QUANTITATIVE MODELLING AND ASSESSMENT OF SURGICAL MOTOR ACTIONS IN MINIMALLY INVASIVE SURGERY

by

Sayra Magnolia Cristancho

B.Sc., Universidad Pontificia Bolivariana Bucaramanga, Colombia, 1999
M.Sc., Universidad de Los Andes, Colombia, 2001

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

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Mechanical Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

November 2008

Abstract

The goal of this research was to establish a methodology for quantifying performance of surgeons and distinguishing skill levels during live surgeries. We integrated three physical measures (kinematics, time and movement transitions) into a modelling technique for quantifying performance of surgical trainees. We first defined a new hierarchical representation called Motor and Cognitive Modelling Diagram for laparoscopic procedures, which: (1) decomposes 'tasks' into 'subtasks' and at the very detailed level into individual movements 'actions'; and (2) includes an explicit cognitive/motor diagrammatic representation that enables to take account of the operative variability as most intraoperative assessments are conducted at the 'whole procedure' level and do not distinguish between performance of trivial and complicated aspects of the procedure. Then, at each level of surgical complexity, we implemented specific mathematical techniques for providing a quantitative sense of how far a performance is located from a reference level:

(1) The Kolgomorov-Smirnov statistic to describe the similarity between two empirical cumulative distribution functions (e.g., speed profiles)
(2) The symmetric normalized Jensen-Shannon Divergence to compare transition probability matrices
(3) The Principal Component Analysis to identify the directions of greatest variability in a multidimensional space and to reduce the dimensionality of the data using a *weight space*.

Two experimental studies were completed in order to show feasibility of our proposed assessment methodology by monitoring movements of surgical tools while: (1) dissecting mandarin oranges, and (2) performing laparoscopic cholecystectomy procedures at the operating room to compare residents and expert surgeons when executing two surgical tasks: exposing Calot's Triangle and dissecting the cystic duct and artery.

Results demonstrated the ability of our methodology to represent selected tasks using the Motor and Cognitive Modelling Diagram and to differentiate skill levels. We aim to use our approach in future studies to establish correspondences between specific surgical tasks and the corresponding simulations of these tasks, which may ultimately enable us to do validated assessments in a simulated setting, and to test its reliability in differentiating skill levels at the operating room as the number of subjects and procedures increase.

Table of Contents

List of Tables

List of Figures

Glossary

Action	Tool motion primitive (e.g., push, pull, sweep, etc)
AIC	Akaike´s Information Criterion
AND	MCMD symbol to indicate that all requirements must be satisfied before proceeding
Atraumatic grasper	Laparoscopic surgical tool for retracting the gallbladder
BIC	Bayesian Information Criterion
Bimanual dexterity	Ability to manipulate objects easily by using coordinated movements between the dominant and the non-dominant hands
Bootstrapping	Method for data resampling with replacement
CA	Cystic Artery
Calot’s triangle	Anatomical structure that contains the cystic artery and the cystic duct
CBD	Common Bile Duct
CBDS	Common Bile Duct Stones
CD	Cystic Duct
CDF	Cumulative Distribution Function
Competence in use of tools	Execution of confident and fluid movements using the surgical tools without damaging any tissue
CTA	Cognitive Task Analysis
Curved dissector	Laparoscopic surgical tool for dissecting anatomical structures
D	KS difference value (0: similar; 1: different)
Decision	MCMD symbol to indicate various possibilities
Difference measure	A score that indicates similarity between subjects’ performances
DoF	Degrees Of Freedom
ETO	Ethylene Oxide sterilization method

FFT	Fast Fourier Transform
Flow of procedure	Appropriate selection and execution of surgical tasks
GB	Gallbladder
HA	Hierarchical Analysis
HMM	Hidden Markov Model
Holding (dwell) time	Time spent at a MM state before transitioning
I	Mutual Information
ICSAD	Imperial College Surgical Assessment System
Intergroup variability	Variation between experiment repetitions from subjects belonging to the different skill level
Intragroup variability	Variation between experiment repetitions from subjects belonging to the same skill level
Intrasubject variability	Variation between experiment repetitions from an individual subject
IP	Information Processing
IRED’s	Infrared Light Emitting Diodes
ISR	Interrupt Service Routine – MCMD symbol to invoke a sub-process while performing a certain process
JSD	Jensen-Shannon Divergence
KL	Kullback-Liebler Divergence
KS	Kolgomorov-Smirnov Statistic
LapChole	Laparoscopic Cholecystectomy
LapCholectomy	Laparoscopic Colectomy
LCG	Laparoscopic Cholangiogram
L-Hook	Laparoscopic surgical tool for dissecting and cauterizing anatomical structures
LUS	Laparoscopic Ultrasonography
MCMD	Motor and Cognitive Modelling Diagram
MIS	Minimally Invasive Surgery

MSD	Mean Square Distance
Option point	MCMD symbol to indicate possibility of ‘jumping’ between parallel branches
OR	MCMD symbol to indicate that a procedure may proceed when at least one requirement is satisfied
OR	Operating Room
OSATS	Objective Structured Assessment of Technical Skills
PC	Principal Component Coefficient
PCA	Principal Component Analysis
Fastrak	Polhemus electromagnetic position tracking system
Process	MCMD symbol to represent a surgical activity
RMS	Root-Mean Square
SMM	Semi-Markov Model
Subtask	Surgical activity that represents manipulation on a single anatomical structure using a single tool (local surgical goal)
Summary measure	Performance measurements using point values such as mean, median, maximum
Task	Surgical activity that represents manipulation on a single anatomical structure using multiple tools (larger surgical goal)
Task analysis	Set of Cognitive Sciences methods to identify key components of complex activities that need to be analyzed when designing training systems
Think-aloud	Task Analysis data acquisition technique to register thought processes while executing a task
TPM	Transition Probability Matrix
Transition	MCMD symbol to link processes and decisions
VR	Virtual Reality

Acknowledgements

I would like to thank all of those who have supported me throughout the course of my PhD. In a very special way to my supervisor Dr. Antony Hodgson for his constant enthusiasm and support since the beginning of my Canadian experience. His valuable feedback, knowledge and expertise contributed the most in the successful achievement of my research goals. I am particularly grateful to Dr. Neely Panton for sharing his extensive knowledge in the area of laparoscopy and for facilitating to me all the related administrative issues of the project. The clinical side of my research was immensely enhanced by his insights, and his professional expertise has become a role model for my future career. I would also like to thank Dr. Karim Qayumi for facilitating all the CESEI resources for the proper development of the project. In addition, special thanks to Drs. Karim Qayumi, Adam Meneghetti and Garth Warnock for their active participation, clinical interpretation and assistance during the OR experiments. Thanks very much to the CESEI team: Vanessa, Ferooz, Marlene, Valentyna for all their support and great company while working with them at CESEI.

Thanks to my Vancouver best friends Jean, Daniel, Jill and Katie, whose friendship I know will remain for a life time. Also to all those St. Johns College friends who taught me that the beauty of Canada resides in the richness of its multicultural society.

In Colombia, very special thanks to Professor Alex Monclou, the boss every worker would like to have. His continuous support since the beginning and until the end of this journey was a main ingredient in the success of this experience. I would like to thank Universidad Pontificia Bolivariana Seccional Bucaramanga as the institution where my professional career has started.

Most of all I would like to thank my family for always knowing exactly the words I needed and sometimes did not want to hear, but which certainly made me grow even more during these PhD years.

Dedication

To those persons who stood by my side
since the beginning of this journey.

Those who never let me step back during
the times in which I was about almost quitting.

Those who always believed that I could do it and
helped me fighting so many times against my own believe.

Those who said me goodbye during an August sunset of 2003
and who greeted me again on a November sunrise of 2008.

Those who I always thought of the last and the first
during every Vancouver night and day.

Those are the persons to whom my heart belongs and
to whom I completely owe the achievement of this dream.

Chapter 1

Introduction

Advances in technology during the second half of the past century changed significantly our conception of general surgical practice. The advent of fiber optics changed the way surgery was being performed and marked a transition towards procedures that seek to avoid large exposure of the patient's inner anatomy [Veelen 2003, Jordan 2000]. Laparoscopic surgery then emerged as a minimally invasive procedure, which is performed using long thin instruments inserted into the body through small incisions in order to operate with minimal damage to healthy tissue.

Reduction in patients recovery time is the major advantage of this type of procedure and has driven an increasing interest in using laparoscopic techniques for a wide range of applications, in spite of the limitations imposed on the surgeon whose motor abilities are hampered due to constraints such as limited degrees of freedom of the surgical tools, loss of depth perception since the 3D surgical field was converted into a two-dimensional viewing of the inner anatomy, increased operative times (approximately 30% longer than standard open procedures) [Berguer 2001], amongst others, as more conventional techniques are switched to laparoscopic ones [Veelen 2003, Nguyen 2001, Berguer 2001]. **Figure 1.1** shows a typical operating room setup at UBC Hospital for laparoscopic procedures.

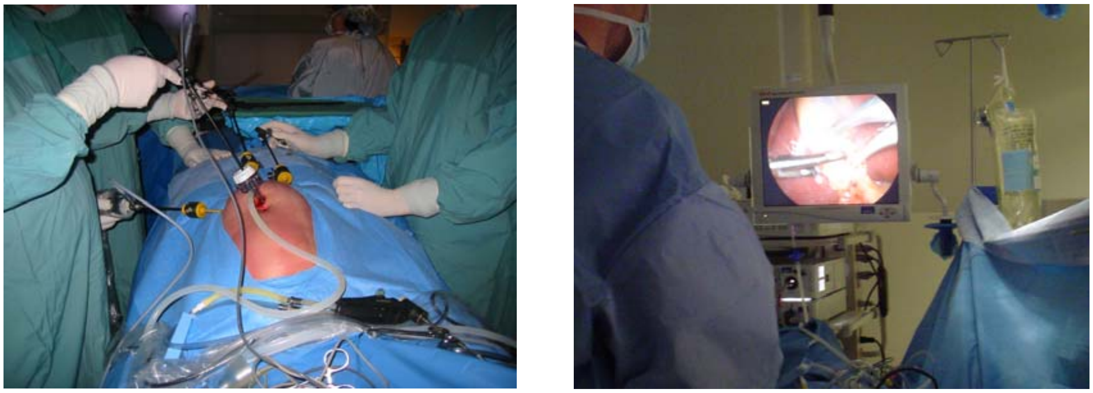


Figure 1.1: Typical operating room set up for laparoscopic procedures at UBC Hospital (Left: insertion of surgical instruments and laparoscope through patient's abdomen. Right: Surgeon's view of inner patient's anatomy).

Because of these technical challenges, it has become more difficult both to acquire and teach minimally invasive surgical skills. A survey conducted in the USA to examine how surgical skills were taught and evaluated in 266 obstetric and gynaecology programs, indicated that most of the residency programs use the operating room (99%) and lectures (88%) for instruction, but only 29% had a surgical curriculum which included bench and animal laboratory training as part of their program. Overall, 79% of programs use subjective evaluation methods to assess skills, which have been shown to often result in poor reliability and validity [Hammond 2006]. However, as constraints on instruction in the operating room increase (e.g., time for OR teaching has dropped by ~20% since 2006 at UBC), the number of medical schools turning towards using simulated surgical scenarios such as animal or cadaveric labs (although there are many outstanding issues – cost, difficulty in reproducing disease, differences with human anatomy, ethical concerns) or physical or virtual reality (VR) simulators, should increase rapidly [Bridges 1999, Babineau 2004, Britt 2007].

Therefore, the constraints imposed on surgeons and the current trend towards reducing working hours and training in the live operating room has led surgical education to face the complex challenge of figuring out how this specialty should be taught and how individual competency should be assessed.

1.1 Monitoring Surgical Training

A survey of surgery program directors carried out in 2001, revealed that 92% of respondents felt there is a need for teaching surgical motor skills outside the operating room [Haluck 2001]. Since then, surgical skill laboratories have been developed by several groups, which designed and built physical and virtual reality (VR) simulators to expose trainees to new skills outside of the operating theatre [Torkington 2001, Gallagher 2001, Scott 2000, Derossis 1998, Rosser 1997]. However, in spite of the fact that physical and VR simulators allow trainers to acquire objective measurements of the trainee and can facilitate the design of step-wise training by controlling or eliminating some variability sources, such as differences in patient's anatomy and disease conditions, there are significant gaps in our understanding of how effective simulators are in developing surgical skill.

There is evidence that people with more advanced surgical skills do better in simulators than novices and that training in simulators improves skill in simulators [Feldmand 2004, Paisley 2001, Gallagher 2001] but there is relatively weak evidence showing that training in simulated environments improves performance in live human surgeries. Grantcharov and Fried found good correlation between virtual reality simulator (MIST-VR and

MISTELS respectively) performances and performance in a pig model of cholecystectomy [Grantcharov 2004, Fried 1999]. They measured individual laparoscopic skills through tasks such as transferring, cutting, clipping, ligating, suturing, and provide scores in terms of speed (faster performance was rewarded with higher scores) and precision (by means of penalty scores per task). An overall score is computed as the difference between the timing score and the precision score. In order to compare performances between the two settings (simulator vs. OR), they tested for correlation between in vitro and in vivo scores [Fried 1999]. While proving initial evidence for transfer of training for basic skills, these studies only provide insights about the final performance at each task but do not allow for tracking performance during the execution of a particular task. In addition, analysis of continuously-acquired measures (e.g., kinematics of the surgical tool) may constitute a better tool for comparing amongst settings than using overall score correlations.

Common problems with studies investigating the issue of transfer of surgical skills include a lack of universal agreement on the most appropriate metrics, lack of a ‘gold standard’ for assessing operating room performance, lack of integration of cognitive and motor skill assessments, and differing skill levels of the participants [Feldman 2004]. Since transference of skill acquisition from simulators to OR has not been completely established, monitoring motor performance in the OR remains the preferred choice of the surgical community [Park 2002]. Currently used performance evaluation methods for the operating room [Alleman 2005, Moorthy 2003, Wanzel 2002, McKenzie 2001, Cao 1999] include direct observation, global assessment and checklists; although these methods have been shown to be effective, they require evaluators to be present in the OR for the entire case in order to track requisite movements and errors and they are time-consuming and therefore costly. Global assessments are not procedure-specific but rate skill using general performance criteria so as to be applicable to different operations without modification. Global assessments are considered the most valid tool for evaluating skill level in the OR at present. Unfortunately global assessments are time consuming, which may decrease the frequency of usage, and they rely on surgeons' opinions, which introduces an element of subjectivity into the assessment which can potentially decrease the reliability of the assessment [Warf 1999, Scott 2000, Smith 2001]. Moreover, this type of evaluation provides limited information for further focused training since it is performed at the 'whole procedure' level and therefore does not distinguish between or focus on specific tasks in the surgery, which the trainer may wish to emphasize. It is also subject to intraoperative variability (i.e., patient's conditions, OR staff, equipment, etc), so reliability is difficult to establish with this type of assessment and multiple assessments may need to be performed to ensure a fair assessment of a trainee's performance [Alleman 2005, Aggarwal 2004].

To overcome the subjectivity disadvantage of checklists, logbooks and direct observation assessments, motion analysis of surgeons' hands or surgical tools has emerged as a promising alternative based on the premise that more experienced or competent surgeons will have greater economy of movements, with fewer wasted motions and greater speed [Moothy 2003, Smith 2001]. Several studies, including some from our laboratory, have shown that it is feasible to acquire such measurements in the operating room using optoelectronic or electromagnetic position tracking systems, although to date such measurements have required dedicated technical support and so have not been used routinely [Aggarwal 2007, Datta 2006, Dosis 2005, Bann 2003, Darzi 2001, McBeth 2002, Kinnaird 2004]. In addition, using tracking equipment produces large amounts of low level data and it is still not clear how such data can be used effectively for assessment and training purposes.

While we do not believe that objective assessment methods will or should replace subjective and nuanced feedback from attending surgeons during the training process, but we do believe they can offer an unbiased evaluation starting-point for evaluation based on quantitative metrics that have the potential to discriminate between skill levels and to detect subtle issues in a given trainee’s surgical technique; we also believe they have the potential to provide specific feedback to the trainee concerning areas in which improvement is needed.

The primary purpose of this thesis, therefore, is to determine whether intraoperatively-acquired quantitative tool movement data can potentially be used to distinguish between levels of training of surgeons learning to perform laparoscopic procedures and to provide insight into specific aspects or elements of their surgical technique that will prove useful for instruction and feedback.

1.1.1 Surgical Assessment Scenarios

To begin to address these questions, we began by trying to understand how surgical educators might wish to use quantitative information if it were available. We began by asking the four surgeon educators involved in this study to outline a number of scenarios in which they might consider using quantitative data and the specific criteria they would like to use in monitoring surgical motor skills in the operating room.

Their input was combined with concepts drawn from the literature [Thomas 2006, Khan 2005]. According to Kahn 2005, areas in which technical skill assessment may be used include evaluating an individual against their peer group to compare relative performance and identify outliers, identifying both those who under perform and so may need extra training as well as those who may be excellent performers within their cohort. In the end, we created a set of 6 scenarios (**Table 1.1**), which described typical uses surgeons might have for a quantified motion assessment system.

We asked the 4 participating surgeons to rate the scenarios from 1 (most important) to 6 (least important). The ordering shown in **Table 1.1** corresponds to the consolidated answers (mean scores) from all surgeons, who were very consistent in the importance given to the first three scenarios. Scenarios 1, 2, 5 and 6 are essentially identical in terms of the analysis required – a single surgeon is to be compared to one or more reference groups. Scenario 3 is similar, but the focus is not on a single surgeon but on the group. Scenario 4 is primarily concerned with group-to-group comparisons. Since 5 of the 6 scenarios rely on comparing an individual to a group, the focus of this study will therefore be on making such comparisons, but we will also consider how to adapt these comparison techniques to support group to group comparisons.

Scenario	S1	S2	S3	S4
Self-monitoring of training Residents are interested in reviewing their own performance to identify particular difficulties and to test themselves against other surgeons' performance. Therefore, each new procedure is analyzed in relation to previous procedures from that surgeon, to their peer group and to expert group performance.	1	2	1	1
Regular monitoring of resident's training Attending surgeon S1 is interested in reviewing the progress of resident R1. Therefore, at intervals S1 analyzes one or more procedures by R1 in relation to previous procedures by R1, by R1's peer group and by an expert group.	2	1	2	2
Annual monitoring of residency program The head of the department of surgery's training program is interested in reviewing the performance of the group of residents from time to time, specifically to identify outliers (either those who are underperforming and may need extra training or those who seem to be excellent performers within their group).	5	3	3	3
Comparison of different training programs An academic society (e.g., SAGES) is interested in assessing quality of training at different geographical regions. Therefore three types of analyses are considered relevant: the performance of residents in region A vs. expert groups from both regions A and B; the performance of residents in region B vs. expert groups from both regions A and B; and the performance of residents in region A vs. the performance of residents in region B.	6	4	5	4
Self assessment and peer assessment (certification) of current practice Expert surgeon S1 wishes to review his/her own or another expert's performance at regular intervals to verify that his/her performance is within expected norms.	3	5	6	5
Writing a reference An attending surgeon S1 is asked to write a reference for resident R1; therefore S1 is interested in reviewing the overall performance of the resident to ensure that he/she is able to perform the basic tasks well and confidently. S1 is particularly interested in identifying common strengths and difficulties apparent during various executions during a period of training.	4	6	4	6

Table 1.1: Surgical scenarios for performance assessment of trainees as described by the four surgeon educators involved in this study. S# indicates surgeon number and rating varies from 1 (most important) to 6 (least important).

1.1.2 Assessment Criteria

We also asked the same group of surgeons to provide a description of the assessment criteria they use or would like to use in the operating room to analyze residents' performance. They agreed on the eight criteria shown in **Table 1.2** similar to those described by Sarker 2006, which we in turn categorized as 'criteria amenable to objective evaluation' and 'criteria to be assessed by other techniques'. We believe that the first two criteria can be assessed by developing a multilevel flowchart of a surgical procedure (see Chapter 2), while the second two criteria can be assessed by evaluating tool movement patterns associated with particular segments of the procedure identified in the flowchart (see Chapter 3).

CRITERIA AMENABLE TO OBJECTIVE EVALUATION
a. Flow of procedure – forward planning (smooth progression without stopping frequently); appropriate selection of subtask order of execution
b. Surgical technique – correct execution of steps in each subtask (subtask-specific checklist); error rate in execution of steps at each subtask
c. Efficiency – use only necessary movements and reasonable amount of time according to procedure difficulty
d. Competence in use of instruments – confident and fluid movements, bimanual dexterity, depth perception (accurate orientation and direction of instruments in the correct plane), appropriate traction, minimal damage to tissue
CRITERIA TO BE ASSESSED BY OTHER TECHNIQUES
e. Organization of the OR – correct equipment and instrument selection, mode, and connection; convenient positioning of equipment to avoid accidents and to facilitate instrument exchange
f. Knowledge of instruments – familiar with names and tasks performed with each instrument; particular selection of instruments
g. Autonomy – appropriate use of assistant, confident decisions with immediate implementation, minimal guidance needed by attending surgeon
h. Quality – achieve desired outcome of the procedure

Table 1.2: Assessment criteria identified by the 4 surgeons involved in this study for describing and measuring laparoscopic surgical performance in the operating room.

1.1.3 Certifying Surgical Practice

A key purpose of objective assessment systems is related to setting standards for promoting trainees. Therefore, data obtained from live surgical practice will hopefully be able to be used to define typical levels of performance for novice, intermediate, and expert surgeons. Reliable means for classifying a given surgeon's typical performance will therefore need to be developed before we can begin to design specific licensing requirements based on quantitative evaluations.

Some authors have observed that measured aspects of operative behaviour can likely be described by some sort of bell curve-type distribution where on the far left would be the 'mavericks' and on the far right would be the surgical geniuses (**Figure 1.2**) [Thomas 2006]. Thomas argued that the aim of assessing surgical competency should be to eradicate the mavericks, to emulate the geniuses and thus to move the whole bell curve to the right by the spreading of best practices (**Figure 1.2**) [Thomas 2006]. To accomplish this goal, we must be able to do what we described in section 1.1.1 – provide a comparison between a given surgeon's motor (movement) behaviours and that of the reference group of surgeons to be emulated and provide instruction to the surgeons so that they can modify their techniques accordingly.

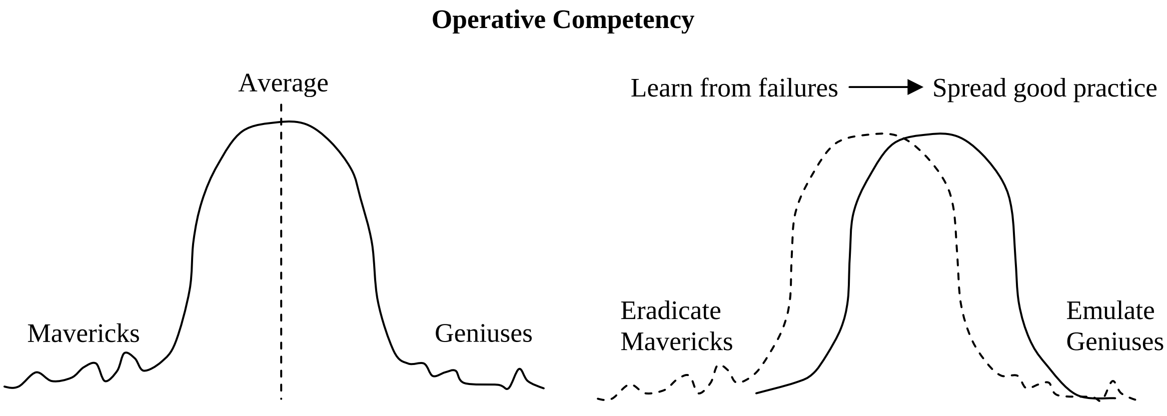


Figure 1.2: Potential normal distribution curve for surgical competence (left) and ideal behaviour of the surgical competence curve after effective training (right) [Thomas 2006].

In the next sections, we will present a literature review of the potential performance measures we might use and how they have been measured and used in the operating room to date. Finally we will introduce our research questions and provide an overview of the rest of the thesis.

1.2 Potential Quantitative Measures

Time (or speed) and accuracy have often been used as measures of performance in the OR [Feldman 2004, Pearson 2002, Risucci 2001] but it is unclear that a task performed quickly does not necessarily mean that it was performed well, so time alone is not a sufficient measure of surgical skill [Datta 2002]. In addition, time is a summary measure and therefore does not provide insight into how a task is executed [Smith 2001]. Other kinds of measures have also been shown to be helpful in making assessments in surgical simulators: kinematics [Aggarwal 2007, Datta 2006, Dosis 2005, Bann 2003, Darzi 2001, Torkington 2001], forces [Rosen 2001, Rosen 2002, Rosen 2006], path length [Ahlberg 2002], distance traveled by the instruments [Hamilton 2002, Jordan 2000], etc; although relatively little work has been done to demonstrate that the patterns found in simulator settings transfer to the live operating room [Hyltander 2002, Satava, 2001].

Amongst the techniques used to measure performance level and differentiate amongst skill groups, the following ones have been published and cited extensively:

a) Physical Trainers:

- Southwestern Center for Minimally Invasive Surgery tasks and ‘Rosser’ tasks: Both task sets include transferring and suturing exercises performed under videoscopic guidance in a trainer box, and performance is assessed by measuring time to execute the tasks [Rosser 1997]

- MISTELS (McGill Inanimate System for Training and Evaluation of Laparoscopic Skills): The original system consisted of 7 tasks performed in a trainer box under videoscopic guidance and each task is scored for precision of performance and speed, with different penalty scores used for each exercise. [Derossis 1998]

b) Virtual Reality Trainers:

- MIST-VR (Minimally Invasive Surgical Trainer Virtual Reality): This VR system is commercially available and has been studied by several groups in Europe and North America. The program consists of 6 tasks of increasing complexity and measurements include time, economy of movement (the distance traveled by the instrument tip past the target), errors, and economy of diathermy use (total burn time) [Hamilton 2002, Jordan 2000]

c) Operating Room Assessment Tools:

- Imperial College Surgical Assessment Device: The ICSAD is a computer program, which processes data from an electromagnetic sensor attached to the surgeon's hand during either simulated or live surgery. This software generates scores of time, number of movements, speed of travel, and distance traveled by each instrument during completion of the task. Besides simulator, this system has also been used in a human operating room, and therefore it is the most extensively studied to date. Some validity studies has been carried out with ICSAD and has shown construct validity[1] as time and number of movements discriminated senior from junior surgeons when executing an open vascular surgical simulation [Aggarwal 2007, Datta 2006, Dosis 2005, Torkington 2001, Smith 2002].

- Advanced Dundee Endoscopic Psychomotor Tester (ADEPT): In this physical simulator, laparoscopic graspers are equipped with sensors to measure angular deviations, and the target plate can also measure errors like excessive force. Various tasks, involving manipulation of switches and dials can be performed under videoscopic guidance and the number of tasks successfully completed, total time, and errors are used as performance measures [Francis 2001, Macmillan, 1999].

- Forces and torques exerted by the tools on operative tissues have also been examined; both in the form of grip force and tool tip forces [De Visser 2002]. Rosen's group at the University of Washington has done extensive work using

[1] Construct Validity differentiates between skill levels. It is the most common validity test applied to surgical simulators, where an expert should show a marked improvement over a novice's performance on analogous tasks.

force/torque signatures measured at the hand/tool interface to evaluate performance in a porcine model. Rosen uses Markov Modeling on the whole procedure force data stream to show the feasibility of correctly classifying surgeons into two experience levels based on the similarity of the models representing a given surgeon's low-level tool-tissue interactions to models derived from reference groups representing the two experience levels. They demonstrated that the forces and torques applied by experts and novices differed, as did the time to complete the procedures [Rosen 2001, Rosen 2002, Rosen 2006]

Although current simulators have been shown to be a valid tool for training novice surgeons in basic psychomotor skills [Park 2002, Grantcharov 2001, Ahlberg 2002, Hyltander 2002] as performed and assessed in a simulator, their ability to provide valuable guidance at more advanced levels of training has not been established. In particular, current simulator technology cannot yet represent the wide range of variability seen in patients in the operating room. In addition, while it is clear that low scores on particular simulated tasks suggest that more practice might be required, there has been virtually no work done on using intraoperatively-acquired data to identify suboptimal performance that can be linked to [Feldman, 2004].

In the present study, we concentrate on integrating three types of physical measures (tool kinematics, time and movement transitioning) into a modelling technique for quantifying performance of surgical trainees while performing at the operating room.

1.3 Quantitative Assessments in the Operating Room

As described above, the vast majority of quantitative studies have been conducted in surgical simulators. To our knowledge, only three main types of performance evaluation methods have been tested at the operating room. The advantages and disadvantages of each are presented so as to assess the potential of each method and the current trend of research in surgical skill assessment.

- Surgical observation relies on expert surgeons' qualitative opinions about trainees' performance during individual procedures. In an attempt to add some objectivity to this method and to provide a framework for describing procedures, Cao 1999 introduced the notion of surgical task decomposition using video analysis to identify activities and motions in endoscopic procedures. This group developed a hierarchical decomposition primarily for the Nissen Fundoplication procedure by decomposing the overall procedure into tasks, then tasks into sub-tasks and sub-tasks into component motions. Measures of time spent to complete each task and subtask, the types of motions and the number of times each motion was used were recorded [Cao, 1999 MacKenzie, 2001]. Major issues in using this technique included logistical problems of scheduling trained observers, patient variability, unreliable evaluation due to observers using subjective criteria, time spent and lack of agreement by reviewers not only during direct observation but also when using video assessment [Alleman 2005].

- The University of Toronto's OSATS (objective structured assessment of technical skills) is to date the only standardized method capable of evaluating procedures in the operating room, although it has been used more extensively in animal and simulator settings; it uses a global rating system, checklists, and time measurements [Wanzel 2002, Martin 1997]. Performance during execution of tasks is assessed using checklists specific to the operation or surgical task and a global rating scale. The global scale is composed of seven variables that represent operative skill. A reviewer uses a 5-point scale to evaluate every variable. The middle and the extreme points are described using the following rubric to help the assessor assigning consistent scores (**Table 1.3**).

	Rating				
Variable	**1**	**2**	**3**	**4**	**5**
Respect for tissue	Often used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage
Time and motion	Many unnecessary moves		Efficient time and motion but some unnecessary moves		Economy of movement and maximum efficiency
Instrument handling	Repeatedly makes tentative or awkward moves with instruments		Competent use of instruments, although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Knowledge of instruments	Frequently asked for the wrong instrument or used an inappropriate instrument		Know the names of most instruments and used appropriate instrument for the task		Obviously familiar with the instruments required and their names
Use of assistants	Consistently placed assistants poorly or failed to use assistants		Good use of assistants most of the time		Strategically used assistants to the best advantage at all times
Flow of operation and forward planning	Frequently stopped operating or needed to discuss next move		Demonstrated ability for forward planning with steady progression of operative procedure		Obviously planned course of operation with effortless flow from one move to the next
Knowledge of specific procedure	Deficient knowledge. Needed specific instruction at most operative steps		Knew all important aspects of the operation		Demonstrated familiarity with all aspects of the operation

Table 1.3: Global Rating Scale from OSATS.

The OSATS offers several benefits toward the assessment of technical competence as it provides standardized assessment criteria, it is portable with good reliability and validity when administered in different medical centers and it could be used to track the progress of individual residents through training and offer valuable feedback for future improvement as the criteria have been shown to discriminate between levels of training [Wanzel 2002, Regehr 1998]. However, important drawbacks in using the OSATS method are the resources (expensive and logistically complex to administer) involved in getting expert surgeons to observe the performance of trainees. Therefore, the surgical community would greatly benefit from evaluation systems capable of assessing technical skills in real time without needing to rely so heavily on expert observers [Sidhu 2004, Moorthy 2003].

- Motion analysis-based systems have recently arisen as a potential alternative to observer-based methods. They rely on the concept that motions become more efficient with level of training, therefore motion measures will reflect skill levels. Darzi’s group at Imperial College London has demonstrated that by using electromagnetic sensors, surgeons’ hand movements could be tracked and analyzed. They have shown a strong correlation between previous laparoscopic experience on a simple task in a box trainer and dexterity in more complex tasks such as laparoscopic cholecystectomy on a porcine model [Darzi 2001]. In those studies, expert surgeons proved to be more economical in terms of the number of movements and more accurate when approaching specific surgical targets, which allowed them to utilize shorter paths [Aggarwal 2007, Datta 2006, Dosis 2005].

We regard this approach as one of the most promising for the future of surgical skill assessment; however, we also believe that some issues need to be addressed before the system will be useful for instruction. Darzi's approach considers major parts of the entire surgical procedure (e.g., Calot's Triangle dissection) without further task decomposition, which therefore, does not allow identifying and distinguishing critical aspects at subcomponents of the procedure. A thorough understanding of the flow of the procedure in terms of a hierarchical and sequential representation of motor and cognitive activities is necessary to (1) identify causes of deviations in the normal path due to individual surgeon's decisions, (2) to describe how the surgical tools are actually used following a standardized and structured framework of the procedure, and (3) to take account of operative variability by allowing for variable weighting on different tasks during a surgical procedure to reflect differences in importance, difficulty or relevance for the current level of surgical training. Decomposing surgical procedure into simpler tasks will also facilitate providing relevant feedback to the trainer by focusing the assessment on problematic areas of the procedure. In addition, it is still an open question whether position data from the back of the surgeon's hand is substantially equivalent to that of the tool itself. Since we would ultimately like to include force data describing tool-tissue interactions, we propose to acquire tool tip data.

Our group has previously demonstrated the feasibility of acquiring intraoperative measures of tool motion. Two previous studies [McBeth 2002, Kinnaird 2004] used a Northern Digital Polaris Hybrid Optical Tracking System to track the 3D position of both active infrared light emitting diodes (IRED's) and passive retro-reflective markers in order to acquire postural data and tool tip trajectories. The major disadvantage of the optical systems is that they suffer from occlusions in the line of sight between the camera and the markers. McBeth found that due to the complicated OR set up, the line-of-sight issue prevented him from obtaining reliable data from some of the procedures; therefore, Kinnaird improved the tracking method by incorporating an electromagnetic system into the tool tracking system. However, this proved to be somewhat cumbersome for the surgeons to use. Nonetheless both studies successfully proved that tracking systems could be used in a reasonably practical manner to record surgeons' motion data while performing live surgical procedures.

1.4 Research Questions

Our approach to assessing surgical motor performance is based on video analysis and position measurements of the surgical tool movements. We use a hierarchical decomposition to represent 3 levels of surgical procedure organization (Task, Subtasks, and Actions) and attach time, kinematic and state transition measures to each node of the Motor and Cognitive Modelling Diagram (MCMD). We use a variety of techniques to evaluate differences between individual surgeons and reference groups and use a dimension reduction technique to make the resulting analysis more comprehensible to the trainees and trainers.

The overall goals of the present research are to layout and evaluate the feasibility of a novel assessment framework and methodology for quantifying and assessing the psychomotor performance of surgeons during live surgeries.

We therefore concentrate on answering the following specific research questions:

1. Can quantitative measures acquired intraoperatively reliably characterize motor performance?
2. Do surgeons at similar stages of training exhibit similar psychomotor patterns?
3. Is there a clear separation of patterns between the extremes of the training spectrum?
4. What data/measures are most useful in separating surgeons along this spectrum?
5. Can a quantitative analysis produce insights useful for instruction?

In implementing our approach, two key elements are developed: (1) a ‘language’ for modeling surgical procedures, and (2) techniques for representing and processing the quantitative data.

For the first part, the methods include extensive observations of the actual laparoscopic cholecystectomy (LapChole) surgical procedure used by attending surgeons at the University of British Columbia, followed by multiple interviews and repeated applications of ‘think-aloud’ techniques in the operating room. This cyclic process and a validation study allow us to define a set of general symbols which are used to construct our MCMD diagram for describing laparoscopic procedures.

For the second part, we track the motions of two surgical tools which are commonly used in the most important phases of the LapChole procedure (Curved Dissector and

Atraumatic Grasper). We use a Polhemus 3SPACE Fastrak 6-dof electromagnetic system to measure tools' position and orientation data. The resulting data stream is segmented into tasks, subtasks, and actions based on the times identified in the video analysis for the MCMD. We then apply Principal Components Analysis (PCA) to extract the dominant contributors to overall variability and to visualize motor performance as function of level of surgical training. In addition, we assign measures of dissimilarity (Kolgomorov-Smirnov measure and Jensen-Shanon Divergence) between kinematics and time profiles and transition probability matrices in order to compare an individual surgeon's performance to that of different reference groups such as peers or experts. These measures provide a more complete representation of the tool use patterns than would be possible if using summary or average measures only.

1.5 Thesis Layout

The present thesis has been structured as follows:

Chapter 2 describes our approach to decomposing laparoscopic surgical procedures by developing a standardized and structured framework to describe the organization and progression of a surgery. This motor MCMD representation is developed in the context of laparoscopic cholecystectomies and we show how the notation is sufficiently rich to represent a second laparoscopic procedure – laparoscopic colectomy.

Chapter 3 presents the proposed general assessment methodology, including extensive descriptions of the implemented data acquisition system and a detailed explanation of how we derived and transformed our performance measures into difference scores at the various hierarchical levels of our MCMD representation.

Chapters 4 and 5 describe the results of implementing our proposed methodology in a physical simulation study (chapter 4) and in an intraoperative study (chapter 5).

Chapter 6 summarizes the findings of the thesis by highlighting the main contributions and proposes future studies using the methodology developed here to test its reliability in comparing large datasets (multiple procedures) of surgeons and to determine any correlation between assessment on simulators and assessment at the operating room.

1.6 Contributions and Significance

Given that most testing scenarios for performance assessment make use of animal models or simulators, we believe that one of the most significant contributions of the present research would be to become the first group in North America to develop and apply an assessment methodology for measuring surgical performance in a human operating room environment. By means of our original motor and cognitive modelling diagram (MCMD) we expect to take account of interprocedure variability at various hierarchical levels by decomposing the procedure into individual tasks to which we would be able to attach performance measures segmented from a continuous data stream, as well as derived difference scores, in order to identify critical aspects of the overall surgery for which instruction should be emphasized. Through the two experimental studies, we intend to demonstrate the feasibility of our methodology by showing that these measures can differentiate between skill levels. In addition, we believe that the ‘difference measure’ concept we introduce here, which produces normalized difference scores lying between 0 (similar) and 1 (different) regardless of the units measured, will contribute to defining scoring scales that the broader surgical education community will find useful and intuitive to use.

A successful demonstration of this quantitative assessment approach could potentially become a standard component of intraoperative surgical skill assessment protocols following two further developments: (1) improvements in automatic data segmentation and analysis to make the method more practical in day-to-day applications, and (2) application to a broader range of subjects from multiple institutions to enable construction of a larger and more reliable dataset.

Chapter 2

Task Analysis in Minimally Invasive Surgery

2.1 Introduction

As described in the previous chapter, the surgical motor performance by a given surgeon during live surgeries can vary significantly from procedure to procedure due to differences in the patient, the surgical team and the equipment available. Relatively random events, such as the occurrence of unexpected bleeding, can produce a significant diversion in the course of the procedure, and differences in patient anatomy can require variations in technique or even different surgical steps to be performed [Ignjatović 2006]. Some patients have aberrant anatomy or other complications (e.g., scarring from previous surgeries) that simply make accomplishing the surgical goal more difficult and time-consuming [Ding 2007, Larobina 2005]. For these reasons, it is difficult to monitor a given surgeon's surgical performance across patients or to compare a given surgeon with a reference group either of peers or experts based on whole task measures alone.

It is therefore necessary to develop a means of representing the flow of surgical procedures that can capture the variations in procedures that exist in real patients. In that way, we can evaluate surgical motor performance at a more detailed level where the tasks being executed are more directly comparable and less affected by inter-patient or inter-procedure variability. For example, if an operation involves a suturing task, it is reasonable to evaluate their suturing ability by focusing on this task, regardless of how much surgical effort was required to get to this stage in each particular patient, rather than trying to infer how well they suture based on measures such as the overall time taken for the procedure (which may include 15 minutes spent trying to deal with unexpected bleeding). Similarly, the way a surgeon moves their tools when performing a blunt dissection task is likely quite characteristic of their current level of surgical training and skill development, regardless of which specific patient they are currently operating on.

To make these ‘in-context’ sorts of comparisons, therefore, we need to develop a ‘language’ for describing the flow of a surgical procedure. At one level, a surgical procedure is simply a long series of tool movements that can be described as a position vs. time history (or force vs. time). However, such a description ignores the cognitive aspects of a surgical procedure; it does not express the goals the surgeon is trying to achieve and does not distinguish between the main tasks they are performing and adaptive responses to unforeseen events such as bleeding, nor does it convey any sense of progress through the procedure. To capture such aspects of surgery, we need to overlay onto these continuous data streams a description of the goals the surgeon has in mind. In short, we need to add meaning and context to the data stream. Therefore, key features required by such a language are the ability to represent various levels of detail, sequencing/flow (including loops and branching/options, as well as order-independent tasks – i.e., those that can be done in any order, but which all must be done before proceeding), decision points, and interruptions (suspend and resume) [Bittner 2004, Berber 2001, Reddick 1993, Cuschieri 1990].

In previous work, some authors have found it helpful to describe a surgical procedure in a hierarchical form. In order to establish differences between novice and expert surgeons in terms of the executed steps during a procedure, Cao 1996 introduced the notion of surgical task decomposition using video. They developed a hierarchical decomposition for the Nissen Fundoplication surgery by decomposing the whole procedure into tasks, then tasks into sub-tasks and sub-tasks into component motions. Later, they proposed representations for inguinal repair and laparoscopic cholecystectomy procedures [Cao, 1999]. Each activity was limited by operational beginnings, endings and target states and a measure of time to complete each activity was used as the performance measure.

Afterwards, McBeth 2002, developed an alternative hierarchical decomposition based on Cao's structure but modified it to improve generality and to incorporate additional kinematic features of low-level tool movements. He defined the following five levels (**Figure 2.1**):

1. The phase level outlines the global goals of the procedure, which are likely to be invariant regardless of who performs the procedure.
2. The stage level outlines local goals required to complete each phase; the stages are usually similar in most patients and are usually carried out in a standard order, although there may be some variation in ordering, depending on the procedure (for example in laparoscopic cholecystectomy surgeons have the choice between

performing an ultrasonography test prior to or during the surgery. At the stage level, the surgeon may have to use multiple tools to accomplish the surgical goal.

3. The task level is similarly defined in terms of a discrete surgical goal, but generally involves the use of only a single tool (or pair of tools in the surgeon's two hands). There is generally the sense that tasks are to be performed sequentially according to a predetermined plan, although McBeth did not explicitly model the flow of a procedure.
4. The subtask level describes the set of sub-goals the surgeon uses a single tool to achieve. At this level, there is more possibility for the sequencing between subtasks to depend on patient-specific factors; there is also more of a notion of cycling between subtasks until the larger goal or sub-goal is achieved, although again such cycling was not explicitly modeled.
5. Finally, actions describe low-level surgical gestures (i.e., individual tool 'movements') such as reaching or sweeping that are used to accomplish a higher-level surgical goal. At this level, there is no notion of a surgical goal or of forward progress; the individual gestures are viewed as states that the surgeon cycles between until the higher-level goal is achieved.

This five-level hierarchical decomposition was designed to provide a foundation for quantitatively analyzing surgeons performing a standardized version of a common minimally invasive procedure (in this case, laparoscopic cholecystectomy).

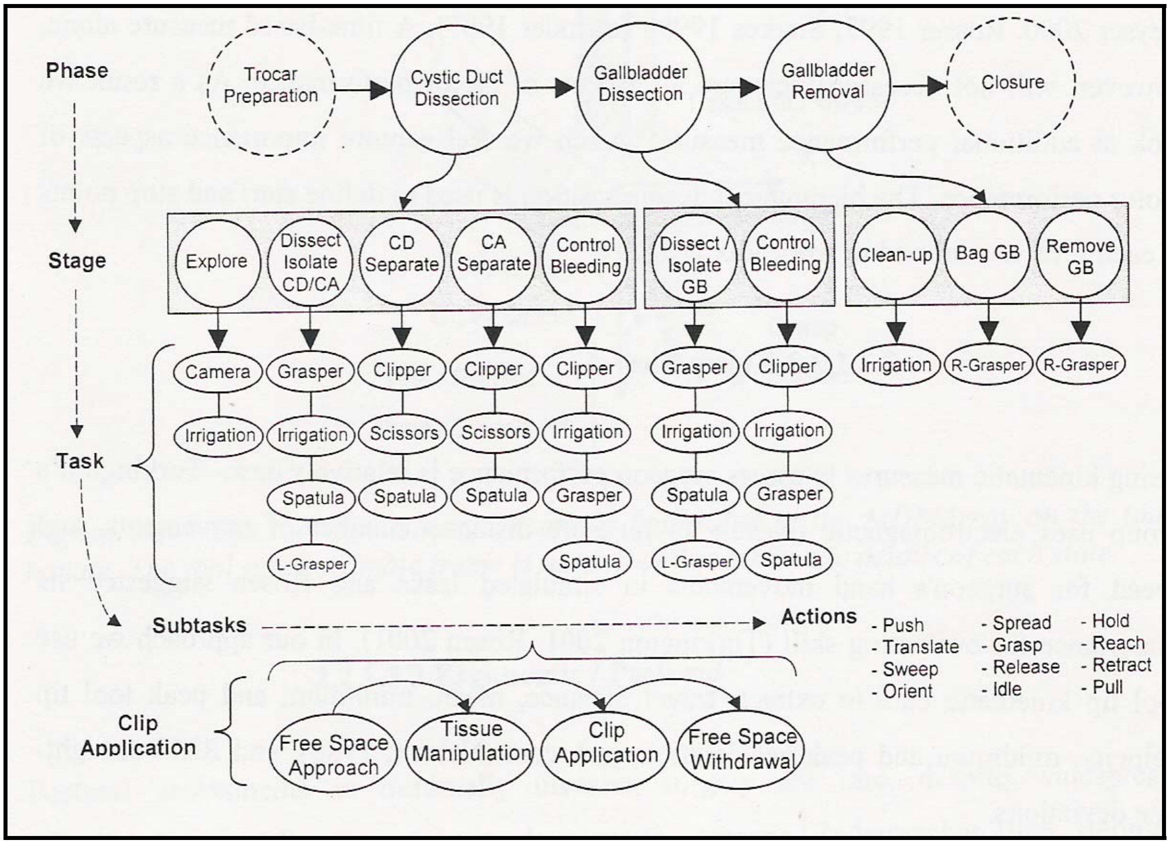


Figure 2.1: Hierarchical decomposition of the laparoscopic cholecystectomy by McBeth 2002.

While helpful in understanding the main tasks that make up selected surgical procedures, both Cao and McBeth's approaches concentrated exclusively on describing motor activities using a hierarchy. Hierarchical decompositions assume that all processes at a given level proceed sequentially and do not reflect the branching and decision points that occur in real procedures (see **Figure 2.2**). Neither representation, therefore, is adequate for representing both motor and cognitive aspects of surgery in a unified framework. Other studies have also tackled the issue of representing minimally invasive procedures (MIS) for assessment purposes; however, as will be explained in the next section, while they have identified some key elements, they have not integrated a consideration of flow and decision-making.

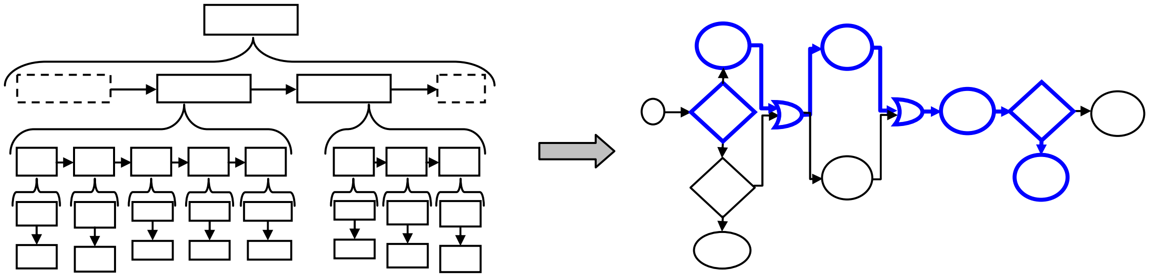


Figure 2.2: The hierarchical decomposition of McBeth does not directly represent event sequences and decision points.

2.1.1 Task Analysis for Representing Minimally Invasive Surgical Procedures

Task analysis is an important tool in the cognitive sciences field, which can be used to identify key components of complex activities, such as surgery, that need to be analyzed when designing training systems. From the technological perspective, the application of this methodology aimed to provide the foundation for constructing suitable settings for the practice of technical skills while taking into account human needs, behaviours and limitations. Potential benefits are reductions in latent human errors caused by lack of experience or cognitive processing limits, which in the surgical setting may lead to undesirable complications for the patient [Stone 2004, McCloy 2001].

An important example of the application of task analysis in the development of surgical simulators is the MIST VR system. An ergonomic evaluation of psychomotor skills was performed in the operating room leading to the definition of a set of simplified tasks which collectively represent the skills required while executing the main steps of a real laparoscopic procedure (e.g., holding tissue, separating tissue and vessels, left hand and right hand instrument control, etc) [Stone 2004, McCloy 2001].

However, since there is a growing need to establish ways to reliably assess surgical performance in real settings, it has become imperative to perform a more extensive analysis of the constituent parts of a procedure. This has led engineers to develop structural methods that provide a standard framework that can objectively describe the flow of the procedure. Although only a few research groups have addressed this issue to date, the current approaches have been well-received [Sullivan 2008, Sarker 2006].

Cao et al 1996 introduced the notion of surgical task decomposition using video analysis to identify activities and motions in endoscopic procedures as a way to establish differences between novice and expert surgeons. They developed a hierarchical decomposition in terms of three levels: tasks, sub-tasks and component motions.

Four basic surgical tasks were identified for the Nissen Fundoplication procedure: dissecting tissue, suturing, tying knots, and cutting sutures, for which high levels of skill are required. Measures of time spent to complete each task and subtask, component motions and the number of attempts for each of the component motions to achieve the task goals, were recorded as performance measures through a qualitative description of the end-effector's movement characteristics and a simple scoring of the number of repeated attempts made by the surgeons.

Suturing was found to be the longest and most involved of the four surgical tasks, followed by tying knots, dissecting tissue, and cutting tissue. They also found that two tasks (cutting sutures and dissecting tissue) shared the same sub-task decomposition (pull taut object and snip object), but were differentiated by the time spent due to the particular subtask requirements and constraints (mainly precision and safety) and the object (tissue or suture) to be divided, which determined the degree of difficulty as expressed by the number of motions performed. Five distinct basic motions were identified: reach & orient, grasp & hold/cut, push, pull, and release

In 1999, Cao et al. extended their hierarchical decomposition to describe other procedures: laparoscopic cholecystectomy, inguinal repair and nissen fundoplication. In this study, they highlighted the importance of performing surgical task analysis for assessing new technology by measuring its impact on surgical skill acquisition and performance.

Observational research was carried out to qualitatively describe the procedures in terms of increasing levels of detail, from high-level surgical steps down to sub-steps, tasks, subtasks, and what they called motions. For laparoscopic cholecystectomy, which is our reference procedure because it is the most commonly performed laparoscopic procedure and the one residents first learn to perform, four surgical steps were identified: 1) prepare patient, 2) isolate gallbladder, 3) remove gallbladder, and 4) close; where 2) isolate gallbladder and 3) remove gallbladder are the major steps. As will be shown in section 2.3, this decomposition matches well with our analysis.

A third study carried out by this group [MacKenzie 2001] and presented as an integration of the two previous ones used a triangle strategy to study user, task and tool in the context of the operating room (OR) environment, in contrast to their previous studies mainly conducted using pig models. At this point they extended and developed a detailed hierarchical framework for Nissen Fundoplication procedures. One interesting issue during this decomposition focused on the comparison of whether or not surgeons divided the short gastrics. Although it is clear that the presence or absence of a certain step could be a way to assess experience, they realized that the hierarchical decomposition approach is limited in its ability to represent the human cognitive processes underlying task performance since it is exclusively based on the observable functional aspects of the task. Therefore we believe that it is necessary to complement it with a representation of the flow of the procedure, which will allow us to represent surgical decision points and variations or adjustments in technique. These are the types of issues that we intend to contribute with our study and that will be explained in detail in section 2.3.

In another approach not based on structured task decomposition, Berber 2001 critically analyzed the intraoperative time utilization for laparoscopic cholecystectomy and identified the most important moments of the procedure. They found that the operation time may be divided into seven parts: trocar entry, laparoscopic ultrasonography, dissection of the triangle of Calot, cholangiogram, dissection of gallbladder, placement of gallbladder in the endobag, and irrigation-suction process & removal of ports. Even though this study did not suggest any hierarchical analysis of the seven identified phases, it served as a basis of comparison for our study. In particular, it suggested to us the need to include two issues we had not originally identified (i.e., laparoscopic ultrasonography and cholangiography) in order to make our representation more general.

Two recent approaches have added additional insights about the importance of having a standard diagrammatic representation for surgical procedures. Weigmman 2007 studied surgical errors and their relationship to flow disruptions in cardiovascular surgery. Based on a conventional surgical protocol, they identified two types of errors: immediate vs. delayed capture errors and found that those errors that were captured immediately were more likely to be detected by the same person who committed the error than were events captured after a delay. In terms of the five categories of flow disruption defined in this study (teamwork, extraneous interruptions, equipment/technology, resources-based issues, supervisory/training-related issues), teamwork/communication accounted for the highest percentage (52%) of occurrence of disruptions and constituted the strongest predictor of surgical errors. Since this study was performed at a high level of surgical description and no exclusion criteria for patients and surgeons were applied, disruptions and errors associated with specific surgical steps could not be identified. However, this study suggested to us the need to include a way to represent flow disruptions [Weigmman 2007].

In order to explore the cognitive side of surgical procedures, Sullivan carried out a study to investigate if cognitive task analysis (CTA) could capture steps and decision points that were not articulated during traditional teaching of a colonoscopy. They created a procedural checklist and a 14-point scale for measuring cognitive demands, and made use of the ‘think-aloud’ technique for recording descriptions from three expert surgeons about what/how they performed. They found that the surgeons omitted explaining more than 50% of the essential steps and critical decisions, which supports the notion that expertise is highly automated and that during difficult cases, surgeons tend to stop explaining because they become worried about committing an error [Sullivan 2008]. The findings of this study also support our contention that it is important to document procedural descriptions from multiple expert surgeons in order to standardize representations of cognitive aspects of surgical procedures in order to help trainers articulate the critical aspects of the surgery prior to going into the OR. Additionally, correlating critical decision points with specific surgical steps would help surgeons to devise alternative strategies for ensuring the patient’s safety.

From a technical point of view and with the aim of providing a qualitative and objective analysis of laparoscopic procedures, Rosen 2001 hypothesized that: 1) haptic information and tool/tissue interactions performed in laparoscopic surgery are skill-dependent, and 2) statistical models (Hidden Markov Models – HMMs) representing these interactions are capable of objectively evaluating laparoscopic surgical skills. The method they chose to use was based on an instrumented grasper equipped with F/T sensors at the hand/tool interface and a standardized seven-step procedure performed on a pig model.

Although a formal hierarchy was not developed, they found 14 unique force/torque signatures representing 14 types of tool/tissue interactions that they grouped into three broader types based on the level of force/torque interaction (**Table 2.1**). They then analyzed the procedure as continuous cycling between the 3 states but did not include any notion of decision points or progression. Those states are similar to the ‘actions’ described by Cao.

Type	State name	State Acron.	Force/Torque						
			Fx	**Fy**	**Fz**	**Tx**	**Ty**	**Tz**	**Fg**
I	Idle	ID							
	Grasping	GR							+
	Spreading	SP							-
	Pushing	PS			-				
	Sweeping (lateral retraction)	SW	±	±		±	±		
II	Grasping – Pulling	GR-PL			+				+
	Grasping – Pushing	GR-PS			-				
	Grasping – Sweeping	GR-SW	±	±		±	±		+
	Pushing – Spreading	PS-SP			-				-
	Pushing – Sweeping	PS-SW	±	±	-	±	±		
	Sweeping – Spreading	SW-SP	±	±		±	±		
III	Grasping – Pulling – Sweeping	GR-PL-SW	±	±	+	±	±		+
	Grasping – Pushing – Sweeping	GR-PS-SW	±	±	-	±	±		+
	Pushing – Sweeping – Spreading	PS-SW-SP	±	±	-	±	±		-

Table 2.1: List of tool/tissue interactions identified by Rosen’s study.

McBeth 2002 (from our lab) developed a second technical approach (i.e., measurement of kinematic and postural data in the live operating room setting) in which, in contrast to Rosen’s study, the data analysis was performed using an organizational structure provided by a hierarchical decomposition. The aim was to introduce a sense of context while describing the procedure in terms of surgical tasks, tool sequences and fundamental tool actions (**Figure 2.1**). This technique was based on the decomposition approach originally described by Cao 1996, but modified to improve generality and to incorporate additional quantitative kinematic features on low-level tool movements. As stated earlier, it was composed of five levels: phase, stage, task, sub-task, and action.

More relevant to the issue of instrument design, Mehta 2001 used motion analysis to explore the specific maneuvers that can be performed with various laparoscopic instruments as well as the sequence in which they were executed in order to reveal patterns of instrument use during procedures. This study developed the notion of individual tool movements as well as the notion of order of execution for describing how a tool is used in terms of those predefined movements.

A list of distinct instrument maneuvers was identified from a consideration of six relatively common types of laparoscopic procedures (**Table 2.2**): cholecystectomy, nisssen fundoplication, adrenalectomy, appendectomy, splenectomy, and nephrectomy.

Maneuver	Operational definition
Retracting with grasping	Maneuvering a tissue or organ that is inside the jaws of the instrument
Retracting without grasping	Maneuvering a tissue or organ while it is not within the jaws of the instrument
Cut ultrasonic	Separating tissue planes using the ultrasonic energy generated by the ultrasonic shears
Dissecting	Separating tissue planes using the blunt end of an instrument
Cutting	Slicing tissue or sutures using sharp scissors
Coagulation	Cauterizing a vessel without cutting
Clipping	Occluding a vessel or connecting latex drains with a metal clip
Irrigation and suction	Clearing the field of view using saline and/or suction
Suturing	Piercing of tissue with the suture needle
Suture tying	Manipulating the needle or free end of a suture to make a knot
Specimen/material removal	Removing an organ, tissue sample, or surgical material
Stapling	Separating tissues with a mechanical stapling device
Cut cautery	Separating tissue planes using electrical cautery

Table 2.2: Operational definition of maneuvers in Mehta's study.

Moreover, Mehta has advocated for the need to know how an instrument is actually used during the surgery and what impact its use has on the flow or dynamics of the procedure, which again suggests the need to integrate a description of flow into the hierarchical decomposition.

With regard to instrument exchange, Mehta found that two and three instrument cycles were used in laparoscopic cholecystectomy for achieving the highest level goals of the procedures (i.e., at the stage level), with the "curved dissector → clipper → scissors" and "hook cautery → suction irrigator" cycles being the most prevalent patterns of instrument exchange.

According to Mehta, one reason often given for why a particular instrument is chosen is the diversity of its functions (multifunctionality) because exchange during laparoscopic procedures is time-consuming and changing instruments disrupts the flow of the procedure which can break the concentration of the surgeons and interfere with their planning of the stages which still lie ahead. Mehta realized that this study was limited by the lack of knowledge regarding the circumstances surrounding instrument and maneuver changes, since the underlying reasons (i.e., cognitive behaviour) for exchanging certain instruments were not addressed. For example, situations such as an accidental rupture of a vessel or the availability of specific instruments during a surgery, which might affect the instrument exchange patterns, were not taken into account in this study. Therefore, he argued for analyzing instrument use in context (i.e., to explicitly identify which tasks are associated with certain patterns and how the flow of the procedure, as indicated by choices made by the surgeon, affects the use of instruments).

These previous studies concentrated primarily on hierarchical representations of MIS tasks and therefore do not allow for describing alternatives in execution plans since the cognitive element of human behaviour was not included. However, some of them recognize the importance of these issues in defining the appropriate context either for tool use analysis or monitoring of surgeons' skill development.

Since our goal at this point is to combine all these aspects of live surgeries in a single framework, it is first necessary to develop an adequate context-based representation for MIS procedures; thus, the purpose of the next section is to describe the tools from educational psychology, which we used to ensure that all key issues were addressed.

2.1.2 Task Analysis Methods

In order to design a framework to represent both motor and cognitive aspects of surgical task performance, we turn to the fields of psychology and education theory where a variety of task analysis methods have been introduced and developed to understand and design training and evaluation processes in a myriad of situations ranging from operating nuclear power stations to aircraft to simple car driving. In these various applications, the goals of applying task analysis methodologies are similar – to identify what the learners need to know, how they should perform or are performing, what skills they need to develop, and how the context may affect their decisions and actions.

Since training goals for laparoscopic cholecystectomy have already been established based on a hierarchical decomposition of activities [McBeth 2002], we are now interested in modelling the manner in which individual surgeons move along the task structure to achieve those goals. Therefore, the representation of three additional elements should be included in the performance analysis: the elemental behaviours involved in performing the procedure, the way surgeons process information as they execute specific tasks, and the influence of context (e.g., unexpected events such as bleeding) on the surgeon's activities.

2.1.2.1 Purpose of Task Analysis

There is no unique definition of *Task Analysis*. Different descriptions depend on the purpose for conducting it, the context in which it is applied and the type of performers involved. Purposes for conducting task analysis include developing job descriptions, designing human-computer interactions or designing different forms of instruction. From an educational perspective, task analysis could be defined as a process to determine statements of learning goals, to describe and prioritize tasks and subtasks that the learner will perform, and to develop assessment methods to determine what actually gets taught or trained while performing a particular activity [Jonassen 1999].

Since learning is a human-centered activity, all learning situations are different according to the different possible contexts; therefore there are many different task analysis methods. According to Jonassen 1999, however, there are 5 main kinds of task analyses (**Figure 2.3**):

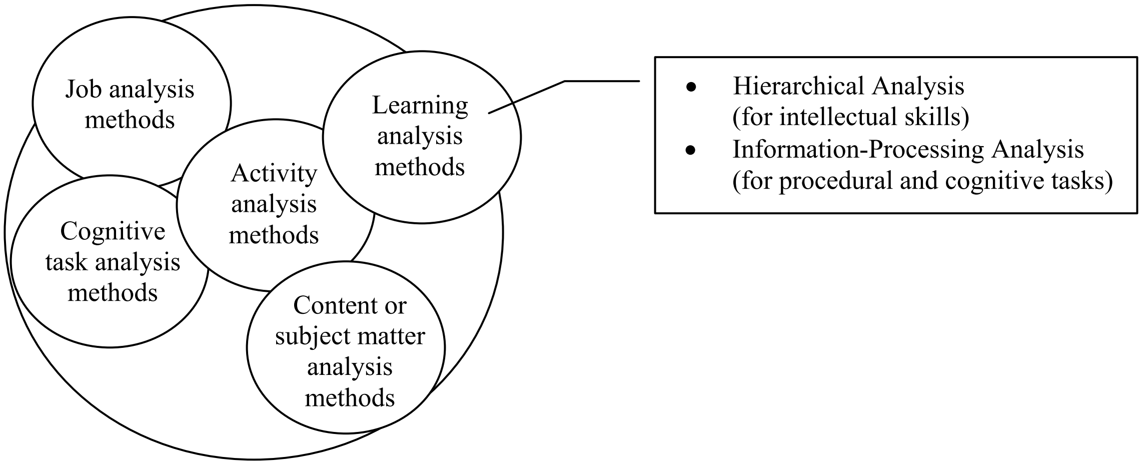

Job analysis methods
Learning analysis methods
Activity analysis methods
Cognitive task analysis methods
Content or subject matter analysis methods
Hierarchical Analysis (for intellectual skills)
Information-Processing Analysis (for procedural and cognitive tasks)

Figure 2.3: Domain of Task Analysis (Jonassen 1999).

From the variety of task analysis techniques that may arise from the previous domain, we focused our approach on two learning analysis methods, which seemed most appropriate for our needs: Hierarchical Analysis and Information-Processing Analysis because they offer complementary largely representations of the tasks and the steps needed to accomplish them. In addition, we did not find either a technique fully adequate by itself and therefore decided to combine and expand them to meet our requirements. This was achieved by analyzing the surgical procedure simultaneously in terms of task complexity and task sequences, as shown in **Table 2.3**.

Hierarchical	Information-processing
Solve the question: "What must the learner know in order to achieve this task?"	Solve the question: "What are the mental and/or physical steps that the learner must go through in order to complete this task?"
Developed from general to specific	Developed step-by-step (It has a start and an end)
Represented in terms of levels of tasks	Represented as a flowchart or an outline
Based on learning taxonomies (from most to least complex)	It is procedural in nature

Table 2.3: Comparative table: Hierarchical Analysis Vs Information-Processing Analysis.

The other methods in **Figure 2.3** concentrate mainly on structuring instruction (i.e., how trainers ought to teach skills) rather than describing learning (i.e., how people process information as they perform tasks), so we did not use them in this particular application.

The main objective of any task analysis method is to identify the most representative activities carried out during the performance of a task. In representing surgical procedures, hierarchical analysis would allow us to define the goal structures for the surgical tasks in terms of levels of complexity by observing and gathering data from video analyses and opinions from expert surgeons. However, since it is also necessary to describe surgeon's performance as a combination of overt (motor) and covert (cognitive) actions, information-processing analysis appears to be a useful way to deal with the sequential (procedural) representation. In this way, it becomes possible to characterize individual performances by identifying particular behaviours and decisions (i.e., to define all the possible routes that a surgeon may follow to achieve the global goals of the procedure).

2.1.2.2 Hierarchical Analysis

Hierarchical Analysis begins by breaking down the activity from more complex to less complex tasks in order to identify the prerequisite skills for adequate performance. For example, in the field of problem-solving, any final decision depends on mastery of certain rules, which in turn, demands mastery of certain concepts which, in turn again are based on knowledge of definitions. In MIS, isolating the Cystic Duct from the Cystic Artery would require surgeons to learn the difference between dissecting and clipping, which in turn also requires them to master tool movements such as pushing, grasping or sweeping. Therefore, each skill builds on simpler skills to form a learning hierarchy [Gagne 1985]. The result of a hierarchical analysis is a tree structure, which portrays the dependencies of the various skills and suggests an order in which they should be acquired.

Hierarchies have been frequently used to represent goal structures as a graphic summary. Typically, the first step is to use existing references such as texts, manuals and videos to construct a comprehensive list of the tasks that make up an activity. Then the definition of the complexity levels allows for grouping of the tasks, which also need to be ordered to show the hierarchical relationships for learning. Finally it is important to determine the hierarchy's accuracy by discussing the results with experts in the topic.

One good example from daily life is the description of the departure process at an airport (**Figure 2.4**).

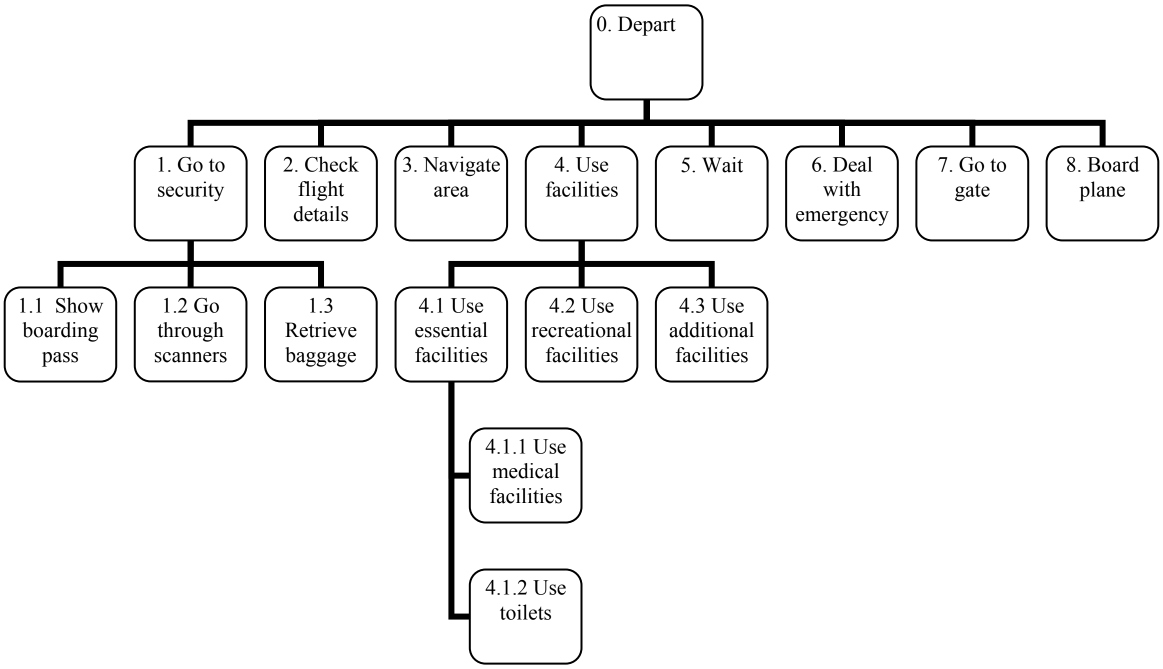


Figure 2.4: Hierarchy analysis of the departure process at an airport (adapted from http://polo.lancs.ac.uk/CDP/Uniport/Dresearch.htm).

This representation highlights the main benefit provided by a hierarchical analysis: decomposition of the tasks from the highest to the lowest level of complexity through an appropriate clustering. It consequently offers a good description of the overall activity (e.g., to assess functionality of surgical tools at different levels of the procedure); however, it does not adequately express the concepts of flow (sequencing) or optionality (decision-making in the face of patient variability or the occurrence of unexpected events), so we need to also consider approaches that include these concepts.

2.1.2.3 Information-Processing Analysis

As a complement to hierarchical analysis and with the purpose of identifying the mental and/or physical steps that a performer needs to go through in order to complete a task, Information-Processing Analysis is commonly used to develop a better description of performance in terms of both procedural (observable) and cognitive (non-observable) behaviours. The aim is to provide a graphical representation of the different routes that may be followed to complete an activity: for this particular case, the flow of the surgical procedure based on a surgeon's actions.

Information-Processing Analysis, also known as procedural task analysis was developed in 1960 when the behaviourist movement aimed to represent human performance as a chain of stimulus-response reactions. Each step of a task was modeled as a response to a given stimulus, which served in turn as a stimulus to the next response step. Performance was primarily described as a linear series of steps. However, it was realized that complex tasks involve decisions, and alternative action sequences. A more complex description of task behaviour was therefore necessary, and the computer programming method of flowcharting was adopted because it allowed for branching, loops, and decision points.

In general, information-processing analysis analyzes goals by describing the sequence of activities that must be executed to complete them. This method breaks up a goal into its component tasks and represents actions, decisions, and paths as a sequence of **observable (motor skills) and non-observable (executor's thought processes) behaviours**. It reveals the individual's overt steps and decisions taken to accomplish a task, as well as the overall executive routine of the procedure as a whole [Jonassen 1999]. This type of analysis is usually represented in the form of a flowchart.

To conduct an information-processing analysis, it is necessary to gather as much information as possible from experts who know how to complete the task. This data acquisition process is mainly implemented by observation (i.e., recording the steps while completing the tasks) and think-aloud techniques (i.e., to register thought processes), which together may reveal all the possible paths through the procedure.

An example of an information-processing analysis for the familiar activity of sending an email (where the component tasks are mostly observable) is shown below (**Figure 2.5**).

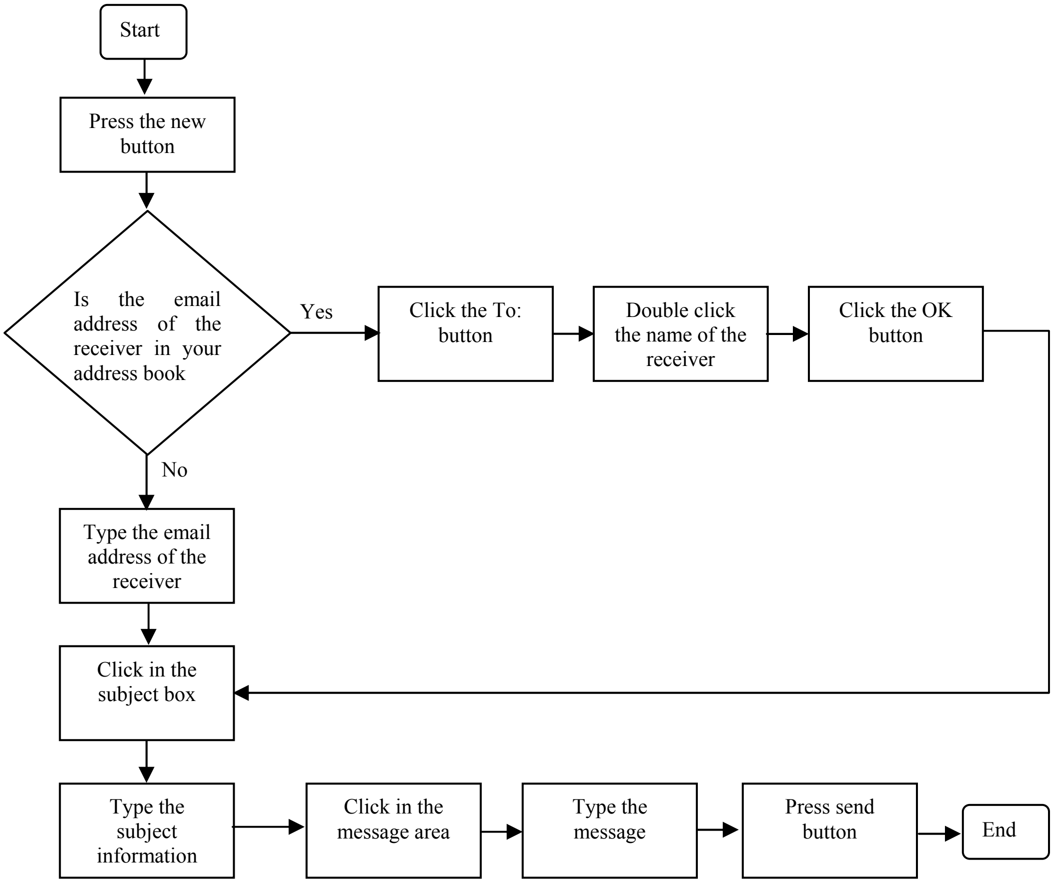

Start
Press the new button
Is the email address of the receiver in your address book
Yes
Click the To: button
Double click the name of the receiver
Click the OK button
No
Type the email address of the receiver
Click in the subject box
Type the subject information
Click in the message area
Type the message
Press send button
End

Figure 2.5: Information-Processing Analysis for sending an email (from INSTRUCTIONAL DESIGN KNOWLEDGE BASE – IDKB, Instructional Technology Program, Graduate School of Education, George Mason University).

This simple example illustrates two key features of value: sequencing and a decision point that leads to two alternative routes through the flowchart.

Often, the decision-making process is hidden from the observer. Consider the simple example of identifying a geometrical shape as shown in **Figure 2.6**.

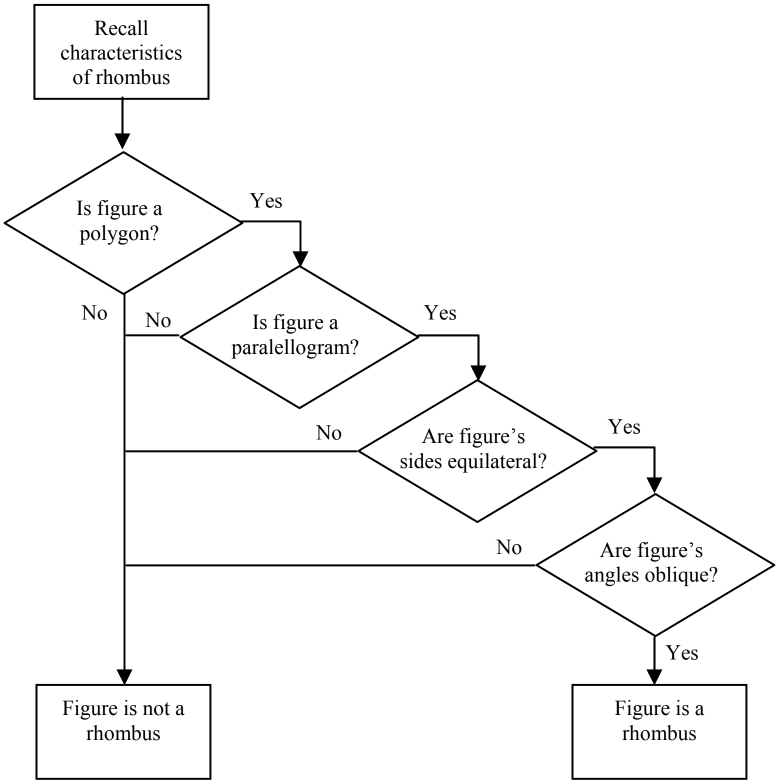

Recall characteristics of rhombus
Is figure a polygon?
Yes
No
No
Is figure a paralellogram?
Yes
No
Are figure's sides equilateral?
Yes
No
Are figure's angles oblique?
Yes
Figure is not a rhombus
Figure is a rhombus

Figure 2.6: Information-Processing Analysis for the identification of a geometrical shape (from INSTRUCTIONAL DESIGN KNOWLEDGE BASE – IDKB, Instructional Technology Program, Graduate School of Education, George Mason University).

