levels for safe probing of menisci were investigated using cadaver samples. A force sensor was placed beneath the samples to measure the applied force. According to the results, the maximum force level that is safe for probing menisci is 8.5 N. The tasks that were studied were probing and lifting the surface of menisci. In addition, the performance of expert and novice subjects in these tasks were compared. It was found in this study and

also in [43] that expert subjects demonstrated higher levels of force. This can be a sign of dexterity and novice trainees might need to be encouraged to incorporate higher levels of force when performing arthroscopic tasks. However, other studies have found a different result, i.e., experts apply lower amounts of force than novice trainees [37, 44]. A shortcoming of the Tuijthof's study, [53], is that measuring force below the sample might reduce the accuracy of force measurement. In addition, the cadaver menisci samples are stiffer than live tissue. This issue was justified by using cadaveric menisci samples of elderly people, which usually represent lower stiffness than that of young people. The force threshold is a reasonable criteria for assessing trainees and providing feedback to them. However, this limits the assessment into only one feature of performance. In order to develop a comprehensive evaluation, other characteristics such as motion and temporal features should be considered. To analyze multiple features of performance, machine learning algorithms have been evaluated in the literature [45]. These algorithms are reviewed in the next paragraph.

Vendula *et al.* [9] published a thorough review of the algorithms that have been used for surgical skills assessment for classification. According to this study, the classifiers can be divided into three groups based on their input: 1) the group that uses extracted features of performance as its input. In this group, the size of the input is fixed for all of the trainees. 2) the group of classifiers that utilize time-series variables as their input, and 3) the group of classifiers that are implemented for use in conjunction with histogram-based and dictionary variables. For the first group of these classifiers, two categories can be defined: discriminative and generative. In the generative models, the probability distribution of parameters is generated based on prior knowledge; then the probability of belonging to each class is calculated for the target data using Bayes' theorem. Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are examples of generative models. In MIS skills assessment, descriptive models are more prevalent. Descriptive classifiers calculate the probability of belonging to each class directly. Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Linear Regression (LR) are examples of algorithms that establish a linear discrimination criteria, and Neural Networks (NN), Support Vector Machines (SVM), and K Nearest Neighbors are examples of algorithms that establish a non-linear discrimination.

HMM and Descriptive Curve Coding (DCC) methods can be used with time-series data as

their inputs [74, 75]. These models are sensitive to the input data. For instance, HMM is very sensitive to the coordinate system when developing the model based on position data. These methods are also not very robust to new samples that were not used in the model development [76].

In applications for which the size of input data is very large, pre-processing techniques are used to extract the most valuable features. Principal Component Analysis (PCA) can be used to extract information and reduce the dimension of data when many variables might represent the same features. The PCA algorithm generates another set of variables that represent the original data's variation. The size of PCA generated variables is the size of the original data. However, in most studies, the first few components can represent more than 75–90% of variation of data. Parts of the information in the original data might be lost in the process of calculating principal components. Nevertheless, the reduced dimension permits easier analysis and visualization of the data [77]. This method has been used in [3, 35, 76] for MIS data processing.

The above-mentioned algorithms have been used in the following surgical skills assessment studies:

- Zirkle *et al.* [78] used LR to evaluate the correlation between experts' opinion with various methods of assessment such as Global Rating Scales, task-based Checklist, and final product analysis.

- Ahmidi *et al.* [35] used HMM and DCC to evaluate performance in septoplasty based on motion of the instruments. The results of this study demonstrated higher accuracy levels for the DCC method. In addition, in [74] a performance index was defined based on similarity of the Markov model of the trainee's with that of the experts and novices based on force/torque data. In this study, the data recorded from 3 novice and 3 expert subjects were used for training the HMMs and the data that was recorded from another group of participants, including 2 novices and 2 experts, was used for evaluation. This method demonstrated 87% accuracy in determining the level of expertise of trainees.

- The SVM method has been used in [35] for classifying surgeons in septoplasty and in [29, 76] for robotic-assisted MIS practice on a simulator.

- In the study by Despinoy *et al.* [29] the KNN method was developed for classifying

surgical gestures performed by the Raven-II robot. In this study, the KNN method, when K was 5, provided superior accuracy to the SVM method.

- LDA is another commonly used algorithm in MIS studies. In [3, 79] this algorithm has been used for classifying trainees who performed laparoscopic tasks on box trainers.

- Richstone *et al.* [80] investigated the metrics developed based on eye movements and evaluated their metrics with the NN and LDA methods. The NN method demonstrated slightly more accurate results.

2.6 Shortcomings of Current Metrics

In 2015, a study was published and discussed by the Association for Surgical Education (ASE) Simulation Committee concerning the challenges in the use of surgical simulators and included possible solutions to these challenges. A lack of optimal assessment methods and constructive feedback were mentioned as two of the major gaps in training with a surgical simulator [81]. Further investigation is required to increase the knowledge on psychomotor skills. There cannot be a trend defined for a particular metric or set of metrics and claim that this trend applies to every task. For example, in many studies it has been shown that experts demonstrate shorter path lengths than novices. However, in a study by Tashiro *et al.* [44] it was noticed that during meniscectomy the expert subjects demonstrated longer path lengths than the novice subjects. By adopting larger movements, the experts make the procedure easier and the resulting higher path length cannot be classified as unnecessary. This study shows that skills assessment techniques should be enhanced and possibly new metrics should be developed that can represent both general and detailed features of proficiency.

While stress can also produce fatigue and might result in negative effects on surgical outcomes, it is outside the scope of this study. In this project, the effect of dexterity on muscle activity is investigated.

Bibliography

[1] P. Modi, A. Hassan, and W. R. Chitwood, "Minimally invasive mitral valve surgery: a systematic review and meta-analysis," *European Journal of Cardio-Thoracic Surgery*, vol. 34, no. 5, pp. 943–952, 2008.

[2] J. D. Mason, J. Ansell, N. Warren, and J. Torkington, "Is motion analysis a valid tool for assessing laparoscopic skill?" *Surgical Endoscopy*, vol. 27, no. 5, pp. 1468–1477, 2013.

[3] T. Horeman, J. Dankelman, F. W. Jansen, and J. J. van den Dobbelsteen, "Assessment of laparoscopic skills based on force and motion parameters," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 805–813, 2014.

[4] N. R. Howells, M. D. Brinsden, R. S. Gill, A. J. Carr, and J. L. Rees, "Motion analysis: a validated method for showing skill levels in arthroscopy," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 24, no. 3, pp. 335–342, 2008.

[5] A. Escoto, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "Force sensing-based simulator for arthroscopic skills assessment in orthopaedic knee surgery." in *Medicine Meets Virtual Reality 19, NextMed*, 2012, pp. 129–135.

[6] A. Escoto, F. Le Ber, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "A knee arthroscopy simulator: design and validation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5715–5718.

[7] E. Boyle, M. Al-Akash, A. G. Gallagher, O. Traynor, A. D. Hill, and P. C. Neary, "Optimising surgical training: use of feedback to reduce errors during a simulated surgical procedure," *Postgraduate Medical Journal*, vol. 87, no. 1030, p. 524–528, 2011.

[8] W. E. Garrett Jr, M. F. Swiontkowski, J. N. Weinstein, J. Callaghan, R. N. Rosier, D. J. Berry, J. Harrast, G. P. Derosa, *et al.*, "American Board of Orthopaedic Surgery practice of the orthopaedic surgeon: Part-II, certification examination case mix," *Journal of Bone and Joint Surgery*, vol. 88, no. 3, pp. 660–667, 2006.

[9] S. S. Vedula, M. Ishii, and G. D. Hager, "Objective assessment of surgical technical skill and competency in the operating room," *Annual Review of Biomedical Engineering*, no. 0, 2016.

[10] B. McCrory, C. A. LaGrange, and M. Hallbeck, "Quality and safety of minimally invasive surgery: past, present, and future," *Biomedical Engineering and Computational Biology*, vol. 6, pp. 1–11, 2014.

[11] A. Insel, B. Carofino, R. Leger, R. Arciero, and A. D. Mazzocca, "The development of an objective model to assess arthroscopic performance," *Journal of Bone and Joint Surgery*, vol. 91, no. 9, pp. 2287–2295, 2009.

[12] C. Tay, A. Khajuria, and C. Gupte, "Simulation training: A systematic review of simulation in arthroscopy and proposal of a new competency-based training framework," *International Journal of Surgery*, vol. 12, no. 6, pp. 626–633, 2014.

[13] J. L. Hodgins and C. Veillette, "Arthroscopic proficiency: methods in evaluating competency," *BMC Medical Education*, vol. 13, no. 1, p. 61, 2013.

[14] Y. Munz, B. Kumar, K. Moorthy, S. Bann, and A. Darzi, "Laparoscopic virtual reality and box trainers: is one superior to the other?" *Surgical Endoscopy and Other Interventional Techniques*, vol. 18, no. 3, pp. 485–494, 2004.

[15] D. Stefanidis, M. W. Scerbo, P. N. Montero, C. E. Acker, and W. D. Smith, "Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: a randomized controlled trial," *Annals of Surgery*, vol. 255, no. 1, pp. 30–37, 2012.

[16] A. Dubrowski, R. Brydges, L. Satterthwaite, G. Xeroulis, and R. Classen, "Do not teach me while I am working!" *The American Journal of Surgery*, vol. 203, no. 2, pp. 253–257, 2012.

[17] K. S. Gurusamy and B. R. Davidson, "Laparoscopic surgical box model training for surgical trainees with no prior laparoscopic experience," *Cochrane Database of Systematic Reviews*, no. 3:Cd010478.48, 2014.

[18] L. M. Sutherland, P. F. Middleton, A. Anthony, J. Hamdorf, P. Cregan, D. Scott, and G. J. Maddern, "Surgical simulation: a systematic review," *Annals of Surgery*, vol. 243, no. 3, pp. 291–300, 2006.

[19] R. Kneebone, "Simulation in surgical training: educational issues and practical implications," *Medical Education*, vol. 37, no. 3, pp. 267–277, 2003.

[20] S. F. Fucentese, S. Rahm, K. Wieser, J. Spillmann, M. Harders, and P. P. Koch, "Evaluation of a virtual-reality-based simulator using passive haptic feedback for knee arthroscopy," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 23, no. 4, pp. 1077–1085, 2015.

[21] P. G. Roberts, P. Guyver, M. Baldwin, K. Akhtar, A. Alvand, A. J. Price, and J. L. Rees, "Validation of the updated ArthroS simulator: face and construct validity of a passive haptic virtual reality simulator with novel performance metrics," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 25, no. 2, pp. 616–625, 2017.

[22] M. Karahan, G. M. M. J. Kerkhoffs, P. Randelli, and G. J. Tuijthof, *Effective Training of Arthroscopic Skills*. Springer Berlin Heidelberg, 2015.

[23] L. H. Poniatowski, J. S. Wolf Jr, S. Y. Nakada, T. E. Reihsen, F. Sainfort, and R. M. Sweet, "Validity and acceptability of a high-fidelity physical simulation model for training of laparoscopic pyeloplasty," *Journal of Endourology*, vol. 28, no. 4, pp. 393–398, 2014.

[24] M. A. Maricic, M. M. Bailez, and S. P. Rodriguez, "Validation of an inanimate low cost model for training minimal invasive surgery (MIS) of esophageal atresia with tracheoe-

sophageal fistula (AE/TEF) repair," *Journal of Pediatric Surgery*, vol. 51, no. 9, pp. 1429–1435, 2016.

[25] S. Rodrigues, T. Horeman, P. Sam, J. Dankelman, J. van den Dobbelsteen, and F.-W. Jansen, "Influence of visual force feedback on tissue handling in minimally invasive surgery," *British Journal of Surgery*, vol. 101, no. 13, pp. 1766–1773, 2014.

[26] M. Uhrich, R. Underwood, J. Standeven, N. Soper, and J. Engsberg, "Assessment of fatigue, monitor placement, and surgical experience during simulated laparoscopic surgery," *Surgical Endoscopy*, vol. 16, no. 4, pp. 635–639, 2002.

[27] G. Islam, K. Kahol, B. Li, M. Smith, and V. L. Patel, "Affordable, web-based surgical skill training and evaluation tool," *Journal of Biomedical Informatics*, vol. 59, pp. 102–114, 2016.

[28] L. S. Feldman, S. E. Hagarty, G. Ghitulescu, D. Stanbridge, and G. M. Fried, "Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents," *Journal of the American College of Surgeons*, vol. 198, no. 1, pp. 105–110, 2004.

[29] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, 2016.

[30] J. E. Bennett, J. V. Carter, C. J. Foy, C. D. Roddb, and R. J. Longmanc, "Unsupervised use of laparoscopic box trainers in a cohort of UK core surgical trainees," *Journal of Surgical Simulation*, vol. 3, pp. 32–42, 2016.

[31] R. A. Pedowitz, G. T. Nicandri, R. L. Angelo, R. K. Ryu, and A. G. Gallagher, "Objective assessment of knot-tying proficiency with the fundamentals of arthroscopic surgery training program workstation and knot tester," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 31, no. 10, pp. 1872–1879, 2015.

[32] S. Srivastava, P. L. Youngblood, C. Rawn, S. Hariri, W. Heinrichs, and A. L. Ladd, "Initial evaluation of a shoulder arthroscopy simulator: establishing construct validity," *Journal of Shoulder and Elbow Surgery*, vol. 13, no. 2, pp. 196–205, 2004.

[33] R. J. Koehler, S. Amsdell, E. A. Arendt, L. J. Bisson, J. P. Bramen, *et al.*, "The arthroscopic surgical skill evaluation tool (ASSET)," *The American Journal of Sports Medicine*, vol. 41, no. 6, pp. 1229–1237, 2013.

[34] A. C. Hoyle, C. Whelton, R. Umaar, and L. Funk, "Validation of a global rating scale for shoulder arthroscopy: a pilot study," *Shoulder & Elbow*, vol. 4, no. 1, pp. 16–21, 2012.

[35] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 6, pp. 981–991, 2015.

[36] X. Sun, S. Byrns, I. Cheng, B. Zheng, and A. Basu, "Smart sensor-based motion detection system for hand movement training in open surgery," *Journal of Medical Systems*, vol. 41, no. 2, pp. 1–13, 2017.

[37] A. L. Trejos, R. V. Patel, R. A. Malthaner, and C. M. Schlachta, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surgical Endoscopy*, vol. 28, no. 7, pp. 2106–2119, 2014.

[38] A. McCarthy, P. Harley, and R. Smallwood, "Virtual arthroscopy training: do the "virtual skills" developed match the real skills required?" *Studies in Health Technology and Informatics*, vol. 62, pp. 221–227, 1998.

[39] K. P. Sherman, J. W. Ward, D. P. Wills, V. J. Sherman, and A. M. Mohsen, "Surgical trainee assessment using a VE knee arthroscopy training system (VE-KATS): experimental results." *Studies in Health Technology and Informatics*, vol. 81, pp. 465–470, 2000.

[40] R. A. Pedowitz, J. Esch, and S. Snyder, "Evaluation of a virtual reality simulator for arthroscopy skills development," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 18, no. 6, pp. 1–6, 2002.

[41] A. H. Gomoll, R. V. O'Toole, J. Czarnecki, and J. J. Warner, "Surgical experience correlates with performance on a virtual reality simulator for shoulder arthroscopy," *The American Journal of Sports Medicine*, vol. 35, no. 6, pp. 883–888, 2007.

[42] A. H. Gomoll, G. Pappas, B. Forsythe, and J. J. Warner, "Individual skill progression on a virtual reality simulator for shoulder arthroscopy a 3-year follow-up study," *The American Journal of Sports Medicine*, vol. 36, no. 6, pp. 1139–1142, 2008.

[43] G. Chami, J. W. Ward, R. Phillips, and K. P. Sherman, "Haptic feedback can provide an objective assessment of arthroscopic skills," *Clinical Orthopaedics and Related Research*, vol. 466, no. 4, pp. 963–968, 2008.

[44] Y. Tashiro, H. Miura, Y. Nakanishi, K. Okazaki, and Y. Iwamoto, "Evaluation of skills in arthroscopic training based on trajectory and force data," *Clinical Orthopaedics and Related Research*, vol. 467, no. 2, pp. 546–552, 2009.

[45] Y. Kassahun, B. Yu, A. T. Tibebu, D. Stoyanov, S. Giannarou, J. H. Metzen, and E. Vander Poorten, "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 4, pp. 553–568, 2016.

[46] S. Smith, A. Wan, N. Taffinder, S. Read, R. Emery, and A. Darzi, "Early experience and validation work with procedicus VA–the Prosolvia virtual reality shoulder arthroscopy trainer." *Studies in Health Technology and Informatics*, vol. 62, pp. 337–343, 1998.

[47] J. P. Bliss, H. S. Hanner-Bailey, and M. W. Scerbo, "Determining the efficacy of an immersive trainer for arthroscopy skills," *Studies in Health Technology and Informatics*, vol. 111, pp. 54–56, 2005.

[48] A. L. Trejos, "A sensorized instrument for minimally invasive surgery for the measurement of forces during training and surgery: development and applications," PhD Thesis, Dept. of Electrical and Computer Engineering, University of Western Ontario, 2012.

[49] T. P. Grantcharov, L. Bardram, P. Funch-Jensen, and J. Rosenberg, "Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills," *The American Journal of Surgery*, vol. 185, no. 2, pp. 146–149, 2003.

[50] A. Alvand, T. Khan, S. Al-Ali, W. Jackson, A. Price, and J. Rees, "Simple visual parameters for objective assessment of arthroscopic skill," *The Journal of Bone & Joint Surgery*, vol. 94, no. 13, pp. e97(1–7), 2012.

[51] F. Cavallo, G. Megali, S. Sinigaglia, O. Tonet, and P. Dario, "A biomedical analysis of a surgeon's gesture in a laparoscopic virtual scenario," *Medicine Meets Virtual Reality, Studies in Health Technology and Informatics*, vol. 119, pp. 79–84, Long Beach, CA, USA, January 24–26, 2006.

[52] N. Hogan and D. Sternad, "Sensitivity of smoothness measures to movement duration, amplitude, and arrests," *Journal of Motor Behavior*, vol. 41, no. 6, pp. 529–534, 2009.

[53] G. J. Tuijthof, T. Horeman, M. U. Schafroth, L. Blankevoort, and G. M. Kerkhoffs, "Probing forces of menisci: what levels are safe for arthroscopic surgery," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 19, no. 2, pp. 248–254, 2011.

[54] A. L. Trejos, R. V. Patel, M. D. Naish, R. Malthaner, and C. Schlachta, "The application of force sensing to skills assessment in minimally invasive surgery," in *IEEE International Conference on Robotics and Automation*.   Karlsruhe, Germany, May 6-10, 2013, pp. 4370–4375.

[55] N. Stylopoulos and K. G. Vosburgh, "Assessing technical skill in surgery and endoscopy: a set of metrics and an algorithm (C-PASS) to assess skills in surgical and endoscopic procedures," *Surgical Innovation*, vol. 14, no. 2, pp. 113–121, 2007.

[56] S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, R. Bardsley, and S. Dawson, "Surgical training system for laparoscopic procedures," US Patent App. 10/797,874, March 10, 2004.

[57] R. Berguer, "Surgical technology and the ergonomics of laparoscopic instruments," *Surgical Endoscopy*, vol. 12, no. 5, pp. 458–462, 1998.

[58] G. Lee, T. Lee, D. Dexter, R. Klein, and A. Park, "Methodological infrastructure in surgical ergonomics: a review of tasks, models, and measurement systems," *Surgical Innovation*, vol. 14, no. 3, pp. 153–167, 2007.

[59] A. Cuschieri, "Whither minimal access surgery: tribulations and expectations," *The American Journal of Surgery*, vol. 169, no. 1, pp. 9–19, 1995.

[60] B. Jonsson, "Quantitative electromyographic evaluation of muscular load during work." *Scandinavian Journal of Rehabilitation Medicine. Supplement*, vol. 6, pp. 69–74, 1978.

[61] S. Maithel, L. Villegas, N. Stylopoulos, S. Dawson, and D. Jones, "Simulated laparoscopy using a head-mounted display vs traditional video monitor: an assessment of performance and muscle fatigue," *Surgical Endoscopy and Other Interventional Techniques*, vol. 19, no. 3, pp. 406–411, 2005.

[62] U. Matern, G. Kuttler, C. Giebmeyer, P. Waller, and M. Faist, "Ergonomic aspects of five different types of laparoscopic instrument handles under dynamic conditions with respect to specific laparoscopic tasks: an electromyographic-based study," *Surgical Endoscopy and Other Interventional Techniques*, vol. 18, no. 8, pp. 1231–1241, 2004.

[63] R. Sancibrian, M. C. Gutierrez-Diez, C. Torre-Ferrero, M. A. Benito-Gonzalez, C. Redondo-Figuero, and J. C. Manuel-Palazuelos, "Design and evaluation of a new ergonomic handle for instruments in minimally invasive surgery," *Journal of Surgical Research*, vol. 188, no. 1, pp. 88–99, 2014.

[64] D. Büchel, R. Mrvik, B. Hallabrin, and U. Matern, "Ergonomics of disposable handles for minimally invasive surgery," *Surgical Endoscopy*, vol. 24, no. 5, pp. 992–1004, 2010.

[65] N. Hubert, M. Gilles, K. Desbrosses, J. Meyer, J. Felblinger, and J. Hubert, "Ergonomic assessment of the surgeon's physical workload during standard and robotic assisted laparoscopic procedures," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 9, no. 2, pp. 142–147, 2013.

[66] I. Suh, M. Mukherjee, D. Oleynikov, and K.-C. Siu, "Training program for fundamental surgical skill in robotic laparoscopic surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 7, no. 3, pp. 327–333, 2011.

[67] K. Narazaki, D. Oleynikov, and N. Stergiou, "Robotic surgery training and performance," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 1, pp. 96–103, 2006.

[68] T. N. Judkins, D. Oleynikov, K. Narazaki, and N. Stergiou, "Robotic surgery and training: electromyographic correlates of robotic laparoscopic training," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 5, pp. 824–829, 2006.

[69] A. Shafti, F. Andorno, N. Marchese, S. Arolfo, A. Aydin, O. Elhage, Y. Noh, H. Wurdemann, A. Arezzo, P. Dasgupta, *et al.*, "Comfort and learnability assessment of a new soft robotic manipulator for minimally invasive surgery," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, Italy, Aug 25–29, 2015, pp. 4861–4864.

[70] N. Quick, J. Gillette, R. Shapiro, G. Adrales, D. Gerlach, and A. Park, "The effect of using laparoscopic instruments on muscle activation patterns during minimally invasive surgical training procedures," *Surgical Endoscopy*, vol. 17, no. 3, pp. 462–465, 2003.

[71] I. H. Suh, C. A. LaGrange, D. Oleynikov, and K.-C. Siu, "Evaluating robotic surgical skills performance under distractive environment using objective and subjective measures," *Surgical Innovation*, vol. 23, no. 1, pp. 78–89, 2016.

[72] A. Keshavarz Panahi and S. Cho, "Prediction of muscle fatigue during minimally invasive surgery using recurrence quantification analysis," *Minimally Invasive Surgery*, vol. 2016: 5624630, 2016.

[73] C. Sewell, "Automatic performance evaluation in surgical simulation," PhD Thesis, Stanford University, 2007.

[74] J. Rosen, B. Hannaford, C. G. Richards, and M. N. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for

evaluating surgical skills," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 579–591, 2001.

[75] J. Rosen, J. D. Brown, L. Chang, M. N. Sinanan, and B. Hannaford, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 399–413, 2006.

[76] M. J. Fard, S. Ameri, R. B. Chinnam, A. K. Pandya, M. D. Klein, and R. D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," vol. 1, no. 18, San Francisco, USA, Oct 19–21, 2016.

[77] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics), 2nd ed.* Springer, New York, 2007.

[78] M. Zirkle, M. A. Taplin, R. Anthony, and A. Dubrowski, "Objective assessment of temporal bone drilling skills," *Annals of Otology, Rhinology & Laryngology*, vol. 116, no. 11, pp. 793–798, 2007.

[79] M. K. Chmarra, S. Klein, J. C. de Winter, F.-W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surgical Endoscopy*, vol. 24, no. 5, pp. 1031–1039, 2010.

[80] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye metrics as an objective assessment of surgical skill," *Annals of Surgery*, vol. 252, no. 1, pp. 177–182, 2010.

[81] D. Stefanidis, N. Sevdalis, J. Paige, B. Zevin, R. Aggarwal, T. Grantcharov, D. B. Jones, Association for Surgical Education Simulation Committee, *et al.*, "Simulation in surgery: what's needed next?" *Annals of Surgery*, vol. 261, no. 5, pp. 846–853, 2015.

Chapter 3

Analysis of Energy-based Metrics for Laparoscopic Skills Assessment

3.1   Introduction

Minimally invasive surgery (MIS) promises certain advantages for patients such as lower pain levels, reduced blood loss and better cosmesis. However, it demands the manipulation of long instruments in difficult positions with limited degrees of freedom. In addition, dealing with the fulcrum effect and a different sense of force compared to conventional open surgery are among the complications of MIS. Consequently, MIS tasks must be practiced repeatedly to achieve mastery prior to performing them in the operating room. Surgical simulators are helpful in providing the opportunity to practice in a safe environment with fewer time constraints. These simulators will be more efficient when equipped with objective assessment methods [1]. Appropriate assessment methods are essential to quantify the level of expertise of trainees, to provide them with feedback about their performance, to allow independent training, and to certify the proficiency level of a resident or a surgeon before operating on a patient [2]. Section 3.1.1 provides a review of the commonly used performance metrics.

3.1.1 Review of Performance Metrics

The metrics proposed for surgical skills assessment can be classified in different ways. Mc-Crory, *et al.* [2] considered three main groups of metrics: 1. the group related to patient safety—e.g., force magnitude, 2. the group related to the success of the procedure—e.g., task outcome metrics, and 3. the group that deals with efficiency—e.g., path length [3,4]. However, the ability of the provided examples to represent safety, success, or efficiency depends on the task and should be further investigated. A more detailed classification was performed in [5] by dividing metrics into temporal, task outcome, motion-based, force-based, and combined metrics, as described bellow:

Temporal metrics are among the most commonly used metrics for assessing the performance of trainees. *Task completion time* has been shown to best discriminate between novice and expert subjects and has been used in [6–9] for skills assessment. Inclusion of this metric in skills assessment enhances discrimination between various levels of expertise [10]. Although experts are faster than novices in performing surgical tasks, a short task completion time does not necessarily mean a superior performance. Consequently, in most studies, *time* has been combined with other performance metrics to provide a more complete assessment. Other examples of temporal metrics include the number of repetitions required to complete a task successfully [3] and hesitation. Hesitation can be determined by the time intervals between subtasks [5] or by short intervals in which the instrument does not move [11].