In this example, the decisions are not directly apparent to an observer, but must be elicited by a think-aloud process so that the observer becomes aware of what criteria the subject used to draw their conclusion. Furthermore, the decisions need not be made in the order presented – the subject may well have first tested whether the sides were equilateral rather than parallel, so this representation does not properly represent all valid paths through the activity.

2.1.3 Finite State Machines and Petri Nets for Representing Task Sequences

In addition to the task analysis methods described above, we considered using Finite State Machines (FSM) or Petri Nets (PN) to model surgical processes.

A finite state machine is a mathematical model that represents how a system can change its state over time as it reacts to internally or externally triggered events. It is composed of states, transitions and actions. A state stores information about the past, i.e. it reflects the input changes from the system start to the present moment. A transition indicates a state change and is described by a condition that would need to be fulfilled to enable the transition. An action is a description of an activity that is to be performed at a given moment. Finite state machines are a subset of Petri Nets in which each transition has exactly one input and one output [Brownlee 2006, Gibson 2000].

Petri Nets are a generalization of Finite State Machines which are particularly well-suited for representing systems in which synchronization, concurrency, communication and resource sharing are important [Bobbio 1990]. A Petri net consists of places (circles), transitions (rectangles), tokens (moving points) and directed arcs (arrows). Arcs connect places and transitions - not places and places or transitions and transitions. The places from which an arc run to a transition are called the input places of the transition; the places to which arcs run from a transition are called the output places of the transition. A transition can only fire when there are tokens in every input place. When it fires, one token is taken from every input place, and every output place from the transition gets a token [Murata 1989, Peterson 1977].

FSMs would not be particularly appropriate for representing surgical flows because there is no obvious way to represent sub-processes that are only invoked when certain circumstances trigger them, such as the 'control bleeding' task other than by explicitly representing these interrupting tasks at every point where they could potentially occur, which would significantly complicate the diagrammatic representation of the main steps of the procedure. We are aiming to provide a more compact representation of the standard steps of a procedure and therefore we are interested in treating unexpected conditions such as 'control bleeding' as sub-routines that may be used or not as the need arises. In addition, Finite State Machines do not allow for simple representations of parallel processes, nor can they represent cognitive elements such as decision points.

It is similarly difficult to use a Petri net representation directly, although it is somewhat more flexible than FSMs. For example, one could represent AND and OR operations by specifying appropriate transition rules. However, while PNs allow for the representation of distributed systems through the notion of multiple tokens, each of which can move independently and simultaneously through different places and transitions in the net, there is no obvious way to implement the notion of a single actor who is limited to moving one token at a time. We therefore believe that a more straightforward and simple representation would be more appropriate for the context of the present research.

However, in future work, it may be possible to borrow the notion of distributed independent tokens to represent the interactions amongst surgeons, nurses, anesthesiologist and any other member of the operating room team.

2.1.4 Description of the Standard LC Surgical Technique

Laparoscopic cholecystectomy (LC) is commonly described as a sequence of six major activities: establishment of pneumoperitoneum, ultrasonography, placement of trocars, isolate gallbladder, remove gallbladder, and closure [Reddick 1993, Cuschieri 1990].

✓ *Establishment of pneumoperitoneum* [Reddick 1993, Cuschieri 1990]. The purpose of this step is to provide space in order to visualize the abdominal cavity with the laparoscope. It is performed after patient preparation, which includes general anesthesia, insertion of catheter and positioning of the patient (**Figure 2.7**).

- Insufflation needle is placed through a small skin incision just above the umbilicus
- The abdomen is filled with 3 to 4 liters of carbon dioxide until the intraabdominal pressure reads 12-14 mm Hg
- Insufflation needle is removed and replaced by a 10mm or 11mm trocar through which a laparoscope with attached camera is inserted to confirm intraperitoneal placement
- Patient is repositioned to allow the abdominal viscera to fall inferiorly, away from the GB

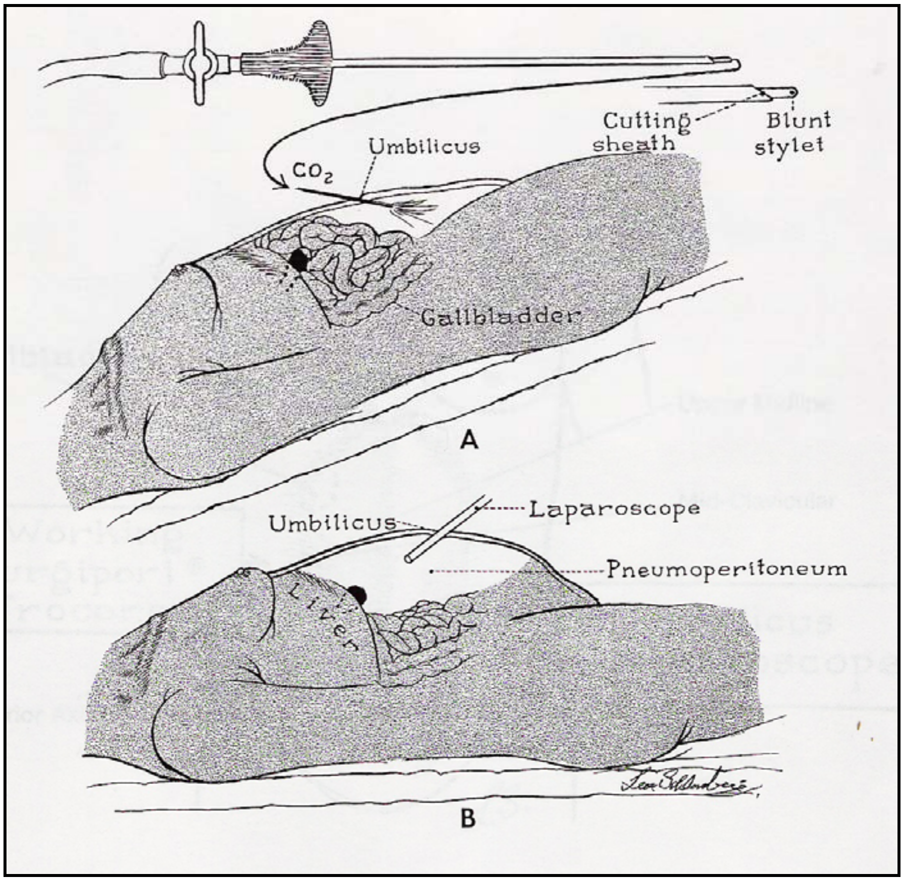


Figure 2.7: Establishment of pneumoperitoneum (from Atlas of Endo Cholecystectomy with Auto Suture instruments. Zucker K., Bailey R. in cooperation with United States Surgical Corporation).

- ✓ *Ultrasonography* [Berber 2001, Reddick 1993, Cuschieri 1990]. Although this step is mostly performed pre-operatively to confirm the presence of gallstones and to detect any dilation of the intrahepatic or extrahepatic bile ducts, it is the surgeon's choice to perform a laparoscopic ultrasonography as a way of assessing the feasibility of continuing laparoscopically. When done as part of the procedure, it is used to improve the safety of laparoscopic cholecystectomy, especially in cases of acute inflammation or distorted anatomy.

✓ *Placement of trocars* [Reddick 1993, Cuschieri 1990]. Accessory trocars are placed using direct laparoscopic guidance to allow insertion of instruments (**Figure 2.8**):

- One 10mm trocar is placed one-third of the distance between the xiphoid and the umbilicus, just to the right of the midline
- Two 5mm trocars are placed two fingerbreadths below the right costal margin, one in the anterior axillary line and the other one in the mid-clavicular line

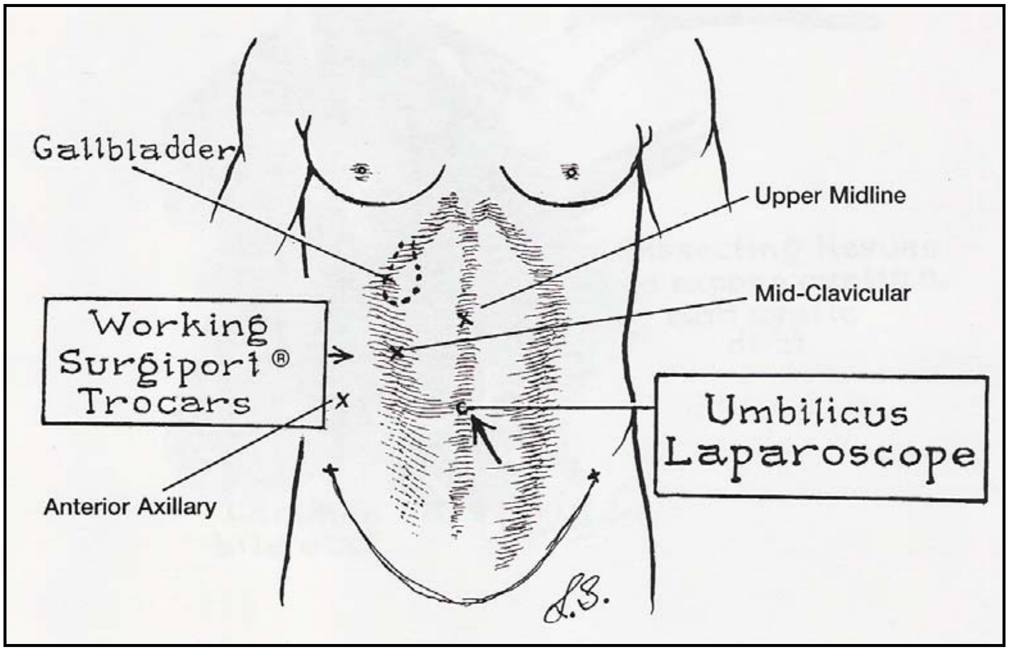


Figure 2.8: Placement of trocars (from Atlas of Endo Cholecystectomy with Auto Suture instruments. Zucker K., Bailey R. in cooperation with United States Surgical Corporation).

✓ *Isolate gallbladder* [Reddick 1993, Cuschieri 1990]. This step is related to the detachment of the gallbladder. It involves the separation of the gallbladder from the anatomical structures that join it to the rest of the body and the dissection of the gallbladder's bed from the liver. In this way, complete separation of the gallbladder from the body is achieved (**Figure 2.9**).

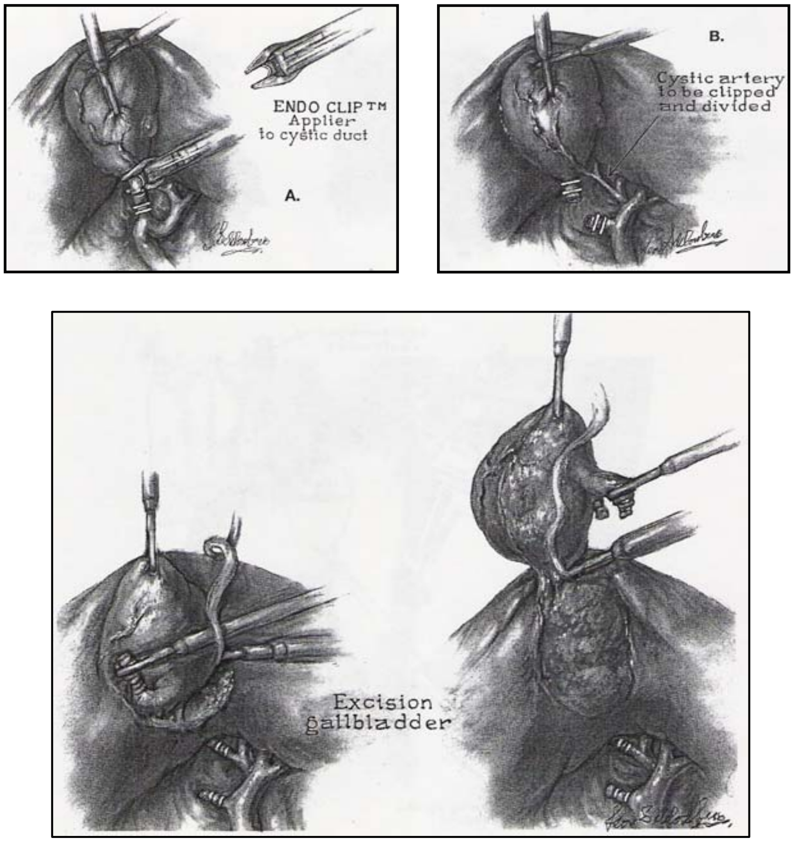


Figure 2.9: Isolate gallbladder (from Atlas of Endo Cholecystectomy with Auto Suture instruments. Zucker K., Bailey R. in cooperation with United States Surgical Corporation).

- ✓ *Remove gallbladder* [Reddick 1993, Cuschieri 1990]. This step consists in the extraction of the GB from the intraabdominal space of the patient, usually through the umbilical incision. It involves a complete clearance of waste materials produced during the dissection of the gallbladder and the optional use of a bag to place the gallbladder before proceeding to extraction (**Figure 2.10**).

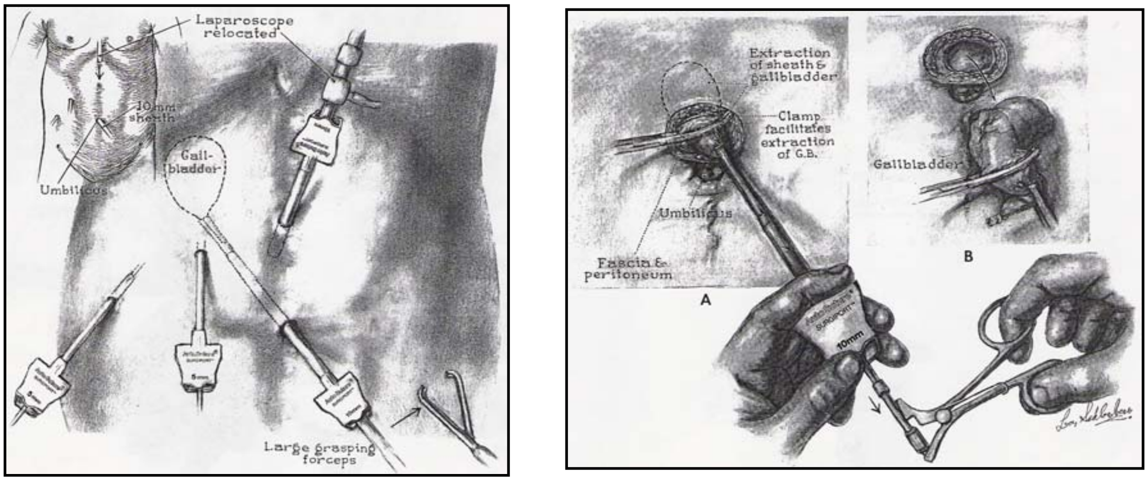


Figure 2.10: Remove gallbladder (from Atlas of Endo Cholecystectomy with Auto Suture instruments. Zucker K., Bailey R. in cooperation with United States Surgical Corporation).

- ✓ *Closure* [Reddick 1993, Cuschieri 1990]. This step consists of the withdrawal of the trocars and the desufflation and suturing of the stab wound.

‘Isolate gallbladder’ and ‘remove gallbladder’ constitute the two central steps of the procedure, which at a more detailed level can be described in terms of five steps for ‘isolate gallbladder’: explore anatomy, isolate CD/CA, separate CD, separate CA, dissect GB; and three steps for ‘remove gallbladder’: clean-up, bag GB, extract GB.

- ✓ *Explore anatomy* [Reddick 1993, Cuschieri 1990]. The objectives of this task include:

  - Detection of inadvertent injuries caused during insufflation and insertion of main trocar/cannula
  - Exclusion of additional unsuspected intra-abdominal pathology

- Assessment of the feasibility of laparoscopic cholecystectomy. This objective involves the assessment of the technical difficulty and safety of gallbladder excision via the laparoscopic route.

✓ *Isolate CD/CA* [Reddick 1993, Cuschieri 1990]. The objective of this task is to expose the cystic duct (CD) and the cystic artery (CA). Dissection in the form of stripping or blunt dissection is used to detach the surrounding tissue from both structures facilitating the correct identification of the biliary tree

✓ *Separate CD* [Reddick 1993, Cuschieri 1990]. The objective of this task is to separate the gallbladder from the cystic duct. It is achieved by applying clips to the duct at the distal and proximal ends to allow for a safe division of this structure.

✓ *Separate CA* [Reddick 1993, Cuschieri 1990]. Similar to the previous activity, the objective of this task is to separate the gallbladder from the cystic artery. It is achieved by applying clips to the artery at the distal and proximal ends to allow for a safe division of this structure.

✓ *Dissect GB* [Reddick 1993, Cuschieri 1990]. The objective of this task is to separate the gallbladder (GB) from its bed that keeps it joined to the liver. Dissection in the form of cauterizing is usually performed to achieve this goal. Careful must be taken to avoid injuries on the gallbladder that may provoke gallstones spillage.

- ✓ *Clean-up* [Reddick 1993, Cuschieri 1990]. A clean dissection and separation of the structures involved in any minimally invasive surgery is necessary to avoid post-operative problems; therefore constant irrigation and suction is performed throughout the procedure. However, for the case of laparoscopic cholecystectomy, clean-up is also a required task that needs to be performed just before the extraction of the gallbladder, which is the reason for situating it at the task level.

- ✓ *Bag GB* [Reddick 1993, Cuschieri 1990]. Bagging the gallbladder is also an optional task. The decision depends on the surgeon's preference and on the availability of resources at the hospital. For the majority, it is considered an adequate activity to prevent gallstone spillage during the extraction phase.

- ✓ *Extract GB* [Reddick 1993, Cuschieri 1990]. Removing the detached gallbladder from the abdomen is the last intraoperative activity in laparoscopic cholecystectomy. This task is performed by grasping the neck of the gallbladder and pulling it gently through the umbilical cannula and out of the abdomen.

2.1.5 Summary

Neither the hierarchical nor the information processing analyses are able, individually, to adequately represent all the concepts we need to represent in modelling a general surgical task. We require that our task analysis approach expresses the concepts of hierarchy, flow and sequencing, explicit and implicit decision-making, freedom to execute selected tasks in arbitrary order, and the need to respond to unexpected events such as bleeding or a cardiac emergency.

Given the various strengths and limitations of the existing hierarchical and information processing analyses, we opted to develop a hybrid approach, which combines their strengths and addresses their limitations. This combined framework, which we call a Motor Cognitive Modelling Diagram (MCMD), is intended to be suitable for describing general surgical tasks in minimally invasive procedures and is presented in section 2.3.

2.2 Protocol

This part of our research is devoted to the application of methods from the cognitive sciences to provide standardized descriptions of minimally invasive procedures able to identify key components of the surgical skills that are relevant during training. The main goal is to develop a method to describe Laparoscopic Cholecystectomy, which integrates motor and cognitive actions so as to enable us to represent the surgical context in which surgical actions and judgements occur.

In order to achieve this goal, we applied concepts from the task analysis literature to implement an appropriate protocol for developing our new representation of surgical procedures (i.e., the MCMD). This representation is based on the standard elements of laparoscopic cholecystectomy and its implementation was developed in consultation with two expert surgeons and one expert from the applied psychology field.

2.2.1 Methods

Jonassen 1999 used the term *task knowledge structures* as a representation of the task knowledge that a user might have in order to perform a task, and created a general protocol for data gathering when attempting to describe complex activities. In accordance with our needs and goals, we have adjusted and implemented the following general methods:

✓ Collect information about the task

- Based on a literature review of surgical techniques, we drafted a preliminary hierarchical decomposition of the laparoscopic cholecystectomy (LC) procedure by identifying the component steps of the surgery and classifying them according to the complexity (i.e., number of surgical goals involved at each step). This was presented in the previous section.

- Observation: We attended and videotaped (10) laparoscopic cholecystectomy procedures and constructed a checklist in order to facilitate the process of verification (i.e., based on literature review about the surgical technique; this list was used as a template to mark the order in which activities were accomplished; see Appendix A). We performed two rounds of video analysis. The first one was concerned with identifying the individual components of the hierarchical levels (i.e, phases, tasks, and subtasks). During the second round, each level was again analyzed to identify the interactions amongst its constituent components (i.e., sequences of activities, deviations in the flow due to decision points, serial and parallel activities, etc).

Furthermore, this observation allowed us to identify the most common difficulties encountered during surgery and how experts deal with them. The analysis of the results provided essential information, which allowed us to update our diagrams.

- Think-aloud: During (6) of the procedures, we asked the surgeon to comment out loud on what he was considering as he performed the surgery. This information was captured using video and audio recordings.
- Interviews: We followed the 'think-aloud' procedures with an interview in which the surgeon reviewed the video with the investigator and added additional comments and insights. A comprehensive analysis of the results was conducted in collaboration with the expert to update the preliminary diagrams, which were again tested in the operating room several times, in addition to the 10 recorded procedures, as a way to perform a follow-up of the think-aloud process.

The general procedure constituted a cycling process of four activities as presented in the following timeline (**Figure 2.11**).

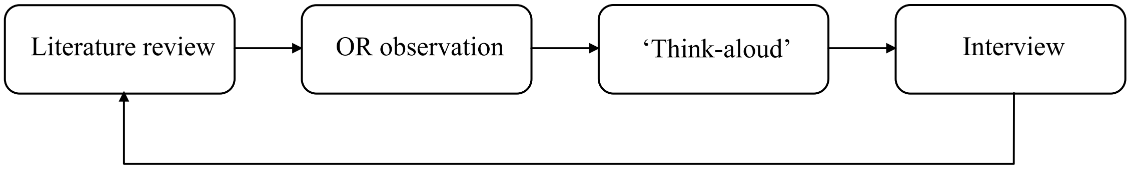


Figure 2.11: Representing the information acquisition process for constructing our MCMD framework.

Finally, we designed a new hierarchical Motor and Cognitive Modelling Diagram (MCMD – a kind of flowchart) to capture the task sequences of laparoscopic cholecystectomies. We started with the symbols of the information processing analysis technique and complemented them with new ones according to our needs. We therefore, created a diagram language composed of six primary symbols: processes, decisions, interrupt service routines (ISRs), options points and AND and OR gates (described in detail in the following section). We then tested and refined them during 8 new cases until no further changes seemed necessary. We did not acquire motor performance measures during these procedures, but the process nodes are designed to contain these measures.

2.3 Results

In this section, we provide a detailed explanation of the notation we developed and present diagrams describing laparoscopic cholecystectomies.

2.3.1 Hierarchical Decomposition

For our hierarchical decomposition, we initially adopted McBeth's approach since, like him, we are also interested in incorporating kinematic features on low-level tool movements for developing our assessment methodology. However, we made a small number of changes in order to maintain the task analysis emphasis on decomposing the procedure in terms of the nature of the performed activities rather than on the type of surgical tool used to achieve a goal. The surgical tool issue will be included as part of our sequential/flow representation described in section 2.3.2.

Our hierarchical decomposition therefore includes 4 levels of complexity defined as follows:

- ✓ Phases: this level describes larger-scale goals (i.e., they may be broken into sub-goals), involving more than one anatomical structure, which are performed using one or more surgical tools.
- ✓ Tasks: this level describes manipulation on a single anatomical structure with one or more surgical tools in order to achieve a larger goal.
- ✓ Sub-tasks: this level corresponds to local goals and it describes elemental surgical activities performed with one surgical tool per hand; it could also introduce the sense of monitoring dominant and non-dominant hand movements separately or simultaneously in order to describe bimanual dexterity. This level also allows representing the notion of parallelism as sets of sequential activities that might represent optional paths during the procedure.
- ✓ Actions: this is a “goal-free” level that describes tool motion primitives (e.g., pushing, sweeping, etc) associated with a single surgical tool, although separate models can be created for the tools used by each hand.

Following the general procedure presented in **Figure 2.11**, we first outlined a preliminary hierarchical diagram based on an extensive review of literature about the surgical technique, which was then iteratively updated after verification with video analysis and interviews with expert surgeons. Most of the phases and tasks included in our diagram corresponded well with the textbook theory and practice of LC surgery; however, we found that there has also been extensive discussion regarding the ultrasonography and cholangiogram steps, which could be performed either before or during the procedure [Patterson 1997].

Most of the literature [Fielding 2002, Siperstein 1999, Reddick 1993, Cuschieri 1990] and the opinions of the surgeons involved in this research, suggest that ultrasonography is a preoperative step. Others [Berber 2001, Siperstein 1999] consider that ultrasonography should be included as an intraoperative step as a means of assessing whether it is feasible to perform a successful laparoscopic procedure. Some people have studied whether intraoperative ultrasonography (non-invasive, no ionizing radiation) has the potential to replace cholangiography in certain cases [Siperstein 1999]. However, cholangiography (the current gold standard for detecting leaks) has a crucial role in patients with suspected common duct injuries, while sonography has little role in this setting, as it cannot identify the site of injury or bile extravasation [Siperstein 1999]. Overall, laparoscopic ultrasonography (LUS) has been shown to be roughly equivalent to intraoperative cholangiography (LCG) in its ability to provide images of the ductal system adequately and efficiently, with some studies favourings LCG and other favouring LUS [Petelin 2002]. Therefore we decided to include both options in our diagram, in order to provide generality in capturing all acceptable behaviours.

As described in section 2.1.3, at the phase level, laparoscopic cholecystectomy is represented as a sequence of six major activities (**Figure 2.12**): establishment of pneumoperitoneum, ultrasonography, placement of trocars, isolate gallbladder, remove gallbladder, and closure.

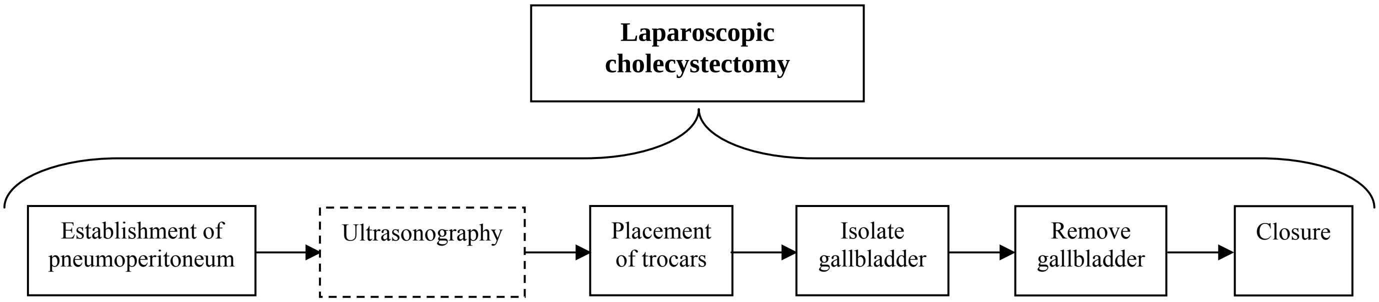


Figure 2.12: Phase level in hierarchical decomposition. Dashed line indicates an 'optional' activity that might be performed either during or before the procedure.

It is important to note that an 'optional' activity usually serves some implicit decision-making goal. That is, it is done to help the surgeon make a decision and may therefore indicate that the surgical situation is not straightforward.

Since at the task level we were mainly interested in analyzing surgical activities which involve the use of MIS tools, and which the instructing surgeons we collaborate with are most concerned with, we provided a hierarchical decomposition for the two central steps of the procedure: 'isolate gallbladder' and 'remove gallbladder'. This level is then composed of eight tasks in total, five for 'isolate gallbladder': explore anatomy, isolate

CD/CA, separate CD, separate CA, dissect GB; and three for 'remove gallbladder': clean-up, bag GB, extract GB (**Figure 2.13**).

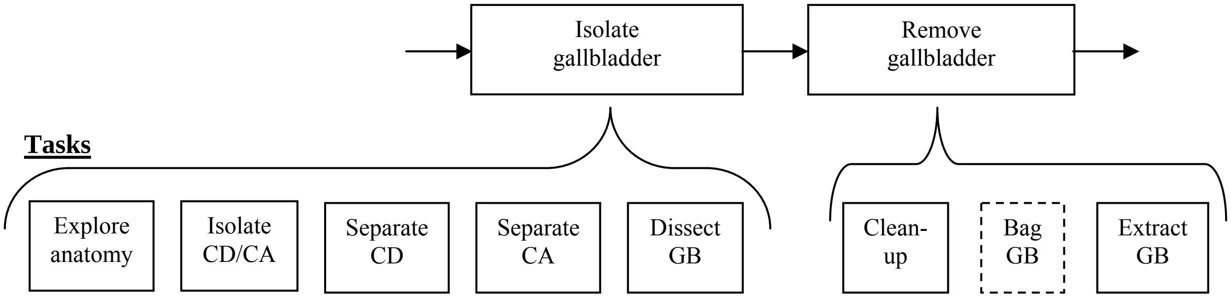


Figure 2.13: Task level in hierarchical decomposition (dashed lines indicate optional tasks).

After videotape analysis and discussions with expert surgeons, we identified four tasks as key components of the procedure to be decomposed at the sub-task level (**Figure 2.14**) since they involve major technical skill proficiency: isolate CD/CA, separate CD, separate CA, and dissect GB. We found that five sub-tasks were appropriate to describe these four mentioned tasks: detach tissue, dissect tissue, cauterize, clip, and divide.

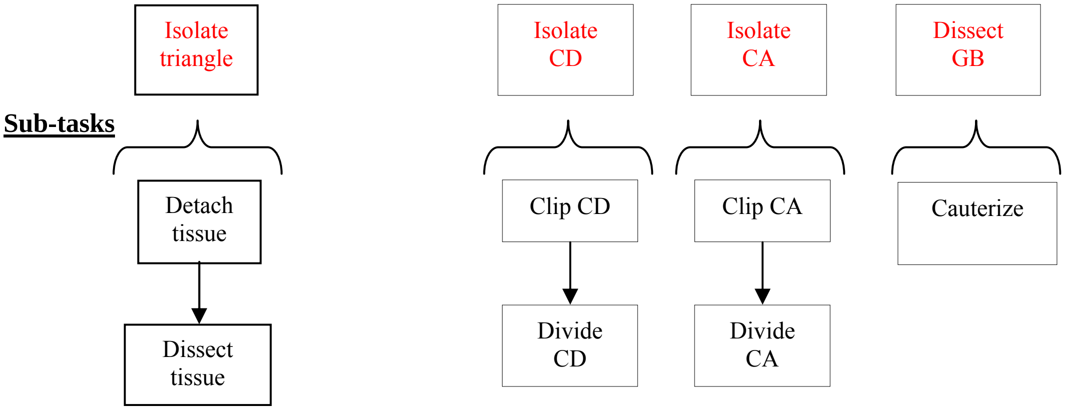


Figure 2.14: Sub-task level in hierarchical decomposition for LC procedures.

Moving into the lowest level of the hierarchy and in order to identify the set of discrete surgical tool movements used to perform any subtask and which are amenable for kinematics assessment, we performed a series of video analyses with expert surgeons across all the recorded procedures. Results from these video analyses indicated that 10 elemental surgical tool motions were sufficient for decomposing our selected subtasks (**Table 2.4**). In comparison to McBeth's approach, we reduced the number of actions as we conceived of 'release' as a jaw opening action, which was included as part of other actions' descriptions. 'Translate' (i.e., X-Y plane) was also eliminated as we conceived of 'translate' as belonging to 'reach', which in our structure represents the motion of the tool in any direction towards a target.

ELEMENTAL MOTION	DESCRIPTION
Push	Repetitive movements of tool into target structure
Pull	Repetitive movements of tool away of target structure
Reach	Movement of tool in any direction towards the target structure
Orient	Rotational movement of the tool
Sweep	Repetitive horizontal movements of the tool to separate tissue
Spread	Repetitive open & close movements of jaws to separate tissue
Grasp & hold	Repetitive open & close movements of jaws attempted to hold the target structure between jaws
Grasp & cut	Repetitive open & close movements of jaws attempted to divide the target structure
Idle	Visible movement of the tool without touching any structure
Out	Removal of tool from patient's abdominal cavity

Table 2.4: Description of the 10 identified elemental motions of the surgical tools used to describe subtasks.

The complete hierarchical decomposition for laparoscopic cholecystectomy implemented from our analysis of the surgical technique, which included literature review, video and audio analyses of real procedures, and discussions with experts, is presented in **Figure 2.15**.

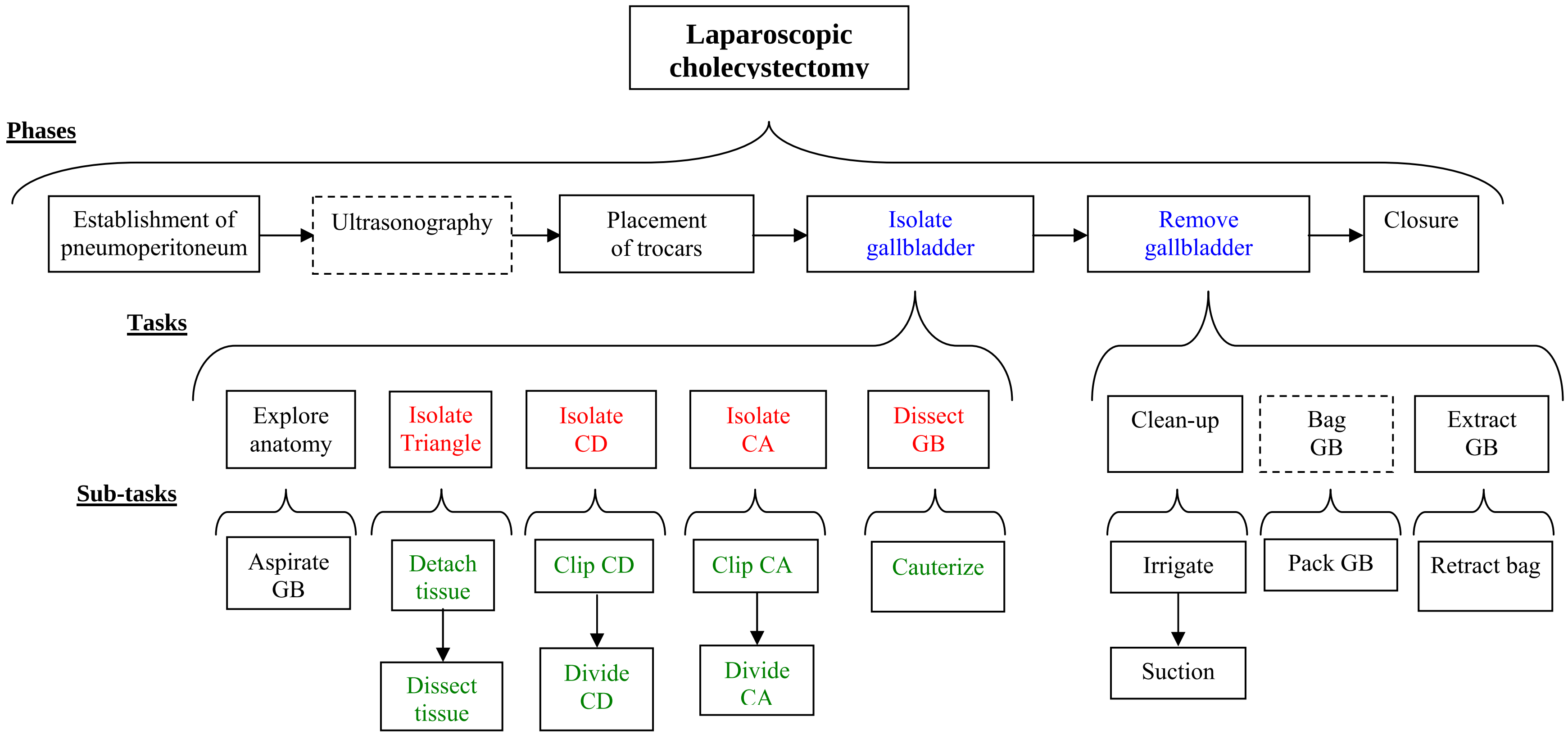


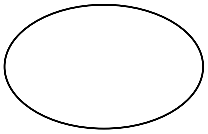
Figure 2.15: Hierarchical decomposition for laparoscopic cholecystectomy. Various colours indicate high demanding phases (blue), tasks (red), and subtasks (green) in terms of technical skills.

2.3.2 Motor and Cognitive Modelling Diagram (MCMD)

Although hierarchical decomposition is a helpful first step in understanding a surgical procedure, it does not incorporate the explicit description of flow and sequencing that is contained in Information Processing analyses. This becomes especially important when the purpose is to describe a surgeon's performance as a set of actions and decisions influenced by the different contexts that are present in the procedure. Therefore, the main contribution of this part of our research focuses on providing a standard notation for modelling surgical performance by means of a new representation of the flow of the surgical procedure in terms of motor and cognitive behaviours.

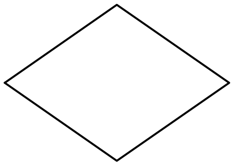
Although the main surgical goal is the same for any Laparoscopic Cholecystectomy procedure, it could be achieved by following different routes depending on the surgeon's experience and the influence of inadvertent or variable factors that might cause a particular diversion. To provide an appropriate framework, we have developed MCMD (Motor & Cognitive Modelling Diagram), which offers a new way of including the possible contexts that lead to changes in routes. MCMD is composed of a set of diagrams that allows for mapping flow between nodes of the hierarchical decomposition. Our set of symbols initially came from the information processing analysis technique but as variant observations were made, other symbols were added and adjusted to our specific needs.

All MCMD's diagrams share the same symbology structure according to the following description:



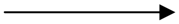
PROCESS:

✓ Takes time
✓ Has properties (e.g., distributions of velocities, forces, etc)



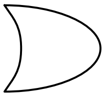
DECISION:

✓ Indicates various possibilities



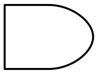
TRANSITION:

✓ Links processes or/and decisions
✓ Defines task sequences
✓ Takes no time



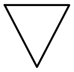
OR:

✓ Indicates that the procedure may proceed when AT LEAST one input requirement is satisfied



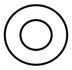
AND:

✓ Indicates that the procedure may proceed ONLY when all input requirements are satisfied



ISR – Interrupt service routine:

✓ Refers to a sub-process that is invoked while performing certain processes (e.g., control bleeding)
✓ Has its own task decomposition and diagram



OPTION POINT:

✓ Exist when there are parallel branches which can be performed in either order
✓ It allows jumping to other available option points

As it will be shown, the flow of the procedure is mainly represented from left to right; however, it is important to note that when executing certain tasks and sub-tasks, the surgeon may note that it is necessary to perform an otherwise unrelated task (e.g., control bleeding, clean-up, and spillage of GB stones). Since those routines need to be executed at specific moments but may be accessed at any time while performing a specific task, we have referred to them as Interrupt Service Routines (ISR) – an analogy to computer processing where the main task is suspended to allow processing of an urgent event. We have found that ISRs considerably simplify the representation because they eliminate the need to include explicit checks for triggering conditions.

In the following, we will present the components of this new representation by relating them to the description of the surgical technique previously presented in section 2.1.3. This description will include MCMD diagrams for tasks, sub-tasks and ISR.

- ✓ *Explore anatomy* [Cuschieri 1990] (**Figure 2.16**). Two important decisions arise during this task:

    - Assessment of the anatomy (D1). At this stage, a preliminary assessment of the general anatomy is performed. The situations that may prompt the surgeon to convert to open at the very beginning of the procedure include: extensive adhesions caused by prior surgery or recurrent attacks of cholecystitis, unusual vascular or ductal anatomy, other unsuspected pathology in the abdomen, and acute inflammation.

- Assessment of the severity of cholecystitis (D2). The decision involved is influenced by the experience of the surgeon in laparoscopic surgery based on situations such as: easy cases, feasible but difficult cases, cases of uncertain feasibility (trial dissection), unsuitable cases (for severe acute cholecystitis in which a decrease of the cholecystitis is carried out by aspiration of the gallbladder fluid contents through healthy fundus).

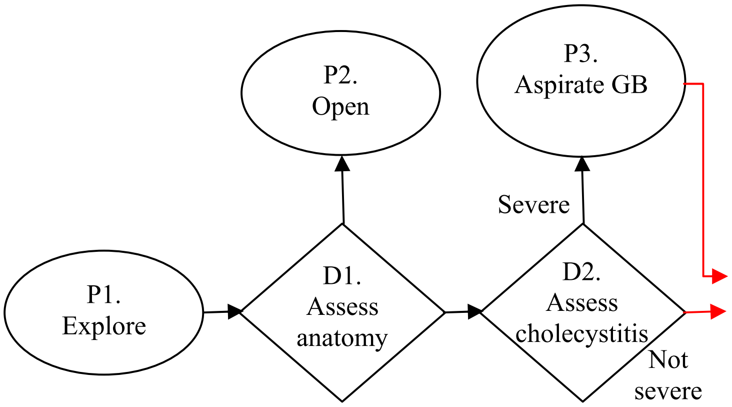


Figure 2.16: MCMD for ***explore anatomy*** at the task level.

✓ *Isolate Triangle of Calot* [Reddick 1993, Cuschieri 1990]. In this task, a thorough appreciation of the anatomy of the cystic pedicle is crucial for the safe dissection of the cystic duct and artery. The cystic pedicle outlines the margins of the triangle of Calot and contains, between its superior and inferior leaves, the cystic duct (usually anteriorly), the cystic artery (above and behind the duct) and the cystic node, which is closely applied to the neck of the GB between the duct and the artery.

As shown in **Figure 2.17**, this task may be reached from one of two different routes as pointed out by the OR symbol: (1) after aspiration of the gallbladder because of severe cholecystitis which enhances difficulty in grasping the fundus of the gallbladder, or (2) after confirming appropriate grasping of the gallbladder due to non-severe cholecystitis.

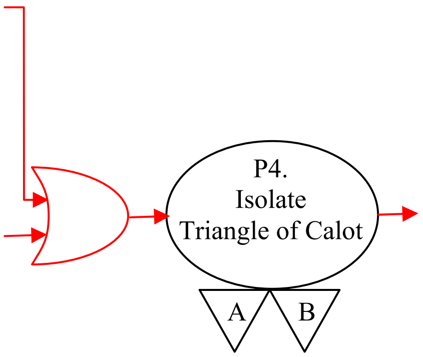


Figure 2.17: MCMD for ***isolate Triangle of Calot*** at the task level.

It is important to note that this particular characteristic of accessing a certain task from different routes provides the notion of context-dependent performance; which in turn, demands a particular representation and modelling. At this stage we then introduced the OR and AND symbols in the subtask representation of the 'isolate triangle of Calot' process, which is achieved as follows (**Figure 2.18**):

- Dissect away the overlying fibroareolar structures from the infundibulum of GB and Hartmann's pouch with blunt stripping action, always starting on the GB and stripping the tissue toward the porta hepatic using either the curved or the L-Hook dissectors (P4.1A and P4.1B). This corresponds to the initial dissection around the neck of the GB in which the peritoneum is lysed.

- Clear and identify the structures contained within the Calot's triangle and its reverse side (D4.1). Calot's triangle is the ventral aspect of the area bounded by the CD, hepatic duct, and liver edge, and its reverse corresponds to the dorsal aspect of this space.
- Identify precisely the junction between the infundibulum and the origin of CD by further blunt dissection, gain as much CD length as possible by stripping away the strands of peritoneal, lymphatic, neural, and vascular tissue from the CD, and create a window to isolate CD from CA using curved dissecting forceps or L-Hook (P4.2A and P4.2B).

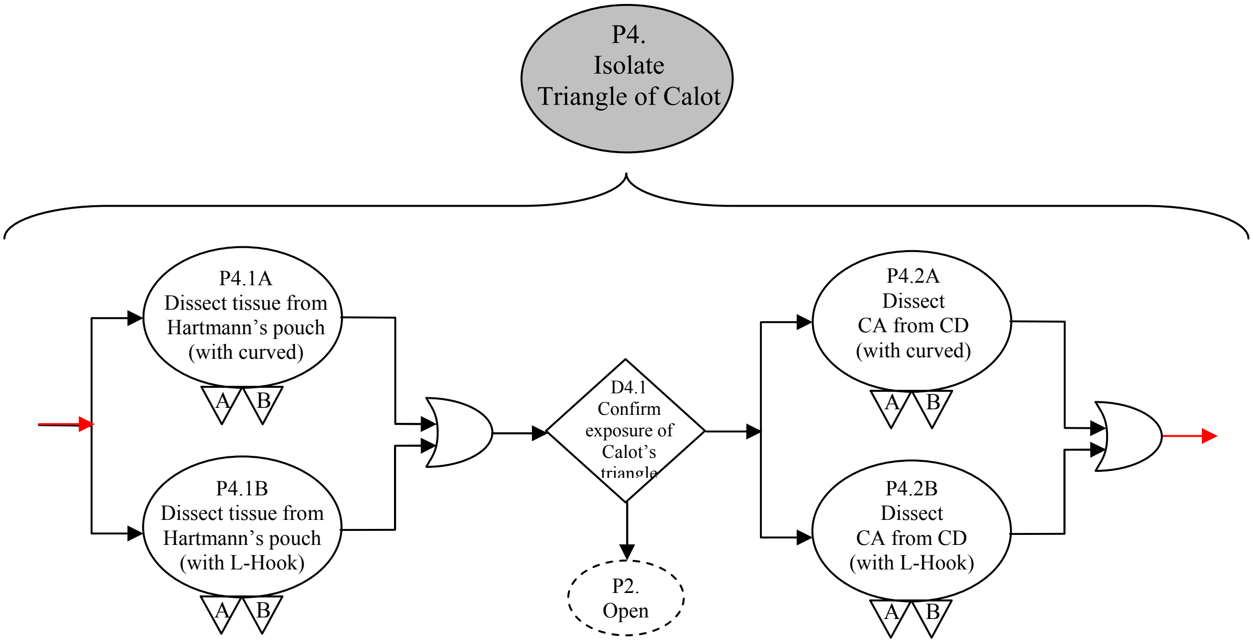


Figure 2.18: MCMD for ***isolate Triangle of Calot*** at the subtask level.

As noted in **Figure 2.17** and **Figure 2.18**, two interrupt service routines (i.e., triangle symbol) were included as part of the description: control bleeding (A) and clean-up (B), due to the presence of dissection activities that may generate considerable amounts of blood or bile spillage[1].

Bleeding occurs either during the dissection of the cystic pedicle or when the GB is detached from the liver; it is the most common cause for enforced conversion. However, three options are available for gaining control laparoscopically (**Figure 2.19**): a) compression, b) coagulation, and c) clipping, followed by irrigation and aspiration (i.e., ISR B) to clear the field. Since moving from one option to another is possible, at this point we introduced the notion of 'jumping' between activities by using the 'option' symbol. If control is not achieved within approximately one to two 2 minutes, or there is a persistent blood spillage, then conversion to open surgery is necessary.

[1] Note that any unexpected event can be handled using this representational structure by defining an 'Undefined ISR'. Thanks to Dr. Elizabeth Croft for this suggestion.

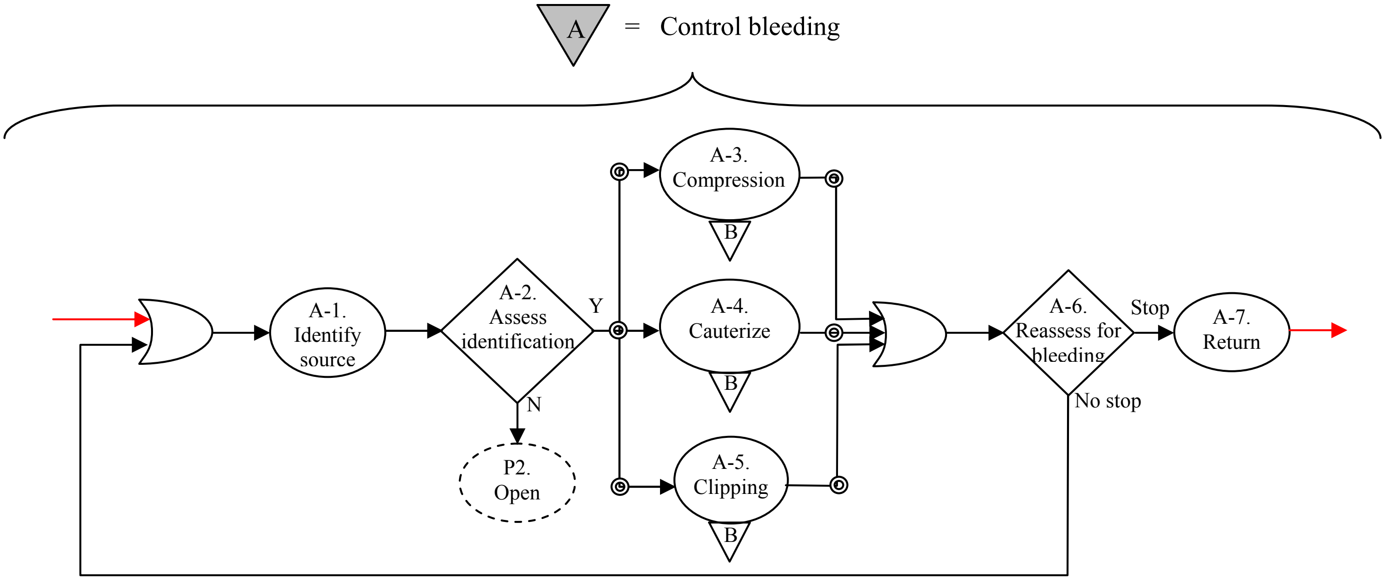


Figure 2.19: MCMD for ISR A: ***control bleeding.***

'Clean-up' (**Figure 2.20**) is an ISR that may be invoked any time the surgeon feels the need to clear the operative field from blood, bile or debris. Continuous irrigation and suction are the actions to perform.

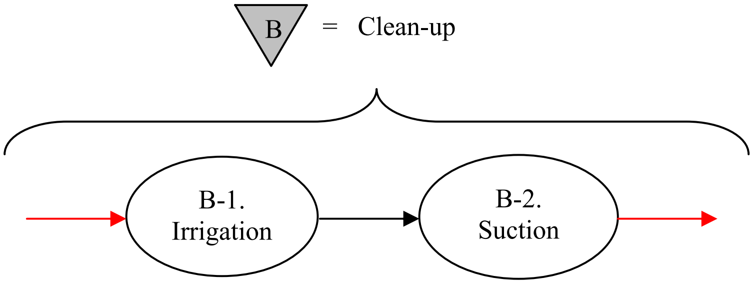


Figure 2.20: MCMD for ISR B: ***clean-up.***

- ✓ *Cholangiogram* [Reddick 1993]. The main objective of this task is to confirm identification of the most important structures involved in dissection of the gallbladder. A cholangiogram is helpful to deal with the most important anomaly: a short CD entering the common hepatic duct in which the common bile duct (CBD) is mistaken for the continuation of the CD; or to assess common bile duct injury.

  The technique involves double-clipping of the gallbladder end of the CD, leaving the medial end patent. A cut is made on the anterior wall of the CD, then a catheter is inserted through the CD into the CBD and a contrast fluid is injected during image intensification to record the early phases of duct filling. If reconfirmation of anatomy is not achieved, conversion to open is necessary. Otherwise, verification of the presence of CBD stones (CBDS) is performed before proceeding. Two options arise for CBDS depending on surgeon's decision: conversion to open or CBDS treatment (i.e, post-operative ERCP-endoscopic retrograde cholagiopancreatography or laparoscopic CBD exploration), (**Figure 2.21**).

  Also, in this task, three different contexts indicate three different possible routes to go through before reaching the following task. The OR symbol points out that following either route is sufficient to move into the following task.

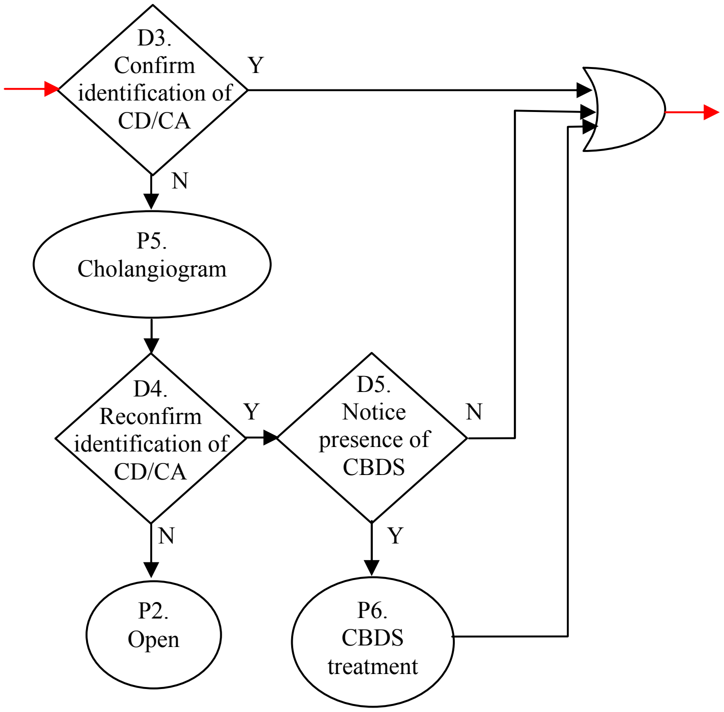


Figure 2.21: MCMD for ***cholangiogram*** at the task level.

✓ *Isolate CD/CA (part I)* [Cuschieri 1990] (**Figure 2.22**). Immediately after the isolation and identification of CD and CA, separation of those structures is performed. The order of execution depends on the surgeon's decision based on anatomical findings; however in the usual anatomic position, the CD is dissected and divided first, as it is the structure appearing most anteriorly in the field. To provide generality in our task-level MCMD representation, we allowed for any order of execution as it is shown in **Figure 2.22**, in which the only condition to proceed into 'dissect GB' (**Figure 2.28**) is set by the AND symbol (i.e., both routes need to be passed through before proceeding).

Due to the possibility of damaging the cystic bile duct (CBD), an inspection is necessary to identify whether or not this complication arose during separation of the CD. If a CBD injury is discovered at this point (D7.4), conversion to open is the appropriate decision to make, otherwise, the procedure may proceed (**Figure 2.22**).

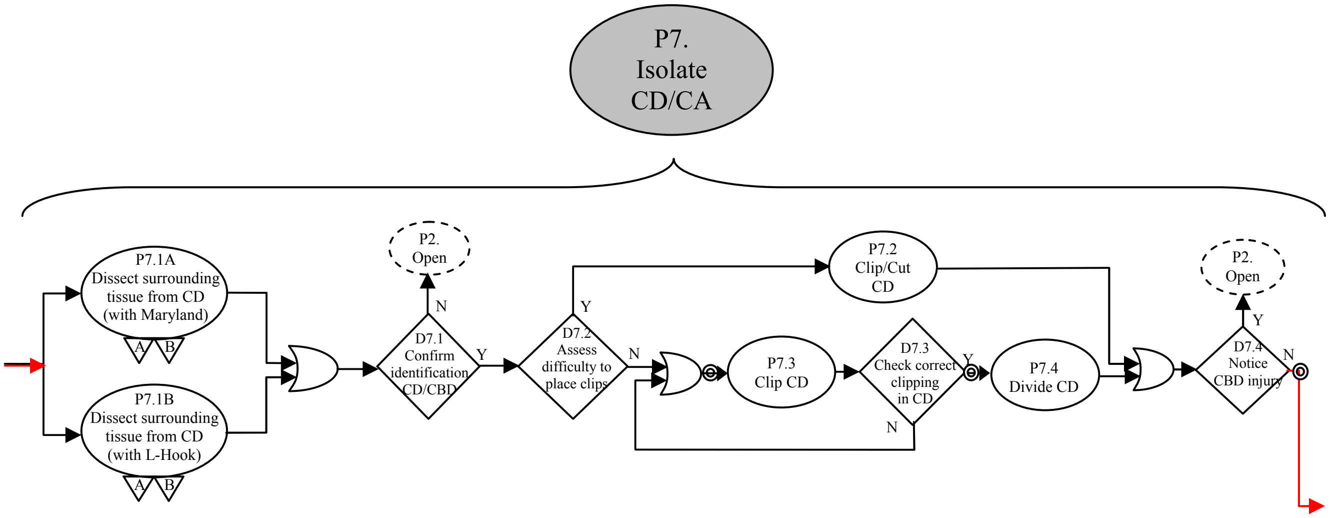


Figure 2.22: MCMD for ***isolate CD/CA (part I)*** at the subtask level.
Refer to figures 2.19 and 2.20 for ISR A and B.

✓ *Isolate CD/CA (part II)* (**Figure 2.23**) [Reddick 1993, Cuschieri 1990]. As mentioned before, there are two options to proceed with the separation of CA: to perform the dissection of CA immediately after mobilization of CD; alternatively, the dissection of the artery is postponed until the CD has been ligated and divided.

The following standard technique is usually applied:

- ‘Isolate CA’ is done by applying tension on the infundibulum of the gallbladder and blunt dissection from the surrounding tissue using either the curved or the L-Hook dissectors (P7.5A and P7.5B).
- Confirm identification of CA due to the possibility of confusing it with the right hepatic artery looping up onto the neck of the gallbladder (D7.5).
- The CA is double clipped both at the patient’s side and at the gallbladder side (P7.6).
- Divide CA (P7.7).

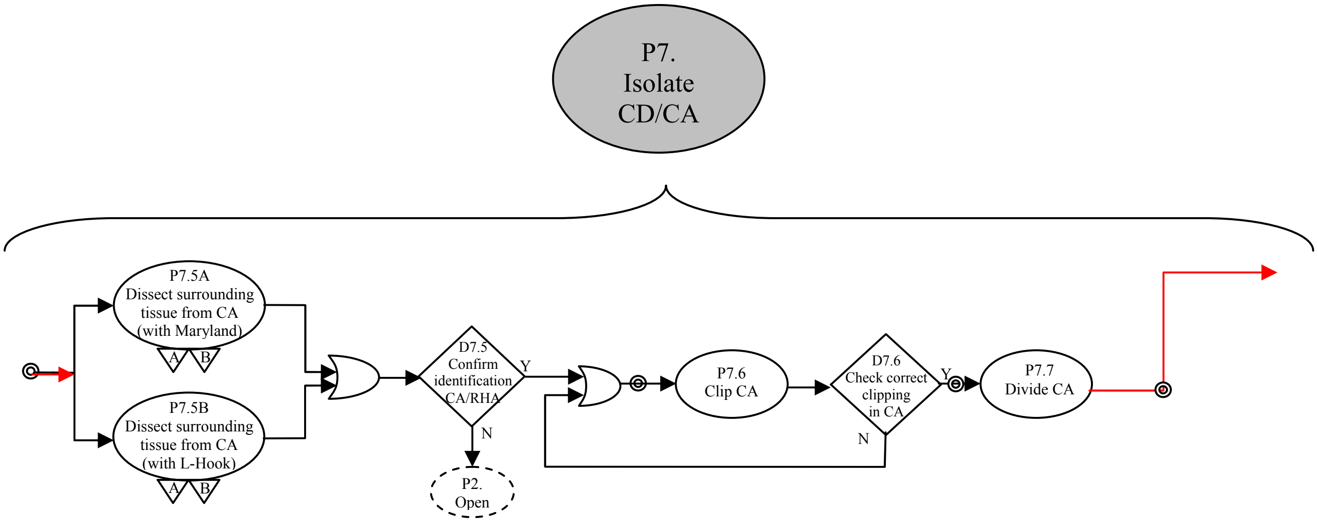

P7. Isolate CD/CA
P7.5A Dissect surrounding tissue from CA (with Maryland)
A
B
P7.5B Dissect surrounding tissue from CA (with L-Hook)
A
B
D7.5 Confirm identification CA/RHA
Y
N
P2. Open
P7.6 Clip CA
D7.6 Check correct clipping in CA
Y
N
P7.7 Divide CA

Figure 2.23: MCMD for ***isolate CD/CA (part II)*** at the subtask level.
Refer to figures 2.19 and 2.20 for ISR A and B.

✓ *Dissect GB* (**Figure 2.24**) [Reddick 1993, Cuschieri 1990]. This task consists of the mobilization of the detached gallbladder from the liver bed. ‘Dissect of GB’ can only start after both separation of CD and CA have been carried out, as indicated by the AND symbol, regardless of the order of execution of these two tasks.

After examining the ligated stumps of the CD and CA to ensure that there is no leakage of either bile or blood, the dissection starts at the gallbladder neck and should proceed along a definite plan towards the fundus as described in the following list of activities:

- The infundibulum is retracted superiorly and laterally as well as distracted anteriorly away from its hepatic bed and dissection of the hepatic fossa is initiated by electrocauterizing either with the curved or the L-Hook dissector (P8.1A and P8.1B).
- Identification of the appropriate plane of dissection (D8.1).
- Separate GB from its bed with electrocautery in sweeping motion creating a horizontal line of dissection (P8.2 and P8.3).

At this task, another interpretation of context arose, in which a decision may lead not only to different types of actions but also to different ways of performing the same action. This is shown in **Figure 2.24** for the decision point “Confirm identification of dissection plane”. The difference between the two displayed routes, lead to the same process “Dissect bed of GB from liver” but under two different contexts. When the dissection plane is not identified, the procedure may proceed but under caution which indicates more mental load on the surgeon. This factor is not present when the plane is effectively identified. We expect that the influence of these two different contexts will produce different performance measurements.

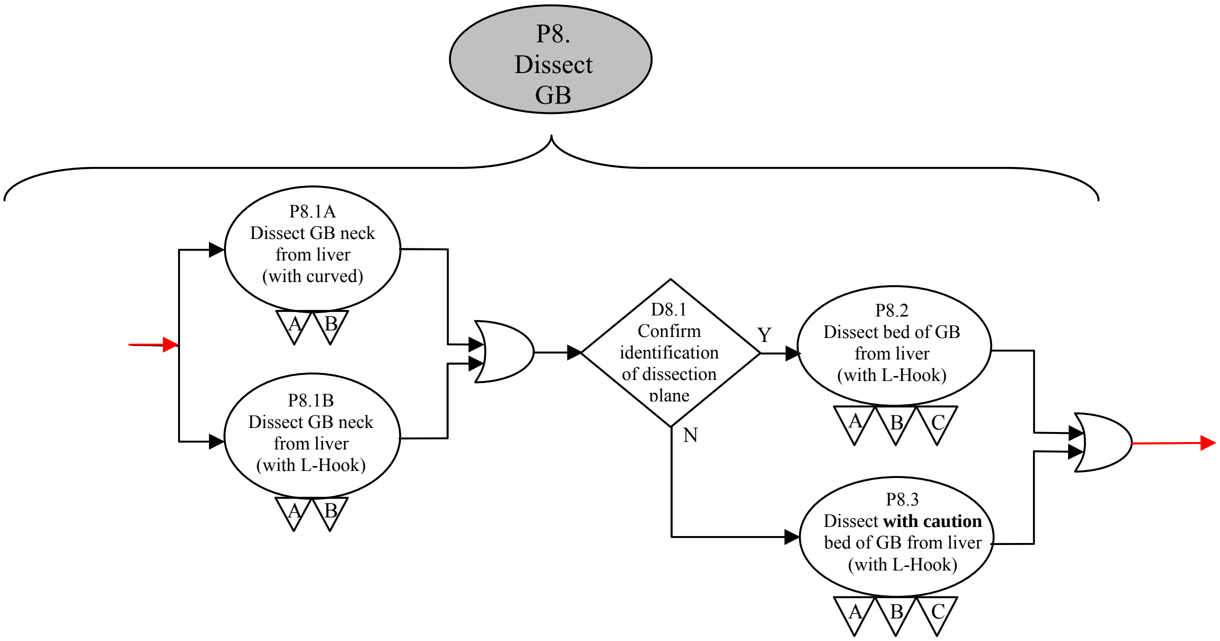


Figure 2.24: MCMD for ***dissect GB*** at the subtask level (Refer to figures 2.19, 2.20 and 2.25 for ISR A, B and C).

As noted in **Figure 2.24**, a third interrupt service routine was included as part of the description: GB stones spillage (C), due to the possibility of damaging the fundus of the gallbladder and having spillage of stones which occurs in approximately 1/3 of the cases.

Situations that may lead to perforation of GB and stone and bile spillage include [McKenzie 2006, Patterson 1997]:

- Damage with the sharp teeth of a grasper instrument or shearing by the back-and-forth traction as GB is moved to enhance exposure.
- GB may be entered inadvertently during its dissection from the liver bed.
- During the force delivery to free a tense GB through a too-narrow umbilical port orifice.

The solution to this problem would be to retrieve all stones immediately, place them in an intraperitoneal specimen bag, and *park* the bag on the liver. Immediately after the GB is dissected off the liver, it should be placed in the specimen bag with the stones and be removed through the umbilical port opening, **Figure 2.25**.

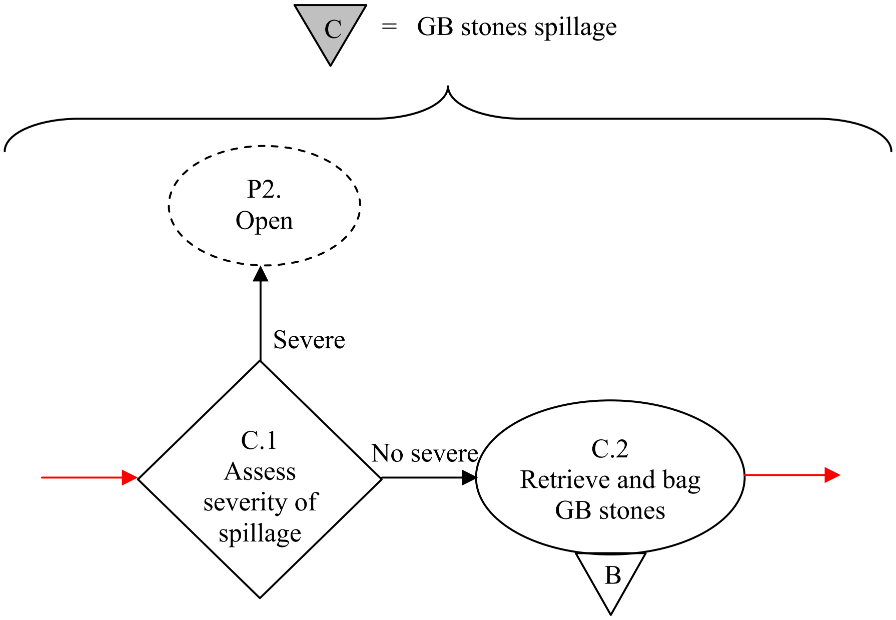


Figure 2.25: MCMD for ISR C: ***GB stones spillage***. Refer to figure 2.20 for ISR B.

✓ *Bag GB* [Reddick 1993] (**Figure 2.26**). The purpose of this optional task is to place the gallbladder into a specimen bag in order to facilitate the extraction process and to avoid spillage of GB stones due to the forces exerted when passing it through the umbilical port.

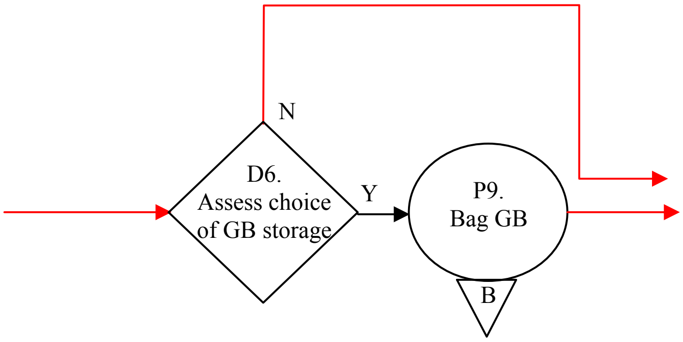


Figure 2.26: MCMD for ***bag GB*** at the task level.

- ✓ *Extract GB* [Reddick 1993]. It corresponds to the removal of the gallbladder from the intraoperative field usually through the umbilicus port. As shown in **Figure 2.27**, this task may be performed after passing through one of two possible routes (i.e., two different contexts), as indicated by the OR symbol: either after dissecting GB or after bagging GB. "Extract GB" constitutes the last intraoperative task.

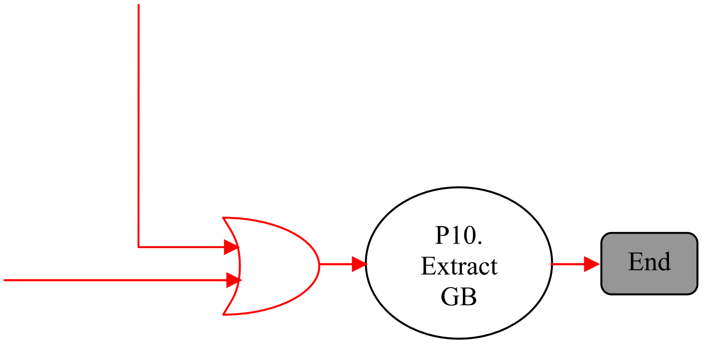


Figure 2.27: MCMD for ***extract GB*** at the task level.

The complete MCMD representation for the task level of the hierarchical decomposition for laparoscopic cholecystectomy is presented in **Figure 2.28** and **Figure 2.29**.

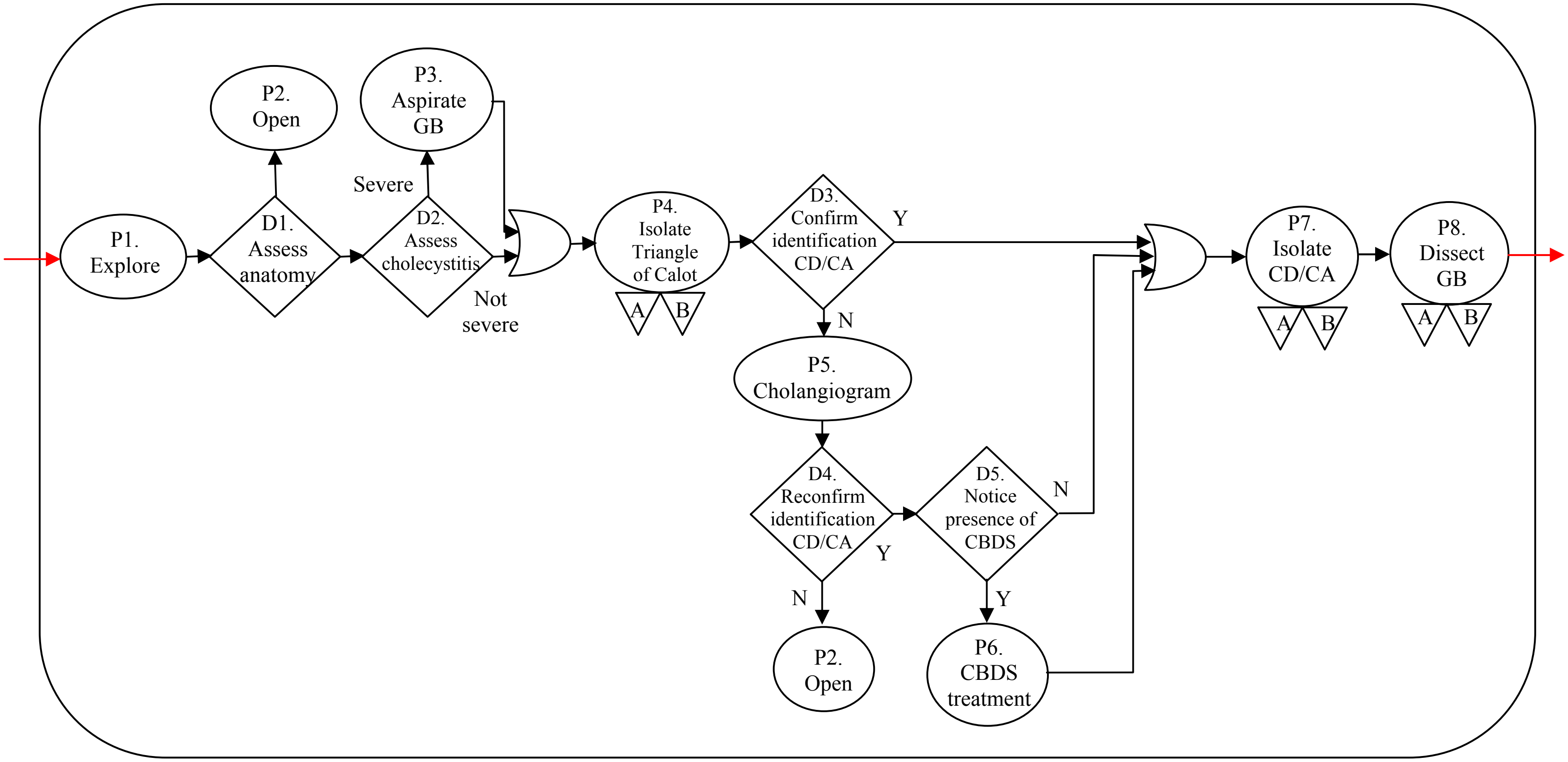


Figure 2.28: Complete MCMD for the task level (Phase: Isolate GB).

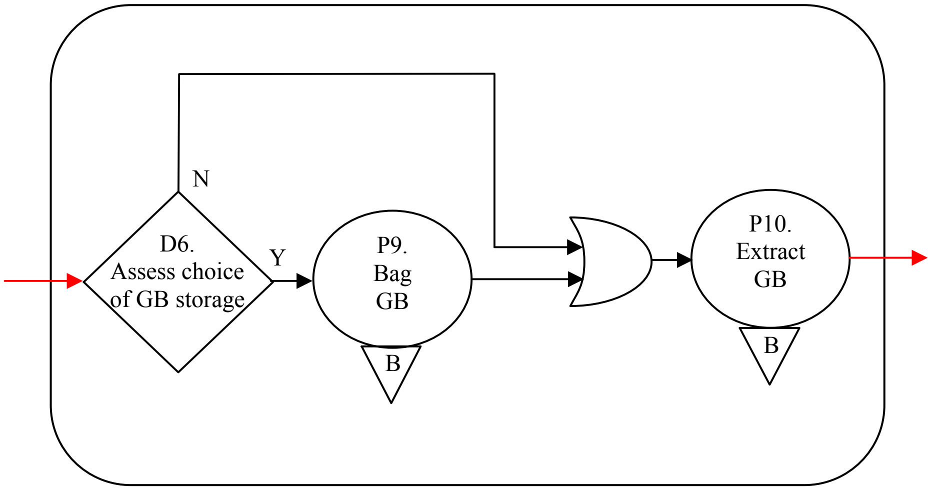


Figure 2.29: Complete MCMD for the task level (Phase: Remove GB).

In defining our MCMD framework we went through various iterations for identifying the necessary set of symbols. We started with the notions from the information processing analysis where processes and decision points were explicitly represented. Then, the need for representing the access to a certain task from different routes (i.e., either by selecting amongst different options or by imposing the requirement of completing previous options before proceeding) led us to include OR and AND symbols. Finally, as the number of iterations in our methodology increased taking us to more detailed levels of description, ‘interrupt service routines’ (ISR) and ‘option points’ were defined in order to account for ‘suspend and resume’ tasks and the possibility of ‘jumping’ between different activities respectively. In the end, we found that these seven symbols were sufficient for providing a thorough motor and cognitive diagrammatic representation for laparoscopic cholecystectomy.

2.4 Validation of Results

In order to provide an argument for using these MCMD symbols to represent other types of MIS procedures, we performed a validation process by creating the corresponding MCMD diagram for Laparoscopic Colectomy procedures (i.e., surgical resection of any extent of the large bowel (colon)).

2.4.1 Procedure

The methods described in section 2.2 were applied as follows.

- We videotaped 6 laparoscopic colectomy (i.e., Sigmoid Colectomies and Right Hemicolectomies) procedures performed by two expert surgeons and manually identified typical surgical tasks and alternatives
- For all of them, we performed a think-aloud process with the surgeons describing performed tasks and decisions in real time
- We then interviewed the surgeons to complement the think-aloud process with any additional information that they considered relevant for the diagram description
- Finally, we used the MCMD symbols to design the corresponding diagrams for both types of colorectal surgeries.

2.4.2 Results

There are three important stages that must be performed for a safe colon resection: Mobilization, Devascularization, and Anastomosis [Zerey 2006, Martel 2006, Finlayson 2005]. The general technique for laparoscopic colectomy involves laparoscopic mobilization and transection of the mesentery and bowel. The anastomosis of the colon can be done either intracorporeally or extracorporeally. Finally, the specimen is removed from the abdomen usually via the same incision through which the anastomosis may be perfomed [Zerey 2006, Martel 2006, Finlayson 2005].

Sometimes, hand-assisted laparoscopic colectomy, which is a hybrid between laparoscopic and open techniques, is used to facilitate the retraction, mobilization, and dissection of the bowel, since the surgeon maintains tactile sensation with the structure [Zerey 2006, Martel 2006].

The hierarchical decomposition for this procedure is presented in **Figure 2.30**. Each task can be decomposed into a set of subtasks, which can be represented with our MCMD representation as shown in Appendix B. As it is a large diagram structure, **Figure 2.31** presents the MCMD corresponding to selected portions of the mobilization and devascularization phases in a sigmoid colectomy procedure. We used the same set of symbols as for LapChole and found that no further symbols were necessary, though the procedural representation was naturally different. This fact highlights the generality of our MCMD symbology and its suitability for providing compact and structured graphical representations for MIS procedures.