Task outcome metrics are developed based on the outcome of task completion regardless of how it is performed. These metrics are defined based on the proposed tasks and are usually evaluated by an external observer. Examples of outcome metrics include the number of instrument collisions, the number of successful identifications of landmarks or structures, and the score assigned to a performance assessment [12]. The number of instrument collisions can also be an indication of safety and can be determined by identifying the number of times a certain force threshold has been exceeded.

Motion-based metrics are defined using instrument and hand position to extract biomechanical parameters [5]. Several metrics are extracted from motion information such as *path length*—the distance that the instrument travels, *speed*—the first derivative of position, *acceleration*—the second derivative of position, and *jerk*—the third derivative of position, which

represents smoothness of motion. Different parameters can be considered to further characterize *speed* and *acceleration* metrics. These include the mean, the peak, the magnitude, and the normalized speed/acceleration [5–7, 13].

Force-based metrics are developed based on force profile of performance [8, 9]. It is essential for trainees to learn to apply sufficient force when needed and to be gentle enough with the tissues. Applying too much force may result in tissue damage; however, applying less force than required may lead to ineffective performance. Force-based metrics include average force, maximum force, the integral of the force, force range, force direction, derivatives of the force, and smoothness of the applied force [3, 5]. These metrics can also be an indication of performance safety and efficiency.

Combined metrics are established by combining different metrics together. These metrics consider multiple characteristics of the performance in one unified metric. Different combined metrics are defined and studied in [5, 8, 14] [15, 16].

The following section describes the development of a new metric for laparoscopic skills assessment.

3.1.2   Using Energy Expenditure for Metric Development

Guthrie's definition of *skill* [17] recognizes *maximum certainty*, *minimum time* and *minimum energy* as features of a skilled performance. *Certainty* has been investigated in qualitative studies [18, 19] and will not be explored here. *Time* has been extensively used for skills assessment but this metric cannot completely represent the level of proficiency of a trainee. *Energy expenditure* has been also considered as a feature of skilled performance in other literature [20, 21]. Elliot, *et al.* [22, 23] indicate that through practice, energy expenditure optimization can be achieved, as well as optimization in accuracy and speed of performance. To reach the minimum energy expenditure, removal of unnecessary and undesirable movements is required [24].

Energy expenditure can be quantified by measuring the heart beat, body temperature, and the rate of oxygen–carbon dioxide exchange [25, 26]. However, for MIS tasks, it is possible to measure force and position information related to energy expenditure at the tip of the instrument. According to Sparrow, *et al.* [27] the human body tends to minimize metabolic energy expenditure in relation to the task to be performed, the environment in which the task is con-

ducted, and the constraints imposed on the performer's action. An important part of energy expenditure minimization relates to receiving appropriate sensory information and adapting motions based on the received information. Unfortunately, the sensory information received by the MIS performer is reduced due to the lack of direct visual and force feedback. Limited degrees of freedom in motion is another constraint that makes energy expenditure optimization challenging.

3.1.3 Objectives

Although various metrics have been proposed and used for MIS skills assessment, as reviewed in Section 3.1.1, a metric that can objectively determine the detailed level of expertise of subjects is still lacking. The goal of this study is to enhance MIS skills assessment by developing objective metrics based on energy expenditure and to validate these metrics.

3.2 Metric Development

Previous studies [5, 7, 9, 28] reported a difference in the velocity and the applied force profiles of novices and experts. Thus, experts use different techniques or movements, coordinating both hands, to perform MIS tasks. The differences observed in velocity and force when experts perform surgical tasks might be due to a different amount of energy expenditure. This information can be incorporated into an energy formula. In this study, the types of energy that can be quantified using force and position information are considered. Energy expenditure, in the form of mechanical energy and work, has been used in human movement studies [29, 30]. However, these forms of energy expenditure are not currently used in surgical skills assessment.

The proposed metrics in this study consist of three components, which are defined based on potential energy, kinetic energy, and work. These components are combined to optimize the ability of the proposed metrics to discriminate various skills levels. Four levels of experience are considered in this study with two levels in each of the novice and expert groups. In this study, it is assumed that the level of experience of the subjects correlates with their level of expertise. However, in some cases, this assumption may not hold for all the subjects. The basic components and the combined method are described in Sections 3.2.1 and 3.2.2, respectively.

3.2.1   Basic Components

Gravitational potential energy is the form of energy that is generated or consumed due to changes in the position of an object in the gravitational field, as follows:

$$E_\text{P} = mgh \ \ [\text{J}],$$

(3.1)

where m is the mass of the surgical instrument [kg], g is the gravitational acceleration [9.8 m/s^2], and h is the height of the tip of the instrument [m]. In laparoscopic skills assessment, the sum of the absolute changes in potential energy is considered:

$$\text{Potential-based component} = \sum_{i=\text{start}}^{\text{end}} |\Delta E_{\text{P}_i}| \ \ [\text{J}],$$

(3.2)

where i represents the index of the sampling time, which ranges from the start to the end of the task.

Kinetic energy is another form of mechanical energy. Here, kinetic energy due to translational velocity is considered, as follows:

$$E_\text{k} = \frac{1}{2}mv^2 \ \ [\text{J}],$$

(3.3)

where v is the translational velocity of the tip of the instrument [m/s]. Similar to potential energy, the sum of the absolute changes in kinetic energy is considered for laparoscopic skills assessment, as follows:

$$\text{Kinetic-based component} = \sum_{i=\text{start}}^{\text{end}} |\Delta E_{\text{K}_i}| \ \ [\text{J}].$$

(3.4)

The kinetic-based and potential-based components were calculated assuming a constant mass for the instruments. However, due to interaction of the instruments with their environment, there may be additional mass carried by the instruments, which increases the total value of these components. Consequently, the potential-based and kinetic-based components represent the minimum amount of change in the potential and kinetic energies.

The amount of *work* done on the surgical simulator through the surgical instruments is another component of the energy-based metric proposed in this study. The total amount of the absolute work during task completion forms this component:

$$\text{Work-based component} = \sum_{i=\text{start}}^{\text{end}} |W_i| \quad [\text{J}], \tag{3.5}$$

where W_i represents the amount of work in each sampling time and can be calculated according to work formula:

$$W = F \cdot d \quad [\text{J}], \tag{3.6}$$

where F and d are vectors of the applied force [N] and displacement [m] at the tip of the instrument.

3.2.2   Combined Energy-based Metrics

Each of the above components represents a part of the change in energy expenditure. In order to study the relationship between the level of experience (LOE) and energy expenditure, *combined metrics* were defined to estimate the minimum energy expenditure. These combined energy-based metrics are the weighted sum of all of the components for the left and right hands.

$$\begin{aligned}
\text{Combined metric} = \ & \alpha_{W_L} \times W_L^{\beta_{W_L}} + \alpha_{W_R} \times W_R^{\beta_{W_R}} + \\
& \alpha_{P_L} \times P_L^{\beta_{P_L}} + \alpha_{P_R} \times P_R^{\beta_{P_R}} + \\
& \alpha_{K_L} \times K_L^{\beta_{K_L}} + \alpha_{K_R} \times K_R^{\beta_{K_R}},
\end{aligned} \tag{3.7}$$

where W_L and W_R represent the work-based components for the left and right hands, P_L and P_R represent the potential-based components for the left and right hands, and K_L and K_R represent the kinetic-based components for the left and right hands.

In this formula, the α values are the corresponding coefficients for each component. Higher values of α mean a higher contribution of the related component. β values are the exponents of each component, where β equal to 1 means a linear relationship between the *combined metric* and the corresponding component. Incorporating exponents into this formula allows non-linear

relationships between LOE and energy-based components to be identified. A positive value of β corresponds to a direct relationship and a negative value reflects an inverse relationship. For instance, α_{W_L} is the coefficient of the work-based metric for the left hand in the combined metric and β_{W_L} is the exponent of this basic component. Due to differences in task requirements, particular combinations of the basic components should be established for each task, such that the difference between the energy expenditure of various levels of experience is maximized.

Optimization methods were investigated to find the coefficients (α) and exponents (β) that maximize the Spearman's rho correlation of the combined metric (Eq. 3.7) with LOE.

Two approaches were explored for developing combined metrics as follows:

3.2.2.1   One-step Combined Metric

In the one-step combined metric, a set of coefficients and exponents is obtained through optimization of the correlation of the combined metric with the four LOE considered in this study.

3.2.2.2   Two-step Combined Metric

The four levels of experience considered in this study can be classified in two main groups of novices (Levels 1 and 2) and experts (Levels 3 and 4). In this approach, the discrimination between different levels of experience is accomplished in two steps. The first step consists of recognizing the main LOE (novice or expert). The optimization is performed for this step to find the set of coefficients and exponents that maximize the correlation with the two main levels of experience. In the second step, different coefficients and exponents are determined to distinguish between LOE 1 and 2 (novice) or between LOE 3 and 4 (expert). Fig. 3.1 outlines this method.

Two methods of optimization were investigated to determine the appropriate set of coefficients and exponents: the Genetic Algorithm (GA) function of the global optimization toolbox and the *fmincon* function of MATLAB (The Mathworks, Inc., Natick, MA). Since each basic component might have a different range, as can be seen in Section 3.4.1.1, a normalization process was implemented before combining these components. This was accomplished by dividing each component by the range of variation of that component. The upper and lower limits for the coefficients were set to +1 and -1 and for the exponents were set to +3 and -3.
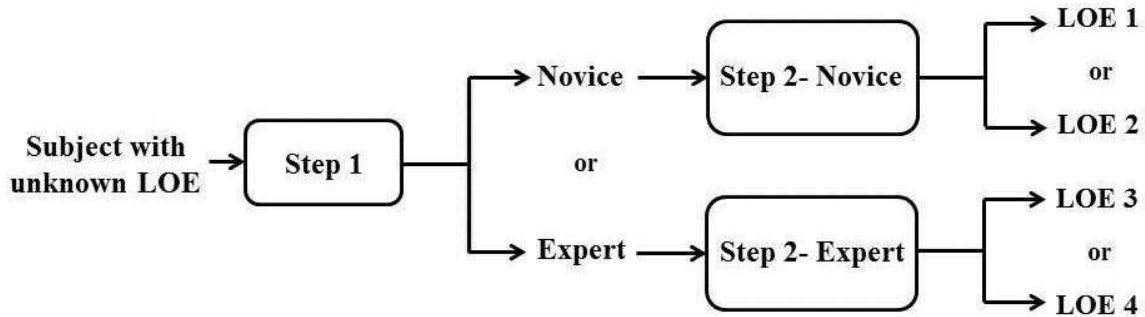
Figure 3.1: Diagram of the two-step combined metric.

3.3 Materials and Methods

The proposed metrics were investigated for two laparoscopic tasks. In this section, the design of the experiment for the data collection and the data processing are described.

3.3.1 Experimental Design

The setup for this experiment is composed of a standard laparoscopic training box. Inside the training box, an ABS plastic frame was placed to hold a soft tissue model made of silicone and foam. A layer of soft rubber was placed on top of the soft tissue model to mimic the skin layers.

For this experiment, 30 subjects with different levels of experience in laparoscopy were recruited. The participants in the study were divided into four levels of experience as follows: LOE 1— with no medical background ($n = 6$), LOE 2— medical students with no surgical training, surgeons with no MIS experience, and postgraduate year (PGY) 2–3 who had no exposure to MIS ($n = 11$), LOE 3— PGY 4–5 and trained fellows ($n = 7$), and LOE 4— expert surgeons ($n = 6$). This division constitutes the reference for evaluating the proposed metrics. All subjects were right-hand dominant.

Each subject was asked to perform a suturing and a knot-tying task in each trial. All participants were asked to repeat these tasks in four trials. In the suturing task, participants were asked to pass a needle through both sides of an incision in the simulated skin. The exact starting point for the suturing task was left to the participant's discretion. The knot-tying task consisted of one double knot and two single knots (Fig. 4.2). The skills required to perform these tasks
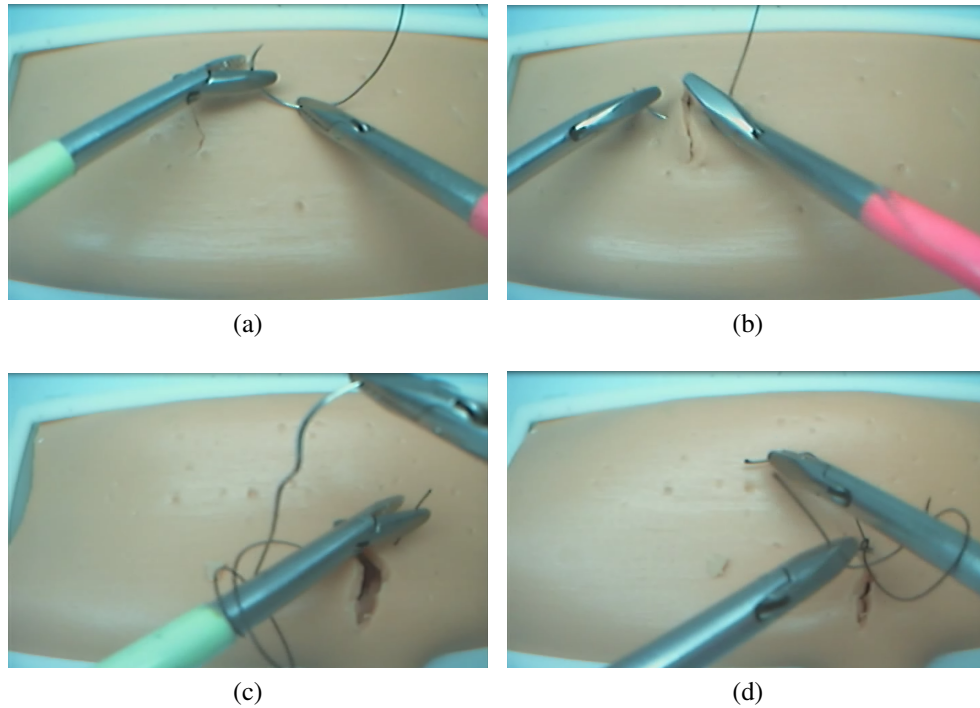
Figure 3.2: The laparoscopic tasks. (a), (b) Passing the needle through the tissue for the suturing task. A double knot (c) and two single knots (d) constitute the knot-tying task.

included proper handling and placement of the needle, and controlling the instrument tip and the suture.

The Sensorized Instrument-based Minimally Invasive Surgery (SIMIS) [31] System was used to perform the tasks [32]. Two SIMIS instruments were used to record forces applied perpendicular to the shaft of the instrument, in two Cartesian directions, and position information of the tip of the instrument in 6 degrees of freedom. The mass of each instrument is 170 g.

3.3.2 Data Processing

Before calculating the proposed metrics, the recorded data for each subject was segmented to isolate the data of the suturing and knot-tying tasks. Time frames for the start and end of the task were identified by video recordings during the experiment. In addition, the intervals where there was an interruption in task completion were determined and removed to focus on the movements related to the task. The recorded data was then low-pass filtered with a second-order Butterworth filter at the cut-off frequency of 40 Hz. This cut-off frequency was

selected by investigating the power spectrum of the force and position signals to ensure that no significant information was lost by filtering and that the data was not affected by high-frequency noise. To identify outliers, the boxplot function of MATLAB was used. The data points outside $[q_1 - w(q_2 - q_1), q_2 + w(q_2 - q_1)]$ were recognized as outliers, where q_1 and q_2 represent 25th and 75th percentiles and w represents the whisker length which was set to 1.72 (equivalent to 3 standard deviations of the data for each LOE). Each of these outliers was then investigated by watching the corresponding video to find the cause of the outlier. For the outlier points that dealt with a reasonable cause, the data point was replaced with the maximum nonoutlier value for the corresponding LOE. The accepted causes were sliding the skin layer out of the tissue model frame or breaking the suture. Among 120 trials executed, 8 trials were recognized for including outlier data for the suturing task and 11 trials were recognized for including outlier data for the knot-tying task.

3.4   Results and Validation

Among the four trials executed in this study for each subject, the data from the first, second, and fourth trials were considered as the training data set. These data were used to determine the coefficients and exponents of the combined metrics. The basic components and the resulting combined metrics obtained from the training data set, as presented in Section 3.4.1, were used to determine the margins of the four levels of experience. Based on these margins, the LOE of a subject with an unknown LOE can be determined. The third trial was considered as the test data set and was utilized in order to validate the proposed metrics. Since the test data consisted of the data that was collected from all of the participants in their third trial, the distribution of LOE over the test data set was the same as the distribution of LOE over the training data. The validation results are presented in Section 3.4.2.

3.4.1   Results of Metric Development

3.4.1.1   Basic Components

The basic components were shown to be successful in discriminating novice and expert subjects for both tasks. This was demonstrated in a previous study considering subjects in two main

levels of novice and expert [32]. However, more detailed classification of subjects would be beneficial in guiding trainees and improving the learning quality. This detailed classification is investigated in the current study. The three basic components of the proposed metrics versus the four LOE for the left and right hands are shown in Fig. 3.3 (suturing task) and Fig. 3.4 (knot-tying task). Since the axial force was not measured in these experiments, the work-based component presented in this study represents the work performed perpendicular to the shaft of the instrument. The maximum correlation was obtained for the work-based component for both tasks. In order to assess the effect of additional mass on the potential-based and kinetic-based metrics caused by the interaction of the instruments with tissue, a sensitivity analysis was performed (Appendix A). The results showed that the possible additional mass does not affect the relationship between these components and the various LOEs. This is likely due to the additional mass having a similar effect on the performance of all the subjects.

The amount of each basic component decreased as the LOE increased for both tasks. The difference between the first two LOE and the last two LOE was significant, while the difference between LOE 1 and LOE 2 and also between LOE 3 and LOE 4 was limited. Consequently, discriminating between subjects with various levels of experience could not be accomplished completely by defining margins for those levels.

3.4.1.2   One-step Combined Metric

Comparing the results obtained from the two optimization methods (fmincon function and GA algorithm) showed that the resulting coefficients and exponents from the GA algorithm demonstrated higher correlations of the combined metrics with LOE. Therefore, the values obtained from the GA algorithm were used for both combined metrics in this study.

Fig. 3.5 shows the results of the one-step combined metric for the suturing and knot-tying tasks. By combining the basic components using the one-step combined metric, the correlation with LOE for the suturing task was improved from -0.459 (the maximum correlation that was obtained for the work-based component for the right hand) to -0.506. For the knot-tying task, the one-step combined metric resulted in an improvement of the correlation with LOE from -0.694 to -0.795. It may be observed in Fig. 3.5 that the overlap between various levels of experience for the one-step combined metric was relatively smaller than the overlap of the
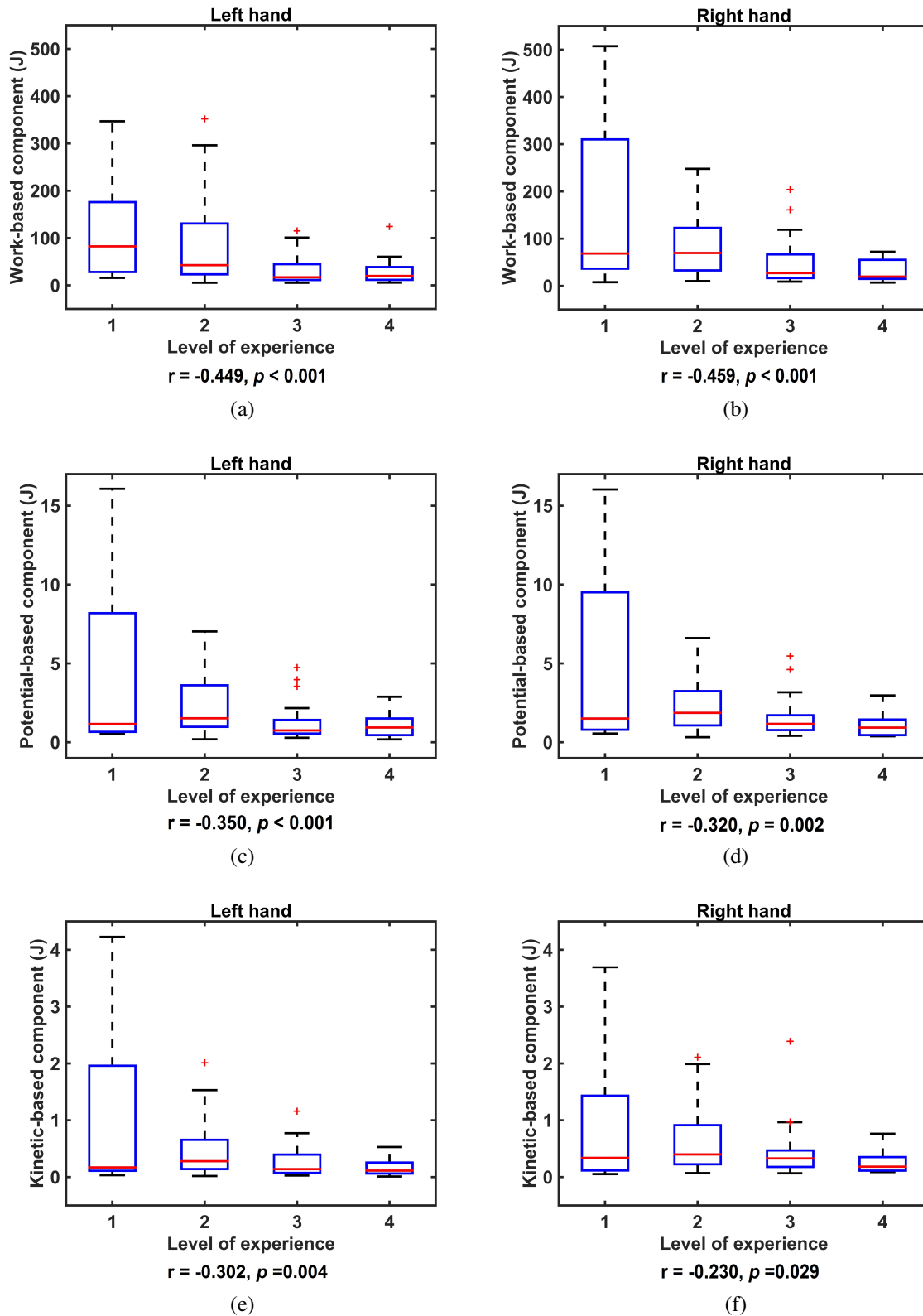
Figure 3.3: Basic components for the suturing task based on the training data set. The Spearman's rho correlation of each basic component with LOE (r) and the corresponding *p* value are shown below each sub-figure. All correlations are statistically significant. The red line in each box indicates the median.
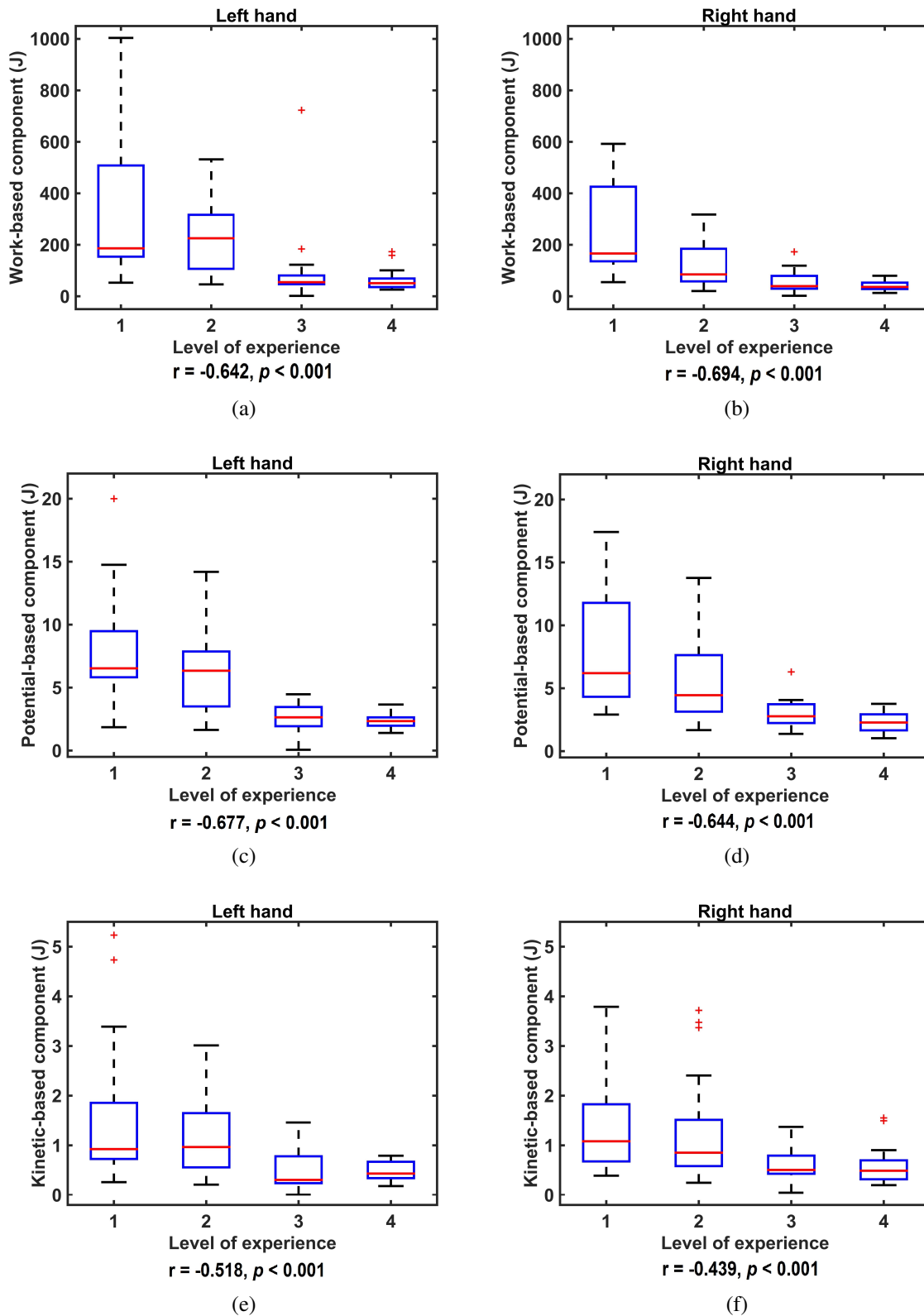
Figure 3.4: Basic components for the knot-tying task based on the training data set. The Spearman's rho correlation of each basic component with LOE (r) and the corresponding *p* value are shown below each sub-figure. All correlations are statistically significant.