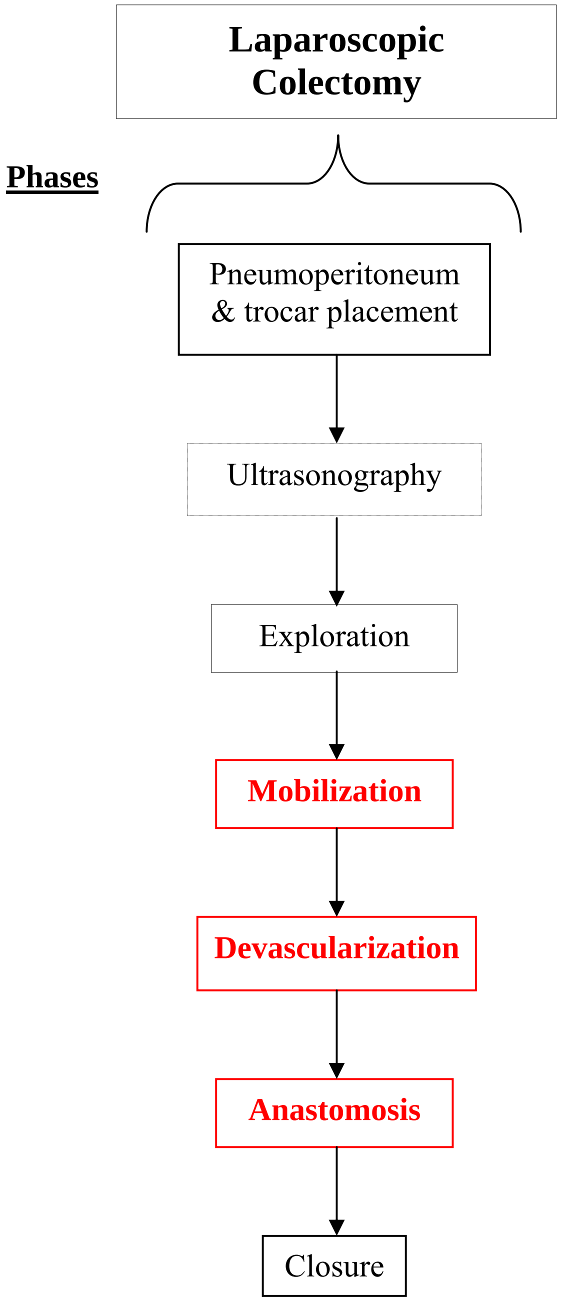


Figure 2.30: Hierarchical decompositions for Laparoscopic Colectomy procedures.

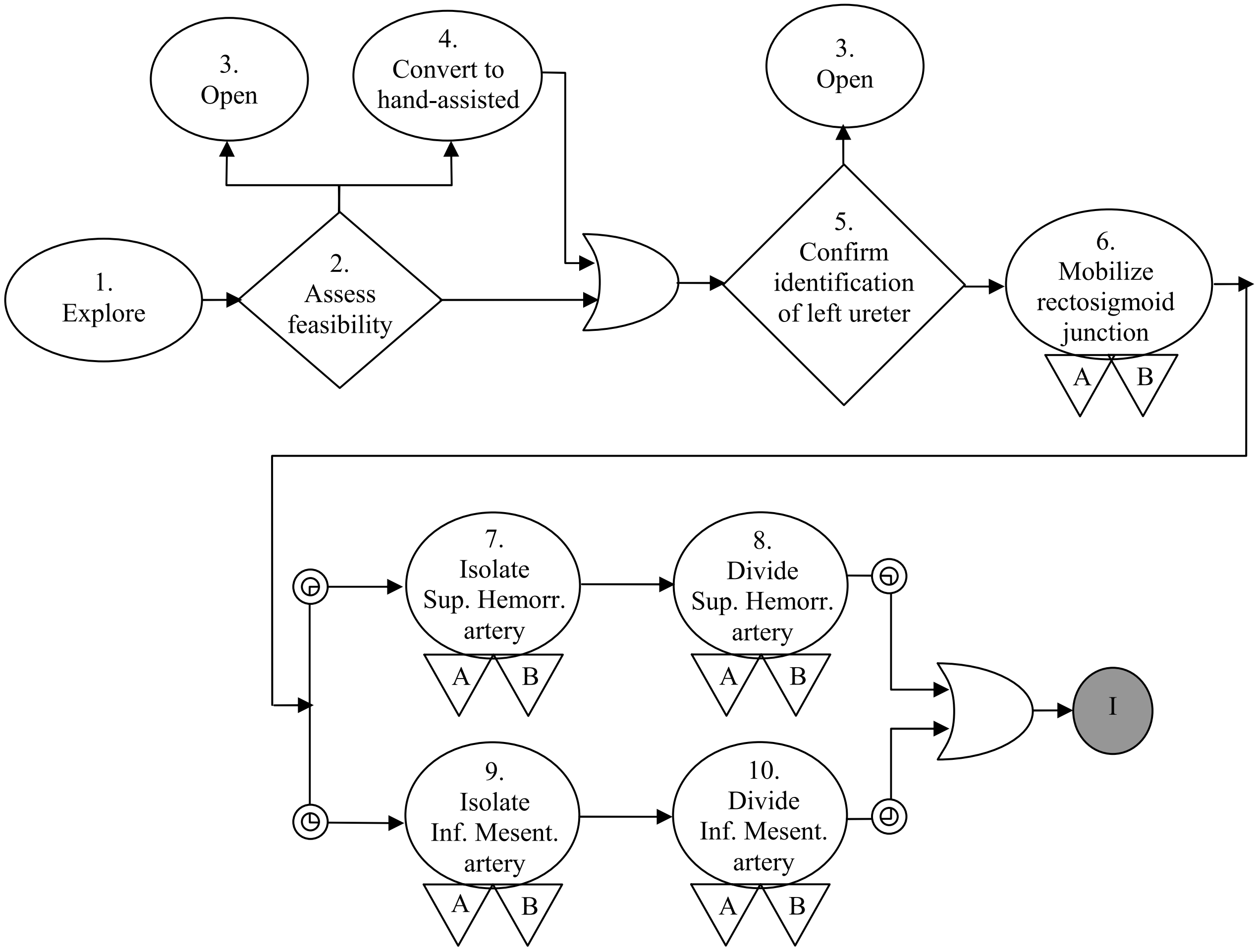


Figure 2.31: MCMD representation for portions of the mobilization and desvascularization phases in a Laparoscopic Sigmoid Colectomy procedure.

2.5 Discussion

In this chapter we have shown how we adapted methods from the cognitive sciences to describe surgical procedures with the aim of identifying key components of the surgical skills, which are important in the training phase. Since minimally invasive surgery requires high technical skill proficiency, we have concentrated on the description of Laparoscopic Cholecystectomy as the most widely practiced procedure in the field of less invasive surgery and one of the earliest ones introduced to trainees. However, for validation purposes, we also developed the graphical description for laparoscopic colon resection surgery.

As shown in **Figure 2.15**, our hierarchical decomposition corresponds well to the results obtained by other research groups [McBeth 2002, Berber 2001, Cao 1996]. It is a four-level representation involving: phases, tasks, sub-tasks, and actions – which forms the basis for our MCMD (Motor and Cognitive Modelling Diagram).

The goal of this part of our research was to provide a method to describe the flow of a surgical procedure by integrating motor and cognitive activities, as a way to represent surgeons' behaviour. Therefore, we developed a symbol notation to represent processes, decision points, transitions, conditions to proceed, and independent routines. The result consisted in a generally left-to-right graphical representation of the flow of the procedure and the possible routes that might be followed to achieve the overall objective (i.e., removal of the gallbladder).

The MCMD constitutes a standard and structured framework for developing objective methods to model surgeon's performance under different surgical contexts (i.e., where 'context' is understood as a group of factors or situations that lead to specific decision and/or actions). It is composed of a set of diagrams for the task and the sub-task levels of the hierarchy. **Figure 2.28** and **Figure 2.29** constitute the representation of the complete task level and **Figure 2.18** (i.e., isolate GB), **Figure 2.22** and **Figure 2.23** (i.e., isolate CD/CA), and **Figure 2.24** (i.e., dissect GB) correspond to the three key components of the procedure. Accordingly, MCMD's notation allowed for representing ISRs (interrupt service routines) as processes that may be executed if necessary (i.e., control bleeding, **Figure 2.19**; clean-up, **Figure 2.20**; and GB stone spillage, **Figure 2.25**).

At the action level, there is no sequence analysis (i.e., no MCMD) since any subtask may be achieved using many different combinations of elemental motions; however, sequence analysis at task & subtask levels are necessary to provide the appropriate context to analyze elemental motions. These elemental motions are in fact the elements that might be quantified using motor performance metrics (e.g., kinematics & forces).

The described results showed the potential of our MCMD to become a standardized notation for modelling laparoscopic surgical procedures in a way that enables an individual to understand differences in surgical performance as deviations from the normal procedural path. Moreover, it serves as a structured framework for including objective performance measures (e.g., time, kinematics, forces) in developing context-based surgical assessment systems. In chapter 3, we will present a new surgical performance assessment methodology we have developed based on this MCDM framework.

Chapter 3

Performance Assessment Analysis for MIS skills

3.1 Introduction

Surgical competence involves numerous elements such as knowledge, judgement, communication, and manual dexterity. Due to the increasing technical difficulties involved in performing more advanced minimally invasive surgical procedures, there is widespread interest in designing objective methods for monitoring skill development in surgeons-in-training which incorporate performance measures such as time, tool kinematics and interaction forces [Seymour 2004, Smith 2001, Rosen 2001]. Such quantitative measures are expected to be useful for several purposes [Khan 2005]:

a) Monitoring of training – to compare one's relative performance with respect to one's peer group to identify particular difficulties and strengths, or to compare performance of residents under one's supervision

b) Comparison of different training programs – to assess effectiveness and quality of training in different programs

c) Selection of candidates – to determine the feasibility of using motor skill measurements to screen candidates for special training programs[1]

[1] This proposed use is likely to be controversial, but is included here for the sake of completeness.

d) Evaluation of new surgical instruments – to assess the impact of newly designed surgical tools on the flow or dynamics of the procedure and consequently on surgeons' performance

However, quantitatively assessing the motor skills of surgeons in the operating room remains problematic and has become an important research topic since current formal structured evaluation methodologies are time consuming and somewhat subjective. Most current approaches rely primarily on comments from the trainees' attending surgeons, which have been shown to be subject to bias [Khan 2005]. Therefore, in order to better monitor the progress of trainees, quantitative and time-efficient methods are required to evaluate the trainees' developing motor skills in the live operative setting [Moorthy 2003]. Furthermore, variability from one procedure to another represents a significant challenge that needs to be addressed while developing these methods [Aggarwal 2007, Datta 2006, Dosis 2005, Bann 2003, Darzi 2001].

Because of this interprocedure variability, most research on technical skill assessment in laparoscopic surgery has been performed on simulators [Aggarwal 2004], where such variability can be eliminated and has focussed on analyzing generic motor skills [Sarker 2006, Taffinder 1998, Martin 1997], but it is not yet clear how relevant such assessments are to skills performed in the operating room. Our group has developed a hierarchical motor/cognitive modelling approach (the MCMD presented in chapter 2) that should enable us to represent live surgical tasks 'in context' and thereby facilitate making comparisons across real procedures and incorporating a variety of objective performance measures [Cristancho 2006].

In this chapter we describe a new methodology based on the Motor Cognitive Modeling Diagram (MCMD; see Chapter 2) for analyzing how surgeons handle their surgical tools. Our main goal is to test whether quantitative data analysis based on performance measures such as holding times, tool kinematics and patterns of movement transitions is able to distinguish skill levels in the OR.

After presenting our proposed methodology in this chapter, in chapters 4 and 5 we will describe its implementation in a simulation study (chapter 4) and in an intraoperative study (chapter 5) which concentrates on answering our five primary research questions presented in chapter 1:

a) Can quantitative measures reliably characterize surgical motor performance?

b) Do surgeons at similar stages of training exhibit similar patterns?

c) Is there a clear separation of patterns across the training spectrum?

d) What data/measures are most useful in separating surgeons along this spectrum?

e) Can a quantitative analysis produce insights useful for instruction?

3.2 Performance Measures

In Chapter 2 we introduced a new diagrammatic representation (MCMD) for Laparoscopic Cholecystectomies, which describes the procedure goals in terms of surgical activities associated with specific motor and cognitive skills. Data from the use of surgical tools may be attached to the activity nodes in the diagram, which will facilitate the analysis of surgical tool use patterns.

Exposing the Calot's Triangle, dissecting the cystic duct and artery (CD/CA), and detaching the gallbladder from the liver were identified by a group of expert surgeons at Vancouver Hospital as the key surgical steps in Laparocospic Cholecystectomy. The surgeons noted that the first two demand the highest levels of technical proficiency and were therefore selected as our main focus for study.

Preliminary video analysis from procedures executed by the expert surgeons involved in this study indicated that a curved dissector and an L-Hook dissector were the tools of choice for the dominant hand, while an atraumatic grasper was the primary tool used in the non-dominant hand (**Figure 3.1**).

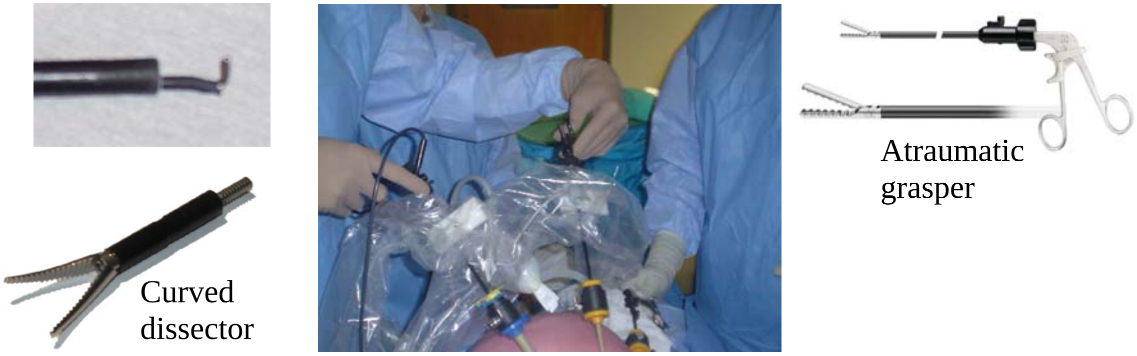


Figure 3.1: Surgical tools used for Laparascopic Cholecystectomy at UBC Hospital

Different surgeons expressed different preferences for using the curved or L-Hook dissectors during the Calot Triangle Exposure and Cystic Duct and Artery Dissection tasks. This preference could be affected by the specific anatomical characteristics of a given patient, but there was an overall preference for using the curved dissector for these two tasks. All surgeons preferred to use the L-hook dissector for the gall bladder detachment task.

In this section, we discuss what kinds of measurements can potentially be acquired in the operating room, how they are related to the assessment criteria identified in Section 1.1.2, and how they can be acquired in the operating room.

3.2.1 Candidate Performance Measures

It is clear from the literature that time has become an important variable for measuring performance and that it can be reasonably easily, though somewhat laboriously, derived from video recordings (Keyser 2000, Fried 1999, Derossis 1998, Starkes 1998, Taffinder 1998). However, by itself, time only partially describes motor performance and does not always correlate strongly with other measures of skilful tool use [Childs 1980].

Previous studies in our lab have demonstrated the feasibility of using tool kinematics to characterize motor performance of an individual surgeon [Kinnaird 2004, McBeth 2002]. Position tracking proved to be practical in the OR environment and kinematics measures were comparatively easily extracted from position data [Torkington 2001, Rosen 2001]. Similarly, the Imperial College Surgical Assessment Device (ICSAD) utilizes an electromagnetic tracker attached to the surgeon's hand to track hand movements on a standardized task [Grober 2003, Smith 2001, Taffinder 1998].

Forces and torques exerted by the tools on operative tissues have also been examined, both in the form of grip force and tool tip forces [de Visser 2002, Morimoto 1997]. Rosen's group at the University of Washington has done extensive work using force/torque signatures to evaluate performance in a porcine model. Our group made similar modifications to a surgical tool and acquired force measurements in live surgeries; however, these modifications resulted in a cumbersome tool for the surgeon and made data gathering more complicated [Kinnaird 2004]. In addition, it is difficult to measure the trocar interaction forces, which introduces a confounding factor that is not easily dealt with.

Since our primary focus is on developing a new method for describing and assessing performance, we concentrated on analyzing more readily obtained measures: time, kinematics and movement transitions. Other measures we might consider, such as tool-tissue interaction forces or physiological measurements such as stress and pulse rate, would potentially involve modifications of the tools or 'interference' with the surgeon and we therefore will defer them to future studies.

3.2.2 Measurement and Preliminary Data Processing

In this section, we concentrate on describing the general methodology of our approach; discussion of methods specific to the simulator and OR experiments (e.g., numbers of subjects and procedures and other details) will be addressed in Chapters 4 and 5, respectively. However, for explanatory purposes we will use the operating room experiment as the context for describing how we acquire measurements and extract the desired information from the whole procedure data stream. In this chapter, we will focus on the data processing techniques used in the following two chapters.

3.2.2.1 Sequencing and Time

It is feasible and practical, though somewhat tedious, to use video recordings to assess the times and state transitions for states at all levels of the hierarchical decomposition diagram described in Chapter 2. In the OR study, we use a video analysis to identify the surgical steps executed by the surgeon based on our MCMD decomposition. We therefore divide the whole video into small video clips representing the surgical *tasks* of the procedures (i.e., ‘Isolate Triangle of Calot’), for which we annotate the start and end times and identify the order of execution.

Individual video clips are then analyzed and segmented in the same manner to separate the tasks into its constituent *subtasks* (i.e., ‘Expose Triangle’ and ‘Dissect CD/CA’). A new set of video clips are then obtained and further decomposed into their component *actions* (i.e., push, pull, reach, orient, sweep, etc). The time records and order of execution for subtasks and actions are also collected. In order to be consistent during the video segmentation process, we assure that the MCMD decomposition at each level of the procedure provides an explicit description of when a task, subtask and action initiates and ends (Chapter 2). **Figure 3.2** presents an example of a video frame for the laparoscopic cholecystectomy procedure that we will use in the OR study. Left and right pictures present the start and end points of the ‘Dissect CD/CA’ subtask respectively. The intra-abdominal image is digitally captured using a Stryker Laparoscopic Camera system and the iMovie software from Apple operating system is used to segment the video clips and to register the time records with a resolution of 33 milliseconds (1 frame, which lasts 1/30th of a second). As this task segmentation process is performed manually, it is exhausting and time consuming; therefore, automatic methods need to be developed in the future to make this methodology more robust and practical.

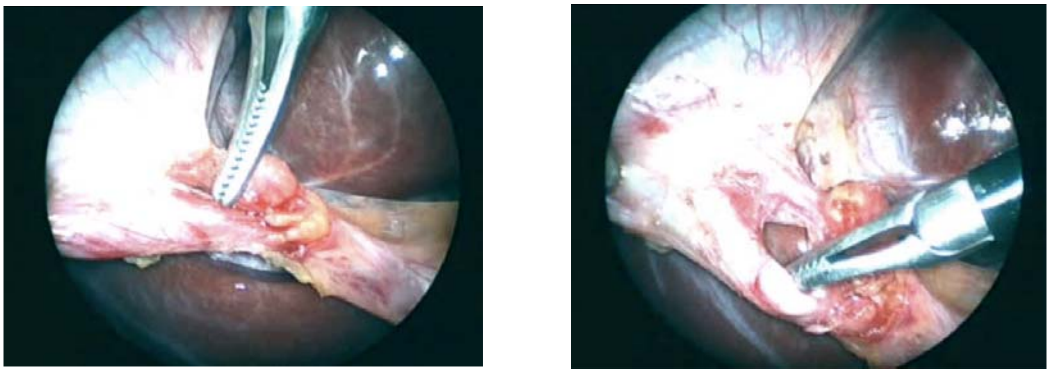


Figure 3.2: Example of start (left) and end (right) points for the 'Dissect CD/CA' subtask during a laparoscopic cholecystectomy procedure. 'Expose Triangle' initiates when the gallbladder is first stretched out and finalizes when the cystic pedicle is identified; 'Dissect CD/CA' initiates when the tip of the tool is first inserted between the two anatomic structures and finalizes when both structures have been completely freed from each other.

While in principle we would want to analyze performance for all tasks and subtasks, in this research we will concentrate particularly on the Expose Calot's Triangle (ECT) and Dissect Cystic Duct and Artery (DCDA) tasks as they include the most demanding activities for surgeons in terms of motor skills.

3.2.2.2 Kinematics

In this section we will describe the position tracking system that we chose and the data processing steps we follow for obtaining the desired performance measures.

Position measurement system:

In our experimental set up, we use an electro-magnetic Polhemus FASTRAK system, which continuously records 3D position and orientation data at 120 Hz from a receiver relative to a transmitter (static accuracy of 0.03 inches RMS for the X, Y, or Z position; 0.15° RMS for receiver orientation [Polhemus 2002]). We provide two custom-designed clips to which the small tracking sensors (approximately 1 cm^3) were attached; at the beginning of the procedure, the surgeon attaches the clips to the surgical tools so that we could track the tools' position (**Figure 3.3**).

As will be explained in Chapter 5, in the OR study, we will be focusing on certain tasks where surgeons are most interested in looking at how residents identify and dissect anatomical structures of the gallbladder and the liver. These tasks can be achieved by using the Curved dissector, the L-Hook dissector or a combination of the two tools. However, given that whenever the L-Hook dissector was used, the position signal was corrupted due to an alteration in the electromagnetic field promoted by the operational feature of this tool, surgeons agree to use the Curved dissector instead. We considered fiber optics as an alternative to resist noise from L-Hook but since we did not have this system available, it was not possible to incorporate it in the present study (Appendix C). However, since the two tasks the surgeons are most interested in could be, and usually are, performed using the Curved dissector, we limit all the surgeons to using that tool in our studies.

Processing data:

There are two key stages of data processing: (1) computing transformations to express the data in an appropriate reference frame, and (2) filtering the data to obtain the desired voluntary movement information.

Since surgical tools may be used at relatively arbitrary orientations relative to the operating table, there is little natural relevance of the table's coordinate frame to that of the surgical tasks. We therefore opt to describe all motions relative to the tool's instantaneous tip location, using the terms lateral, vertical and axial to describe the motions.

In order to obtain the location and orientation of the tool tip at every instant, we locate an 'instantaneous' inertial reference frame at the sensor position and report the tip motion with respect to it (**Figure 3.3**). Therefore, the 'instantaneous' reference frame at the sensor position changes at every sensor reading with respect to the global frame, but remains fixed with respect to the tool tip movement.

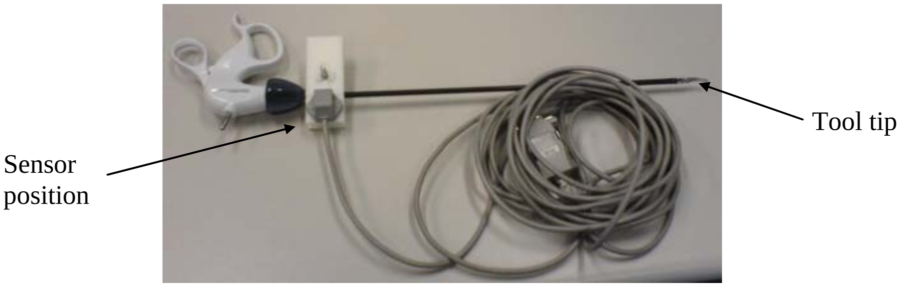


Figure 3.3: Tool tip and sensor locations. The tool tip motion is reported with respect to an instantaneous inertial reference frame is located at the sensor position

Therefore, by obtaining the location of the sensor with respect to the transmitter (global frame) at every instant (i) and performing a calibration process (described below) to compute the relative position of the tip with respect to the sensor position (x_t^s), we calculate the position of the tip and consequently its motion with respect to the global frame at every instant (Equation 3.1, t: tip location; s: sensor reference frame; G: global reference frame).

$$(x_t^G)_i = (T_s^G)_i x_t^s \qquad \text{Eq. 3.1}$$

Afterwards, we are able to compute the instant-by-instant (i) motion of the tip in terms of the instantaneous sensor inertial frame (Equation 3.2).

$$\left(\overset{\bullet}{x_t^s}\right)_i = \left(T_s^G\right)_i^{-1}\left(\overset{\bullet}{x_t^G}\right) \qquad \text{Eq.3.2}$$

In order to find the calibration transformation matrix between the sensor and the tip of the tool (i.e., to find the position of the tip with respect to the sensor, x_t^s), we first estimate the tip location as the center of a sphere obtained by moving the tool handle in circles about a fixed tool tip location (we use a testing table and define an origin point for positioning the transmitter; then in the center of the table, we place a wooden cube with a hole through which the tip of the tool is inserted in order to fix its location with respect to the transmitter). We then use the collected sphere data and apply a non-linear least squares optimization approach in order to find the location of the tip (center of the sphere (x,y,z)) with respect to the global frame (Equation 3.3).

$$(\hat{x}_t^G)_i = (T_s^G)_i \, \hat{x}_t^s \quad \text{Eq. 3.3}$$

Initial estimated value $\hat{x}_t^s$: (0, 30, 0) cm

Optimization problem: Minimize C($\hat{x}_t^s$) where $C = \sum_i \left[\left(x_t^G \right)_i - \left(\hat{x}_t^G \right) \right]^2$

In general, we are not only interested in position but also in velocity, acceleration and jerk[2] in each of the cartesian directions (**Figure 3.4**).

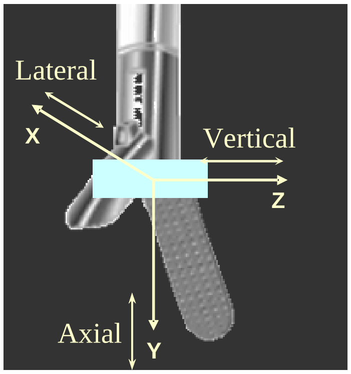


Figure 3.4: Tool tip cartesian directions

Since we are making discrete time position measurements, we need to design an appropriate filtering and differentiation process to estimate the position derivatives. We use a Generalized Cross Validation (GCV) technique to find an optimal smoothing parameter for fitting a spline (i.e., a smoothed polynomial) to the position data set [Hodgson 1994, Dohrmann 1988]. Then the derivatives of the fitted polynomial result in estimates of velocity, acceleration and jerk.

[2] We measured position and extracted derivatives up to jerk measures, since further derivatives implied adding noise to data.

A frequency spectral analysis is also completed using a Fast Fourier Transform (FFT) to show signal power levels of the position signal and higher order derivatives. **Figure 3.5** shows the FFT spectrum for the raw position data in the axial direction derived from all subjects during execution of the 'peeling' subtask from the simulator experiment described in Chapter 4. No frequency content was found near 60Hz, which indicates that the system did not suffer from undersampling problems.

This task, which involves only hand movements, should only produce voluntary frequencies below 15Hz [Zhang 2005, Raethjen 2000]; however, our FFT analysis shows higher frequencies between 20 and 40Hz. We suspected that these are likely due to structural resonances in the tool, so we computed the frequency of the first mode of vibration for the surgical tool (Curved dissector) assuming that it behaves as a simple cantilever beam. The natural frequency for the beam and the structural parameters of the tool are as follows:

$$f_n = \frac{C_n}{2\pi}\sqrt{\frac{EI}{\mu L^4}}$$

f_n = natural frequency in cycles per second (Hz)

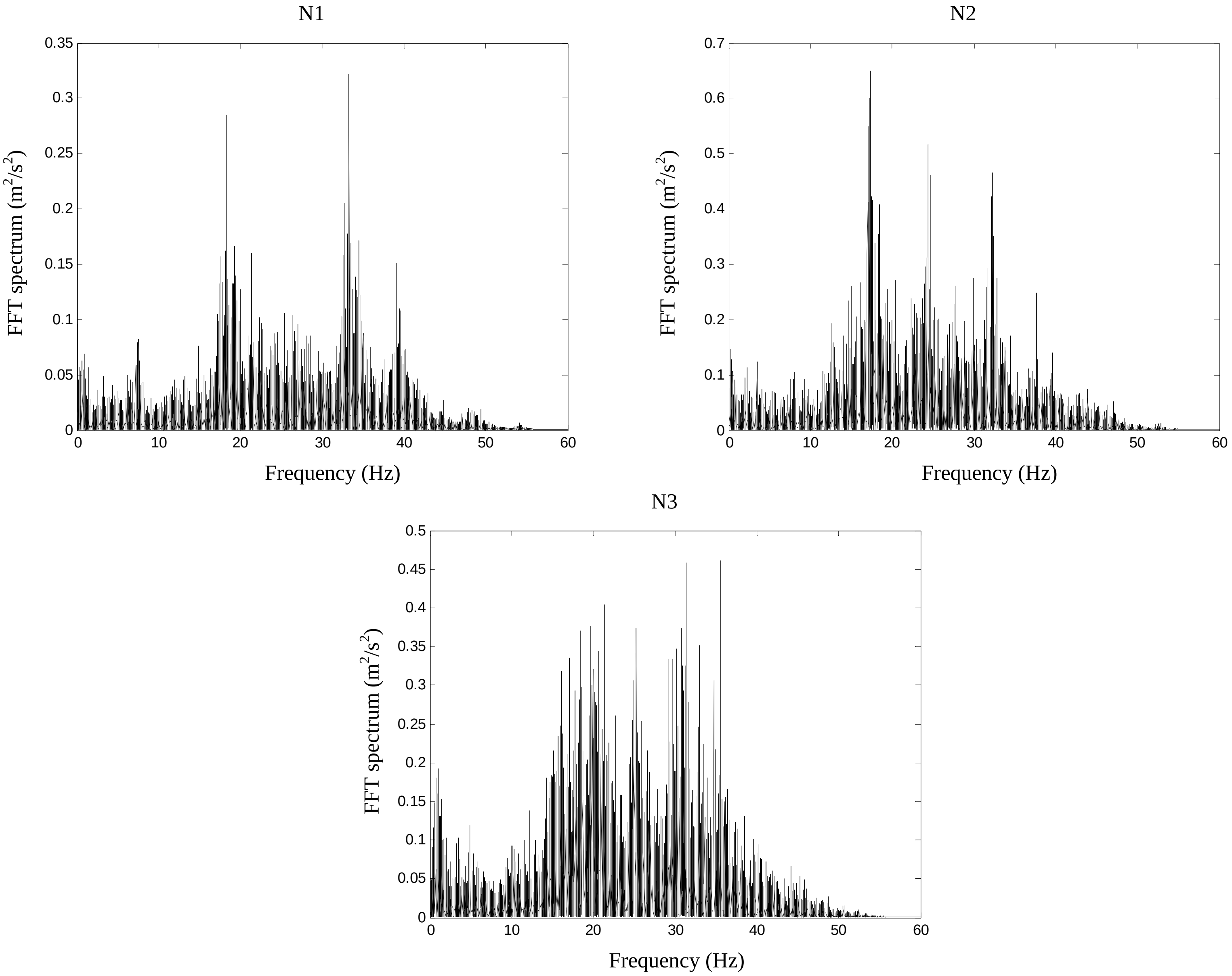
E = modulus of elasticity of the material

I = moment of inertia

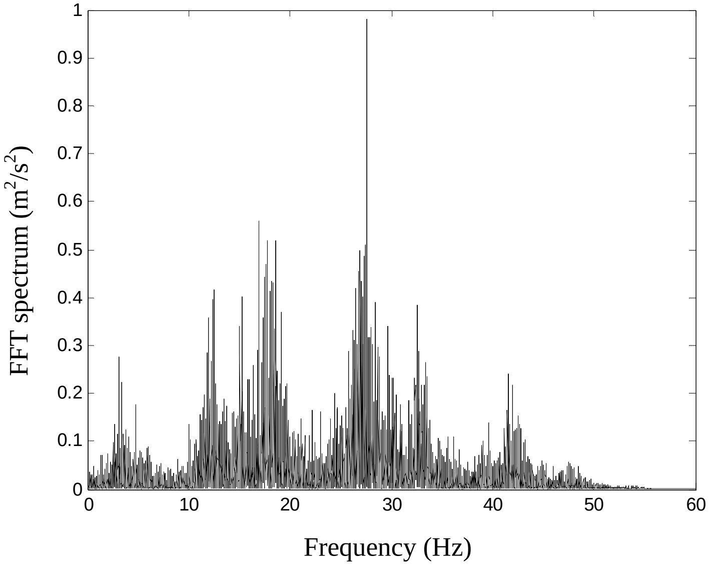
L = length

μ = mass per unit length

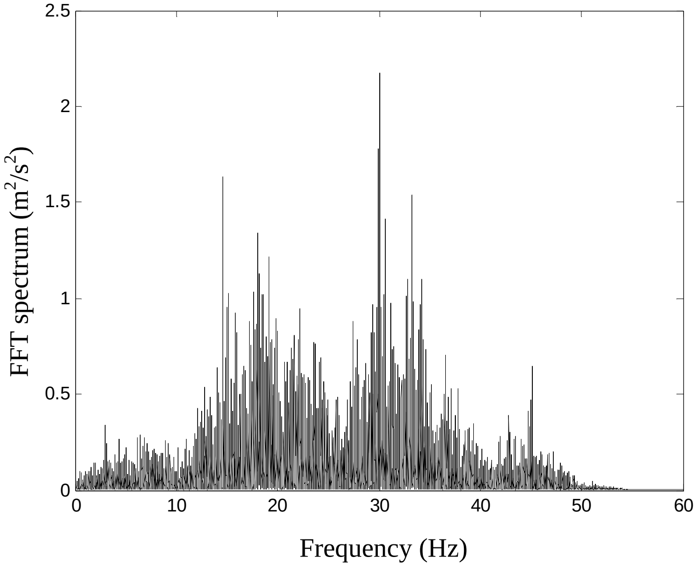
C_n = coefficient for the different resonant modes (C_1 = 3.52, C_2 = 22.4, C_3 = 61.7)


N1
FFT spectrum (m^2/s^2)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
Frequency (Hz)
N2
FFT spectrum (m^2/s^2)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
Frequency (Hz)
N3
FFT spectrum (m^2/s^2)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
Frequency (Hz)

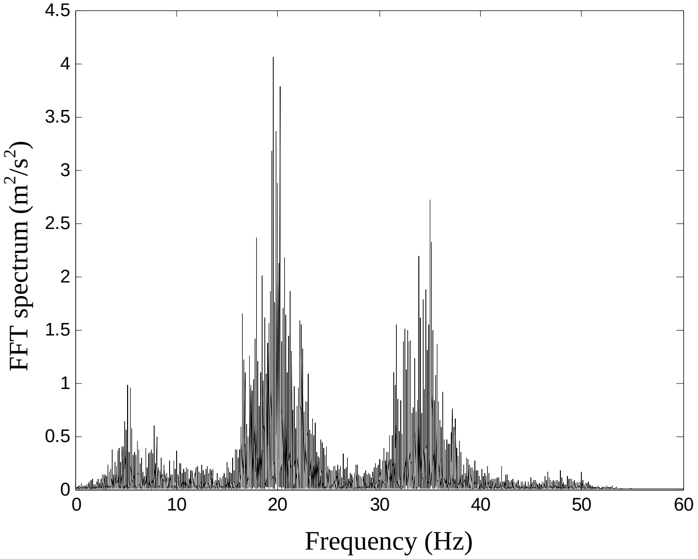
NovT1



NovT2



NovT3



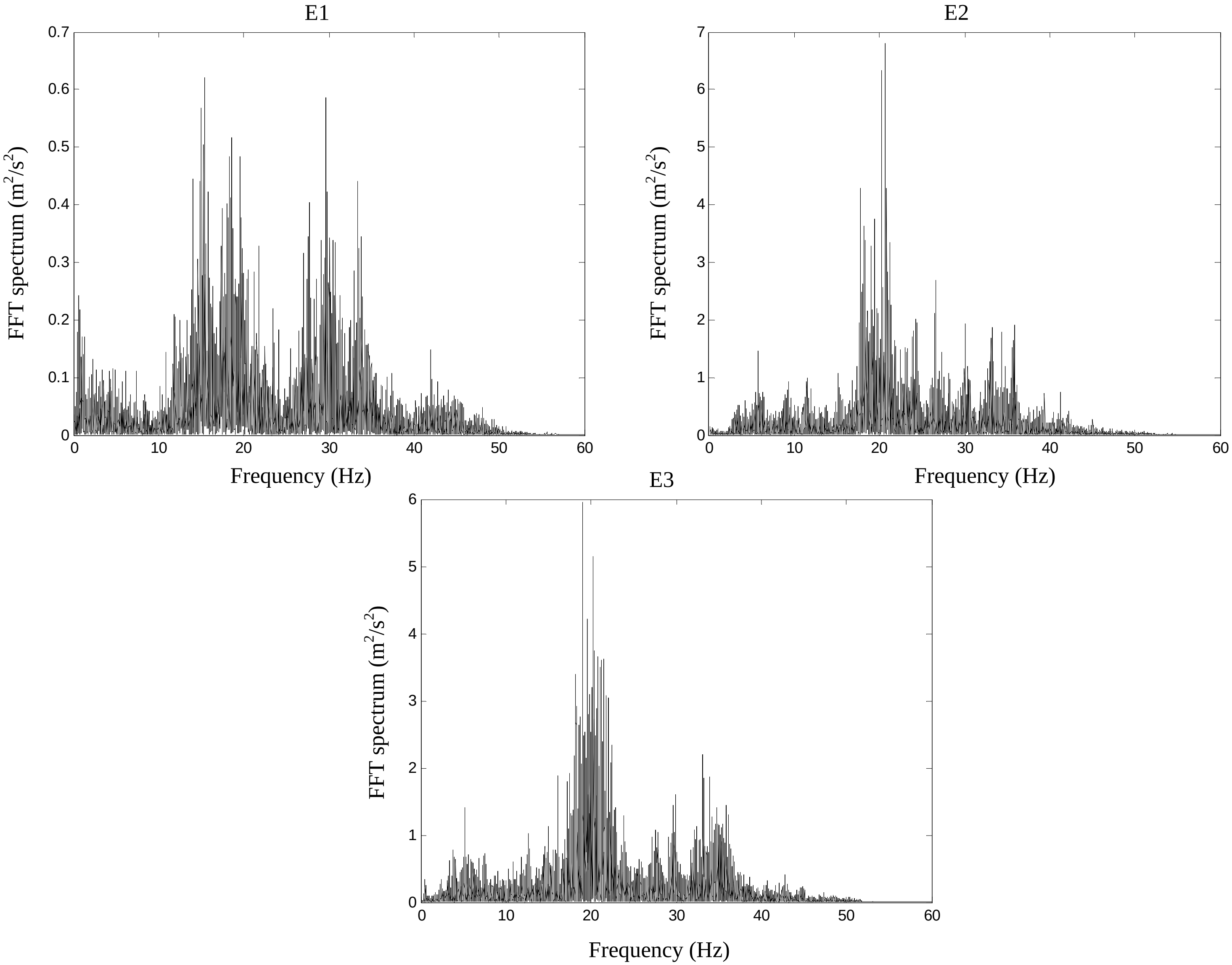


Figure 3.5: Samples of the FFT spectrum for position data in the axial direction for all subjects (N: novices, NovT: novices-with-training, and E: experts) while executing ‘peeling’ during the mandarin experiment described in Chapter 4.

Parameters for Curved dissector:

L = 32cm = 0.32 m

DO (outer diameter) = 5mm = 0.005 m

DI (inner diameter) = 3.5 mm = 0.0035 m

ρ (stainless steel) = 8000 Kg/m^3 → μ = A . ρ = π . r^2 . ρ = 0.1574 Kg/m

E (stainless steel) = 193 x 10^9 Pa

I = π/64 (DO^4 - DI^4) = 0.005π/64 = 2.33 x 10^{-11} m^4

C_1 = 3.52

Hence the first mode of vibration of the Curved dissector is located around ***f*** f_1 **= 29 Hz.** This estimate suggests that structural resonances may well be the source of these higher frequency components in the position signal. Ideally, we would like to filter out these higher frequencies and retain the lower frequency voluntary signals, but, given the relative proximity of the resonance frequency to the voluntary frequencies, we decided to apply a filtering stage with a cutoff around 20Hz by using a modified GCV approach. In order to obtain an appropriate smoothing parameter, we use some of the acquired data as training data and apply a straight GCV algorithm that produces a very low B value (B=10^{-11}), which is then used to filter all the computed kinematics data. **Figure 3.6** shows the FFT spectrum for the post-filtered axial velocity for one typical novice (N1) and one typical expert (E2).

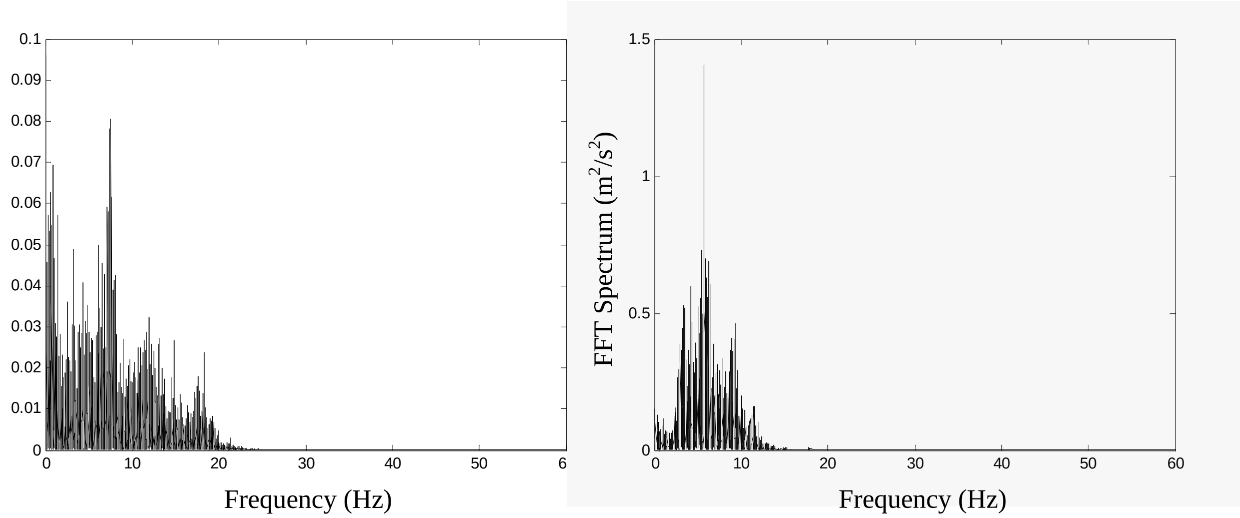


Figure 3.6: Samples of the FFT spectrum after GCV filtering (smoothing parameter B=10-11) for one typical novice (left) and one typical expert (right) while executing 'peeling' during the mandarin experiment

The power spectrums indicate similarity in the frequency component for which the stronger peaks occurs (about 10Hz) but some differences in the magnitude of the power content (Novice in the range of 0.08m^2/sec^2; and expert in the range of 1.5m^2/sec^2), which indicates that the expert is likely moving significantly faster than the novice but at roughly the same frequencies (i.e., moving with larger amplitudes).

3.2.3 Link Between Criteria and Metrics

As part of the collaborative work with the group of surgeons involved in our research, we are committed to present our results in an accessible and comprehensible way for follow-up and feedback purposes. On the basis of the assessment criteria and performance metrics identified previously (Section 1.1.2), we focus our methodology on analyzing two features of strong interest to surgeons:

1. Competence and coordination (bimanual dexterity) in using surgical tools

2. Flow of procedure described as selection of surgical steps and order of execution

Table 3.1 shows how these criteria can be linked to measures of time, kinematics and transitions.

	MEASURE			
CRITERIA	MCMD	Time	Kinematics	Transitions
Competence		✓	✓	✓
Coordination (Bimanual Dexterity)		✓	✓	✓
Flow of procedure	✓	✓		

Table 3.1: Criteria/Metric combinations based on the definition of surgical scenarios and assessment criteria provided in Chapter 1

3.3 Proposed Performance Assessment Methodology

As presented in Chapter 2, our hierarchical representation decomposes larger surgical goals (tasks) into local goals (subtasks) and at the very detailed level into individual movements (actions). Tool use (i.e., competence and coordination) is assessed for each state at each level in the MCMD by describing the time spent in each state, the kinematic profiles (in the form of probability distributions), and patterns of tool use (in the form of transition probabilities between states). Our assessment methodology then builds on this structured data in order to provide an appropriate context for each surgical activity. **Figure 3.7** illustrates the general flow of computations: we begin at the highest levels of the MCMD by computing summary (i.e., descriptive) measures for the surgical tasks, the subtasks and the actions, as indicated by the descending blue line. Next, we compute difference measures at each level, as indicated by the horizontal arrows. Finally, we propagate difference measures from the lowest level to higher ones, as indicated by the green arrow.

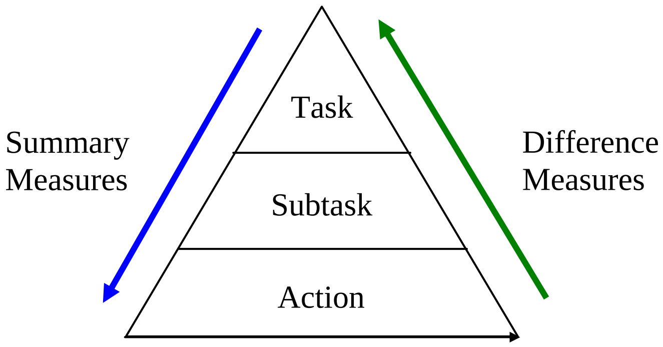
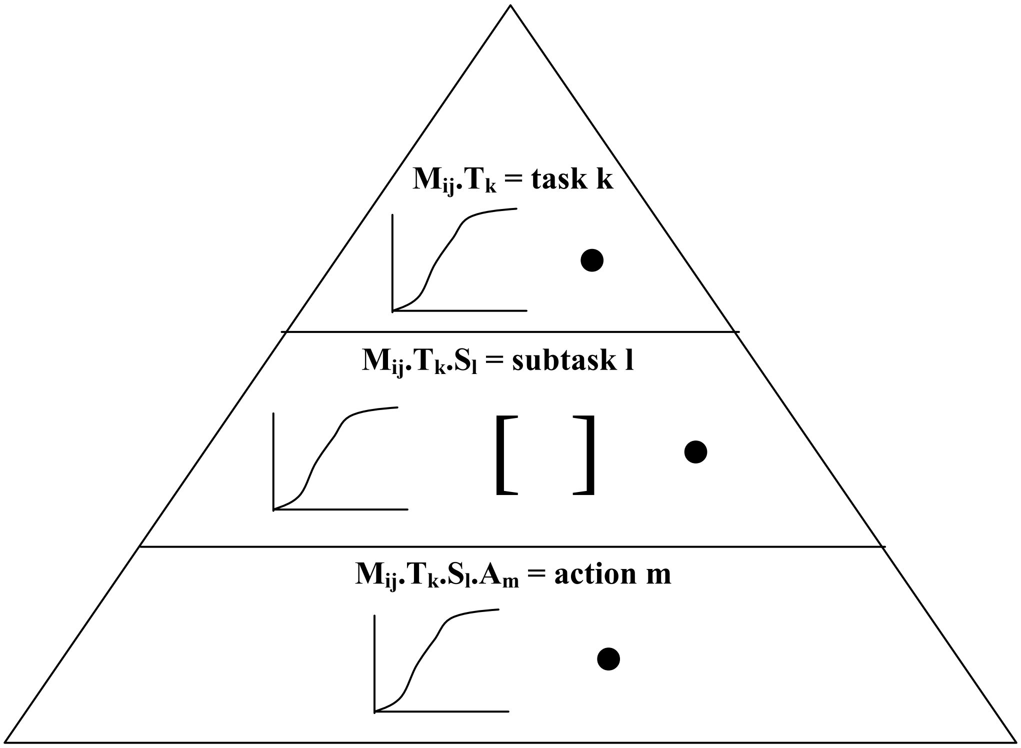


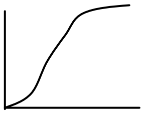
Figure 3.7: Representing downwards and upwards flow of quantitative computations of surgical performance based on the MCMD hierarchical description for minimally invasive surgical procedures. Horizontal arrows indicates that difference measures are computed at the various levels and then propagated up as indicated by the green arrow.

More specifically, from top to bottom in **Figure 3.7**, we use *summary (descriptive) measures* (e.g., average speed, average time) to represent performance at higher levels and to localize possible sources of differences between subjects. From bottom to top in **Figure 3.7**, we use *difference measures* computed at the lowest level in order to perform explicit and quantitative comparisons between subjects, which are then propagated up in the hierarchy to characterize differences at various points of the procedure.

In order to keep track of the different calculations we perform at every level and to be explicit in their description, **Figure 3.8** shows our selected symbology for representing our *descriptive measures* at the task, subtask and action levels for the MCMD from subject i during procedure j, which we will designate as M_{ij}, and across all his/her consolidated procedures, which we will designate as $M_{i\bullet}$.

$\underline{M_{ij}}$ – MCMD for subject i, procedure j



Type of representation	Type of variables
Cumulative Distribution Function (CDF) 	$\underline{t}$: time $\underline{v}$: velocity (l: lateral, a: axial, v: vertical) $\underline{a}$: acceleration (l: lateral, a: axial, v: vertical) $\underline{J}$: jerk (l: lateral, a: axial, v: vertical)
Transition probability matrix (TPM) [ ]	TP_{mn}: transition probability from state m to state n
Point / summary measures ●	$\bar{\underline{t}}$: average time $\bar{\underline{v}}$: average velocity (l: lateral, a: axial, v: vertical) $\bar{\underline{a}}$: average acceleration (l: lateral, a: axial, v: vertical) $\bar{\underline{J}}$: average jerk (l: lateral, a: axial, v: vertical)

M_i• – consolidated across all procedures from subject i

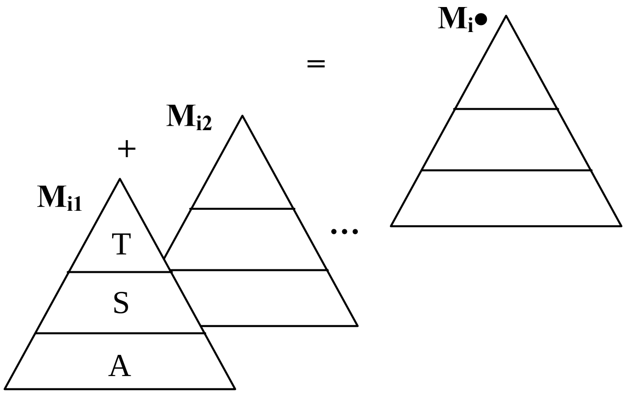


Figure 3.8: Diagrammatic scheme for computing descriptive measures for subject i during procedure j (i.e.,M_{ij}) and across all consolidated procedures (indicated by M_i•, which is formed by concatenating all samples from the different procedures). An underline indicates vectors either of discrete time ($\underline{t}$) and kinematic ($\underline{v}$, $\underline{a}$, $\underline{J}$) measures or average time ($\underline{\bar{t}}$) and average kinematic ($\underline{\bar{v}}$, $\underline{\bar{a}}$, $\underline{\bar{J}}$) measures during single executions[3]. All kinematic measures are acquired along three directions of movement: lateral, axial, and vertical. Other measures could potentially be used (eg., range of movements, forces, etc). TP_{mn} indicates the probability from transitioning from movement m to movement n; every subtask is composed of 10 movements (ie., actions) which are cycled through to achieve the subtask

Figure 3.9 shows the associated symbology for representing *difference measures* when comparing between subjects i and q, which we will designate as D_{iq}. Another possible comparison would be between a subject and a reference group; in this case, we use a capital letter to designate a group, so the notation would become D_{iQ} (ie., comparing subject i and reference group Q).

[3] It is important to note that a variety of summary measures can be used such as mean, median, any other percentile or maximum. In our approach we did not specifically used the maximum as it is a single data point highly variable and therefore unreliable.

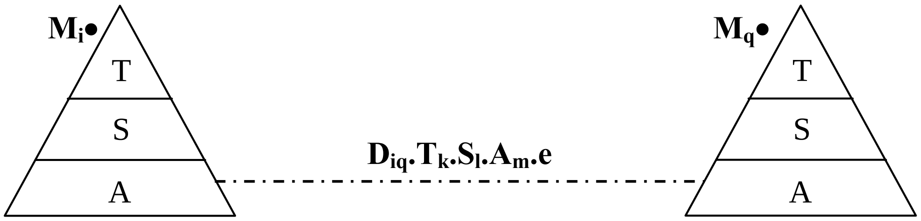


Figure 3.9: Diagrammatic scheme for computing difference measures for some representative parameters (e.g., time, speed) when comparing subjects i and q (i.e.,D_{iq}) across all consolidated procedures for each subject (M_i• and M_q•). For example, $\mathbf{D_{E1R1}.T_1.S_2.A_5.t}$ corresponds to a difference measure computed based on the time distributions for action 5, subtask 2, and task 1 between subject E_1 (i.e., Expert 1) and subject R_1 (i.e., Resident 1)

This computation scheme is based on three analysis stages: *Acquisition, Description,* and *Evaluating Differences*. The next two sections will outline the data processing for the *Description* and *Evaluating Differences* stages. First, we describe how we quantitatively represent at each level of our hierarchy the three previously-described surgical assessment criteria (**Table 3.1**). Then we describe how we use difference measures to perform comparisons between subjects, as well as how we propagate the results of low-level analyses up to higher levels so that it becomes useful for instruction. As described at the beginning of the chapter, there are external factors, particularly in the intraoperative setting, that influence motor performance and which cannot be controlled by the surgeons; therefore, we also outline our approach to determining how repeatable our measures are for individual subjects across multiple procedures.

3.3.1 Describing Data

In this section we describe the computations we perform to describe and summarize the performance of each surgeon during each procedure at the task and subtask level and at the action level in relation to the three surgical assessment criteria identified earlier: flow, competence, and coordination.

3.3.1.1 Task/Subtask Level

Here we introduce how we assess the criteria of flow, competence and coordination at the task and subtask levels using the selected performance measures (**Figure 3.8**). There are no substantial differences in how we analyze performance at these two levels.

3.3.1.1.1 Flow

Motor Cognitive Modelling Diagram (MCMD)

In terms of our MCMD representation, flow can be described as the order of execution of the tasks and subtasks in the MCMD. We assign a new MCMD (Mij) to each subject (i) and procedure (j). This MCMD is a data structure, which consists of several constituent elements as outlined below. In general, we will use dot notation to indicate an element of an MCMD (eg, $M_{ij}.T_1$ refers to task 1 of the data structure M_{ij}).

Flow in a given MCMD is represented as an ordered list of states in the MCMD, which might include the routing to any optional path deviations such as the interrupt service routines (ISR) as described in Chapter 2. For example, in the laparoscopic cholecystectomy procedure used in our OR study, the MCMD is composed of four standard stages: trocar preparation, isolate gallbladder, remove gallbladder, and closure. For each procedure we first identified the set of tasks executed in order to perform the gallbladder separation. As each task constitutes a larger goal in the procedure, which is further decomposed into a set of subtasks, we then proceed to identify in the video recording not only the type of subtask but also the order of execution. The order of subtask execution is stored as a vector, S, of the subtask node numbers visited during a particular procedure. Thus, element $M_{ij}.T_k.S_l$ represents the subtask l of the task k executed by subject i during procedure j.

Time

Once executed tasks and subtasks are identified using the MCMD, the time spent in each state during each entry can be identified in a video analysis by manually determining the start and end points according to previously defined criteria. For example, in the OR study, we used the surgical definitions from Chapter 2 for each subtask to set its corresponding initial and final limits: ‘Expose Triangle’ begins when the gallbladder is first stretched out and ends when the cystic pedicle is identified, and ‘Dissect CD/CA’ begins when the tip of the tool is first inserted between the two anatomic structures and ends when both structures have been completely freed from each other. These times are stored as a vector t associated with each node in the MCMD; we use $M_{ij}.T_k.S_l.t$ to represent the vector of times associated with subtask l and task k (**Figure 3.8**).

3.3.1.1.2 Competence and Coordination

Kinematics

For describing competence in use of tools and bimanual coordination, kinematics of the tool movement become important measures as velocity, acceleration and jerk profiles are unique to every execution and every subject. In our approach, we use two types of kinematics measures: summary measures to represent a performance in terms of a single value (e.g., mean velocity during a particular subtask), and detailed measures to describe trends during a subtask execution in the form of Cumulative Distribution Functions (CDFs) (**Figure 3.10**).

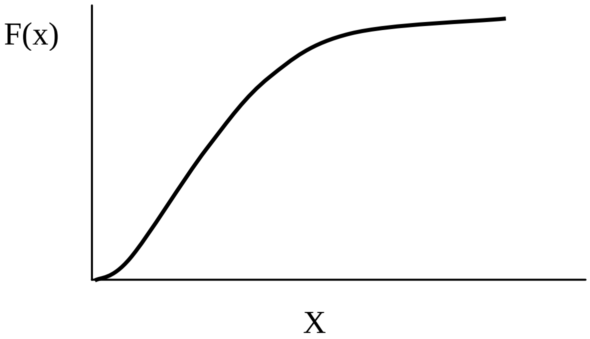


Figure 3.10: Representation of a Cumulative Distribution Function. A CDF describes the probability distribution of a real-valued random variable X and is given by F(x) = P(X < x)

Average and CDF kinematics can be separately computed for each subtask executed either by the dominant or the non-dominant hand. Therefore, following our dot notation, we use for example, $M_{ij}.T_k.S_l.\bar{v}_{ad}$ to represent the average velocity in the axial direction for the dominant hand associated with subtask *l* of task *k* in the MCMD for subject i and procedure j. The same notation is used for the other performance variables described in **Figure 3.8**.

For representing bimanual coordination, we use two types of measures to compare tool kinematics distributions from dominant and non-dominant hands: Mutual Information as a measure of independence and the Kolgomorov-Smirnov (KS) statistic as a measure of dissimilarity.

Mutual Information (I) measures the dependence between two distributions. In our context, it would provide an indication of the information that dominant hand (X) and non-dominant hand (Y) distributions share while executing a subtask (i.e., how much knowing one of the variables reduces the uncertainty about the other) and is expressed in terms of the entropies of the marginal distributions H(x) and H(y) and the joint distribution H(x,y).

$$I(x,y) = H(x) + H(y) - H(x,y)$$

$$H(x) = -\sum_{x} p(x) Ln(p(x)) \qquad H(y) = -\sum_{y} p(y) Ln(p(y)) \qquad H(x,y) = -\sum_{x,y} p(x,y) Ln(p(x,y))$$

In order to also provide an intuitive score between 0 and 1 to represent either existence or absence of independence, we use a normalized variant of the mutual information known as the *symmetric uncertainty measure* (U) [Witten 2005].

$$U(x,y) = 2 \times \frac{I(x,y)}{H(x) + H(y)}$$

We employ the Kolgomorov-Smirnov (KS) statistic to represent the degree of asymmetry in use of the dominant and non-dominant hand (Section 3.4.1). KS quantifies the discrepancy between two distributions in terms of the maximum vertical distance D between CDFs (**Figure 3.11**).

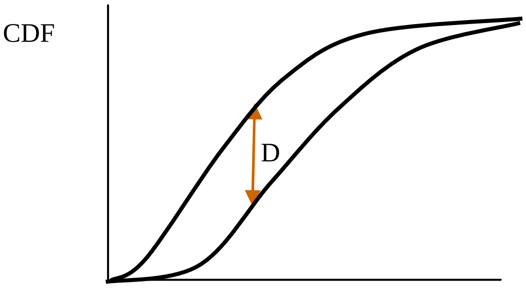


Figure 3.11: The K-S statistic. D represents the maximum vertical distance between two distributions

This is a normalized measure, which varies between 0 (meaning that distributions are similar) and 1 (meaning the distributions are different). In this way, we can compare the velocity profiles of the two hands of a subject (represented as $M_{ij}.T_k.S_l.\bar{v}$ vectors) when executing a particular subtask and determine if there are significant differences in speed. We will see later that the KS statistic is also very helpful in expressing differences in CDFs between subjects.

3.3.1.2 Action Level

At the action level, there is no sense of progression as transitions between states occur because these states represent brief surgical actions that are repeated and cycled between multiple times to achieve a surgical goal. Such transitions can therefore occur bidirectionally and any given state can be visited many times during execution of a subtask. We therefore do not represent transitions as a route through the MCMD, but as a Markov-type model, as described below.

3.3.1.2.1 Flow

Markov Modelling

At the action level, we monitor flow by representing transitions from one state to another. In our approach, we decompose each subtask into a set of 10 movements (‘pull’, ‘push’, ‘reach’, ‘orient’, ‘sweep’, ‘spread’, ‘grasp & hold’, ‘grasp & cut’, ‘idle’, ‘out’) as described in Chapter 2. Transitions between these 10 states can be represented using a Markov modelling scheme (**Figure 3.12**).

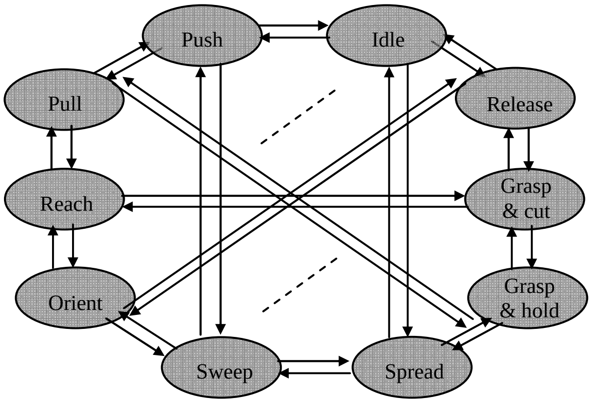
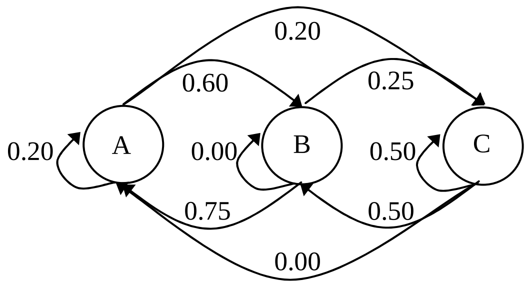


Figure 3.12: State-transition diagram for representing flow at the action level. Each state is characterized by probability value, which represents the possibility of transitioning from one state to another

A *Markov process* is a process in which the probability of moving from state 'a' to state 'b' depends (only) on the previous *n* states visited. The process is called an *order n* model where *n* is the number of states affecting the choice of next state. The simplest Markov process is a first order process, where the choice of state is made purely on the basis of the previous state. In our situation, the future evolution of the process (i.e., the movement that will be chosen by the surgeon) is determined by the description of the present state or movement. For a first order process with 's' states, there are s^2 transitions between states if one assumes that it is possible for any one state to follow another. Transitions between actions are represented by computing the probability P_{ab} of transitioning from state 'a' to state 'b' as the ratio between the number of a→b transitions and the total number of transitions from state 'a' to any other state. The s^2 probabilities may be collected together into a state transition matrix, which represents the overall behaviour of the system (**Figure 3.13**) [Howard 1971].



		End states		
		A	B	C
Start states	A	0.20	0.60	0.20
	B	0.75	0.00	0.25
	C	0.00	0.50	0.50

Figure 3.13: Example of a state transition diagram and probability transition matrix

All subtasks share the same diagram structure but differ in the transition probability matrices computed for each subtask and subject ($M_{ij}.T_k.S_l$.TPM = transition matrix for subtask l).[4]

Time

For every visit to each Markov state, a unique visit duration time is determined using video segmentation as described earlier to identify start and end points of each action and this duration is associated with the state. After visiting a state ‘a’ on multiple occasions, a vector of time measurements $M_{ij}.T_k.S_l.A_m$.t, which we call the holding time vector, is computed for action m, subtask l, task k. We then fit analytical distributions to these holding time vectors to derive parameters describing the holding time distributions (see Section 3.4.2 for more details about this computation).

[4] We actually use a semi-Markov model because the simplest Markov models make the assumption that dwell (or holding) times in a given state follow exponential distributions; this assumption does not match our data well, so we use modified models which allow us to introduce more accurate dwell-time distributions (this is described in more detail in section 3.4.2). At this stage, it is sufficient to note that we use transition probabilities between states to model flow at the action level.

3.3.1.2.2 Competence and Coordination

Kinematics

Based on the previous time records, kinematics data (particularly velocity) in the three cardinal directions is then segmented for each action visit. Execution speed is then computed and every visit is represented as a set of speed data points v = $|\underline{v}|$, which are stored as a vector $M_{ij}.T_k.S_l.A_m v$ associated with action m in the state-transition diagram (**Figure 3.8**). Hence, besides holding times CDFs, we also use speed CDFs to describe every action in the state transition diagram of **Figure 3.12**.

3.3.2 Evaluating Repeatability

As we have pointed out above, one of the most significant issues in making intraoperative measures is the existence of external factors that influence motor performance and that are outside of the surgeons' control; therefore, it is important to assess the repeatability of the measurements at each level of the MCMD for each subject.

We have devised three ways for computing repeatability, depending on the type of measure being assessed:

(a) When we have measures represented as distributions such as speed profiles, the Kolgomorov-Smirnov statistic (KS) is useful to compute differences as it describes the similarity between two empirical cumulative distribution functions by measuring (on a scale from 0 to 1) the maximum vertical distance between the two profiles (Section 3.4.1). Low KS values indicate similarity.

(b) When we have measures represented as probabilities, we use the symmetric normalized Jensen-Shannon Divergence as a similarity measure for comparing transition probability matrices, which individual rows are basically probability distributions of the states transitioning behaviour in a state transition diagram

(c) When we have multiple and different types of measures (e.g., time, velocity, acceleration, jerk) to consider simultaneously – perhaps across multiples scenarios (e.g., ‘peel’ and ‘detach’, or ‘isolate’ and ‘dissect’ subtasks), a multidimensional space is required for the corresponding representation. We use the Principal Component Analysis (PCA) to identify the directions of greatest variability in this multidimensional space and to reduce the dimensionality of the data by representing it in a low dimensional *weight space*. Repeatability is then reported as the standard deviation of distances with respect to a mean position in the weight space. A thorough explanation of this technique is provided in Section 3.4.4 and specific details of its implementation, for the two different experimental scenarios, is described in Chapters 4 and 5.

3.3.3 Evaluating Differences

In the present section we conceptually describe the computational framework we have developed for assessing differences across subjects at the subtask and action levels and how we propagate the extracted measures upwards in the hierarchical representation.

Since we will have available numerous different types of performance measures, it will be infeasible and overwhelming to present the reviewer with many direct comparisons in the units associated with each independent performance measure (e.g., velocity, force, time), especially since we do not know *a priori* which measurements will be most useful in drawing distinctions between surgeons (or indeed whether a single measure will provide as much discrimination as a combination of measures). We therefore need a mechanism for consolidating the various difference measures in a dimensionless form and which enables us to learn which combination of parameters provides the greatest discriminatory value. A multiple-parameter comparison (as opposed to a parameter-by-parameter comparison) would therefore help in compacting the relevant information and providing instructional insights when looking at the overall performance of a subject.

For training and analytical purposes, the user will be looking at summary measures at higher levels and will then drill down to look at more details. There are two distinct issues here: (1) different performance measures have different units, which means that we need a method for converting differences into normalized values, and (2) there are potentially many different performance measures, which means that we need a technique for automatically or semi-automatically identifying those measures or combination of measures which contribute most to differentiating between subjects and groups. The first issue motivates our use of Kolmogorov-Smirnov statistics and the Jensen Shannon divergence measure, both of which can naturally convert differences in dimensional variables to dimensionless quantities without needing to specify arbitrary reference values for normalization, while the second motivates our use of Principal Component Analysis (PCA) to identify linear combinations of variables which contribute the most to data variance and thereby enable us to effectively reduce the overall dimensionality of our data set. The details of how to compute the various measures will be explained in Section 3.4.

3.3.3.1 Conceptual Computational Framework

The primary goal of our analysis is to be able to compare the performances of individuals to their peer group, to experts, and to other groups along the skill development continuum. As described in **Figure 3.8**, we might have three types of variables available: point estimates, cumulative distribution functions (CDFs), and transition probability matrices. Fundamentally, for any comparison we might make, in our studies we have devised two ways of describing differences:

(1) Evaluating Distances Between Points in Weight Space Based on Descriptive Variables: When simultaneous consideration of multiple measures is required, we make comparisons in a multidimensional ‘performance space’ after processing the space using Principal Components Analysis, which allows us to assess differences across a whole range of parameters (e.g., point measures of time, velocity, force, pulse rate, etc) simultaneously by transforming the data into a normalized ‘weight’ space (as described in section 3.4.4). Differences are calculated by measuring the distance from a subject’s position to a group’s center (every subject is represented as a data point in the PCA plot) and applying ANOVA tests to evaluate the null hypothesis that the subject’s underlying data distribution does not differ from the group. In terms of the other types of variables we deal with, CDFs and transition matrices become difficult to include in a PCA analysis because each is composed of so many data points. The transition matrices (size 10x10) are the most problematic variables because they consist of 100 values and cannot be easily summarized. However a plausible solution for CDFs would be to discretize them as percentiles (i.e., 10%, 15%,

20%, etc), which would produce a manageable number of variables to include in a PCA analysis.

(2) Computing Differences Directly Followed By Dimension Reduction Using PCA: Alternatively we can compute comparisons directly (i.e., not in weight space). In this case, CDFs and transition matrices become easier to deal with, as there are statistical measures such as the Kolgomorov-Smirnov statistic and the Jensen-Shannon Divergence that allow us to compare distributions and transition probabilities (Section 3.4.1, 3.4.3). T-statistics can also be used to directly compare point estimates. The result of such comparisons is a vector of difference measures (D-measures) for a given action or subtask between a given subject and either another individual or an averaged set of data from a defined group of subjects.

In contrast to (1) where we are representing individual subjects as individual data points in the PCA plot before computing differences, the method of directly computing differences considers pairs of subjects, as the result is a difference measure between two individuals or between an individual and a group. The advantage of this approach is greater ease in incorporating CDFs and transition probability matrix difference measures, but the computational complexity of adding new data increases in proportion to the amount of existing data since new comparisons can potentially be made to all existing subjects in the data base. In contrast, the computational effort in using method 1 is independent of the amount of existing data.

3.3.3.1.1 Task/Subtask

At the task/subtask level, we represent performances in terms of summary measures such as average time and average velocity (all measures available at the subtask level are presented in **Figure 3.8**), which are then grouped and simultaneously analyzed using PCA (Section 3.4.4).

In order to plot data from multiple subjects and procedures, key descriptive measures for a given task or subtask are entered into a large matrix in which each row represents data from a single subject and procedure, as show in **Table 3.2**. For each subject 'i' and each procedure 'j' (M_{ij}), we defined multi-element vectors consisting of various measures $M_{ij}.T_k.S_l.r_r$ (r: time, velocity, acceleration, jerk, etc) for subtask l. All subjects can then be grouped into a gxh matrix (g: $\sum i \sum j$, total number of experiment repetitions (rows); h: number of dimensions (columns)).

Subject / procedure	**Subtask l**			**Subtask l+1**		
S11	$\mathbf{M_{11}.T_k.S_l.r_1}$	**…**	$\mathbf{M_{11}.T_k.S_l.r_r}$	$\mathbf{M_{11}.T_k.S_{l+1}.r_1}$	**…**	$\mathbf{M_{11}.T_k.S_{l+1}.r_r}$
S21	$\mathbf{M_{21}.T_k.S_l.r_1}$	**…**	$\mathbf{M_{21}.T_k.S_l.r_r}$	$\mathbf{M_{21}.T_k.S_{l+1}.r_1}$	**…**	$\mathbf{M_{21}.T_k.S_{l+1}.r_r}$
…	**…**	**…**	**…**	**…**	**…**	**…**
Sij	$\mathbf{M_{ij}.T_k.S_l.r_1}$	**…**	$\mathbf{M_{ij}.T_k.S_l.r_r}$	$\mathbf{M_{ij}.T_k.S_{l+1}.r_1}$	**…**	$\mathbf{M_{ij}.T_k.S_{l+1}.r_r}$

Table 3.2: Data arranged for PCA analysis at the subtask level. Each entry represents a summary measure 'r'such as average time, average velocity, average force, etc for every subject and procedure

Note that all measures are summary measures. CDFs can be represented as values at pre-selected percentile values (e.g., the 25th, 50th, 75th). The number of columns will vary depending on how many measures are used. More than one task or subtask can be considered simultaneously by adjoining the corresponding matrices. Once data for multiple subjects is entered, the dimensionality of the data set can normally be reduced by applying a Principal Component Analysis and the results for each row can be plotted in a lower dimensional ‘weight space’ – in practice, we have often found that 2D is acceptable (**Figure 3.14**). Different executions by particular subjects or groups can then be easily visualized and measures of proximity of a subject to another subject or to different groups can be assessed.

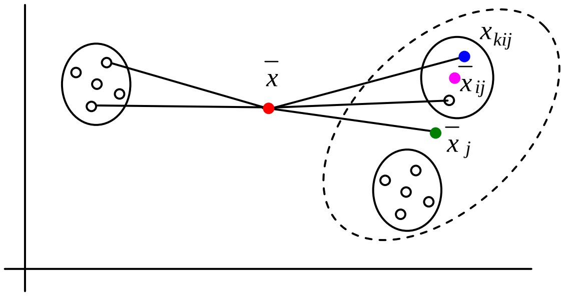
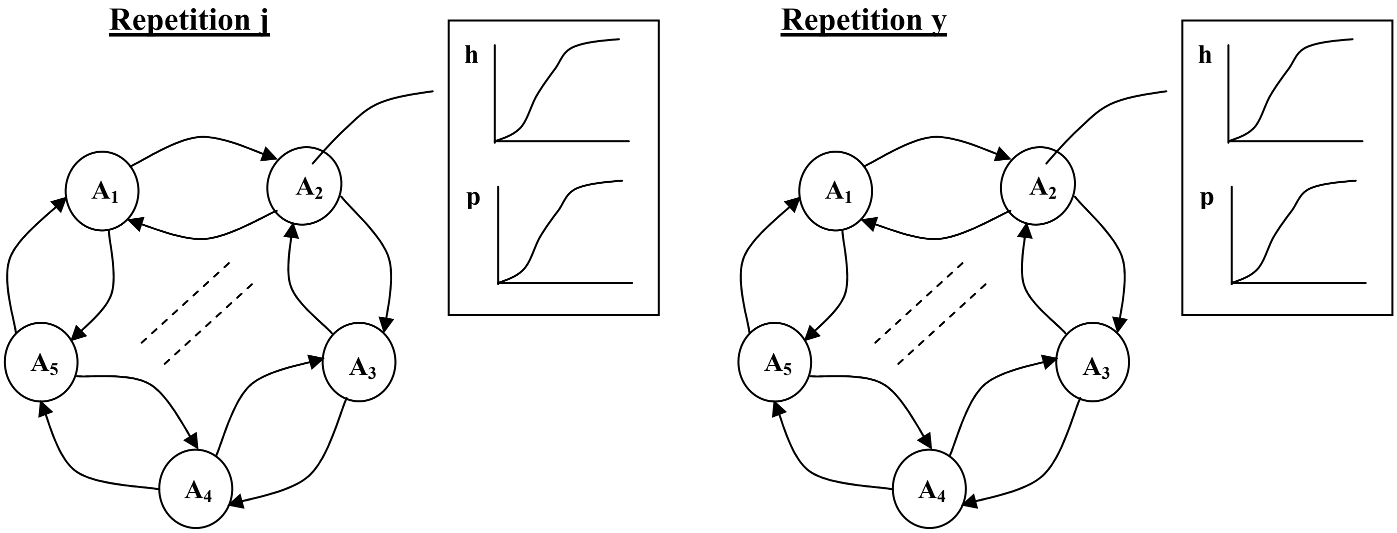


Figure 3.14: Example of intrasubject, intra- and intergroup trial positioning in the PCA weight space (k : procedure #; i : subject #; j : group #). Horizontal and vertical axes represent the 1st and 2nd principal components or directions of the maximum variability in the data (Section 3.4.4)

Special considerations: Before describing how we compute differences at the action level, it is necessary to outline the structural representation of a subtask. For explanatory purposes herein, we concentrate on representing how an individual i (e.g., a resident) is evaluated against another subject q (e.g., an expert).

For any subject i the following data is available for any two procedures j and y **(Figure 3.15**):



$$N_{ij} = M_{ij}.T_k.S_l.N = \left[M_{ij}.T_k.S_l.A_1.n,\ \ M_{ij}.T_k.S_l.A_2.n,\ \ M_{ij}.T_k.S_l.A_3.n,\ \ M_{ij}.T_k.S_l.A_4.n,\ \ M_{ij}.T_k.S_l.A_5.n\right]$$

$$N_{iy} = M_{iy}.T_k.S_l.N = \left[M_{ij}.T_k.S_l.A_1.n,\ \ M_{iy}.T_k.S_l.A_2.n,\ \ M_{iy}.T_k.S_l.A_3.n,\ \ M_{iy}.T_k.S_l.A_4.n,\ \ M_{iy}.T_k.S_l.A_5.n\right]$$

$$W_{ij} = M_{ij}.T_k.S_l.W = \left[M_{ij}.T_k.S_l.A_1.w,\ \ M_{ij}.T_k.S_l.A_2.w,\ \ M_{ij}.T_k.S_l.A_3.w,\ \ M_{ij}.T_k.S_l.A_4.w,\ \ M_{ij}.T_k.S_l.A_5.w\right]$$

$$W_{iy} = M_{iy}.T_k.S_l.W = \left[M_{iy}.T_k.S_l.A_1.w,\ \ M_{iy}.T_k.S_l.A_2.w,\ \ M_{ij}.T_k.S_l.A_3.w,\ \ M_{iy}.T_k.S_l.A_4.w,\ \ M_{iy}.T_k.S_l.A_5.w\right]$$

Figure 3.15: Diagrammatic and conceptual representation of the quantitative data available at the action level for one subject when executing a specific subtasks multiple times. The circled diagram is a state transition diagram representing the movements or actions available at each subtask.

Every diagram corresponds to the same executed subtasks for two different repetitions j and y. Each state in the diagrams is a symbolic representation of an individual movement (e.g., push, pull, orient, etc) to which holding time (ht) and kinematic measures (K: velocity, acceleration, jerk) profiles can be attached. The N and W vectors contain information about number of visits to (N) or amount of time in (W) each state. Additionally each model produces a transition probability matrix TPM_{ij} or TPM_{iy}. The diagrams and the variables attached constitute Semi-Markov model representations for each subject/subtask (Section 3.4.2).

In order to represent a 'mean' performance for subject M$_i$, we merge across all executed repetitions of the same subtask:

$$\begin{aligned}
&\mathrm{TPM}_i \equiv \mathrm{TPM}_{ij} + \mathrm{TPM}_{iy} + \ldots \\
&\mathrm{N}_i \equiv \mathrm{N}_{ij} + \mathrm{N}_{iy} + \ldots \\
&\mathrm{W}_i \equiv \mathrm{W}_{ij} + \mathrm{W}_{iy} + \ldots \\
&\mathrm{ht}, \mathrm{K} \equiv \mathrm{concatenate} \cdot \mathrm{vectors}
\end{aligned}$$

3.3.3.1.2 Action

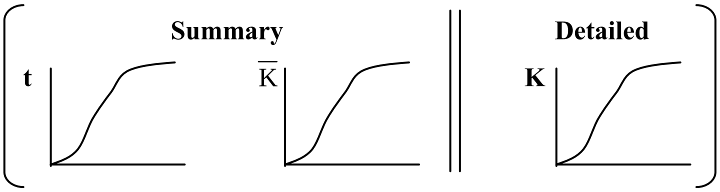
At the action level, we use the following direct comparison methods:

- Comparing TPM matrices: Since the TPM matrices are estimates of the probability distributions of the states' transitioning behaviour (in our application, the size of the TMPs is 10x10 because there are 10 actions), we use the Jensen-Shannon divergence JSD (Section 3.4.3) to compare TPM matrices derived from our pre-defined Semi-Markov models. However since all states were not visited equally often, the standard JSD metric might underestimate the importance of frequently visited states (i.e., all transition probability values in a TPM are estimated on an equal footing, no matter how many times each state was visited); therefore, we modified the metric by applying it to versions of the matrices normalized ($\mathrm{T\tilde{P}M}_i$) by the means of the N vectors in order to increase the weighting on frequently visited states.

$$\mathrm{T\tilde{P}M}_i = \mathrm{diag}(\mathrm{N}_i) \times \mathrm{TPM}_i(\mathrm{row} \cdot \mathrm{normalized})$$

As a result, our difference measure for TPM matrices from subjects ‘i’ and ‘q’ is computed as $\Delta\left(\tilde{TPM}_i \middle\| \tilde{TPM}_q\right)$ using the Jensen-Shanon Divergence measure explained in Section 3.4.3.

- Comparing vectors of holding time (ht) and kinematics (K) distributions: For each state (i.e., movement), we have attached holding times ($M_{ij}.T_k.S_l.A_m t$) and kinematics ($M_{ij}.T_k.S_l.A_m v$) profiles in the form of CDFs. Holding times measure the duration of each visit and kinematics describe the tool motion during the visit. We also define both summary values and detailed measures for the kinematics data (K: velocity, acceleration, jerk). A summary measure is computed once per state visit and typically represents an average (eg, average velocity) during the visit; the collection of all summary measures across all visits to a state can be described using a cumulative distribution function. A detailed measure will typically be the collection of all data points measured during a single visit and cumulative distribution functions are used to concatenate all the individual measurements acquired during all the visits.



We use D values from the Kolmogorov-Smirnov statistic (Section 3.4.1) to measure differences between cumulative distribution functions. To compare two subjects, ‘i’ and ‘q’, we form a D_{iq} matrix (composed of submatrices D_s and D_d) from all D metrics (D_{iq}^{t}, $D_{iq}^{\overline{K}}$, D_{iq}^{K}) across states and performance measures **(Figure 3.9)**.

$$
\begin{aligned}
D_{iq} &= subject.i \quad vs \quad subject.q \\
D_{iq}^{t} &= D_{iq}.T_k.S_l.A_m.t \\
D_{iq}^{\overline{K}} &= D_{iq}.T_k.S_l.A_m.\overline{p} \\
D_{iq}^{K} &= D_{iq}.T_k.S_l.A_m.p
\end{aligned}
$$

$$
\mathrm{D}_{\mathrm{iq}} = \underset{\xrightarrow[\text{Performance measures}]{}}{\begin{bmatrix} \mathrm{D}_{\mathrm{iq}}^{\mathrm{t}} & \mathrm{D}_{\mathrm{iq}}^{\overline{\mathrm{K}}} \,\|\, \mathrm{D}_{\mathrm{iq}}^{\mathrm{K}} \\ \vdots & \vdots \,\|\, \vdots \\ \vdots & \vdots \,\|\, \vdots \end{bmatrix}} \Big\downarrow \text{States} \quad \mathrm{D}_{\mathrm{iq}} = \left[\mathrm{D}_{\mathrm{summary}} \,\|\, \mathrm{D}_{\mathrm{det\,ailed}}\right] = \left[\mathrm{D}_{\mathrm{s}} \,\|\, \mathrm{D}_{\mathrm{d}}\right]
$$

Special considerations: In computing these D measures at the action level, we realize that different subjects may use the ten different fundamental surgical actions in vastly different proportions. In particular, surgeons in training may not avail themselves of as broad a range of surgical actions as experts; indeed, we observed that on some occasions individual surgeons rarely if ever used certain actions. We therefore have to take account when computing some D values of the fact that some comparisons are made on the basis of large numbers of samples from both subjects and may therefore be considered reliable difference measures, while other comparisons may be made on the basis of few samples from one or both subjects and so are far less reliable and should be given less weight in the analysis **(Figure 3.16).**

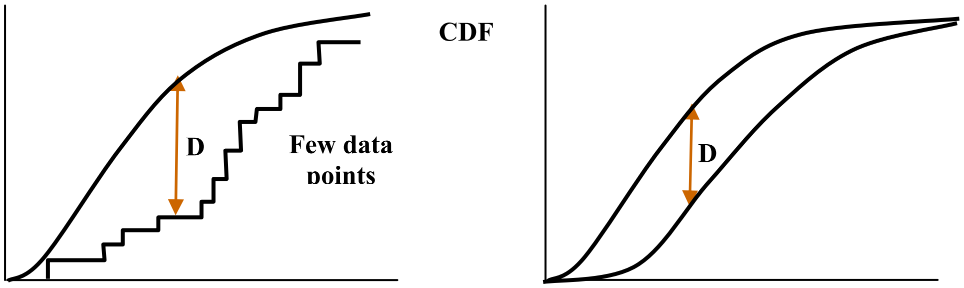


Figure 3.16: Schematic representation of difficulty in computing reliable D measures when number of points in a particular distribution varies widely

Because the number of points in each distribution might vary widely due to limitations in the amount of available data, we weighted D_s and D_d in proportion to the square root of the largest number of samples, n_s in either of the two distributions used to compute each D value. This method emphasizes states which at least one subject visits frequently. However, expert surgeons sometimes believe that a less frequently used movement (action) is actually more critical for performing a good surgery, in which case the weighting could also be adjusted to incorporate this expert knowledge. We therefore allow for incorporation of 'a priori belief' (b=1: average importance; b=0: no importance; b>1: greater importance) based on surgeons' opinions and used this weighting in addition to n_s to adjust the weighting of the D matrix:

$$\tilde{D}_a, \tilde{D}_d = \frac{\mathrm{diag}\left(\sqrt{\underline{n}\underline{b}}\right)}{\max(n_s)\max(b_s)} \times D_a, D_d$$

$$\tilde{D}_{iq} = \left[\tilde{D}_a \Vert \tilde{D}_d\right]$$

3.3.3.2 Bootstrapping Approach for Computing Confidence Level

In our applications, difference scores, such as D values derived from the K-S statistic, are computed based on two empirical CDFs (C_1, C_2), each of which has a finite number of samples in them, so the value (D_{12}) we find is itself an estimate of the true difference between the two sets. In order to test hypotheses and to recognize difference values, which indicate statistically significant differences, we need to assign confidence intervals to our estimates. Since the statistics we use to evaluate differences are rarely well-approximated as normally distributed, we use a bootstrapping approach in which we synthetically resample the data sets a large number of times (according to the techniques described by [Efron, 1986]) and re-compute the resulting *D* values. Hence we are able to establish a distribution of *D* values as shown in **Figure 3.17** and the 5% and 95% confidence bounds can be easily identified on the cumulative distribution curve.

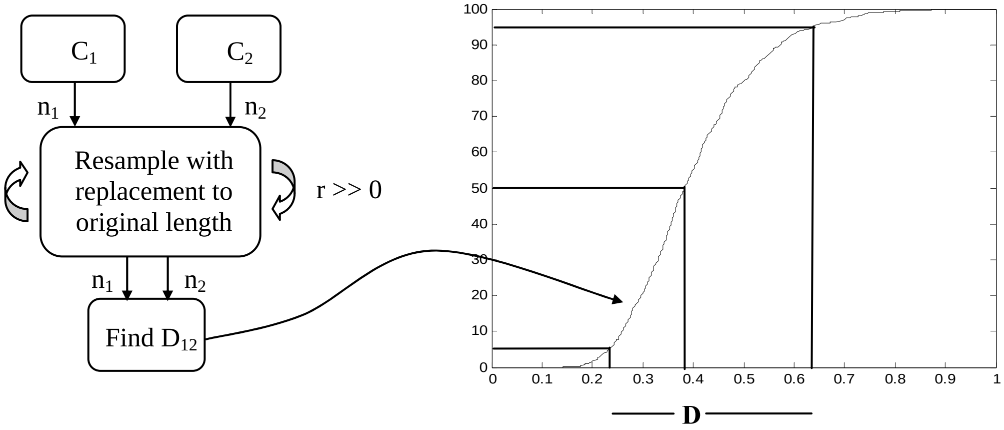


Figure 3.17: Resampling cumulative probability distributions of two finite data sets to establish confidence bounds on an estimate for D. The right illustration depicts the 5% and 95% confidence bounds on the estimate for D

3.3.3.3 Propagating Difference Measures Upward

Once we have computed difference measures at the most detailed level of the MCMD (ie, the action level), we will have a large number of difference measures to deal with since numerous measures $D_{iq}.T_k.S_l.A_m.r$ (r: time, velocity, acceleration, etc.) are computed for each of the ten action states. The surgical trainer needs to be able to identify which of these measures indicate important differences between the subject under consideration and the reference group, and this data needs to be summarized in a useful way when it is consolidated into summary measures associated with the corresponding subtask at the next level in the MCMD. In this section we discuss how we propagate this collection of D-measures upwards **(Figure 3.7**).

Subtask Level

As described in section 3.3.3.1.2, at the action level, we compute difference (D) measures derived from comparing movement transitions, time, and kinematics profiles for each of the 10 predefined actions. D-values might be computed for subject-to-subject, subject-to-group, or group-to-group comparisons. We follow the same strategy for propagating them upwards; however, for explanatory purposes we use subject-to-subject comparison herein.

At the action level, every comparison is represented as a collection of D measures for each action arranged as a single row. In order to propagate them to the subtask level, we simply concatenate all rows in order to form a *k* x *w* matrix, where k corresponds to the number of subject-to-subject comparisons $M_i$vs.M_q (where M_i and M_q can be from the same group –

eg, E: Experts, E_I vs. E_J; R: Residents, R_I vs. R_J – or from different groups – R_I vs. E_J) and w corresponds to the number of extracted performance measures.

$$\begin{bmatrix} R_1 \text{vs.} E_1 \\ R_1 \text{vs.} E_2 \\ \vdots \\ R_3 \text{vs.} E_3 \end{bmatrix} = \left[\begin{array}{c||ccc||ccc} \text{JSD}_{R_1E_1} & D^t_{R_1E_1}(\text{push}) & \cdots & D^t_{R_1E_1}(\text{out}) & D^v_{R_1E_1}(\text{push}) & \cdots & D^v_{R_1E_1}(\text{out}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & & & \\ \text{JSD}_{R_3E_3} & D^t_{R_3E_3}(\text{push}) & & D^t_{R_3E_3}(\text{out}) & D^v_{R_3E_3}(\text{push}) & & D^v_{R_3E_3}(\text{out}) \end{array}\right]$$

For each subject-to-subject comparison (ie, each row), we use 1 JSD value; 10 D values (per variable) for average time, and average velocity, acceleration, and jerk profiles; 10 D values (per variable) for detailed velocity, acceleration, and jerk profiles (as described in Section 3.3.3.1.2). Afterwards, the matrix is processed using Principal Component Analysis (PCA) in order to reduce dimensionality so that each row can be plotted in a low dimensional ‘weight space’ and measures of proximity between subjects’ pairs can be implemented.

Task Level

In propagating difference measures from the subtask to the task level, we follow the same approach as before but instead of concatenating D values for actions, we use D values for subtasks. The same PCA analysis is carried out at this level.

3.4 Mathematical Formulation of Difference Measures

In the previous section we presented the conceptual framework we used in computing difference values, but did not fully describe the various difference measures we used. In this section, we provide further details on the following mathematical methods used:

- The Kolmogorov-Smirnov (KS) statistic for representing and computing differences from time and kinematics cumulative probability distributions

- Semi-Markov models for representing movement transitions and the Jensen-Shannon Divergence (JSD) measure as a way of computing differences between transition matrices.

- Principal Component Analysis as a method for reducing the dimensionality of our difference measures and for assessing differences across multiple measures simultaneously.

3.4.1 Kolmogorov-Smirnov Statistic for Time and Kinematics Profiles Comparison

Whenever we have available vectors (X) of data such as a set of 100 velocity points or a set of 120 time measures representing 120 visits to a particular action, we may visualize them as histograms showing what proportion of data points fall into each of several specified categories (see Figure 3.19 left); these histograms may be replotted as cumulative distribution functions CDFs, which can be thought of as the integral of the distribution plotted in the histogram **(Figure 3.18**).

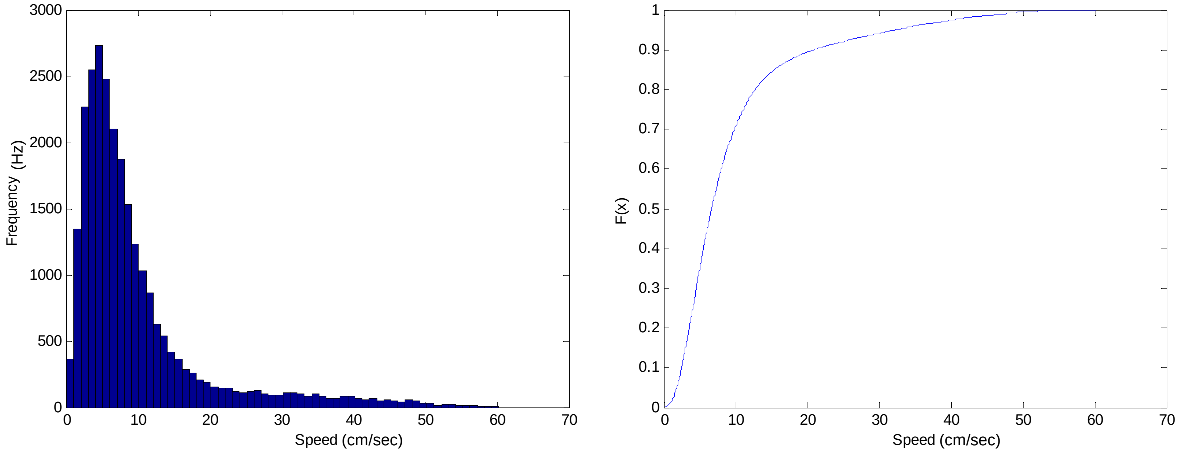


Figure 3.18: Example of a vector of speed values represented as a histogram (left) and as a cumulative distribution function (right). A CDF describes the probability distribution of a real-valued random variable X and is given by F(x) = P(X < x)

Since drawing a histogram of experimental data always requires putting the data into bins, and since the discretization resulting from selecting the bin size always results in the loss of information, we opt to use cumulative distribution functions instead. CDFs are also useful because the median is immediately apparent, while it is only approximately shown in a histogram. Another important reason for using CDFs rather than histograms is the possibility of using the Kolmogorov-Smirnov (KS) test for assessing differences between distributions. The KS statistic characterizes the difference between two distributions as the maximum vertical difference (D) between CDFs (**Figure 3.19**). It varies between 0 (similar) and 1 (different), which is helpful because this difference measure is intrinsically normalized; no special scaling or other treatment of the original data is required.

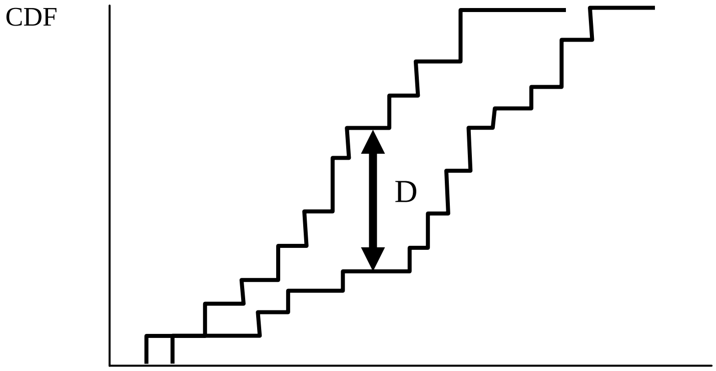


Figure 3.19: The KS statistic (D is the maximum vertical distance between two cumulative distribution functions). D varies between 0 (similar) and 1 (different)

The KS statistic is normally used to test whether CDFs drawn from different groups defined in a particular experiment are statistically distinct from one another. There are statistical tests available to calculate the p-value for comparing two different distributions [Hodgson 2002, Drew 2000]. While more commonly used tests such as Student's t-test and the F test detect differences only in the average or the variance of two groups, respectively, the KS test picks up any sort of difference in the CDFs of two groups, whether differences in the mean, in the variance or in the shape of the distributions (though in general it will not be as sensitive as an explicitly parametric test on distributions which match the assumptions behind the parametric analysis) [Von Mises, 1964].

In addition to performing hypothesis testing, we often want to use the KS statistic directly as a measure of difference (eg, to assess the extent of difference between how a particular subject performs a surgical task and how a reference group performs the same task). In these circumstances, we need to assign a confidence interval to the measure, as mentioned above in section 3.3.3.2; this is slightly different from the standard use of the Kolmogorov-Smirnov statistic, where one wishes to test the hypothesis that the two distributions are different. In this latter case, there is an established procedure for calculating the p value. However, to assign a confidence interval to the statistic directly, there is no simple approach that we know of which would work for arbitrary distributions. We therefore chose to use a approach in which we synthetically resample the data sets a large number of times [Efron, 1986] and re-compute the resulting *D* values in terms of a distribution where the 5%-95% confidence bounds can be easily identified on the resulting cumulative distribution curve.

In addition, to assess the significance of the particular D values we find, we also compute typical D values for intragroup comparisons and for intergroup comparisons between the most widely spread groups in our dataset. This allows us to place a particular difference measure on a scale representing the smallest and largest D values we could reasonably come across. The wider the range of possible D values and the tighter the confidence interval, the more reliable the estimate is. We are then interested in weighting the D values in order to reduce the influence of those unreliable D measures, particularly as the amount of data (number of columns) increases in the PCA analyses. We propose relating the weight 'w' to some function of D_{max} and D_{min}, such that it goes down as D_{min} rises or D_{max} decreases. A candidate function would be w(r), where $r = D_{max} / D_{min}$.

In summary, we use D values either for computing differences or to test null hypotheses.

3.4.2 Markov Models for Action Level Analysis

At the action level of the MCMD, which describes motion primitives (e.g., pushing, pulling, sweeping, etc), there is no notion of a surgical goal; therefore, the motion primitives are viewed as states that the surgeon cycles between until the higher-level goal is achieved **(Figure 3.20)**. These movements do not have an order of execution and happen almost at random, (i.e., the surgeon would have a difficult time explaining why they decided to move from a pushing action to a sweeping action to a grasping action; they would likely say that these are simply the actions that are required to accomplish the surgical goal, but they would not say to a resident that there is an explicit order of sweep, push, grasp, etc. required to accomplish the goal; rather, they would tend to instruct the resident by saying something like "dissect this structure until this other structure is visible"); therefore, we represented them on an interconnected diagram or state transition diagram and used transition probabilities between states to model flow at the action level.

Dr. Jacob Rosen at the University of Washington has been using Markov models for characterizing magnitudes of forces and torques (F/T) measured at the human/tool interface in order to distinguish between surgical skill levels when performing laparoscopic cholecystectomies on pigs. Initially, a set of 14 tool/tissue interactions were identified by visual inspection and the magnitudes of the force/torque profiles were used to classify the 14 interactions into 3 groups (low, medium, high) of magnitudes, which constituted a 3-state Markov model. Individual models were trained for each surgical step and each skill group and a statistical distance (based on the parameters μ and σ of each F/T distribution at each state) was then used to determine how far a subject is from the expert group [Rosen 2006, 2001]. Based on this work and other applications of Markov models to modeling human tasks, we considered whether some variant of Markov models was appropriate for this project.

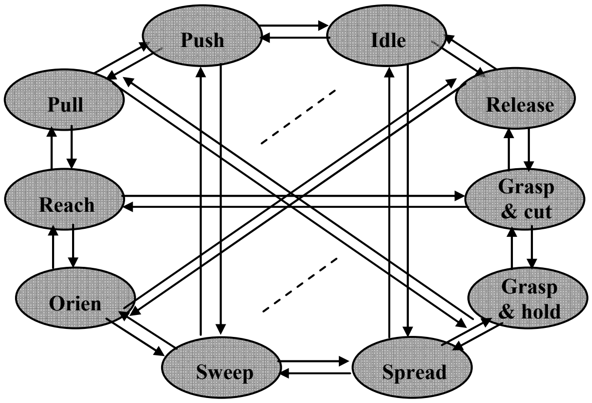


Figure 3.20: State-transition diagram for the action level defined in our application. 10 discrete tool movements describes this level and transitioning between states occurs probabilistically

Markov models (MMs) can be generally classified as continuous or discrete. Continuous CMMs model the tendency to leave a state as constant over time. This is known as the 'race condition' of the CMM and leads to dwell times in a state that are characterized by exponential distributions. In our model each state has an inherent cognitive load and requires completion of a finite task; these factors influence the execution of the movement such that the probability of leaving the state soon after entry is low and increases over time until it is greatest near the expected mean residency time, after which it again drops off. The exponential distribution is therefore not a good model of our situation [Meyn 1993, Doob 1953]. Discrete MMs exhibit similar behaviour; transitions occur at discrete intervals and the probability of leaving is constant during each interval; this produces a dwell time distribution that is best represented as a geometric distribution [Meyn 1993, Doob 1953], which again does not match the expected dwell time distributions in our experiments **(Figure 3.21)**.

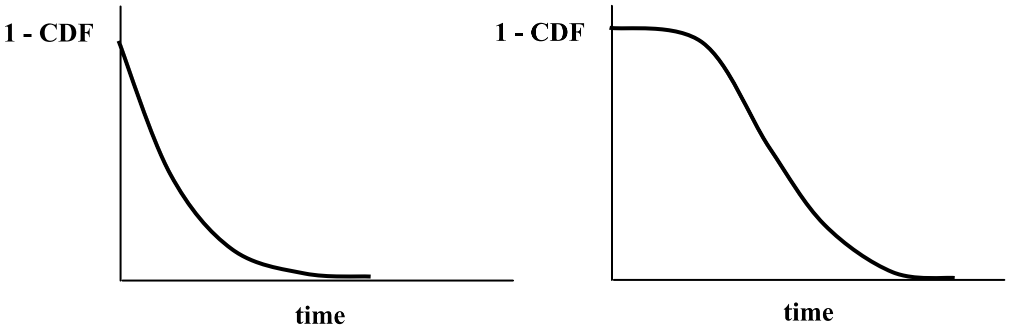


Figure 3.21: Complement of the Cumulative Distribution Function for dwell times (probability of remaining in the state after a time t): (left) the exponential and geometric behaviour of classical Markov Models, and (right) the behaviour of a human when executing a task with inherent cognitive load as that of a surgeon during surgical procedures

We therefore decided to use semi-Markov models to represent the non-exponentially-distributed holding times in each state we expect to see. A semi-Markov model is a process whose successive state occupancies are governed by the transition probabilities of a Markov process, but whose stay in any state is described by a random variable that depends on the state presently occupied and on the state to which the next transition will be made [Howard 1971].

The assumptions embodied in Semi-Markov Models are that transition probabilities only depend on the current state (so they have the same transition matrix representation as a classical Markov model) and that the holding time distribution could be a function of the subsequent state (ie, there could be a longer expected time if the next state indicated that the surgeon thought this was a difficult case) [Haverkot 2001]. We used this at the task and subtasks levels as well.

To test our assumption that holding times in the action states were not well modelled as exponential distributions, we performed a distribution fitting test using three types of distributions: Exponential, Weibull, and Lognormal. These last two were chosen as candidate distributions because they have been used in other human task applications to model task completion times, particularly for very quick tasks [Zhang 2007, Murthy 2003, Giuntini 2000]. We used holding time data from the most executed movements (about 4 actions) per subtask (Expose Triangle and Dissect CD/CA) while performing one selected procedure for each one of the 3 experts and 3 residents who performed the experiment described in Chapter 5. In total, we performed 12 distribution comparisons per subject (4 actions x 3 distributions).

Using the Distribution Fitting Tool from Matlab, which performs a non-linear least square fit, we obtained the parameters for each fitting based on the actual vectors of times for each action (**Figure 3.22**). These vectors are obtained from measuring time spent during each visit to the state; if a state is visited 100 times; we get a 100-element vector with 100 holding time values. Using the Kolgomorov-Smirnov statistic (Section 3.4.1, we found that in all cases either a lognormal or a weibull distribution provided a better model for the holding time distributions than an exponential distribution (Appendix D).

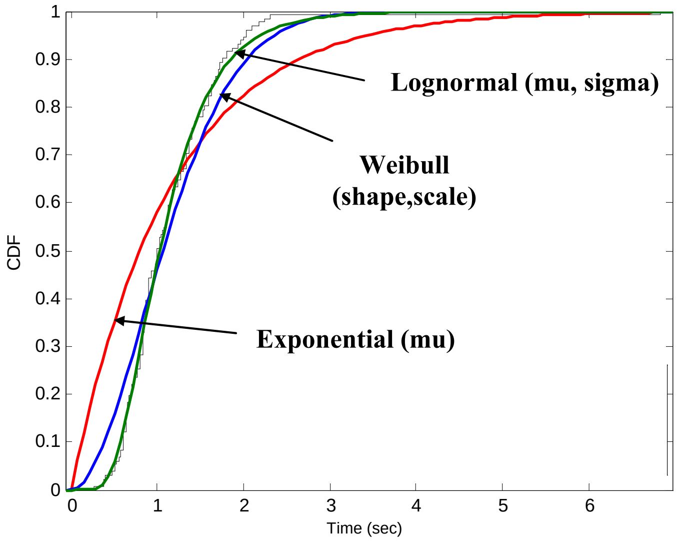


Figure 3.22: Example of the distribution fitting process with the three selected distributions (Exponential, Lognormal, and Weibull) performed on time measurements, which represented time (in seconds) spent at each action during each visit. The degree of fit was evaluated based on the D values between each distribution and the experimental data.

The primary contributions we seek to provide with regard to the use of Markov-like models for modeling surgical procedures are (1) to recognize that dwell times in surgical states are not well-modeled by the exponential dwell time assumptions of the standard Markov model and (2) to provide a different metric for comparing different Markov models which is more intuitive and does not require a hidden discrete Markov model.

3.4.3 Jensen-Shannon Divergence as a Similarity Measure for Markov Models

The pattern of tool movements executed by each subject is modelled by a semi Markov process. In our assessment methodology we need to compute difference measures for the different elements of a semi Markov model. Here we consider how to compute a difference measures for a pair of transition probability matrices.

Several measures have been proposed as metrics for Markov models [Bicego 2004, Qian 2003, Wu 2001, Lyngso 1999]. However, all of them deal with measuring similarity of sequences, and therefore are inherently related to Hidden Markov Models, where the states are not directly apparent and the transition probabilities not explicitly available. Some of the concepts are based on co-emission probabilities of various HMMs, or on using word frequencies (based on the principle that similar sequences (or time series) share similar words for the case of DNA analysis).

Since we have the transition probability matrix directly available to us, we should compute the difference measure directly on these matrices [Qian 2003].

In consultation with Professor Kevin Murphy from the Computer Science Department at UBC, we decided to use the Jensen-Shannon Divergence (JSD) as our similarity measure for the probability distributions derived from the transition matrices in our Semi-Markov model approach. The JSD is derived from the Kullback-Leibler Divergence (KL), which is a standard information-theoretic measure of the dissimilarity of two probability functions [Dagan 1999], and is defined as (Equation 3.4, ‘i’ refers to a discrete probability function):

$$KL(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \quad \text{Eq. 3.4}$$

$$JSD(p\|q) = 0.5KL\left(p\middle\|\frac{p+q}{2}\right) + 0.5\left(q\middle\|\frac{p+q}{2}\right) \qquad \text{Eq. 3.5}$$

It is important to note that KL is not a distance metric itself since it is not symmetric and moreover, does not satisfy the triangle inequality. On the other hand, the JSD (Equation 3.5) is a symmetrized version of the KL divergence. Additionally, it is always well defined and bounded, and its square root is a true metric for the probability distributions space (i.e., its square root is symmetric, null only when the probability distributions coincide and it verifies the triangle inequality) [Majtey 2005, Endres 2003, Lee 1999].

Since we would prefer to use difference which range from 0 to 1 (Section 3.3.2), we first find the maximum possible value for JSD and then normalized it as follows:

$$\mathrm{JSD}(p\|q) = \sum_{i=1}^{N}\left(p_i \log\left(\frac{2p_i}{p_i + q_i} \right) + q_i \log\left(\frac{2q_i}{p_i + q_i} \right)\right)$$

$$\mathrm{JSD}(p_i\|q_i) = \underbrace{(p_i + q_i)\log 2}_{\geq 0} + \underbrace{p_i \log\left(\frac{p_i}{p_i + q_i} \right)}_{\leq 0} + \underbrace{q_i \log\left(\frac{q_i}{p_i + q_i} \right)}_{\leq 0}$$

When p and q are two distinct deterministic distributions:

$$\sqrt{\mathrm{JSD}(p\|q)}_{\max} = \sqrt{2\log 2}\text{ ; so we define } \Delta_{pq} = \frac{\sqrt{\mathrm{JSD}(p\|q)}}{\sqrt{\mathrm{JSD}(p\|q)}_{\max}}$$

Therefore, the properties of our metric are:

- $\Delta_{pq} \geq 0$

- $\Delta_{pq} = \Delta_{qp}$

- $0 \leq \Delta_{pq} \leq 1$ with $\Delta_{pq} = 0$ if and only if $p = q$

- $\Delta_{pq} + \Delta_{qr} \geq \Delta_{pr}$

For the case when $p_i = 0$, computation of the logarithm is not possible, but since $\lim_{p_i \to 0} p_i \log(p_i) = 0$, it is still possible to compute the JSD. JSD divergence is applied as a similarity measure for comparing T matrices, which individual rows are basically probability distributions of the states transitioning behaviour (Section 3.4.2).

3.4.4 Principal Component Analysis for Separating Skill Levels

In our approach, we deal with multiple performance measures (e.g., time, velocity, acceleration, jerk) across multiple tasks (e.g., ‘peel’ and ‘detach’, or ‘isolate’ and ‘dissect’ tasks), which make our data set not only multidimensional in nature, but multidimensional with high dimensionality (up to dozens of more dimensions). Since we are aiming to convey results to surgeons, we need to be able to express the patterns in the data as intuitive representations of subjects’ behaviours. Thus, we use Principal Component Analysis to dramatically reduce the dimensionality of the data by projecting it into a ‘weight space’, as described below. We have found that we can often capture upwards of 90% of the variability in the data with as few as 2 or 3 dimensions.

3.4.4.1 Principal Component Analysis

The method is based on expressing the data in terms of a weighted sum of the eigenvectors of the covariance of the data. In this way, data points are remapped into a weight space which represents the contributions from each of the eigenvectors; therefore these vectors represent the directions of greatest variance in the original data (see **Figure 3.23).**

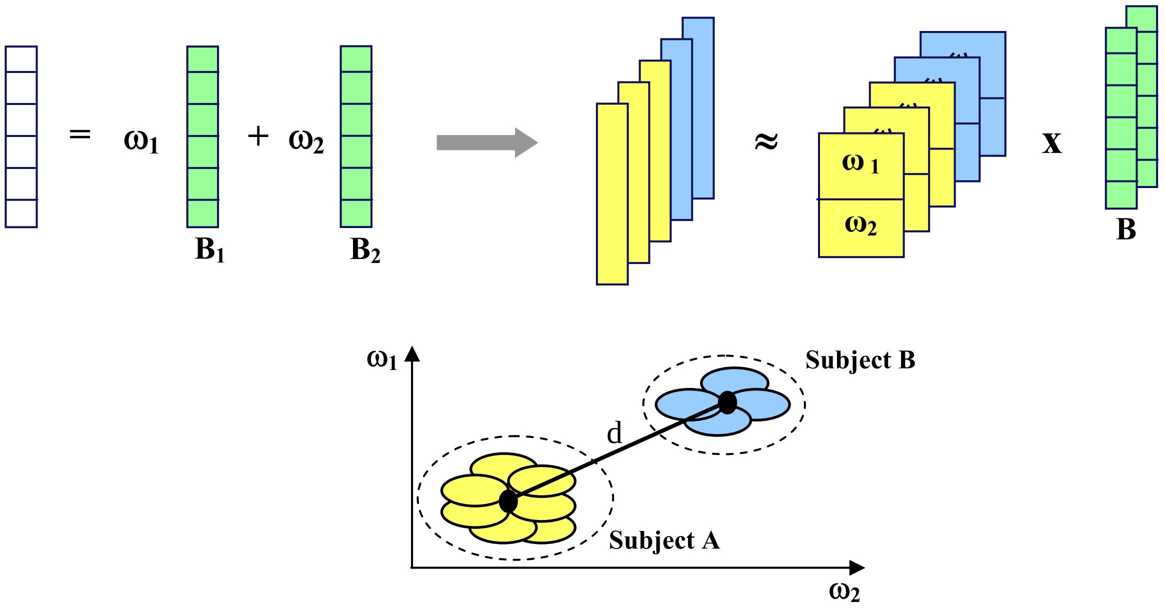


Figure 3.23: Graphical description of the transformation of the data into the PCA space. PCA is a weighting space in which a set of data might be represented in terms of a vector of weights (i.e., principal components, shown in yellow and blue), which multiply eigenvectors (shown in green) which correspond to the directions of the maximum variability in the original high dimensional data space [Jolliffe 2002].

Since the various elements of the data vectors may have different units and may be of markedly different magnitudes, it is standard practice to normalize each element by the standard deviation of the set of all corresponding elements across all samples and to subtract the global mean [Johnson 2002, Jackson 1991]. With the PCA, the eigenvectors are meaningful and provide information about the principal correlations between variables. By observing which elements of the principal eigenvector have the largest values, one gains insight into which variables or combination of variables make the dominant contribution to variability in the data set.

3.4.4.2 Determining Number of Dimensions to Include in Analysis

The PCA finds the directions of greatest variance in the data set. By selecting a small number of high variance dimensions, we can capture most of the variance in the data with only a small number of variables. Therefore, it is necessary to make a principled choice of how many dimensions to work with.

Some traditional methods involve plotting the eigenvalues against the number of principal components and finding the "elbow" on the graph (i.e., the point where the graph of eigenvalues levels out), or limiting the variance accounted for by a set of PCs [Jolliffe 2002]. However, since those methods often rely on a visual heuristic, we propose using a more objective condition based on the reasoning behind the Principle of Parsimony, which is defined as the conceptual tradeoff between the squared bias and variance vs. the number of model parameters. Model selection methods such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) implicitly employ some notion of this tradeoff [Burnham 2004, Burnham 2002, Breiman 1992]. We then define a trade-off between the error expressed in terms of the variability accounted for and the complexity as the number of dimensions increases. We use the unexplained variability represented by each component or the variability explained by all components beyond the one being considered (100% - VAF) as an error measure, and the number of PCs as a complexity measure. We then define a dimensionality-decision criterion, which is an analog of an information criterion, to be the sum of the error and complexity measures and plot this against the number of principal components used. We chose the projection on the PC axis of the minimum point from the trade-off variable to be the optimal number of principal components to retain **(Figure 3.24**).

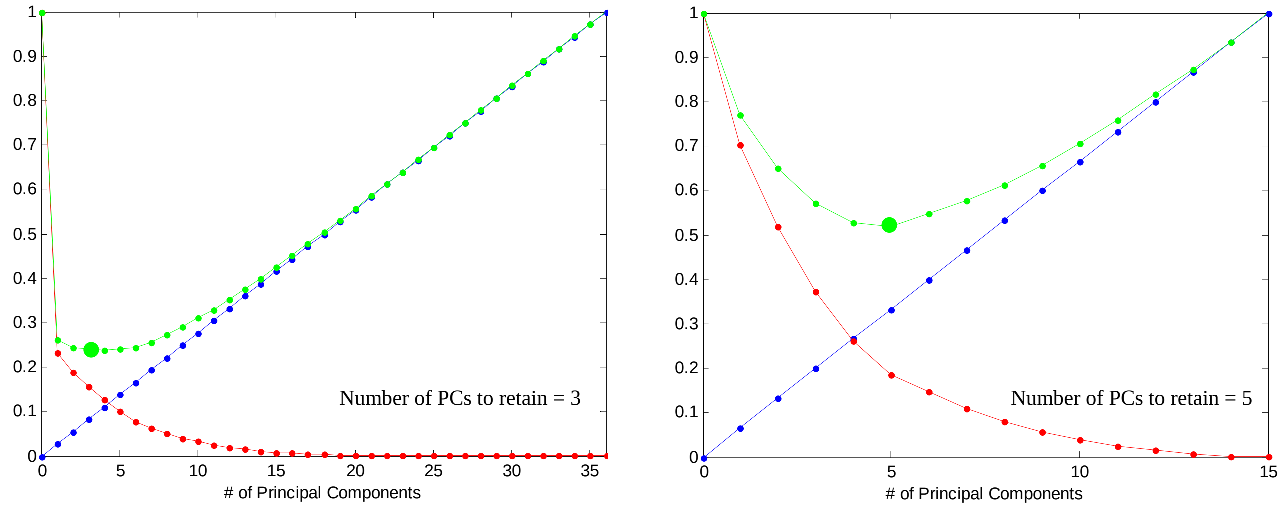


Figure 3.24: Examples of proposed method for selecting number of PCs in Simulator an OR studies (blue: complexity measure; red: (1 - VAF); green: sum of the two parameters); large dot: optimal point

In the studies reported later (Chapters 4 and 5), we perform PCA on high dimensional data sets (~36 and 15 variables respectively). The figure above shows the fraction of variance unaccounted for in these data sets as a function of the number of PCs retained vs. the complexity associated with including more dimensions. These plots show that a minimum of the sum of these two values occurs at a dimension of about 3-5, at which point 80-90% of the total data variance is accounted for. We typically present plots showing the two principal components, which often accounts for 70-80% of the total variance, for purposes of visualization, though we normally retain as many principal components in the analysis as are specified by the analysis described above.

3.4.4.3 Dealing With Missing Data

As explained in Section 3.3.3.1.2, some surgeons may not take advantage of the full repertoire of behaviours (e.g., actions) allowed for in the MCMD and therefore some elements of the matrices may be missing data. According to the literature [Jolliffe 2002, Johnson 2002, Jackson 1991] the most common ways of dealing with such situations are:

- Deleting entirely any observation (i.e., a row) for which at least one of the variables has a missing value or replace it with zero

- Replacing missing values for a variable by the mean of all values in the corresponding row of the variable (i.e., imputation)

We have two types of PCA matrices: (1) based on difference measures (i.e., JSD and Dvalues) where rows correspond to pairs of subjects being compared and columns correspond to performance measures, and (2) based on actual measurements (i.e., holding times and average kinematics) for each subtask where rows correspond to individual subjects and columns correspond to the type of measurement. Consequently, in our case, the first method is not applicable since ignoring whole rows means ignoring comparison possibilities across subjects; moreover, replacing missing data with zeros can significantly distort the results of the PCA. In our application given that every row corresponds to a set of measures from a same single subject or from a same subject-to-subject pair, we therefore implemented the second method which would not alter the PCA analysis significantly.

3.4.4.4 Assessing Repeatability of PCA Analysis

When computing a PCA for a given set of data, it is reasonable to ask how stable the resulting eigenvectors are likely to be to resampling. Two main approaches have been developed to discuss what it is known as the ‘sensibility’ of PCA [Jolliffe 2002]:

- Index of repeatability of PC directions, proposed by Dudzinski [Dudzinski 1975]. They examined how much the directions of the principal components from samples of different sizes differ from those of the population from which the samples were derived. In this method, for each component of interest the angle between the vector of coefficients in the population and in the sample is calculated; then a repeatability index is defined ad-hoc as the proportion of times in repeated samples that this angle has a cosine greater than 0.95. Although this index was custom-designed for that particular application, one generalization would be to report the RMS angular deviation itself.

- Criterion of stability for eigenvectors, proposed by Daudin [Daudin 1988]. They suggested a function to measure the distance between subsets of k PCs in order to choose how many PCs to retain.

We built on the ideas of Daudin 1988 to compute a stability index for subsets of PCs using a bootstrap approach [Besse 1992, Daudin 198, Besse 1988].

Given a data matrix $X^{(P)}$ of size mxn, where m = number of samples and n = number of variables, and (P) represents the population of data obtained experimentally, a PCA analysis produces an mxm matrix, $U^{(P)}$, where each column represents one of the eigenvectors, and a transformed data matrix $Z^{(P)} = U^{(P)'} X^{(P)}$ that is also mxn, where each column represents the weight applied to the first PCA eigenvector.

$X^{(P)}$ is then resampled to get a particular (re)sample X, where the individual rows of X can be drawn from any row in $X^{(P)}$, so X is a row-shuffled version of $X^{(P)}$. After performing a PCA on X, a new set of eigenvectors U and a transformed sample Z are obtained.

For the stability computation, we then have:

$X^{(P)}$: original sample

$U^{(P)}$: PCA eigenvectors

$Z^{(P)}$: transformed original sample

X : resampled data

U : PCA eigenvectors computed from resampled data

The transformed sample Z is obtained as Z = U' * $X^{(P)}$ – i.e., we use the resampled eigenvectors, but process the original data $X^{(P)}$ rather than the resampled data X; this enables us to compare how variable the transformed data is under different PCA eigenvector estimates.

The variability A_k for the E_k subspace composed of the first k eigenvectors in the resampled data relative to the $E_k^{(P)}$ subspace composed of the first k eigenvectors in the original data, is then defined, following [Daudin 1988], as:

$$A_k = \sum_{i,j=1}^{k} \vartheta_{ij}{}^2$$

where ϑ_{ij} = $\mathrm{corr}^{(P)}(Z_i^{(P)}, Z_j)$ – i.e., the correlation in the original dataset between the ith variable of the transformed original data and the jth variable of the transformed resampled PCA.

If U = $U^{(P)}$, then the correlation coefficients will all be 1. If it differs, then there will be some divergence. If the first eigenvector does not shift much, then Z_1 will be approximately equal to $Z_1^{(P)}$ and the correlation will be high.

The stability measure 'S' and the stability index 'S_index' (0 means full stability; 1 means no stability) are subsequently computed as:

$$S = E(A_k - k)^2$$

$$S\_index = S/k \qquad 0 \leq S\_index \leq 1$$

The results from the implementation of these computations are presented in Chapter 4.

3.5 Overview of Experiments

Our main goal in this thesis is to test whether quantitative data analysis based on performance measures such as holding times, kinematics and patterns of movement transitions is able to distinguish between skill levels in the operating room (OR).

After consulting surgeons at Vancouver Hospital who are responsible for training residents, the research team decided to concentrate on analyzing performance during laparoscopic cholecystectomy procedures. We began by developing the MCMD described earlier and then developed the analytical approach outlined in this chapter.

In the remaining chapters, we apply this methodology in two experimental conditions: (1) a physical surgical simulator, and (2) the live operating room.

For the first scenario, a physical simulator, the task was to dissect 2-3 mandarin oranges. Three groups of subjects representing three different skill levels participated in this study. We applied our proposed assessment methodology and were specifically interested in evaluating if (1) intrasubject repeatability was good, (2) scores for trainees with similar skill levels were similar, and (3) scores for trainees at different stages were significantly different. We presumed that if these conditions were met, the technique would be worth testing in the live operative setting.

In a second stage, we proposed moving into a real surgical setting in which the surgical task is less standardized and more subject to interprocedure variability. For this second experiment, we monitored movements of a curved dissector and an atraumatic grasper during 18 laparoscopic cholecystectomy procedures. From the tools' positions, we extracted our performance measures and applied our methodology to compare residents and expert surgeons executing two key surgical tasks: exposing Calot's Triangle and dissecting the cystic duct and artery (CD/CA).

The results for these two experiments are presented in Chapter 4 (Simulator Study) and Chapter 5 (OR Study). **Figure 3.25** shows a summary of the main steps for the implementation of our methodology.

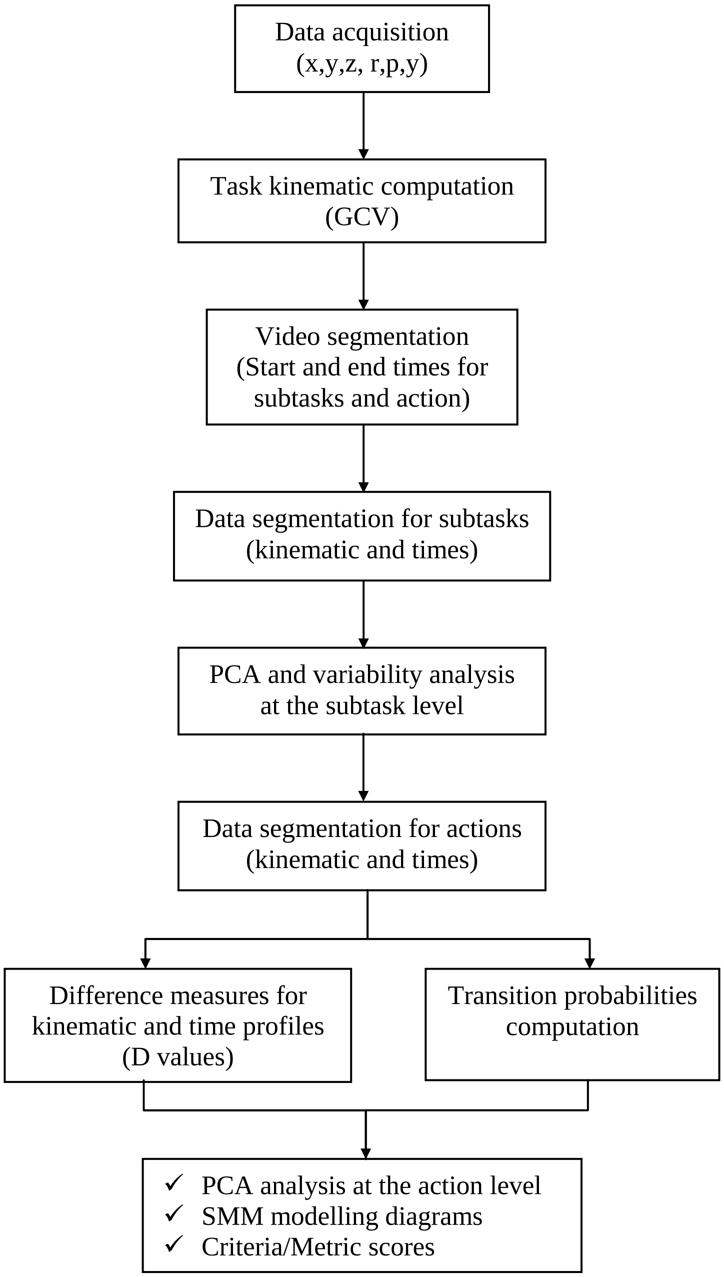


Figure 3.25: Summary diagram of the main steps of the proposed methodology for assessing surgical performance at the operating room. This framework was applied for both the Simulator and the OR studies.

Chapter 4

Feasibility of Using Kinematics Performance Measures to Monitor Laparoscopic Skills

4.1 Introduction

The first attempts to train laparoscopic skills outside the operating room used didactic lectures with some hands-on simulator practice (synthetic and porcine models) to allow trainees to acquire skills in a controlled environment, which is especially effective for developing basic skills [Fried 1999, Derossis 1998, Rosser 1997,]. However, this method has been criticized for being unrealistic; therefore, the idea of more realistic pilot-like training was developed and various types of virtual reality (VR) simulators were proposed such as the MIST-VR, LapSim, Xitact LS500, LapMentor, ProMIS, and Reachin Laparoscopic Simulators which involve more realistic tasks and more objective assessment modules in the form of graded exercises at different skill levels using objective performance measures [Aggarwal 2004].

Practicing skills on VR simulators has shown positive learning curves for trainees who improved their performances until they matched that of the expert surgeons on the same simulated tasks [Gallagher 2002]. Using the MIST-VR tasks (touching, grasping, transferring, and applying diathermy to virtual spheres and cubes within a computer-generated wire frame), Gallagher's group showed that on average novices' performances converged to that of an expert for these basic tasks after four or five training sessions as simulated tasks are likely still not sufficiently representative of the actual operating room and the evidence of effective transfer of skills is still somewhat under-developed. In fact, what is desirable is to see a speed on the OR learning curve and therefore, it is still necessary to measure the effectiveness of simulators before fully incorporating them into the general surgical curriculum [Gallagher 2002, Smith 2001]. Few, if any, would claim that measures of performance in a simulator accurately reflect intraoperative skills, so it is still necessary to conduct intraoperative assessments. Indeed, quantitative methods for assessing intraoperative performance will eventually be needed to demonstrate that performance on a simulator can be considered equivalent to that in the OR.

Despite the ongoing discussion about the efficacy and transferability of simulator training, it remains comparatively expensive to run intraoperative tests, so it is often most cost-effective to test new skill assessment methods on simulators prior to their application in the operating room environment.

In the previous chapter, we outlined an analytical approach to deal with surgical motor skill assessment. We integrated three types of physical measures (kinematics, time and movement transitions) into a modelling technique for quantifying performance of surgical trainees. We first created a hierarchical representation to decompose larger surgical goals into clearly identifiable tasks amenable to being monitored by our measures. Then, at each level of surgical complexity, we implemented specific mathematical techniques to derive intuitive scores for providing a quantitative sense of how far a performance is located from a reference level (i.e., expert surgeons group or a peer group).

Before testing this methodology in the operating room, where the experimental complexity and variability is relatively high, we decided to first use data from a physical surgical simulator in order to make sure that our assumptions are reasonably plausible, and to verify that our data acquisition and processing methods work well in differentiating skill levels. Therefore, our specific objectives at this stage were: (1) to acquire motor performance data on a simulated surgical task (performed in a dry lab setting on inanimate anatomical models) from three sets of subjects representing different stages of training: novices, mid-stage trainees, and experts; and (2) to test whether or not our proposed analytical method is able to reliably distinguish between these three groups of subjects.

More specifically, we were interested in evaluating if (1) intrasubject repeatability is good, (2) scores for trainees with similar skill levels are similar, and (3) scores for trainees at different stages are significantly different. If these conditions are met, the technique will be worth testing in the live operative setting.

In this study, we focused on two levels of analysis: First, we conducted an investigation of average kinematics to determine whether it is possible to find differences between subjects; and second, we assessed details at the action level to determine if our analytical approach was able to identify where in the surgical process any such differences arose.

4.2 Methods

After consulting with the attending surgeons involved in this study, we determined that the task of peeling a mandarin orange and separating the segments would require the principal surgical skills they were interested in assessing.

In this section, we describe the details of the 'mandarin' experiment in terms of the number of subjects and skill groups recruited, the equipment set up and the specific analytical methods used to compare subjects' performances at the subtask and action levels.

4.2.1 Participants

We recruited three sets of subjects to represent different stages of training: novices (represented by three graduate students with no specific surgical training), novices with training (represented by three graduate students who received a half-hour of training from an expert surgeon), and experts (represented by three attending surgeons). All subjects signed consent forms as requested by UBC Research Ethics Board (Appendix E) in order to respect confidentiality and to assure that the data acquired will be only used for research purposes relating to the present study.

4.2.2 Experiment Setup

We simulated a surgical dissection task by asking participants to use laparoscopic tools to peel and separate the segments of two to three mandarin oranges placed in a training box.

The movements of the laparoscopic tool for the dominant hand were tracked using an electro-magnetic Polhemus sensor, which continuously recorded 3D position and orientation data at 120 Hz while the task was being executed (static accuracy of 0.03 inches RMS for the X, Y, or Z position; 0.15° RMS for receiver orientation [Polhemus 2002]). As described in chapter 3, we calibrated the tool by estimating the tip location as the center of a sphere described when moving the tool handle in circles about a fixed tool tip location. We used the collected sphere data and apply a non-linear least squares optimization approach in order to find the location of the tip with respect to the global frame.

In addition, the execution of the task was recorded on videotape so that the investigator could later correlate the movement patterns with discrete phases of the task execution (**Figure 4.1**). Polhemus system was first initialized followed by the video system 10 seconds later and during the pre-processing stage, we eliminated the first 10 seconds of Polhemus data so as to establish a synchronization point between both systems.

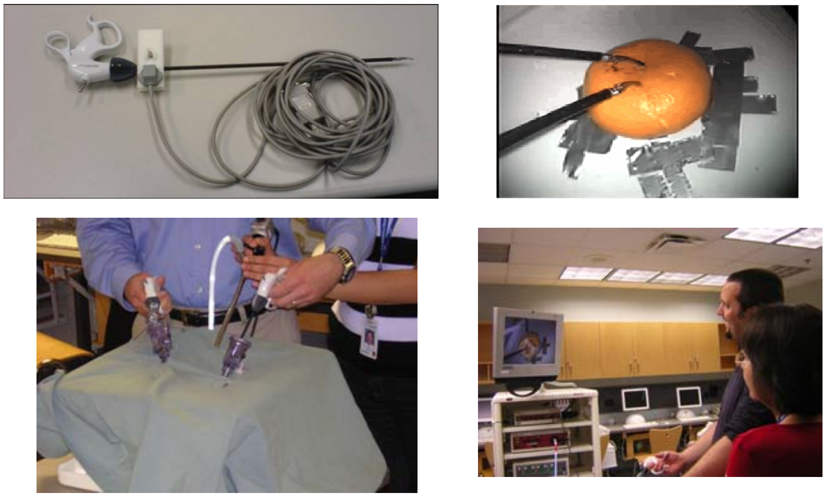


Figure 4.1: Equipment setup for Mandarin Experiment

Following initial instruction in the task and demonstration by the investigator, subjects performed the task at their own pace. Afterwards, the acquired data was processed (Section 3.2.2.2) to extract kinematic measures from the tool movements.

4.2.3 Analytical Methods

Similarly to the surgical representation we have provided in Chapter 2, the mandarin dissection task was described using our MCMD structure (**Figure 4.2**). Four subtasks were identified: (A) Explore, (B) Peel skin, (C) Detach segments, (D) Place segments, with B and C as the most challenging tasks. For the detailed analysis, each subtask was correspondingly represented as combinations of 10 tool movements following the conventions defined in Chapter 2.

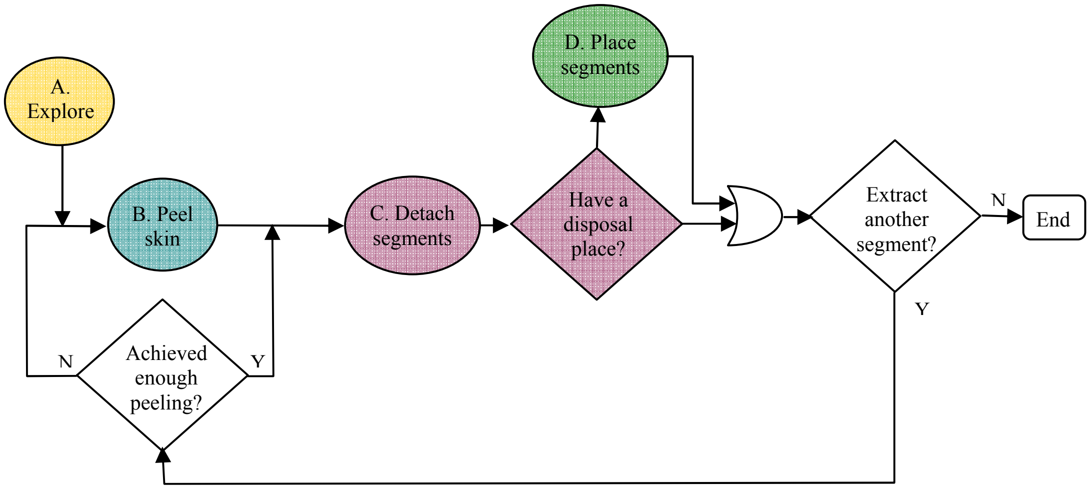


Figure 4.2: MCMD for the Mandarin Dissection Task

We performed two separate analyses: (1) Subtask level: examination of the 3D average kinematics of the surgical tool movement when executing the two main experimental subtasks (Peel Skin and Detach Segment) to identify broad differences; and (2) Action level: decomposition and analysis of individual subtasks using our pre-defined set of ten actions (e.g., push, sweep, spread, etc) based on our assessment methodology (Chapter 3) in order to identify at a detailed level the sources of any detected differences between groups. We followed the flow of steps summarized in **figure 3.25** from Chapter 3 to present the corresponding methods and results for the Mandarin experiment.

4.2.3.1 Subtask Level

The position data from the mandarin dissection task was initially segmented into its four subtasks by manually identifying the start and end points using video analysis. ‘Peel skin’ consisted of removing some mandarin skin to allow for detaching the inner segments. It begins when the tip of the tool first breaks in the skin and ends when at least one segment is completely uncovered. ‘Detach segment’ consists of isolating each segment by separating the neighbouring segments’ walls. It begins when the tip of the tool is first inserted between two walls and ends when the segment has been completely freed from the rest of the mandarin. The video clips were collected for further decomposition at the action level and the time records were used to segment the corresponding position data streams. Afterwards, the 3D velocities of the tool motion during each segmented subtask were derived by differentiation using a generalized cross validation (GCV) algorithm (Section 3.2.2.2).

For each subject ‘i’ and procedure (i.e., mandarin) ‘j’ (S_{ij}), we defined a 6-element vector consisting of the average tooltip velocities in each of the three cardinal directions (i.e., l:lateral, a:axial, v:vertical) for each of the two main subtasks. Although we could analyze each subtask separately, herein we grouped them so as to provide a performance description of the overall task. All subjects were then grouped into a nxm matrix (n:

$\sum i \sum j$, total number of experiment repetitions; m: number of dimensions, m=6 for this application). All subtasks ('peel' or 'detach') repetitions during one single procedure were concatenated in single vectors and each entry in the **Table 4.1** corresponds to the average of the consolidated data for all subtasks in a given procedure.

Subject / trial	B. Peel skin			C. Detach segment		
S11	$V_{11}l_B$	$V_{11}a_B$	$V_{11}v_B$	$V_{11}l_C$	$V_{11}a_C$	$V_{11}v_C$
S21	$V_{21}l_B$	$V_{21}a_B$	$V_{21}v_B$	$V_{21}l_C$	$V_{21}a_C$	$V_{21}v_C$
…	…	…	…	…	…	…
Sij	$V_{ij}l_B$	$V_{ij}a_B$	$V_{ij}v_B$	$V_{ij}l_C$	$V_{ij}a_C$	$V_{ij}v_C$

Table 4.1: Velocity data arrange for PCA analysis at the subtask level

As described in section 3.4.4.1, we then normalized the data by dividing each element in a column by the column's standard deviation (defined across all trials and subjects), and used Principal Components Analysis (PCA) to extract the dominant contributors to overall variability to simplify the presentation of the data to the trainer. By analogy to the display in **Figure 3.23**, the PCA produces an m column matrix, B, of eigenvectors of length m, a weight matrix ω (nxm), where each row represents a remapping of the original data into a weight space such that the original data can be reconstructed as $\sum \omega_i B_i$, and the variance explained by the corresponding principal component.

We then plotted the unaccounted variance against the complexity measure (proportional to number of dimensions) to select the appropriate number of dimensions to retain in this new weight space (Section 3.4.4.2). The aim is to reduce the dimensionality of the data in such a way to facilitate the visual presentation of the data (in this case from 6D to 2D, **Figure 3.23**).

In this weight space, each subject's execution is represented as a data point and groups of points correspond to repetitions from a single subject or from subjects belonging to the same skill level. It is in turn possible to use the concept of distance to assess variation within and between individuals and groups. We separately evaluated the two null hypotheses that the three groups all have the same means, and that the nine subjects all have the same mean. We reported the ratio of mean square distance (MSD) in the PCA weight space from the mean position of all trials executed by a specific group or subject to the MSD from the global mean position to describe variability for specific groups or subjects. We statistically evaluated our comparison using a one-way ANOVA (see Appendix G) test for group comparison (intergroup DoF=2; intragroup DoF=19; α=0.05) and a one-way ANOVA test for subject comparison (intersubject DoF=8; intrasubject DoF=13; α=0.05).

To evaluate whether or not group performance measures shift progressively in the weight space as surgical skill level increases, as illustrated in **Figure 4.3**, we tested the null hypothesis that the intergroup distances between the novices and the other two groups was zero and that distances between each subject and the mean expert position are equal. We therefore computed the distances from each subject and to the center of the experts' group (selected as the reference level) and used a non-parametric test (Mann-Whitney) to show statistical significance (p-value < 0.05).

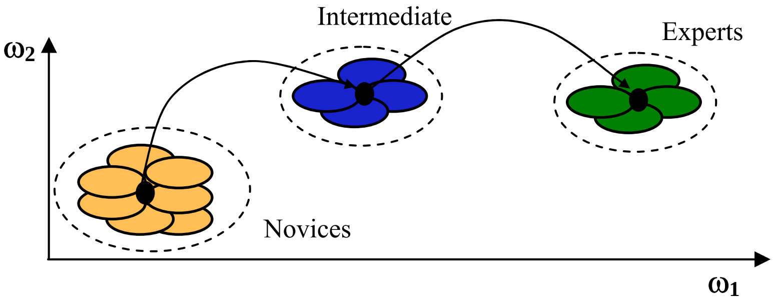


Figure 4.3: Schematic representation of hypothesized shift in group performance measures with increasing skill level (shown by arcs)

We hypothesize that plots of the extracted principal components and the derived variation measures will show consistency in individual performance, similarity amongst individuals at similar levels of training and distinctions between the different groups of subjects when comparing them across various executed subtasks (i.e., 'Peel Skin' and 'Detach Segment'). To test our three hypotheses, we computed the contributions to total variability from intrasubject, intragroup (i.e., equivalent stages of training) and between group variations and report these numbers as percentages of total variation from the global data mean.

4.2.3.2 Action Level

Moving further down in the hierarchy, we also compared performances of single subtasks by decomposing them into 10 characteristic actions: push, pull, reach, orient, sweep, spread, grasp&hold, grasp&cut, idle, and out. The process of segmenting and obtaining the times records for these actions was achieved by identifying the start and end points through video analysis according to the action definitions provided in Chapter 2. For each subtask repetition during each experiment execution (S_{ij}), Excel templates were used to record the timing information and to compute the matrix of action transitions. **Table 4.2** presents an example of how data is processed in Excel to obtain the list of transitions and holding times for one subtask execution.

	TIME IN			**TIME OUT**				
ACTION	**min**	**sec**	**30ths**	**min**	**sec**	**30ths**	**Transition**	**HT (sec)**
Reach	0	0	0	0	4	20	**Reach-Push**	4.67
Push	0	4	20	0	5	0	**Push-Reach**	0.33
Reach	0	5	0	0	6	22	**Reach-Out**	1.73
Out	0	6	22	0	8	26	**Out-Idle**	2.13
Idle	0	8	26	0	9	19	**Idle-Out**	0.77
Out	0	9	19	0	10	1	**Out-Reach**	0.40
Reach	0	10	1	0	10	22	**Reach-Out**	0.70
…								

Table 4.2: Decomposition of a particular subtask into its corresponding set of executed actions. Start and end times for each action were registered and every action transition was obtained

Using the list of time in and time out values, we segmented the kinematics signal of the subtask and were therefore able to associate segments of the kinematics data with each action. At this level, we characterized each action using distributions of holding times and kinematic measures (both summary and detailed, Section 3.3.3.1.2), and we computed the transition matrices representing the state transitions.

To evaluate the same hypotheses described above for the subtask level that there were no intergroup differences, we first computed direct differences between two subjects as it is easier to compare distributions and transition matrices this way (section 3.3.2). For comparing performances between two subjects M_{r1} and M_{e1} (resident #1 vs. expert #1), we defined: (a) $M_{r1}.T_k.S_l.A_m v$[1] and $M_{e1}.T_k.S_l.A_m v$ as the group of kinematic (e.g., velocity) distributions for all executed actions; (b) $M_{r1}.T_k.S_l.A_m t$ and $M_{e1}.T_k.S_l.A_m t$ as the

[1] Represents vector of velocity measures for m action during execution of l subtask and k task

group of holding time distributions for all executed actions; and (c) TMP_{r1} and TMP_{e1} as the action transition probability matrices. We then computed difference measures for the various corresponding kinematics ($M_{r1}.T_k.S_l.A_m v$ vs. $M_{e1}.T_k.S_l.A_m v$) and holding time ($M_{r1}.T_k.S_l.A_m t$ vs. $M_{e1}.T_k.S_l.A_m t$) distributions using the Kolgomorov-Smirnov statistic (D measure); and difference measures for the transition matrices using the Jensen-Shanon divergence (JSD measure) after modelling the system as a Semi-Markov process (Section 3.4.3).

Across all subject comparisons, we obtained a v x w matrix, where v corresponds to the number of subject comparisons (M_{r1} vs. M_{e1}) and w to the number of extracted performance measures.

$$\begin{bmatrix} M_{r1} vs. M_{E1} \\ M_{r1} vs. M_{E2} \\ \vdots \\ M_{r3} vs. M_{E3} \end{bmatrix} = \left[\begin{array}{c||c||cc} JSD_{r1e1} & 9D_{r1e1}.T_k.S_1.A_m.t & 27D_{r1e1}.T_k.S_1.A_m.(\bar{v},\bar{a},\bar{j}) & 27D_{r1e1}.T_k.S_1.A_m.(v,a,j) \\ \vdots & & \vdots & \vdots \\ JSD_{r3e3} & 9D_{r3e3}.T_k.S_1.A_m.t & 27D_{r3e3}.T_k.S_1.A_m.(\bar{v},\bar{a},\bar{j}) & 27D_{r3e3}.T_k.S_1.A_m.(v,a,j) \end{array}\right]$$

For the mandarin experiment at the action level there were 36 subject-to-subject comparisons: n(n-1)/2 = 3 intragroup comparisons for novices, nn(nn-1)/2 = 3 intragroup comparisons for novices-with-training, e(e-1)/2 = 3 intragroup comparisons for the attendings, nxnn = 9 intergroup comparisons between novices and novices-with-training, nxe = 9 intergroup comparisons between novices and attendings, nnxe = 9 and intergroup comparisons between novices and novices-with-training. We also computed 64 difference (D) measures for each row: 1 JSD value; 9 D values (per variable) for average time, and summary velocity, acceleration, and jerk profiles (36 total); 9 D values (per variable) for detailed velocity, acceleration, and jerk profiles (27 total)[2]. We then used PCA as described in section 3.4.4 to reduce the dimensionality of this data.

4.2.3.3 Stability of Principal Components

For assessing the stability of principal components we implemented a bootstrapping approach to create resamples of the original dataset and followed the approach described in section 3.4.4.4 in order to study the stability of the principal components derived from the PCA analyses.

Our original data matrix, $X^{(P)}$, was a 21x6 matrix corresponding to 21 subjects and 6 velocity measures at the subtask level, and a 64x36 matrix corresponding to 36 subject-to-subject pairs and 64 difference measures (D) at the action level. After applying PCA on $X^{(P)}$, we obtain:

- a 6x6 $U^{(P)}$ (for subtask level) or a 36x36 $U^{(P)}$ (for action level) matrix containing the PCA eigenvectors of the original data $X^{(P)}$, and
- a 21x6 $Z^{(P)}$ (for subtask level) or a 64x36 $Z^{(P)}$ (for action level) matrix containing the transformed original data

We then create 100 resamples of X by bootstrapping $X^{(P)}$ 100 times and apply PCA on each resampled dataset, which produces the corresponding 100 U matrices for the resamples.

[2] We reduced our set of actions to 9 since ‘Grasp&Cut’ was not applicable to this experiment.

For computing the variability A_k for the E_k subspace composed of the first k eigenvectors in the resampled data around the $E_k^{(P)}$ subspace composed of the first k eigenvectors in the original data, we estimate the 100 transformed Z data matrices as Z = U' * $X^{(P)}$ and compute the 100 A_k values and the stability index S for each E_k subspace (see section 3.4.4.4 for details).

4.3 Results

In this section we present the results from implementing our proposed methodology to the Mandarin experiment at two levels of analysis.

4.3.1 Subtask Analysis

4.3.1.1 Overview of Kinematic Data

Figure 4.4 shows samples of cumulative distributions of velocity components for three subjects (E1, NovT1, Nov1) and the corresponding root-mean-square values, which are the actual variables that are arranged into a matrix and used in the PCA analysis. It is apparent that for both subtasks, there is significant separation between the subjects with the novices performing at a slower pace than the novices-with-training and the experts. NovT subjects seem to use middle range velocities with is consistent with the idea that improvement in motor skills is related to progressive training.

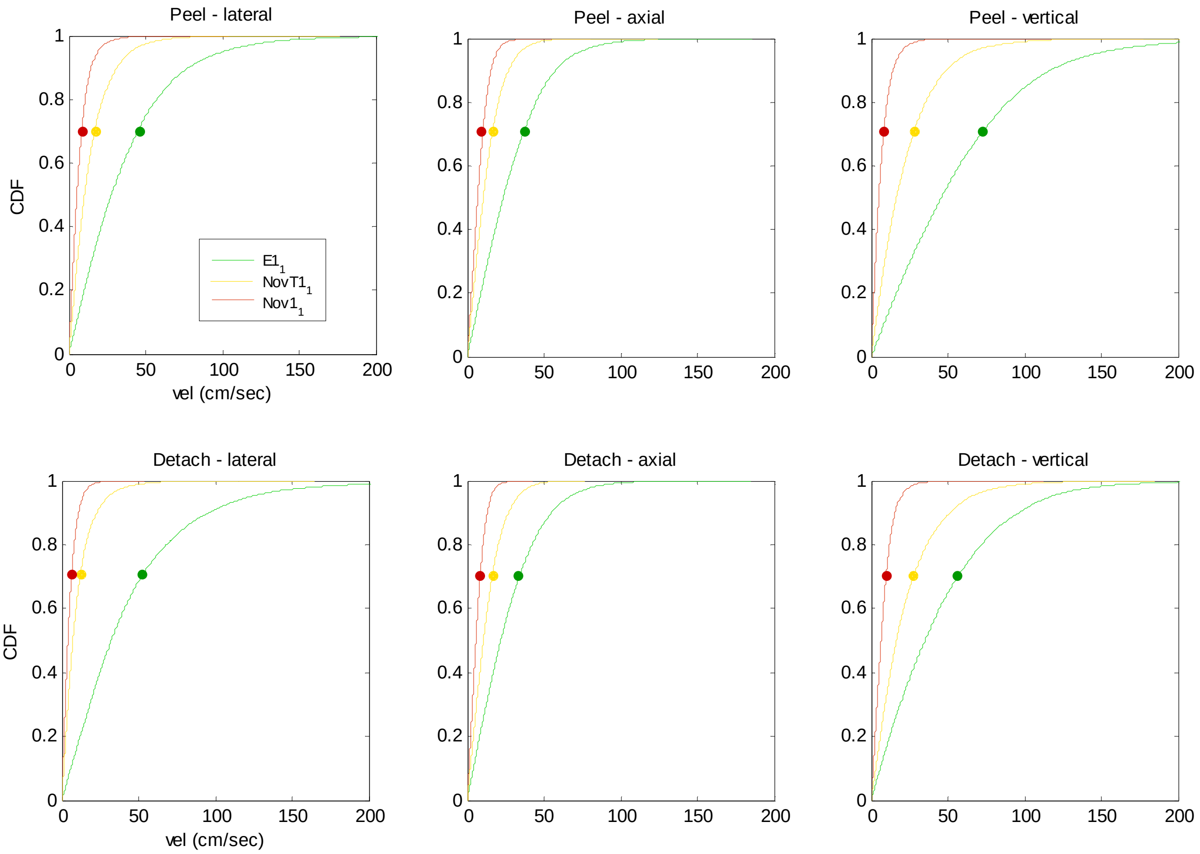


Figure 4.4: Samples of Cumulative Distribution Functions (CDF) for velocity components in different directions for two subtasks for a representative subject from each group. Colored dots indicate mean velocity for each subject, direction and task

4.3.1.2 PCA Analysis

At first we analyzed the influence of time and velocity and therefore defined a 8D data set composed of 6 rms (root-mean-square) velocity components (lateral, axial, and vertical velocities for 'Peel skin' and 'Detach segment' respectively) and the average time spent at each subtask. Since velocities seemed to be the dominant differentiator

(Appendix F), we reduced the dataset and only concentrated on analyzing velocities in a 6D space.

We performed a PCA analysis and by applying our PC selection method (Section 3.4.4.2) found that by retaining only the first two principal components it was possible to explain ~80% of the variance across all subjects and subtasks (**Figure 4.5**).

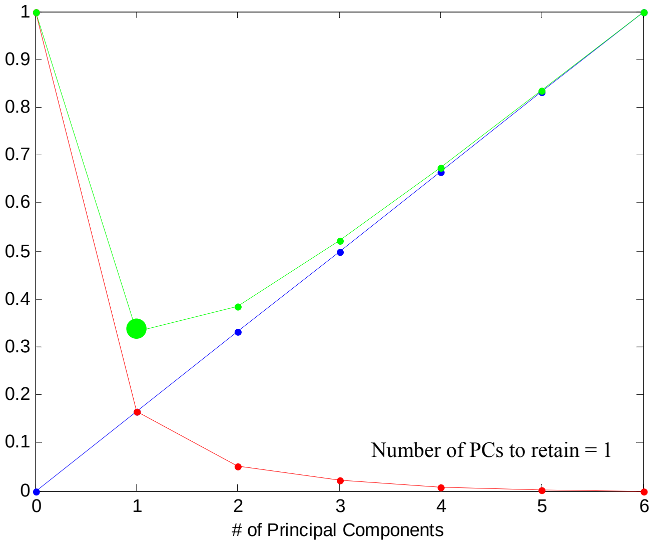


Figure 4.5: Selecting number of Principal Components for 6D dataset blue: complexity measure; red: variance unaccounted for; green: dimensionality-decision criteria (sum of the two parameters); optimum indicated by large green dot

Analysis of the coefficients and signs of the first two principal components (**Figure 4.6** and **Figure 4.7**) showed: (1) that in the first principal eigenvector, all velocity variables covary, which indicates that the primary feature characterizing the subjects is the overall speed with which they move the tool; and (2) that the second eigenvector displays a contrast between the vertical and lateral velocity components. This effect was more prevalent in the 'detach' subtask as indicated by the larger PC value there.