(a)

(b)
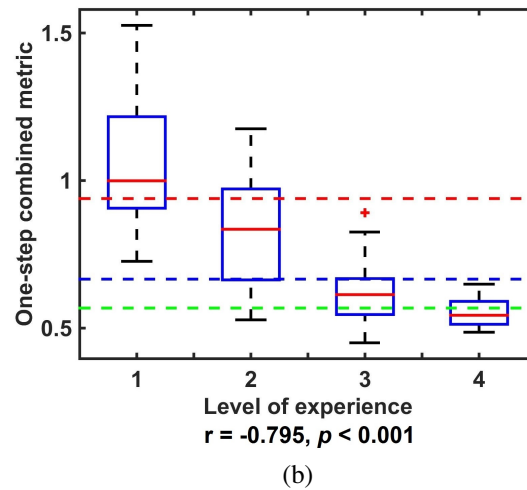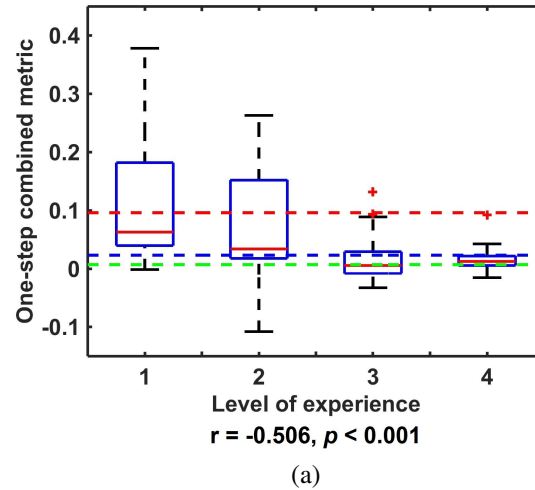
Figure 3.5: The one-step combined metric for (a) the suturing task and (b) the knot-tying task for the training data. Red, blue, and green dashed lines represent the lower margin of LOE 1, LOE 2, and LOE 3, respectively.

basic components (as seen in Figs. 3.3 and 3.4).

The margins of variation for each LOE were also derived from the results of the training data set for all of the metrics presented in this study. For instance, the lower margin of LOE 1, which was also the upper margin of LOE 2, was the average of the 25th percentile of LOE 1 subjects and the 75th percentile of the subjects in LOE 2 for each metric. These margins for the one-step combined metric are shown by dashed lines in Fig. 3.5. The area above the red dashed line indicates the area of variation for LOE 1, the area between the red and blue dashed lines relates to LOE 2, the area between the blue and green dashed lines defines LOE 3, and the area below the green dashed line indicates LOE 4.

(a)

r = -0.905, p < 0.001

(b)
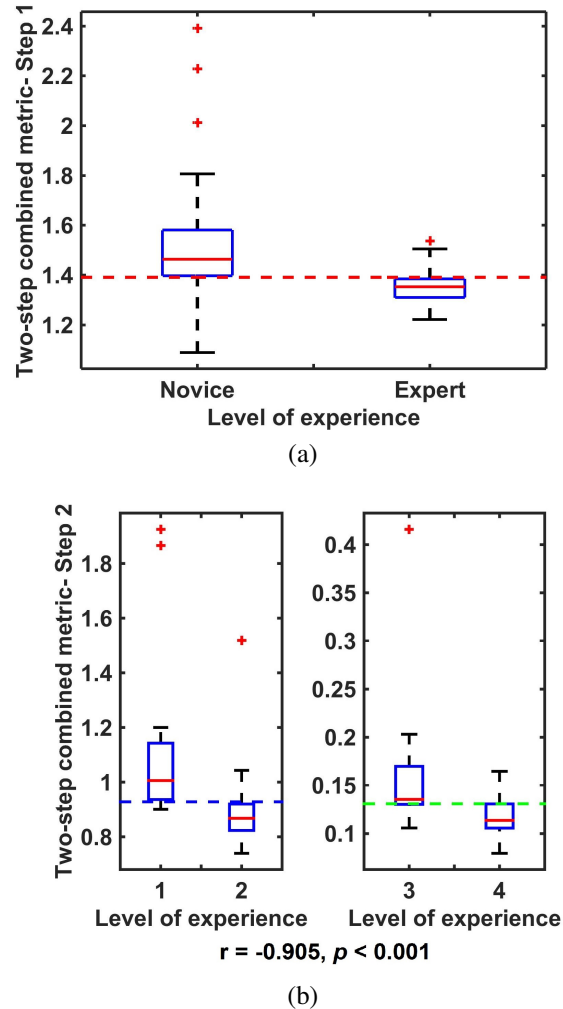
Figure 3.6: The two-step combined metric for the suturing task for the training data: (a) Step 1 for two main LOE and (b) Step 2 for detailed LOE for the novice and expert groups. Red, blue, and green dashed lines represent lower margins for the novice, LOE 1, and LOE 3, respectively.

3.4.1.3   Two-step Combined Metric

Figs. 3.6 and 3.7 show the results of the two-step combined metric for the suturing and knot-tying tasks. The correlation with LOE was calculated for this metric based on the results of the second step. For the suturing and knot-tying tasks, the correlations with LOE were -0.905, which are considerably higher than the corresponding values obtained from the individual components and the one-step combined metric.

The margins of variation for this metric were also calculated using the same method explained in Section 3.4.1.2. The red dashed lines in Figs. 3.6(a) and 3.7(a) show the margin

(a)

(b)
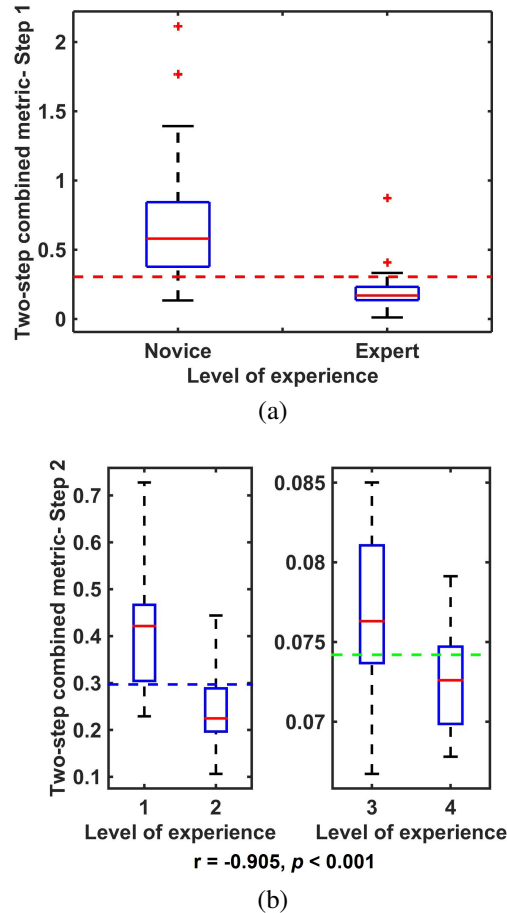
Figure 3.7: The two-step combined metric for the knot-tying task for the training data: (a) Step 1 for two main LOE and (b) Step 2 for detailed LOE for the novice and expert groups. Red, blue, and green dashed lines represent lower margins for the novice, LOE 1, and LOE 3, respectively.

for discriminating between experts and novices. The blue and green lines in Figs. 3.6(b) and 3.7(b) indicate the margins for differentiating between LOE 1–2 and between LOE 3–4.

3.4.2   Validation

In this part of the study, it was assumed that the LOE of the subjects were unknown. The margins extracted in Section 3.4.1 were used in MATLAB to determine the LOE of subjects based on the test data set. Afterwards, the determined levels of experience were compared to the true levels of experience to investigate the accuracy of each metric. Thus, the validation was performed blindly.

The results of the one-step combined metric for the test data set are shown in Fig. 3.8 along
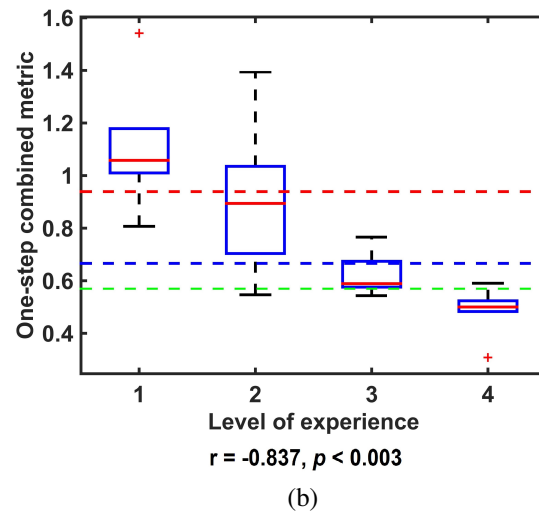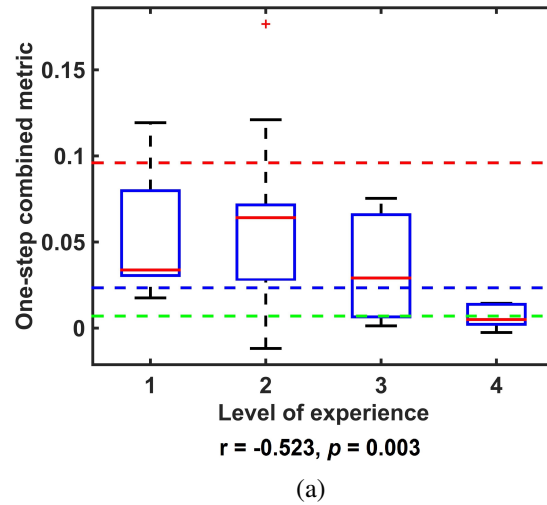
Figure 3.8: The one-step combined metric for the test data set for (a) the suturing task and (b) the knot-tying task. The margins specified in this figure are obtained from the training data set.

with the margins for each LOE. The two-step combined metric for the test data set related to the suturing and knot-tying tasks are presented in Figs. 3.9 and 3.10, respectively. Due to variations in the performance of the subjects in each trial, the margins obtained based on the training data set do not necessarily match the values that could be obtained from the test data set.

The accuracy, defined as the total number of correct identifications over the total number of subjects expressed as a percentage, is represented in Fig. 3.11 for all of the energy-based metrics. The basic components provided a maximum accuracy of 30% and 37%, for the su-

(a)

$r = -0.860, p < 0.001$

(b)
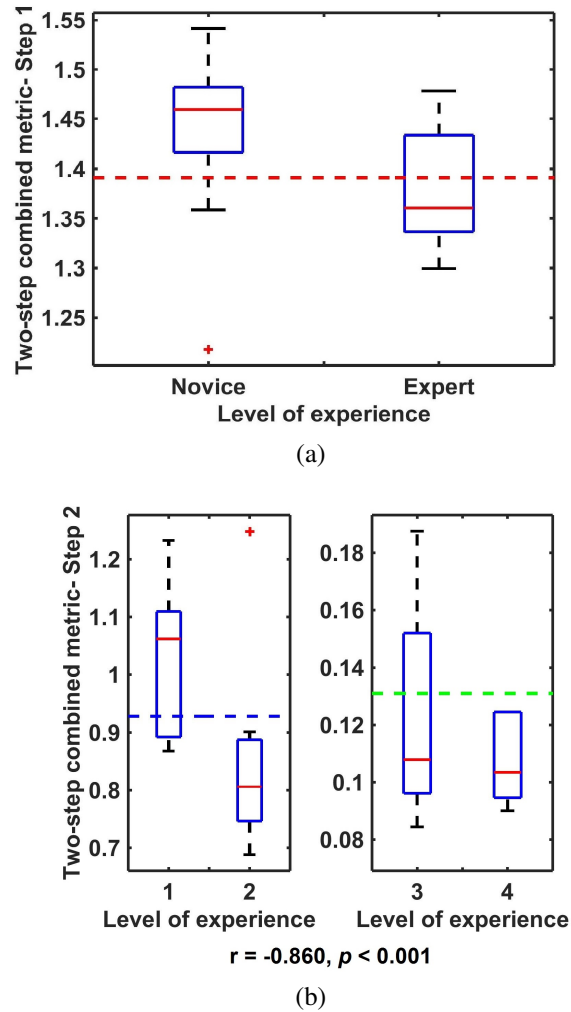
Figure 3.9: The two-step combined metric for the test data set for the suturing task: (a) Step 1 and (b) Step 2. The margins specified in this figure are obtained from the training data set.

turing and knot-tying tasks, respectively. Using the combined metrics, the discrimination is significantly improved compared to the basic components. The one-step combined metric can accurately identify the LOE of 14 and 18 subjects (47% and 60%) for the suturing and knot-tying tasks. Using the two-step combined metrics, the LOE of 20 subjects (67%) was properly recognized for both tasks. In addition, 29 subjects (97%) were identified within ±1 level of the correct LOE for the knot-tying task. For the suturing task, the number of participants identified within ±1 level of the correct LOE was 28 (93%) for both combined metrics.

These combined metrics were compared with two metrics commonly used in MIS skills assessment: path length and task completion time (Fig. 3.11). The margins of variation for
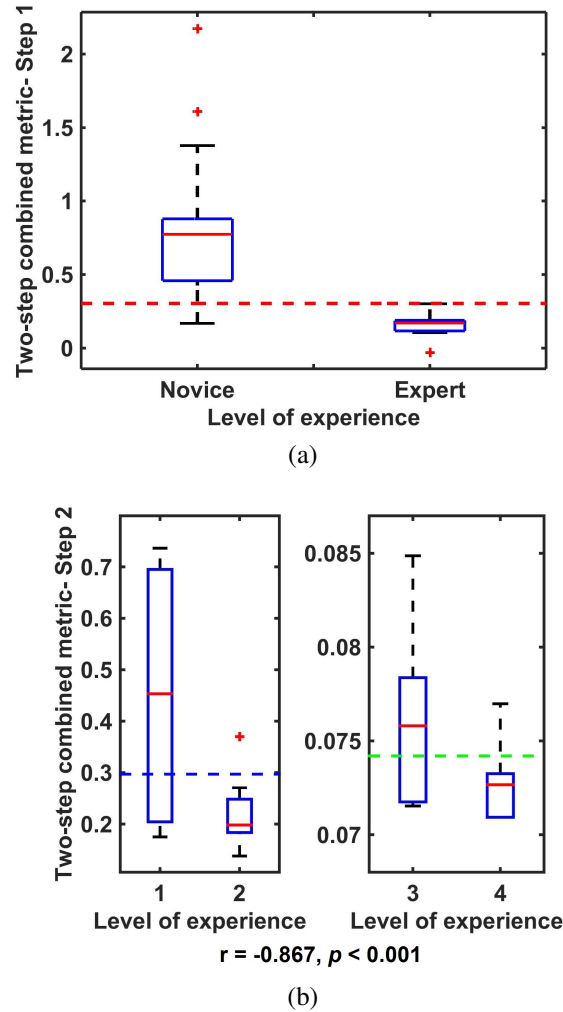
Figure 3.10: The two-step combined metric for the test data set for the knot-tying task: (a) Step 1 and (b) Step 2. The margins specified in this figure are obtained from the training data set.

these metrics were determined based on the training data and using the same method explained in Section 3.4.1.2. These margins were utilized for LOE determination of the subjects based on the test data. For the suturing task, path length of the right hand demonstrates better accuracy in identifying LOE than the basic and the one-step combined metrics. However, the two-step combined metric provided superior accuracy compared to the path length and time. For the knot-tying task, both one-step and two-step combined metrics provided more accurate identifications than path length and time.

The correlation with the LOE for the basic components, and for path length and time, calculated for the test data set, are shown in Tables 3.1 and 4.1, respectively. The maximum

amount of correlation among the path length and task completion time for the suturing task was obtained for time: -0.369 (Table 4.1). Among the basic components for the suturing task, the highest correlation was obtained for the work-based component for the right hand: -0.582 (Table. 3.1). The correlation for the one-step combined metric was -0.523 (Fig. 3.8 (a)). The highest correlation was obtained for the two-step combined metric, which was -0.860 (Fig. 3.9).

For the knot-tying task, the maximum correlation of the path length and the task completion time was obtained for time: -0.754 (Table 4.1). This correlation was smaller than the one for the work-based component (the maximum among the basic components), which was -0.791 (Table 3.1). This was also smaller than the correlations of the one-step and two-step combined metrics with LOE, which were -0.837 (Fig. 3.8 (b)) and -0.867 (Fig. 3.10), respectively.

Table 3.1: The correlation with LOE and the corresponding p values for the basic components applied to the test data set. The statistically significant correlations ($p < 0.05$) are displayed in bold.

Task	Basic component	Left hand	Right hand
Suturing	Potential-based component	$-0.292, p = 0.118$	$-0.260, p = 0.165$
	Kinetic-based component	$\mathbf{-0.367, p = 0.046}$	$-0.280, p = 0.133$
	Work-based component	$-0.326, p = 0.079$	$\mathbf{-0.582, p < 0.001}$
Knot-tying	Potential-based component	$\mathbf{-0.741, p < 0.001}$	$\mathbf{-0.713, p < 0.001}$
	Kinetic-based component	$\mathbf{-0.651, p < 0.001}$	$\mathbf{-0.388, p = 0.003}$
	Work-based component	$\mathbf{-0.791, p < 0.001}$	$\mathbf{-0.684, p < 0.001}$

Table 3.2: The correlation with LOE and the corresponding p values, for path length for the left and right hands (PL_L, PL_R), and time for the test data set. The statistically significant correlations ($p < 0.05$) are displayed in bold.

Task	PL_L	PL_R	Time
Suturing	-0.318, 0.087	-0.345, 0.062	**-0.369**, **0.045**
Knot-tying	**-0.747**, **<0.001**	**-0.745**, **<0.001**	**-0.754**, **<0.001**

3.5 Discussion

In this study, the accuracy in determining LOE and the Spearman's rho correlation were used to assess several proposed energy-based metrics and compare them with commonly used metrics for minimally invasive skills assessment.

The basic components (potential-based component, kinetic-based component, and work-based component) showed a decreasing trend as the LOE increased for both the suturing and knot-tying tasks. This indicates that as the LOE increases, subjects become more efficient in performing the task and expend less energy. In this study, the work-based component represented the amount of work that was produced perpendicular to the shaft of the instrument. Incorporating the axial force into the work-based metric can enhance the discriminatory capability of this component.

The combination of these components was analyzed in the one-step combined metric. Using this metric, accuracies of 47% and 60% in recognizing LOE were obtained for suturing and knot-tying tasks. However, due to overlaps between the one-step combined metric for the different levels of experience, it was not possible to locate margins between different LOE to allow for complete differentiation between all subjects. In terms of correlation with LOE, the one-step combined metric demonstrated smaller correlation with LOE than the basic component of work for the right hand for the test data of the suturing task. This reduced correlation is most likely due to the optimization weights of the combined metric based on the training data. These weights do not necessarily provide higher correlations for every set of data. This issue might be mitigated by using larger data sets for the optimization process and determin-
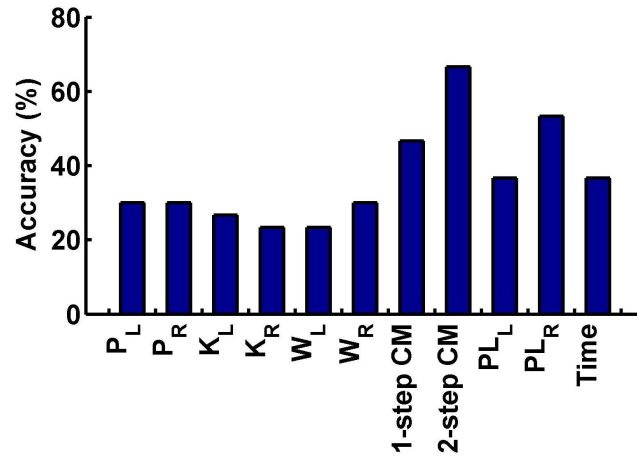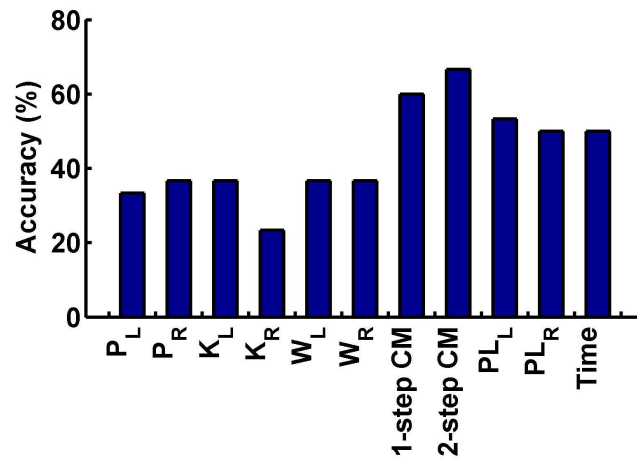
Figure 3.11: Accuracy of the energy-based metrics and path length for the left hand (PL_L), path length for the right hand (PL_R), and task completion time for the test data set. (a) The suturing task and (b) the knot-tying task.

ing a metric that can compensate for the variations in task executions. However, the one-step combined metric for the knot-tying task and the two-step combined metrics for both tasks also provided higher correlations with LOE than the basic components for the test data.

Decomposing the discrimination into two steps facilitated the determination of appropriate coefficients and exponents for each main LOE. This metric resulted in 67% accuracy for both the suturing and knot-tying tasks. In addition, higher amounts of correlation with LOE were obtained using the two-step combined metric for the training and test data sets compared to other energy-based metrics and the time and path length. The superior performance of the two-

step combined metric is due to using different combinations of the basic components for each main LOE (novice and expert). The pattern of expending different forms of energy (potential energy, kinetic energy, and work) among novice subjects can be different from that of expert subjects. Another set of weights for combining these components is required to demonstrate the major differences between the two main levels of experience.

Among the four trials that were performed by each participant, the third trial was used to evaluate the performance of the proposed metrics. This trial was chosen because it represented a trial in which the subjects were already familiar with the setup and the tasks, but had not yet performed at their best. In other words, this trial represents the average state of their performance during the four trials. To further investigate the effect of using this third trial, an evaluation of the proposed metrics was performed by selecting the test trial randomly between the second and fourth trials, and also between the third and fourth trials. The first trial was excluded from the randomization to avoid the effect of the learning curve. The results showed the same trend and similar values when using a random trial for testing.

Comparing the energy-based metrics proposed in this study with two commonly used metrics in MIS skills assessment (path length and task completion time), demonstrated superior performance for the one-step and two-step combined metrics for the knot-tying task, and superior performance of the two-step combined metric for the suturing task.

The superior results obtained for the knot-tying task— better accuracy and correlation with LOE—is likely due to the difficulty of the task. The knot-tying task consists of three knots, which requires a larger number of movements compared to the suturing task. The higher complexity of this task more clearly demonstrates the difference in subject proficiency. Another point that might affect the results of this study was that the level of expertise of the subjects in our experiment was determined based on their experience in MIS. However, having more experience does not necessarily result in demonstrating higher level of expertise. The motor skills of subjects might influence their performance more than their direct experience in MIS. It should be noted that the developed metrics in this study can be used for providing feedback to trainees about their level of expertise at the end of the task. Developing instructions for trainees based on these metrics requires additional research.

The results obtained in this study are not at a level that they can be relied upon blindly for

automatic discrimination of skill level. However, the energy-based metrics demonstrated a better performance than the commonly used metrics and can be used for enhancing these metrics. In order to improve the proposed energy-based metrics, additional investigation should be performed. Using a larger number of subjects would provide the opportunity to develop a more comprehensive metric in terms of the weights of the combined metrics. A larger number of subjects would also help to define more accurate margins for each LOE and provide increased robustness to the different possible ways of executing the same task. The weight factors that were used in this study were determined based on performance of right-handed subjects. Investigation of the performance of left-handed subjects is also required to establish an appropriate set of weight factors for them. Overall, there is a trade-off between speed, energy expenditure, and accuracy of performance. This trade-off should be further investigated to obtain more knowledge about how different subjects deal with task requirements. Consequently, there cannot be a single set of weights to discriminate subjects ranging from those with no expertise to expert subjects. At each main LOE, such as novice or expert, there should be a different set of weights to establish the appropriate combination of components.

3.6 Conclusion

In this study, novel metrics were proposed based on analyzing energy expenditure. The three components of these metrics were potential energy, kinetic energy, and work. Two methods of using these components for determining combined metrics were proposed in this study and were tested on a data set recorded for two laparoscopic tasks performed by subjects of various levels of experience. In conclusion, the accuracy of the one-step combined metric in identifying LOE was 47% and 60% for the suturing and knot-tying tasks, respectively. The two-step combined metric demonstrated an accuracy of 67% for both tasks. The metrics proposed here, reflect the efficiency of the performance. In these metrics, different aspects of the subjects' performance, such as motion of the instrument and the amount of force applied to the tissue, are considered for both hands. These metrics provide an objective method for assessing the LOE of subjects, can be computed automatically, and can be used for other tasks and other MIS applications. However, for each task, a particular combination of these components should be established due to different task requirements.

Bibliography

[1] H. Abboudi, M. S. Khan, O. Aboumarzouk, K. A. Guru, B. Challacombe, P. Dasgupta, and K. Ahmed, "Current status of validation for robotic surgery simulators–a systematic review," *BJU international*, vol. 111, no. 2, pp. 194–205, 2013.

[2] B. McCrory, C. A. LaGrange, and M. Hallbeck, "Quality and safety of minimally invasive surgery: Past, Present, and Future," *Biomedical Engineering and Computational Biology*, vol. 6, pp. 1–11, 2014.

[3] M. Karahan, G. M. M. J. Kerkhoffs, P. Randelli, and G. J. Tuijthof, *Effective Training of Arthroscopic Skills*.   Springer Berlin Heidelberg, 2015.

[4] T. Horeman, K. Sherman, and G. J. Tuijthof, "What measures represent performance?" in *Effective Training of Arthroscopic Skills*.   Springer, 2015, pp. 125–140.

[5] A. L. Trejos, R. V. Patel, R. A. Malthaner, and C. M. Schlachta, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surgical Endoscopy*, vol. 28, no. 7, pp. 2106–2119, 2014.

[6] S. Estrada, M. K. O'Malley, C. Duran, D. Schulz, and J. Bismuth, "On the development of objective metrics for surgical skills evaluation based on tool motion," in *IEEE International Conference on Systems, Man and Cybernetics*.   San Diego, California, USA, October 5-8, 2014, pp. 3144–3149.

[7] N. R. Howells, M. D. Brinsden, R. S. Gill, A. J. Carr, and J. L. Rees, "Motion analysis: a validated method for showing skill levels in arthroscopy," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 24, no. 3, pp. 335–342, 2008.

[8] T. Horeman, J. Dankelman, F. W. Jansen, and J. J. van den Dobbelsteen, "Assessment of laparoscopic skills based on force and motion parameters," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 805–813, 2014.

[9] A. Escoto, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "Force sensing-based simulator for arthroscopic skills assessment in orthopaedic knee surgery." in *Medicine Meets Virtual Reality 19, NextMed*, Newport Beach, California, USA, February 2012, pp. 129–135.

[10] T. Horeman, S. P. Rodrigues, F. W. Jansen, J. Dankelman, and J. J. van den Dobbelsteen, "Force parameters for skills assessment in laparoscopy," *IEEE Transactions on Haptics*, vol. 5, no. 4, pp. 312–322, 2012.

[11] B. Rohrer, S. Fasoli, H. I. Krebs, R. Hughes, B. Volpe, W. R. Frontera, J. Stein, and N. Hogan, "Movement smoothness changes during stroke recovery," *The Journal of Neuroscience*, vol. 22, no. 18, pp. 8297–8304, 2002.

[12] V. Datta, M. Mandalia, S. Mackay, A. Chang, N. Cheshire, and A. Darzi, "Relationship between skill and outcome in the laboratory-based model," *Surgery*, vol. 131, no. 3, pp. 318–323, 2002.

[13] J. D. Mason, J. Ansell, N. Warren, and J. Torkington, "Is motion analysis a valid tool for assessing laparoscopic skill?" *Surgical Endoscopy*, vol. 27, no. 5, pp. 1468–1477, 2013.

[14] A. L. Trejos, R. V. Patel, M. D. Naish, R. Malthaner, and C. Schlachta, "The application of force sensing to skills assessment in minimally invasive surgery," in *IEEE International Conference on Robotics and Automation*.   Karlsruhe, Germany, May 6-10, 2013, pp. 4370–4375.

[15] N. Stylopoulos and K. G. Vosburgh, "Assessing technical skill in surgery and endoscopy: a set of metrics and an algorithm (c-pass) to assess skills in surgical and endoscopic procedures," *Surgical Innovation*, vol. 14, no. 2, pp. 113–121, 2007.

[16] S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, R. Bardsley, and S. Dawson, "Surgical training system for laparoscopic procedures," US Patent App. 10/797,874, March 10, 2004.

[17] E. R. Guthrie, *The Psychology of Learning (Rev. ed.)*.   New York: Harper and Row, 1952.

[18] S. M. Cristancho, M. Vanstone, L. Lingard, M.-E. LeBel, and M. Ott, "When surgeons face intraoperative challenges: a naturalistic model of surgical decision making," *The American Journal of Surgery*, vol. 205, no. 2, pp. 156–162, 2013.

[19] P. H. Patel, "Pressures to 'measure up' in surgical training: Managing one's impression and managing one's patient," Ph.D. dissertation, Dept. of Medical Science, University of Toronto, 2014.

[20] M. D. Robb, *The Dynamics of Motor-skill Acquisition.*   Englewood Cliffs, NJ: Prentice Hall.l, 1972.

[21] F. T. Oliveira, D. Elliott, and D. Goodman, "Energy-minimization bias: compensating for intrinsic influence of energy-minimization mechanisms," *Motor Control*, vol. 9, no. 1, pp. 101–114, 2005.

[22] D. Elliott, L. E. Grierson, S. J. Hayes, and J. Lyons, "Action representations in perception, motor control and learning: implications for medical education," *Medical Education*, vol. 45, no. 2, pp. 119–131, 2011.

[23] D. Elliott, S. Hansen, J. Mendoza, and L. Tremblay, "Learning to optimize speed, accuracy, and energy expenditure: a framework for understanding speed-accuracy relations in goal-directed aiming," *Journal of Motor Behavior*, vol. 36, no. 3, pp. 339–351, 2004.

[24] R. A. Schmidt and C. A. Wrisberg, *Motor Learning and Performance: A Situation-based Learning Approach.*   Champaign, IL: Human Kinetics, 2008.

[25] K. Tsurumi, T. Itani, N. Tachi, T. Takanishi, H. Suzumura, and H. Takeyama, "Estimation of energy expenditure during sedentary work with upper limb movement." *Journal of Occupational Health*, vol. 44, no. 6, pp. 408–413, 2002.

[26] A. M. Swartz, L. Squires, and S. J. Strath, "Energy expenditure of interruptions to sedentary behavior," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 8, no. 69, pp. 1–7, 2011.

[27] W. Sparrow and K. Newell, "Metabolic energy expenditure and the regulation of movement economy," *Psychonomic Bulletin & Review*, vol. 5, no. 2, pp. 173–196, 1998.

[28] A. Escoto, F. Le Ber, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "A knee arthroscopy simulator: design and validation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, July 3-7, 2013, pp. 5715–5718.

[29] D. Sha, C. R. France, and J. S. Thomas, "Mechanical energy expenditures and movement efficiency in full body reaching movements," *Journal of Applied Biomechanics*, vol. 26, no. 1, pp. 32–44, 2010.

[30] K. E. Gordon, D. P. Ferris, and A. D. Kuo, "Metabolic and mechanical energy costs of reducing vertical center of mass movement during gait," *Archives of Physical Medicine and Rehabilitation*, vol. 90, no. 1, pp. 136–144, 2009.

[31] A. L. Trejos, R. V. Patel, M. D. Naish, A. C. Lyle, and C. M. Schlachta, "A sensorized instrument for skills assessment and training in minimally invasive surgery," *Journal of Medical Devices*, vol. 3, no. 4, p. 041002, 2009.

[32] B. Poursartip, M.-E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Energy-based metrics for laparoscopic skills assessment." in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 16-20 August, 2016.

Chapter 4

Energy-Based Metrics for Arthroscopic Skills Assessment

The material presented in this chapter has been published in Sensors, vol. 17, no. 8 (2017): 1808.

4.1 Introduction

In this chapter, energy-based metrics are investigated for arthroscopy and these metrics are evaluated for classifying trainees into their level of expertise. Section 4.1.1 provides an introduction of arthroscopic skills assessment and related works and Section 4.1.2 describes the objectives of this chapter.

4.1.1 Skills Assessment in Minimally Invasive Surgery

Surgical simulators are now being used for training and assessment purposes in various surgical fields including arthroscopy. The advantage of using these simulators in training programs consists of unrestricted practice time, lower cost compared to cadaver models, the opportunity for independent learning, and decreasing the risk to patients in the operating room [1]. The suitability of these simulators for training and assessment purposes not only depends on a realistic design and efficient use of the simulator, but also depends on the assessment method that is incorporated into the simulator to evaluate the proficiency levels of users. Objective assessment

methods are essential in evaluating residents and surgeons before entering the operating room to increase safety of patients. Traditionally, skills assessment is performed by expert evaluators using Global Rating Scales (GRS) for scoring [2]. The Global Rating Scale for Shoulder Arthroscopy (GRSSA) is an example of a GRS developed for shoulder arthroscopy [3]. However, these methods are subjective and the results among different evaluators are inconsistent.

A clear definition of proficiency in minimally invasive surgery is not provided in the literature. Many studies demonstrate that a higher level of expertise is associated with a shorter task completion time [4, 5]. However, faster performance may result in reduced quality. Motion is another parameter that has been analyzed for skills assessment. Several metrics such as path length, velocity, and jerk are defined based on motion information [6, 7]. The amount of force applied to the target tissue has also been considered as a representation of skill proficiency [8–10]. These metrics have shown high correlation with the level of expertise. However, the currently available metrics do not address all the needs and the appropriate combination of these metrics should be investigated to enhance surgical skills assessment.

Minimum energy expenditure has been identified as a feature of general motor skills [11]. Elliot, *et al.* [12, 13] demonstrated that practicing a physical task reduces energy expenditure. Analysis of energy expenditure based on instrument kinetic energy was investigated for skills assessment in [14] in the form of Integral of Acceleration Vector (IAV). In our previous study, energy expenditure was introduced for laparoscopic skills assessment [15, 16]. In the current study, another metric is added to the previously developed energy-based metrics, the proposed metrics are normalized, and the resulting metrics are studied for arthroscopic skills.

In order to incorporate performance metrics into surgical simulators, the criteria of expertise should be defined based on the performance of subjects with various levels of expertise. Knowing these criteria, the level of expertise of a new trainee can be determined. Machine learning algorithms are helpful in defining these criteria. As different performance metrics may demonstrate various distributions over levels of expertise, an appropriate classifying algorithm is needed for each metric or for each combination of metrics. Several machine learning algorithms have been investigated in the literature. For example, the Support Vector Machine (SVM) is a classifying algorithm used to explore motion patterns in [17]. An accuracy of 91% was obtained in this study. Linear Discriminant Analysis (LDA) is another classifier used

in [18,19]. In [18], LDA was used to evaluate the combination of time, force, and motion-based metrics and 100% accuracy in classifying subjects into two groups (experts and novices) was achieved. LDA was also utilized in [20] to investigate eye metrics, which were developed based on pupillary and eye movements, and provided 91.9% accuracy. The use of Neural Networks (NN) was also explored in this study, resulting in 92.9% accuracy using the same eye metrics. In the current study, various methods of combining the normalized energy-based metrics using machine learning algorithms are investigated.

4.1.2   Objectives

Although energy expenditure was investigated in our previous studies for two laparoscopic tasks, its applicability in different areas of MIS has not been explored sufficiently. The goal of this study was to introduce and evaluate normalized energy-based metrics for basic arthroscopic tasks. Evaluating the combination of these metrics with various classifiers was also among the objectives of this study.

4.2   Methods

To accomplish the aforementioned objectives, a series of experiments were performed. In Section 4.2.1 the experimental protocol for data collection is explained. The normalized energy-based metrics, the classifiers that were used with these metrics, and the validation procedure are presented in Sections 4.2.2, 4.2.3, and 4.2.4, respectively.

4.2.1   Experimental Design

A sensorized physical shoulder simulator was used in this study for investigating three arthroscopic tasks: two probing tasks and a grasping task. The shoulder simulator was developed at Canadaian Surgical Technologies and Advanced Robotic (CSTAR), and its face and construct validity were demonstrated in [21]. The simulator and the accompanying video tower are shown in Fig. 4.1 (a). The first probing task, Task 1, consisted of pressing a switch at the top and another switch in the middle of the glenoid (Fig. 4.2 (a)). The second probing task,

(a)
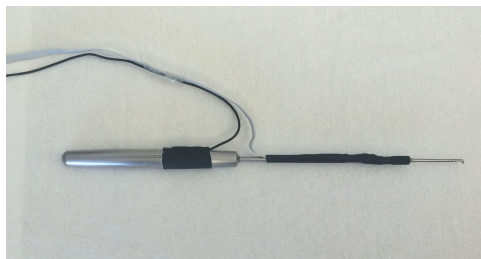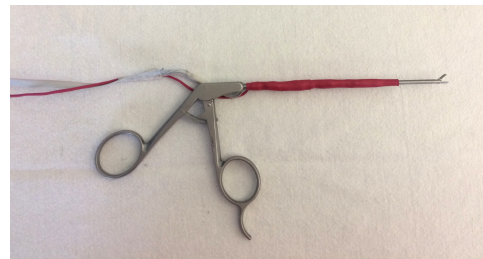
(b)   (c)

Figure 4.1: Shoulder simulator and video tower (a), the sensorized arthroscopic probe (b), and the sensorized arthroscopic grasper (c).
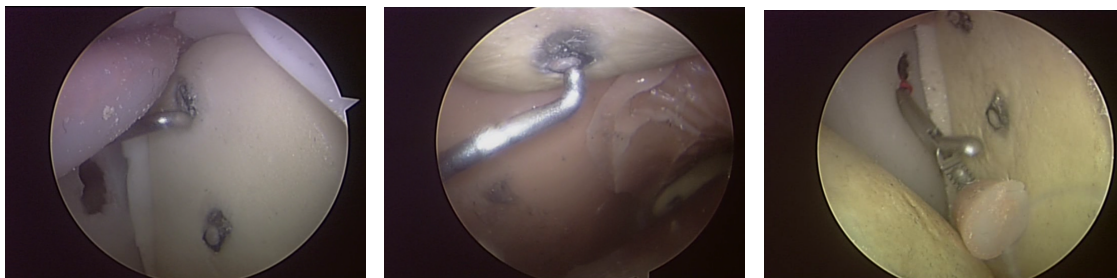
(a)   (b)   (c)

Figure 4.2: The arthroscopic tasks investigated in this study: a) Task 1, b) Task 2, and c) Task 3.

Task 2, consisted of pressing a switch underneath the acromion and another switch underneath the coracoid process (Fig. 4.2 (b)). The switches used in the probing tasks were top-actuated switches with the operating force of 1 *N*. The successful probing of each switch was indicated by the illumination of an LED located close to the base of the simulator and was also indicated in the integrated graphical user interface of the system. The grasping task, Task 3, involved grasping and removing a loose body made of silicone from the joint capsule (Fig. 4.2 (c)). For all three tasks, the arthroscopic instruments were held in the left hand and the arthroscope was held in the right hand. Prior to the start of the procedure, the arthroscope was placed such that the video provided an appropriate view of the target area. The instrument was placed outside of the simulator at the opening of the appropriate portal for the task. A sensorized arthroscopic probe and a sensorized arthroscopic grasper were used for the probing and grasping tasks, respectively (Fig. 4.1 (b, c)). These sensorized instruments were capable of measuring bending forces applied at the tip of the instrument and tracking the position of the tip of the instrument in 6 degrees of freedom (DOF). A set of four strain gauges were attached to the shaft of the instruments for force sensing, connected in a half-bridge II Wheatstone bridge. A 6 DOF position sensor (Aurora mini 6 DOF sensor, Northern Digital Inc. (NDI), Waterloo, ON, Canada), coupled to an electromagnetic position tracking system (Aurora v2, Northern Digital Inc. (NDI), Waterloo, ON, Canada), was also embedded in the shaft of each instrument and the camera [8]. The sampling frequency for both force and position data was 20 Hz. These data were then low-pass filtered with a 12 Hz cut-off frequency.

In this study, 26 participants were divided into two levels of expertise: novice (*n*=18) and expert (*n*=8). This grouping was performed based on each subject's experience in arthroscopic surgery. The novice group consisted of subjects with no surgical training, orthopaedic residents, and non-orthopaedic surgeons without scoping experience. The expert group consisted of orthopaedic fellows and fellowship-trained orthopaedic surgeons. No exclusion criterion was applied for recruitment of the participants. Human Research Ethics Board approval was obtained prior to the start of the experiments.

4.2.2   Metrics

The use of energy expenditure in the form of mechanical energy, including potential energy and kinetic energy, and work was proposed in our previous study for laparoscopic suturing

and knot-tying tasks [15]. Work (W) is generated due to a force that causes a displacement. Potential energy (E_P) is due to the position of instruments in a gravitational field and kinetic energy is due to the velocity of instruments. The energy based metrics were defined as the total work and the sum of the changes in potential energy and the sum of the changes in kinetic energy when performing a task [15]. The kinetic energy was considered due to the translational velocity of the instrument. In the current study, two forms of kinetic energy are considered: translational kinetic energy (E_{TK})—due to translational velocity, and rotational kinetic energy (E_{RK})—due to rotational velocity. The rotational kinetic-based metric is calculated according to the following formula:

$$E_{RK} = \int_0^T \frac{d(\omega_x^2 + \omega_y^2 + \omega_z^2)}{dt}\, dt, \tag{4.1}$$

where T is task completion time, and ω_x, ω_y, and ω_z are rotational velocities about x, y, and z axes. If the same instrument is used to perform a task by all of the subjects, the mass of the instrument and the moment of inertia can be removed from the equations, as they would contribute the same scaling factors to the metrics of all of the subjects. In this study, the same instruments were used and the energy-based metrics did not include the mass and moment of inertia of the instruments.

Interpreting the values of the defined metrics is not possible without knowing the amount of energy expenditure corresponding to the ideal performance. In this study, an expert arthroscopist was asked to perform the tasks of this study with the same conditions as all other subjects. This expert arthroscopist had performed well over 2500 arthroscopic interventions and was also an expert with the simulator, due to her contributions to the design of the simulator and the experiment. This trial was recorded without previous practice on the same day in order to be consistent with all of the other subjects. The energy-based metrics that were calculated based on her performance were considered as the ideal metric values. Each energy-based metric was divided by the corresponding ideal value and the resulting metrics are referred to as normalized energy based metrics (W_N, $E_{P\text{-}N}$, $E_{TK\text{-}N}$, $E_{RK\text{-}N}$). In other words, the normalized metrics indicate the performance of a subject relative to the ideal performance.

As the arthroscope was not sensorized with force sensors, the work-based metric was not calculated for the arthroscope. Consequently, four metrics were calculated for the instrument ((W_N, E_{P-N}, E_{TK-N}, E_{RK-N}) and three metrics were calculated for the arthroscope (E_{P-N}, E_{TK-N}, E_{RK-N}).

4.2.3   Trainee Classification

In order to accurately determine the level of expertise of trainees, a classifier should be trained with data from subjects at various levels of expertise. The classifiers should be able to accurately determine the level of expertise of subjects based on their performance metrics. The metrics used in this study were the normalized energy-based metrics as inputs to four classifiers: SVM, K-nearest neighbors (KNN), NN, and LDA. All of the energy-based metrics have been included in the analysis without any exclusions.

In the SVM classifier, the input data is mapped onto another feature space by a kernel function. Then the optimum hyperplane that separates the data in the mapped feature space is determined [22]. The *fitcsvm* function of MATLAB with a linear kernel function was used to establish the SVM classifier.

KNN performs the classification based on K points that lie nearest to the test data point. The test point is assigned to the class with the highest posterior probability of class membership. This is computed as K_i/K, where K_i is the number of points of Class i that lie nearest to the test point. As K increases the borders of each class become smoother, and as it decreases fine variations in each class can be determined. The choice of a large K reduces sensitivity to noise [22]; however, due to the small sample size of the current study, the choice of a large K was not possible. Considering a maximum of 6 valid trials for experts (as described in Section 4.3), K was assigned a value of 3 in this study.

NN were also investigated through the Neural Network Toolbox of MATLAB. As suggested in the literature, the maximum number of hidden layer nodes should be N/d, where N is the length of the training data and d is the number of input nodes [23]. For all three tasks of this study, the network structure consisted of 3 input nodes when the energy-based metrics of the arthroscope were considered, 4 input nodes when the energy-based metrics of the instrument were considered, and 7 input nodes when the energy-based metrics of both of the instrument

and the arthroscope were considered. In addition, one hidden layer with 3 nodes and one output node were specified in the network structure. This structure reduces computational cost and the possibility of overfitting. The training data were divided into two subsets: 70% for network training and 30% for training validation. The optimization of the weights and bias was performed by the Levenberg-Marquardt backpropagation algorithm. The target matrix was set to 1 for novices and 2 for experts. The output of the NN model was then rounded to assign the test data point to its corresponding group.

In the LDA algorithm, the multi-dimensional feature matrix is projected into one dimension by multiplying the feature matrix by a weight vector. This weight vector is determined in a manner that maximizes the separation of class means and minimizes interclass variance [22]. The *fitcdiscr* function of MATLAB was used to implement the LDA classifier.

4.2.4 Validation

4.2.4.1 Leave-one-subject-out Cross-validation

Validation of the proposed metrics and the combination of these metrics with the above-mentioned classifiers was performed through a leave-one-subject-out (LOSO) cross-validation technique. In this technique, the data is partitioned into two sets: a *test* set, consisting of one subject, and a *training* set, consisting of all subjects except the test subject. The validation procedure is repeated with different test subjects until all the subjects have been in the test group once [17, 18, 22, 24]. The level of expertise of the test subject is determined in the validation procedure, assuming that his/her level of expertise is unknown. The determined level of expertise is then compared to the level of expertise of subjects based on their experience in arthroscopic surgery.

The performance of these classifiers in combination with the normalized energy-based metrics was quantified through four measures: accuracy—ratio of the total number of correct identifications to the total number of subjects, precision—ratio of the number of experts classified as expert to the number of subjects classified as expert, recall— ratio of the number of expert subjects classified as experts to the total number of experts, and F1 score which is defined as:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.2}$$

Mistakenly classifying experts as novices indicates that they require more practice, however, wrongly classifying novices as experts can result in safety issues for patients. Consequently, it is very important to investigate the ability of the assessment method to correctly classify experts, which can be evaluated by precision and recall measures. The F1 score is the harmonic mean of precision and recall, which is the appropriate method of calculating the average of parameters that are represented as percentages. In other words, the F1 score demonstrates the balance between precision and recall [25, 26].

The performance of the energy-based metrics is also compared to the combination of task completion time, path length, and maximum bending force. This combination is evaluated using the LOSO cross validation for all of the classifiers that are investigated in this study.

4.2.4.2 Computation Time

In order to compare the computation times of the classifiers, the running time for training the classifiers and testing of all the subjects in the cross-validation was measured. The stopwatch timer of MATLAB was employed for the three tasks of this study and the mean and standard deviation values were calculated. Statistical analysis was also performed to investigate the difference between the classifiers in terms of the running time. All computations were implemented on a PC running Windows 7 with a 3.40 GHz Intel(R) Core(TM) i7-3770 CPU and 8 GB RAM.

4.3 Results

The recorded data were explored to remove any erroneous data from the analysis. The data sets that contained significant interruptions in the recording were excluded from the study. These interruptions could happen due to limited range of position tracking, sensitivity of the position tracking system to ferromagnetic metal, or a disconnection in the force sensing circuit. Therefore, the number of subjects for which valid data were recorded varied in different tasks.

Table 4.1: Number of subjects with valid data from the instrument, the arthroscope, and both the instrument and the arthroscope for the three studied tasks.

Task	Instrument			Arthroscope			Instrument and athroscope		
	Novices	Experts	Total	Novices	Experts	Total	Novices	Experts	Total
1	16	4	20	15	5	20	14	4	18
2	18	5	23	17	6	23	15	5	20
3	16	4	20	12	4	16	12	4	16

Similarly, for analysis of both hands together, the subjects whose data from either the instrument or the arthroscope was not valid were excluded. Table 4.1 shows the number of subjects with valid data from the instrument, the arthroscope, and both the instrument and arthroscope.

The experimental design of this study required holding the arthroscope in an appropriate position at the beginning of the task. In Task 1, both switches were clearly visible in front of the camera at the beginning. However, subjects were allowed to move the arthroscope as required, e.g., to zoom in on the switch or find the instrument tip. In Task 2, the switch underneath the acromion was clearly visible at the beginning, but to have an appropriate view of the switch underneath the coracoid process, subjects needed to navigate around the coracoid. In Task 3, the arthroscope was located in a position that showed the loose body, but it could be re-positioned by the subject as needed. Although the main part of the task was supposed to be completed by manipulating the instrument, the use of the arthroscope was affected by the expertise of the subjects as well. Fig. 4.3 provides a comparison of the changes in the displacement and angle of the arthroscope in 6 DOF between a random novice and a random expert during a 15 second time frame. In this figure, the same vertical limits are applied for both the novice and expert subjects to provide a clear comparison, i.e., 10 cm for displacement and $20°$ for angle. More fluctuations and changes in position and angle of the arthroscope can be seen for the novice subject compared to the expert one. These fluctuations result in a higher energy expenditure by the novice subjects.
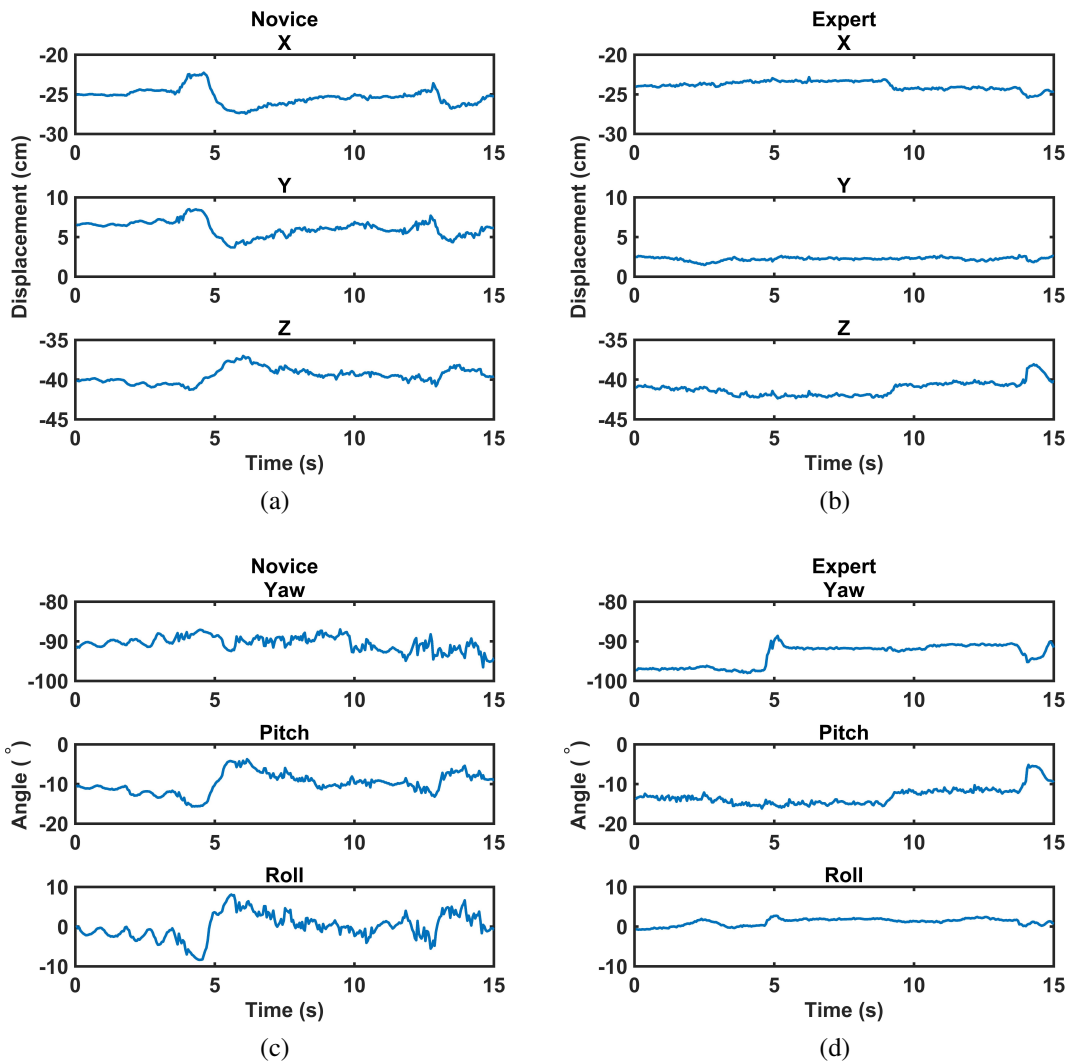
Figure 4.3: The arthroscope's tip displacement (a,b), and angle (c,d) for a random novice and expert subject over the same time duration.