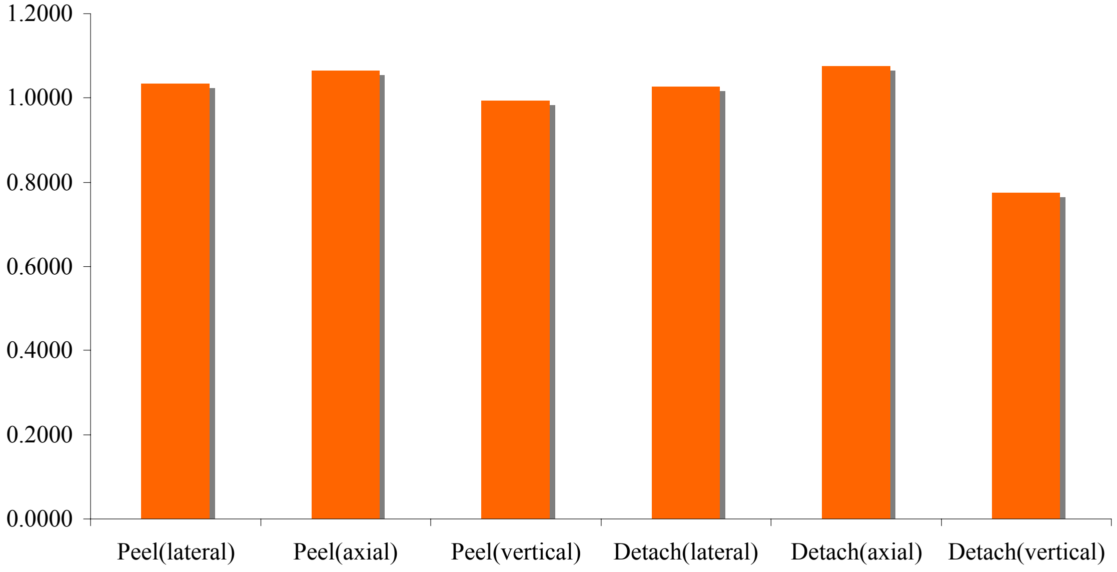


Figure 4.6: Normalized coefficients for PC1 in 6-D data set composed of kinematic parameters[3]

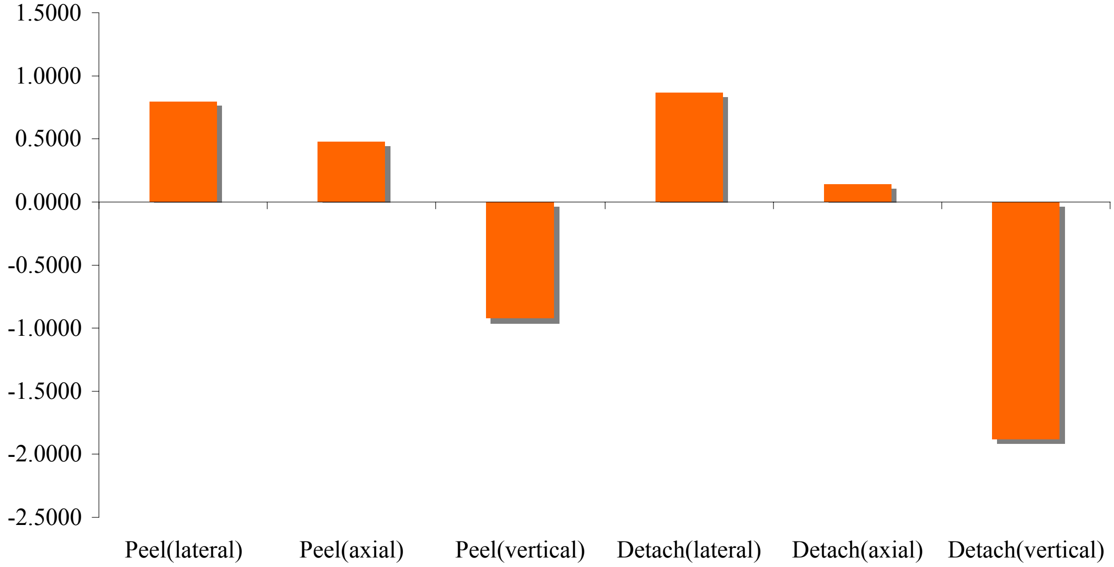


Figure 4.7: Normalized coefficients for PC2 in 6-D data set composed of kinematic parameters[4]

[3] The principal components were normalized (multiplied by $\sqrt{\#\,\mathrm{of}\cdot\mathrm{PCs}}$)

The position of each trial in PCA-weight space is shown in **Figure 4.8**. It appears that intrasubject repeatability is generally high, that the data from subjects of comparable training level is in relatively close proximity to one another, and that there are significant variations between groups. It also appears from the plot that variation in the 2nd direction is primarily due to variation amongst experts. In the next section, we test these observations statistically.

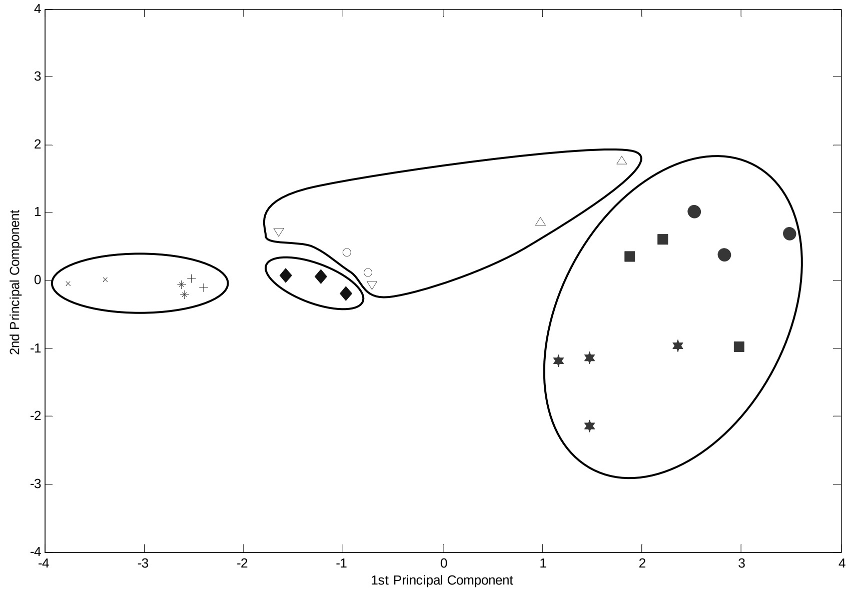


Figure 4.8: Cross-plot of the first 2 principal components for the three subject groups tested. [x,+,*]: novices (Nov); [▽,O,△]: novices with training (NovT); [(●◆)■✱]: experts (Exps); [◆]: expert #1 (E#1) measured while instructing NovT group; [●]: expert #1

Figure 4.9 (pieplot) shows the contributions to total variability for the velocity 6-D analysis. The low values of intrasubject and intragroup variability support the qualitative observations that the greatest contributor to overall variability is difference in degree of

[4] The principal components were normalized (multiplied by $\sqrt{\#\,\mathrm{of}\cdot\mathrm{PCs}}$)

training. The one-way ANOVA test implemented for group comparisons (intergroup DoF=2; intragroup DoF=19; α=0.05), showed that there is detectable intergroup variation and detectable intragroup (subjects within a group) variation.

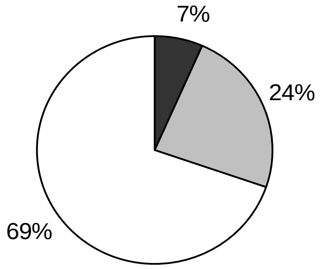


Figure 4.9: Contributions to total variability. [■]: intrasubject variability; [▒]: intragroup variability; [ ]: intergroup variability (one-way ANOVA F=38.37; Fcrit.=3.52 at α=0.05)

Additionally:

a) **Figure 4.10** shows that intrasubject variability is lowest with the novices and increases with experience level. The one-way ANOVA test implemented for subject comparisons showed there is detectable intersubject variation and detectable intrasubject (trials within a subject) variation. (intersubject DoF=8; intrasubject DoF=13; F=5.96; $F_{crit.}$=2.77 at α=0.05;)

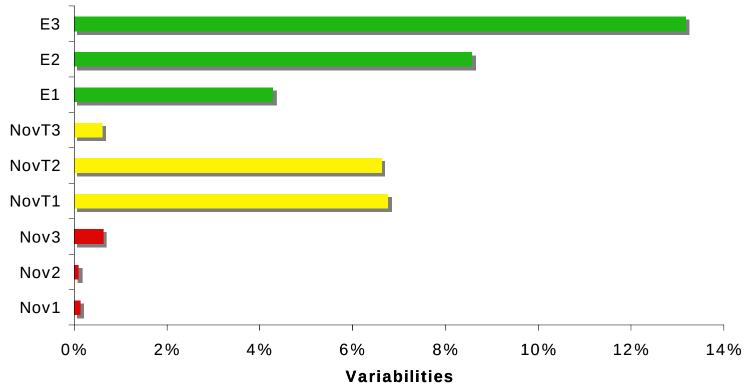


Figure 4.10: Intrasubject variabilities for the three skill levels

b) The root-mean square (RMS) distances between different groups are considerably larger than the RMS distances within groups for the subjects included in this experiment (Mann-Whitney test, p-value = 0.049); this is consistent with the idea that training changes motor patterns (**Figure 4.11**).

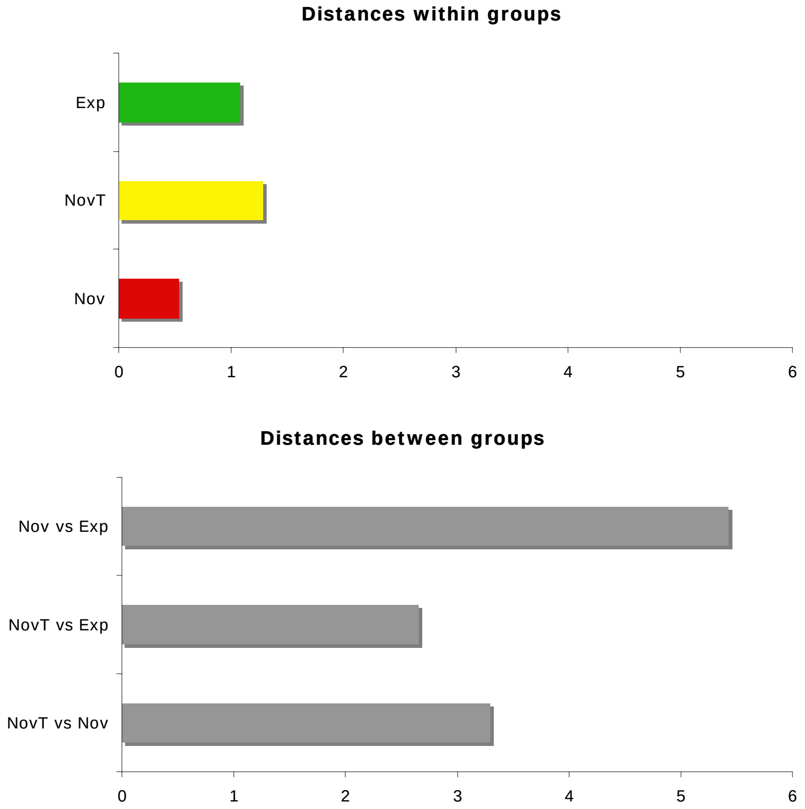


Figure 4.11: Distances (average) within groups and distances between groups

c) The distances from each subject to the mean expert position is shown in **Figure 4.12**. Mann-Whitney test (p-value = 0.049) showed that distances between novices and experts' center are considerably larger than distances between novices-with-training and experts' center and distances between experts and experts' center.

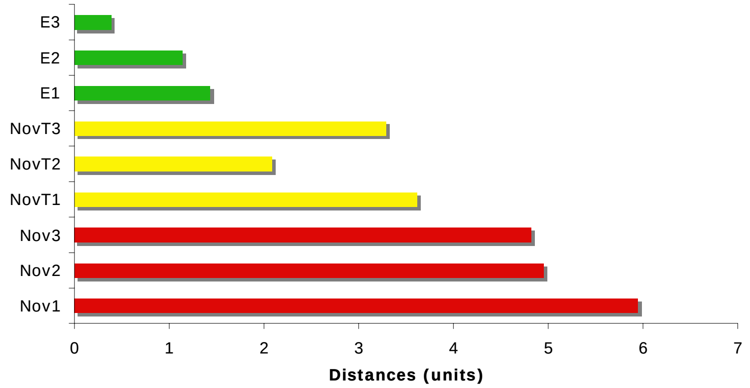


Figure 4.12: Distances between each subject's center and experts' group center

We also noticed in **Figure 4.8** that expert #1's performance while instructing was more similar to that of the trainee group than to the expert group or to his own typical performance. The RMS distance of these instructional trials was 1.62 to the centre of the novices-with-training group and 3.63 to the centre of the experts' group, which suggests that this surgeon changed his performance to more closely match the capabilities of those he was instructing.

4.3.2 Action Analysis

4.3.2.1 Overview of Kinematic Data

Figure 4.13 and **Figure 4.14** show samples of cumulative distributions of velocities for 9 actions and three subjects (E1, NovT1, Nov1) while executing each subtask (Peel and Detach). Subjectively, there appears to be high intrasubject consistency across experiment repetitions. This is confirmed (**Figure 4.15** and **Figure 4.16**) by the low K-S values between intrasubject pairs of distributions form different task repetitions for the peel and detach subtasks.

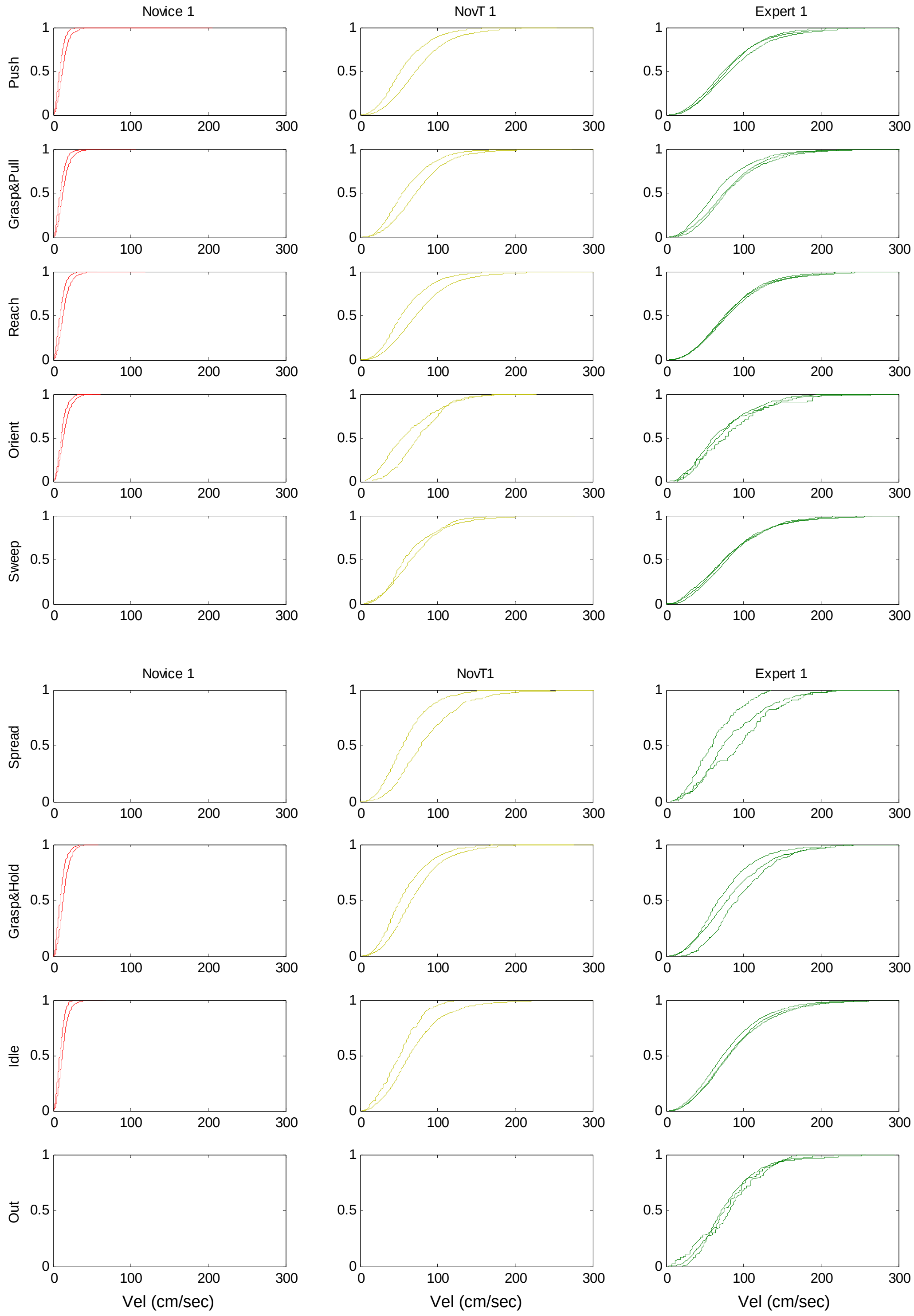


Figure 4.13: Samples of CDFs for velocity components (horizontal axes) at each action during 'Peel' for representative subjects from the different groups. Empty plots indicate subject did not use the specified action in performing the task.

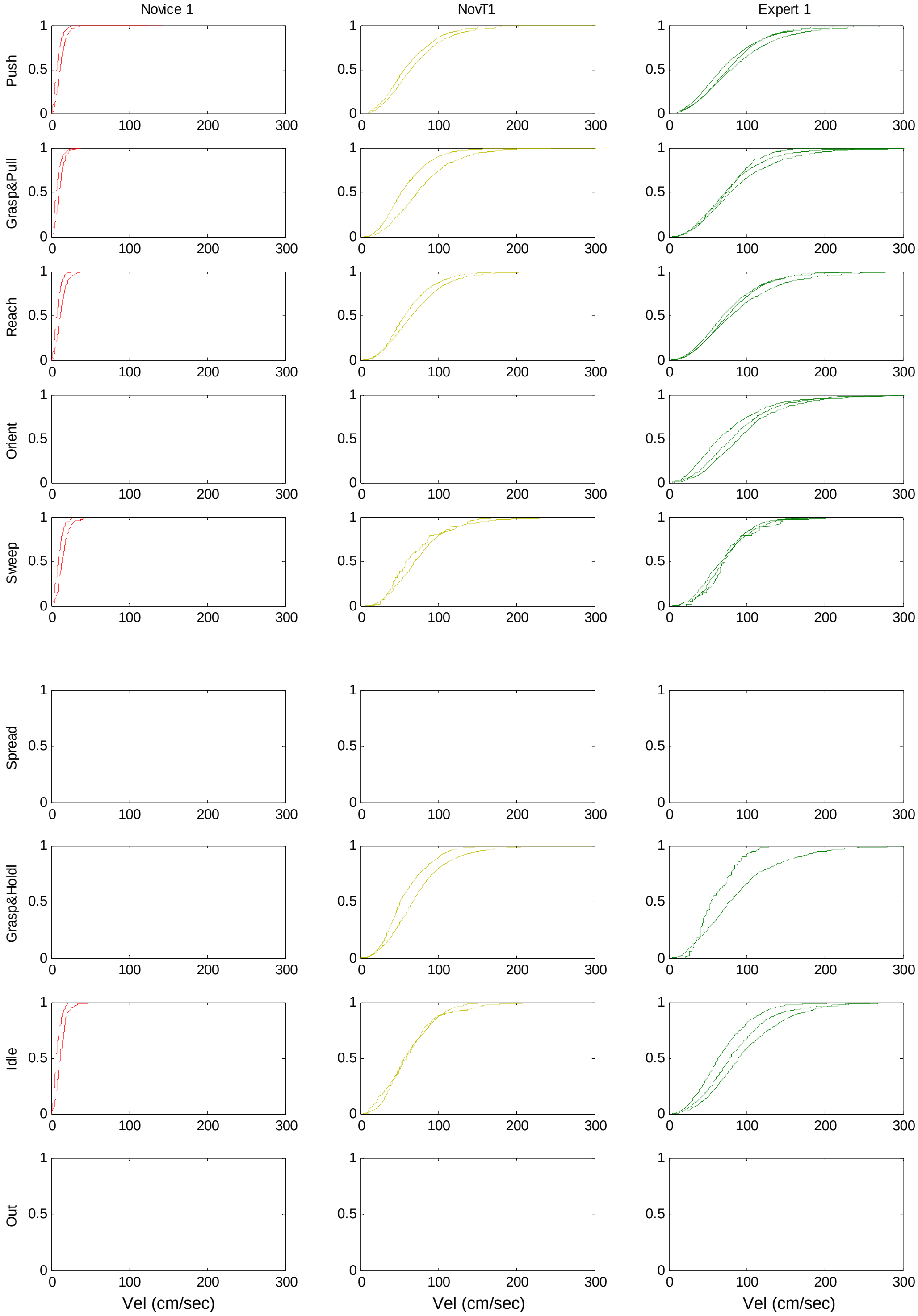


Figure 4.3: Samples of CDFs for velocity components (horizontal axes) at each action during 'Detach' for representative subjects from the different groups Empty plots indicate subject did not use the specified action in performing the task.

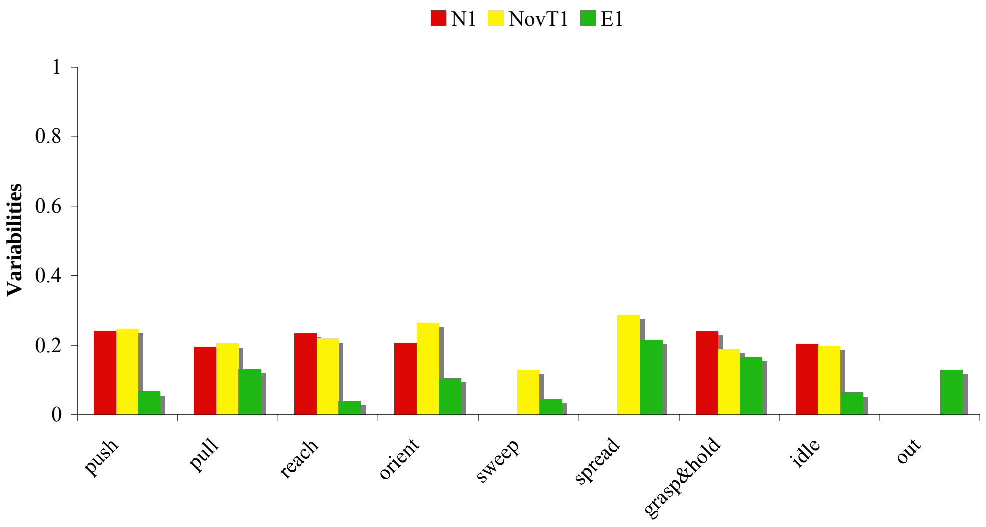


Figure 4.15: Consistency measure (mean D values) for actions executed by one novice, one novice with training and one expert while in ‘Peel’ subtask

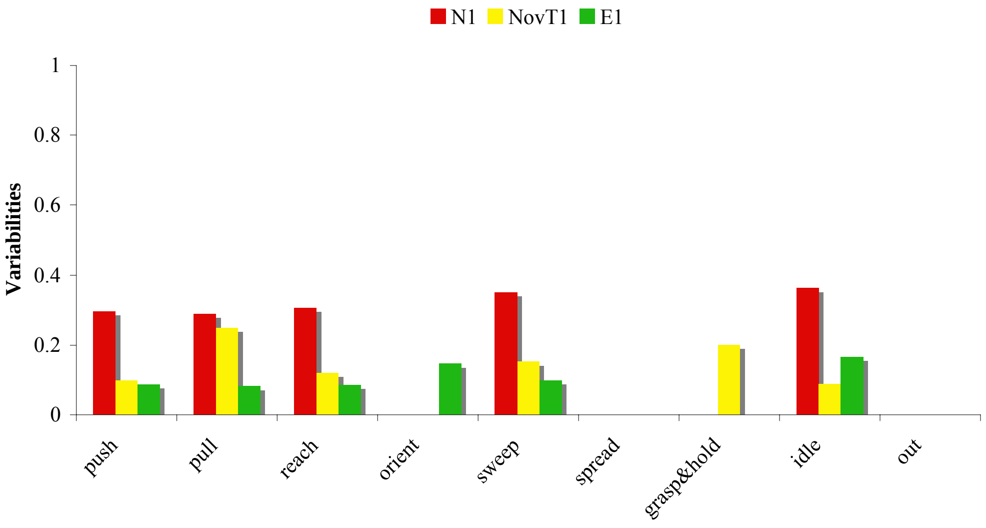


Figure 4.16: Consistency measure (mean D values) for actions executed by one novice, one novice with training and one expert while in ‘Detach’ subtask

We also found, using paired t-tests, that intersubject differences within a group were significantly larger than intrasubject differences, and that intersubject differences between groups were significantly larger than intersubject differences within groups (**Figure 4.17**).

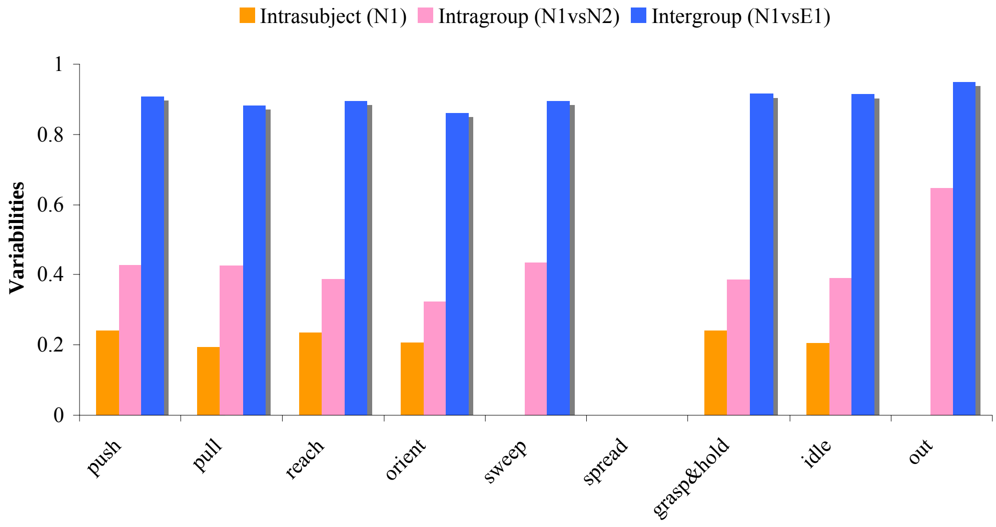


Figure 4.17: Intrasubject, intra- and intergroup comparisons (mean D values) for actions in 'Peel' subtask. Paired t-test between intrasubject (orange) and intragroup (pink) D values: p-value = 0.003 Paired t-test between intrasubject (orange) and intergroup (blue) Dvalues: p-value < 0.0001

4.3.2.2 PCA Analysis

Applying our PC dimension selection method, we found that the first three principal components should be retained to explain most of the variance across all subjects and subtasks (**Figure 4.18**). However, since there was little difference in the dimensionality-decision criteria for two and three principal components, we decided, for ease of visualization and interpretation, to present the data in a two-dimensional space (i.e., retaining more than 75% of the information), though we computed all distances in the 3D space weight.

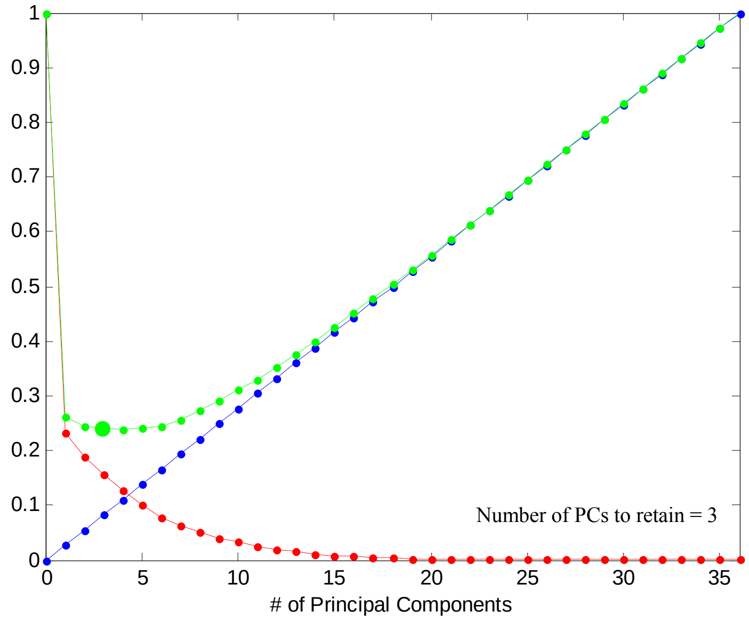


Figure 4.18: Selecting number of Principal Components for 36D dataset. Blue: complexity measure; Red: variance unaccounted for; Green: dimensionality-decision criteria (sum of the two parameters)

The distribution of data points on the PCA plot in **Figure 4.19** shows that differences in intragroup (pairs of subjects from the same skill level) and intergroup (pairs of subjects from different skill levels) comparisons are primarily differentiated along the horizontal axis. Intragroup comparisons are located on the left while comparisons between novices and experts are located on the right (see circled points). Data points in between are representations of comparisons of novices-with-training to either experts or novices. It also appears that comparisons between novices (N) and novices-with-training (NovT) are generally closer to the intragroup comparisons, while novices-with-training (NovT) vs. experts (E) appear to be closer to the comparisons between novices and experts.

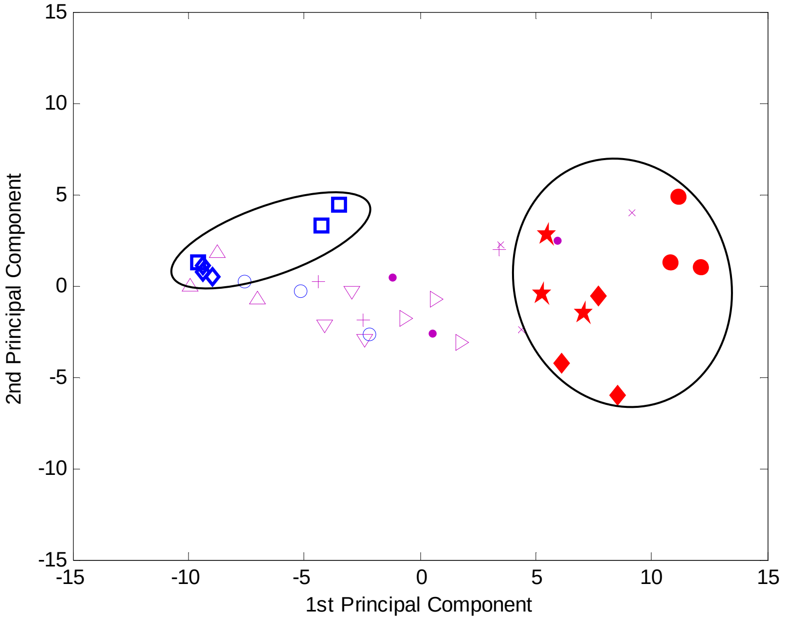


Figure 4.19: Cross-plot of the first 2 principal components for the 36 subject pairs tested.
Left circle - [◇]: Experts vs. Experts, [□]: Novices vs. Novices, [○]: NovT vs. NovT
Right circle - [●,◆,✶]: Novices (N1,N2,N3) vs. Experts (E1,E2,E3)
In between - [△,▷,▽]: (NovT1,NovT2,NovT3) vs. (N1,N2,N3) and
[×,+,•]:(NovT1,NovT2,NovT3) vs. (E1,E2,E3)

Interpretation of principal component eigenvectors (Appendix F) indicated that times did not contribute strongly to differentiating between subjects or groups. **Figure 4.20** to **Figure 4.22** show representation of 63 components of the corresponding first 3 eigenvectors at the action level:

a) The first eigenvector suggests that difference measures derived from kinematics $\left[Dvalues(\bar{v},\bar{a},\bar{j}) \quad Dvalues(v,a,j)\right]$ provide most of the differentiation and that holding time contributions play relatively smaller role (**Figure 4.20**)

b) The second eigenvector shows that kinematics Dvalues provide major contrast between [*Grasp&Pull* and *Orient*] vs [*Push* and *Sweep*] (i.e., contrast of how selected movements outperformed) [5] (**Figure 4.21**)

c) The third eigenvector shows *Grasp&Hold* and *Idle* with higher coefficient values over the other actions (i.e., contrast between the less action oriented movements). Moreover, PC3 gave more importance to average vel ($\bar{v}$) over the other kinematic measures (**Figure 4.22**)

[5] Since time did not contribute strongly to the analysis, we did not present its D value in figures 4.21 and 4.22

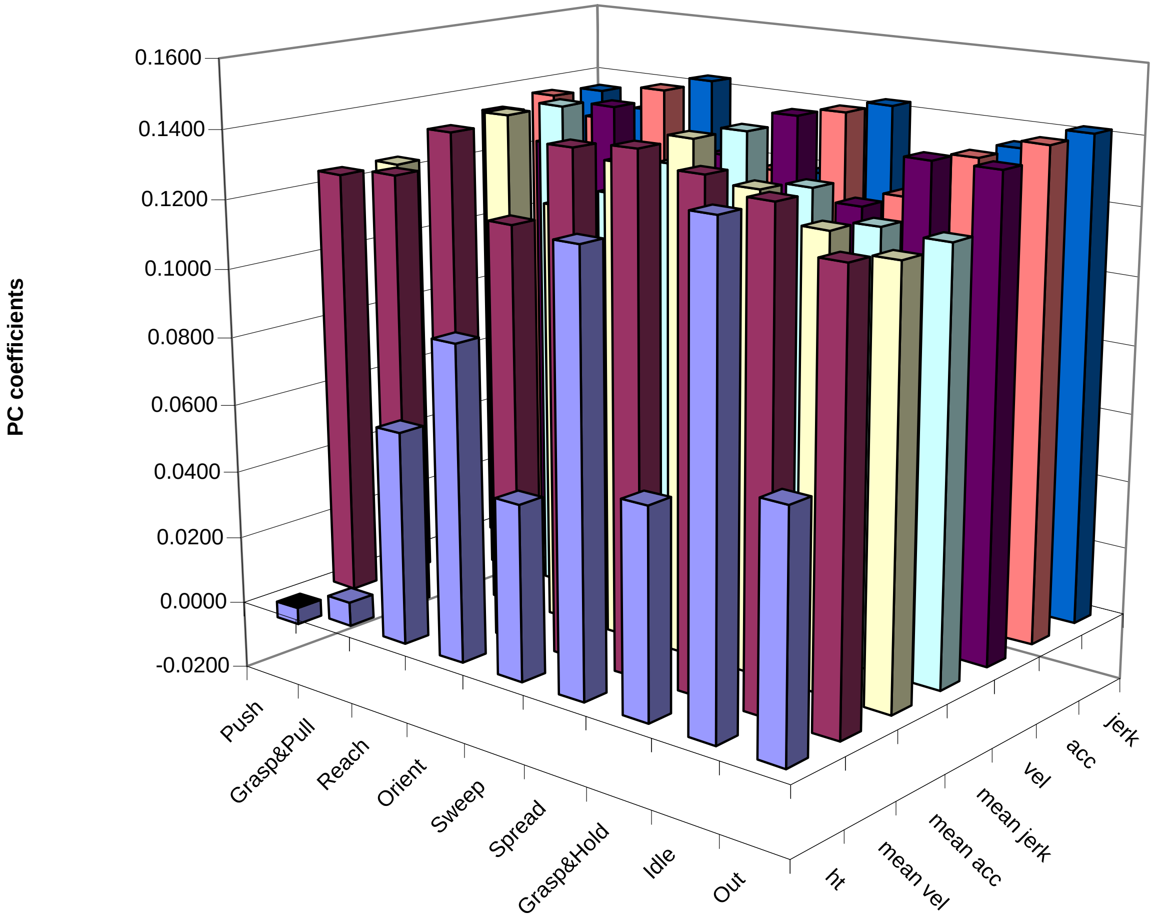


Figure 4.20: Coefficients for PC1

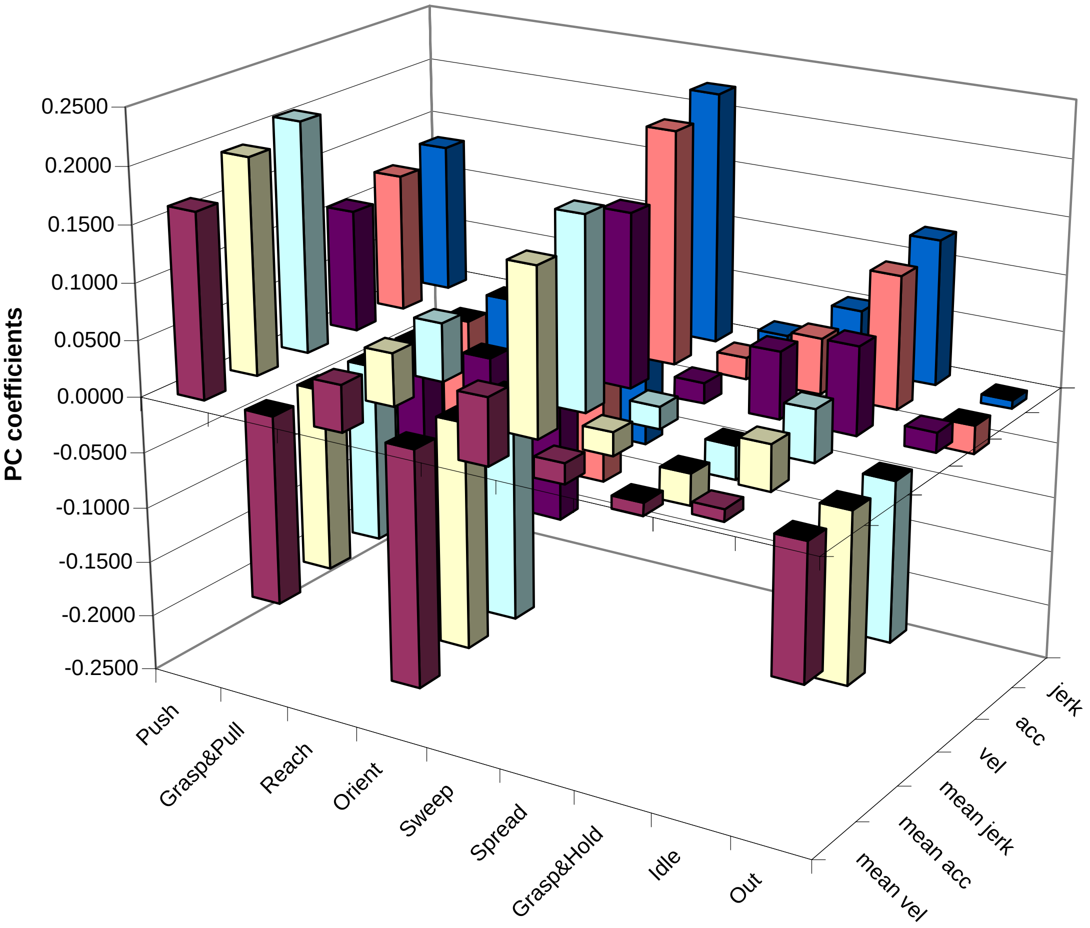


Figure 4.21: Coefficients for PC2

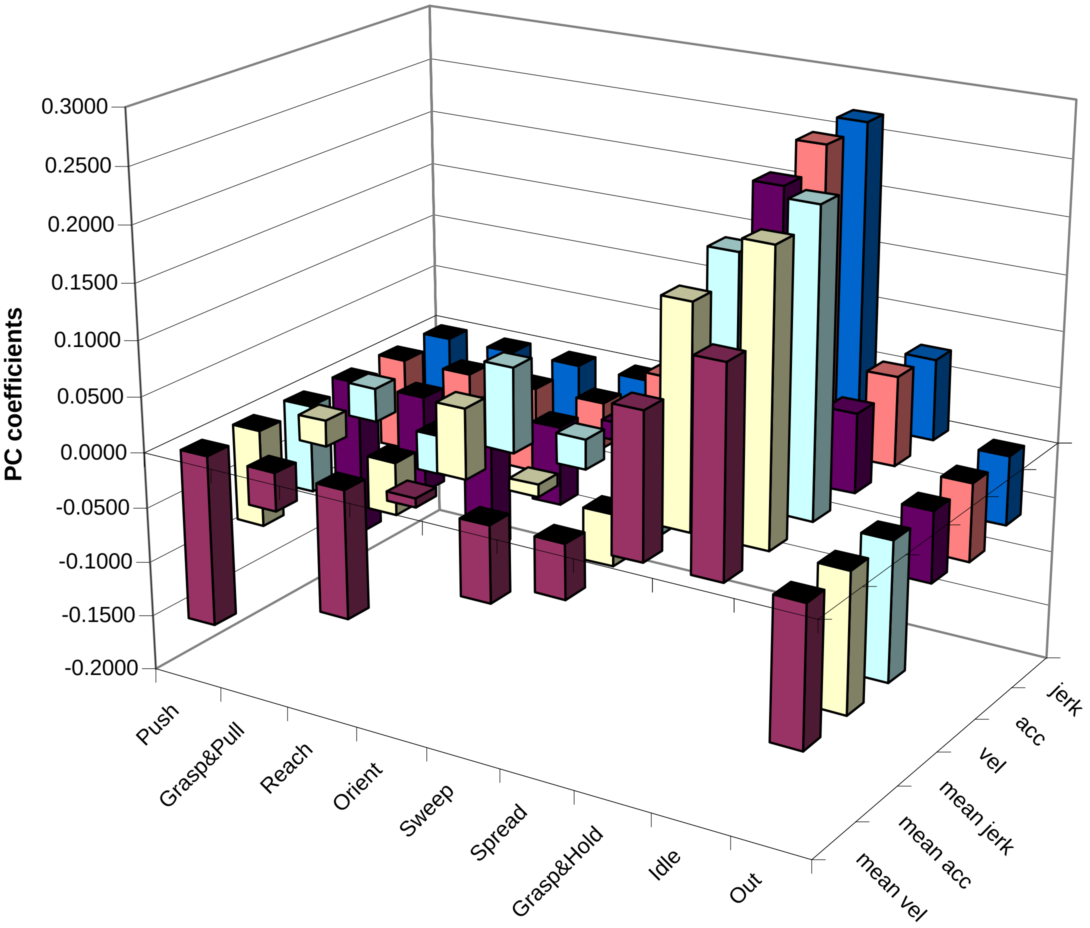


Figure 4.22: Coefficients for PC3

To test the hypothesis that subjects in different groups can be reliably distinguished form one another, we computed distances between the points on **Figure 4.19**: between each novice and expert pair and between subjects in the same group (**Figure 4.23**). A Mann-Whitney test (p=0.01) showed that the novice-to-novice and expert-to-expert comparisons (i.e, subject pairs from the same skill level) could be distinguished from the novice-to-expert comparisons.

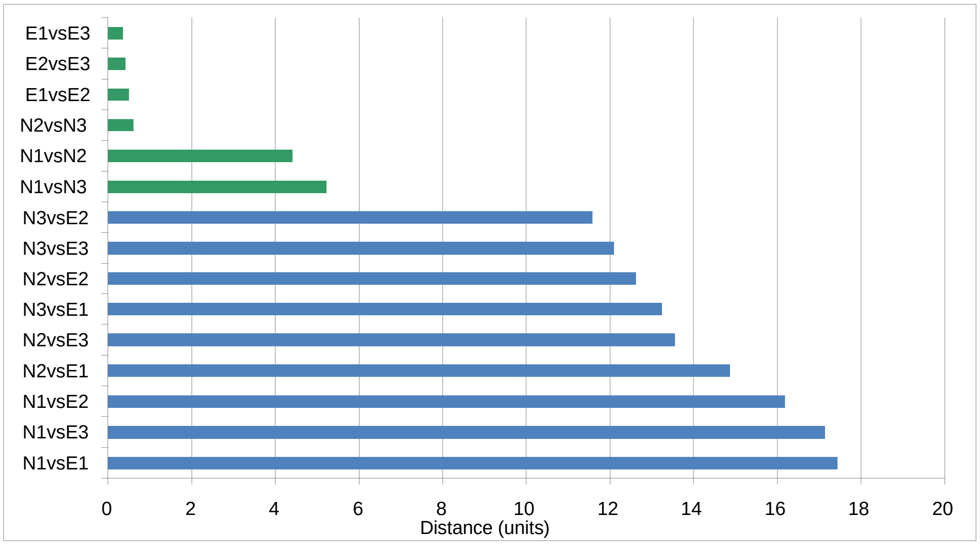


Figure 4.23: Distances from each novice and expert pair [□,●,◆,✶] to intra-experts' center (+) in the PCA space of **Figure 4.** (Mann-Whitney test, p=0.01)

4.3.3 Analysis of Stability in Principal Components

Since the previous PCA analysis revealed the importance of kinematics as a reliable performance measure, we performed a variability test in order to investigate how sensitive our conclusions are to random variations in our experimental data; therefore, we checked to see if the eigenvectors we obtained from the PCA analysis were reasonably stable as more repetitions of experiments are simulated. Using the methodology explained in Section 4.2.3.3, we made the following observations:

a) **Figures 4.24 and 4.25**: The graphical representation of A_k shows that as the number of dimensions (k) increases, the E_k subspace becomes progressively fully stable (A_k converges to k for k = 3 at the subtask level and k = 14 at the action level).

b) **Figures 4.26 and 4.27**: The plots of the stability index 'S_index' shows that a stability above 80% is achieved at k=2 and k=3 for the subtask and action levels respectively, which corresponded well to the values estimated using our PC selection method analysis.

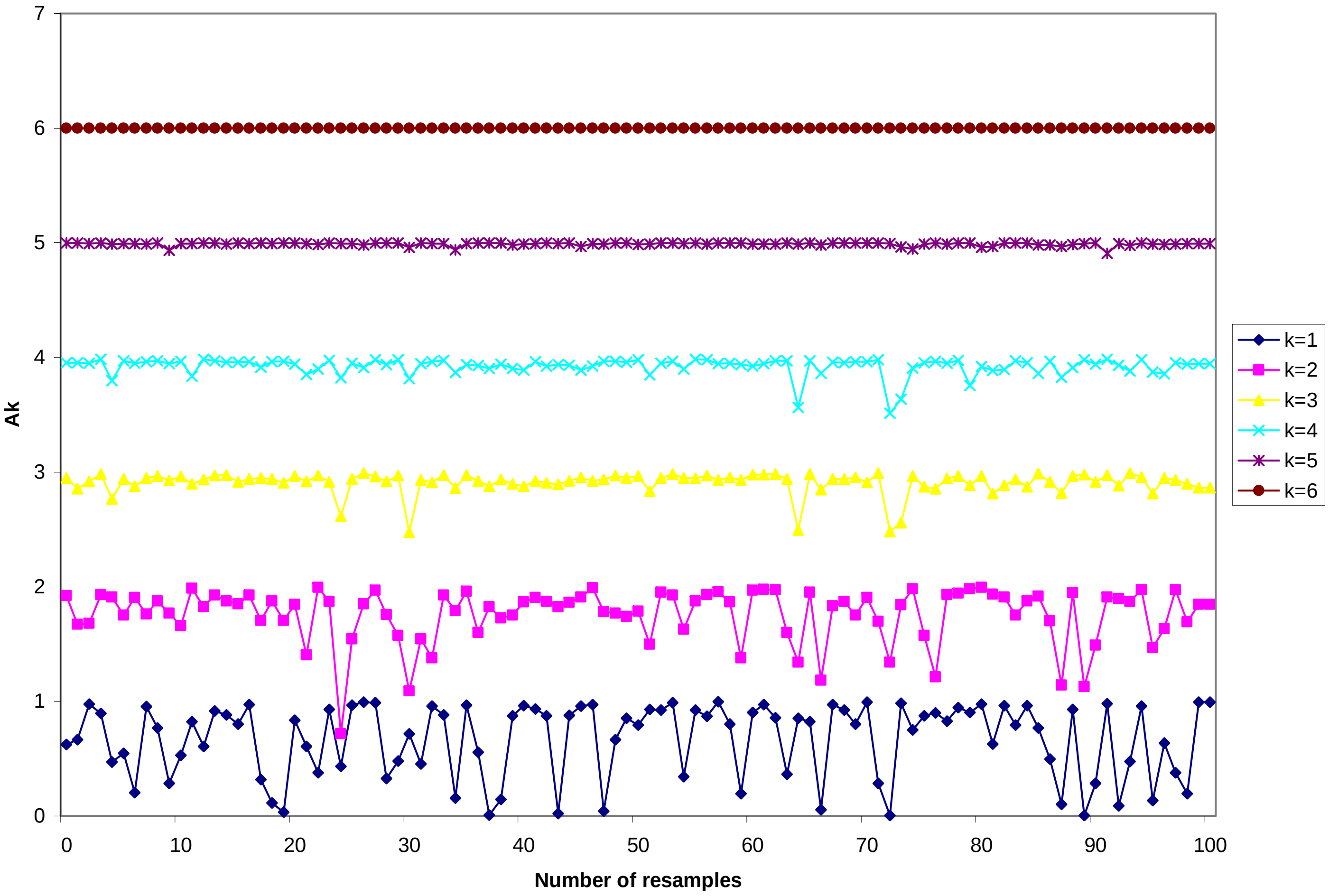


Figure 4.24: Variability of subspace E_k around $E_k^{(P)}$ (composed of the k first eigenvectors) – subtask level

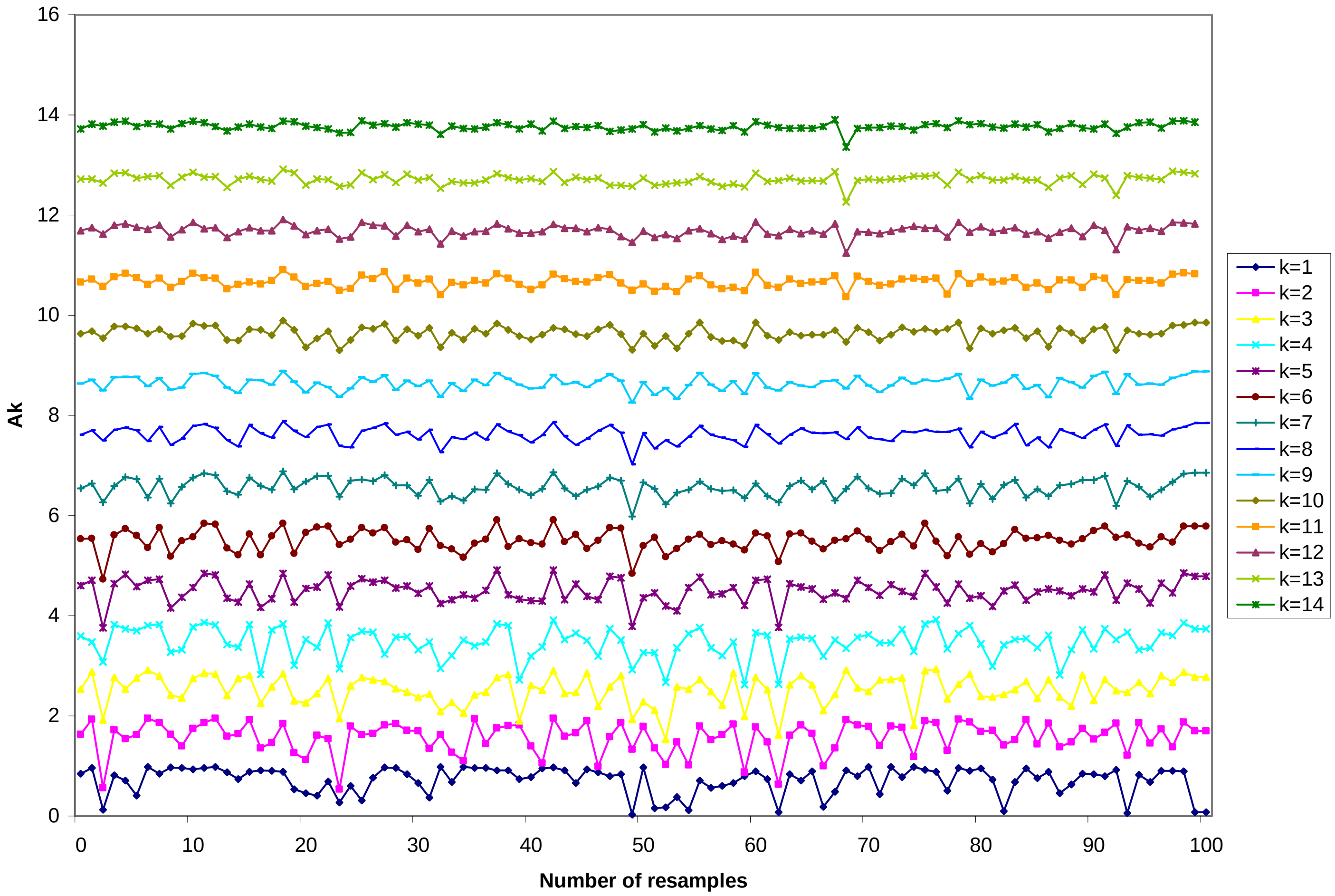


Figure 4.25: Variability of subspace E_k around $E_k^{(P)}$ (composed of the k first eigenvectors) – action level (plotted up to 14 dimensions; the subspaces are roughly fully stable from k = 14)

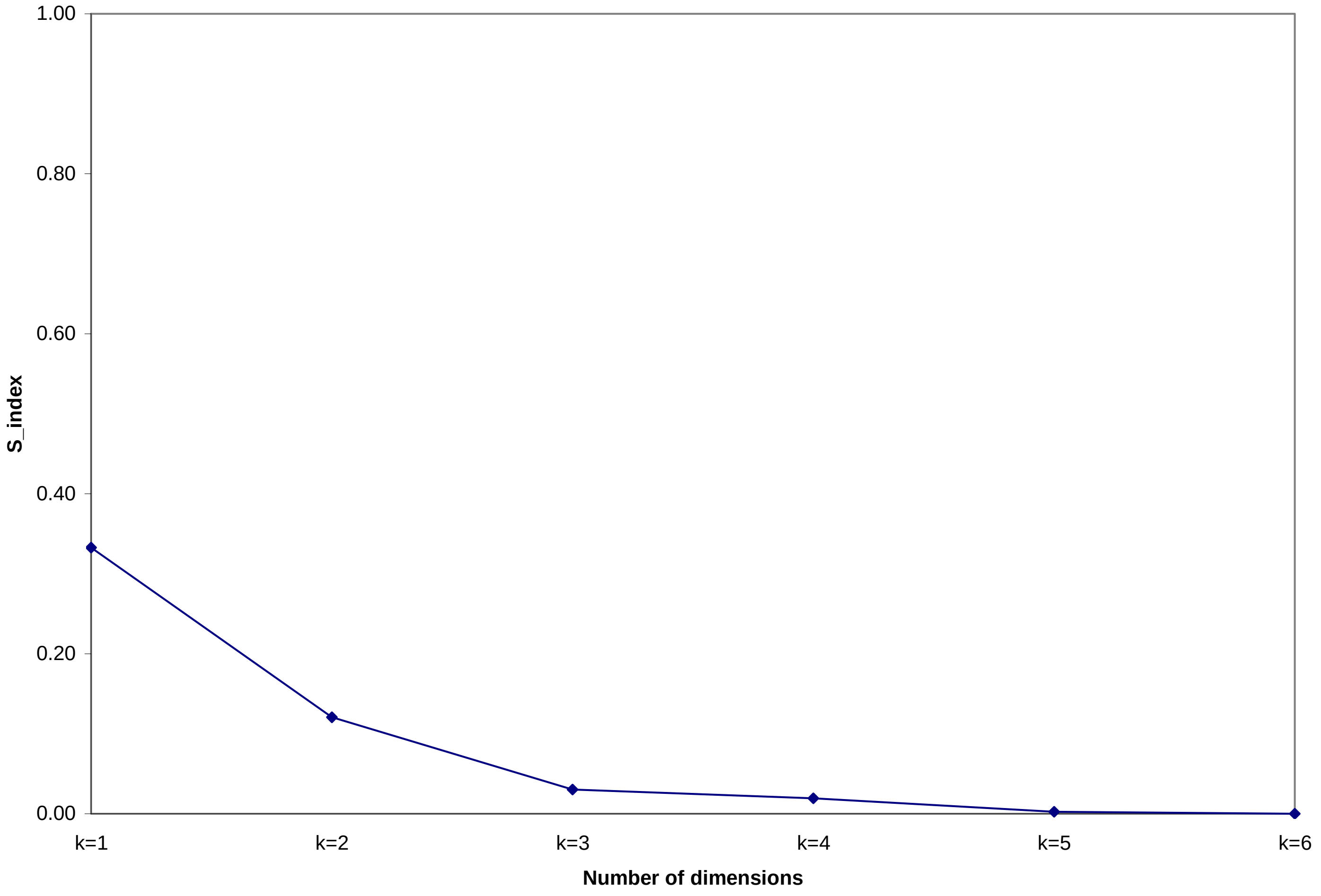


Figure 4.26: Stability indices when increasing subspace dimension (k: # of eigenvectors) – subtask level

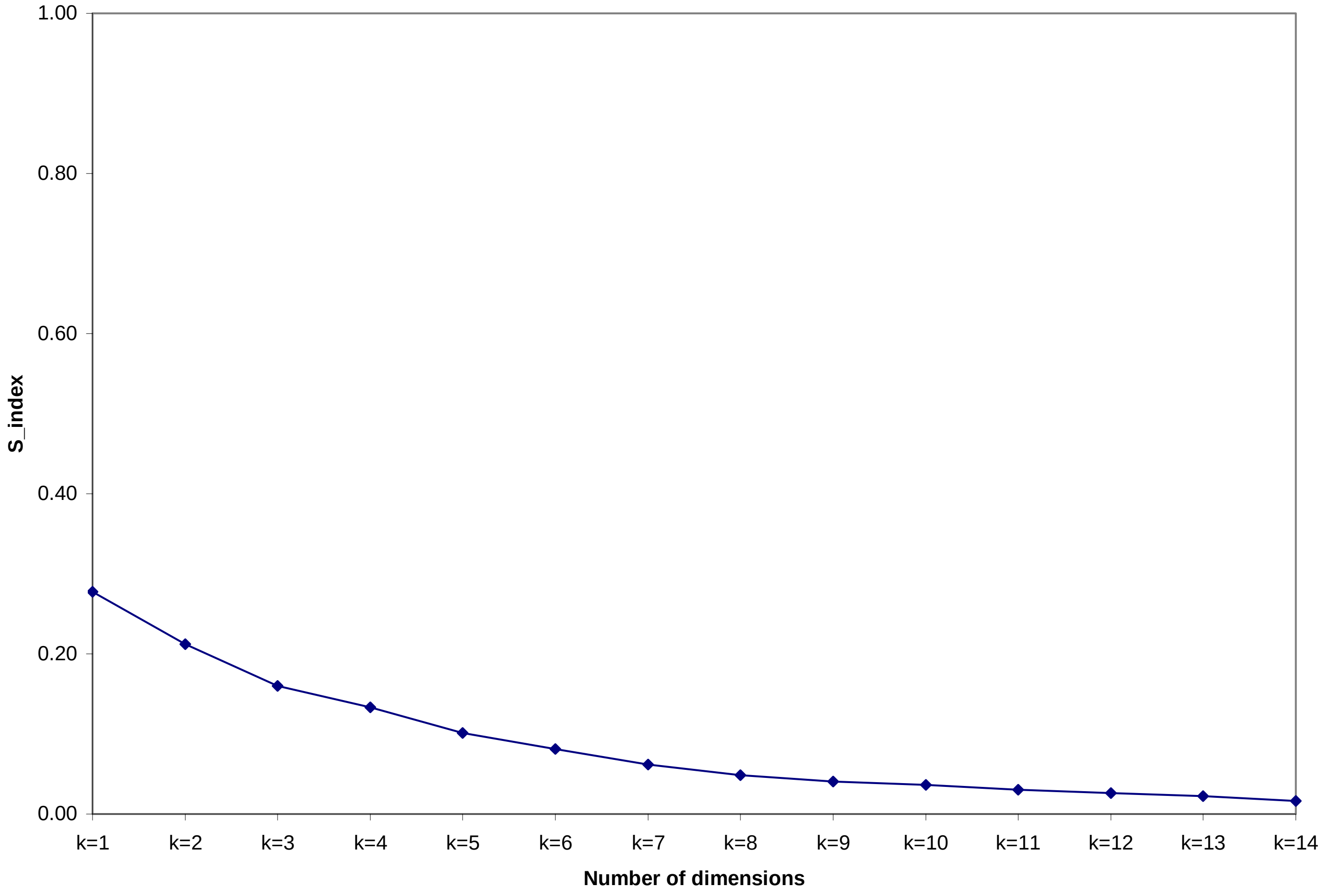


Figure 4.27: Stability indices when increasing subspace dimension (k: # of eigenvectors) – action level (plotted up to 14 dimensions; the subspaces are roughly fully stable from k = 14)

4.4 Summary

In chapter 3, we outlined an analytical approach to deal with surgical motor skill assessment which integrated three types of physical measures (kinematics, time and movement transitions) into a modelling framework for quantifying performance of surgical trainees. The primary purpose of the surgical simulated experiment described in the present chapter was to test whether or not our proposed analytical method was able to reliably distinguish between three groups of subjects representing different stages of training: novices, mid-stage trainees, and experts, and to demonstrate the reliability of our measurement equipment and our selected performance measures.

We simulated a surgical dissection task by asking participants to use laparoscopic tools to peel and separate the segments of two to three mandarin oranges placed in a training box.

The movements of the laparoscopic tool for the dominant hand were tracked using a magnetic sensor, which continuously recorded 3D position and orientation data at 120 Hz while the task was being executed. This position tracking system was selected to overcome the 'line-of-sight' issue in spite of its limited accuracy when compared with the optoelectronic system. In addition, the execution of the task was recorded on videotape and time records were obtained so to correlate the movement patterns with discrete phases of the task execution. Afterwards, kinematics data was derived by differentiation using a generalized cross validation algorithm.

We decomposed the dissection task into 4 subtasks: 'explore', 'peel skin', 'detach segment', and 'place segment' and concentrated on analyzing performance for 'peel skin' and 'detach segment' as the most descriptive subtasks when considering physical performance measures.

At the subtask level, we applied Principal Component Analysis (PCA) over multi-element vectors consisting of the average execution times and tooltip average velocities in each of the three cardinal directions (l:lateral, a:axial, v:vertical) for each of the two subtasks. In this weighted space, each subject's execution was represented as a data point and groups of points corresponded to repetitions from a single subject or from subjects belonging to the same skill level. Therefore, PCA representation in this application allowed for grouping subjects according to the technical proficiency levels perceived by our measuring system and therefore, we used the concept of distance to measure group membership (**Figure 4.3**).

Examination of the PC coefficients indicated that times did not provide much information to the analysis since their contributions were considerably lower than those provided by velocities. This suggested that kinematics perform better than time in differentiating subjects' performances at the subtask level. Additionally, we found that while the first PC separates skill levels, the second PC has to do with representing differences among the experts group (**Figure 4.8**).

We computed the ratio of mean square distance (MSD) in the PCA weight space from the mean position of all trials executed by a specific subject or group to the MSD from the global mean position to describe variability for specific subjects and groups. The low values of intrasubject (7%) and intragroup (24%) variability supported the qualitative observations that the greatest contributor to overall variability was difference in degree of training. Moreover, we found that the intersubject variability for each group (Experts: 22%; Novices with training: 31%; Novices: 5%) seems larger after training, which might suggest that training could enable operators to try more flexible strategies. The distances between different groups were considerably larger than the distances within groups, which is consistent with the idea that training might change motor patterns.

At the action level, we also compared performances of single subtasks by decomposing them into their characteristic actions as a set of 10 elemental tool tip motions: push, pull, reach, orient, sweep, spread, grasp&hold, grasp&cut, idle, out. Using video analysis and the previously defined start and end points, we derived a list of action transitions with the corresponding time spent at a specific action before transitioning to another (i.e, holding times). The time records were then used to segment the kinematics signal of the subtask into the kinematics data for each action. We then characterized every action using distributions of holding times and kinematics and used Kolgomorov-Smirnov statistic (D) to measure differences from these two parameters. Additionally, we computed transition probability matrices and used Jensen-Shanon divergence (JSD) to define a third difference measure. All JSD and D values for every action were grouped into a 64-element matrix and PCA was then applied to test the hypothesis that in this weight space our difference measures were able to provide skill level separation when comparing subjects from the same group (e.g., experts vs. experts) and subjects from different groups (e.g., novices vs. experts).

The distribution of data points on the PCA space showed that differences in intragroup (pairs of subjects from the same skill level) and intergroup (pairs of subjects from different skill levels) comparisons are primarily differentiated along the horizontal axis and a distance measure quantitatively demonstrated that novice-to-novice and expert-to-expert comparisons (i.e, subject pairs from the same skill level) were located close together in one group while novice-to-expert comparisons belonged to a different group (**Figure 4.19**). Analysis of the PC coefficients indicated that out of the three performance measures considered (time, kinematics, transitions), difference measures derived from kinematics provide most of the differentiation between skill groups at the action level as well.

In addition, a study on the variability of our principal components as the number of experiments increase, demonstrated that our PCA analysis is generally stable for the number of eigenvectors we are retaining at both subtask and action levels.

Although current simulators have been shown to be a valid tool for training novice surgeons in basic psychomotor skills [Park 2002, Grantcharov 2001, Ahlberg 2002, Hyltander 2002] as performed and assessed in a simulator, their ability to provide valuable guidance at more advanced levels of training has not been established. We believe that by using the MCMD we have developed to describe laparoscopic surgical procedures into a standardized framework, it will be possible to focus analytical attention on specific surgical tasks and therefore potentially to establish stronger correspondences between selected surgical tasks and the corresponding simulations of these tasks, which may ultimately enable us to do validated assessments in a simulated setting.

The most common performance measures used in simulators to discriminate senior from junior surgeons include time, number of movements, number of errors, path length, distance and travelled by the instrument [Datta 2006, Ahlberg 2002, Torkington 2001, Smith 2002, Macmillan 1999; Francis 2001]. Tool-tissue interaction forces and tool kinematics have also been explored respectively by Rosen's group at the University of Washington and our group, but they are not yet in widespread use due to the complexities involved in modifying surgical instruments to accept the necessary sensors [Rosen 2006, Rosen 2002, Rosen 2001, Kinnaird 2004, McBeth 2002]. For the present study we implemented three measures of performance: time, pattern of movement transitions and tool kinematics, which were attached to the nodes of our MCMD in order to provide data representation for isolated tasks so as to identify specific sources of performance differences between subjects. Computation and analysis of intuitive difference measures (JSD and D values) have clearly shown that across subjects, tool kinematics data show detectable differences among skill levels. In addition, we have introduced a principal component analysis (PCA), not yet explored by other surgical performance assessment studies, which has allowed us to perform simultaneous analysis of multiple measures by reducing the dimensionality of the data. It has also proved to be useful and practical in determining intrasubject, intragroup, and intergroup variabilities.

On the basis of the presented results which clearly show differentiation between skill levels, we decided to move into an intraoperative pilot study. In Chapter 5 we will describe the corresponding implementation and results for assessing performance between expert surgeons and residents during laparoscopic cholecystectomy procedures.

Chapter 5

Intraoperative Study

5.1 Introduction

In the previous chapter we presented the results of applying our assessment methodology in a physical simulation. By measuring the kinematics of the surgical tool motion, time and tool movement transitions, we found the kinds of intrasubject consistency in performance, intersubject similarity and intergroup differences that we would need to in order to use this technique clinically. This justifies moving on to a clinical trial in order to investigate whether intra- and inter-subject variabilities in the operating room (OR) setting follow similar patterns.

In contrast to the controlled simulator environment, assessing skills in the OR adds more variables such as variations in the patient's anatomic features, and the experience level of the surgical team, or unanticipated equipment problems, which to a greater or lesser extent might influence surgeons' performance and therefore affect the reliability of an intraoperative assessment system.

Standard and widely accepted skill evaluation methods for use in the operating room include direct observation assessment and checklists, both time consuming and subject to bias. Objective systems for this environment are under development and to our knowledge Dr. Ara Darzi's group at the Imperial College and our group are the only two approaches currently using data from the operating room for testing assessment methodologies [Aggarwal 2007].

As described in the previous chapter, we have found that our methodology can differentiate between different skill or experience levels in a physical simulator [Cristancho 2007]. The purpose of the study reported in this chapter is to see whether it can differentiate between skill levels in a live OR setting. To that end, we acquired intraoperative data from two sets of subjects representing the extreme two stages of training: attending surgeons with extensive laparoscopic experience and residents just learning laparoscopic cholecystectomies.

Our primary hypotheses are similar to these tested in the physical simulator study:

(1) the intrasubject variability of extracted measures of tool use patterns will be less than the intragroup and intergroup variations in these measures, and

(2) the extracted patterns of the surgical tool movements will enable us to distinguish between trainees at different stages of their training when performing in the intraoperative setting.

In this study, we also include measurements from both the dominant and the non-dominant hand in order to establish if dexterity in using the non-dominant hand helps differentiates between groups. We therefore used two techniques for representing bimanual coordination: a measure of dependence and a measure of differentiation between the velocity distributions of both hands when executing individual subtasks.

The results of this study will hopefully provide us with a foundation to conduct a larger intraoperative study in the future to test if the proposed methodology may be practically and usefully incorporated into the evaluation process of the surgical residency curriculum.

5.2 Protocol

To determine if our quantitative analysis technique provides repeatable results in the operating room that can be used to monitor development of surgical motor skills during training, we assessed several surgeons in the early stages of their training and several attending surgeons because these two groups represent the widest possible separation of skill levels that we can observe in the OR performing whole laparoscopic cholecystectomy procedures. To evaluate the repeatability of our technique, we assessed each surgeon on three different occasions.

5.2.1 Participants

We recruited two sets of subjects to represent different stages of training: residents (represented by three 4th year surgical residents, i.e., mid-stage of training, but they are beginning to perform laparoscopic cholecystectomies), and experts (represented by three attending surgeons).

Both the Vancouver Coastal Health Authority and the UBC Research Ethics Board granted ethics approval to this study (Appendix E). Residents were protected from coercion by ensuring that their clinical supervisors were not involved in the participation request process and so would not know which particular residents are invited to participate in the study, or what the reply of any individual resident was. We were not able to protect against supervisors knowing the identity of residents who accepted the invitation to participate because their surgeries would be instrumented and monitored and the presence of the equipment was obvious. Residents were, however, protected from potential judgment because clinical supervisors were not be shown data with particular residents' identifying information attached.

5.2.2 Experimental Setup

We observed surgeons in the operating room performing 3 Laparoscopic Cholecystectomies (LC) per subject using standard surgical tools. LCs were chosen because it is one of the earliest procedures that a resident is introduced to the beginning of their training, it is the most commonly performed laparoscopic procedure [Tendick 2000], and it has become an 'index' operation for ongoing assessment of laparoscopic skills.

Using the position measurement system and techniques described in Section 3.2.2.2, we extracted kinematic measures which characterize the movements made by the surgeon (eg, velocities, accelerations, and jerks). In addition, we recorded the video from the laparoscopic camera, which provides a view from inside the body, and we used this video to manually segment the surgical tasks, subtasks and actions as described in Chapter 2. **Figure 5.1** shows an overview of our equipment setup in the operating room.

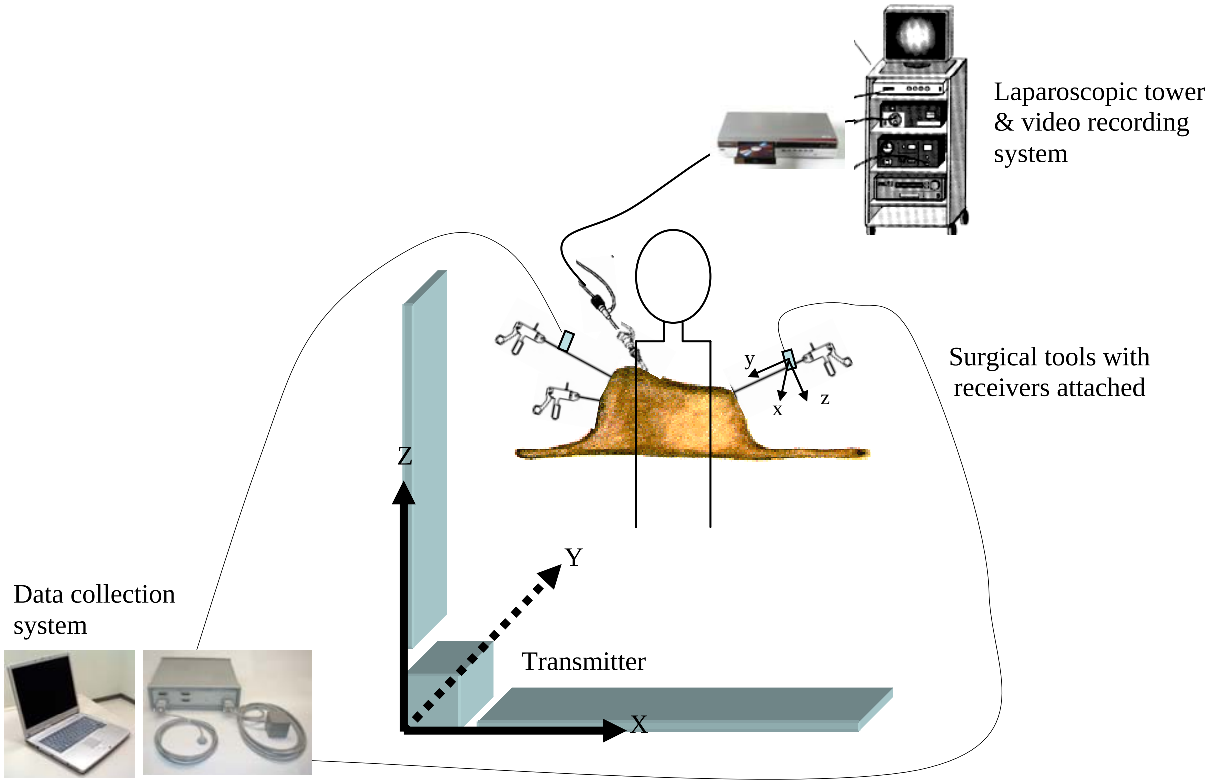

Laparoscopic tower & video recording system
Surgical tools with receivers attached
y
x
z
Z
Y
X
Transmitter
Data collection system

Figure 5.1: Overview of the operating room setup

All equipment used was approved by the Biomedical Engineering Department at UBC Hospital, and sterilized using ETO (Ethylene Oxide), where appropriate (**Figure 5.2**).

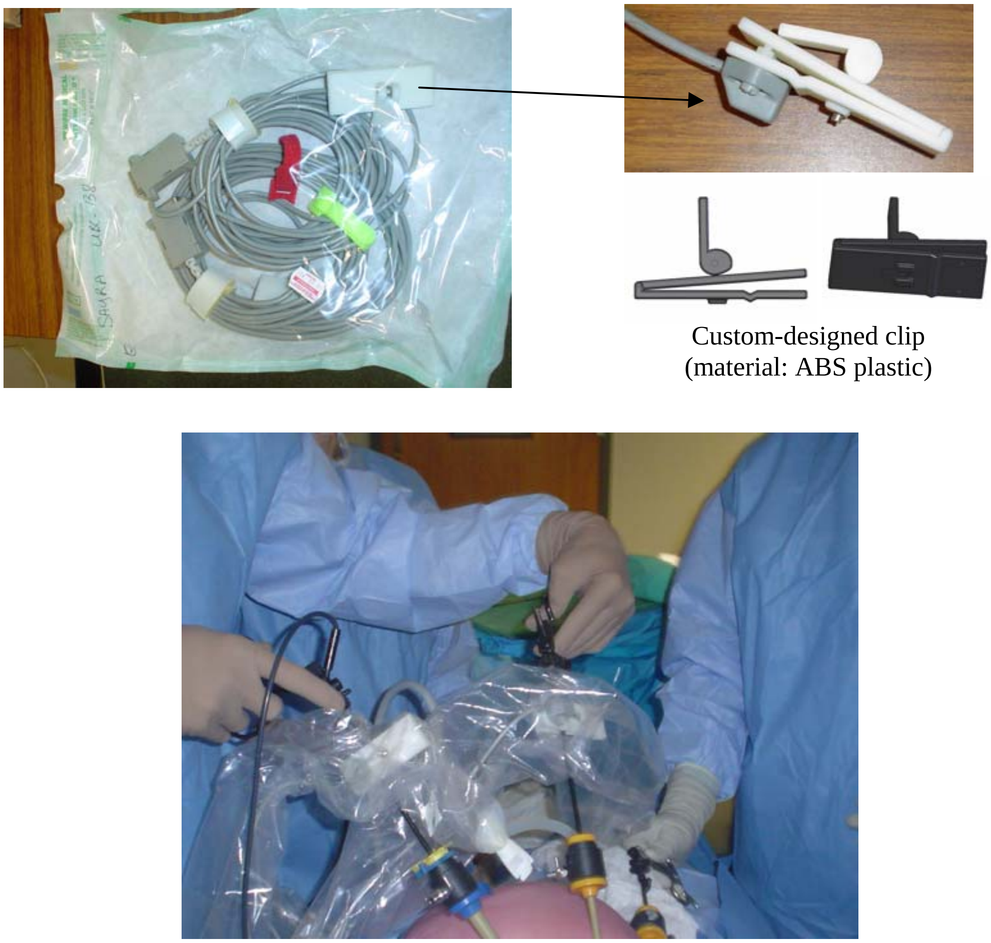

Custom-designed clip
(material: ABS plastic)

Figure 5.2: Sterilized sensors as delivered by SPD department and when used in actual procedures

5.2.3 Analytical Methods

Following the MCMD surgical description provided in Chapter 2, three hierarchical levels were defined for the OR study. For this study, we focused on a single task ('Isolate Cystic Duct / Cystic Artery', **Figure 5.3**) because the two key subtasks, 'Expose Triangle' and 'Dissect CD' were identified by the expert surgeons as the most demanding steps of the procedure in terms of the surgical dexterity required, we therefore further focused our analysis on these two subtasks.