4.3.1   Energy-based Metrics and Normalized Energy-based Metrics

The valid data were used to calculate energy-based metrics for the left hand (holding the proper instrument for the task) and the right hand (holding the arthroscope). As can be seen in Fig. 4.4, the amount of energy expenditure for the experts was considerably lower than that for the novices. As seen in this figure, Task 2 required higher levels of energy than Tasks 1 and 3. This is due to the position of the switches, which required more effort, even by experts. Tasks 1 and 3 required similar ranges of energy in terms of potential energy, translational kinetic energy,
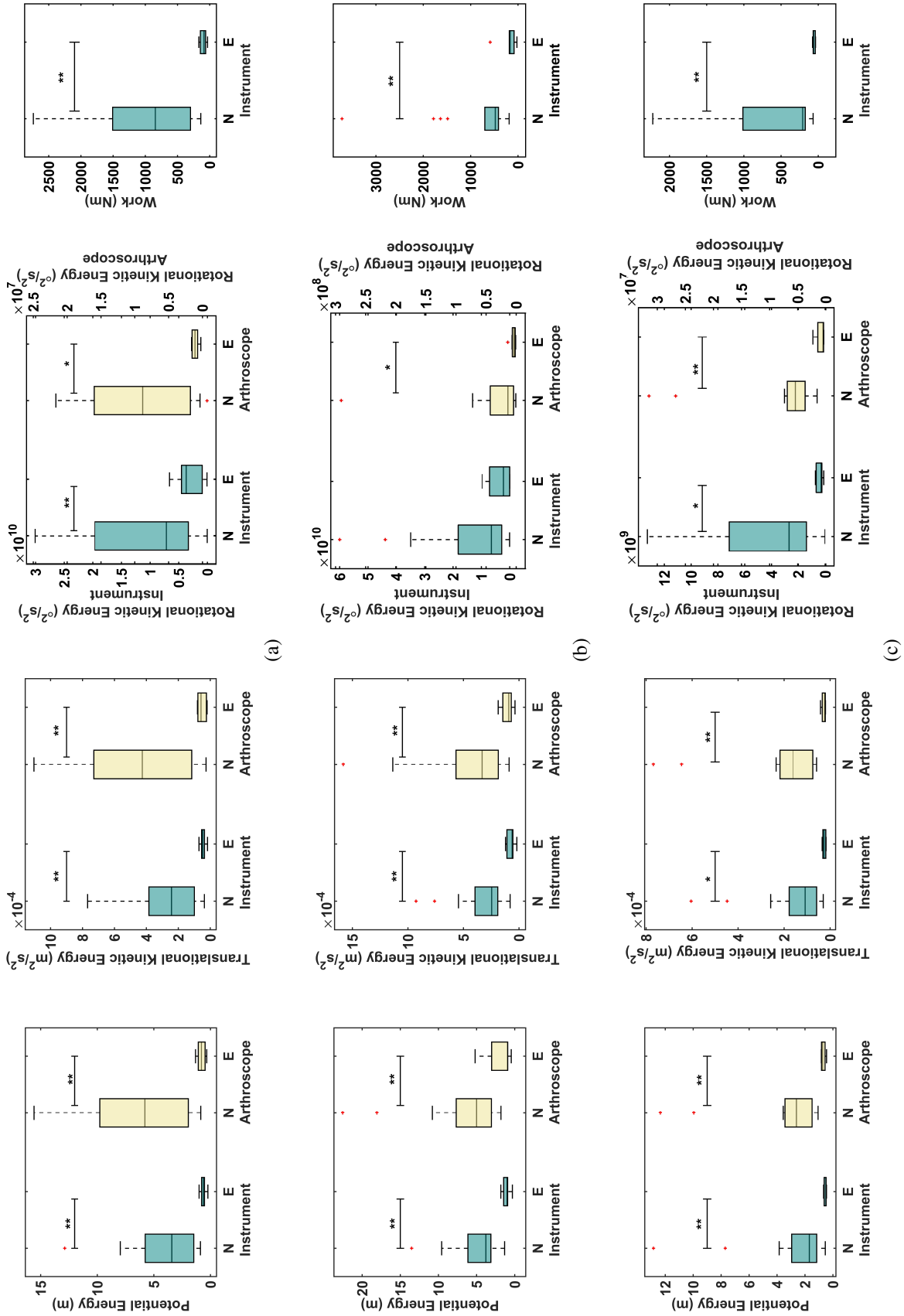
Figure 4.4: Energy based metrics for the instrument (left) and arthroscope (right). a) Task 1, b) Task 2, and c) Task 3. In this figure, ** indicates a statistically significant difference with p value less than 0.01 and * indicates a statistically significant difference with p value less than 0.05.

and work. However, the required amount of rotational energy for Task 3 was considerably less than the corresponding value for Task 1. The Probing tasks required manipulation of the probe in certain angles to successfully press the switches, which was not required in Task 3. Regarding the outliers in Fig. 4.4, the videos of subjects who were recognized as outliers were inspected to find any external reason that might affect their performance. As these outlier points were not related to a reasonable cause, they were included in the analysis.

The normality of the results for each metric was analyzed using the Shapiro-Wilk test through the Statistical Package for the Social Sciences, Version 24 (SPSS, Chicago, IL, USA). The normality test was rejected for some of the energy-based metrics in different tasks. The metrics with a normal distribution were analyzed using the Independent-Sample t test and the metrics with non-normal distribution were analyzed using the Mann-Whitney U test of SPSS. The statistical analysis showed a significant difference between the two levels of expertise for all the energy based metrics except rotational kinetic energy of the instrument for Task 2. These metrics were then normalized with respect to the corresponding values of the ideal performance of each task as was explained in Section 4.2.2. The mean and standard deviation of the resulting metrics, the normalized energy-based metrics, are shown in Table 4.2. Statistical analysis was also performed on these metrics using the Independent-Sample t test or the Mann-Whitney U test depending on whether the data presented a normal or non-normal distribution. The metrics with a normal distribution are marked by an asterisk in the p value columns of Table 4.2. For most of the normalized energy-based metrics, the mean values of the expert group were close to 1 and there was a significant difference between the expert and the novice groups. The small variance among the expert group demonstrates the similarity of the performance of the expert subjects to the ideal performance. The only metric that had a mean value considerably higher than 1 was the rotational kinetic energy for Task 1. This can be due to the unfamiliarity of the subjects with the appropriate angle of holding the instrument when pressing the switches. This metric decreases significantly in Task 2.

4.3.2   Validation

The accuracy of classification using the normalized metrics and the investigated classifiers are shown in Fig. 4.5. Overall, considering the metrics of the arthroscope as the only inputs to the

Table 4.2: The mean and standard deviation of the normalized energy-based metrics for the novice and expert groups and the corresponding p values. The statistically significant p values are shown in bold. All of the metrics, except $E_{\text{RK-N}}$ of the instrument and the arthroscope for Task 2, which are indicated with an asterisk, demonstrated statistically significant differences between novices and experts. The metrics with a normal distribution are marked with † in the p value column.

	Metric	Level	Task 1		Task 2		Task 3	
			Mean ± SD	p value	Mean ± SD	p value	Mean ± SD	p value
Instrument	$E_{\text{P-N}}$	Novice	5.87 ± 4.47	**0.001**	5.15 ± 3.26	**<0.001**	5.22 ± 5.55	**0.002**
		Expert	0.88 ± 0.44		1.18 ± 0.60		1.10 ± 0.17	
	$E_{\text{TK-N}}$	Novice	6.24 ± 4.65	**<0.001†**	5.92 ± 4.12	**0.001**	6.92 ± 6.98	**0.022**
		Expert	1.07 ± 0.51		1.37 ± 0.74		1.84 ± 1.29	
	$E_{\text{RK-N}}$	Novice	145.35 ± 127.06	**0.014†**	8.05 ± 10.15	0.199 *	5.10 ± 3.33	**0.003 †**
		Expert	50.56 ± 29.24		2.41 ± 2.65		1.95 ± 0.78	
	W_{N}	Novice	29.43 ± 22.71	**<0.001†**	9.92 ± 10.16	**0.007**	7.64 ± 7.89	**0.001**
		Expert	3.49 ± 1.32		2.68 ± 2.52		0.53 ± 0.21	
Arthroscope	$E_{\text{P-N}}$	Novice	11.89 ± 9.35	**0.001†**	7.07 ± 6.10	**0.024**	8.34 ± 8.02	**0.001**
		Expert	1.59 ± 0.79		2.34 ± 1.93		2.23 ± 1.56	
	$E_{\text{TK-N}}$	Novice	10.72 ± 8.74	**0.011**	6.44 ± 5.47	**0.002**	8.38 ± 8.56	**0.001**
		Expert	1.40 ± 0.77		1.58 ± 0.79		1.84 ± 1.88	
	$E_{\text{RK-N}}$	Novice	17.23 ± 23.34	**0.042**	10.95 ± 19.03	0.062 *	10.46 ± 11.53	**0.002**
		Expert	1.56 ± 0.79		1.44 ± 1.36		1.09 ± 1.10	
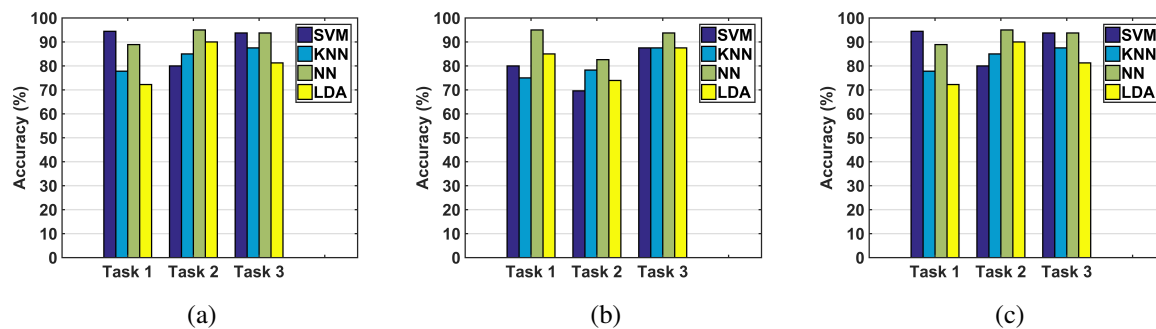
Figure 4.5: Accuracy using a) only the instrument's metrics, b) using only the arthroscope's metrics, and c) using metrics of both the instrument and the arthroscope.

classifiers provides lower accuracy levels than incorporating the metrics of the instruments in the classification. In addition, the NN method demonstrated higher accuracy levels compared to the other classifiers. Accuracy, precision, recall, and F1 score, for using the normalized energy-based metrics of both hands, including the metrics of the instruments and the arthroscope together, are shown in Table 4.3. Although the results were superior for the instrument

Table 4.3: Accuracy, precision, recall, and F1 score as percentages when the normalized energy-based metrics of both the instrument and the arthroscope are used as inputs of the classifiers.

Task	Classifier	Accuracy	Precision	Recall	F1 score
	SVM	94.44	80.00	100.00	88.89
	KNN	77.78	50.00	75.00	60.0
1	NN	88.89	75.00	75.00	75.00
	LDA	72.22	42.85	75.00	54.55
	SVM	80.00	57.14	80.00	66.67
	KNN	85.00	75.00	60.00	66.67
2	NN	95.00	100.00	80.00	88.89
	LDA	90.00	80.00	80.00	80.00
	SVM	93.75	80.00	100.00	88.89
	KNN	87.50	66.67	100.00	80.00
3	NN	93.75	80.00	100.00	88.89
	LDA	81.25	60.00	75.00	66.67

only, the inclusion of both hands was considered to be the broader use of the metrics and the corresponding results are reported to allow for comparison of the different classifiers. The NN method provides the highest accuracy for nearly all of the tasks and different input metrics. NN also demonstrate precision levels higher than 75%.

Temporal, motion-based and force-based metrics were calculated in a previous study for the same data set [27]. The results of [27] showed statistically significant differences between the experts and novices for most of the investigated metrics. The performance of the classifiers in conjunction with task time, path length for both the instrument and the arthroscope, and maximum bending force were evaluated and the results are presented in Table 4.4. As can be seen, the results that were obtained using the normalized energy-based metrics provide superior accuracy, precision, and recall in a larger number of conditions of using different classifiers and tasks. However, for some of the conditions, such as using NN for Tasks 2 and 3, both the energy-based metrics and the non-energy metrics provide similar accuracy levels.

The running times were also measured for different tasks and the mean and standard deviations are represented in Table 4.5 for various classifiers. As can be seen, NN require the maximum running time among the four classifiers investigated in this study. The difference between these running times was investigated using Kruskal-Wallis test, followed by post-hoc

Table 4.4: Accuracy, precision, recall, and F1 score as percentages when task time, path length of both the instrument and the arthroscope, and maximum bending force of the instrument are used as inputs to the classifiers.

Task	Classifier	Accuracy	Precision	Recall	F1 score
	SVM	66.67	25.00	25.00	25.00
	KNN	61.11	0.00	0.00	0.00
1	NN	83.33	66.67	50.00	57.14
	LDA	88.89	75.00	75.00	75.00
	SVM	80.00	66.67	40.00	50.00
	KNN	75.00	50.00	20.00	28.57
2	NN	95.00	100.00	80.00	88.89
	LDA	80.00	100.00	20.00	33.33
	SVM	81.25	60.00	75.00	66.67
	KNN	87.50	75.00	75.00	75.00
3	NN	93.75	100.00	75.00	85.71
	LDA	87.50	100.00	50.00	66.67

Table 4.5: Mean and standard deviation of the running time for different classifiers.

Classifier	SVM	KNN	NN	LDA
Running time (s) (Mean ± SD)	0.969 ± 0.028	0.866 ± 0.039	3.290 ± 0.452	1.015 ± 0.033

tests. The results of statistical analysis showed that the running time of NN is significantly different from that of the KNN, SVM, and LDA with the following p values, respectively: <0.001, 0.001, and 0.044. In addition, the running time of KNN and LDA were also significantly different (p=0.001).

4.4   Discussion

The goal of this study was to develop new metrics for arthroscopic skills assessment and evaluate the use of these metrics with different classifiers to determine a subject's level of expertise. The results of this study are discussed in detail in the following sections.

4.4.1   Normalized Energy-based Metrics

All energy-based metrics showed higher levels of energy expenditure for novices compared to experts. This is due to a larger number of movements of the instrument or the arthroscope and higher levels of applied forces that were unnecessary for completion of the task. These unnecessary forces and movements can be due to lack of appropriate control over the instrument or the arthroscope. The tasks we studied were designed to focus on the performance of the instrument. However, it was noticed that there were significant differences in manipulating the arthroscope between the experts and novices. The unnecessary arthroscope movements may have been generated as a result of motor overflow, which can occur in effortful actions [28, 29]. It was also observed in [30] that an unsuccessful navigation in cadaver models using an arthroscope generates large number of fluctuations in the applied force. The arthroscopic tasks studied here were comprehended as complicated motor activities for many of the novices. The statistical significant difference between novices and experts and the approximately similar accuracy levels that various classifiers provided for each task, support the presence of a strong relationship between the normalized energy-based metrics and level of expertise. The comparison between the energy-based metrics and the combination of time, path length, and maximum force showed that higher accuracy levels can be achieved for all three tasks studied using energy-based metrics in conjunction with some of the classifiers such as SVM.

4.4.2   Instrument, Arthroscope, or Both?

The maximum accuracy of 95% was obtained for all three input conditions, Fig. 4.5. However, the overall accuracy levels for different tasks were lower when the arthroscope's metrics were the only inputs to the classifiers. Regarding the arthroscope's metrics, it should be considered that these metrics were developed based on the motion parameters only and the work-based metric was not calculated. This indicates the importance of measuring force for surgical skills assessment, which is in accordance with the results found in [18, 31]. The inferior results of skills assessment based on the arthroscope's metrics can be due to the absence of a work-based metric in the assessment or because of the secondary role of the arthroscope in performing the

tasks. However, for other tasks that require further navigation of the arthroscope, more accurate identification might be obtained by incorporating the arthroscope's metrics.

4.4.3 Classifiers

The classifiers investigated in this study are among the machine learning algorithms that do not require heavy computations. These classifiers provided approximately similar results. However, the KNN and LDA have demonstrated the minimum accuracy and precision among the classifiers used. The LDA reduces the dimension of the input data and in this procedure tries to maximize the distance between the mean values of the two groups. However, the difference between the mean values of the two groups is not usually the best criterion of discrimination. Since normality is among the assumptions of the LDA, another reason for the low accuracy of this classifier may be the non-normal distribution of some of the normalized energy-based metrics. The KNN classifiers do not require a particular distribution of the samples, but has shortcomings such as sensitivity to the local structure of the data and the curse of dimensionality. In addition, the performance of KNN is affected by the value of K, which in our study was limited due to the limited number of experts.

SVM and NN provided promising results. The range of accuracy of NN was 89%–95%. In this study, a simple configuration was considered for the NN to prevent overfitting. This method is robust to an increase in the number of inputs and is also capable of learning non-linear relationships. However, a dependency on the initial conditions and a large computational burden can be cited as disadvantages of this method. SVM provides a unique solution for classification and offers a reasonable computational time. This method provides the highest accuracy levels (95%) but when considering the arthroscope's metrics for Task 2, SVM did not demonstrate a high accuracy.

The results of our study are comparable to the results of previous studies in surgical skills assessment. According to our results, the groups of novices and experts can be discriminated with 95% accuracy, which is slightly higher than the results reported in [17, 20] (92%) and is slightly lower than the results of [18] (100%). However, it should be noted that these results also depend on the tasks studied, the diversity of subjects, and data recording methods. The results of our method, which are also close to the accuracy level of previous studies, demonstrate the high potential of the proposed metrics and classifiers for surgical skills assessment.

4.4.4 Tasks

In this study, two probing tasks (Task 1 and Task 2) were investigated in different shoulder locations. The two non-significant differences between novices and experts were found for the normalized rotational kinetic energy for Task 2. The difficult posture required to press the switches in this task increased the complexity of the task, even for some of the expert subjects. This task might be valuable for studies that also investigate intermediate levels of expertise. Task 1 and Task 3 demonstrated suitable levels of difficulty for distinguishing the two levels of expertise. However, performing Task 3 in a wet environment—closer to a real surgical condition— can possibly increase the difficulty of this task by impacting the degree of visibility of the anatomical structures.

To summarize, the energy-based metrics were analyzed for the first time for arthroscopic tasks. In addition, a new energy-based metric, rotational kinetic energy, was proposed and evaluated. In this study, the role of the arthroscope was secondary relative to the role of the other instrument in completing the tasks. However, it was shown that even for the arthroscope, there was a significant difference between experts and novices in terms of the energy-based metrics. The normalization of the metrics provided additional information about variation in performance in the novice and expert groups. Furthermore, various machine learning algorithms were evaluated in conjunction with the normalized metrics to establish the appropriate combination of the proposed metrics, and their performances were evaluated by implementing various measures. Although this study uses some of the energy-based metrics that were introduced in our earlier study, several new aspects of their use have been investigated for the first time and have been modified to improve the quality of skills assessment. In addition to the above novelties and in comparison with other studies in the area of surgical skills assessment, this study evaluates various machine learning algorithms for the normalized energy-based metrics and for arthroscopic tasks.

4.5   Conclusions and Future Work

This study proposed novel performance metrics based on normalized mechanical energy and work. The incorporation of these metrics for arthroscopic skills assessment was studied. For

this purpose, various machine learning classifiers were investigated, among which Support Vector Machines (SVM) and Neural Networks (NN) demonstrated high discrimination capabilities. The validation results showed that these metrics are capable of differentiating between novices and experts with 95% accuracy. It was also demonstrated that the work-based metrics can enhance the accuracy of classification. Consequently, it is recommended that force sensing is incorporated into data recording system to establish a more accurate assessment method. Overall, our results show that normalized energy based metrics can enhance arthroscopic skills assessment. The normalization of the metrics using ideal performance metrics allows trainees to compare their performance with the ideal performance.

One of the future works of this study is to record further performance data for the arthroscopic tasks. Larger numbers of samples would provide more comprehensive models of performance at each level of expertise. In particular, more data related to expert performance can be used to further refine the criteria of expertise. Investigating the use of these metrics for finer classification of the levels of expertise, including intermediate levels, is another future direction of this study. In addition, the appropriate form of using the energy-based metrics for providing feedback and the methods for presenting this data to trainees need to be explored as part of future work.

Bibliography

[1] P. G. Roberts, P. Guyver, M. Baldwin, K. Akhtar, A. Alvand, A. J. Price, and J. L. Rees, "Validation of the updated arthros simulator: face and construct validity of a passive haptic virtual reality simulator with novel performance metrics," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 25, pp. 616–625, 2016.

[2] C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surgical Endoscopy*, vol. 25, no. 2, pp. 356–366, 2011.

[3] A. C. Hoyle, C. Whelton, R. Umaar, and L. Funk, "Validation of a global rating scale for shoulder arthroscopy: a pilot study," *Shoulder & Elbow*, vol. 4, no. 1, pp. 16–21, 2012.

[4] Y. Tashiro, H. Miura, Y. Nakanishi, K. Okazaki, and Y. Iwamoto, "Evaluation of skills in arthroscopic training based on trajectory and force data," *Clinical Orthopaedics and Related Research*, vol. 467, no. 2, pp. 546–552, 2009.

[5] K. D. Martin, P. J. Belmont, A. J. Schoenfeld, M. Todd, K. L. Cameron, and B. D. Owens, "Arthroscopic basic task performance in shoulder simulator model correlates with similar task performance in cadavers," *The Journal of Bone & Joint Surgery*, vol. 93, no. 21, p. e127, 2011.

[6] A. Escoto, F. Le Ber, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "A knee arthroscopy simulator: design and validation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan*, 3–7 July, 2013, pp. 5715–5718.

[7] I. Oropesa, P. Sánchez-González, P. Lamata, M. K. Chmarra, J. B. Pagador, J. A. Sánchez-Margallo, F. M. Sánchez-Margallo, and E. J. Gómez, "Methods and tools for objective

assessment of psychomotor skills in laparoscopic surgery," *Journal of Surgical Research*, vol. 171, no. 1, pp. e81–e95, 2011.

[8] A. Escoto, A. L. Trejos, M. D. Naish, R. V. Patel, and M.-E. LeBel, "Force sensing-based simulator for arthroscopic skills assessment in orthopaedic knee surgery." in *Medicine Meets Virtual Reality 19, NextMed*, 2012, pp. 129–135.

[9] P. Puangmali, K. Althoefer, L. D. Seneviratne, D. Murphy, and P. Dasgupta, "State-of-the-art in force and tactile sensing for minimally invasive surgery," *IEEE Sensors Journal*, vol. 8, no. 4, pp. 371–381, 2008.

[10] N. Cutler, M. Balicki, M. Finkelstein, J. Wang, P. Gehlbach, J. McGready, I. Iordachita, R. Taylor, and J. T. Handa, "Auditory force feedback substitution improves surgical precision during simulated ophthalmic surgery," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 2, pp. 1316–1324, 2013.

[11] E. R. Guthrie, *The Psychology of Learning (Rev. ed.)*. New York: Harper and Row, 1952.

[12] D. Elliott, L. E. Grierson, S. J. Hayes, and J. Lyons, "Action representations in perception, motor control and learning: implications for medical education," *Medical Education*, vol. 45, no. 2, pp. 119–131, 2011.

[13] D. Elliott, S. Hansen, J. Mendoza, and L. Tremblay, "Learning to optimize speed, accuracy, and energy expenditure: a framework for understanding speed-accuracy relations in goal-directed aiming," *Journal of Motor Behavior*, vol. 36, no. 3, pp. 339–351, 2004.

[14] F. Cavallo, S. Sinigaglia, G. Megali, A. Pietrabissa, P. Dario, F. Mosca, and A. Cuschieri, "Biomechanics–machine learning system for surgical gesture analysis and development of technologies for minimal access surgery," *Surgical Innovation*, vol. 21, no. 5, pp. 504–512, 2014.

[15] B. Poursartip, M.-E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Energy-based metrics for laparoscopic skills assessment." in *38th Annual International Conference of*

the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, Florida, USA, Orlando, Florida, USA, 16-20 August, 2016, pp. 2648–2651.

[16] B. Poursartip, M.-E. LeBel, R. Patel, M. Naish, and A. L. Trejos, "Analysis of energy-based metrics for laparoscopic skills assessment," *IEEE Transactions on Biomedical Engineering*, 2017. Available online: http://ieeexplore.ieee.org/document/7932145 (accessed on 19 May 2017).

[17] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 6, pp. 981–991, 2015.

[18] T. Horeman, J. Dankelman, F. W. Jansen, and J. J. van den Dobbelsteen, "Assessment of laparoscopic skills based on force and motion parameters," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 805–813, 2014.

[19] M. K. Chmarra, S. Klein, J. C. de Winter, F.-W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surgical Endoscopy*, vol. 24, no. 5, pp. 1031–1039, 2010.

[20] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye metrics as an objective assessment of surgical skill," *Annals of Surgery*, vol. 252, no. 1, pp. 177–182, 2010.

[21] L. C. McCracken, "Development of a physical shoulder simulator for the training of basic arthroscopic skills," *Master's Thesis, Biomedical Engineering Program, Western University*, 2015.

[22] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, New York, 2nd ed 2007.

[23] S. Xu and L. Chen, "A novel approach for determining the optimal number of hidden layer neurons for FNN's and its application in data mining," in *In 5th International Conference*

on Information Technology and Applications (ICITA), Cairns, Queensland, Australia, 23-26 June, 2008, pp. 683–686.

[24] C. E. Reiley and G. D. Hager, "Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill," in *M2CAI Workshop. MICCAI, London*, 2009.

[25] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, 2016.

[26] R. Kumar, A. Jog, A. Malpani, B. Vagvolgyi, D. Yuh, H. Nguyen, G. Hager, and C. C. G. Chen, "Assessing system operation skills in robotic surgery trainees," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 8, no. 1, pp. 118–124, 2012.

[27] B. Poursartip, L. McCracken, A. Escoto, R. Patel, M. LeBel, A. Trejos, and M. Naish, "Development and evaluation of a sensorized shoulder simulator," in *Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on*.   IEEE, 2017, pp. 1–4.

[28] J. A. Bodwell, R. K. Mahurin, S. Waddle, R. Price, and S. C. Cramer, "Age and features of movement influence motor overflow," *Journal of the American Geriatrics Society*, vol. 51, no. 12, pp. 1735–1739, 2003.

[29] P. K. Addamo, M. Farrow, J. L. Bradshaw, and N. Georgiou-Karistianis, "Relative or absolute? implications and consequences of the measures adopted to investigate motor overflow," *Journal of Motor Behavior*, vol. 43, no. 3, pp. 203–212, 2011.

[30] T. Horeman, G. Tuijthof, P. Wulms, G. Kerkhoffs, R. Gerards, and M. Karahan, "A force measurement system for training of arthroscopic tissue manipulation skills on cadaveric specimen," *Journal of Medical Devices*, vol. 10, no. 4, p. 044508, 2016.

[31] A. L. Trejos, R. V. Patel, R. A. Malthaner, and C. M. Schlachta, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surgical Endoscopy*, vol. 28, no. 7, pp. 2106–2119, 2014.

Chapter 5

Muscle Activity Analysis for Surgical Skills Assessment

5.1 Introduction

Minimally invasive surgery (MIS) requires the handling of long instruments in difficult surgical postures. It was shown in [1] that the physical workload of MIS is considerably higher than that required for open surgery. Learning MIS requires learning how to manipulate the instruments and adjust hand motions. It was shown in [2] that surgeons with higher levels of experience demonstrate lower levels of fatigue after performing a series of laparoscopic tasks.

The effect of training level on muscle activity has been studied previously [3–5]. The results of these studies showed that as skill level increases, muscle activity reduces and trainees learn to recruit their muscles in a more efficient manner. However, finding appropriate features of muscle activity as measured through Electromyography (EMG) has not been investigated in these studies. In addition, the work to date has only considered the effect of training for robotic surgery. The use of muscle activity features to distinguish between subjects with various levels of expertise has not been explored for manual MIS. Specifically for arthroscopy, muscle activity has not been studied to the best of our knowledge.

Robotic and manual surgery share several characteristics, such as manipulating long instruments through small portals and an indirect sense of contact force with tissue. In manual surgery, the surgeon receives no assistance, such as tremor compensation or increased de-

grees of freedom, which results in a more challenging surgical condition than robotics-assisted surgery. Consequently, muscle activity requirements might be higher for manual surgery.

Section 5.1.1 provides a review of various EMG features that have been defined in previous studies for different applications of EMG processing.

5.1.1 EMG Features

EMG features can be divided into three groups: *frequency-domain* features, *time-domain* features, and *time-frequency-domain* features [6]. Each group of EMG features is explained in the following sections.

5.1.1.1 Frequency-domain Features

Analyzing the frequency content of EMG signals is a useful approach for investigating muscle activity. Overall, frequency-domain features represent muscle fatigue and muscle recruitment and are extracted from the power spectrum of the signal. Various features have been reviewed in [6, 7]. The most widely used features are described here.

Mean frequency   The mean frequency or central frequency (f_c) is the weighted mean of the frequency from the EMG signal, in which the weights for each frequency are defined as the power densities at the corresponding frequency, divided by the sum of the power densities for the total range of frequencies [6, 8], as follows:

$$f_c = \frac{\sum_{j=1}^{N} f_j P_j}{\sum_{j=1}^{N} P_j},$$

(5.1)

where P_j and f_j are the jth samples of power density and frequency, respectively, and N is the total number of points in the power spectrum of the signal.

Median frequency   The median frequency (f_{med}) of the power spectrum is the frequency at which the sum of the power density from 0 to f_{med} equals the sum of the power density from f_{med} to the maximum frequency of the power spectrum, as follows:

$$f_{\text{med}} : \sum_{j=1}^{f_{\text{med}}} P_j = \sum_{j=f_{\text{med}}}^{f_{\text{max}}} P_j, \tag{5.2}$$

where P is power density, and f_{med} and f_{max} are the median and maximum frequencies of the power spectrum. A higher median frequency is associated with faster performance and less muscle fatigue [3].

Mean power   The average of the power spectrum density is proposed in the literature as a feature of EMG signals [6].

Total power   The sum of power density or energy of the power spectrum was one of the frequency-domain features used in [6].

Bandwidth   Frequency bandwidth is the difference between the maximum and the minimum frequency at which the power spectrum is half of its maximum level, which provides the 3 dB bandwidth of the signal. Bandwidth is proportional to the range of muscles that are involved in the performance [3].

Spectral moments   Moments of the power spectrum are other features of EMG signals that have been used. The first, second, and third moments have been recognized as the most informative moments [6], as follows:

$$SM1 = \sum_{j=1}^{N} f_j P_j,$$

$$SM2 = \sum_{j=1}^{N} f_j^2 P_j,$$

$$SM3 = \sum_{j=1}^{N} f_j^3 P_j,$$

Variance of central frequency   This feature is calculated based on the power spectrum of the signal according to the following formula [6,9]:

$$\text{VCF} = \frac{1}{\sum_{j=1}^{N} P_j} \sum_{j=1}^{N} P_j (f_j - f_c)^2,$$ (5.3)

where f_c is the mean frequency and is calculated according to Eq. 5.1.

5.1.1.2   Time-domain Features

Time-domain features depend on the amplitude of the signal. There are four main categories that can be specified for these features as described below:

a) Direct dependence on signal amplitude   The following features belong to this group:

- IEMG—Integral of the absolute value of the EMG amplitude [4, 10].

- MAV or EMGR—Average of the absolute value of the EMG amplitude. This metric was named Mean Absolute Value (MAV) in [6] and was called EMG Rate (EMGR) in [4].

- MAV1 and MAV2—Weighted mean absolute value. For instance, a weight of 0.5 for the beginning and end of the task and a weight of 1 for the rest of the task can be considered [6]. Two versions of this feature are defined in [6, 7], as follows:

$$\text{MAV1} = \frac{1}{N} \sum_{i=1}^{N} w_i |x_i|;$$

$$w_i = \begin{cases} 1, & \text{for } 0.25N \le i \le 0.75N \\ 0.5, & \text{otherwise.} \end{cases}$$

$$\text{MAV2} = \frac{1}{N} \sum_{i=1}^{N} w_i |x_i|;$$

$$w_i = \begin{cases} 1, & \text{for } 0.25N \le i \le 0.75N \\ \frac{4i}{N}, & \text{for } i < 0.25N \\ \frac{4(i-N)}{N}, & \text{otherwise,} \end{cases}$$

where x is EMG signal amplitude, and N is the number of samples of the EMG signal.

- RMS—Root Mean Square of the EMG amplitude [11].

- VAR—Variance of the EMG signal (VAR) [6, 7].

- TM—The nth order temporal moment of the signal. In different studies, n has been a number between 1 and 5 [12], as follows:

$$\text{TM}_n = \left| \frac{1}{N} \sum_{j=1}^{N} x_j^n \right|, \tag{5.4}$$

where x_j is the ith sample of the EMG signal.

- LOG—Log-detector, which is based on the logarithm of the amplitude. This feature was defined in [13] as follows:

$$\text{LOG} = e^{\frac{1}{N} \sum_{i=1}^{N} \log(|x_i|))} \tag{5.5}$$

b) Slope of the amplitude and changes in the sign of the signal   Zero Crossing (ZC) and Slope Sign Change (SSC) are examples of the features in this group. These features usually include a threshold criteria to avoid the effect of noise on the feature:

- ZC is the number of times that the amplitude of the EMG signal changes from negative to positive or vice versa.

- SSC is the number of times that the amplitude changes more than a certain threshold twice in three consecutive samples, as follows:

$$\text{SSC} = \sum_{i=2}^{N-1} [f[(x_i - x_{i-1}) \times (x_i - x_{i+1})]];$$

$$f(x) = \begin{cases} 1, & \text{for } x > \text{threshold} \\ 0, & \text{otherwise.} \end{cases}$$

These features represent the frequency content of the signal in the time domain [6, 7].

c) **Amplitude during the task**   A histogram is a feature that is derived based on the distribution of different levels of EMG amplitude during the task. In this feature, the total range of muscle activity is divided into a specified number of bins; and the number of data samples that are placed in each bin are calculated [6, 14].

d) **Time-series models of the signal**   Autoregressive (AR) coefficients and Cepstral coefficients are examples of this group of features. These features indirectly depend on the amplitude.

- In AR models, each sample of EMG is modeled by a linear combination of the previous data points with added white noise [6, 11, 14]:

$$x_i = \sum_{p=1}^{P} a_p x_{i-p} + w_i, \tag{5.6}$$

where x_i is the ith sample of the EMG signal, a_p is the pth coefficient of the AR model, P is the order of the model, and w is white noise. The order of the AR model was suggested in [6, 11] to be 4 for EMG signals.

- Cepstrum coefficients have been recognized as strong features of EMG signals [11]. Cepstrum coefficients can be extracted from AR coefficients as follows:

$$c_n = \begin{cases} -a_1 & \text{if } n = 1 \\ -a_n - \sum_{k=1}^{n-1} \left(1 - \frac{1}{n}\right) a_n c_{n-k} & \text{if } n > 1 \end{cases}, \tag{5.7}$$

where c_n and a_n are the nth cepstrum and AR coefficients.

In a study by Tkach *et al.* [13], the robustness of various time-domain features to disturbances were investigated with respect to the accuracy of classifying EMG signals. According to this study, AR coefficients and cepstrum coefficients are the most robust time-domain features.

5.1.1.3 Time-frequency-domain Features

Time-frequency domain features include the features that are defined based on the Short Time Fourier Transform (STFT) and the Wavelet Transform. These features represent the power spectrum over time. It was found in [15] that the wavelet coefficients are more appropriate than the STFT for investigating EMG features. Wavelet coefficients can be obtained by the following formula:

$$W(s, \tau) = \frac{1}{\sqrt{s}} \int f(t) \psi^* \left(\frac{t - \tau}{s} \right) dt, \tag{5.8}$$

where s and τ are scaling and transition parameters, t is time, $f(t)$ is the EMG signal at each instant of time, and ψ is the mother wavelet. Various mother wavelets such as Daubechies4, Cauchy, and Morelet have been used in previous studies [7, 16].

Further assessment ability can be achieved by incorporating hand movements in the assessment to interpret surgical gestures. Hand gestures can be investigated through analysis of the acceleration of the hands while performing surgical tasks. In Section 5.1.2, the metrics that were developed based on acceleration data are discussed briefly.

5.1.2 Acceleration-based Metrics

Acceleration metrics are mainly defined as the following:

- Mean acceleration—the average of acceleration magnitude during task performance

- Maximum/peak acceleration—the maximum value of the magnitude of the acceleration

- Consistency of acceleration—determined by calculating the standard deviation of the acceleration

- Integral of Acceleration Vector (IAV)—this metric is calculated according to the following formula [17]

$$IAV = \sum_{i=\text{start}}^{\text{end}} \sqrt{(a_x^2) + (a_y^2) + (a_z^2)}, \tag{5.9}$$

where a_x, a_y, and a_z are acceleration in x, y, and z directions.

5.1.3   Objectives

The hypothesis explored in this chapter was that muscle activity can be an indicating factor of surgical dexterity. The objective of this study was to explore various features of EMG signals and to identify the features that demonstrate the highest correlation with surgical psychomotor skills. This took place in the context of evaluating the use of these features for the development of metrics for surgical skills assessment. In addition, hand movements were explored in this study, as described in the following section.

5.2   Materials and Methods

The experimental design for recording a data set for arthroscopic task performance, the data recording system, and the data processing methods are explained in this section.

5.2.1   Experimental Design

The same experimental design that was explained in Chapter 4 was used for data collection. A brief summary is provided herein. Three arthroscopic tasks, including two probing tasks and one grasping task (Fig. 5.1(b, c, d)), were performed on a physical shoulder simulator (Fig. 5.1(a)). These tasks required pressing two top-actuated switches positioned at the middle and at the top of the glenoid (Task 1), pressing a similar switch underneath the acromion and another switch underneath the coracoid process (Task 2), and grasping a loose body and removing it from the joint capsule (Task 3). The probing and grasping tasks were performed using an arthroscopic probe and an arthroscopic grasper, respectively. The probe/grasper were held in the left hand and an arthroscope was held in the right hand for all of the participants. In this study, 26 participants, consisting of 18 novices and 8 experts, performed the tasks. The novice subjects did not have previous experience with scoping or any other MIS procedure. The appropriate view of the arthroscope was set for the novice subjects before starting the task. The expert group consisted of arthroscopic surgeons and orthopaedic fellows. Each participant performed a pre-practice test, which involved performing all three tasks. Then, they were

(a)

(b)   (c)   (d)
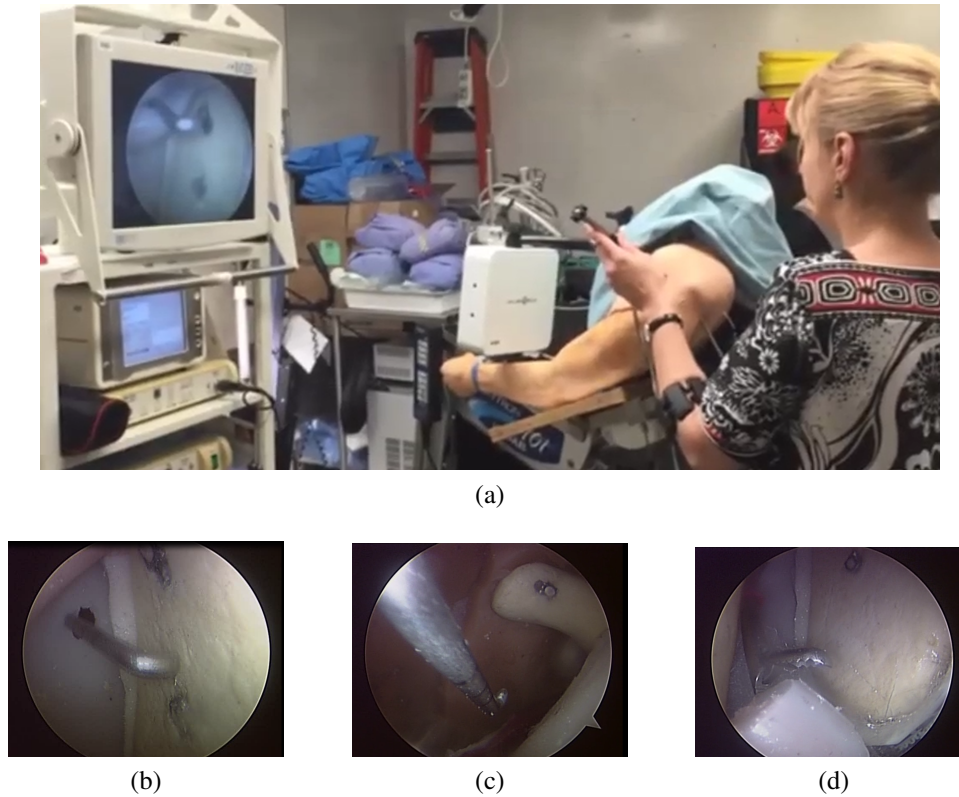
Figure 5.1: The experimental setup (a), and the three tasks of this study: b) Task 1, c) Task 2, and d) Task 3.

allowed to practice these tasks for up to 30 minutes. Finally, a post-practice test was done to evaluate the improvement of participants when performing these tasks. In these trials, 4 experts performed the tasks in the post-practice trial. The post-practice test for experts was performed without practicing. This experimental procedure was approved by the Human Research Ethics Board of Western University.

During the trials, muscle activity and acceleration of the forearm were recorded and analyzed. Before starting the tasks, the resting muscle activity of each participant was recorded for normalization so that the signals could be normalized. The data recording system is explained in the following subsection. Sensorized arthroscopic instruments were used to collect the applied force at the tip of the instrument and the position of the tip of the instrument. These data were investigated and the results have already been described in Chapter 4.

(a)   (b)   (c)
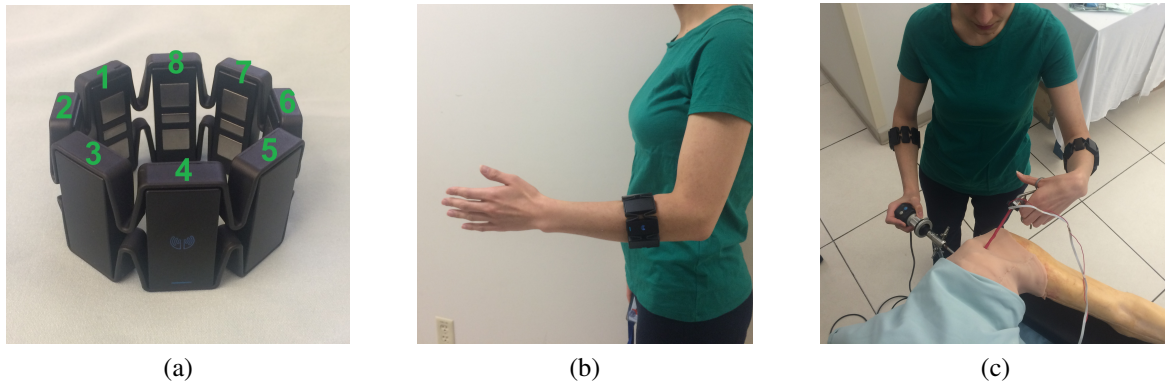
Figure 5.2: EMG and acceleration recording system: (a) Myo channel assignment, b) positioning the armband with respect to the epiconyle bone, and c) recording data while performing arthroscopic tasks.

5.2.2   Data Recording System

In order to record muscle activity, Myo armbands (Thalmic Labs, Waterloo) were used. These armbands have recently been utilized primarily in studies that are related to gesture recognition [18–21]. Myo armbands have 8 separate EMG sensors. The EMG signals of each sensor are referred to as Channel 1 to Channel 8. The assignment of channel numbers to each sensor is shown in Figure 5.2 (a). An anatomical landmark was used to align the armbands and maintain a consistent measurement for all subjects. The armbands were placed on the widest part of the forearm and the EMG sensor with the Myo logo was aligned with the lateral epicondyle bone (Fig. 5.2(b)). The LED below the logo was placed towards the distal side of the arm. The recorded muscle activity is the result of activation of the major forearm muscles. The approximate assignment of channels to forearm mucles is as follows:

– Channel 1: pronator teres and/or brachioradialis,

– Channel 2: brachioradialis,

– Channel 3: extensor carpi radialis,

– Channel 4: extensor digitorium,

– Channel 5: extensor carpi ulnaris

– Channel 6: flexor carpi ulnaris

– Channel 7: flexor digitorium

– Channel 8: flexor carpi radialis and/or pronator teres.

However, corsstalk might affect the measured signals and due to slight movement of the forearm, these assignments might shift around about one channel. The sensor assignments were determined based on the instructions provided in [22]. The EMG signals captured by these sensors are converted to unitless values in the range of -128 to +128. These data were transferred to a desktop computer through the Bluetooth protocol and were recorded at a frequency of 200 Hz [23]. In addition to EMG sensors, each armband is equipped with an Inertial Measurement Unit (IMU), which records acceleration data at 50 Hz. These data can be used for tracking the orientation of the forearm. Software simultaneously records muscle and spatial data from two armbands, corresponding to the left and right hands.

5.2.3   Data Processing

Before extracting features of the EMG signals, it was necessary to pre-process the signals to remove noise and obtain a smooth signal. According to [24], the presence of very low frequencies in the EMG signal, which lie in the range of 1–5 Hz, is usually due to low frequency noise and should be removed by filtering. Zero offset removal and high pass filtering of the signal, with a cut-off frequency of 5 Hz, were first performed. Following that, the signal was smoothed by a moving average method. In this method, a linear envelope/window moves along the signal, and the RMS value that is divided by length of the window is calculated for the data points that place inside the window [25]. The window length determines the level of smoothing: larger windows result in smoother signals; however, some information might be lost by applying large windows. Smaller windows preserve the shape and information of the signal, but do not provide considerable smoothing. Small size windows, such as 20 ms, are appropriate for studying fast movements and large windows, such as 500 ms, are preferred for studying stationary motions. Overlapped windows combine the benefits of these two types of windows. In this study, 75% overlap was established. To determine an appropriate window length, windows with various lengths from 20 ms to 100 ms were investigated for a data sample. As can be seen in Fig. 5.3, a window length of 40 ms provides appropriate smoothness and can also track rapid changes in the amplitude of the signal. All of the signal processing and subsequent feature extraction were performed in MATLAB.

(a)
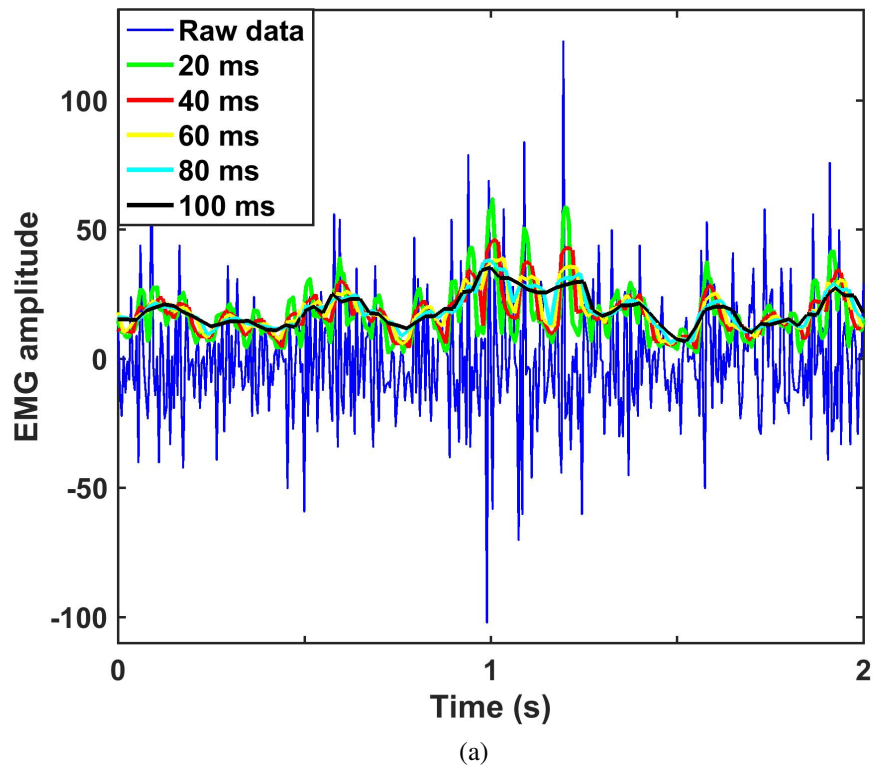
Figure 5.3: Mean RMS of a sample EMG signal for various lengths of linear envelope.

5.3   Metric Development and Evaluation

In this section, the development of metrics based on muscle activity and hand movement is described. Sections 5.3.1 and 5.3.2 describe the details that were considered for using EMG features as performance metrics and the pre-processing of the acceleration data. Sections 5.3.3 and 5.3.4 describe the method of combining the features that showed the best results and the evaluation method, respectively.

5.3.1   EMG Features as Performance Metrics

The EMG features and the acceleration metrics reviewed in Sections 5.1.1 and 5.1.2 were considered as potential performance metrics and their ability to distinguish between subjects with various levels of expertise was evaluated. All of these features were calculated for the three tasks performed in this study, for the eight EMG channels and one IMU for each hand.

Normalization of the EMG amplitude with respect to each subject's minimum muscle activity was performed prior to calculating these features. Most of the features did not require specific settings. However, the following details were considered for the histogram, AR, cepstrum, and wavelet coefficients.

- Histogram: the range of EMG signals that are recorded with the Myo armbands is between -128 and +128. In the current study, the absolute value of the EMG signals was considered and the range of 0 to 128 was divided into 8 bins with an equal length of 16.

- AR coefficients: as was suggested in [6, 11], the order of the AR model was four.

- Cespstrum coefficients: similar to the AR coefficients, an order of four was used for this feature. In addition, the first coefficient was excluded from the analysis, as this coefficient provides the same information as the first coefficient of the AR model.

- Wavelet coefficients: a Daubechies4 mother wavelet was used and the mean value of the first five wavelet coefficients was calculated.

5.3.2   Acceleration-based Metrics

The acceleration signals were prepared for further analysis by first calculating the square root of the sum of the squares (SRSS), which allows the magnitude of the acceleration to be investigated. Following that, the acceleration-based metrics reviewed in Section 5.1.2 were calculated for the left and right hands.

5.3.3   Combination of EMG-based Metrics and Acceleration-based Metrics for Classification

The optimal performance metric based on hand movements and muscle activation should incorporate various features of motion and EMG signals. The superior features for the purpose of surgical skills assessment were identified in a two-step method. First, the overall pattern of change in the correlation with level of expertise was evaluated for the eight channels of the Myo armbands. The channels that showed the highest correlation over all of the features

were selected. Second, for those channels, the metrics that demonstrated a statistically significant correlation with level of experience, with the magnitude higher than 0.5, were extracted. Support Vector Machines (SVM) were used to develop a model of proficiency based on EMG features and acceleration metrics. This classifier demonstrated a robust, fast, and unique classification according to the results of Chapter 4. A Gaussian Radial Basis Function (RBF) was used for the kernel function of the SVM.

5.3.4   Metric Evaluation

As mentioned in Section 5.3.3, the correlation of each metric with level of experience was calculated. The Spearman's rho measure was selected to quantify the correlation. In addition, the Mann-Whitney U test was performed on all of the metrics to explore their ability to differentiate between novices and experts. The Statistical Package for the Social Sciences, Version 24 (SPSS, Chicago, IL, USA) was used for analysis. An additional statistical analysis was performed to examine the ability of these features to differentiate between pre-tests and post-tests. This analysis was performed using the General Linear Model/Repeated Measures test of SPSS. In addition, the classification and the suitability of the combination of these features for use in skills assessment was evaluated through a leave-one-subject-out (LOSO) cross validation. Four measures were employed to quantify the validation: accuracy, precision, recall, and F1 score. These measures and the LOSO cross validation method are explained in Section 4.2.4.1. Separate evaluations were performed based on the metrics for the left and right hands.

5.4   Results and Discussion

All of the metrics that were reported in [6] were investigated in the current study. However, many of the proposed metrics did not show a significant correlation with level of experience. In this section, the results of each group of metrics are explained, and the results of the classification are reported. Two approaches were considered for evaluation of the metrics: 1) investigating their ability to show a difference between experts and novices, and 2) investigating their ability to show a difference between pre-test and post-test performance. Overall, for Task 1, Channels 1, 2, 4, 7, and 8 demonstrated higher correlations with level of experience.

For Task 2, Channels 1, 2, 4, and 5 were the most powerful sensor locations; and for Task 3, Channels 1, 2, 3, 5, 7, and 8 provided the best results in terms of correlation.