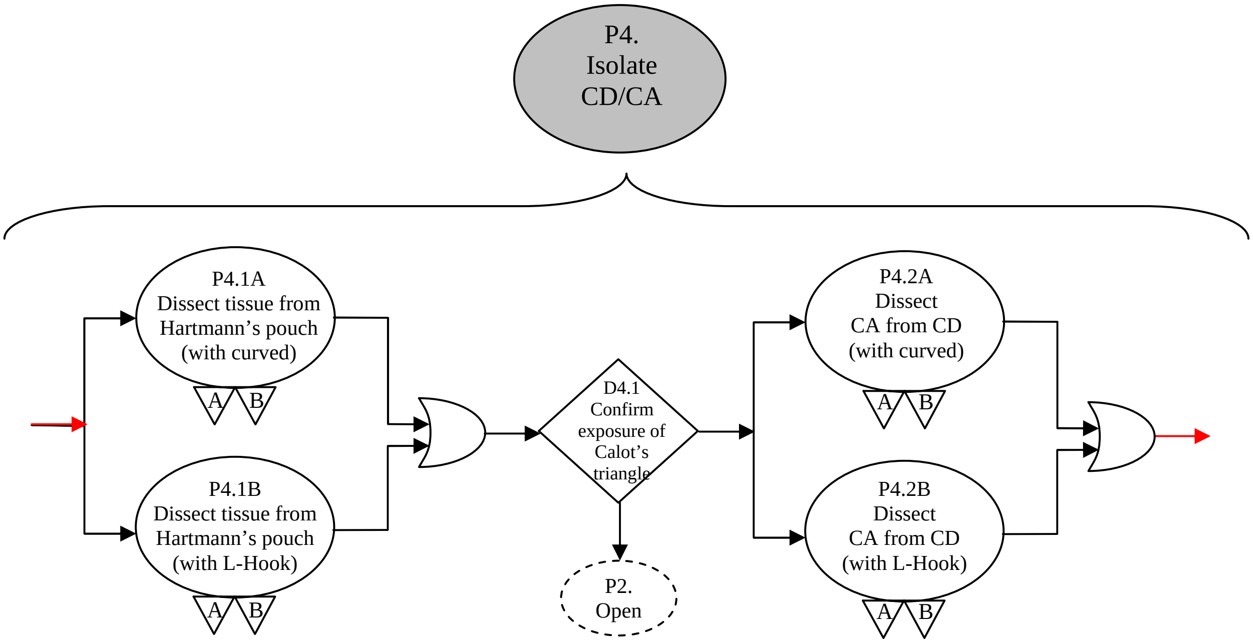


Figure 5.3: MCMD decomposition for the surgical task 'Isolate CD/CA'

As with the mandarin experiment described in the previous chapter, we performed two separate analyses: (1) Subtask level: examination of the 3D average kinematics of the surgical tool movement when executing the two main surgical subtasks (Expose Triangle and Dissect CD/CA) to identify broad differences in execution; and (2) Action level: decomposition and analysis of individual subtasks into our pre-defined set of ten actions (e.g., push, sweep, spread, etc) using our assessment methodology (Chapter 3) to localize any sources of differences in motor performance. We follow the flow of steps summarized in figure 3.26 from Chapter 3 in presenting the corresponding methods and results for this OR study.

5.2.3.1 Subtask Level

The position data from the two main subtasks 'Expose Triangle' and 'Dissect CD/CA' was separated from the whole procedure data stream by manually identifying the start and end points using video analysis. 'Expose Triangle' consists of retracting the gallbladder and dissecting some of the surrounding tissue so as to open the cystic pedicle space and to identify where the cystic duct and the cystic artery lie. It begins when the gallbladder is first stretched out and ends when the cystic pedicle is identified. 'Dissect CD/CA' consists of identifying and isolating the cystic duct from the cystic artery by dissecting the surrounding tissue. It begins when the tip of the tool is first inserted between the two anatomic structures and ends when both structures have been completely freed from each other. The video clips were collected for further decomposition at the action level and the timing records were used to segment the data streams. Afterwards, the 3D kinematics of the tool motion during each segmented subtask were derived by differentiation using a generalized cross validation algorithm (GCV) (Section 3.2.2.2).

For each subject 'i' and procedure 'j' (S_{ij}), we first defined a 3-element (for each individual subtask) or 6-element (when analyzing both subtasks simultaneously) vector consisting of the tooltip velocities in each of the three cardinal directions (i.e., l:lateral, a:axial, v:vertical) for each of the two subtasks ('Expose Triangle' and 'Dissect CD/CA'). All subjects were then be grouped into a nxm matrix (n: $\sum i \sum j$, total number of procedures; m: number of dimensions, m=3 or m=6). All subtask repetitions during a single procedure are concatenated into single vectors and each row entry in the **Table 5.1** corresponds to averages of the consolidated data.

Subject / trial	Expose Triangle			Dissect CD/CA		
S11	$V_{11}l_B$	$V_{11}a_B$	$V_{11}v_B$	$V_{11}l_C$	$V_{11}a_C$	$V_{11}v_C$
S21	$V_{21}l_B$	$V_{21}a_B$	$V_{21}v_B$	$V_{21}l_C$	$V_{21}a_C$	$V_{21}v_C$
…	…	…	…	…	…	…
Sij	$V_{ij}l_B$	$V_{ij}a_B$	$V_{ij}v_B$	$V_{ij}l_C$	$V_{ij}a_C$	$V_{ij}v_C$

Table 5.1: Velocity data arranged for PCA analysis at the subtask level

In a second version of the analysis, we also included average time execution for each subtask into the rows of table 5.1, thereby increasing the number of columns to m=4 for individual subtasks or m=8 for simultaneous analysis of both subtasks. This was done to assess if duration of subtask execution contained any diagnostically-useful information.

Due to the large amount of data available in the data stream, we decided to additionally perform a more detailed look at the data in order to determine if adding more information to the analysis other than averages would provide more discriminatory power for differentiating between skill groups on either one or both surgical subtasks. We therefore, obtained samples from the velocity profiles at every 5th percentile over the middle 50% of the range for each movement direction and each individual subtask. For example, the data table for a **detailed analysis** (25th to 75th percentiles) of the **dominant hand** movement in the **axial direction** while executing **'Expose Triangle'** became as is shown in **Table 5.2**.

Subject / trial	25th	30th	35th	…	…	75th
S11	$V_{11}a_{25th}$	$V_{11}a_{30th}$	$V_{11}a_{35th}$			$V_{11}a_{75th}$
S21	$V_{11}a_{25th}$	$V_{21}a_{30th}$	$V_{11}a_{35th}$			$V_{11}a_{75th}$
…	…	…	…	…	…	…
Sij	$V_{11}a_{25th}$	$V_{ij}a_{30th}$	$V_{11}a_{35th}$			$V_{11}a_{75th}$

Table 5.2: Data arranged for percentile analysis of velocity profiles of the dominant hand movement in the axial direction while Exposing Calot's Triangle

We restricted our detailed analysis to values located from the 25th percentile to the 75th percentile as it would correspond to the median area of the data set and we expect that most of the measures that would represent a surgeon's performance will fall within this range. We looked at the percentile full range (from 5th to 100th) and found out that it did not show significant differences in the variability analysis with respect to the 25th to 75th range[1].

In order to facilitate the presentation of results, we separated our analysis into two categories as presented below.

Analysis of AVERAGE data:

(a) Average tool tip velocities for "Expose triangle"

(b) Average tool tip velocities for "Dissect CD/CA"

(c) Average tool tip velocities and time for "Expose triangle"

(d) Average tool tip velocities and time for "Dissect CD/CA"

[1] The fractions of variability (intrasubject, intragroup, and intergroup) were essentially independent of the range of data used and there was very little variation in the locations of the points along the first PC direction (Appendix H).

Analysis of DETAILED data:

(a) 25th to 75th percentiles, at 5 percentile increments, of velocity profiles for each movement direction (lateral, axial, vertical) analyzed separately during 'Expose Triangle'

(b) 25th to 75th percentiles, at 5 percentile increments, of velocity profiles for each movement direction (lateral, axial, vertical) analyzed separately during 'Dissect CD/CA'

(c) 25th to 75th percentiles, at 5 percentile increments, of velocity profiles for all movement directions analyzed altogether (Expose triangle)

(d) 25th to 75th percentiles, at 5 percentile increments, of velocity profiles for all movement directions analyzed altogether (Dissect CD/CA)

Once the data from each category is arranged into the corresponding matrix form (see **Table 5.1** and **Table 5.2**), we normalized it by dividing each element by the column standard deviation (defined across all procedures and subjects), and used Principal Components Analysis (PCA, as described in Section 3.4.4) to extract the dominant contributors to overall variability to simplify the presentation of the data to the trainer. We applied our dimensionality-decision criterion to select the appropriate number of dimensions to retain in the new space provided by PCA (Section 3.4.4.2).

To test the hypothesis that level of skill development is apparent in intraoperatively acquired quantitative measurements, we computed the contributions to variability in the PCA weight space (defined as the mean squared distance of points in the weight space relative to the global mean position across all subjects and groups) due to intrasubject, intragroup and intergroup variability. We evaluated the null hypothesis that residents and experts all have the same means using a nested ANOVA test ($F_{critical}$ = 7.71, α = 0.05, intragroup DoF = 4, intergroup DoF = 1) (Appendix G).

As this was a pilot study, we were not yet sure how much variability to expect in surgical performance between groups. Our preliminary study indicated very good intra-subject repeatability and reasonable intragroup consistency, but it was initially unknown whether these results would hold for the live operating room situation.

Since in the OR study we monitored both dominant and non-dominant hands, we also investigated bimanual coordination by defining two types of measures based on the kinematics profiles for each subtask. We first tested the hypothesis that the velocity distributions of the two hands are independent of one another (or, conversely, the extent to which joint interactions are needed to explain the overall distribution). A normalized variant of the Mutual Information between two distributions (Section 3.3.1.1.2) was computed for 'Expose Triangle' and 'Dissect CD/CA' separately for each of the 18 recorded procedures.

A second measure for bimanual coordination was defined to test the assumption that lack of experience would result in the subject concentrating on single hand movements at a time (i.e., faster movements with the dominant hand are accompanied by slower movements of the non-dominant hand). We therefore used the Kolgomorov-Smirnov statistic to compute differences between the speed distributions of both hands derived from execution of each subtask and reported the corresponding D values. We used

Mann-Whitney tests to evaluate if the distributions of coordination difference measures differ between groups.

5.2.3.2 Action Level

Following the same methodology as in the physical simulator, we compared performances of single subtasks by decomposing them into their characteristic actions as a set of 10 elemental tool tip motions: push, pull, reach, orient, sweep, spread, grasp&hold, grasp&cut, idle, out. The process of segmenting and obtaining the times records for these actions was achieved by identifying the start and end points through video analysis according to the action definitions provided in Chapter 2. For each subtask repetition during each procedure (S_{ij}), Excel templates were used to register this information temporally and to derive the list of action transitions with the corresponding time spent at a specific action before transitioning to another (i.e, holding times). **Table 5.3** presents an example of how data is processed in Excel to obtain the list of transitions and holding times for one subtask execution.

	TIME IN			**TIME OUT**				
ACTION	**min**	**sec**	**30ths**	**min**	**sec**	**30ths**	**Transition**	**HT (sec)**
Reach	0	0	0	0	4	20	**Reach-Push**	4.67
Push	0	4	20	0	5	0	**Push-Reach**	0.33
Reach	0	5	0	0	6	22	**Reach-Out**	1.73
Out	0	6	22	0	8	26	**Out-Idle**	2.13
Idle	0	8	26	0	9	19	**Idle-Out**	0.77
Out	0	9	19	0	10	1	**Out-Reach**	0.40
Reach	0	10	1	0	10	22	**Reach-Out**	0.70
…								

Table 5.3: Decomposition of a particular subtask into its corresponding set of executed actions. Start and end times for each action were registered and every action transition was obtained.

Using the list of time in and time out values, we segmented the kinematics signal of the subtask and were able to separate the kinematics data for each action. At this level, we characterized every action using distributions of holding times and kinematics, and we computed a matrix describing the transition patterns.

We computed direct differences between two subjects as it is easier to compare distributions and transition matrices this way (section 3.3.2). For comparing performances between two subjects M_{r1} and M_{e1} (resident #1 vs. expert #1), we defined: (a) $M_{r1}.T_k.S_l.A_m v$[2] and $M_{e1}.T_k.S_l.A_m v$ as the group of kinematic (e.g., velocity) distributions for all executed actions; (b) $M_{r1}.T_k.S_l.A_m t$ and $M_{e1}.T_k.S_l.A_m t$ as the group of holding time distributions for all executed actions; and (c) TMP_{r1} and TMP_{e1} as the action transition probability matrices. We then computed difference measures for the various corresponding kinematics ($M_{r1}.T_k.S_l.A_m v$ vs. $M_{e1}.T_k.S_l.A_m v$) and holding time ($M_{r1}.T_k.S_l.A_m t$ vs. $M_{e1}.T_k.S_l.A_m t$) distributions using the Kolgomorov-Smirnov statistic (D measure); and difference measures for the transition matrices using the Jensen-Shanon divergence (JSD measure) after modelling the system as a Semi-Markov process (Section 3.4.3).

Across all subject comparisons, we obtained a vxw matrix, where v corresponds to the number of subject comparisons (M_{r1} vs. M_{e1}) and w to the number of extracted performance measures.

[2] Represents vector of velocity measures for m action during execution of l subtask and k task

$$\begin{bmatrix} M_{r1} vs. M_{E1} \\ M_{r1} vs. M_{E2} \\ \vdots \\ M_{r3} vs. M_{E3} \end{bmatrix} = \left[\begin{array}{c||c||cc} JSD_{r1e1} & 9D_{r1e1}.T_k.S_l.A_m.t & 27D_{r1e1}.T_k.S_l.A_m.(\bar{v},\bar{a},\bar{j}) & 27D_{r1e1}.T_k.S_l.A_m.(v,a,j) \\ \vdots & & \vdots & \vdots \\ & & & \\ JSD_{r3e3} & 9D_{r3e3}.T_k.S_l.A_m.t & 27D_{r3e3}.T_k.S_l.A_m.(\bar{v},\bar{a},\bar{j}) & 27D_{r3e3}.T_k.S_l.A_m.(v,a,j) \end{array}\right]$$

For the OR experiment at the action level there were 15 subject-to-subject comparisons: n(n-1)/2 = 3 intragroup comparisons for residents, e(e-1)/2 = 3 intragroup comparisons for the attendings, and nxe = 9 intergroup comparisons between residents and attendings. We also computed 71 D measures for each row: 1 JSD value; 10 D values (per variable) for average time, and summary velocity, acceleration, and jerk profiles (40 total); 10 D values (per variable) for detailed velocity, acceleration, and jerk profiles (30 total). Given the large number of elements in each row, we again used Principal Components Analysis to reduce the dimensionality of the difference matrix and a Mann-Whitney test to test the null hypothesis that there is no distinction between intragroup (same skill level) and intergroup (different skill levels) comparisons.

We also used Semi-Markov models to represent the transitions between action states and the times spent in each of them. As discussed in section 3.4.2, we performed a distribution fitting test using 3 candidate parametric distributions: Exponential, Weibull, and Lognormal to determine which kind of distribution best modeled the holding time distribution of each state; the Weibull and the Lognormal proved to offer best fits than the Exponential.

In estimating the confidence intervals for the parameters of the selected parametric distributions, we used a bootstrapping approach by resampling with replacement 1000 times our original time vectors and calculating the corresponding distribution parameters for each set. The experimental set of parameters was sorted and the percentile quartile method with a 95% confidence level (α=0.05) was applied to obtain each parameter confidence interval (Equation 4.1, Appendix D.1 and D.2)

$$CI=[N(1-\alpha)^{th}, (N*\alpha)^{th}] \qquad \text{Eq. 4.1}$$

We then implemented the corresponding Semi-Markov models for each subject / each subtask (Appendix D.3) by additionally computing the movement transition probabilities and the corresponding confidence intervals. Transition probabilities (P_{ij}) were obtained from the ratio between the number of transitions from action 'i' to action 'j' by the total number of transitions from action 'i' to all other possible actions. To establish the confidence interval for the transition probabilities we used Equation 4.2 for α=0.05 [Walpole 2007]:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad \text{Eq. 4.2}$$

Where $\hat{p}$ is the estimated transition probability, and n is the number of transitions used to calculate the transition probability, and q=1-p.

We used the parameters of the holding time distributions and the movement transition probabilities with their corresponding confidence intervals for characterizing the Semi-Markov diagrams of every subject's subtask execution.

5.3 Results

In this section, we describe the results for implementing our proposed methodology to compare surgeon's performance in the operating room. Two variability analyses are reported in order to find out if there is separation of patterns across the training spectrum and to determine which data/measures are most useful in separating surgeons along this spectrum. Additionally, we present the implementation of a Semi-Markov modelling approach to describe motor behaviour at the action level and the results for an initial approximation to characterize and measure bimanual coordination in this particular surgical context.

5.3.1 Variability Analysis at the Subtask Level

Based on the kinematics profiles from our subtasks of interest: Expose Triangle and Dissect CD/CA, we applied Principal Component Analysis to study: (1) if one type of measure (i.e., average or detailed in the form of percentiles) is more capable of discriminating between groups than the other; and (2) if either one or both subtasks carry information about group differentiation.

Figure 5.4 shows samples of cumulative distributions of velocity in the axial direction for all procedures (left) and one representative procedure (right) per subject (R: residents; E: experts). It appears that there is high repeatability in individuals as all residents and experts' executions are roughly together, with the residents being apparently more consistent across procedures. As presented on the left side plots, it also seems that there is a more distinct separation of groups during the 'Expose Triangle' subtask than in the 'Dissect CD/CA' where various procedures from residents (reddish) and experts (greenish) appear to overlap more.

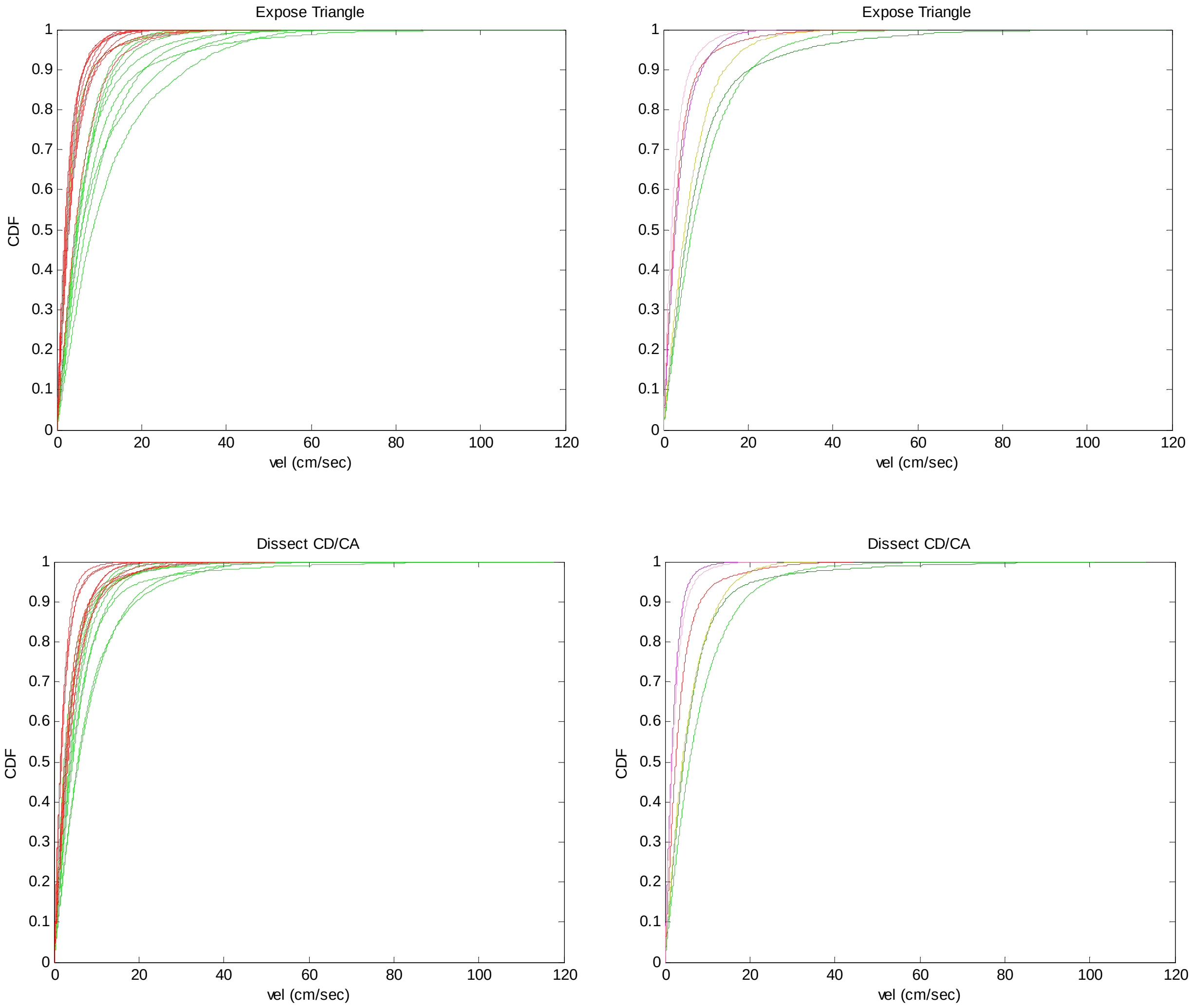


Figure 5.4: Samples of Cumulative Distribution Functions (CDF) for velocity in the axial direction for the two subtasks for all procedures (left) and one representative procedure (right) from each subject. Reddish colours correspond to residents and greenish colours correspond to experts

For the analysis based on measure averages, we computed the root-mean-square values of the velocity profiles and arranged various matrices to be used in the PCA analysis (**Table 5.4**). The total number of rows corresponds to the total number of recorded procedures (n=18 for this study) and the total number of columns corresponds to the 3 velocity components – lateral, axial, and vertical – during ‘Expose triangle’ or ‘Dissect

CD/CA' subtasks. With average times spent during execution of each subtask are included, the number of columns becomes 4.

Type of average analysis	PCA matrix size
(a) Average tool tip velocities for "Expose triangle"	18x3
(b) Average tool tip velocities for "Dissect CD/CA"	18x3
(c) Average tool tip velocities and time for "Expose triangle"	18x4
(d) Average tool tip velocities and time for "Dissect CD/CA"	18x4

Table 5.4: PCA matrix sizes for various types of average analysis

By applying our PC selection method (Section 3.4.4.2) we found that retaining only the first principal component explains most of the variance across all subjects and subtasks. **Figure 5.5** shows the PC selection results for the cases shaded in **Table 5.4**. However, as in the Mandarin experiment, we used the first two PCs when plotting our data in order to gain further insight in the interpretation of the data.

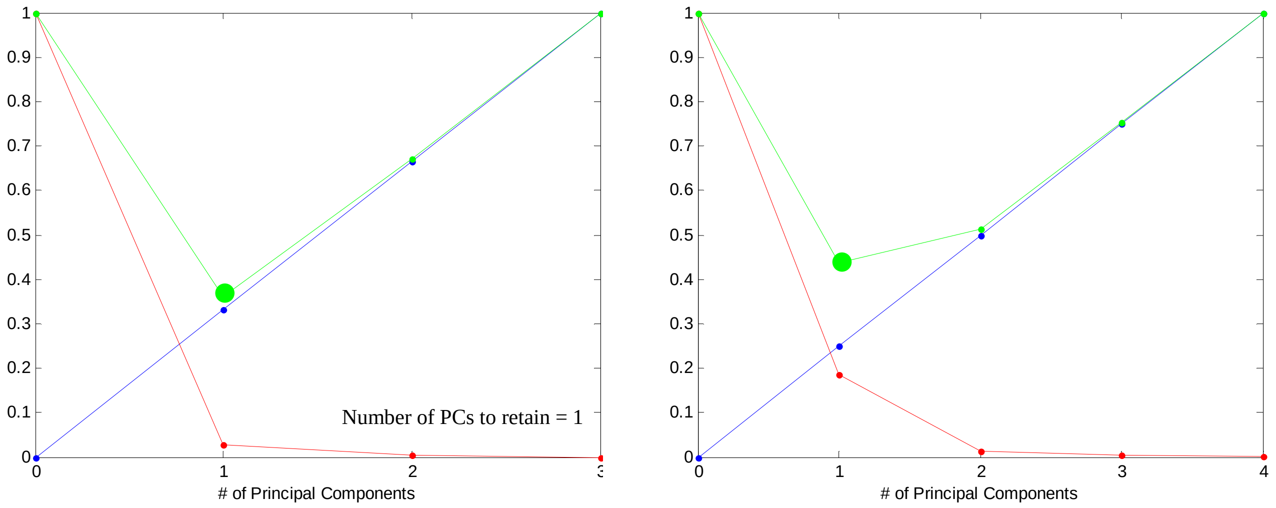
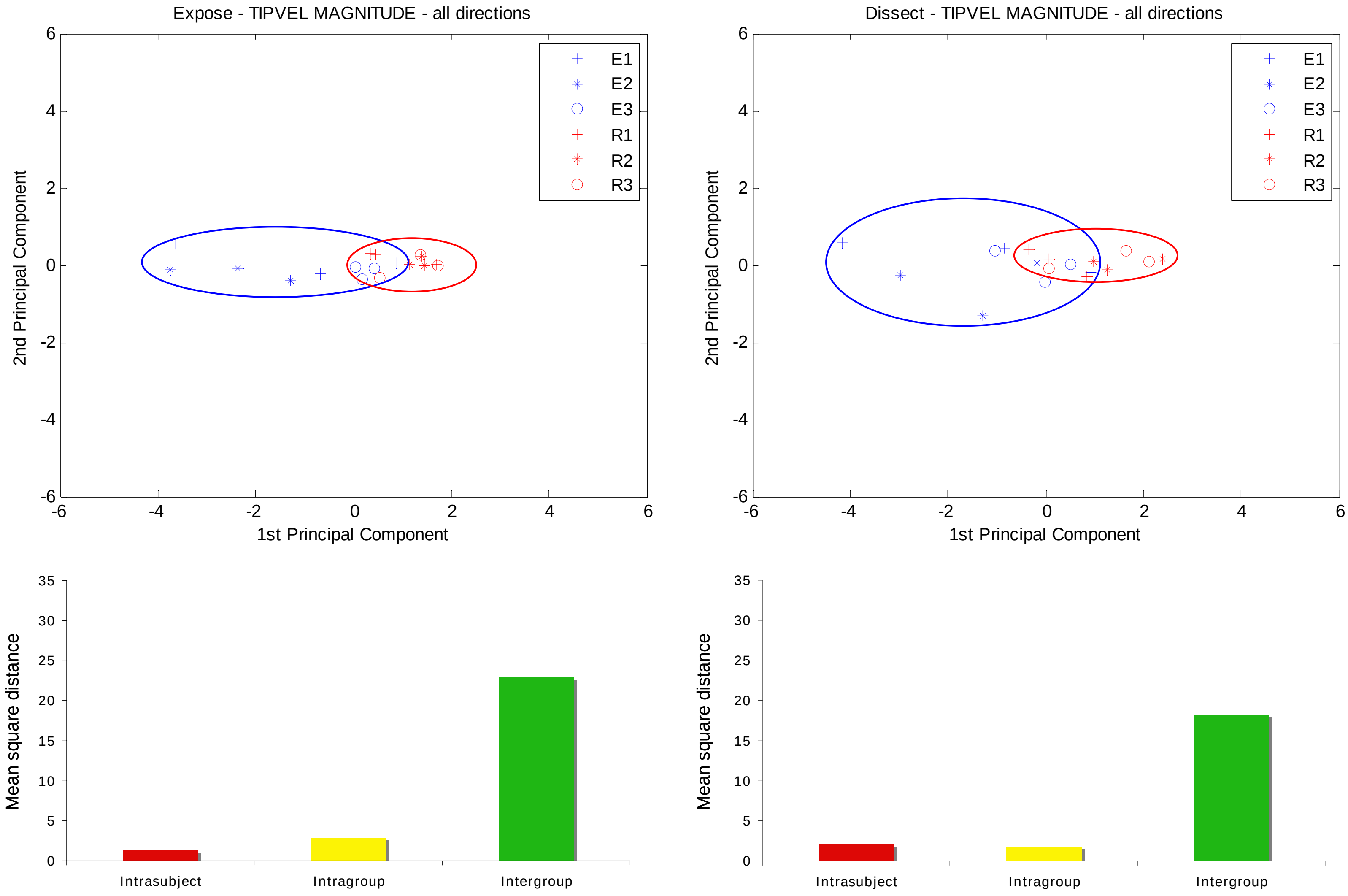


Figure 5.5: Selecting number of Principal Components for cases (a), and (c) in table 5.4. Blue: complexity measure (# of PCs); Red: variance unaccounted for; Green: dimensionality-decision criteria (sum of the two parameters); optimal indicated by large green dot

Figure 5.6 and **Figure 5.7** show the variability results for the various cases listed in **Table 5.4** when monitoring the dominant hand. A nested ANOVA (intragroup DoF = 4, intergroup DoF = 1) showed significant contributions of group membership (p<0.05) in differentiating skill levels. Note that generally residents are located on the right hand side of the plot while experts are located on the left hand side along the first principal component. This apparent separation is also highlighted, as the intergroup component was considerably higher than the intrasubject and intragroup components in all four cases.

Analysis of average tip velocity

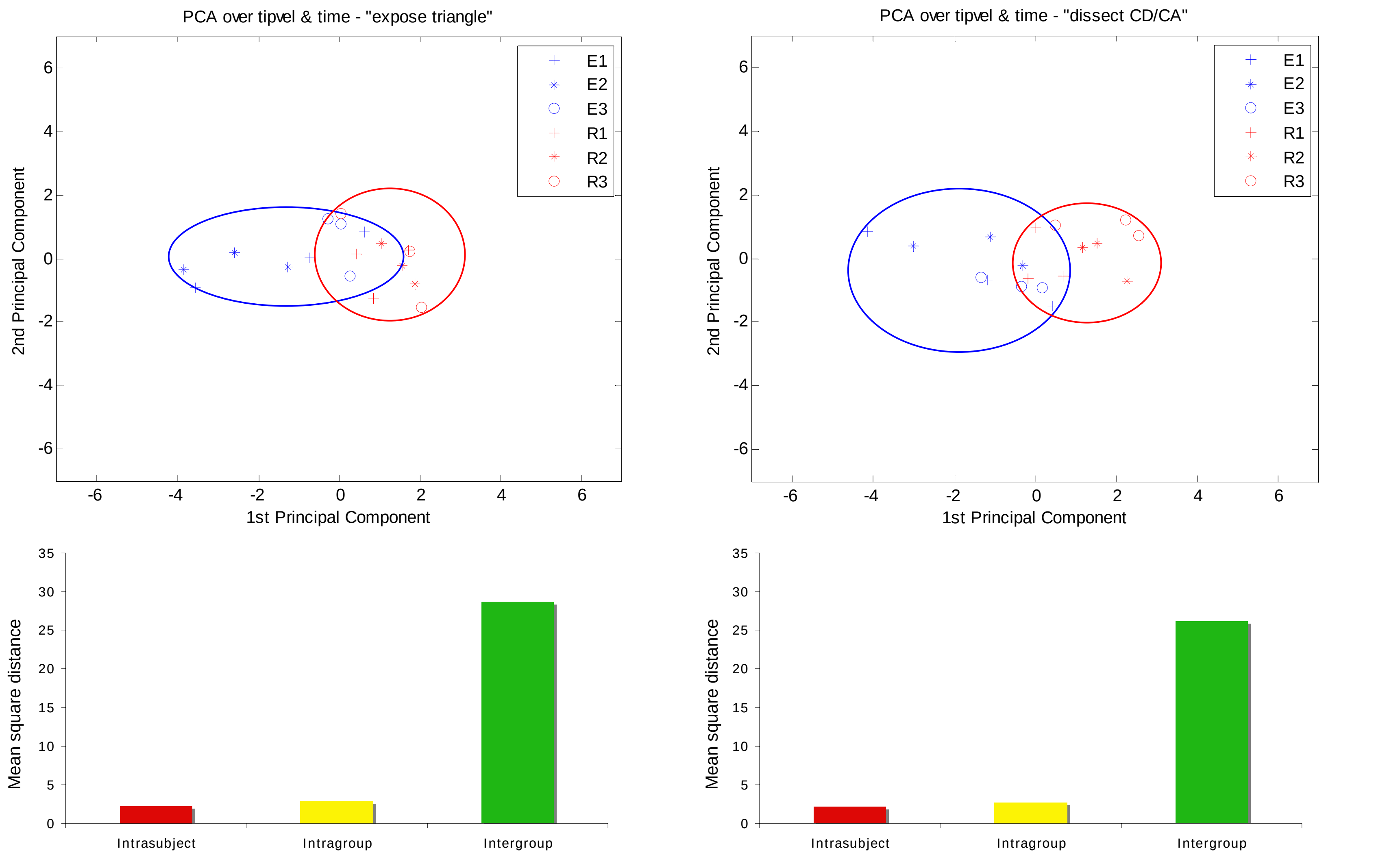


F = 7.78; p = 0.047*

F = 10.28; p = 0.03*

Figure 5.6: PCA variability analysis (Dominant hand) – Average tool tip velocity (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Analysis of average tip velocity and times



F = 9.96; p = 0.03*

F = 9.64; p = 0.04*

Figure 5.7: PCA variability analysis (Dominant hand) – Average tool tip velocity and times. (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

To determine whether the details of the velocity distributions convey more nuanced information that can be used to discriminate between groups, we extracted values from the velocity profiles at every 5th percentile from 25th to 75th percentile and again constructed the matrices be used in the PCA analysis (**Table 5.5**). The total number of rows corresponds to the total number of recorded procedures (n=18 for this study) and the total number of columns corresponds to 11 values for lateral, axial, and vertical movements individually for each subtask, or 33 values when analyzing all the three movement directions at the same time for each subtask.

Type of average analysis	**PCA matrix size**
(a) 25 to 75 percentiles of velocity profiles for each movement direction (lateral, axial, vertical) during 'Expose Triangle'	18x11
(b) 25 to 75 percentiles of velocity profiles for each movement direction (lateral, axial, vertical) during 'Dissect CD/CA'	18x11
(c) 25 to 75 percentiles of velocity profiles for all movement directions concatenated together (Expose triangle)	18x33
(d) 25 to 75 percentiles of velocity profiles for all movement directions concatenated together (Dissect CD/CA)	18x33

Table 5.5: PCA matrix sizes for various types of detailed (i.e., percentile) analysis

According to our selection method (**Figure 5.8** for shaded cases on **Table 5.5**), one and two principal components were sufficient to represent the variance across the data set for all cases in table 5.5. Since we have previously shown in the Mandarin experiment that a 2D description is effective to visualize differences; we then chose to use a 2D representation in order to be consistent in presenting results.

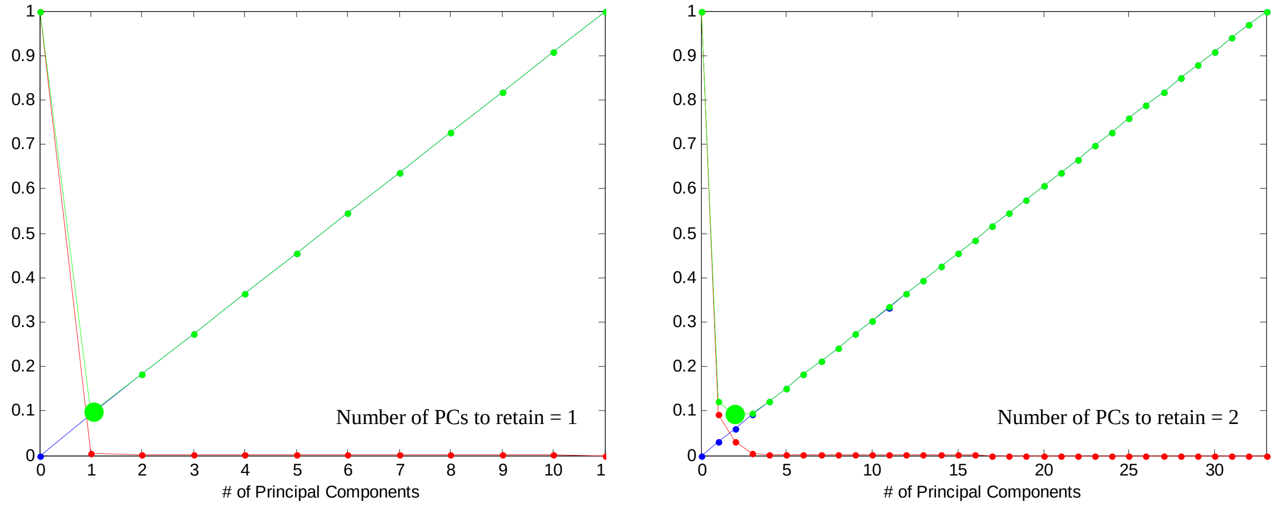
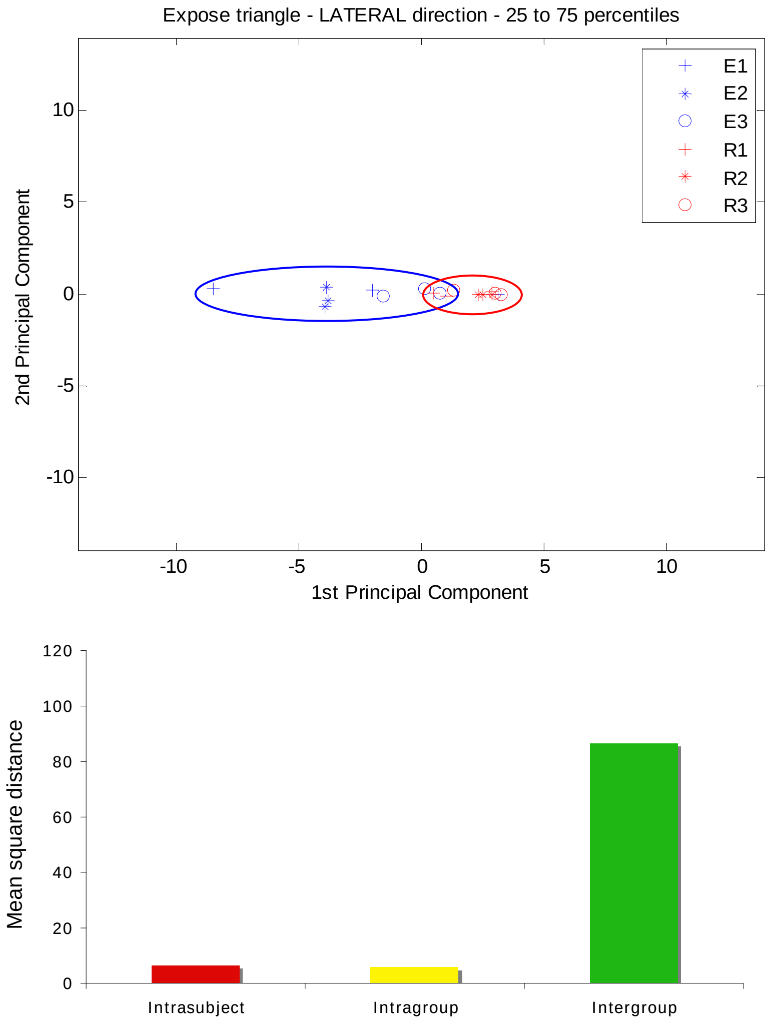


Figure 5.8: Selecting number of Principal Components for cases (a), and (c) on table 5.5. Blue: complexity measure (# of PCs); Red: variance unaccounted for; Green: dimensionality-decision criteria (sum of the two parameters); optimal indicated by large green dot

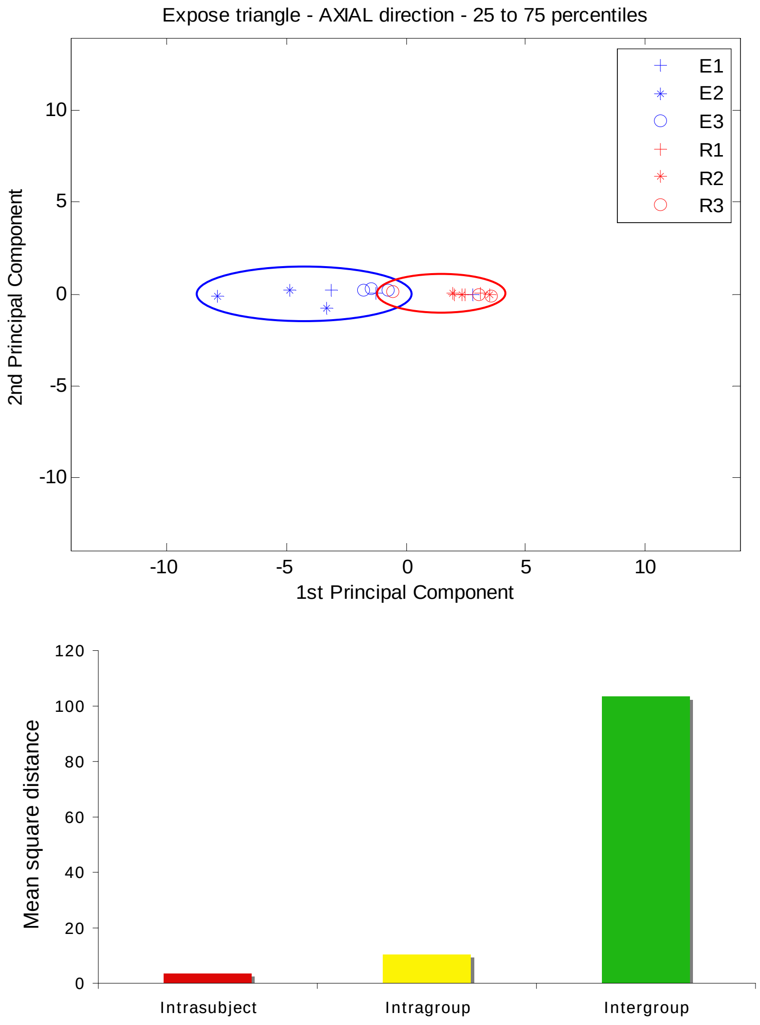
Figures 5.9 to **5.11** show the variability results for the detailed analyses listed in **Table 5.5** when monitoring the dominant hand. Although group separation was significant for both subtasks, the intergroup mean square (MS) was consistently higher in the ‘Expose Triangle’ subtask than in the ‘Dissect CD/CA’ subtask.

We repeated the same ‘average’ and ‘detail’ processes previously described for analyzing the non-dominant hand and the corresponding results are presented in **Figure 5.12** to **Figure 5.16**. In this analysis separation between groups was not significant for any case; however, p-values for all cases in the ‘Expose Triangle’ subtask were consistently lower than for the ‘Dissect CD/CA’ subtask. Additionally, the MS values for the three variability components in ‘Dissect CD/CA’ were similar, while for ‘Expose Triangle’ the intergroup component was noticeably greater than the other two.

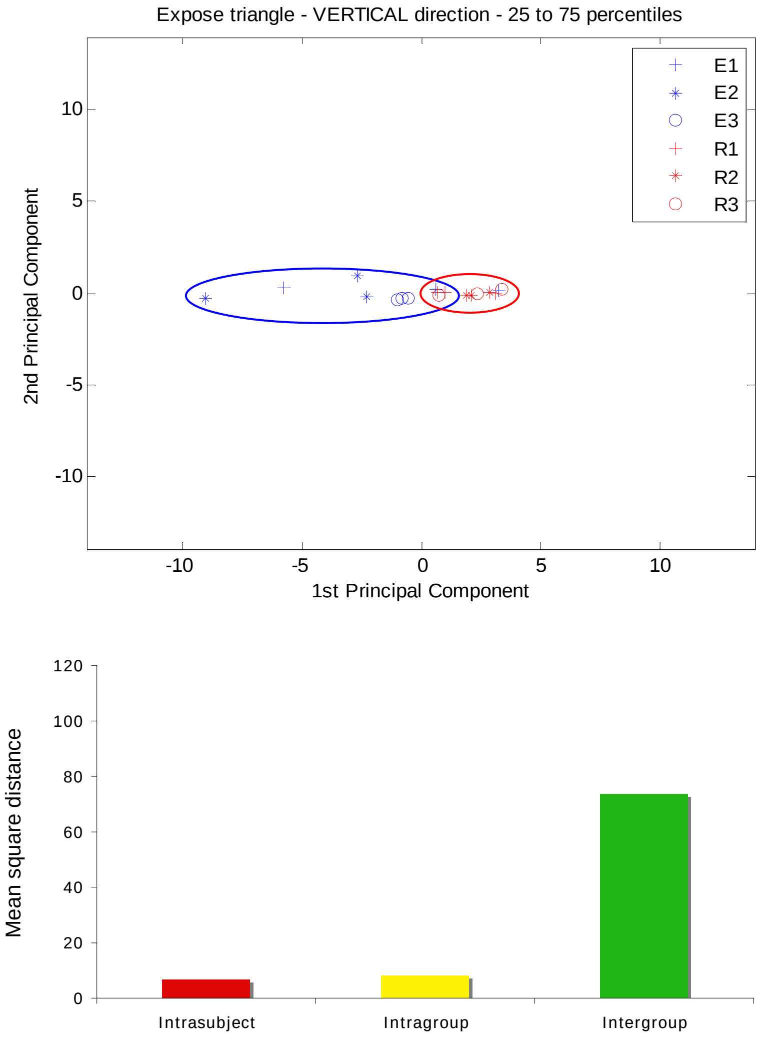
Analysis of detailed (25th to 75th percentiles) velocity profiles for individual movement directions during ‘Expose triangle’



F = 14.94; p = 0.02*



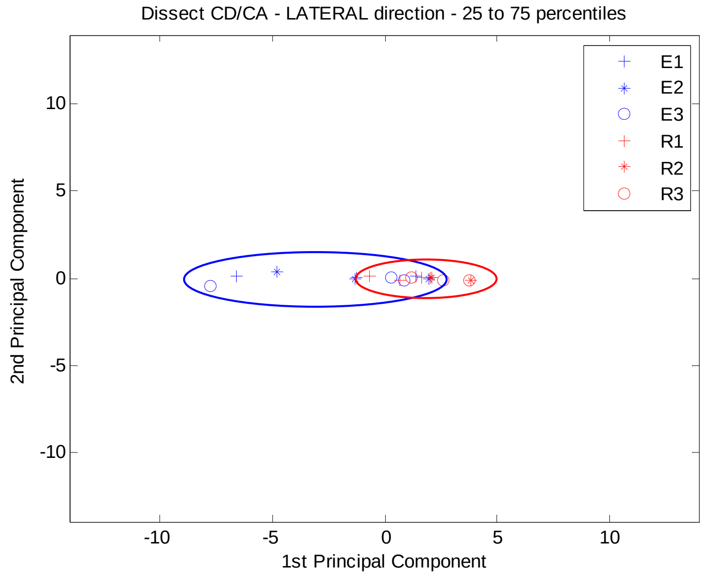
F = 10.01; p = 0.034*

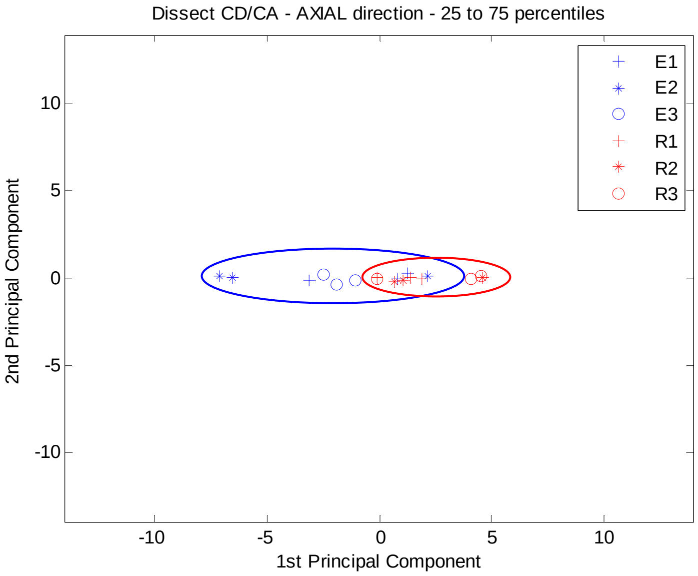


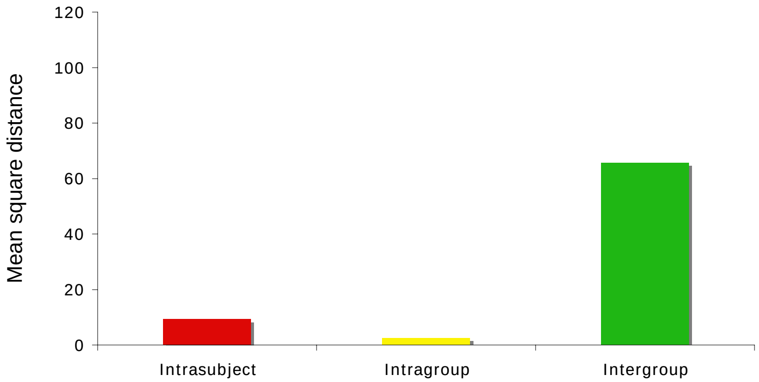
F = 9.01; p = 0.04*

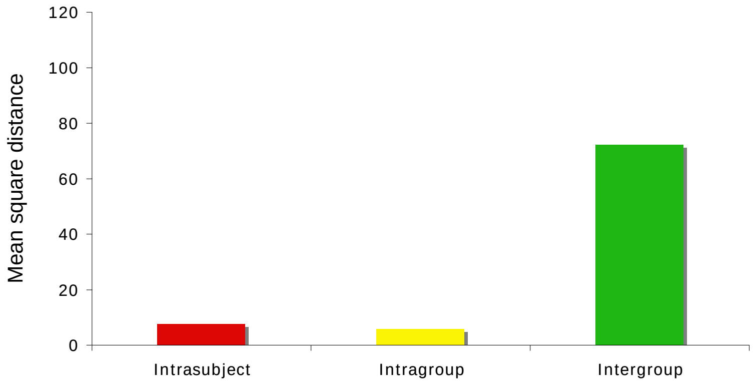
Figure 5.9: PCA variability analysis (Dominant hand) – 25th to 75th percentiles (individual directions ‘Expose triangle’) (* indicates statistical significant) Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Analysis of detailed (25th to 75th percentiles) velocity profiles for individual movement directions during ‘Dissect CD/CA’


Dissect CD/CA - LATERAL direction - 25 to 75 percentiles

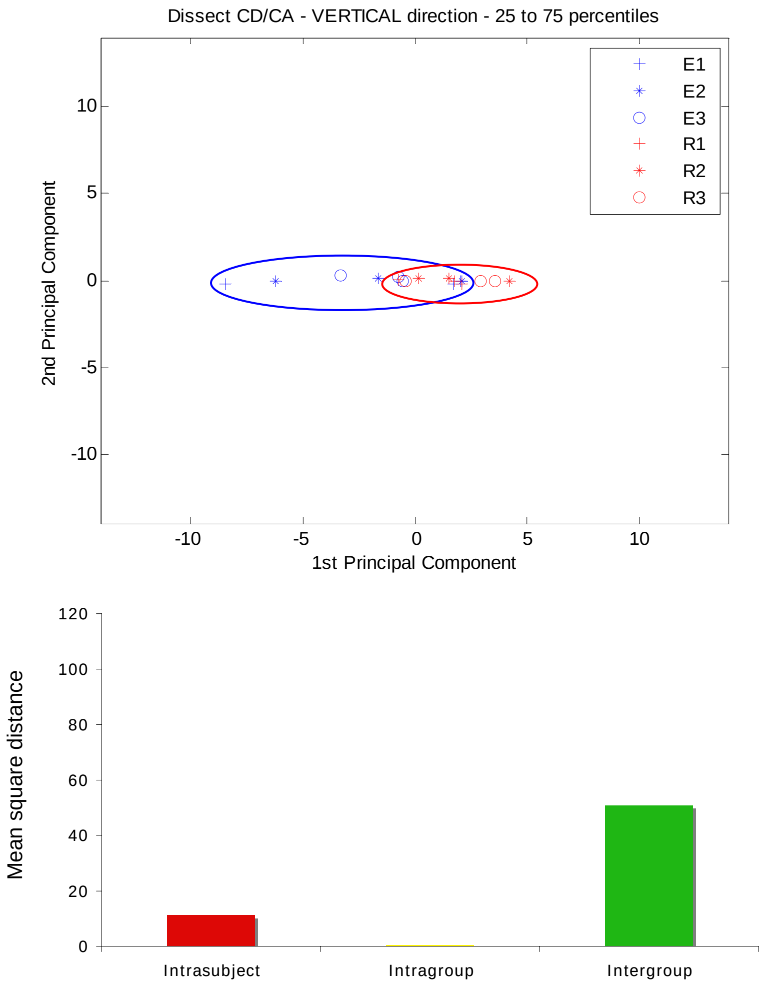

Dissect CD/CA - AXIAL direction - 25 to 75 percentiles





F = 26.67; p = 0.01*

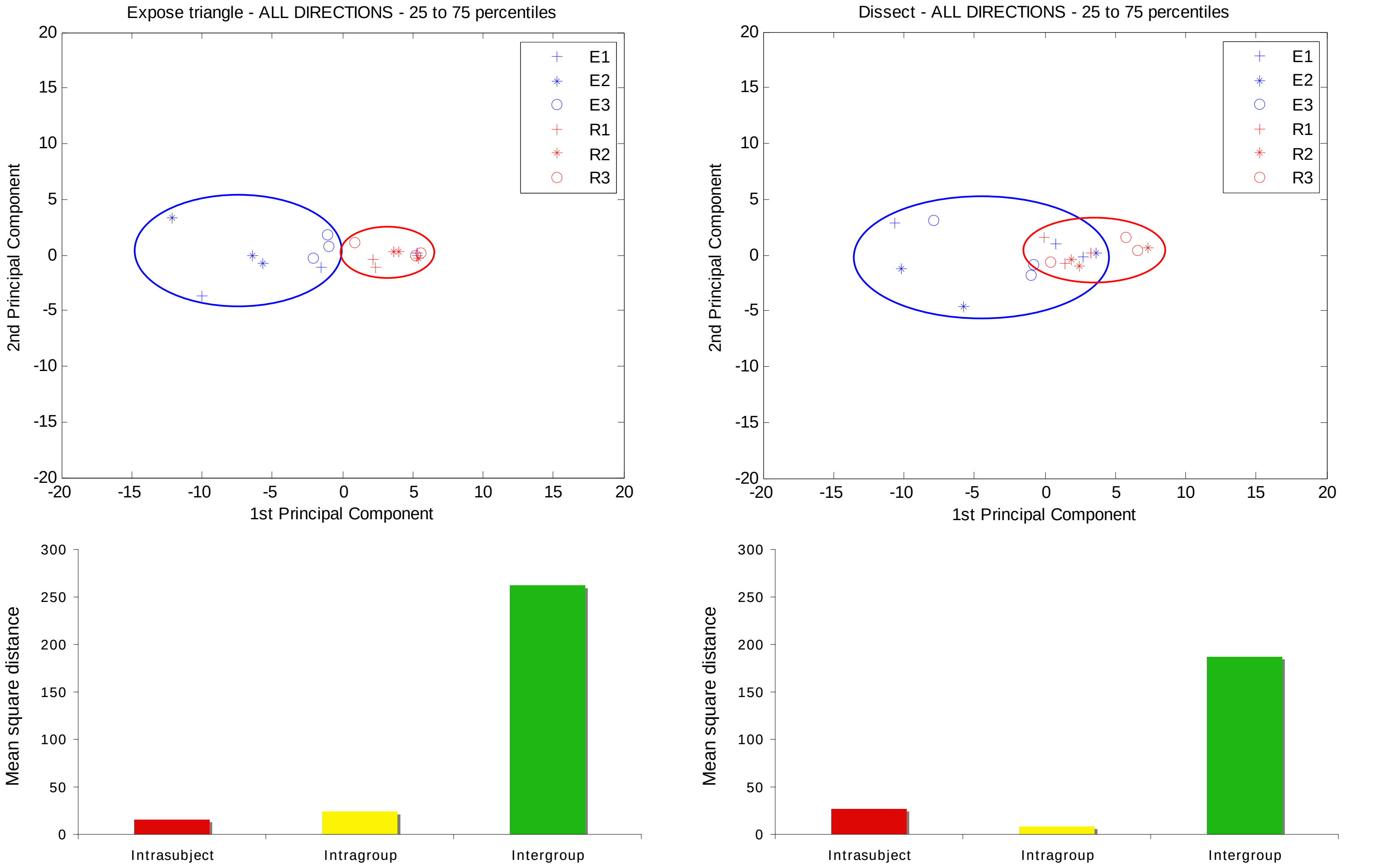
F = 12.52; p = 0.02*



F = 90.15; p = 0.001*

Figure 5.10: PCA variability analysis (Dominant hand) – 25th to 75th percentiles (individual directions ‘Dissect CD/CA’) (* indicates statistical significant) Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Analysis of detailed (25th to 75th percentiles) velocity profiles including all movement directions during each individual subtask



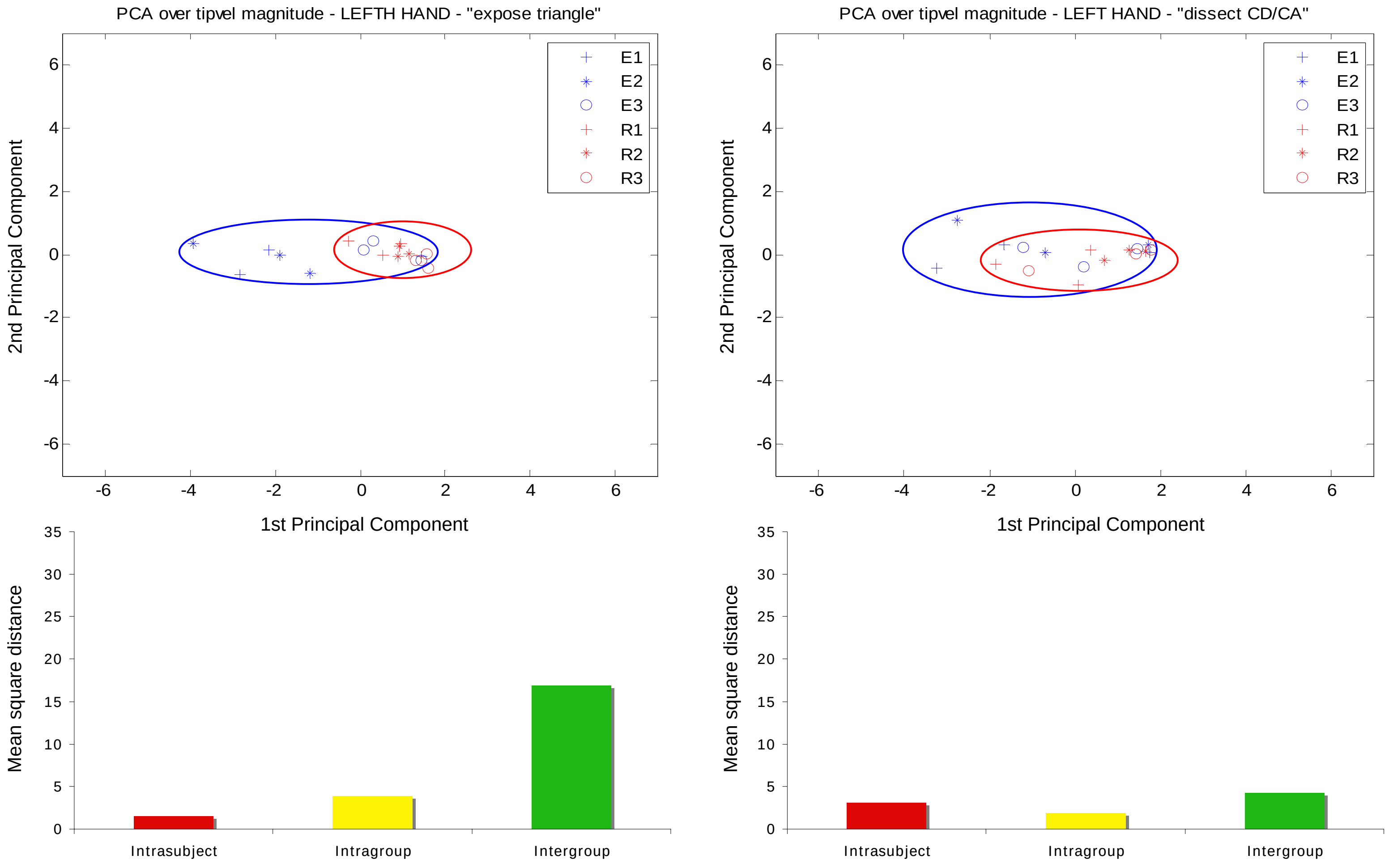
F = 10.94; p = 0.03*

F = 22.48; p = 0.01*

Figure 5.11: PCA variability analysis (Dominant hand) – 25th to 75th percentiles (ALL directions; individual subtasks) (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

NON-DOMINANT HAND (left hand for all subjects) VARIABILITY RESULTS

Analysis of average tip velocity

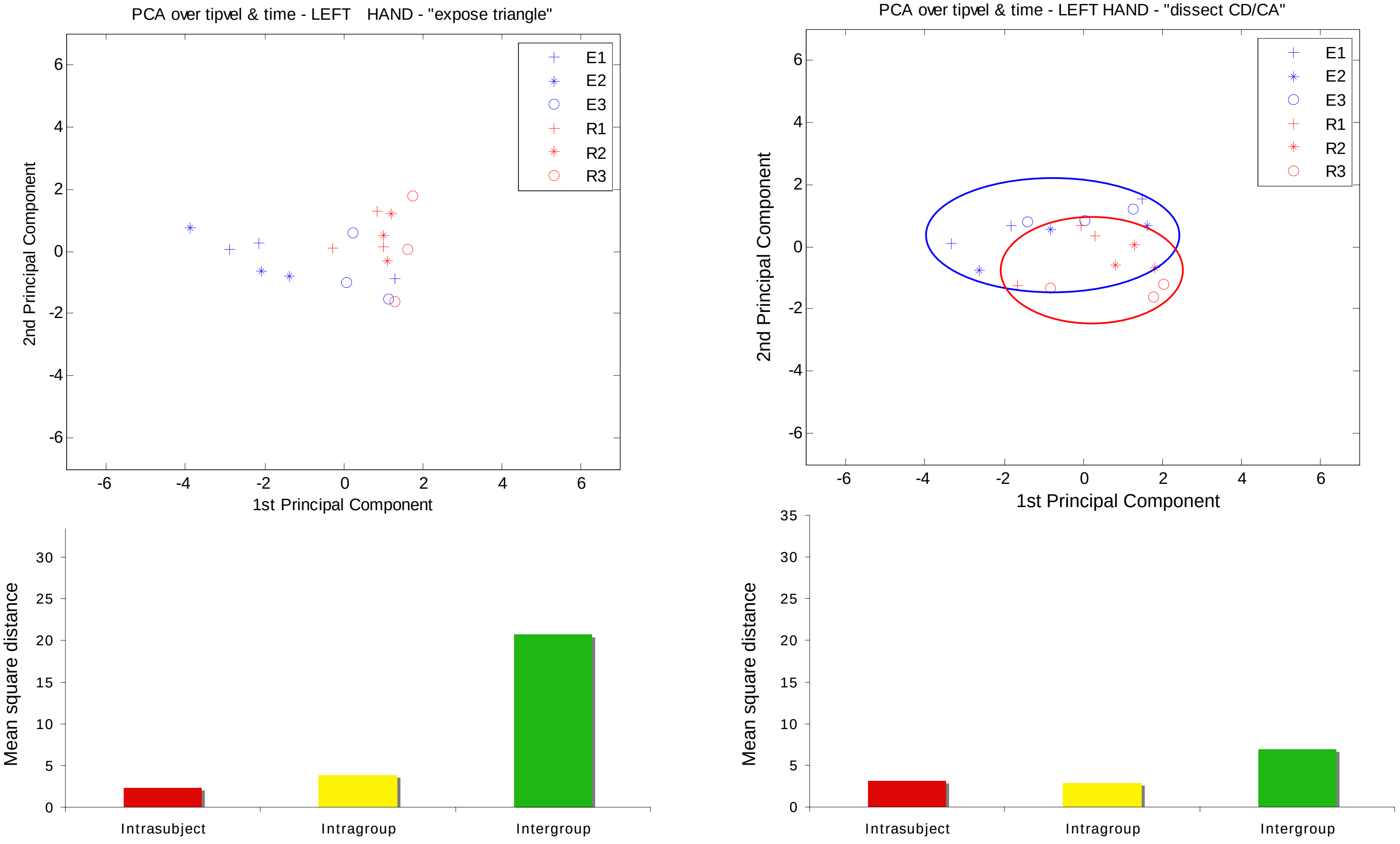


F = 4.35; p = 0.11

F = 2.30; p = 0.20

Figure 5.12: PCA variability analysis (Non-dominant hand) – Average tool tip velocity (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Analysis of average tip velocity and times

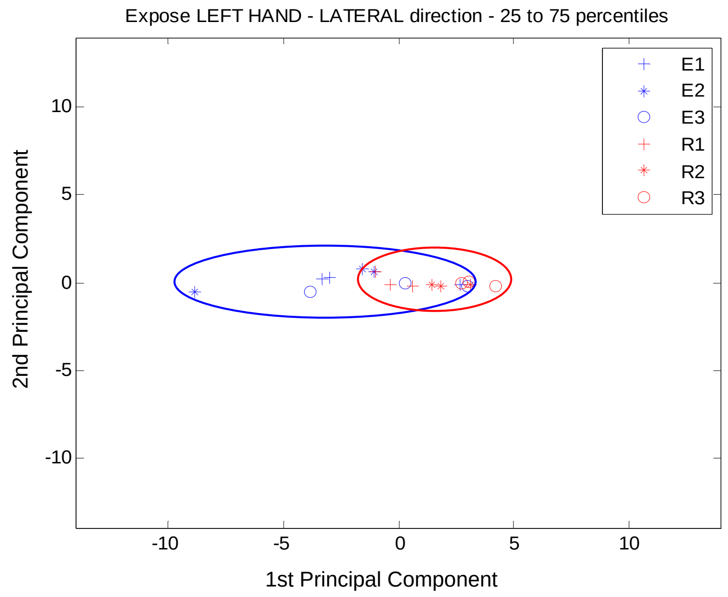


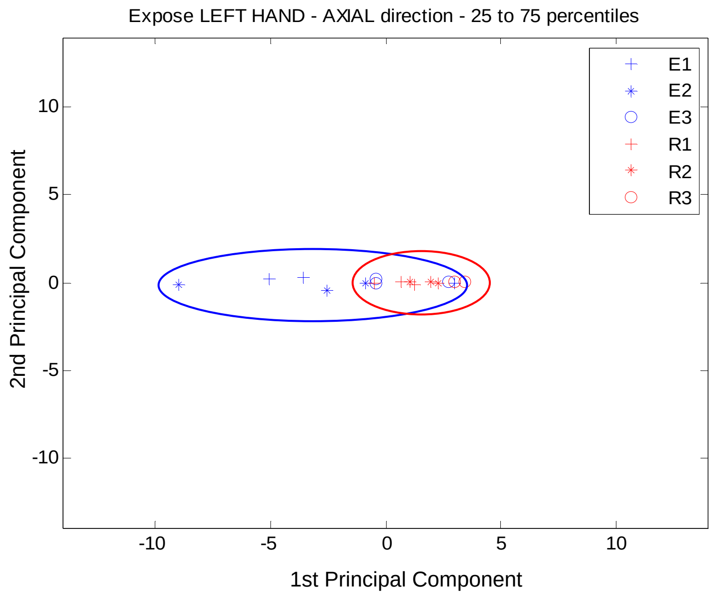
F = 5.41; p = 0.08

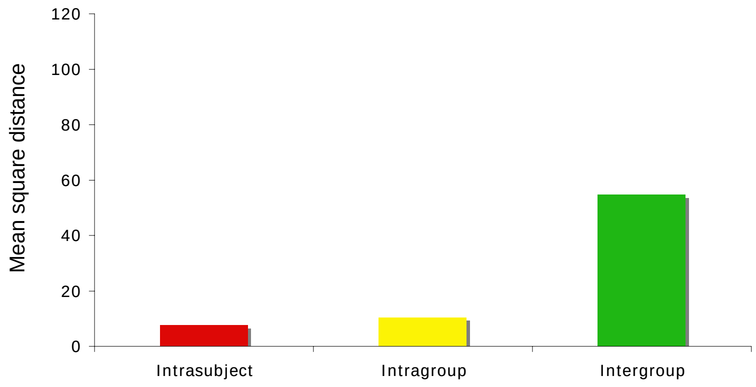
F = 2.41; p = 0.20

Figure 5.13: PCA variability analysis (Non-Dominant hand) – Average tool tip velocity and times (* indicates statistical significant) Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

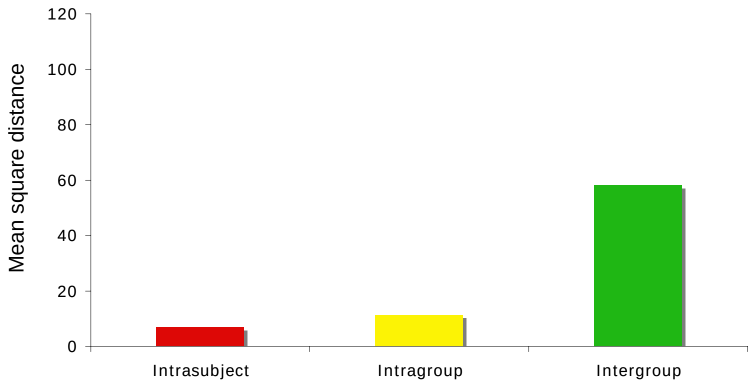
Analysis of detailed (25th to 75th percentiles) velocity profiles for individual movement directions during ‘Expose triangle’


Expose LEFT HAND - LATERAL direction - 25 to 75 percentiles

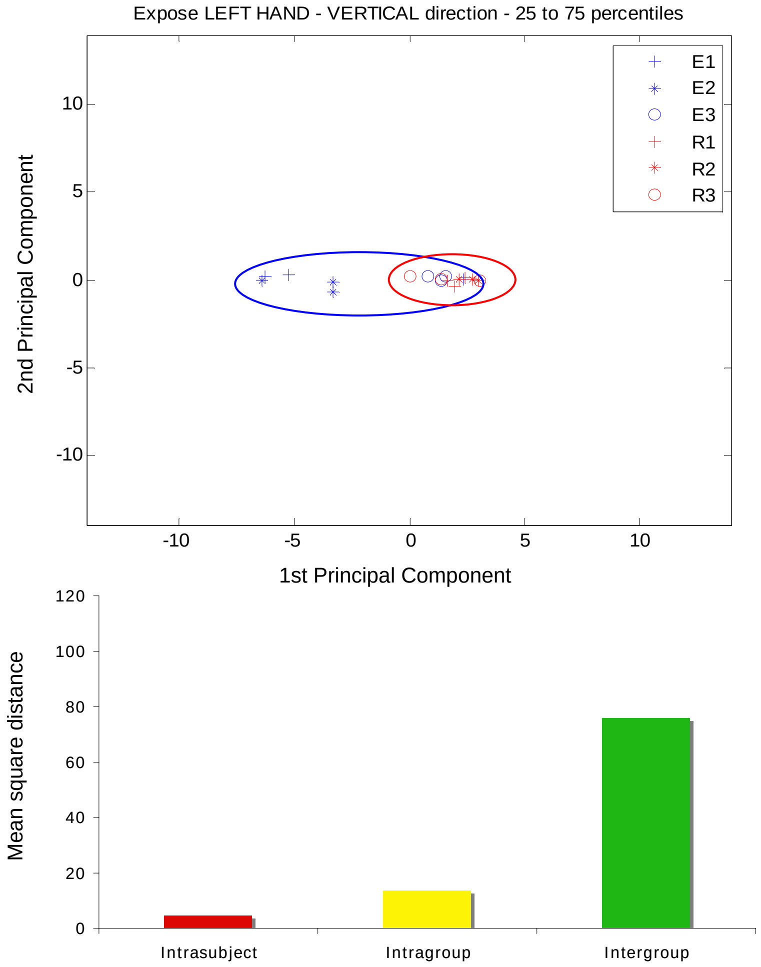

Expose LEFT HAND - AXIAL direction - 25 to 75 percentiles



F = 5.26; p = 0.08



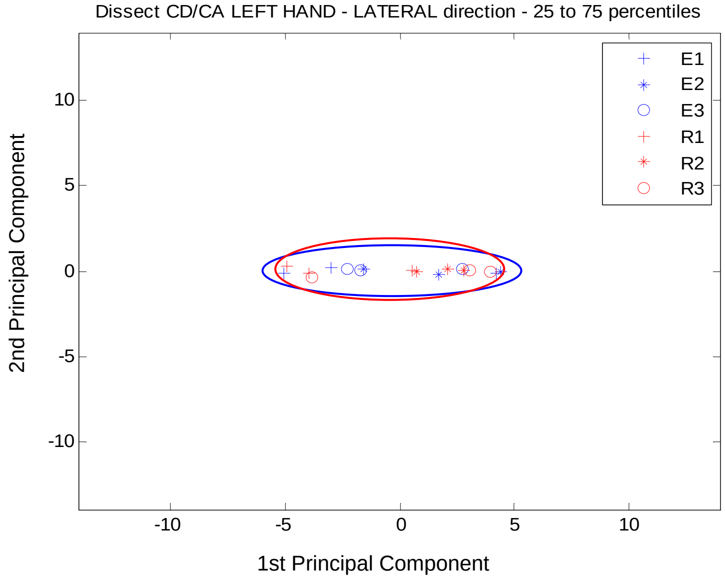
F = 5.12; p = 0.09

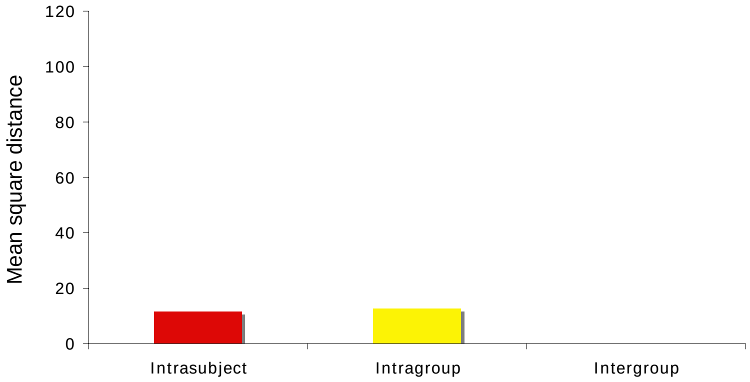


F = 5.58; p = 0.08

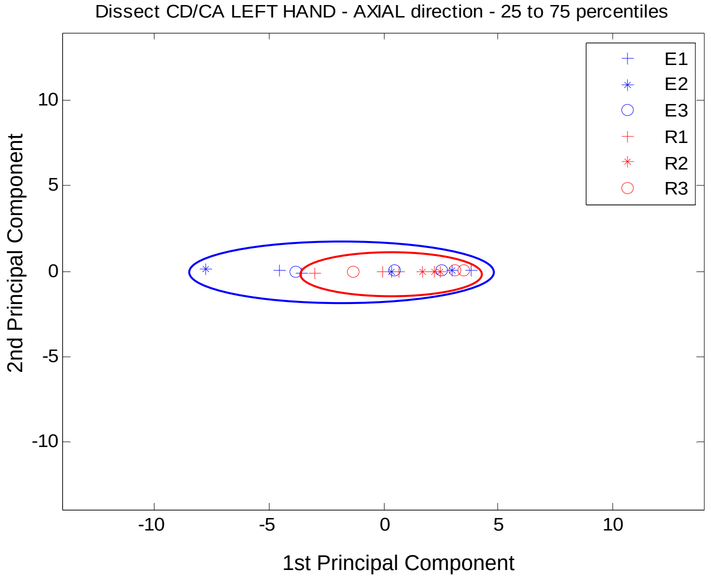
Figure 5.14: PCA variability analysis (Non-dominant hand) – 25th to 75th percentiles (individual directions ‘Expose triangle’) (* indicates statistical significant) Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

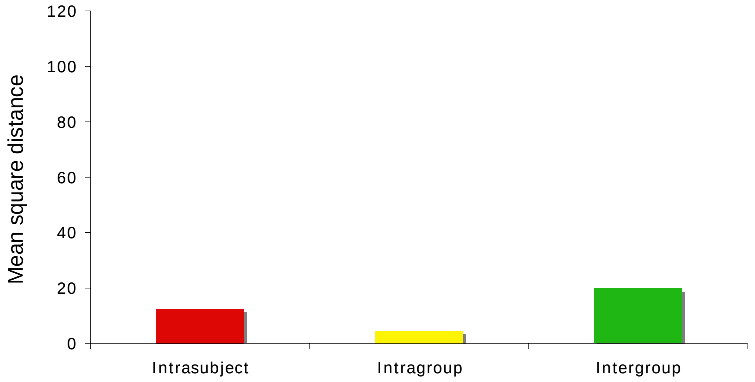
Analysis of detailed (25th to 75th percentiles) velocity profiles for individual movement directions during ‘Dissect CD/CA’



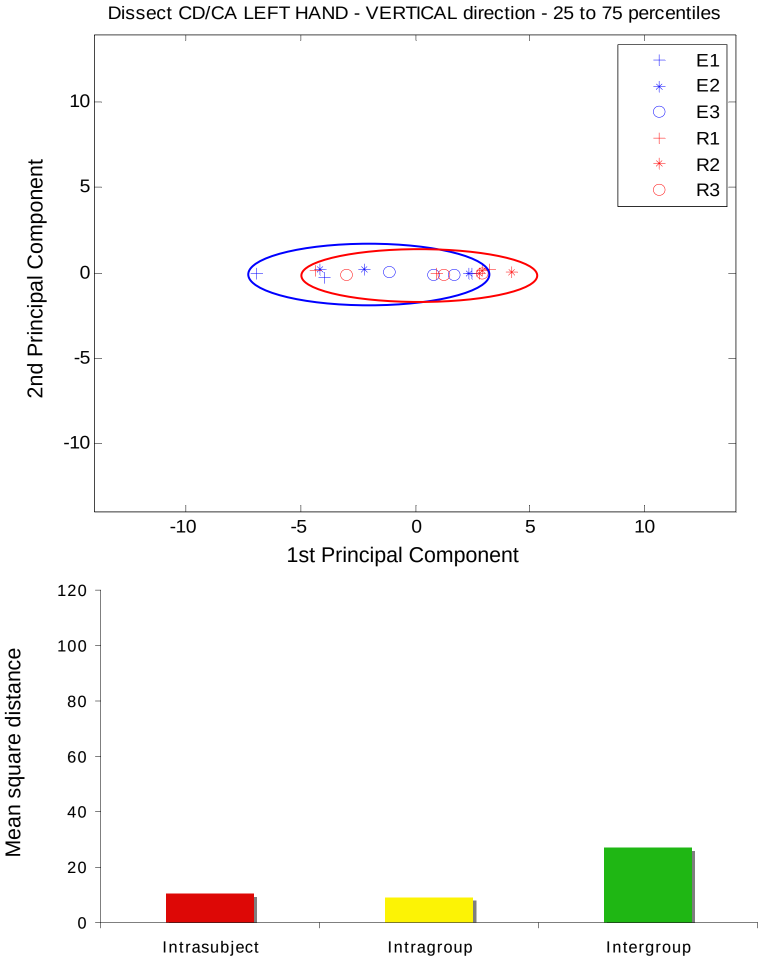


F = 0.004; p = 0.95





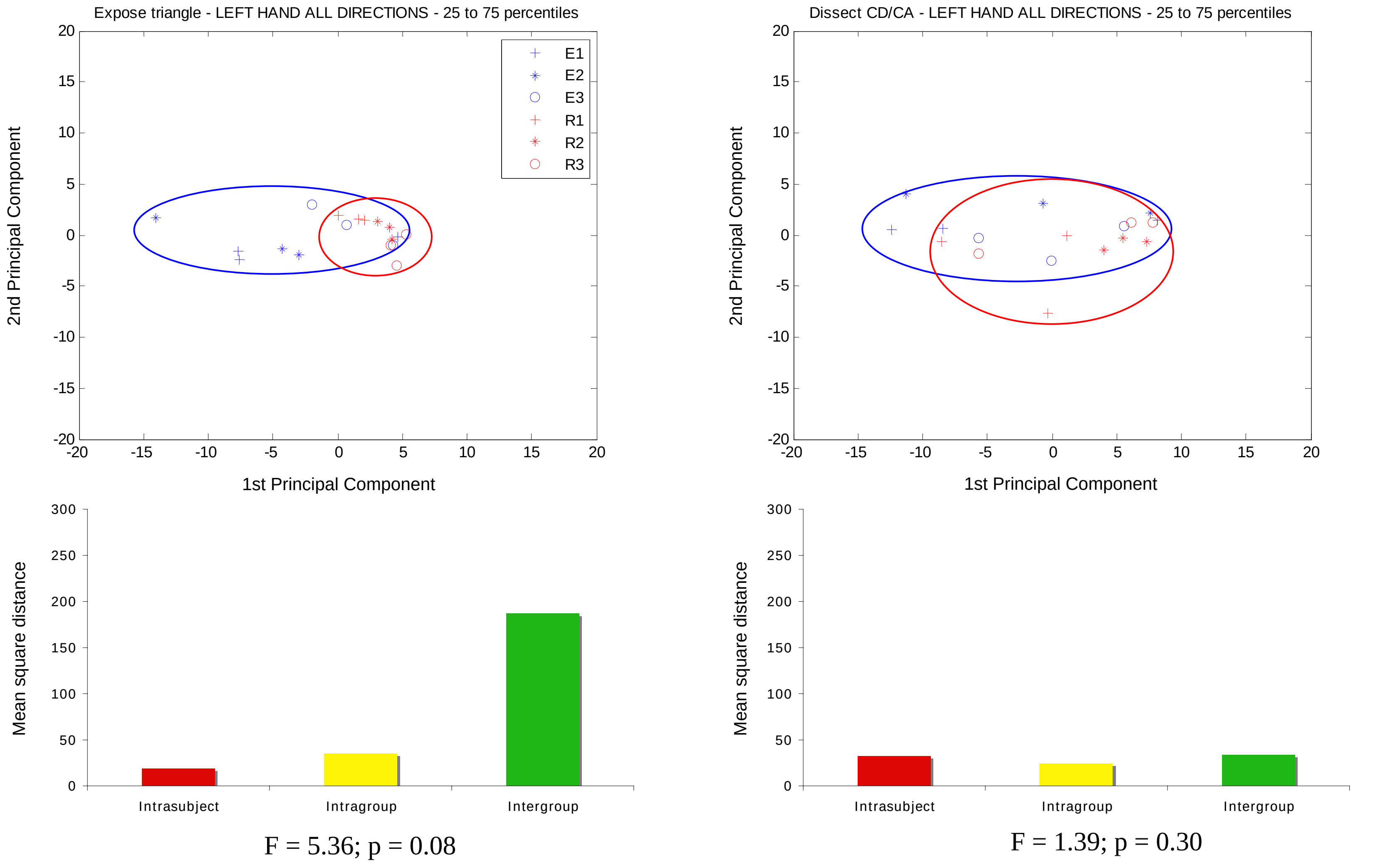
F = 4.37; p = 0.10



F = 2.99; p = 0.16

Figure 5.15: PCA variability analysis (Non-dominant hand) – 25th to 75th percentiles (individual directions ‘Dissect CD/CA’) (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Analysis of detailed (25th to 75th percentiles) velocity profiles including all movement directions during each individual subtask



F = 5.36; p = 0.08

F = 1.39; p = 0.30

Figure 5.16: PCA variability analysis (Non-dominant hand) – 25th to 75th percentiles (ALL directions; individual subtasks) (* indicates statistical significant)
Top: the positions of trials in the PCA space; Bottom: Sum of squares of each variability component

Our variability analysis at the subtask level indicated that:

(1) The first PC explained the majority (> 80%) of the variation, and there was little contribution of the second principal component.