5.4.1 Time-domain EMG-based Metrics

Among the time-domain features that were reviewed in Section 5.1.1.2, AR coefficients, histogram, cepstral coefficients, and SSC demonstrated superior differentiation ability between different levels of experience. The results of these metrics are explained in detail in the following paragraphs. In addition, ZC and AAC showed significant differences, but since a consistent result was not found for these metrics, they have not been discussed in this section.

AR coefficients   AR coefficients demonstrated a high correlation with level of expertise for all three tasks of this study. The AR coefficients are plotted in Fig. 5.4 for the novice and expert subjects and for the most powerful channels of each task. As can be seen, AR-1 and AR-4 demonstrated higher values for the expert subjects for Tasks 1 and 2. AR-2 and AR-3 demonstrated smaller values for the experts for Tasks 1 and 2. The trend of AR coefficients for Task 3 is the opposite of the trend that exists in Tasks 1 and 2, i.e., experts produce higher coefficients for Tasks 2 and 3, and produced smaller coefficients for Task 3. This inverse trend might be due to the different requirement of this task, which was working with the grasper instrument versus holding a probe. As muscle contraction affects the EMG spectrum, it also alters the AR coefficients [14]. Consequently, these coefficients indicate the state of muscle contraction. Interpreting AR-1 for Task 3 is not easy, since there is a large overlap between the results of experts and novices. The best correlation with level of expertise for this meatric was -0.767 for AR-1 for Task 1.

Histogram   The EMG signal histogram, normalized over the minimum muscle activity of the related subject, is plotted in Fig. 5.5 for a random expert and a random novice subject. As can be seen, the distribution of most of the Myo channels are different for a novice and an expert subject. This difference is more significant for Channels 1, 2, 3, 5, 6, and 8. According to these results, novice subjects demonstrated minimum muscle activity for a considerable duration of the task. In addition, a significant difference was found between the novices and experts for
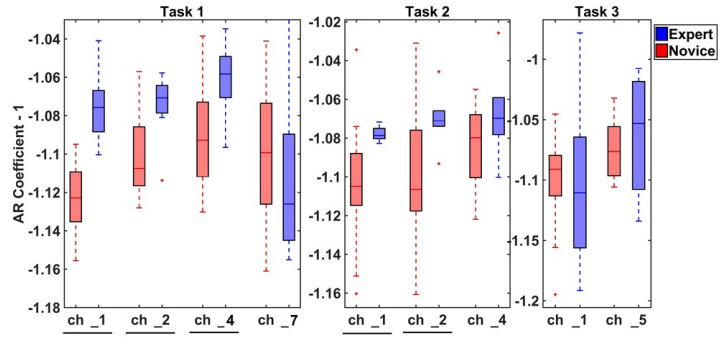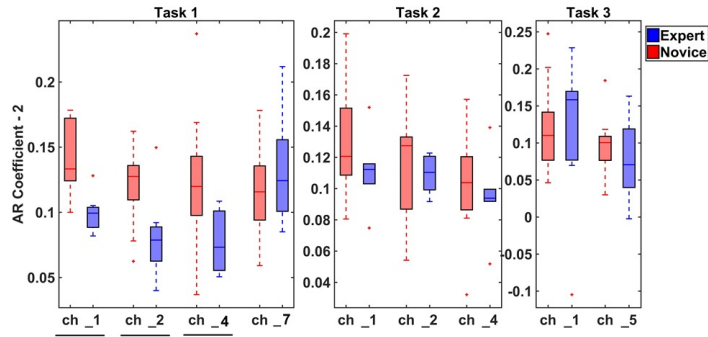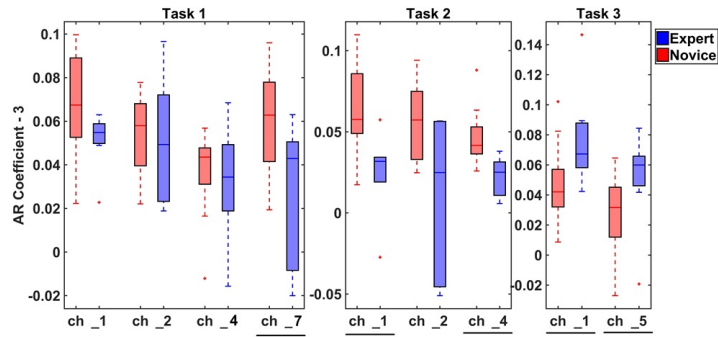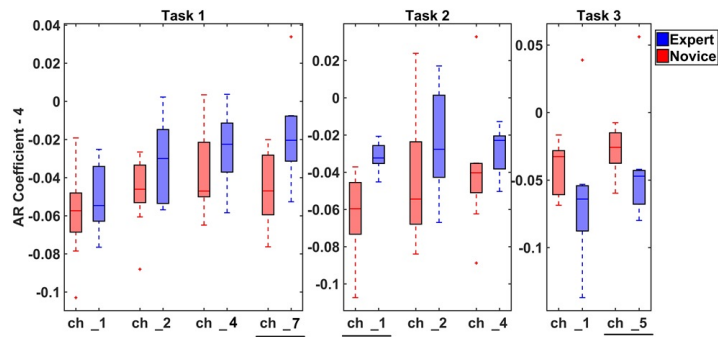
Figure 5.4: AR coefficients for Tasks 1, 2, and 3, for the Myo channels specified in the horizontal axis. The statistically significant features are shown with a line beneath the channel number.

(a)

(b)
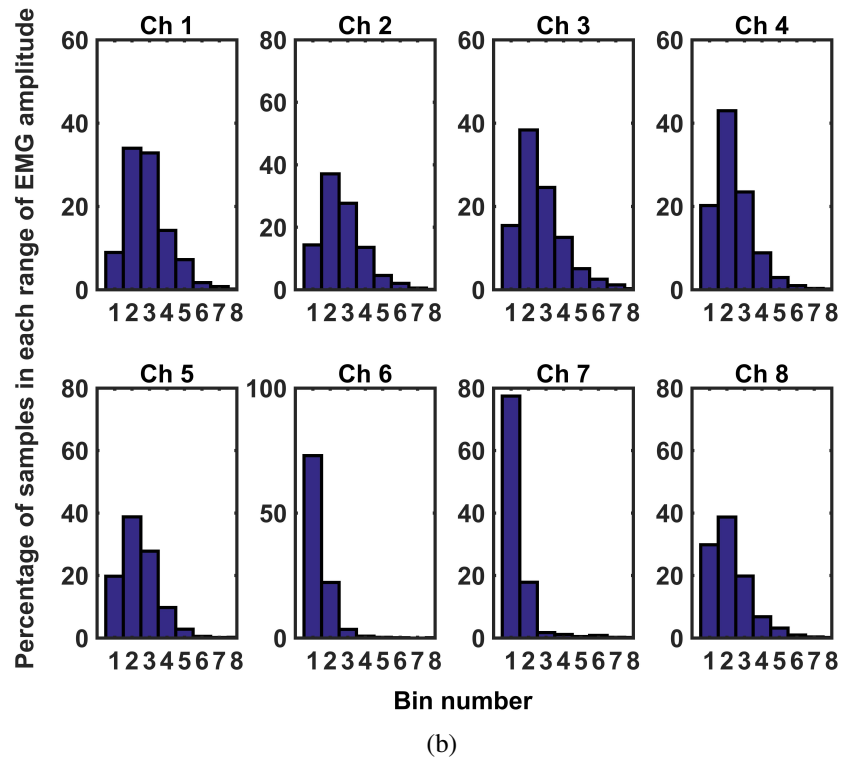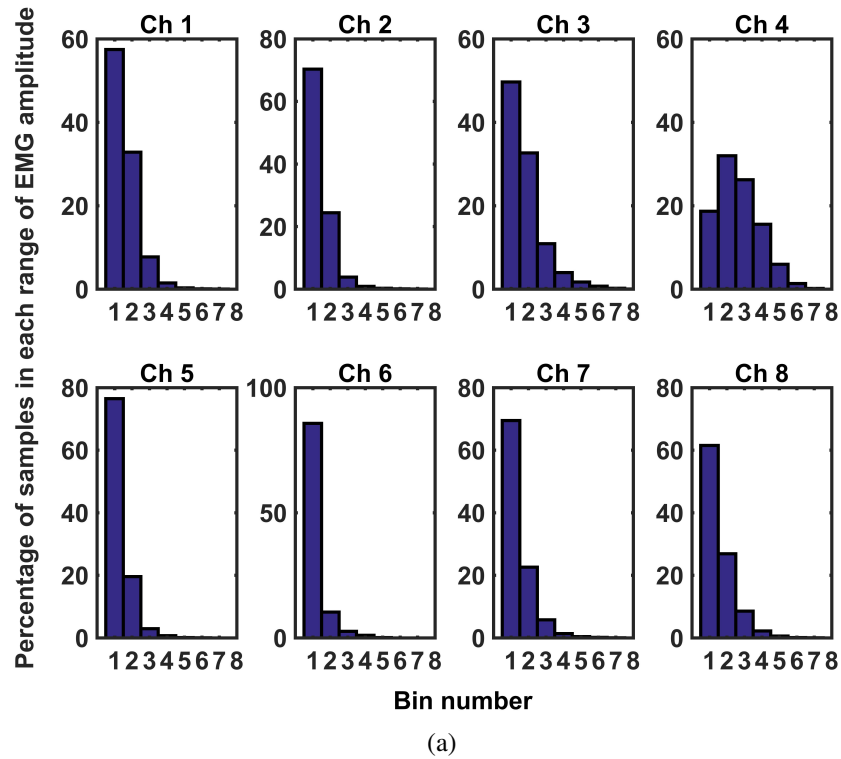
Figure 5.5: Percentage of data samples in each range of EMG amplitude, considering the left hand, for Task 1 for a) a novice and b) an expert subject.

the highest range of amplitude, which was 113–128. In this range, experts demonstrated a larger number of samples, i.e., a positive correlation was obtained. Regarding the difference between pre-test and post-test, the histogram showed significant differences in Tasks 1 and 3.

Cepstral coefficients   This feature also demonstrated high correlations with level of expertise. As the cepstral coefficients are calculated based on the AR coefficients, high correlations were expected for these features. In some cases, cesptral coefficients improved the amount of correlation that was obtained by the AR model. The cepstral coefficients calculated for the left hand for the most significant channels are shown in Table 5.1.

Table 5.1: Spearman's rho correlation with level of experience for cepstral coefficients, considering the left hand, for the Myo channels with higher levels of correlation. C1–C4 stand for the first to the 4th order cepstral coefficients. Statistically significant values are shown in bold.

Task	Myo Channel	C1	C2	C3	C4
1	1	**-0.767**	**0.734**	0.400	-0.117
	2	**-0.517**	**0.484**	0.0834	-0.234
	4	**-0.501**	**0.534**	0.100	-0.317
	7	0.184	-0.134	**0.484**	**-0.551**
2	1	**-0.511**	0.322	**0.587**	**-0.738**
	2	**-0.492**	0.303	0.435	-0.341
	3	-0.302	0.114	**0.662**	-0.416
3	1	0.086	-0.124	**-0.482**	0.395
	5	-0.124	0.124	**-0.445**	**0.470**

SSC   This metric showed a significant difference between experts and novices (Table 5.2). All of the 8 sensors of the armband showed significant correlations for Task 3. No difference between pre- and post-tests was found in terms of SSC.

Overall, the higher amplitude of EMG signals represents a higher amount of muscle recruitment and increased strain. On the other hand, low amplitude EMG signals are not necessarily associated with less effort. Small EMG amplitude might be due to exhaustion and decreased muscle power. This effect produces additional complexity in processing EMG signals in sur-

Table 5.2: Spearman's rho correlation with level of experience for SSC, considering the left hand (Task1, Task 2, Task 3). Statistically significant values are shown in bold.

Myo channel	SSC
1	(-0.417, **-0.445**, **-0.507**)
2	(**-0.484**, **-0.487**, **-0.519**)
3	(-0.417, -0.264, **-0.507**)
4	(-0.367 -0.222 **-0.667**)
5	(-0.300 -0.236 **-0.469**)
6	(-0.417, -0.278, **-0.531**)
7	(**-0.567**, -0.361, **-0.556)**
8	(**-0.701** , **-0.445**, **-0.605**)

gical performance. For the subjects that required relatively longer time to complete the task, the reduced amplitude of EMG might be misinterpreted with efficiency in muscle recruitment. Consequently, features that do not depend on the amplitude of the signal should be included in the analysis.

5.4.2   Frequency-domain EMG-based Metrics

The frequency-domain features, explained in Section 5.1.1.1, were calculated for the data set recorded in this study. The MNP, TTP, f_c, and f_{med} demonstrated the highest correlation with level of experience among the frequency-domain features. The results of these metrics are provided in the following paragraphs. Other frequency-domain features did not provide a consistent trend for all of the tasks. Overall, the number of significant correlations in frequency domain features were larger for Task 3.

MNP   Mean power is one of the frequency-based metrics that demonstrated significant correlations with the level of experience for all three tasks of this study (Table 5.3). However, this metric showed a significant difference between pre-test and post-test only for Task 3. The results of pre- and post-tests for Task 3 are shown in Table 5.4.

Since Task 3 was the easiest task, novice subjects might have a significant improvement in their post-test performance. In particular, at the post-test of Task 3, the novice subjects had the highest amount of experience in manipulating arthroscopic instruments and the highest famil-

iarity with the shoulder model. It can be inferred from these results that the MNP metric can show the difference between levels of expertise in cases where there exists a large difference between the performance of different groups.

TTP   This metric showed significant correlations for Task 1 and Task 3 (Table 5.3).

f_c & f_{med} Mean and median frequency demonstrated an increasing trend with the increase in level of experience for most of the Myo channels. These metrics showed a significant difference between experts and novices for a few channels of Tasks 2 and 3 (Table 5.3). Low frequencies of action potential of muscles can be associated with continued stress on the muscles [26].

In terms of bandwidth, a negative correlation was demonstrated for most of the channels for the tasks of this study. This observation indicates that expert subjects might recruit a smaller range of muscle fibers; however, for this metric few significant correlations with the level of experience were obtained.

5.4.3   Time-frequency-domain EMG-based Metrics

Four levels of wavelet decomposition were investigated in this study. The variation among different level coefficients for various channels was relatively large. However, the detailed coefficient 2 and 4 (CD-2 and CD-4) demonstrated significant correlations with level of expertise for Task 1 and 3 for the specified channels in Table 5.5. The CD2 was also able to identify the difference between pre-test and post-tests for Tasks 2 and 3 (Table 5.6).

Table 5.3: Spearman's rho correlation with level of experience for the left hand (Task1, Task 2, Task 3). Statistically significant values are shown in bold.

Myo channel	MNP	TTP	f_c	f_{med}
1	(**0.567**, 0.361, **0.494**)	(**-0.517**, -0.389, **-0.445**)	(0.267, -0.057, 0.074)	(0.234, 0.076, 0.111)
2	(**0.601**, 0.278, **0.445**)	(-0.384, -0.403, -0.321)	(0.384, 0.246, **0.482**)	(0.317, 0.284, **0.482**)
3	(**0.601**, **0.445**, **0.494**)	(-0.184, -0.153, -0.235)	(0.100, 0.378, **0.445**)	(0.050, 0.341, **0.408**)
4	(**0.517**, **0.431**, **0.482**)	(-0.217, -0.139, **-0.408**)	(0.300, -0.227, 0.062)	(0.367, -0.132, 0.099)
5	(**0.584**, **0.487**, **0.593**)	(-0.267, 0.083, -0.222)	(-0.075, -0.227, -0.012)	(-0.133, -0.189, -0.099)
6	(**0.484**, **0.639**, **0.667**)	(**-0.450**, -0.056, -0.284)	(-0.083, 0.303, 0.321)	(-0.017, 0.246, 0.259)
7	(0.417, **0.584**, **0.544**)	(-0.35, -0.014, -0.284)	(0.317, **0.492**, 0.222)	(0.384, **0.473**, 0.074)
8	(**0.584**, 0.320, 0.346)	(**-0.534**, -0.348, **-0.507**)	(0.234, -0.095, 0.049)	(0.133, -0.057, 0.148)

Table 5.4: Mean and standard deviation of the MNP metric for pre- and post-tests of novice subjects, for their left hand. The results of Myo channels with statistically significant difference between pre- and post-tests are shown in bold.

Myo Channel	Pre-test mean ± StDev	Post-test mean ± StDev	p value
1	0.084 ± 0.056	0.105 ± 0.058	0.076
2	0.051 ± 0.023	0.070 ± 0.043	**0.045**
3	0.073 ± 0.051	0.099 ± 0.071	0.078
4	0.102 ± 0.061	0.139 ± 0.080	**0.028**
5	0.082 ± 0.061	0.104 ± 0.066	**0.046**
6	0.084 ± 0.056	0.110 ± 0.072	**0.046**
7	0.079 ± 0.051	0.114 ± 0.074	0.054
8	0.133 ± 0.117	0.161 ± 0.118	0.189

5.4.4 Acceleration

Comparing the mean value of acceleration for pre- and post-tests, the novice group showed smaller mean values of acceleration in the post-test. However, this difference was only significant for Task 3. Range of acceleration, maximum acceleration, and consistency of acceleration demonstrated significant improvements for the post-test compared to the pre-test (Fig. 5.6).

The only significant difference between experts and novices was found for Task 1, for both of the left and right hands (Table 5.7). It was noticed that for Task 2 and Task 3, there was a large standard deviation in the expert group with some experts demonstrating lower levels of acceleration metrics, while some experts showed the maximum value among all of the subjects. Since no reasonable cause was identified for excluding them, these samples were included in the analysis. Further investigation of hand movement is required to obtain more knowledge in this regard. A possible solution could be segmenting various parts of the task to recognize the reason behind this variation in the expert group.

Comparing the acceleration metrics with the EMG metrics indicates that trainees can improve their hand motions rapidly in consecutive trials. However, for learning fine gestures, which might influence EMG activity, a larger number of practice sessions is required.

Table 5.5: Spearman's rho correlation with level of experience for CD-4 and CD-2 coefficients of the wavelet transform, considering the left hand for Tasks 1 and 3. Statistically significant values are shown in bold.

Task	Myo channel	CD-4	CD-2
	1	0.117	**-0.551**
1	3	**-0.534**	-0.317
	4	**-0.701**	0.167
	3	**0.432**	0.210
3	5	**0.432**	0.296
	8	-0.358	**-0.457**

Table 5.6: Mean and standard deviation of the CD-2 wavelet coefficient metric for pre- and post-tests of novice subjects, for their left hand. The results of Myo channels with statistically significant difference between pre- and post-tests are shown in bold.

Task	Myo Channel	Pre-test mean $\pm$ StDev	Post-test mean $\pm$ StDev	p value
2	5	3.781 $\pm$ 14.472	-4.554 $\pm$ 17.971	**0.007**
	8	-1.264 $\pm$ 31.863	3.070 $\pm$ 24.612	**0.046**
3	1	22.041 $\pm$ 49.611	-14.853 $\pm$ 63.825	**0.036**

Table 5.7: Acceleration metrics for Task 1. N stands for novices and E stands for experts.

Hand	Group	Acc. Mean (m/s) mean $\pm$ StDev	p value	Acc. Range (m/s) mean $\pm$ StDev	p value	Acc. Max (m/s) mean $\pm$ StDev	p value	Acc. StDev (m/s) mean $\pm$ StDev	p value
Left	N	1.018 $\pm$ 0.022	**0.031**	0.625 $\pm$ 0.217	**0.010**	1.437 $\pm$ 0.162	**0.006**	0.043 $\pm$ 0.015	0.197
	E	1.00 $\pm$ 0.014		0.404 $\pm$ 0.145		1.246 $\pm$ 0.146		0.035 $\pm$ 0.014	
Right	N	0.975 $\pm$ 0.010	1.000	0.524 $\pm$ 0.225	**0.019**	1.375 $\pm$ 0.190	**0.025**	0.050 $\pm$ 0.029	0.133
	E	0.974 $\pm$ 0.008		0.2744 $\pm$ 0.186		1.188 $\pm$ 0.172		0.026 $\pm$ 0.019	
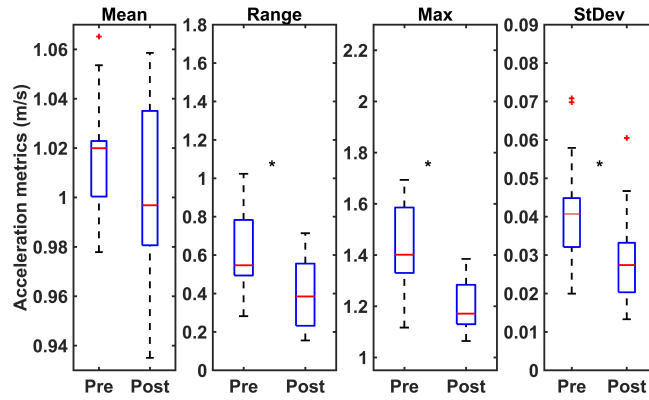
(a)
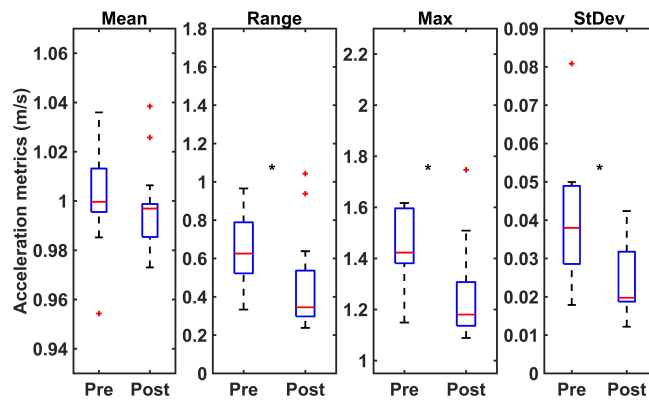
(b)
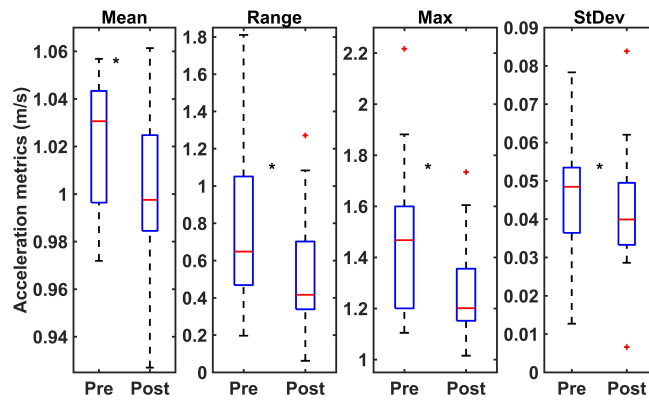
(c)

Figure 5.6: Acceleration-based metrics, including mean, range, maximum (max), and standard deviation (StDev) for pre- and post-tests of the novice subjects, for their left hand: a) Task 1, b) Task 2, and c) Task 3. The metrics with a statistically significant difference between pre- and post-tests are indicated with a * sign.

5.4.5   Feature Selection

The correlation of features with level of experience was calculated for the eight channels from the Myo armbands. Investigating these channels for the time and frequency domain metrics, resulted in choosing different channels for each task. For Task 1, Channels 1, 2, 4, 7, and 8 were selected for the left hand. For Task 2, Channels 1, 2, 4, and 5 were included in the analysis. For Task 3, Channels 1, 2, 3, 7, and 8 were the superior EMG channels. The features that demonstrated correlations higher than 0.5 were included in the final feature set. The number of features that were finally extracted for Tasks 1, 2, and 3 were 38, 19, and 15 for the left hand and were 23, 15, and 14 for the right hand.

5.4.6   Classification

Due to frequent activation of the sleep mode of the Myo armbands, when there was no gesture, the right hand data for some of the subjects were excluded from the analysis. The number of subjects whose data was excluded for Tasks 1, 2, and 3 was 9, 10, and 3, respectively. In order to provide more detailed analysis, a separate evaluation of subject classification was performed for the left and right hands. The results of the classification for the left and right hands separately are presented in Table 5.8. As can be seen, higher levels of accuracy were obtained for the left hand. This was expected as the main part of the task was performed with the left hand and also the numbers of valid data samples were larger, which increases the training power of the SVM model. However, the right-hand metrics also provided accuracy levels higher than 82%. This indicates that the performance of the hand that is indirectly involved in the task might also play an important role in identifying the level of expertise of trainees.

The first two tasks of this study were similar, in the sense that both of them were probing tasks and that the subjects had to press top-actuated switches. The level of difficulty of Task 2 was perceived to be the highest in the opinion of the participants. The results of cross validation demonstrated 100% accuracy for this task. The third task was perceived the easiest task among the tasks studied in this project, according to the participant's impression. The lowest accuracy was obtained for this task. From these results, it can be concluded that the more difficult the

task, the more suitable the EMG and acceleration metrics are for skills assessment.

Table 5.8: Accuracy, precision, recall, and F1 score as percentages based on Myo features for the left and right hands.

Task	Hand	Accuracy (%)	Precision(%)	Recall(%)	F1 score(%)
1	Left	90.48	85.71	85.71	85.71
	Right	82.35	83.33	71.43	76.92
2	Left	100.00	100.00	100.00	100.00
	Right	87.50	83.33	83.33	83.33
3	Left	88.00	100.00	57.14	72.73
	Right	82.61	62.50	83.33	71.43

Regarding the statistical analysis performed in this chapter, a correction to the level of significance to account for multiple comparisons was applied. To this end, Bonferroni and Benjamini-Hochberg correction methods were utilized for p value correction. Using these methods, the correlation with the level of experience was found to be significant for the AR coefficients, cepstral coefficients, wavelet coeffficients, MNP, f_c, f_{med}, and SSC. However, the number of significant correlations was reduced when considering the corrected p values. In addition, classification was repeated using the significant features according to the corrected p values as inputs to the SVM model. It was found in this analysis that classification accuracy was reduced by excluding the EMG features with non-significant high correlations (higher than 0.5). The reduced accuracy in classification shows that valuable information exists in the features that were ignored due to not meeting the corrected p value criteria. Adjusting the level of significance decreases the probability of Type I error (false positive) and increases the probability of Type II error (false negative). In this study, the consequence of performing a Type I error means that no difference is found between subjects with various levels of experience; however, the consequence of making a Type II error results in neglecting the information about

surgical proficiency that might exist in some of the EMG features. The reduced accuracy of classification when correcting the *p* value supports the possibility of making a Type II error. In this study, the objective was to explore a variety of EMG features and recognize the features that are related to surgical proficiency. Consequently, the statistical analysis was performed considering a 0.05 level of significance instead of the corrected value. More powerful statistical analysis is possible with the availability of a larger sample size. This will be pursued in our future work.