(2) The nested ANOVA tests for both ‘Expose Triangle’ and ‘Dissect CD/CA’ subtasks when monitoring the dominant hand showed a significant intergroup contribution to variability, which indicates that the velocity measure is able to distinguish between residents and experts

(3) This result is consistent with the data points shown in the PCA plots where it seems that there is consistent separation between the residents at the right side of the plot and the experts spread over a wider range on the left side of the plot

(4) In contrast, the test failed to find such a distinction between the residents and experts when monitoring the non-dominant hand

5.3.2 Variability Analysis at the Action Level

At the action, we investigated whether: (1) there is any significant variation in movement behaviour for factors considered one at a time from procedure to procedure; and (2) if a PCA analysis including all of our difference measures provides group separation when comparing subjects from the same group (i.e., experts vs experts; residents vs residents) and subjects from different groups (experts vs residents).

5.3.2.1 Variation From Procedure to Procedure

To test the first hypothesis, we compared the kinematics profiles of each movement across all the procedures performed by one resident and one expert (**Figure 5.17**) using the Kolmogorov-Smirnov statistic (D-value), which measures the discrepancy between two empirical distributions (Section 3.4.1) on a 0 to 1 scale. It is interesting to note that while the resident seems consistent across the three procedures, the expert varied widely from a ‘fast’ procedure to a ‘slow’ one. We believe that variations in the procedure difficulty might have an influence here since training guidelines only allow the residents to handle relatively simple cases while the experts can deal with cases of any degree of difficulty.

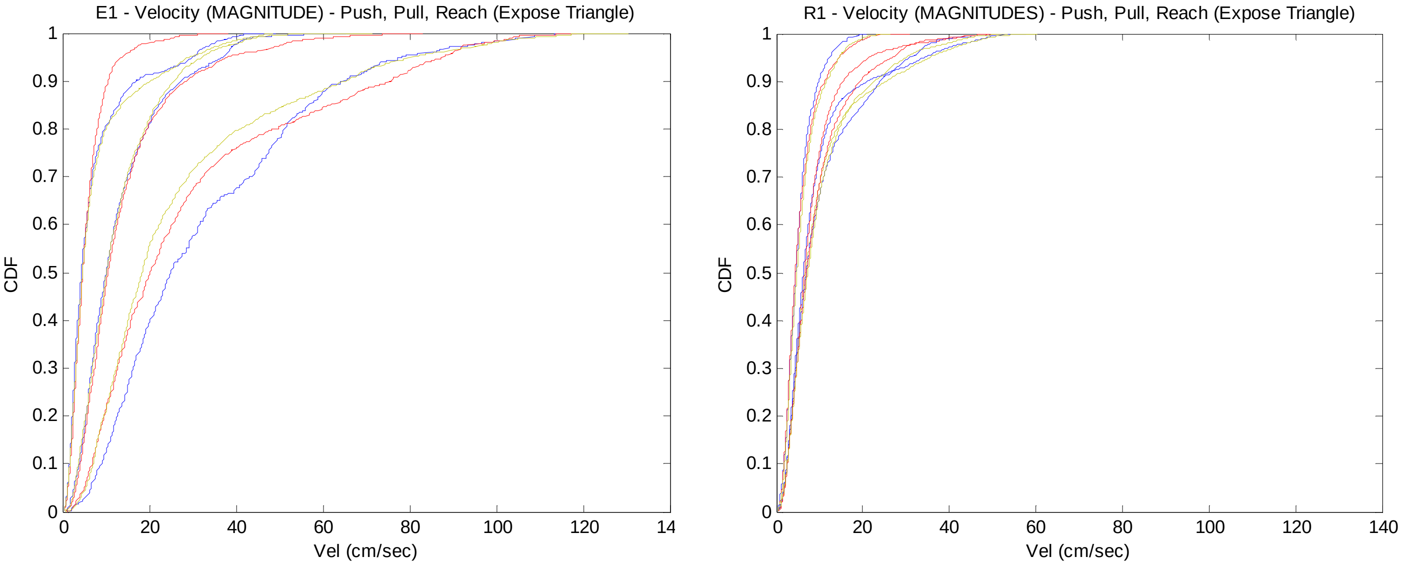


Figure 5.17: Samples of velocity profiles (CDF) for E1 and R1 across 3 procedures each and for 3 types of movements during ‘Expose Triangle’ (blue: Push; red: Pull; yellow: Reach)

Figure 5.18 shows intrasubject, intragroup, and intergroup difference measures from comparing tip velocity profiles of the 10 actions using the Kolgomorov-Smirnov statistic. In general, a visual inspection indicates that there is more intrasubject variation amongst the experts, and the intergroup variation is modestly greater than the intragroup variation.

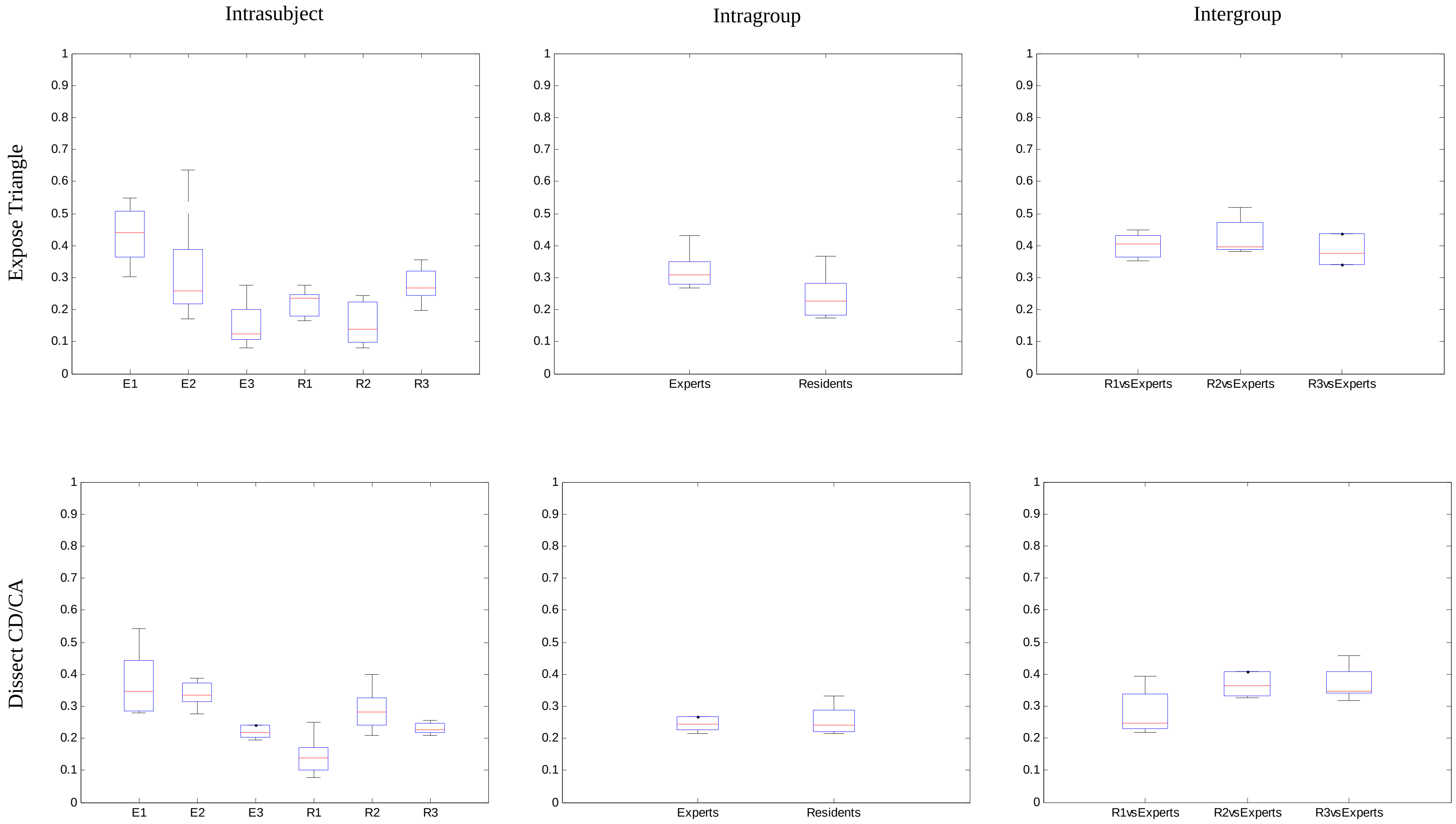


Figure 5.18: Ranges of D-values from comparing tip velocity profiles for the 10 actions during each subtask (red lines indicate mean Dvalue)

Figure 5.19 and **5.20** present difference measures computed from comparing time profiles and transitions for the 10 actions using the Kolgomorov-Smirnov and the Jensen-Shannon Divergence statistics respectively. Both show that there is no detectable difference in intra- and inter-group variability, which suggests that time or transition matrices alone are relatively weak predictors of group membership.

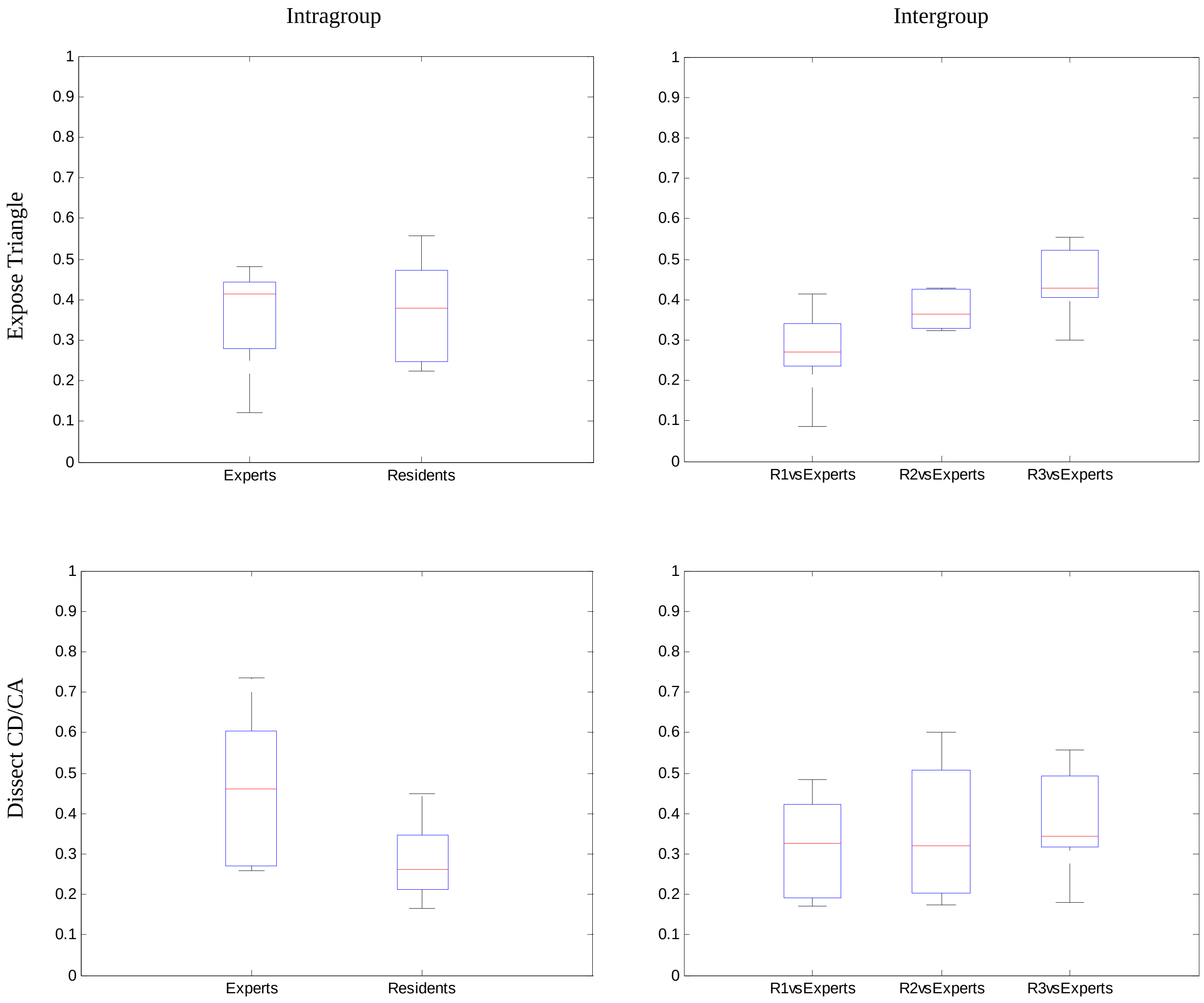


Figure 5.19: Ranges of D-values from comparing time profiles for the 10 actions during each subtask (red lines indicate mean Dvalue)

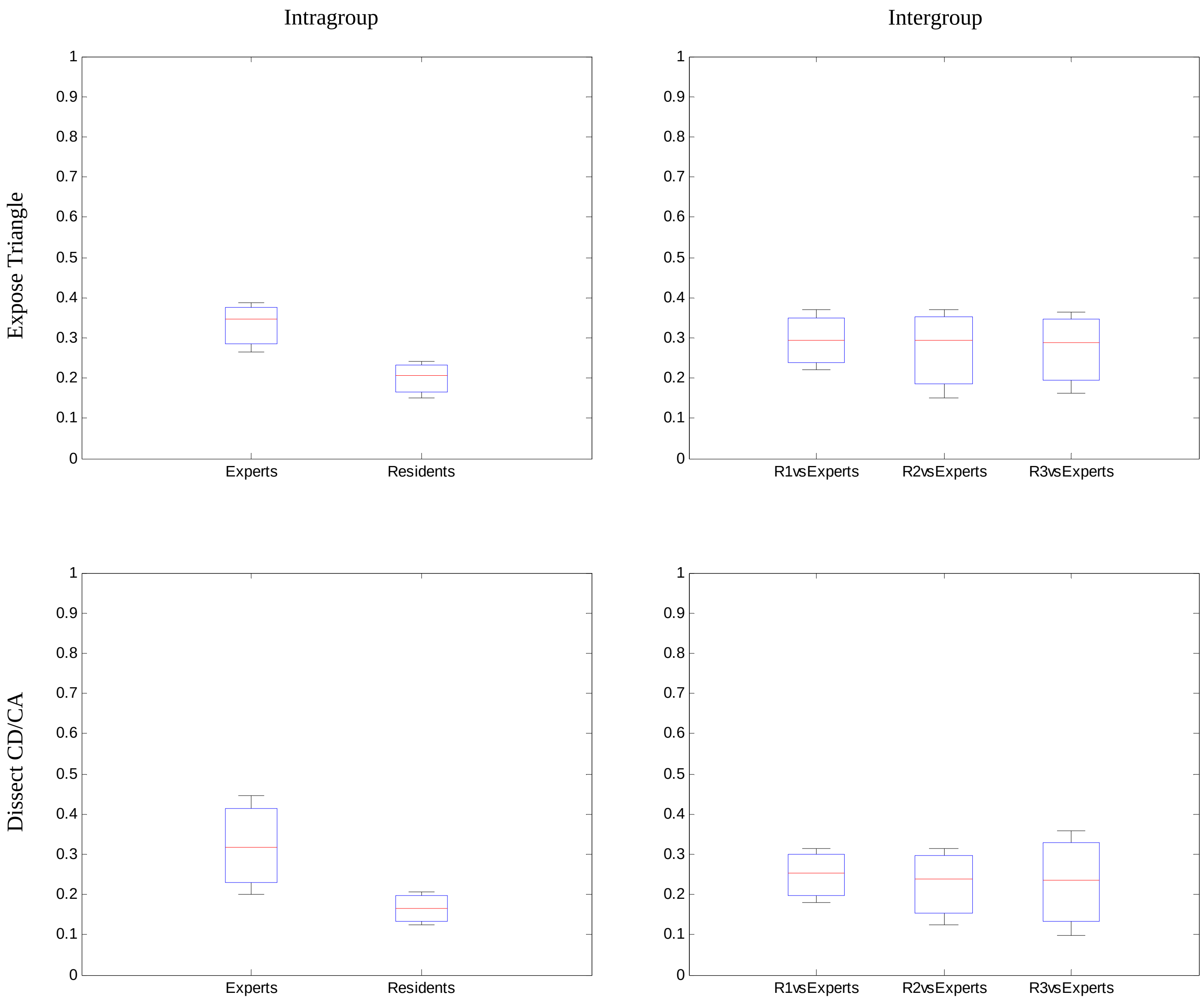


Figure 5.20: Ranges of JSD values from comparing movement transition matrices during each subtask (red lines indicate mean JSD value)

Procedure to procedure variability across subjects and groups in terms of D values (for kinematics and time profiles) and JSD values (for movement transitions) showed kinematics as the performance measure providing difference values for individual resident-to-experts comparison above the average expert-to-expert comparison level, i.e., > 0.3 (**Figure 5.18**, right and middle plots respectively). **Figure 5.19** and **Figure 5.20** did not show the same feature for comparing time and movement transitions. However, it is important to note that the reference levels (i.e., red lines in **Figure 5.18** to **Figure 5.20**) correspond to average of difference values among all procedures irrespective of the difficulty of the cases.

In order to quantify the differences among procedures in terms of patient anatomy conditions (**Table 5.6**), we asked two experienced laparoscopic surgeons to jointly assess the difficulty of each case by examining the recorded videos and reporting a score between 1 (least difficult) and 5 (most difficult). Both evaluators were blinded as to the identities of the operating surgeons.

Score	**Patient-related conditions**
1 (least difficult)	Pelvic surgery Petite patient Narrow costal margin
5 (most difficult)	Chronic cholecystitis Obesity Fatty liver Foregut surgery Bleeding dyscrasia Acute cholecystitis Choledocholithiasis

Table 5.6: Patient conditions associated to procedure difficulty (provided by the evaluators)

	Subject E1	**Subject E2**	**Subject E3**	**Subject R1**	**Subject R2**	**Subject R3**
Procedure 1	1	3	5	4	4	2
Procedure 2	2	2	3	5	1	2
Procedure 3	1	1	3	3	3	1

Table 5.7: Procedure difficulty scores from video analysis

As presented in **Table 5.7** and considering that it was not possible to incorporate exclusion criteria for the study due to the limited number of procedures available, difficulty scores spanned across the predefined scale (from 1 to 5), which added an additional factor affecting subjects comparisons. We believe that this issue was evident by the relative wide ranges of difference values, especially some corresponding to the experts group, which served as our reference level (**Figure 5.18** to **Figure 5.20**)

5.3.2.2 PCA Analysis

To test the second hypothesis that a PCA analysis of all 71[3] measures considered simultaneously will be able to provide intergroup separation, we constructed the following matrix (each row corresponds to a single subject-to-subject comparison) and then we performed a PCA analysis.

$$\begin{array}{cccc} \Big[JSD & Dvalues(ht) & Dvalues(\bar{v},\bar{a},\bar{j}) & Dvalues(v,a,j)\Big] \\ \text{nx1} & \text{nx10} & \text{nx30} & \text{nx30} \end{array}$$

[3] 1 JSD value; 10 D values (per variable) for average time, and summary velocity, acceleration, and jerk profiles (40 total); 10 D values (per variable) for detailed velocity, acceleration, and jerk profiles (30 total)

where n=15[4] (number of subject-to-subject comparisons). In selecting the number of principal components for dimensionality reduction, we found that between 4 and 5 components struck the best balance between accounting for most of the variability in the data sets and the number of dimensions needed (**Figure 5.21**).

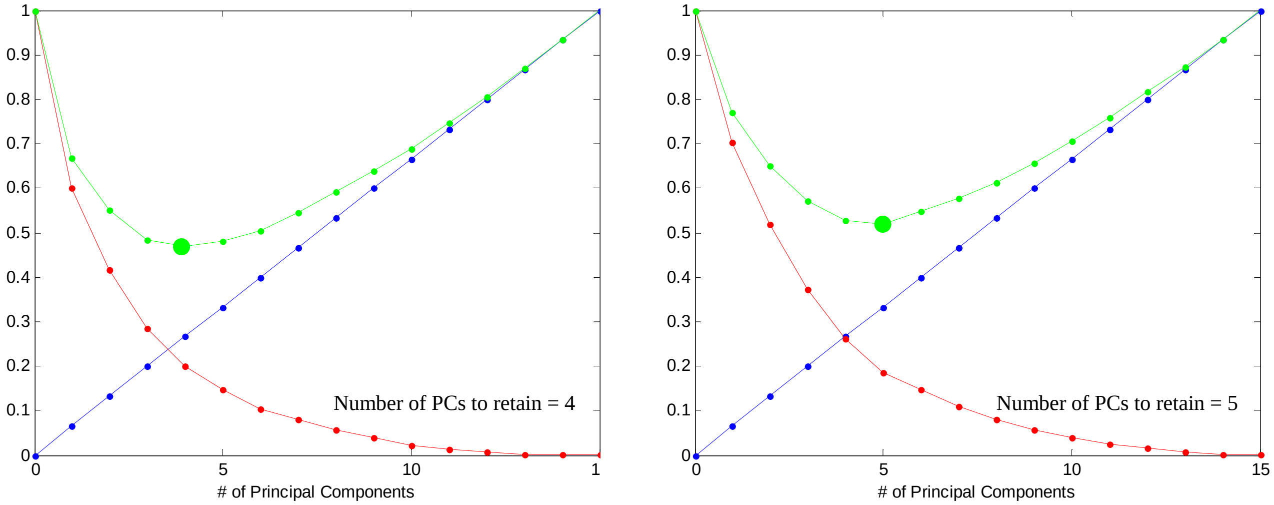


Figure 5.21: Selecting number of Principal Components for each subtask at the action level; blue: complexity measure; red: variance unaccounted for; green: dimensionality-decision criteria (sum of the two parameters); optimal indicated by large green dot. Left: 'Expose triangle'; Right: 'Dissect CD/CA'

In order to test if comparing subjects belonging to the same skill level and subjects from different levels provides differentiation between the two types of comparisons (groups) at each subtask, we plotted the first two principal components for each subtask to facilitate the graphical representation of the data (**Figure 5.22** and **Figure 5.23**) and then

[4] r(r-1)/2 = 3 intragroup comparisons for residents, e(e-1)/2 = 3 intragroup comparisons for the attendings, and rxe = 9 intergroup comparisons between residents and attendings.

computed a Mann-Whitney test to test the hypothesis that intragroup variability was less than intergroup variability[5].

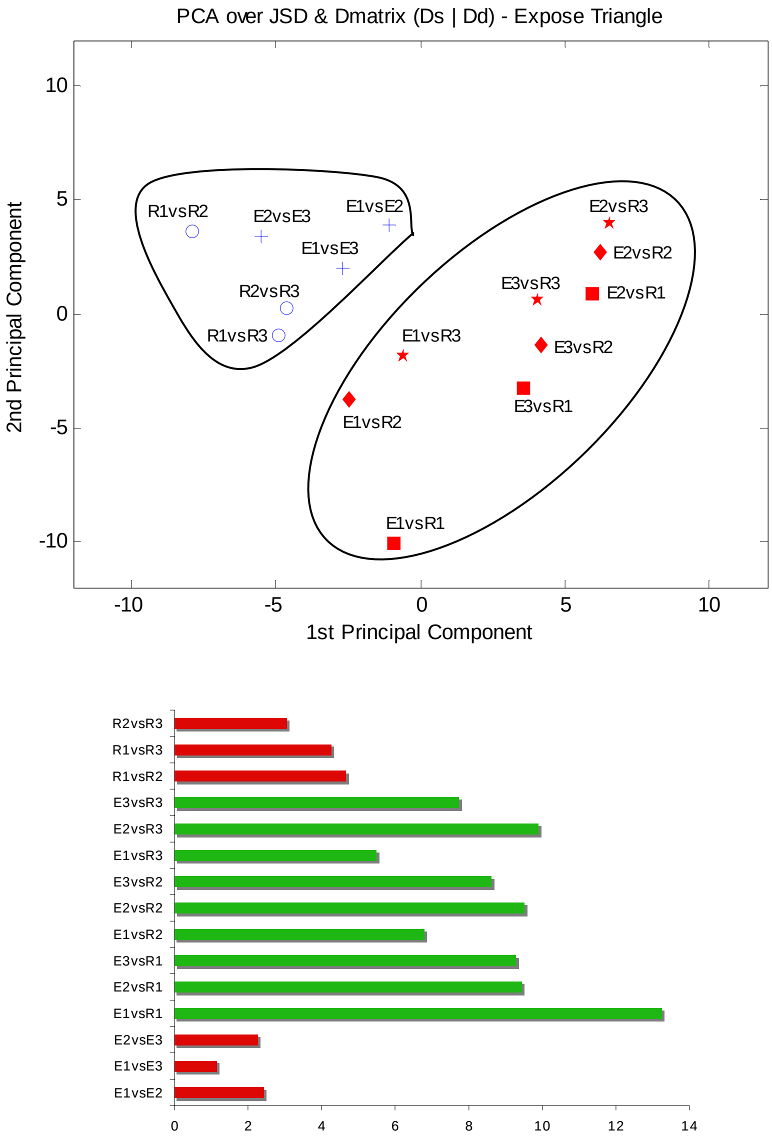


Figure 5.22: PCA analysis for 'Expose Triangle' at the action level

[5] Distances were computed in a 5D space.

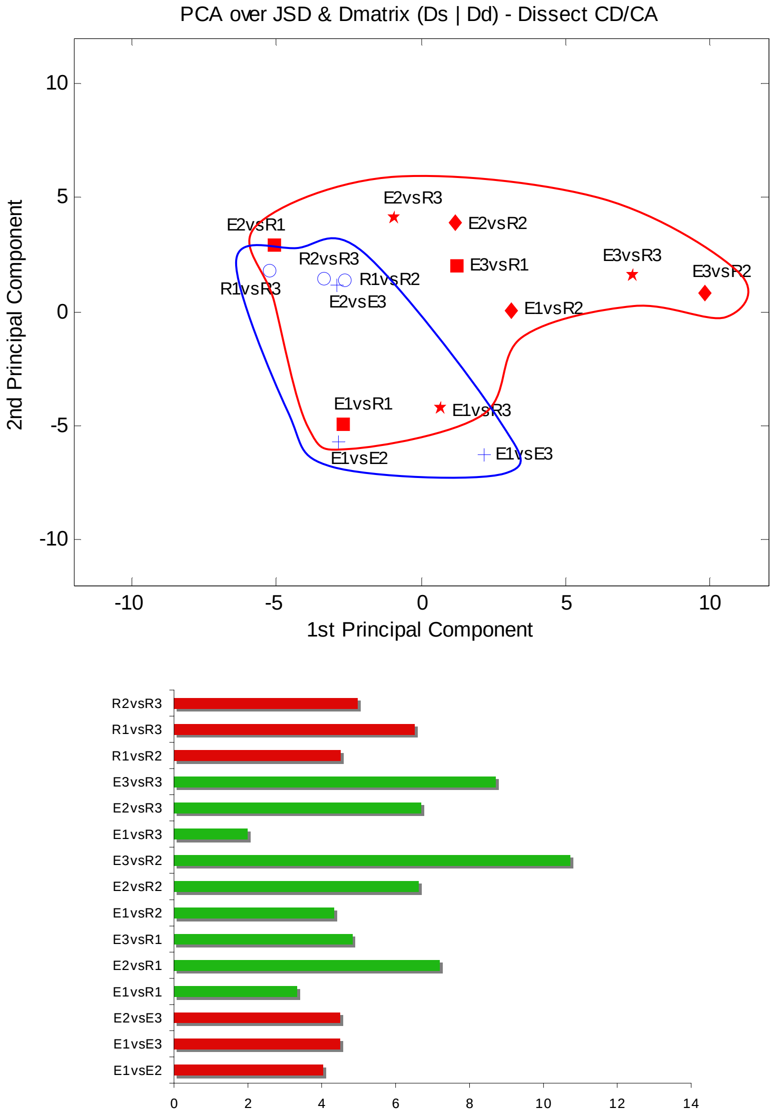


Figure 5.23: PCA analysis for 'Dissect CD/CA' at the action level

The Mann-Whitney test indicated that performance during 'Expose Triangle' (p-value = 0.001) clearly differentiates the set of comparisons among peers (i.e., Expert-to-Expert or Resident-to-Resident) from comparisons among subjects belonging to different groups (i.e., Expert-to-Resident), but not for the 'Dissect CD/CA' subtask (p-value = 0.35). This is somewhat different from the PCA analysis at the subtask level where the test is significant for separating between groups in both subtasks; however, there appears to be some intermingling of subjects from the two groups in the 'Dissect CD/CA' subtask, which could be related with the lack of group differentiation found at the action level in the single measure comparisons.

5.3.3 Semi-Markov Modelling of Surgical Motor Performance

Following the theoretical arguments provided in Section 3.4.2 for choosing a semi-Markov approach to model action transitions and average times to complete individual actions, we first demonstrate that exponential distributions do not appropriately model state holding times in the operative setting.

Figure 5.24 shows an example of comparing the empirical cumulative distribution function for resident R1 executing the 'Spread' action during 'Dissect CD/CA' against three parametric distributions: Exponential, Weibull, and Lognormal.

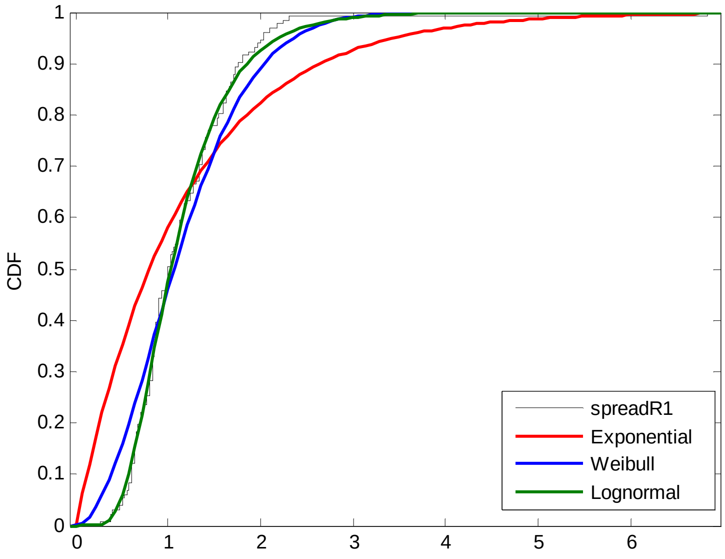


Figure 5.24: Distribution fitting example for Spread-R1 during Dissect CD/CA

Figure 5.25 presents the set of 4 actions most used by R1 during 'Dissect CD/CA' with their corresponding D values from the distribution fitting procedure explained in Section 3.4.2. The exponential distribution generated the highest D value which indicated that the time profiles for each action are more accurately represented using either a Weibull or Lognormal characteristic. In our model implementation, we decided to select the distribution with the lowest D value amongst the three to represent holding time parameters.

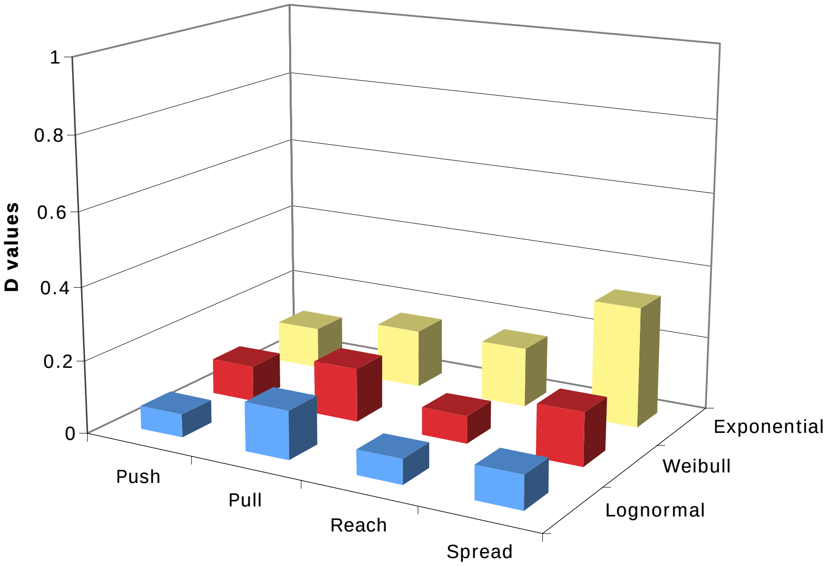


Figure 5.25: D values for the distribution fitting of holding time profiles of the set of 4 actions most used by resident R1 during Dissect CD/CA

Based on the previous selection, we built the corresponding Semi-Markov models for each subject / each subtask (Appendix D.3) by computing the movement transition probabilities and the confidence intervals for both model parameters according to the methods outlined in Section 5.2.3.2.

Figure 5.26 and **Figure 5.27** show the corresponding Semi-Markov models for subjects E1 and R1 for each individual subtask. This representation only includes transitions (represented by arcs) with confidence values less than the computed transition probability (P_{ij}) value as the significant ones and reveals that E1 effectively only used four actions, while R1 required seven movements to complete the same subtask.

The average time to execute a specific action can be computed from the parameters of either a Lognormal or a Weibull distribution as shown in Equations 4.3 and 4.4 and this average time is included along with the transition probability as the state parameters in the SMM representations.

$$t = \exp(m) \qquad \text{m:log location (Lognormal)} \quad \text{Eq. 4.3}$$

$$t = a*\text{gamma}(1+1/b) \quad \text{a: scale; b: shape (Weibull)} \quad \text{Eq. 4.4}$$

‘Expose Triangle’ SMM models for dominant hand (P_{ij} & ht):

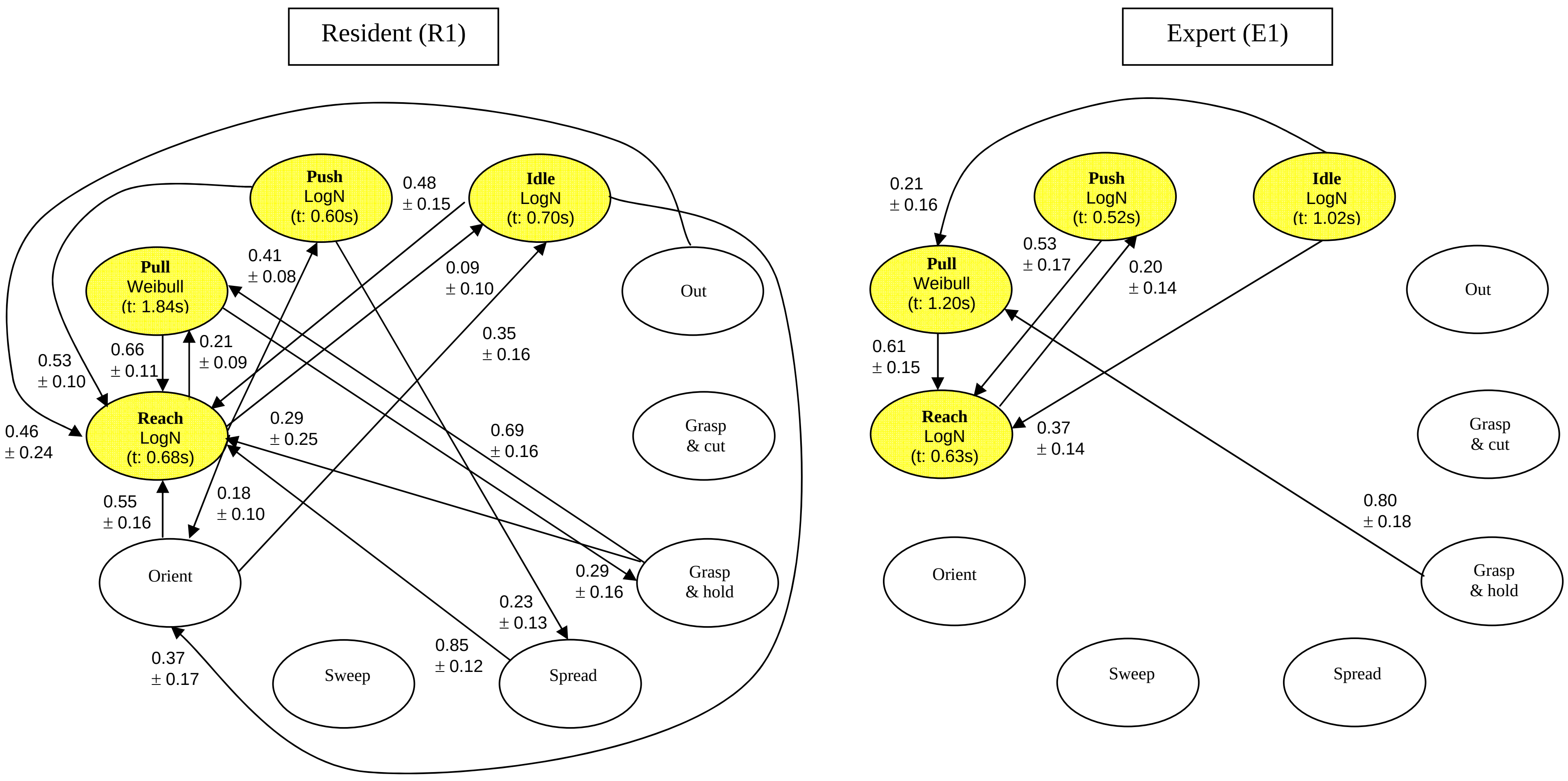


Figure 5.26: Semi-Markov models for subjects E1 and R1 during ‘Expose Triangle’ (models include transitions (represented by arcs) with confidence values less than the computed transition probability (P_{ij}) value)

‘Dissect CD/CA’ SMM models for dominant hand (P_{ij} & ht):

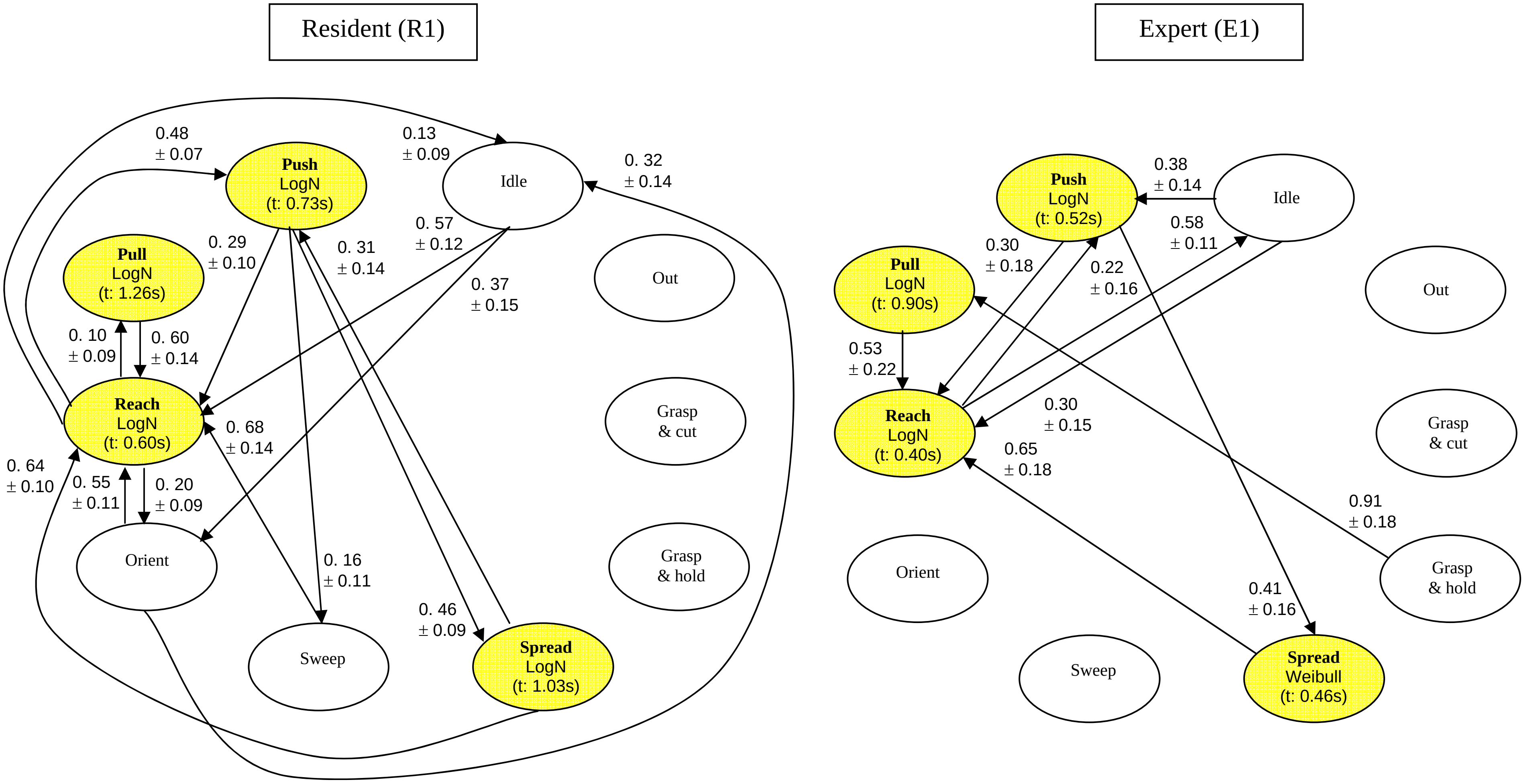


Figure 5.27: Semi-Markov models for subjects E1 and R1 during ‘Dissect CD/CA’ (models include transitions (represented by arcs) with confidence values less than the computed transition probability (P_{ij}) value)

5.3.4 Representing Bimanual Dexterity

Since we monitored movements from both dominant and non-dominant hands, it was also of interest to explore whether measures of bimanual coordination can differentiate amongst surgical skill levels. **Figure 5.28** presents cross-plots of right hand (x-axis) vs. left hand (y-axis) speed data for one selected case for each expert and each resident included in the study.

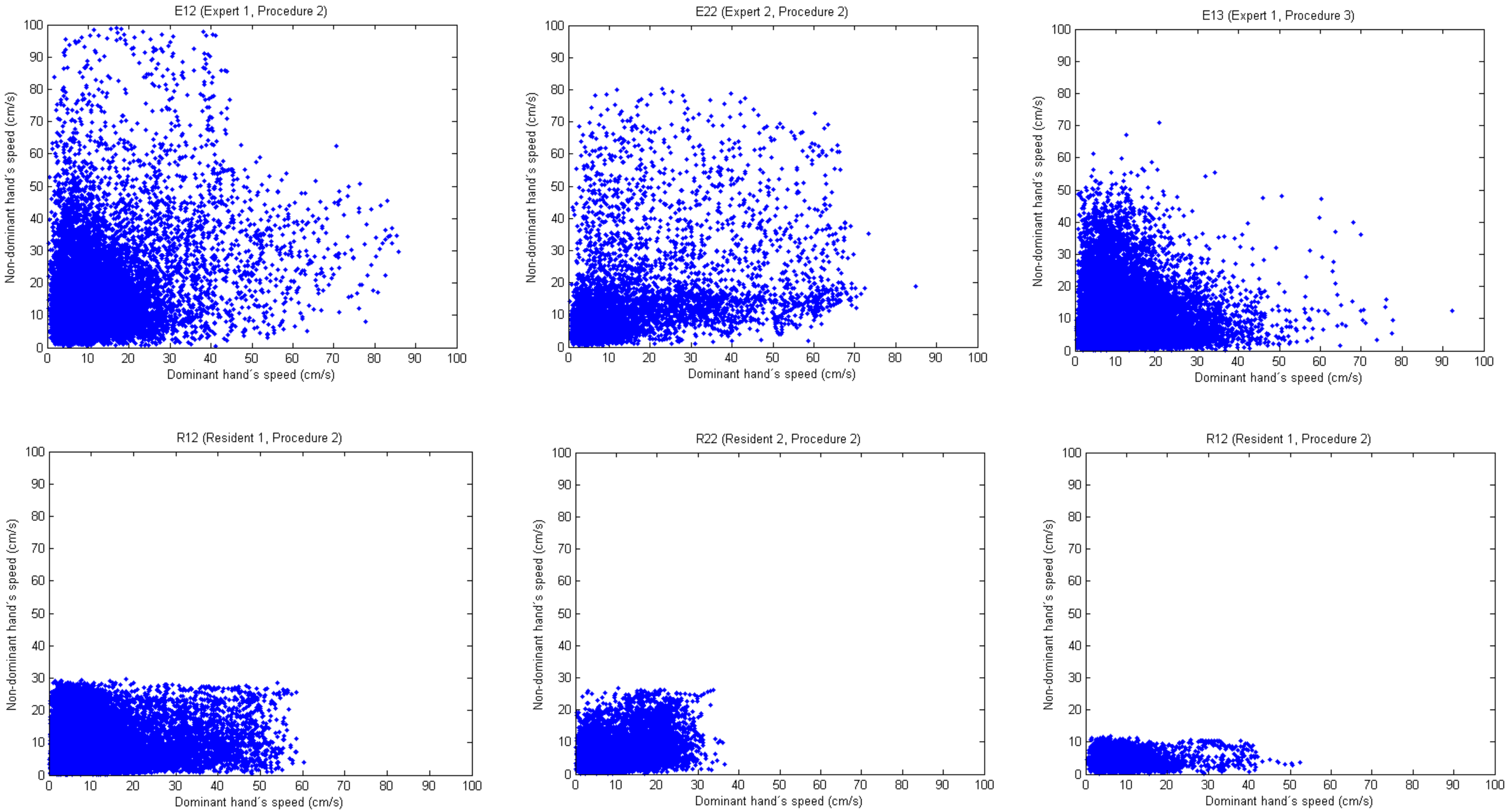


Figure 5.28: Cross-plots of speed data for right hand vs. left hand movements during ‘Expose Triangle’ subtask. Top row includes selected cases for each of the three expert surgeons and bottom row includes the corresponding ones for residents.

A visual inspection seems to indicate that while experts tend to move both hands at a similar speed, residents tend to move the right hand (being the dominant hand for all subjects) more quickly than the left hand. In order to quantitatively assess bimanual coordination, we first derived kinematics profiles from dominant and non-dominant tool movements and computed the mutual information of the two distributions as a measure of the degree of dependency between them.

Figure 5.29 shows the reported mutual information values for the 18-recorded procedures. Values close to zero indicate that the two distributions are independent.

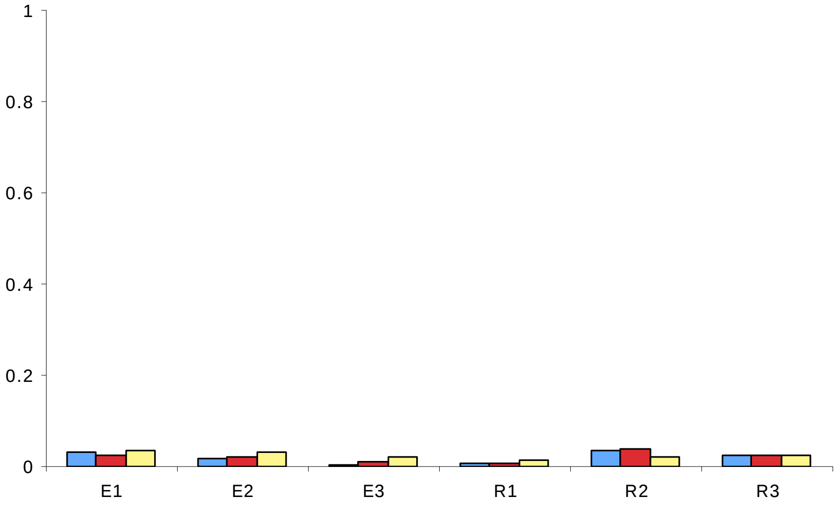
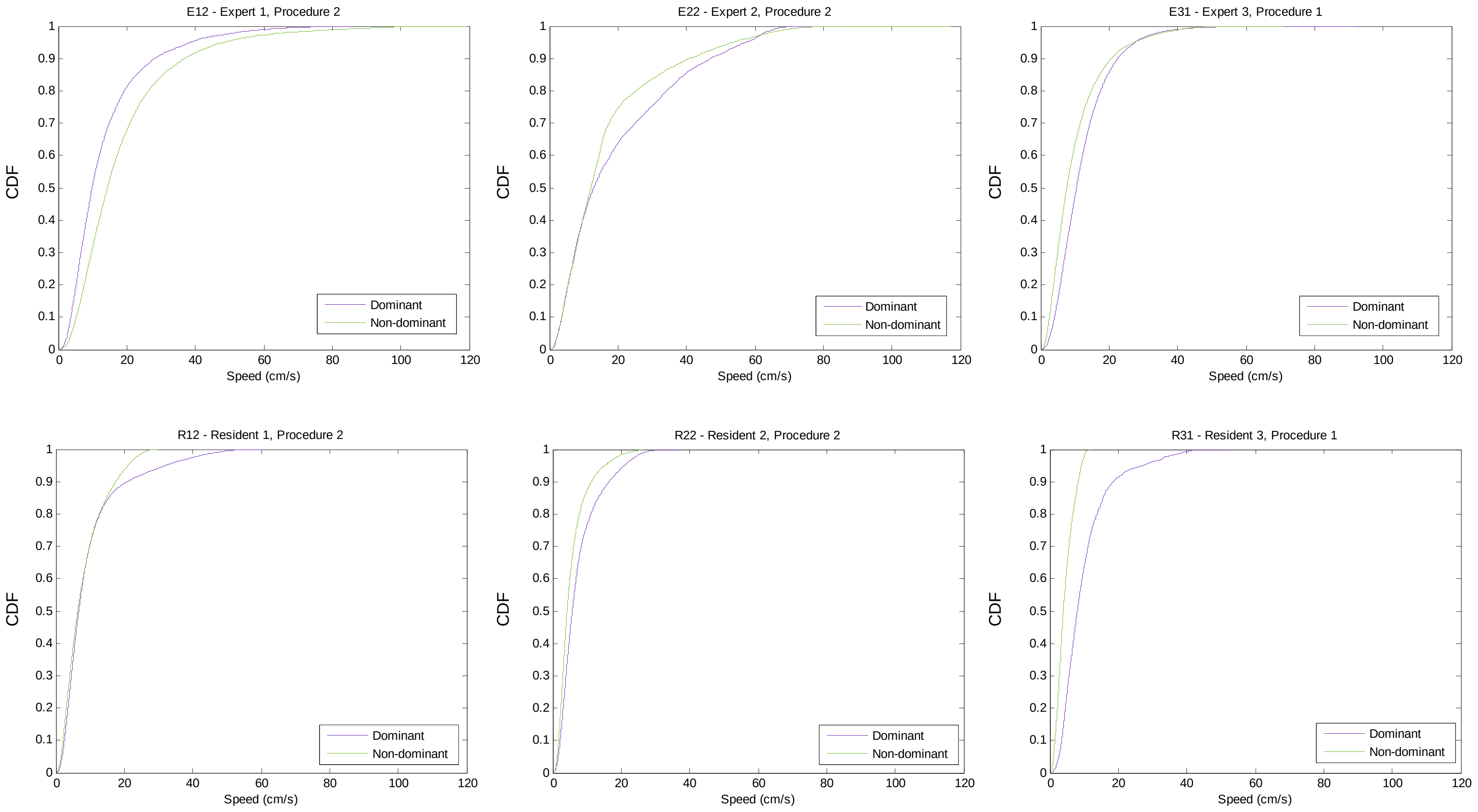


Figure 5.29: Mutual Information between dominant and non-dominant speed distributions during 'Expose Triangle' subtask across three procedures per subject (E: Experts; R: Residents)

The mutual information values indicated almost no dependence between dominant and non-dominant speed distributions. In addition, a Mann-Whitney test comparing the mutual information values for all subjects showed that this measure did not provide significant group differentiation between expert and residents (p-value = 0.45) in terms of the degree of dependency between both hands.

A second measure of bimanual coordination was applied which is based on comparing directly how similar the two speed distributions are by using the Kolgomorov-Smirnov (KS) statistic. In the same way as for mutual information, values close to zero in the KS computation indicate similarity. **Figure 5.30** presents the cumulative distribution functions for dominant and non-dominant hand movements for the selected cases of **Figure 5.28** and **Figure 5.31** show the corresponding values from the KS computation which represent the degree of asymmetry in use of the dominant and the non-dominant hands.


E12 - Expert 1, Procedure 2

E22 - Expert 2, Procedure 2

E31 - Expert 3, Procedure 1

R12 - Resident 1, Procedure 2

R22 - Resident 2, Procedure 2

R31 - Resident 3, Procedure 1

Figure 5.30: Cumulative distribution functions for dominant and non-dominant speed distributions during 'Expose Triangle' subtask. Top row includes selected cases for each of the three experts and bottom row includes selected cases for each of the three residents

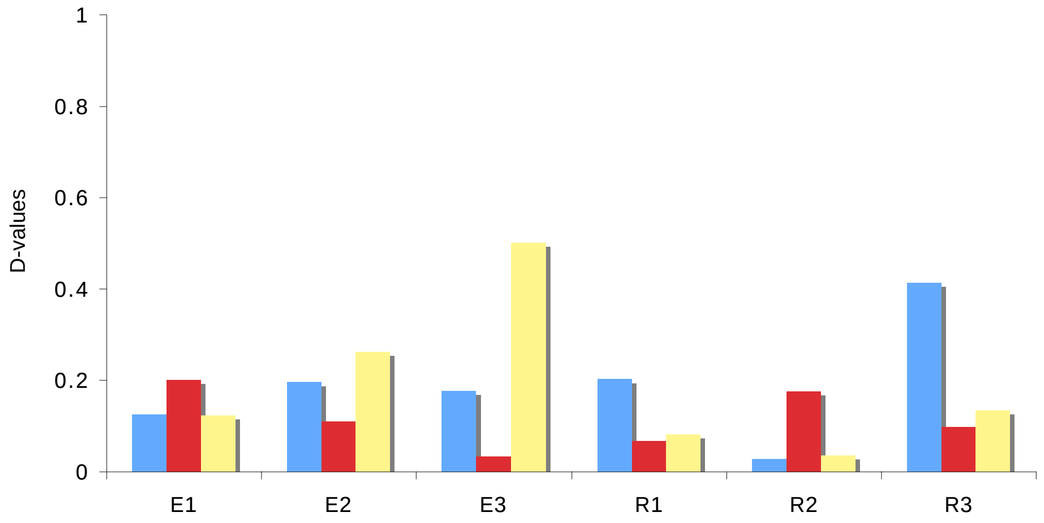


Figure 5.31: Kolgomorov-Smirnov statistic (D values) from comparing dominant and non-dominant speed distributions during 'Expose Triangle' subtask across three procedures per subject (E: Experts; R: Residents) (blue: procedure #1, red: procedure #2, yellow: procedure #3)

Except for one procedure from E3 (procedure #3) and one from R3 (procedure #1), D values for all subjects indicated similarity between dominant and non-dominant speed distributions, which did not support the initial visual inspection from **Figure 5.28**. We believe that this is due to the much higher concentration of data points around the lower left corner relative to the right upper corner of the cross-plots; in effect, even if there are differences, there are so few occasions when these differences occur that they do not achieve statistical significance. In addition, the Mann-Whitney test comparing the computed D values did not show significant differentiation between groups (p-value = 0.23).

5.4 Summary

The primary purpose of this study was to evaluate whether our proposed quantitative assessment methodology based on three performance measures (time, kinematics and movement transitioning of the surgical tool tip) was able to quantify motor aspects of live surgical performance and to use these measures to distinguish between trainees at different levels of development when working in the operating room environment. We acquired intraoperative data from two sets of subjects representing the two extreme ends of the surgical skill spectrum: Residents and Experts, and selected the laparoscopic cholecystectomy as our baseline procedure for study, as it is the most commonly and well-defined minimally invasive procedure [Tendick 2000]. The position of the tool tip was recorded by using an electromagnetic measurement system because it does not suffer from line-of-sight problems, which allows for continuous data recording and velocity data was then derived by differentiation. A laparoscopic camera was also used to record the intrabdominal view of the surgery and a video analysis was performed to identify the times and movement patterns from the discrete phases of the procedure execution.

Using our MCMD approach from Chapter 2, 'Expose Triangle' and 'Dissect CD' were identified by the expert surgeons as the most demanding steps of the procedure in terms of the surgical dexterity required; therefore, in the present study we focused on analyzing performance during these two subtasks. The Curved and L-Hook dissectors were the tools of choice for the dominant hand, while an atraumatic grasper was used for the non-dominant hand. However, due to the susceptibility of the electromagnetic sensor to the electrical noise produced by the L-Hook dissector during cautery, the expert surgeons involved in this study agreed to limit themselves to using the curved dissector for the dominant hand. Fiber optic position sensors may be sufficiently accurate and sufficiently resistant to the interference caused by cautery to allow them to be used for future studies of this type using an L-Hook dissector.

At the subtask level we derived multi-element vectors consisting of the average execution time, the average tooltip velocities, and detailed samples from the velocity profiles every 5th percentile in each of the three cardinal directions (lateral, axial, vertical) for each of the two subtasks (‘Expose Triangle’ and ‘Dissect CD/CA’). We then used Principal Components Analysis (PCA) to extract the dominant contributors to overall variability and computed the ratio of mean square distance (MSD) in the PCA weight space from the mean position of all procedures executed by a specific subject or group to the MSD from the global mean position to describe variability for specific subjects and groups.

We found that separation between groups was significant for both subtasks while performing with the dominant hand and that the intergroup mean square (MS) was higher in the ‘Expose Triangle’ subtask than in the ‘Dissect CD/CA’ subtask. Analysis of the non-dominant hand indicated that separation between groups was not significant for any case.

Additionally, in an attempt to explore whether measures of bimanual coordination could differentiate amongst surgical skill levels, we reported the mutual information and the Kolgomorov-Smirnov statistic between the kinematics distributions from dominant and non-dominant tool movements; however, tests of significance indicated that neither of these two measures provided significant group differentiation between expert and residents. A recent study on bimanual coordination for simulated surgical tasks (e.g., transferring pegs) used a similar approach based on the concept of the phase portrait and hypothesized that they would find a significant difference whenever there is out of phase movement, which implies high velocity of one tool when the velocity of the other tool is low [Narazaki 2007]. However, the differences they report were small and only present in two of the tasks, and the standard deviations across populations were large relative to the purported intergroup differences. Small differences will make it difficult to track development in bimanual skill in one subject or even to reliably differentiate within a group. There is little evidence to date for large (or even any) bimanual coordination differences between groups, though Narazaki's approach does suggest that some small difference may exist.

At the action level, we compared performances of single subtasks by decomposing them into their characteristic actions as a set of 10 elemental tool tip motions: push, pull, reach, orient, sweep, spread, grasp&hold, grasp&cut, idle, out, represented as Semi-Markov models.

Using video analysis and the previously defined start and end points, we derived a list of action transitions with the corresponding time spent at a specific action before transitioning to another (i.e, holding times). The time records were then used to segment the kinematics signal of the subtask into the kinematics data for each action. We then characterized every action using distributions of holding times and kinematics and used the Kolgomorov-Smirnov statistic (D) to measure differences in these two parameters. In addition, we computed transition probability matrices and used the Jensen-Shanon divergence (JSD) to define a third difference measure. We found that kinematics measures produced difference values for intergroup comparisons which were significantly greater than the average expert-to-expert comparison level, though the degree of difference or separation was considerably less than we found in the physical simulation. We believe that the significant variability in the complexity of cases studied likely contributed to this interprocedural variability, which makes it more challenging to assess differences in skill level based on intraoperatively-acquired data.

All JSD and D values for every action were then grouped into a multi-element matrix and a PCA analysis was then performed to test the hypothesis that in this weight space our difference measures were able to provide skill level separation when comparing subjects from the same group (e.g., experts vs. experts) and subjects from different groups (e.g., novices vs. experts). At the action level, we found that data from "Expose Triangle" subtask clearly differentiates the set of comparisons amongst peers from comparisons amongst subjects belonging to different groups, though data from the "Dissect CD/CA" subtask does not. This is somewhat different from the PCA analysis at the subtask level where the test is significant for separating between groups in both subtasks. This suggests that low level behaviour may be more subject to individual patient differences, while higher live movement characteristics may be more reliable.

Previous studies using motion analysis-based systems to track surgeons' hand movements have shown significant differences (for time taken, total path length, and number of movements) between two groups of surgeons while performing laparoscopic cholecystectomies [Aggarwal 2007, Datta 2006]. We regard this approach as one of the most promising for the future of surgical skill assessment; however, we also believe that some issues that we addressed in the present study need to be included before the system will be useful for instruction.

Darzi’s approach considers major parts of the entire surgical procedure (e.g., Calot’s Triangle dissection) without further task decomposition. Using the MCMD we have developed to isolate selected surgical tasks, rather than looking at an undifferentiated stream of data, will facilitate: (1) identifying causes of deviations in the normal path due to individual surgeon’s decisions, (2) describing how the surgical tools are actually used following a standardized and structured framework of the procedure, and (3) taking into account the operative variability by allowing for variable weighting on different tasks during a surgical procedure to reflect differences in importance, difficulty or relevance for the current level of surgical training. Besides the parameters used by [Aggarwal 2007, Datta 2006], our methodology also demonstrated the feasibility of acquiring and using the kinematics (velocity, acceleration and jerk profiles) of the surgical tools for differentiating among skill groups when performing at the operating room.

While it is clear that low scores on particular simulated tasks suggest that more practice might be required, there has been virtually no work done on using intraoperatively-acquired data to identify suboptimal performance [Feldman 2004]. In our approach, differences in performance for individual portions of the procedures were described in terms of intuitive scores (i.e., 0: similar; 1: different), which would facilitate providing specific and relevant feedback to trainees concerning areas in which improvement is needed. In addition, simultaneous analysis of multiple measures by means of a dimensionality reduction technique (i.e., PCA) proved to be useful and practical in determining intrasubject, intragroup, and intergroup variabilities.

Furthermore, when comparing the results from the two experimental studies we implemented, it was also evident that differences identified in a simulated experiment seem not to be completely transferable to the operating room scenario, where the conditions are not as controllable as in the simulation. While we found significant differences among groups for all the tasks considered in the simulator, in the live surgeries, performances of trainer and trainees diverged only for one of the surgical tasks.

Chapter 6

Discussion & Future Work

6.1 Introduction

The goal of the research presented in this thesis was to establish a methodology for quantifying performance of surgeons and distinguishing skill levels during live surgeries. We have integrated three types of physical measures (kinematics, time and movement transitioning) into a modelling technique for quantifying performance of surgical trainees. We first created a hierarchical representation to decompose larger surgical goals into clearly identifiable tasks amenable to being monitored by our measures. Then, at each level of surgical complexity, we implemented specific mathematical techniques to derive intuitive scores for providing a quantitative measure of how far a performance is located from a reference level (e.g., a group of expert surgeons). To show the reliability of the established performance parameters, we also implemented various statistical methods to measure repeatability across subjects and groups.

Two experimental studies were completed in order to show the feasibility of our proposed assessment methodology: (1) performance in a physical surgical simulator, and (2) in the operating room. We therefore concentrated on answering the following specific research questions:

1. Can quantitative measures acquired intraoperatively reliably characterize motor performance?
2. Do surgeons at similar stages of training exhibit similar psychomotor patterns?
3. Is there a clear separation of patterns between the extremes of the training spectrum?
4. What data/measures are most useful in separating surgeons along this spectrum?
5. Can a quantitative analysis produce insights useful for instruction?

For the simulator scenario, the task was to dissect 2-3 mandarin oranges; this was performed by three groups of subjects representing three different skill levels. We applied our proposed assessment methodology and evaluated if (1) intrasubject repeatability was good, (2) scores for trainees with similar skill levels were similar, and (3) scores for trainees at different stages were significantly different. We presumed that if these conditions were met, the technique would be worth testing in the live operative setting.

In a second stage, we moved into the less controlled environment of real surgical procedures. For this second experiment, we monitored movements of a curved dissector and an atraumatic grasper during 18 laparoscopic cholecystectomy procedures performed by two sets of three surgeons – one set of residents and one of attending surgeons. From the tools' positions, we extracted various performance measures and applied our methodology to compare residents and expert surgeons executing two key surgical tasks: exposing Calot's Triangle and dissecting the cystic duct and artery (CD/CA).

Results from these two studies demonstrated the ability of our methodology to differentiate skill levels and we therefore plan to use this system in future studies for the purpose of measuring motor performance both in simulators and in the operating room, and for developing a database of performance measures of surgeons at various skill levels for reference purposes.

6.2 Review of Present Research

In this section, we will summarize the main results and conclusions derived from answering the proposed research questions through our two experimental studies.

6.2.1 Experience With Data Acquisition System

The data acquisition system proved to be an efficient set up to be used in the operating room (OR) where space and time constraints must be dealt with. The computer station, which was the only additional element to the normal OR set up, occupied only 60x60cm^2. The custom-designed clip for attaching the sensor to the tool was found to be practical for the surgeons since it only required a two minute calibration process before each procedure and eliminated the problem of unbalanced loading experienced with the instrumented surgical tool used in previous studies in our lab [Kinnaird, 2004]. In general, the surgeons expressed no significant concerns with our data acquisition set up except that they were not able to use the L-Hook dissector due to the interference introduced by the cautery.

Other issues we had to deal with during the data acquisition phase of this study included the introduction of new operating room scheduling policies which reduced the OR time available for the attending surgeons participating in the study, complicated procedure scheduling and extended the period of data collection (the time needed to collect the minimum of 18 procedures we required). In addition, the video segmentation process was exhausting and time consuming since it was performed manually; each one-hour procedure required approximately 8-10 hours of segmentation work.

6.2.2 Proposed Performance Assessment Methodology

Our assessment methodology starts by defining a new hierarchical representation (MCMD) for laparoscopic procedures, which decomposes larger surgical goals (tasks) into local goals (subtasks) and at the very detailed level into individual movements (actions). To our knowledge, this is the first performance assessment study to include a explicit cognitive and motor diagrammatic representation that enables the investigator to take account of the operative variability; most previous intraoperative assessments are conducted at the 'whole procedure' level and do not distinguish between performance of more or less challenging elements of the overall procedure.

The proposed methodology proved to be feasible for differentiating surgical skill levels, however it is not yet practical in terms of the data handling as the video segmentation was performed manually. Therefore more effort would be needed before it could potentially be integrated in day-to-day applications to provide feedback in real time.

6.2.3 Physical Simulator (Mandarin) Experiment

The primary purpose of the surgical simulated experiment was to test whether or not our proposed analytical method was able to reliably distinguish between three groups of subjects representing different stages of training: novices, novices following explicit instruction, and experts. We simulated a surgical dissection task by asking participants to use laparoscopic tools to peel and separate the segments of two to three mandarin oranges placed in a training box. The movements of the laparoscopic tool for the dominant hand were tracked while the task was being executed.

At the subtask level, we constructed six-element vectors consisting of the tooltip average velocities and used Principal Components Analysis to reduce the dimensionality of this data to 2. In the resulting 2D weight space, we showed that we could readily differentiate between different technical proficiency levels and an analysis of the PCA eigenvectors suggested that velocity information was a more significant contributor to making distinctions between groups than time. The low values of intrasubject (7%) and intragroup (24%) variability indicated that the greatest contributor to overall variability was difference in degree of training, which is consistent with the idea that level of training should be visible in the movement patterns.

At the action level, we applied PCA to a high-dimensional (64x36) data set based on difference measures extracted in each of the nine fundamental surgical actions studied, as well as on a difference measure based on the transition probability matrix. We again found that we could clearly distinguish between skill levels. Analysis of the PC coefficients indicated that difference measures related to tool tip velocities provided the greatest degree of differentiation between skill levels.

In order to answer our research questions, we found that the PCA technique applied over the three performance measures used in this study (time, tool kinematics, movement transitions) allowed for representing and grouping subjects according to the technical proficiency levels, and for using the concept of distance to measure group membership. Moreover the PCA technique suggested that tool kinematics perform better than the other two measures in differentiating subjects' performances at the subtask level. The variability analysis indicated that intrasubject repeatability was generally high, that the data from subjects of comparable training level was in relatively close proximity to one another, and that there were significant variations between groups, which showed separation of skill levels between the extremes of the training spectrum. These findings comprise the most important early test of a proposed assessment technique.

6.2.4 Operating Room Study

The simulator study established the feasibility of using our proposed methodology to differentiate amongst different skill levels in a simulated setting; this justified testing our approach in the operating room environment. We therefore acquired intraoperative data from two sets of subjects representing the two extreme stages of training: Residents and Experts. We selected the laparoscopic cholecystectomy as our baseline procedure for study as it is one of the most commonly performed and studied minimally invasive procedures. It is also one of the first procedures a resident surgeon learns to perform so it enables us to study surgeons in training at the earliest possible point in their training.

Using our MCMD approach, the 'Expose Triangle' and the 'Dissect CD' subtasks were identified by the expert surgeons as the most demanding steps of the procedure in terms of the surgical dexterity required; therefore, in the OR study we focused on analyzing performance during these two subtasks. The surgeons recommended the curved dissector and the atraumatic grasper as the tools of choice for the dominant hand and non-dominant hand, respectively.

In answering our research questions, the same methodology used in the simulation study was applied for analyzing performances at the subtask and action levels. An ANOVA test at the subtask level indicated that intergroup differences were significant for both subtasks when monitoring the dominant hand. In contrast, at the action level, we found that any intergroup differences in performance during the 'Dissect CD/CA' subtask did not reach statistical significance, while the 'Expose Triangle' subtask did exhibit significant differences.

Analysis of the relative contributions of time, kinematic and transition probability measures to the dominant eigenvectors in the PCA analysis showed that kinematic measures provided the strongest differentiation between groups, though the degree of difference or separation was considerably less than we found in the physical simulation. We found little evidence for any bimanual coordination differences between groups; this is consistent with a recent study which showed that differences which might exist in certain simulated surgical tasks such as transferring pegs are comparatively small [Narazaki, 2007]. Moreover, in our approach, differences in performance for individual portions of the procedures were described in terms of intuitive scores (i.e., 0: similar; 1: different), which would facilitate providing specific and relevant feedback to trainees concerning areas in which improvement is needed.