5.5   Conclusions

This study evaluated the use of EMG and forearm movements in arthroscopic skills assessment. The use of these metrics was evaluated to differentiate between expert and novice subjects and also to track the improvement of trainees over two trials. According to the results, surgical proficiency affects muscle activity. EMG features such as AR coefficients, histogram, wavelet coefficients, and MNP demonstrate various EMG features related to surgical skill. It was also found that practice has a significant effect on forearm motions, which can be detected by acceleration metrics.

In conclusion, Myo armbands are easy to use devices that can be incorporated into surgical assessment systems. However, the activity of other groups of muscles, such as trapezius, biceps, and triceps, should be investigated in future studies.

Bibliography

[1] R. Berguer, "Surgical technology and the ergonomics of laparoscopic instruments," *Surgical Endoscopy*, vol. 12, no. 5, pp. 458–462, 1998.

[2] M. Uhrich, R. Underwood, J. Standeven, N. Soper, and J. Engsberg, "Assessment of fatigue, monitor placement, and surgical experience during simulated laparoscopic surgery," *Surgical Endoscopy*, vol. 16, no. 4, pp. 635–639, 2002.

[3] T. N. Judkins, D. Oleynikov, K. Narazaki, and N. Stergiou, "Robotic surgery and training: electromyographic correlates of robotic laparoscopic training," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 5, pp. 824–829, 2006.

[4] K. Narazaki, D. Oleynikov, and N. Stergiou, "Robotic surgery training and performance," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 1, pp. 96–103, 2006.

[5] I. Suh, M. Mukherjee, D. Oleynikov, and K.-C. Siu, "Training program for fundamental surgical skill in robotic laparoscopic surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 7, no. 3, pp. 327–333, 2011.

[6] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for emg signal classification," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7420–7431, 2012.

[7] E. J. Rechy-Ramirez and H. Hu, "Bio-signal based control in assistive robots: a survey," *Digital Communications and Networks*, vol. 1, no. 2, pp. 85–101, 2015.

[8] M. A. Oskoei and H. Hu, "Support vector machine-based classification scheme for myoelectric control applied to upper limb," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1956–1965, 2008.

[9] S. Du and M. Vuskovic, "Temporal vs. spectral approach to feature extraction from prehensile emg signals," in *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration*, Las Vegas, Nevada, US, Nov 1–3, 2004, pp. 344–350.

[10] H.-P. Huang and C.-Y. Chen, "Development of a myoelectric discrimination system for a multi-degree prosthetic hand," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Detroit, Michigan, US, 1999, pp. 2392–2397.

[11] R. Boostani and M. H. Moradi, "Evaluation of the forearm EMG signal features for the control of a prosthetic hand," *Physiological Measurement*, vol. 24, no. 2, pp. 309–319, 2003.

[12] G. N. Saridis and T. P. Gootee, "EMG pattern analysis and classification for a prosthetic arm," *IEEE Transactions on Biomedical Engineering*, vol. 29, no. 6, pp. 403–412, 1982.

[13] D. Tkach, H. Huang, and T. A. Kuiken, "Study of stability of time-domain features for electromyographic pattern recognition," *Journal of Neuroengineering and Rehabilitation*, vol. 7, no. 1, pp. 1–13, 2010.

[14] M. Zardoshti-Kermani, B. C. Wheeler, K. Badie, and R. M. Hashemi, "EMG feature evaluation for movement control of upper extremity prostheses," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 4, pp. 324–333, 1995.

[15] M. R. Canal, "Comparison of wavelet and short time Fourier transform methods in the analysis of EMG signals," *Journal of Medical Systems*, vol. 34, no. 1, pp. 91–94, 2010.

[16] R. B. Graham, M. P. Wachowiak, and B. J. Gurd, "The assessment of muscular effort, fatigue, and physiological adaptation using EMG and wavelet analysis," *PloS One*, vol. 10, no. 8, p. e0135069, 2015.

[17] F. Cavallo, S. Sinigaglia, G. Megali, A. Pietrabissa, P. Dario, F. Mosca, and A. Cuschieri, "Biomechanics–machine learning system for surgical gesture analysis and development

of technologies for minimal access surgery," *Surgical Innovation*, vol. 21, no. 5, pp. 504–512, 2014.

[18] A. Boyali, N. Hashimoto, and O. Matsumoto, "Hand posture and gesture recognition using myo armband and spectral collaborative representation based classification," in *IEEE 4th Global Conference on Consumer Electronics (GCCE)*, Osaka, Japan, Oct. 27–30, 2015, pp. 200–201.

[19] M. Sathiyanarayanan and S. Rajan, "Myo armband for physiotherapy healthcare: A case study using gesture recognition application," in *8th International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, India, Jan. 5–10, 2016, pp. 1–6.

[20] F. Haque, M. Nancel, and D. Vogel, "Myopoint: Pointing and clicking using forearm mounted electromyography and inertial motion sensors," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea, Apr. 18–23, 2015, pp. 3653–3656.

[21] M. Chmielewski and M. Nowotarski, "Sensor-based supporting mobile system parkinson disease clinical tests utilising biomedical and RFID technologies," in *MATEC Web of Conferences*, vol. 125, no. 05003.   EDP Sciences, 2017, pp. 1–11.

[22] E. Criswell, *Cram's introduction to surface electromyography*, Jones & Bartlett Publishers, ,Sudbury, MA 2010.

[23] "MyoCraft: EMG in the Bluetooth Protocol." [Online]. Available: http://www.thalmic. com/myo/.htm

[24] D. Robbins, "An introduction to EMG signal processing using Matlab and Microsoft Excel," in *Applications, Challenges, and Advancements in Electromyography Signal Processing*.   IGI Global, 2014, pp. 95–112.

[25] P. Konrad, *The ABC of EMG*.   Noraxon, Scottsdale, 2005, vol. 1.

[26] S. Koneczny and U. Matern, "Instruments for the evaluation of ergonomics in surgery," *Minimally Invasive Therapy & Allied Technologies*, vol. 13, no. 3, pp. 167–177, 2004.

Chapter 6

Concluding Remarks and Future Work

The main goal of this thesis was to improve motor skills assessment in arthroscopy. In Chapter 2, the current state of arthroscopic skills assessment methods was reviewed. In addition, studies in other areas of MIS, which can be related to arthroscopy and/or for which there exists the potential to adapt them to arthroscopy, were reviewed. The important shortcomings of these studies were the lack of an appropriate assessment method that investigates various features of performance, from kinematic to physiological parameters, instead of limiting the assessment to commonly used metrics such as task completion time, path length, etc.

In this thesis, performance metrics were developed inspired by the concept of energy expenditure by using the measured contact force and tip position of the instruments. These metrics are named energy-based metrics; however, the metrics do not represent the amount of energy expenditure. They were inspired by definition of potential energy, kinetic energy, and work. Accurate measurement of energy expenditure requires more complicated analysis to investigate motions of the instruments at different points, an accurate centre of mass calculations, and force sensing along the shaft of the instrument. This is left as future work. The results of this thesis demonstrated the ability of energy-based metrics for arthroscopic and laparoscopic skills assessment. In addition, optimization methods, such as GA and machine learning algorithms, can be used to define expertise criteria and to differentiate between trainees with various levels of expertise. In addition, the muscle activity and the movement of the forearm were investigated to explore the relationship between various features of EMG signals and surgical dexterity. These features showed different muscle activity for experts and novices and

were capable of classifying trainees with 100% accuracy for one of the arthroscopic tasks.

Overall, to develop a comprehensive assessment method, various parameters obtained by measuring instrument motion, hand motion, contact force between the instrument and the surgical setup, and muscle activity should be included. The metrics proposed in this thesis can be calculated automatically in a reasonable amount of time. These metrics can be used to evaluate and determine the proficiency levels of trainees, provide feedback and, consequently, enhance the effectiveness of surgical simulators. Although these metrics were mainly investigated for arthroscopy, they can be adapted to other surgical procedures.

6.1   Contributions

This thesis demonstrated the benefits of incorporating energy expenditure and muscle activity into surgical assessment methods, and their high ability to identify fine variations in performance with different levels of dexterity. The main contributions of this project are as follows:

- In this thesis, novel objective performance metrics were proposed based on mechanical energy expenditure and work. The basic energy expenditure metrics were potential energy, translational kinetic energy, rotational kinetic energy, and work. These metrics showed statistically significant differences between experts and novices for arthroscopic and laparoscopic tasks.

- Another contribution of this thesis was an optimized two-step method for combining the basic energy-based metrics and the use of this two-step method for trainee classification. Specific combinations of the basic metrics were established for differentiating between 1) novices and experts, 2) two sub-levels of novices, and 3) two sub-levels of experts. In this method, the difference between novices and experts was maximized in the first step, and the difference between sub-levels of novices and experts was maximized in the second step. The results showed that the two-step method can increase accuracy when determining the level of expertise of trainees. Dividing this process into two steps allows exploring the fine variations in metrics caused by differences in subjects' detailed levels of expertise.

- Various machine learning algorithms were explored for use in conjunction with the basic energy-based metrics. The goal of using these algorithms was to explore the non-linear dynamics that exist between the energy-based metrics as a function of the various levels of expertise. When exploring arthroscopic skills, the NN and SVM methods were able to more accurately identify levels of expertise, with levels of accuracy as high as 95%.

- Performing a thorough analysis of various EMG features for arthroscopic skills assessment was another significant contribution of this project. The results of this thesis showed that muscle activity can be an indication of surgical proficiency and increase the accuracy of surgical skills assessment. Muscle activity had not been utilized for classifying trainees into their level of expertise. In particular for arthroscopy, these parameters have not been investigated in any related study.

- This thesis explored the use of Myo armbands for surgical skills assessment for the first time in arthroscopic surgery. The data recorded using these armbands and the processing technique proposed here, demonstrated the ability of using muscle activity signals to differentiate between two levels of expertise with 100% accuracy. The results of this study support the idea of performing surgical skills assessment using instruments that are located outside of the surgical site or outside the patient's body. A limiting factor in assessing surgical expertise using parameters such as applied forces is that these assessment methods cannot be transferred to the operating room, due to special preparation and safety issues for patients, or they require specific setups for use on cadaver models. The use of these armbands and EMG-based metrics allows evaluation of transfer validity of the surgical skills learned on simulators to the OR.

  Overall, the methods that were investigated in this project advance our knowledge of the characteristics of dexterous performance and add another perspective to quantifying surgical proficiency.

6.2   Future Work

Various studies can be performed to continue this work in future. Future work can include improving the sensing system, acquiring larger sets of data, and using the current metrics to guide trainees. In the following paragraphs, these ideas are explained in detail.

- One of the improvements in the proposed metrics can be to include axial force when calculating energy. Many parts of the tasks of this thesis required applying bending forces, which are were perpendicular to the shaft of the instrument. However, including axial force in the calculations of the proposed metrics will enhance accuracy of energy expenditure calculation and might result in improved performance metrics. For instance, the Fiber Bragg Sensorized arthroscopic instruments [1,2], that are capable of measuring axial force might assist in the future.

- Recording and analyzing muscle activity of other groups of muscles that might be involved in MIS performance can increase our knowledge of surgical postures and muscle recruitment of experts. The additional muscle groups can include trapezius, deltoids, biceps, triceps, and thenar eminence. Although state-of-the-art features proposed for EMG signals were investigated in this project, analyzing EMG signals through other processing techniques, such as fractal patterns of muscle activity or modeling the signal with HMM, can provide additional information.

- The proposed energy-based metrics were evaluated for four levels of experience for the suturing and knot-tying tasks. However, for the arthroscopy experiment, two levels of experience were investigated. This was due to the small number of experts and intermediates that are available in London, Ontario. Recording a larger number of data samples from participants with an intermediate level of experience can provide additional information on the use of these metrics for arthroscopic skills assessment.

- Investigating the relationship between energy expenditure calculated based on the measured data at the tip of the instrument and the EMG-based metrics will also be beneficial in investigating the efficiency of performance.

- Developing an automatic feedback method based on the proposed metrics and evaluating its effect on learning surgical skills is also an important step for enhancing surgical training.

- Another important point of contact between the instrument and the surgical setup is at the instrument portal. Developing a sensorized trocar or a sensorized instrument capable of recording force along its shaft, will help to acquire more knowledge on the expended energy while performing a surgical task.

- Finally, investigating more complex tasks, such as Bankart suture, might be beneficial for investigating the difference between detailed levels of expertise. For such complex tasks, it is necessary to divide the performance into its sub-tasks. Separating a of a task into smaller portions, such as finding the tip of the instrument, reaching a target, and pressing a switch, can also be helpful for simpler tasks.

Bibliography

[1] B. Poursartip, D. Yurkewich, M.-E. LeBel, R. V. Patel, A. L. Trejos, and M. D. Naish, "Performance evaluation of a sensorized arthroscopic grasper," in *ASME 2015 Dynamic Systems and Control Conference (DSCC)*, Columbus, Ohio, USA, Oct. 28–30, 2015, pp. 1–5.

[2] D. S. Yurkewich, A. Escoto, A. L. Trejos, M.-E. LeBel, R. V. Patel, and M. D. Naish, "Low-cost force-sensing arthroscopic tool using threaded fiber Bragg grating sensors," in *5th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, Sao Paulo, Brazil,, Aug. 12–15, 2014, pp. 28–33.

Appendix A

Sensitivity Analysis of Energy-based Metrics to Additional Mass

The mass in the formulas used to compute the kinetic/potential energy was considered to be a constant equal to 170 g, which is the mass of each instrument. The mass of the needle, which is less than 5 g, was negligible compared to that of the instrument. However, the interaction of the instruments with the setup, directly or through the needle contact, might increase the effective mass of the instruments and as a consequence increase kinetic/potential energy. In this appendix, the kinetic/potential energy, which was calculated based on the constant mass of the instruments, is called the minimum kinetic/potential energy. The following investigation was performed to estimate the effect of the possible additional mass on these metrics. To start, it was assumed that if the instrument's velocity or vertical displacement were small, the effect of additional mass on kinetic or potential energy, respectively, would be negligible. Moreover, if the additional mass was not significant, its effect would be negligible as well. The existence of the additional mass can be shown by the force at the tip of the instrument. Hence, the intervals in which the force was beyond a certain threshold were considered as those in which an additional mass was being carried. During those intervals, if the velocity/vertical displacement was higher than their corresponding thresholds, then the corresponding data could have a significant impact on the kinetic/potential energy (Fig. A.1).

Based on the characteristics of the sensorized instruments, specifically their maximum error and the coupling effect of grasping on bending forces [1], a value of 0.5 N was considered
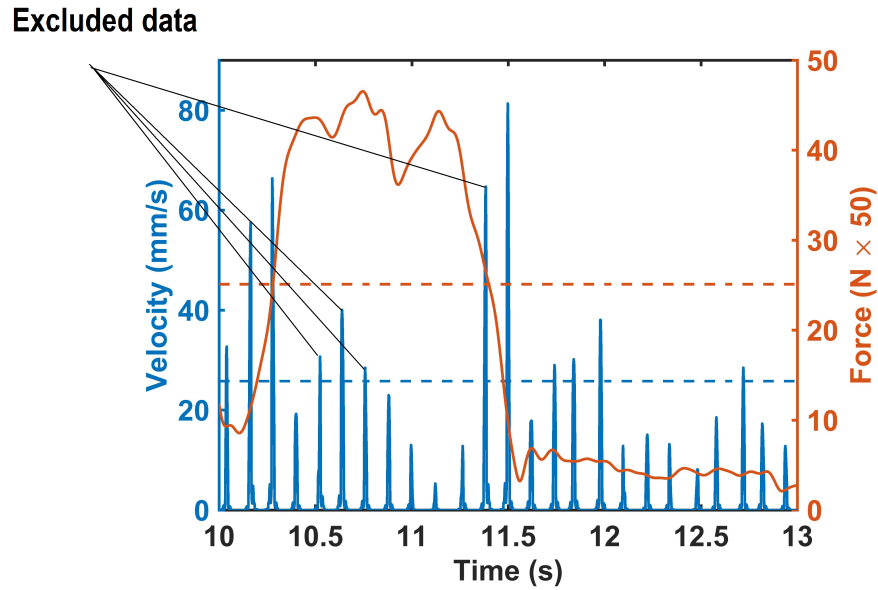
Figure A.1: Sample exclusion criteria. The data points that occur when the force is higher than the red dashed line and velocity is higher than the blue dashed line can have a considerable effect on kinetic energy and were excluded from the data for partial energy-based metrics.

for the force threshold. To avoid the effect of drift on force analysis, the minimum amount of force for each trial was added to the threshold of force for that trial. The vertical displacement or velocity that can produce 1% of the maximum potential energy or maximum kinetic energy were considered as the threshold for vertical displacement or velocity. Consequently, the threshold of vertical displacement was considered to be $0.01 \times \Delta Z_{max}$. Since kinetic energy is proportional to squared velocity, the threshold for velocity was considered to be $0.1 \times V_{max}$.

Using the above-mentioned criteria, the percentage of the data that might be affected by a varying amount of mass was calculated for each trial. The maximum values among the 120 trials for each task and both the left and right hands are shown in Table A.1. The different percentages of the affected data for kinetic and potential energy are due to having different criteria for each of these energy-based metrics. Altogether, the number of affected data samples is small enough to ignore. To confirm this statement, the affected parts of data were removed and the kinetic and potential-based metrics were recalculated. Herein, these metrics are named partial kinetic/potential-based metrics. The correlation of these metrics with the 4 LOEs are shown in Table A.1. To compare the partial metrics with the minimum kinetic/potential-based

metrics, correlations of these metrics with the 4 LOEs are also shown in Table A.1. As can be seen in the table, removing the part of the data associated with the additional mass does not change the relationship between these metrics and the level of experience. A sensitivity analysis was performed by varying the force threshold level by up to 1 N. Using higher thresholds than 0.5 N results in more similar correlations with LOE for the partial and minimum kinetic/potential-based metrics. Based on the above analysis, extracting the parts of data that correspond to possible additional mass does not affect the relationship between these metrics and the LOEs.

Table A.1: Sensitivity of kinetic-based and potential-based metrics to additional mass. The force threshold was considered to be 0.5 N + the minimum force of the trial.

Metric	Task	Hand	Maximum percentage of affected data among 120 trials	Correlation of the partial energy-based metrics with LOE (r, p value)	Correlation of the minimum energy-based metrics with LOE (r, p value)
Kinetic energy	Suturing	Left	6.49%	-0.310, 0.003	-0.302, 0.004
		Right	4.65%	-0.179, 0.090	-0.230, 0.029
	Knot-tying	Left	4.47%	-0.478, <0.001	-0.518, <0.001
		Right	4.01%	-0.484, <0.001	-0.439, <0.001
Potential energy	Suturing	Left	14.04%	-0.351, <0.001	-0.350, <0.001
		Right	12.29%	-0.324, 0.002	-0.320, 0.002
	Knot-tying	Left	12.15%	-0.703, <0.001	-0.677, <0.001
		Right	10.43%	-0.651, <0.001	-0.644, <0.001

Bibliography

[1] A. Trejos, R. Patel, M. Naish, A. Lyle, and C. Schlachta, "A sensorized instrument for skills assessment and training in minimally invasive surgery," *Journal of Medical Devices*, vol. 3, no. 4, pp. 041 002:1–12, 2009.

Appendix B

Ethics Approval

The experimental procedure of Chapters 4 and 5 was approved by the Human Research Ethics Board at Western University as shown in the following figure.

Western Research

Research Ethics

Western University Health Science Research Ethics Board
HSREB Delegated Initial Approval Notice

Principal Investigator: Dr. Ana Luisa Trejos
Department & Institution: Unknown,London Health Sciences Centre

HSREB File Number: 106105
Study Title: Motor Learning in Arthroscopic Surgery
Sponsor: Ontario Centres of Excellence Inc.

HSREB Initial Approval Date: March 03, 2015
HSREB Expiry Date: March 03, 2016

Documents Approved and/or Received for Information:

Document Name	Comments	Version Date
Western University Protocol		2015/02/25
Recruitment Items	Email	2015/02/25
Letter of Information & Consent		2015/02/25

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above named study, as of the HSREB Initial Approval Date noted above.

HSREB approval for this study remains valid until the HSREB Expiry Date noted above, conditional to timely submission and acceptance of HSREB Continuing Ethics Review. If an Updated Approval Notice is required prior to the HSREB Expiry Date, the Principal Investigator is responsible for completing and submitting an HSREB Updated Approval Form in a timely fashion.

The Western University HSREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use Guideline for Good Clinical Practice (ICH E6 R1), the Ontario Personal Health Information Protection Act (PHIPA, 2004), Part 4 of the Natural Health Product Regulations, Health Canada Medical Device Regulations and Part C, Division 5, of the Food and Drug Regulations of Health Canada.

Members of the HSREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Ethics Officer to Contact for Further Information

This is an official document. Please retain the original in your files.

(a)

Figure B.1: Ethics approval from the Human Research Ethics Board at Western University.

Curriculum Vitae

Education

Ph.D. candidate 01/2014 - 12/2017

Electrical and Computer Engineering Department

University of Western Ontario, London, ON

Major: Biomedical Engineering
Dissertation: Motor Skills Assessment in Arthroscopic Surgery by Processing Kinematic, Force, and Bio-signal Data
Supervisors: Drs. Ana Luisa Trejos, Michael Naish, and Rajni Patel

Master of Science 09/2009 - 11/2011

Biomedical Engineering Department

Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

Major: Bioelectric
Dissertation: Modeling the Encoding and Retrieval of Memory Using Hippocampus Microcircuit
Supervisor: Dr. Shahriar Gharibzadeh

Bachelor of Science 09/2004 - 07/2009

Biomedical Engineering Department

Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

Major: Bioelectric
Dissertation: Build and Test of an Artificial Heart Rate Meter Based on a Tactile Sensor
Supervisor: Dr. Siamak Najarian

Publications

- **B. Poursartip**, M. E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Muscle activity analysis for surgical skills assessment" (in preparation)

- **B. Poursartip**, M. E. LeBel, L.C. McCracken, A. Escoto, R. V. Patel, M. D. Naish, and A. L. Trejos,"Energy-based metrics for arthroscopic skills assessment," *Sensors* 17, no. 8 (2017): 1808. .

- **B. Poursartip**, M. E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Analysis of energy-based metrics for laparoscopic skills assessment," *IEEE Transactions on Biomedical Engineering* (2017). Available online: http://ieeexplore.ieee.org/document/ 7932145 (accessed on 19 May 2017).

- **B. Poursartip**, L. C. McCracken, A. Escoto, R. V. Patel, M. E. LeBel, A. L. Trejos, and M. D. Naish, "Development and evaluation of a sensorized shoulder simulator," *In 30th Annual IEEE Canadian Conference on Electrical and Computer Engineering*, Windsor, Canada, May 2017.

- L. C. McCracken, A. L. Trejos, M. E. LeBel, **B. Poursartip**, A. Escoto, R. V. Patel, and M. D. Naish, "Development of a physical shoulder simulator for the training of basic arthroscopic skills," Accepted for publication in *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2017, DOI: 10.1002/rcs.1868.

- **B. Poursartip**, M.-E. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Energy-based metrics for laparoscopic skills assessment." *In 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, August, 2016.

- **B. Poursartip**, D. Yurkewich, M.E. LeBel, R.V. Patel, A.L. Trejos, and M.D. Naish, "Performance evaluation of a sensorized arthroscopic grasper," *In ASME 2015 Dynamic Systems and Control Conference (DSCC)*, Columbus, Ohio, October 2015.

- **B. Poursartip**, D. Yurkewich, M.E. LeBel, R.V. Patel, A.L. Trejos, and M.D. Naish, "Performance evaluation of a custom designed interrogator for an arthroscopic grasper," Poster presented at *London Health Research Day*, London, ON, Canada, February 2015.

Teaching assistance

- Control Systems   Fall 2017
- Introduction to Electrical Engineering   Winter 2017
- Control Systems   Fall 2016
- Introduction to Signal Processing   Winter 2016
- Electric Circuits   Fall 2015

- Introduction to Electrical Engineering   Winter 2015
- Engineering Mathematics   Winter 2009

Research experience

Western University, CSTAR   01/2014 - present

Graduate Research Assistant;

- Analyzed motion, force, and electromyography data for developing performance metrics
- Performed statistical analyses on the variety of the calculated performance metrics
- Applied various machine learning algorithms for surgical skills assessment
- Developed novel performance metrics based on energy expenditure
- Sensorized instruments with Fiber Bragg Grating sensors and calibrated them
- Designed experiments and prepared the required ethics application
- Performed experiments by recruiting subjects with various levels of surgical experience
- Cooperated in developing a sensorized shoulder simulator and the experimental setup

Amirkabir University of Technology   09/2009 - 11/2011

Graduate Research Assistant;

- Developed a program to model hippocmpal CA1 region using NEURON software
- Measured recall performance in different scenarios
- Incorporated neuron's phase relationships with respect to theta rhythm in the model
- Developed a MATLAB code to analyze structure of the neuronal model
- Compared the structure of the neuronal model with small world networks

Amirkabir University of Technology   09/2009 - 11/2011

Undergraduate Research Assistant;

- Building an electronic board for use with a tactile sensor using Altuim Designer software to print the board map
- Developing an AVR code to program a microcontroller
- Simulating the results with the Proteus software

Professional experience

Kavandish System Company, Tehran, Iran   11/2011 - 12/2013

R&D Engineer

- Head of Quality Assurance: evaluated compatibility of the company's product—Electrosurgical Generator— with standards of medical electrical equipment such as IEC 60601-1, IEC 60601-2-2, and IEC 60601-1-2, and performed risk management

- Prepared clinical evaluation of the company products, electrosurgical generators, based on previous literature, previous use of the equipment, and hospital surveys
- Managed the applications for acquiring national certificates and CE marking
- Developed technical files and user manuals for the produced electrosurgical generators

Honors and awards

- Graduate Student Travel Award, Electrical and Computer Engineering Department, Western University, 2017
- 2nd Outstanding Presentation in ECE 3MT competition, Western University, 2016
- Outstanding Presentation in Graduate Symposium, Western University, 2015
- NSERC CREATE Scholarship in Computer-Assisted Medical Interventions (CAMI), Western University, 2014-2015
- Admitted in National Math Olympiad, 2002

Community involvement

Advising experience

- Supervised a summer student on modifying surgical simulators and data processing
- Served as a Robot Design Judge at FIRST Robotics Competition

Reviewer

- Technical reviewer, IEEE Transactions on Biomedical Engineering (TBME)
- Technical reviewer, IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)
- Technical reviewer, ASME Dynamic Systems and Control Conference (DSCC)

Professional memberships

- Member, Institute of Electrical and Electronic Engineers (IEEE)
- Member, IEEE Engineering in Medicine and Biology Society
- Member, IEEE Women in Engineering