Taken together, these observations suggest that there is good potential for discriminating between skill levels in both simulated and live operative settings, although the existing performance measures may not yet be sufficiently sensitive to enable fine discrimination across the spectrum of skill development. Further measures related to tissue-handling behaviours (eg, forces applied) and a more direct assessment of the quality of the results of the tissue interactions may well be required to achieve finer discrimination.

6.3 Contributions

This thesis describes several important contributions:

1. **Design of a New Graphical Language For Describing Surgical Flow:** By combining two task analysis techniques (Hierachical Analysis and Information Processing Analysis) we developed our motor and cognitive modelling diagram (MCMD). This new graphical representation of surgical procedures, which includes conventional symbols from flowchart diagrams and Boolean logic diagrams, together with new symbols needed to describe surgical events, enables us to model both motor and cognitive aspects of surgery in a unified diagram. Using the MCMD, we can record and analyze differences in surgical sequences selected by surgeons during different procedures.

2. **Design of a Hierarchical Framework for Representing Quantitative Data in Context and Performing Similarity Analysis Between Subjects and Groups:** Our motor and cognitive modelling diagram (MCMD) enables us to combine quantitative performance measures on a sample-by-sample basis with information related to the flow of the surgical procedure and to organize it in a hierarchical form. This provides a mechanism for concentrating the trainee's and the trainer's attention on key

elements of the procedure where differences in performance might occur and enables the analysis to take account of inter-procedural variability.

3. **Demonstration that Quantitative Performance Measures Can Differentiate Between Skill Levels in Both Simulated and Live Settings:** Our methodology also demonstrated the feasibility of acquiring and using measures of surgical tool motions to differentiate amongst surgeons with different skill levels when performing in the operating room. Differences in performance for individual portions of the procedures were described in terms of intuitive scores and simultaneous analysis of multiple measures by means of a dimensionality reduction technique (i.e., PCA) proved to be useful and practical in determining intrasubject, intragroup, and intergroup variabilities.

4. **Demonstration of Differences in Sensitivity Between Simulators and the Live Operative Environment:** While the results from the two experimental studies showed that our approach could distinguish between skill levels in both simulated and live surgical settings, the discrimination seemed stronger in the simulated setting, likely due to the greater interprocedural variations in the operating room. Our technique showed the potential for semi-automatically identifying which combinations of performance measures offer the most discrimination between subjects and groups in the OR setting, which provides guidance for choosing metrics to be obtained in simulators and focuses attention on the most critical performance measures to be evaluated when investigating transference of skill between simulators and the OR

6.4 Limitations

The following limitations were encountered during the course of the present thesis:

- Sample size – This study was designed primarily as an evaluation of the feasibility of using a new representational structure for analyzing surgeries. We therefore studied a relatively small number of subjects, which prevented us from making any claim that we have obtained a reasonable representation of the range of the population the sample has been drawn from; we would not be surprised to find future samples that lie considerably outside the range initially found. We are therefore not able to make general claims about these populations (ie, of the form 'expert surgeons behave this way').

- Type of tracking system – We opted to use an electromagnetic tracking system because of its high update rate, reasonable accuracy, and low profile, which made it relatively easy to attach to existing surgical instruments. However, because it is based on sensing electromagnetic fields, it is susceptible to the electrical noise produced by cautery tools, which prevented us from monitoring some of the tools (e.g., L-Hook dissector) commonly used by surgeons.

  In addition, the impact of the accuracy of the tracking system could not be directly evaluated - magnetic trackers are susceptible to distortion when in proximity with metal objects and we had no 'gold standard' against which to evaluate absolute accuracy in the operating room (simulators can place 'virtual target points', but these are not available in the OR). However, since most of the performance measures we used were not position-dependent, but based more on position derivatives, we expect that absolute position errors should have little impact on our conclusions.

In terms of the coordinate frame, it is also important to note that our current convention is orientation-independent and therefore free of orientation errors that might be introduced if we used a body-oriented frame.

- Kinematics measurement only – In this study, we restricted ourselves to recording kinematic measures. Force measures were not included in the present study due to the complexity of designing and building an instrument that could be used in the OR without hampering the execution of a procedure (our group does have prior experience with such instruments [Kinnaird 2004], but concluded that the current instruments were unsuitable for this study). Therefore, an improvement in instrumentation is required before we are able to obtain force measurements and include them in our analytical framework as an additional performance measure.

  In addition, we did not directly assess the quality of the surgical tasks and subtasks, nor did we track surgical outcome measures. This prevented us from establishing relationships between development of motor skills and surgical outcome. Quality could potentially be assessed by implementing image-processing techniques for analyzing the videos of specific steps in the procedure (e.g., to determine the extent of burnt area when detaching the gallbladder from the liver bed). However since there are significant technical challenges involved in doing this automatically, a reasonable first step in future studies would be to incorporate manual assessment by means of rating scales and expert surgeons analyzing the video.

- Only Monitoring Curved Dissector – As discussed above, the incompatibility of cautery and our electromagnetic tracker meant that we were restricted to assessing surgical tasks that could be performed using the curved dissector, rather than an L-hook cauterizing dissector (ie, the ‘Expose triangle’ and ‘Dissect CD/CA’ tasks, but not the ‘Dissecting GB from the liver bed’ task, which is normally performed using the L-Hook dissector). This issue also prevented us from being able to monitor preferred surgical practice in a situation where the L-hook dissector was the preferred tool.

- Residents not included in simulator study – We did not include residents in the simulator study and so were not able to determine whether differentiation between skill levels in the operating room is greater or less than in the simulator. Our original intent was to test feasibility of distinguishing skill levels in the simulator, so we chose people at the extreme ends of the skill spectrum. We acknowledge that it would have been useful to include residents as part of the simulator study, but since this was intended to be a rapid proof-of-concept test and there would likely have been delays introduced due to the more complex resident recruitment protocols required, we decided to perform this study with graduate students serving as the novice and novice-with-training groups.

- Type of simulation task – Given the physical nature of the mandarin dissection task, we were able to obtain actual tool movement data for testing the feasibility of our methods. However, this task was not designed to directly correspond to any particular surgical task, but was chosen simply to require related surgical skills. We therefore cannot directly evaluate whether or not the simulator task is a realistic approximation of a target live surgical task.

- Manual video segmentation - The video segmentation process was exhausting and time consuming since it was performed manually as automatic tools are still not available; each one-hour procedure required approximately 8-10 hours of segmentation work.

6.5 Related Work

In developing systems for training and assessing surgical skills, four elements need to be addressed before the final implementation of those systems into the surgical curriculum: (1) defining and standardizing the performance metrics, (2) differentiating amongst skill levels, (3) providing effective feedback, and (4) deciding upon the outcome measures to be achieved [Satava, 2004].

At present, testing technical skills has been performed using simulators, animal and human operating rooms, and have included objective and subjective methods [Britt 2007, Aggarwal 2004, Moorthy 2002]. Objective or quantitative methods have been mostly developed and used in simulator contexts, with only two approaches currently tested in the operating room: the Imperial College Surgical Assessment Device (ICSAD) and the BlueDRAGON [Rosen 2006, Aggarwal 2007, Datta 2006, Dosis 2005, Darzi 2001].

Both approaches have mainly concentrated on differentiating between skill levels by using distinct performance measures; however the issue of providing specific feedback by decomposing and analyzing individual portions of the procedure has not been addressed yet. Darzi pointed out that the reason why they could not find differences in all tasks was possibly because each surgeon used the surgical technique they are most confident with, and therefore he claims there is a need for developing tools for representing different surgical techniques (i.e., representing flow of procedure and types of surgical tools used) [Datta 2006].

In addition to differentiating skill levels in both the simulator and the operating room settings, we believe that our approach complements previous work such as ICSAD and BlueDRAGON by addressing the issue of providing feedback that is of value in training and evaluating surgeons. Our MCMD enables us to identify specific points in the procedure where differences happen and to track them down in the hierarchy to the level where surgical tool movements (e.g., push, pull, reach) are described. In this way, a trainee is able to identify if his/her performance is different from that of an expert because they chose a different path (at the task and subtask levels) or used different tool movements (at the action level).

Moreover, previous work in simulators and in the OR have so far performed individual analysis for each performance measure with no intuitive scores that would help to facilitate interpretation by the surgical trainer [Rosen 2006, Aggarwal 2007, Datta 2006, Dosis 2005, Darzi 2001, Gallagher 2004, Fried 1999, Scott 2000]. Thus, the PCA feature of our framework developed for implementing simultaneous analysis of multiple measures has also complemented previous studies by enabling us to automatically identify those combinations of performance metrics which provide the greatest ability to discriminate between skill levels.

We therefore believe that our framework has gone beyond differentiating amongst skill levels as demonstrated by other approaches, and has contributed significantly to the development of systems for training and assessing surgical skills by providing (1) a tool for graphically representing the surgical flow, and (2) an analytical scheme for including various performance measures and deciding which measures are most useful in discriminating surgical levels.

6.6 Recommendations For Future Studies

As this is a newly proposed assessment methodology, several aspects of the approach warrant further investigation to assess and maximize its overall reliability and clinical utility. To make the data acquisition system and process more practical, we recommend the following adjustments and enhancements to our setup and approach:

- Find a way to make measurements of the L-Hook dissector (cauterizing tool) to provide more flexibility in the dissection task, since this is a common and even preferred surgical option. This would require us to replace the electromagnetic tracker with a fibreoptic-based system which will be more resistant to the interference problems created by the cautery unit's operation.

- Expand the set of physical measures monitored by incorporating force sensing and scene perception (recommended by the attending surgeons) to determine if the non-dominant hand is providing appropriate traction, if the dominant hand is applying appropriate forces to the tissues, and if the anatomical structures are appropriately exposed. The force sensing is relatively challenging because the force sensor would need to be mounted near the tool tip, so the tool itself would need to be redesigned to incorporate such a sensor. The scene perception software would likely also be a challenging project.

- Design an integrated data acquisition and calibration program capable of acquiring data from multiple sensors and simultaneously registering the intra-abdominal view from the laparoscopic video system

- Develop an automatic movement segmentation method to enhance the objectivity of the assessment system and improve the data post-processing. An interesting approach has been proposed by Dr. Allison Okamura’s group at Johns Hopkins University, which uses automatic techniques based on Hidden Markov Models (HMM) for detecting and segmenting raw motion data from a surgical task to produce a labelled sequence of surgical gestures [Lin 2006, Lin 2005, Murphy 2004, Murphy 2003]. Using simulated surgical tasks executed with the daVinci system, Murphy 2004 applied Linear Discriminant Analysis (LDA) to separate the surgical motions and used statistical methods such as HMM to perform the recognition step [Murphy 2004]. Although this approach has only been tested in simulated environments, it seems to be a promising approach to segmenting motion data from live surgeries.

- Evaluate the feasibility of using a vision system for monitoring the organization of the OR – correct equipment and instrument selection, mode, and connection; convenient positioning of equipment to avoid accidents and to facilitate instrument exchange. These criteria were identified as essential by the 4 surgeons involved in this study for ensuring a safe procedure and cannot be assessed by the tool tip measures presented in this thesis.

- Develop techniques to automatically assess the quality of the various surgical steps involved in a procedure. This is a virtually unexplored area of research.

- Acquire data from both the hand and the tool to compare the results from our study and Darzi's group findings (see Chapter 1)

In the longer term, the following questions need to be addressed:

(1) How many procedures do we need to record from an individual surgeon to reduce patient/procedural variability to a nominal level?

(2) What is the minimum number of performance variables needed to get a representative measure of a surgeon's skill level?

(3) Is it possible to find better measures for reliably representing and assessing bimanual coordination in the human operating room?

(4) Can performance assessment in the simulator correlate with technical performance in the actual surgical setting?

As the main goal of the present research was to prove the feasibility of our new performance assessment methodology based on quantitative measures, we suggest a two-fold study for the next stage:

(1) Testing its reliability in differentiating skill levels in the operating room based on an increased number of participants and procedures

(2) Determining (if any) the correlation between assessment on simulators and assessment in the operating room

For the first part, we propose using the same data acquisition and processing protocol presented in this thesis to create a database of residents and expert surgeons executing multiple procedures; we would also document any characteristics of the patient which might affect the procedure's difficulty, which would potentially allow us to partition the data according to difficulty level to improve repeatability and enable us to explore the influence of case difficulty on surgical tool movement patterns. These databases would allow us to better understand the range of performance at a given skill level and the extent of shifts in performance with differing degrees of surgical experience.

As surgical simulators develop to emulate particular surgical tasks, we propose to directly compare movement patterns executed by surgeons in the operating room and on the corresponding simulated surgical task to determine if any differences detected are within the range of normal variation expected when only operative situations are examined. If so, and if the differences on the simulated tasks between groups are similar to those found in the live surgical tasks, we would have a basis for claiming that these simulations are acceptable substitutes for the live operating room for the purposes of training and evaluation.

Another interesting topic for future studies would be to evaluate skills across different types of surgeries – at the subtask level, essentially the same processes (dissection, suturing, etc.), can be performed in different surgical procedures with different task level goals. By explicitly identifying some tasks which are considered to be common across different surgical procedures, we can evaluate the hypothesis that performance by a single surgeon on common subtasks is sufficiently similar, independent of the overall surgical procedure being performed, to allow an assessment of surgical skill on that class of subtask. Addressing this issue would then require us to utilize a ‘subroutine comparison’ by means of detecting at the subtask level the common subtasks (e.g., tissue dissection) and directly comparing a subject’s motor performance measures for the corresponding subtasks independent of the type of surgery. If ranges of motor skills are similar, then monitoring performance on a given type of procedure would provide insight into how well a surgeon would likely execute the same subtask in another surgical context.

6.7 Conclusion

As a final point, we do not regard objective assessment methods as a substitute for the attending surgeon in the training process; instead, we believe they can offer a valuable evaluation starting-point based on quantitative metrics that have the potential to discriminate between skill levels. In this research, we developed a new assessment methodology for quantifying surgeons’ performance during key portions of the procedure, as identified using our MCMD, despite the fact that it is not yet practical due to its computational complexity. As more work is done in standardizing and simplifying this approach, its ability to compare a resident’s performance with respect to a group either of their peers or of expert surgeons will allow us to draw finer distinctions such as whether or not a given resident is keeping up with his/her year level. In addition, by identifying the points of greatest difference, attending surgeons can eventually use this information to provide more specific feedback to the trainee and better monitor their progress through their training program.

Bibliography

Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, Bello F, Darzi A (2007) An Evaluation of the Feasibility, Validity, and Reliability of Laparoscopic Skills Assessment in the Operating Room. Ann Surg 245: 992-999

Aggarwal R., Moorthy K., Darzi A (2004) Laparoscopic skills training and assessment. Br J Surg 91: 1549-1558

Ahlberg G., Heikkinen T., Iselius L., Leijonmarck C.E, Rutqvist J., Arvidsson D (2002) Does training in a virtual reality simulator improve surgical performance? Surg Endosc 16: 126-129

Alleman A (2005) Have You Wondered About Your Colleague's Surgical Skills? Am J Med Qual 20: 78-82

Babineau T, Becker J, Gibbons G, Sentovich S, Hess D, Robertson S, Stone M (2004) The "Cost" of Operative Training for Surgical Residents. Arch Surg 139: 366-370

Bann S., Khan M., Darzi A (2003) Measurement of surgical dexterity using motion analysis of simple bench tasks. World J Surg 27: 390-394

Berber E., Engle K.L., Garland A., String A., Foroutani A., Pearl J.M., Siperstein A.E (2001) A critical analysis of intraoperative time utilization in laparoscopic cholecystectomy. Surg Endosc 15: 161-165

Berguer R., Forkey D.L., Smith W.D (2001) The effect of laparoscopic instrument working angle on surgeons' upper extremity workload. Surg Endosc 15: 1027-1029

Berguer R., Smith W.D., Chung Y.H (2001) Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. Surg Endosc 15: 1204-1207

Besse P (1992) PCA stability and choice of dimensionality. Stat Probab Lett 13: 405-410

Besse P., Caussinus H., Ferre L., Fine J (1988) Principal component analysis and optimization of graphical displays. Statistics 19: 301-312

Bicego M., Murino V., Figueiredo M (2004) Similarity-based classification of sequences using hidden Markov models. Pattern Recognit 37: 2281-2291

Bittner R (2004) The standard of laparoscopic cholecystectomy. Langenbecks Arch Surg 389: 157-163

Bobbio A (1990) System Modelling with Petri Nets. Systems Reliability Assessment. Proceedings of the Ispra Course 103-143

Breiman L (1992) The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. J Am Stat Assoc 87: 738-754

Bridges M, Diamond D (1999) The Financial Impact of Teaching Surgical Residents in the Operating Room. Am J Surg 177: 28-32

Britt L (2007) Simulation training: what are real questions that must be answered? Am J Surg 194:220

Brownlee J (2006) Finite State Machines. AIdepot

Burnham K.P., Anderson D.R (2002) Model selection and multimodel inference: a practical information-theoretical approach. New York: Springer-Verlag 2nd Edition

Burnham K.P., Anderson D.R (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33: 261-304

Cao C.G.L., MacKenzie C.L (1996) Task and motion analyses in endoscopic surgery. ASME IMECE Conference Proceedings: 5th Annual Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, Atlanta, Georgia, 583-590

Cao C.G.L., MacKenzie C.L., Ibbotson J.A., Turner L.J., Blair N.P., Nagy A.G. (1999) Hierarchical decomposition of laparoscopic procedures. In Proc. Medicine Meets Virtual Reality: 7. IOS Press, Amsterdam

Childs J.M (1980) Time and error measures of human performance: a note on Bradley's Optimal-Pessimal Paradox. Hum Factors 22: 113-117

Cristancho S., Hodgson A., Panton N., Meneghetti A., Qayumi K., (2006) Assessing Cognitive & Motor Performance in MIS for Training & Tool Design. Stud Health Technol Inform 119: 108-113

Cristancho S., Hodgson A., Panton N., Meneghetti A., Qayumi K., (2007) Feasibility of using intraoperatively-acquired quantitative kinematic measures to monitor development of laparoscopic skill. Stud Health Technol Inform 125: 85-90

Cuschieri A., Berci G (1990) Laparoscopic biliary surgery. London: Blackwell Scientific

Dagan I, Lee L, Pereira F (1999) Similarity-based models of word cooccurrence probabilities. Machine Learning, 34: 43–69

Darzi A., Datta V., Mackay S. (2001) The challenge of objective assessment of surgical skill. Am J Surg 181: 484-486

Datta V, Bann S, Mandalia M, Darzi A (2006) The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg 192: 372–378

Datta V., Chang A., Mackay S., Darzi A (2002) The relationship between motion analysis and surgical technical assessments. Am J Surg 184: 70-73

Daudin J.J., Duby C., Trecourt P (1988) Stability of principal component analysis studied by the bootstrap method. Statistics 19: 241-258

De Visser H., Heijnsdijk E.A., Herder J.L., Pistecky P.V. (2002) Forces and displacements in colon surgery. Surg Endosc 16:1426-1430

Derossis A.M., Bothwell J., Sigman H.H., Fried G.M. (1998) The effect of practice on performance in a laparoscopic simulator. Surg Endosc 12: 1117-1120

Ding Y., Wang B., Wang W., Wang P., Yan J (2007) New classification of the anatomic variations of cystic artery during laparoscopic cholecystectomy. World J Gatroenterol 13: 5629-5634

Dohrmann, C., Busby, H., Trujillo, D. (1988). Smoothing noisy data using dynamic programming and generalised cross-validation. J Biomech Eng 110: 37–41

Doob J.L (1953) Stochastic Processes. New York: John Wiley and Sons

Dosis A., Aggarwal R., Bello F., Moorthy K., Munz Y., Gillies D., Darzi A (2005) Synchonized video and motion analysis for the assessment of procedures in the operating theater. Arch Surg 140: 293-299

Drew J, Glen A, Leemis L (2000) Computing the cumulative distribution function of the Kolmogorov–Smirnov statistic. Comput Stat Data Anal 34: 1–15

Dudzinski M.L., Norris J.M, Chmura J.T., Edwards C.B.H (1975) Repeatability of principal components in samples: normal and non-normal data sets compared. Multivariate Behav Res 10: 109-118

Efron B., Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1: 54-77

Endres D, Schindelin J (2003) A new metric for probability distributions. IEEE Trans Inf Theory 49: 1858-1860

Feldman L.S., Sherman V., Fried G.M. (2004) Using simulators to assess laparoscopic competence: ready for widespread use? Surgery 135:28-42

Fielding G (2002) The case for laparoscopic common bile duct exploration. J Hepatobiliary Pancreat Surg 9: 723-728

Finlayson E., Nelson H (2005) Laparoscopic colectomy for cancer. Am J Clin Oncol 28: 521-525

Francis N.K., Hanna G.B., Cuschieri A (2001) Reliability of the Advanced Dundee Endoscopic Psychomotor Tester for bimanual tasks. Arch Surg 182: 30-33

Fried G.M., Derossis A.M., Bothwell J., Sigman, H.H (1999) Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. Surg Endosc 13: 1077-1081

Gagne R.M (1985) The conditions of learning and theory of instruction. New York: Holt, Rinehart and Winston, 4th Edition

Gallagher A.G., Lederman A.B., McGlade K., Satava R.M., Smith C.D. (2004) Discriminative validity of the minimally invasive surgical trainer in virtual reality (MIST-VR) using criteria levels based on expert performance Surg Endosc 24: 660-665.

Gallagher AG, Satava RM (2002) Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. Surg Endosc 16:1746–1752

Gibson M.A. (2000) Computational Methods for Stochastic Biological Systems. PhD thesis, California Inst. Technology

Giuntini R.E (2000) Mathematical characterization of human reliability for multi-task system operations. IEEE International Conference on Systems, Man, and Cybernetics 2: 1325-1329

Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P (2004) Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 91: 146-150.

Grantcharov TP, Rosenberg J, Pahle E, Funch-Jensen P (2001) Virtual reality computer simulation. Surg Endosc 15: 242-244

Grober E.D., Hamstra S.J., Wanzel K.R., Reznick R.K., Matsumoto E.D., Sidhu R.S., Jarvi K.A (2003) Validation of novel and objective measures of microsurgical skill: hand-motion analysis and stereoscopic visual acuity. Microsurgery 23: 317-322

Haluck R.S., Marshall R.L., Krummel T.M., Melkonian M.G (2001) Are surgery training programs ready for virtual reality? A survey of program directors in general surgery. J Am Coll Surg 193: 660-665

Hamilton E.C., Scott D.J., Kapoor A., Nwariaku F., Bergen P.C., Rege R.V., et al (2001) Improving operative performance using a laparoscopic hernia simulator. Am J Surg 182: 725-728

Hammond I (2006) Training, assessment and competency in gynaecologic surgery. Best Practice & Research Clinical Obstetrics and Gynaecology 20: 173–187

Haverkot B (2001) Markovian models for performance and dependability evaluation. Lectures on formal methods and performance analysis. Book chapter. Springer publisher, 1st Edition: 38-83

Hodgson A.J (1994) Considerations in applying dynamic programming filters to the smoothing of noisy data. J Biomech Eng 116: 528-531

Hodgson, A.J. & McBeth, P.B (2002) Comparing motor performance on similar tasks in different settings: statistical characteristics of a nondimensional difference measure. Internal document.

Howard R.A (1971) Dynamic Probabilistic Systems. New York: John Wiley & Sons

Hytlander A., Lilegren D., Rhodin P.H., Lonroth H (2002) The transfer of basic skills learning in a laparoscopic simulator to the operating room. Surg Endosc 16: 1324-1328

Ignjatovic D., Zivanovic V., Vasic G., Kovacevic-Mcilwaine I (2006) Cystic artery anatomy characteristics in minimally invasive surgical procedures. Acta Chir Iugosl 53: 63-66

Jackson J. E (1991) A User's Guide to Principal Components. Wiley Series in Probability and Mathematical Statistics.

Johnson R. A., Wichern D.W (2002) Applied Multivariate Statistical Analysis. Prentice Hall 5th Edition

Jolliffe IT (2002) Principal Component Analysis. Second Edition. New York: Springer Publisher

Jonassen D., Tessmer M., Hannum W (1999) Task analysis methods for instructional design. London: Lawrence Erlbaum Associates Publishers.

Jordan J.A., Gallagher A.G., McGuigan J., McGlade K., McClure N (2000) A comparison between randomly alternating imaging, normal laparoscopic imaging, and virtual reality training in laparoscopic psychomotor skill acquisition. Am J Surg 180: 208-211

Keyser E.J., Derossis A.M., Antoniuk M., Sigman H.H., Fried G.M (2000) A simplified simulator for the training and evaluation of laparoscopic skills. Surg Endosc 14: 149-153

Khan M., Snelling A., Tiernan E (2005) The need for technical skills assessment in surgery. Int J Surg 3: 83-86

Kinnaird K (2004) A multifaceted quantitative validity assessment of laparoscopic surgical simulators. Master's thesis. Department of Mechanical Engineering, University of British Columbia

Larobina M., Nottle P (2005) Extrahepatic biliary anatomy at laparoscopic cholecystectomy: Is aberrant anatomy important? ANZ J Surg 75: 392-395

Lee L (1999) Measures of distributional similarity. Proceedings of the 37th ACL

Lin H., Shafran I., Murphy T., Okamura A., Yuh D., Hager G (2005) Automatic detection and segmentation of robot-assisted surgical motions. Int Conf Med Image Comput Assist Interv 8: 802-810

Lin H., Shafran I., Yuh D., Hager G (2006) Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg 11: 220-230

Lyngso R., Pedersen C., Nielsen H (1999) Metrics and similarity measures for hidden Markov models. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology 178-186

MacKenzie C.L., Ibbotson J.A., Cao C.G.L., Lomax A.J (2001) Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Min Invas Ther & Allied Technol 10: 121-127

Macmillan A.I.M., Cuschieri A (1999) Assessment of innate ability and skills for endoscopic manipulation by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. Am J Surg 177: 274-277

Majtey AP, Lamberti PW, Prato DP (2005) Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. Phys Rev A 72: 052310-1 - 052310-6

Martel G., Boushey R.P (2006) Laparoscopic colon surgery: past, present and future. Surg Clin N Am 86: 867-897

Martin J.A., Regehr G., Reznick R., MacRae H., Murnaghan J., Hutchison C., Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84: 273-278

McBeth P (2002) A Methodology for Quantitative Performance Evaluation in Minimally Invasive Surgery. Master's thesis. Department of Mechanical Engineering, University of British Columbia

McCloy R., Stone R. (2001) Science, medicine and the future: Virtual reality in surgery. BMJ 323: 912-915

McDonald J (2008) Handbook of Biological Statistics. University of Delaware.

McKenzie S., Schwartz R (2006) The management of bile duct injuries occurring during laparoscopic cholecystectomy. Curr Surg 63: 20-23

Mehta N.Y., Haluck R.S., Frecker M.I., Snyder A.J (2002) Sequence and task analysis of instrument use in common laparoscopic procedures. Surg Endosc 16: 280-285

Meyn S.P., Tweedie R.L. (1993) Markov Chains and Stochastic Stability. London: Springer-Verlag, 1993. Online: http://decision.csl.uiuc.edu/~meyn/pages/book.html . Second edition to appear, Cambridge University Press, 2008.

Milne A.D., Chess D.G., Johnson J.A., King G.J.W. (1996) Accuracy of an electromagnetic tracking device: a study of the optimal operating range and metal interference. J Biomech 29:791-3

Moorthy K, Munz Y, Sarker S, Darzi A (2003) Objective assessment of technical skills in surgery. BMJ 327: 1032-1037

Morimoto A.K., Foral R.D., Kuhlman J.L., Zucker K.A., Curet M.J., Bocklage T., MacFarlane T.I., Kory L (1997) Force sensor for laparoscopic Babcock. Stud Health Technol Inform 39: 354-361

Murata T (1989) Petri nets: Properties, analysis and applications. Proceedings of the IEEE. 77 (4): 541 - 580

Murphy T (2004) Towards objective surgical skill evaluation with hidden Markov model-based motion recognition. Master’s thesis, Johns Hopkins University, Baltimore, Maryland, USA

Murphy T., Vignes C., Yuh D., Okamura A (2003) Automatic motion recognition and skill evaluation for dynamic tasks. Eurohaptics

Murthy D.N., Xie M., Jiang R (2003) Weibull Models. Wiley Series in Probability and Statistics

Narazaki K., Oleynikov D., Stergiou N (2007) Objective assessment of proficiency with bimanual inanimate tasks in robotic laparoscopy. J Laparoendosc Adv Surg Tech 17: 47-52

Nguyen N.T., Ho H.S., Smith W.D., Philipps C., Lewis C., De Vera R.M., Berguer R (2001) An ergonomic evaluation of surgeons’ axial skeletal and upper extremity movements during laparoscopic and open surgery. Am J Surg 182: 720-724

Nixon M.A., McCallum B.C., Fright W.R., Price N.B (1998) The effects of metals and interfering fields on electromagnetic trackers. Presence 7: 204-218

Paisley A.M., Baldwin P.J., Paterson-Brown S (2001) Validity of surgical simulation for the assessment of operative skill

Park AE, Witzke D (2002) The surgical competence conundrum. Surg Endosc 16: 555-557

Patterson E., Nagy A (1997) Don't cry over spilled stones? Complications of gallstones spilled during laparoscopic cholecystectomy: case report and literature review. Can J Surg 40: 300-304

Pearson A.M., Gallagher A.G., Rosser J.C., Satava R.M (2002) Evaluation of structured and quantitative training methods for teaching intracorporeal knot tying. Surg Endosc 16: 130-137

Petelin J.B (2002) Surgical management of common bile duct stones. Gastrointest Endosc 56: S183-S189

Peterson J (1977) Petri Nets. ACM Computing Surveys. 9 (3): 223 - 252

Polhemus Incorporated (2002) 3Space Fastrak User's Manual. OPM00PI002. Colchester, Vermont U.S.A.

Qian Y., Jia S., Si W (2003) Markov model based time series similarity measuring. Proceedings of the 2nd International Conference on Machine Learning and Cybernetics 278-283

Raethjen J., Pawlas F., Lindemann M., Wenzelburger R., Deuschl G (2000) Determinants of physiologic tremor in a large normal population. Clin Neurophysiol 111: 1825-1837

Reddick E.J., Saye W.B., Corbitt J (1993) Atlas of laparoscopic surgery. New York: Raven Press

Regehr G., Szalay D., Reznick R (1998) Forced choice rankings of clinical performance: a validation tool. In: Melnick DE, ed. The Eighth International Ottawa Conference on Medical Education and Assessment Proceedings, July 12-15. Philadelphia.

Risucci D., Geiss A., Gellman L., Pinnard B., Rosser J (2001) Surgeon-specific factors in the acquisition of laparoscopic surgical skills. Am J Surg 181: 289-293

Rosen J, Brown J, Chang L, Sinanan M, Hannaford B (2006) Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Trans Biomed Eng 53: 399-413

Rosen J., Hannaford B., Richards G., Sinanan M. (2001) Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng 48: 579-591

Rosen J., Solazzo M., Hannaford B., Richards G., Sinanan M. (2002) Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Comput Aided Surg 7: 49-61

Rosser J.C. Jr., Rosser L.E., Savalgi R.S (1997). Skill acquisition and assessment for laparoscopic surgery. Arch Surg 132: 200-204.

Sarker S.K., Chang A., Vincent C., Darzi A (2006) Development of assessing generic and specific technical skills in laparoscopic surgery. Am J Surg 191: 238-244

Satava R.M (2001) Accomplishments and challenges of surgical simulation. Surg Endosc 15: 232-241

Satava R.M (2004) Disruptive visions: surgical education. Surg Endosc 18: 779-781

Scott D, Rege R, Bergen P, Guo W, Laycock R, Tesfay S, Valentine J, Jones D (2000) Measuring Operative Performance after Laparoscopic Skills Training: Edited Videotape versus Direct Observation. J Laparoendosc Adv Surg Tech A 10:183-190

Scott D.J., Rege R.V., Bergen P.C., Guo W.A., Laycock R., Tesfay S.T., Valentine R.J., Jones D.B (2000) Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech 10: 183-190

Seymour N.E., Gallagher A.G., Roman S.A., O'Brien M.K., Anderson D.K., Satava R.M (2004). Analysis of errors in laparoscopic surgical procedures. Surg Endosc 18: 592-595

Sidhu R.S., Grober E.D., Musselman L.J., Reznick R.K (2004) Assessing competency in surgery: Where to begin? Surgery 135: 6-20

Siperstein A., Pearl J., Macho J., Hansen P., Gitomirsky A., Rogers S (1999) Comparison of laparoscopic ultrasonography and fluorocholangiography in 300 patients undergoing laparoscopic cholecystectomy. Surg Endosc 13: 113-117

Smith C.D., Farrell T.M., McNatt S.S., Metreveli R.E (2001) Assessing laparoscopic manipulative skills. Am J Surg 181: 547:550

Smith C.D., Farrell T.M., McNatt S.S., Metreveli R.E. (2001) Assessing laparoscopic manipulative skills. Am J Surg 181: 547-50

Starkes J.L., Payk I., Hodges N.J (1998) Developing a standardized test for the assessment of suturing skill in novice microsurgeons. Microsurgery 18: 19-22

Stone R., McCloy R (2004) Ergonomics in medicine and surgery. BMJ 328: 1115-1118

Sullivan M., Ortega A., Wasserberg N., Kaufman H., Nyquist J., Clark R (2008) Assessing the teaching of procedural skills: can cognitive task analysis add to our traditional teaching methods? Am J Surg 195: 20-23

Taffinder N., Sutton C., Fishwick R.J, et al (1998) Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. Stud Health Technol Inform 50: 124 –30

Tendick F (2000) A virtual environment testbed for training laparoscopic surgical skills. Presence 9: 236-255

Thomas W (2006) Teaching and assessing surgical competence. Ann R Coll Surg Engl 88: 429-432

Torkington J, Smith S.G.T, Rees B.I., Darzi A (2001) The role of the basic surgical skills course in the acquisition and retention of laparoscopic skill. Surg Endosc 15: 1071-1075

Veelen M.A., Nederlof E.A.L., Goossens R.H.M., Schot C.J., Jakimowicz J.J (2003) Ergonomic problems encountered by the medial team related to products for Minimally Invasive Surgery. Surg Endosc 17: 1077-1081

Vincent C., Moorthy K., Sarker S., Chang A., Darzi A (2004) Systems approaches to surgical quality and safety. Ann Surg 239: 475-482

Von Mises, R (1964) Mathematical Theory of Probability and Statistics. Academic Press.

Wagner A., Schicho K., Birkfellner W., Figl M., Seemann R., Konig F., Kainberger F., Ewers R (2002) Quatitative analysis of factors affecting intraoperative precision and stability of optoelectronic and electromagnetic tracking systems. Med Phys 29: 905-912

Walpole R.E., Myers R.H., Myers S.L (2007) Probability and statistics for engineers and scientist. Prentice Hall 8th Edition

Wanzel KR, Ward M, Reznick RK (2002) Teaching the surgical craft: from selection to certification. Curr Probl Surg 39: 574–659

Warf B.C., Donnelly M.B., Schwartz R.W., Sloan D.A (1999) Interpreting the judgment of surgical faculty regarding resident competence. J Surg Res 86: 29-35

Weigmman D.A., ElBardissi A., Dearani J.A., Daly R., Sundt T.M (2007) Disruptions in surgical flow and their relationship to surgical errors: An exploratory investigation. Surgery 142: 658-665

Witten, I.H., Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufmann

Wu T., Hsieh Y., Li L (2001) Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics 57: 441-448

Zerey M., Burns J., Kercher K., Kuwada T., Heniford T (2006) Minimally invasive management of colon cancer. Surg Inn 13: 5-15

Zhang J., Chu F (2005) Real-time modeling and prediction of physiological hand tremor. IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings (ICASSP'05) 5: 645-648

Zhang L., He X., Dai L., Huang X (2007) The simulator experimental study on the operator reliability of Qinshan nuclear power plant. Reliability Engineering & System Safety 92: 252-259

APPENDIX A

VIDEO VERIFICATION – Checklist

Surgeon: ________________________________ Trial #: ______

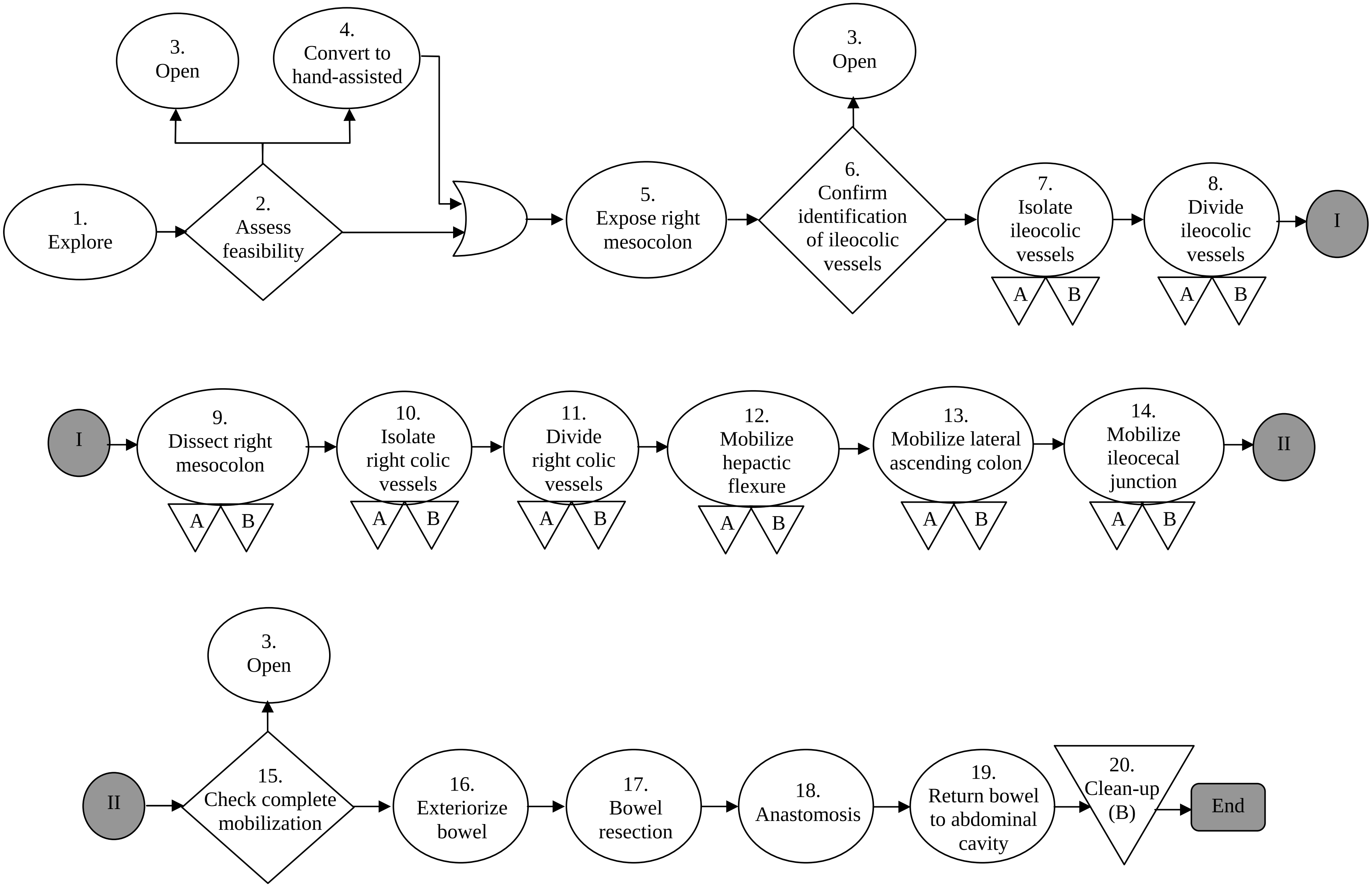
Date of surgery: __________________________

• **Explore:**
• **Assess anatomy:**
Open
Assess cholecystitis
• **Assess cholecystitis:**
Aspirate GB
Isolate CD/CA
• **Isolate CD/CA:**
Control bleeding
Confirm identification of CD/CA
• **Confirm identification of CD/CA:**
Cholangiogram
Reconfirm identification: Open
Notice CBDS
Assess possible CBDS
• **Notice presence of CBDS:**
Open
LCBE
Assess possible CBD injury
• **Assess possible CBD injury:**
Open
Separate CD or CA
• **Separate CD:**
Control bleeding
Separate CA
Dissect GB
• **Separate CA:**
Control bleeding
Separate CD

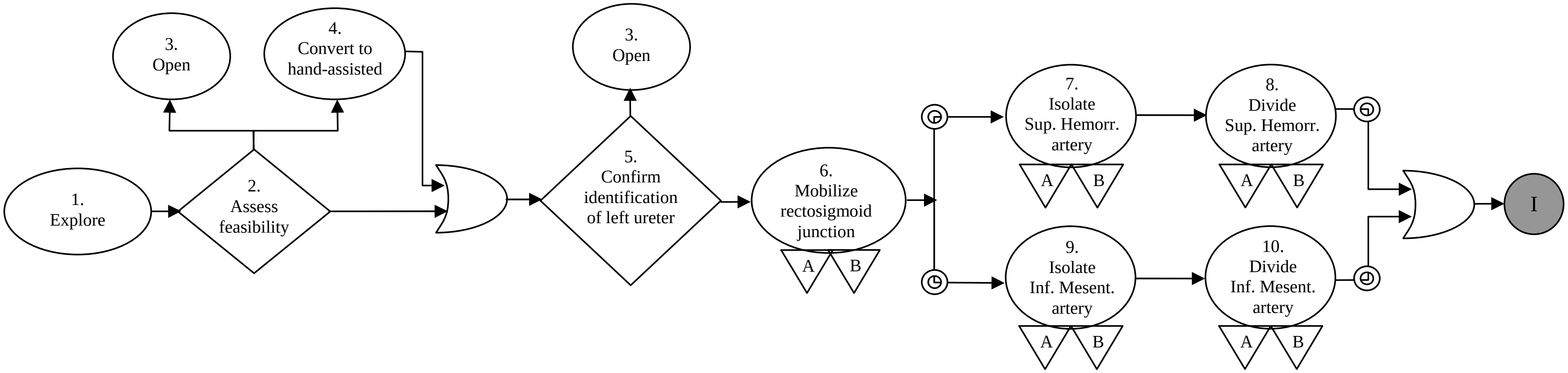
Dissect GB
• **Dissect GB:**
Control bleeding
Clean-up
• **Clean-up:**
Control bleeding
Bag GB
Extract GB
• **Bag GB:**
Clean-up
Extract GB
• **Extract GB:**
End

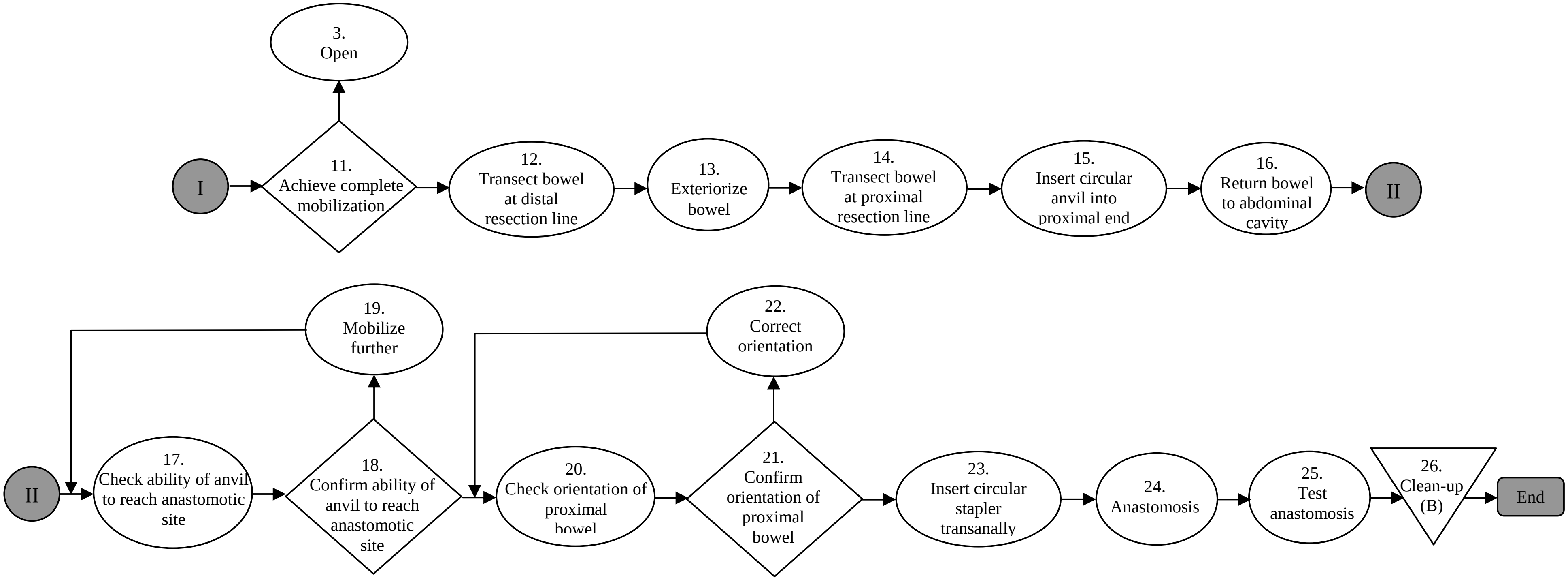
APPENDIX B

MCMD for Laparoscopic Right Hemicolectomy – Task Level


1. Explore
2. Assess feasibility
3. Open
4. Convert to hand-assisted
5. Expose right mesocolon
6. Confirm identification of ileocolic vessels
3. Open
7. Isolate ileocolic vessels
A
B
8. Divide ileocolic vessels
A
B
I
I
9. Dissect right mesocolon
A
B
10. Isolate right colic vessels
A
B
11. Divide right colic vessels
A
B
12. Mobilize hepactic flexure
A
B
13. Mobilize lateral ascending colon
A
B
14. Mobilize ileocecal junction
A
B
II
II
15. Check complete mobilization
3. Open
16. Exteriorize bowel
17. Bowel resection
18. Anastomosis
19. Return bowel to abdominal cavity
20. Clean-up (B)
End

MCMD for Laparoscopic Sigmoid Colectomy – Task Level




3.
Open
I
11.
Achieve complete mobilization
12.
Transect bowel at distal resection line
13.
Exteriorize bowel
14.
Transect bowel at proximal resection line
15.
Insert circular anvil into proximal end
16.
Return bowel to abdominal cavity
II
19.
Mobilize further
22.
Correct orientation
II
17.
Check ability of anvil to reach anastomotic site
18.
Confirm ability of anvil to reach anastomotic site
20.
Check orientation of proximal bowel
21.
Confirm orientation of proximal bowel
23.
Insert circular stapler transanally
24.
Anastomosis
25.
Test anastomosis
26.
Clean-up (B)
End

APPENDIX C

Alternative Position Tracking Systems

Our group has previously used two types of position tracking systems: optical (Northern Digital Polaris System) and electromagnetic (Polhemus 3SPACE Fastrak). It has been shown that use of optoelectronic tracking often misses significant segments of data, which reduces the completeness and validity of the data records [McBeth 2002]. Optoelectronic systems are also sensitive to ambient light from OR lamps, although this is a far less significant problem [Wagner 2002]. On the other hand, while electromagnetic tracking eliminates the line-of-sight requirement, it suffers from lower accuracy and is more subject to distortion than the optical one. Milne evaluated the accuracy of electromagnetic trackers under the effect of different metals as those commonly present at the operating room. This group reported that the accuracy quoted by the product manual (2.5mm RMS) could only be achieved in the operating room with a transmitter-receiver distance between 22.5cm and 64cm [Milne 1996]. This same relationship between metal effects and transmitter-receiver separation was reported by Nixon et al [Nixon 1998].

Recently, fiber optic-based systems have been introduced for position tracking to be used in applications that are not possible with conventional electrical based sensors due to measurement requirements such as extreme temperature, small size, high sensor count, or high electromagnetic energy or radiation environments. In June 2007, a shape-sensing optical fiber ‘smart fiber’ was introduced for minimally invasive surgery applications, specifically to be integrated into Intuitive Surgical's products, which includes the da Vinci®

Surgical System[1]. Although it is a promising system for our application, the system was unavailable at the start of our research.

Figure C.1 shows a spider plot for comparing the three tracking systems being considered in terms of the technical features such as accuracy, update rate, workspace, compactness, immunity to electromagnetic (EM) noise and tracking continuity.

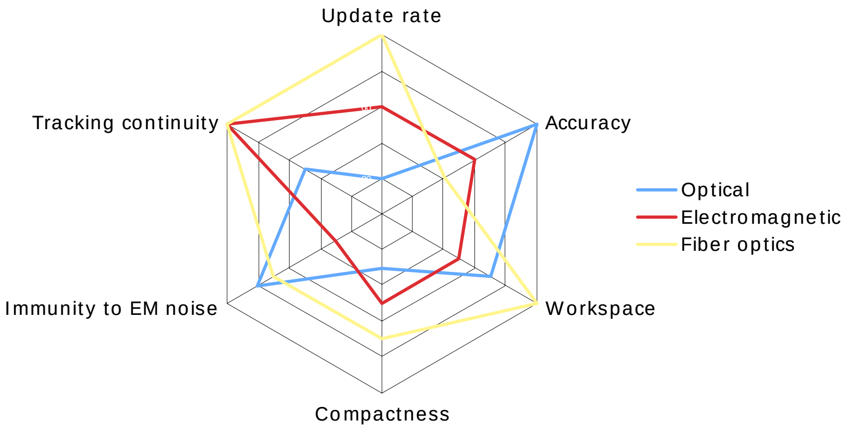


Figure C.1: Comparison plot of three position-tracking systems (optical, electromagnetic, and fiber optics) in terms of six technical features (accuracy, update rate, workspace, compactness, immunity to EM noise, and tracking continuity)

Given that we are looking at general profiles, high accuracy is not so important in this context while continuous tracking is required; therefore, in our approach we used the Polhemus 3SPACE Fastrak 6-dof electromagnetic system to obtain uninterrupted data streams.

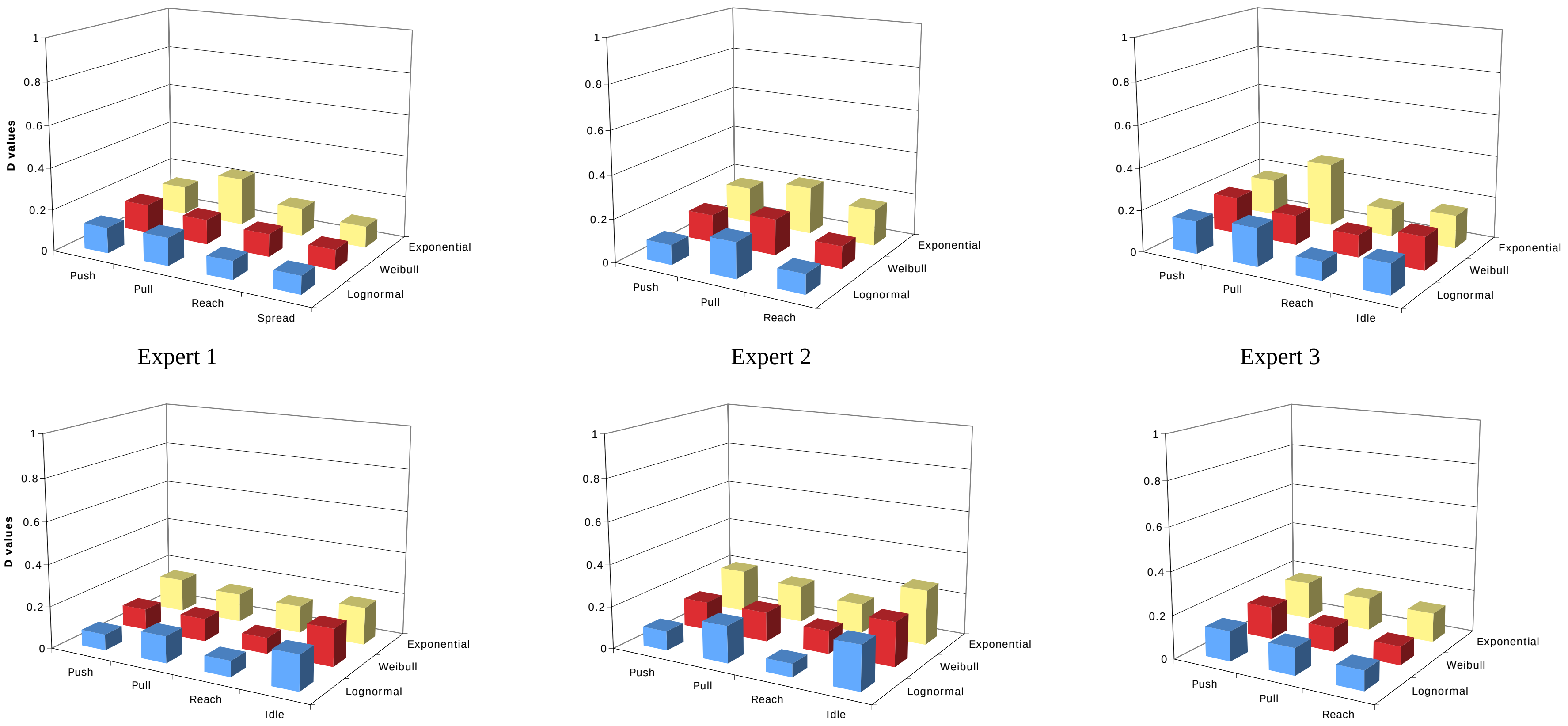
[1] http://www.lunainnovations.com/products/dss.htm (June 2007)

APPENDIX D

Semi-Markov Modelling

D.1 Distribution fitting using D values for holding time profiles of most used actions per subtask

Expose Triangle



Expert 1

Expert 2

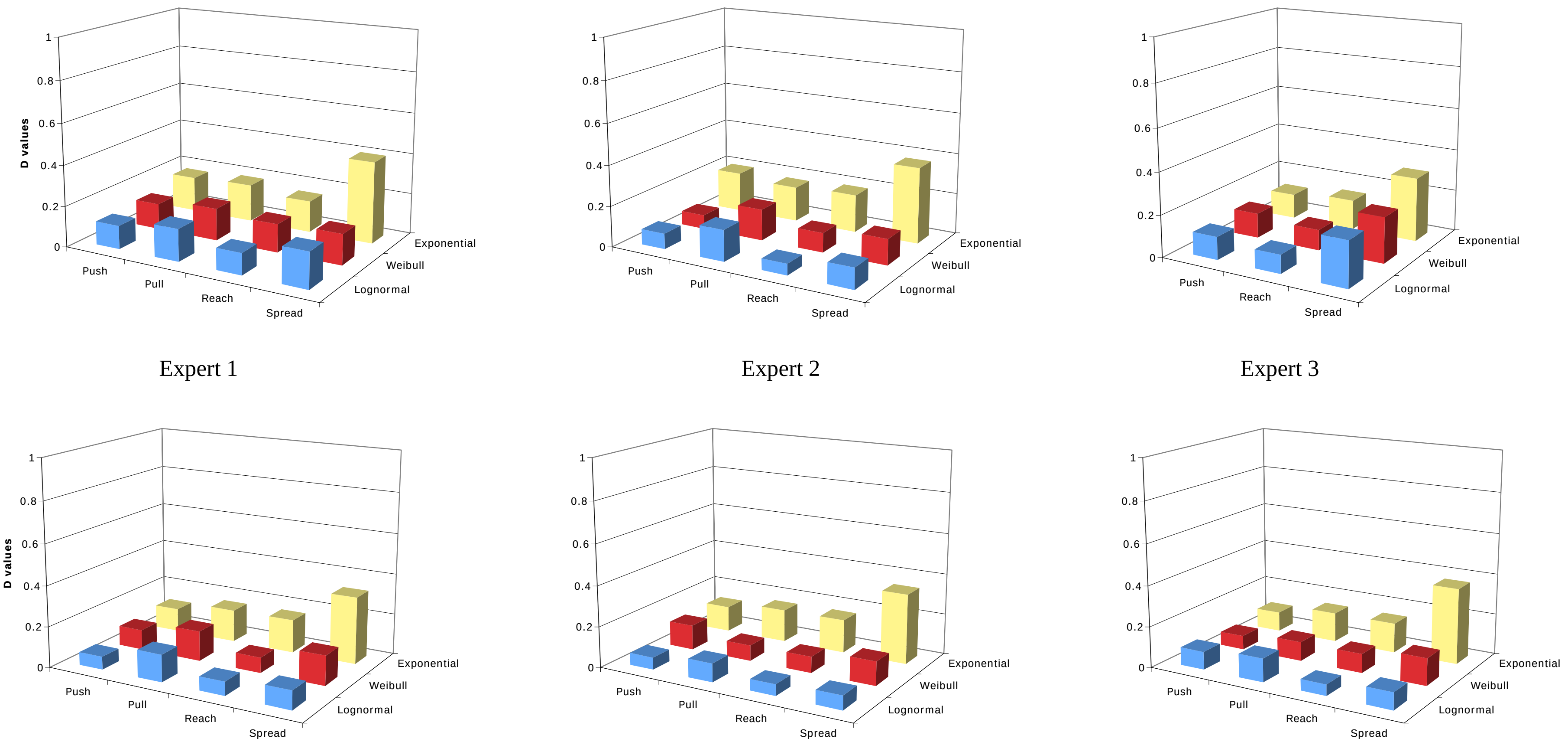
Expert 3

Resident 1

Resident 2

Resident 3

Dissect CD/CA



Expert 1

Expert 2

Expert 3

Resident 1

Resident 2

Resident 3

D.2 Parameter and confidence interval estimation of holding time distributions

In our model implementation, we selected the distribution with the lowest D value from above to represent holding time profiles and implemented a bootstrapping approach for computing the corresponding confidence intervals for the distribution parameters.

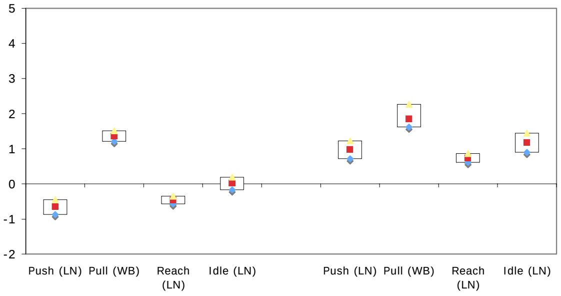
Assuming that the distribution fit h_i uses two parameters (a_i, b_i), we first replicated 1000 times the original holding time data as h^*_i (sampling with replacement) using a_i, b_i. Therefore if $h_i = [h_{i1}\ h_{i2}\ h_{i3} \ldots h_{iN}]'$ (column vector) with length N, we created 1000 vectors of length N each one represented as $[h^*_i(1)\ h^*_i(2) \ldots h^*_i(k) \ldots h^*_i(1000)]$. For each $h^*_i(k) = [h^*_{i1}\ h^*_{i2} \ldots h^*_{iN}]'$, we computed $a^*_i(k)$, $b^*_i(k)$ in order to obtain sets of parameters $a^*_i = [a^*_i(1)\ a^*_i(2) \ldots a^*_i(k) \ldots a^*_i(1000)]'$ and $b^*_i = [b^*_i(1)\ b^*_i(2) \ldots b^*_i(k) \ldots b^*_i(1000)]'$. We finally used the distributions of a^*_i, b^*_i and the percentile quartile method to estimate the corresponding confidence intervals ($\alpha = 0.05$):

$$\text{Confidence level} = 100(1\text{-}2\alpha) = 95\%$$

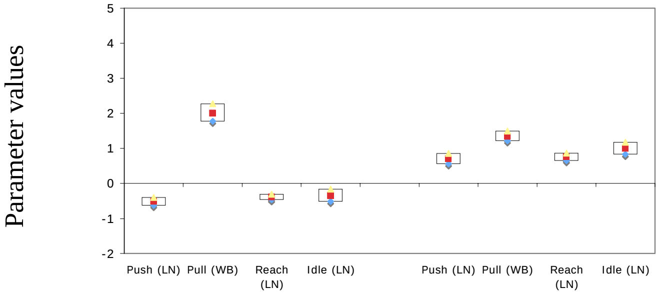
$$\text{CI}=[L(1\text{-}\alpha)^{\text{th}}, (L+\alpha)^{\text{th}}] \text{ with } L=1000;$$

The following plots show the estimated parameters and confidence intervals for the parameters of the selected distributions (LN: Lognormal; WB: Weibull) for the most used actions at each subtask. Parameter ‘a’ is presented on the right side and parameter ‘b’ on the left side of each individual plot. For lognormal, a: mean (mu); b: standard deviation (sigma). For Weibull, a: shape; b: scale.

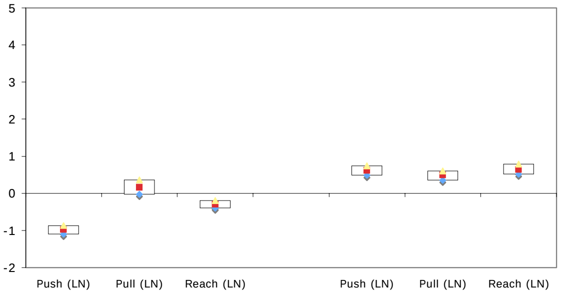
Expose Triangle



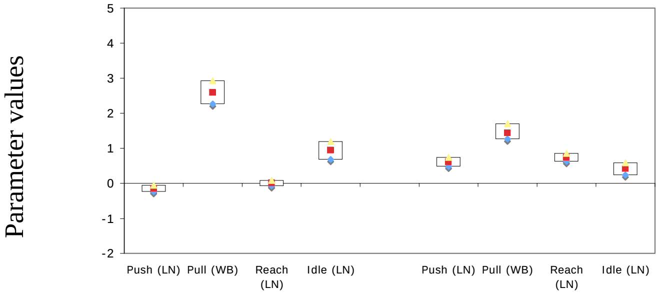
Expert 1



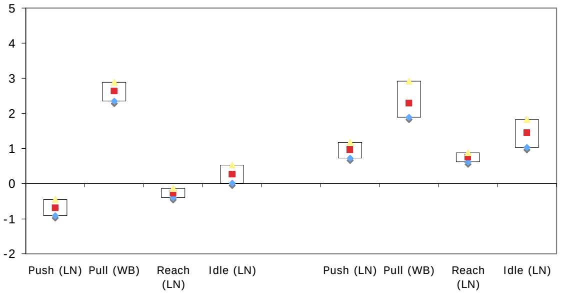
Resident 1



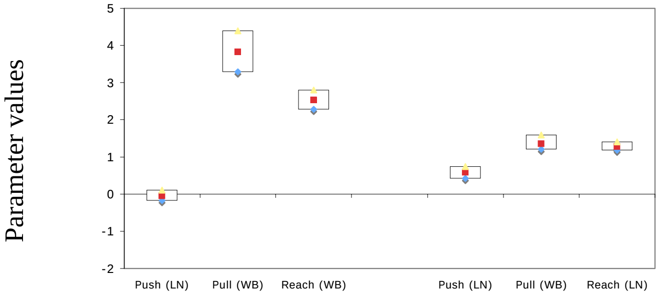
Expert 2



Resident 2

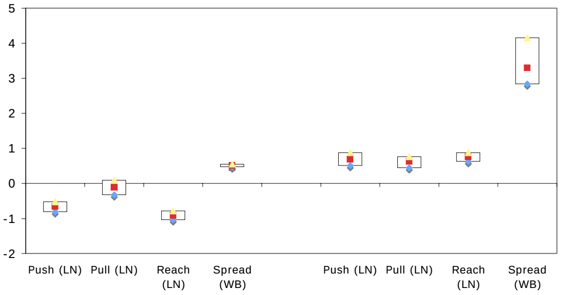


Expert 3

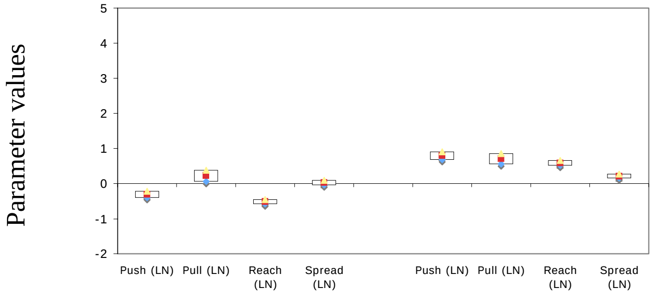


Resident 3

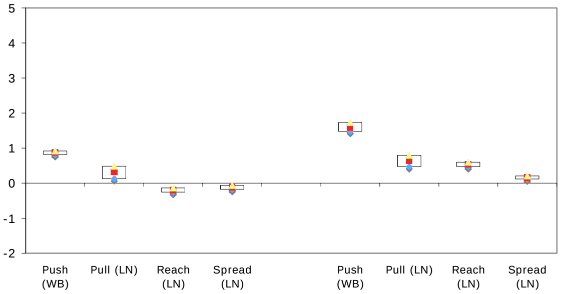
Dissect CD/CA



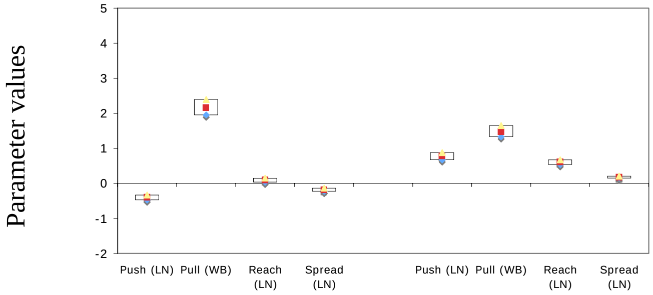
Expert 1



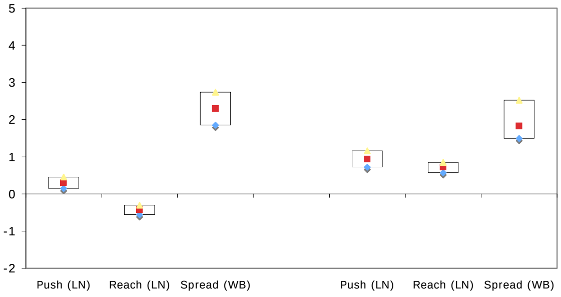
Resident 1



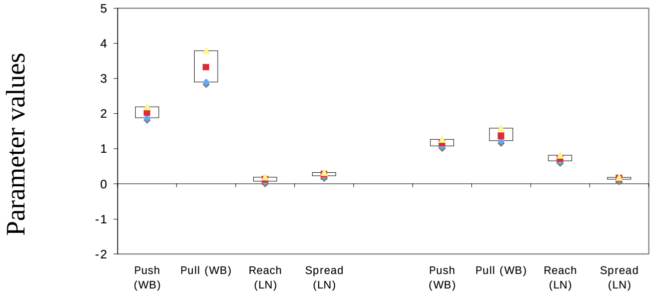
Expert 2



Resident 2



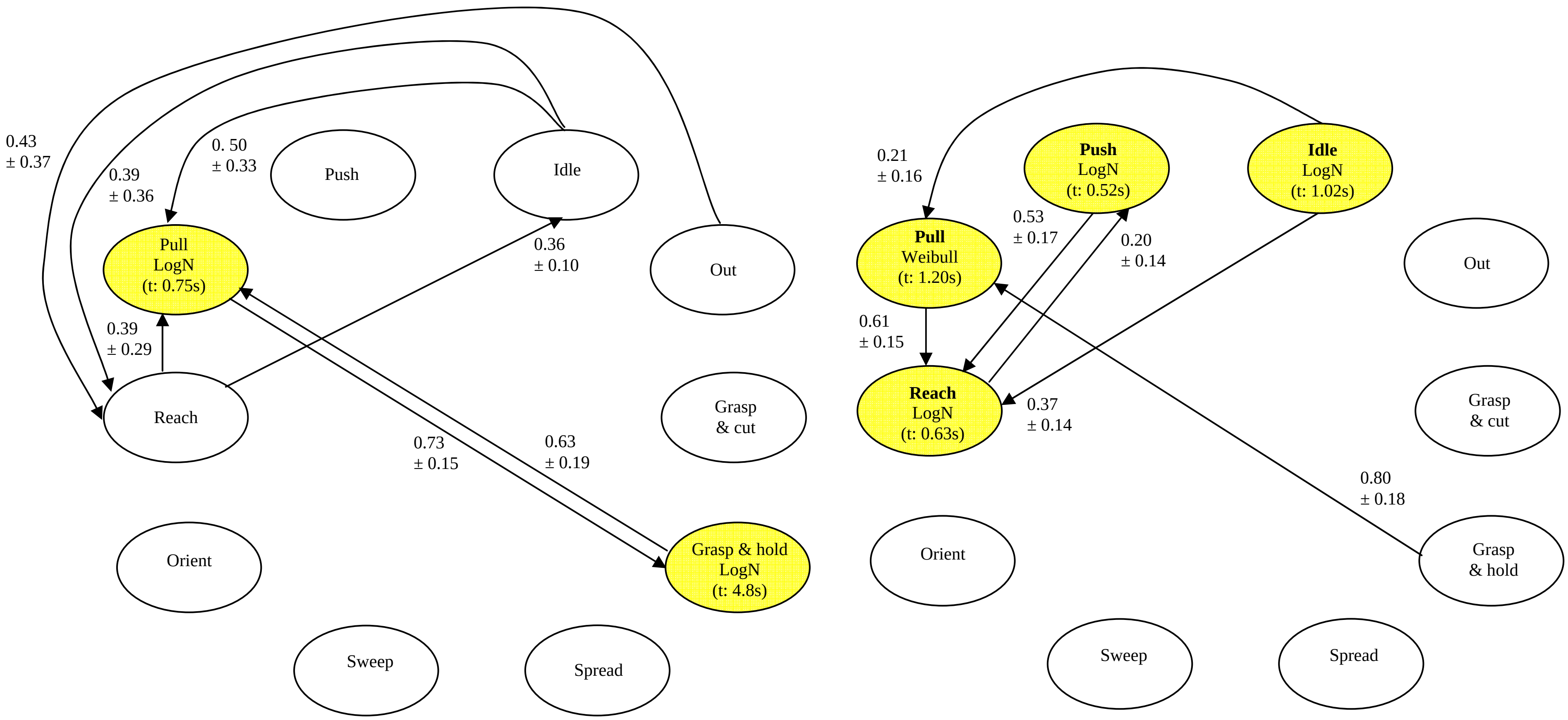
Expert 3



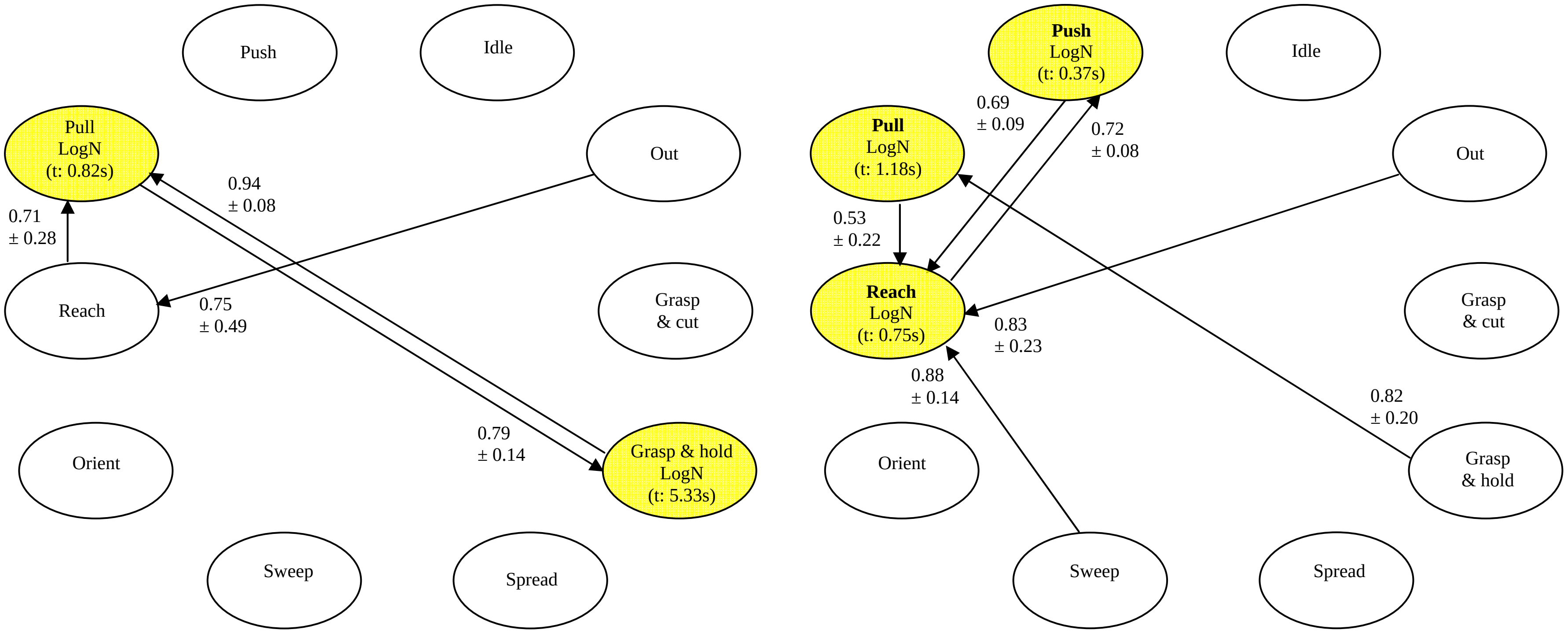
Resident 3

D.3 Semi-Markov models for each subject/subtask

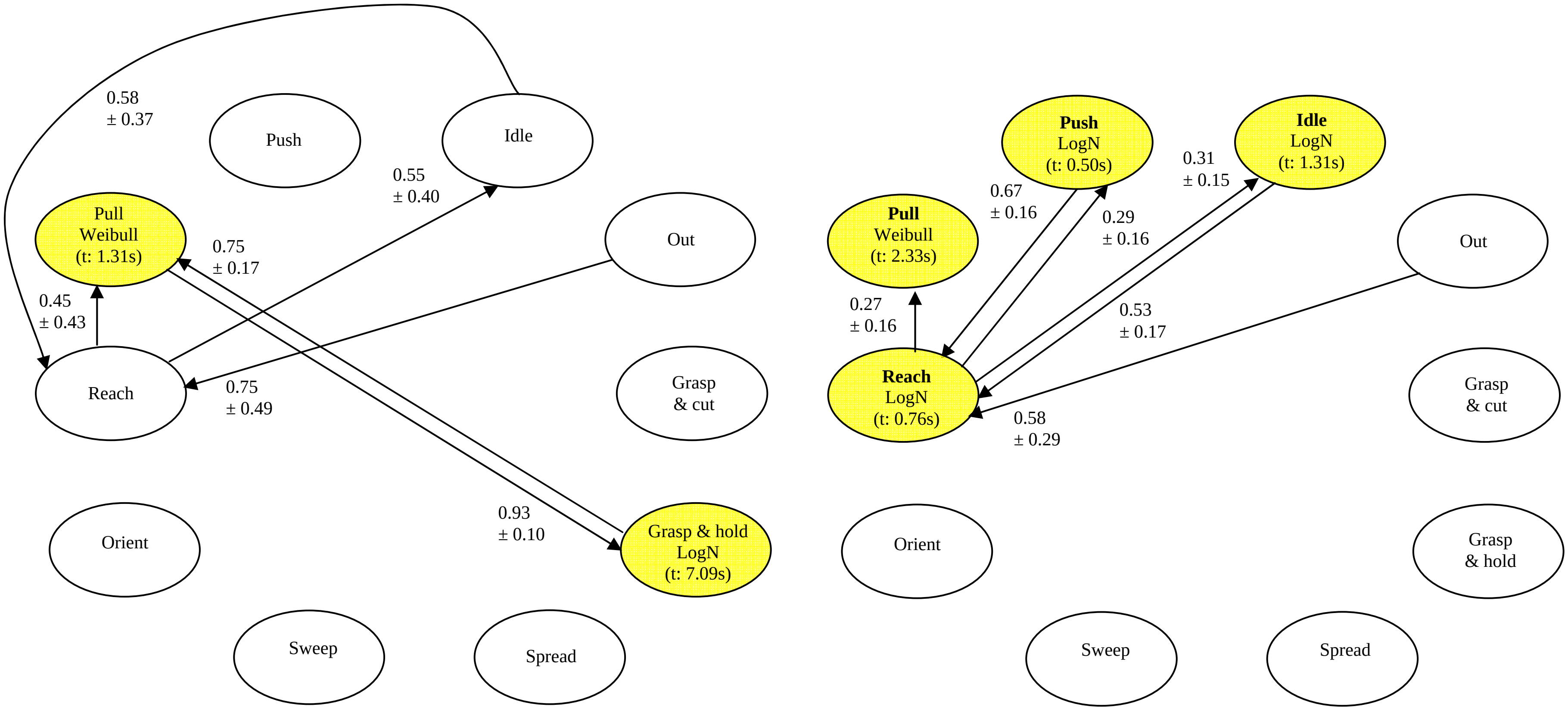
Expose Triangle



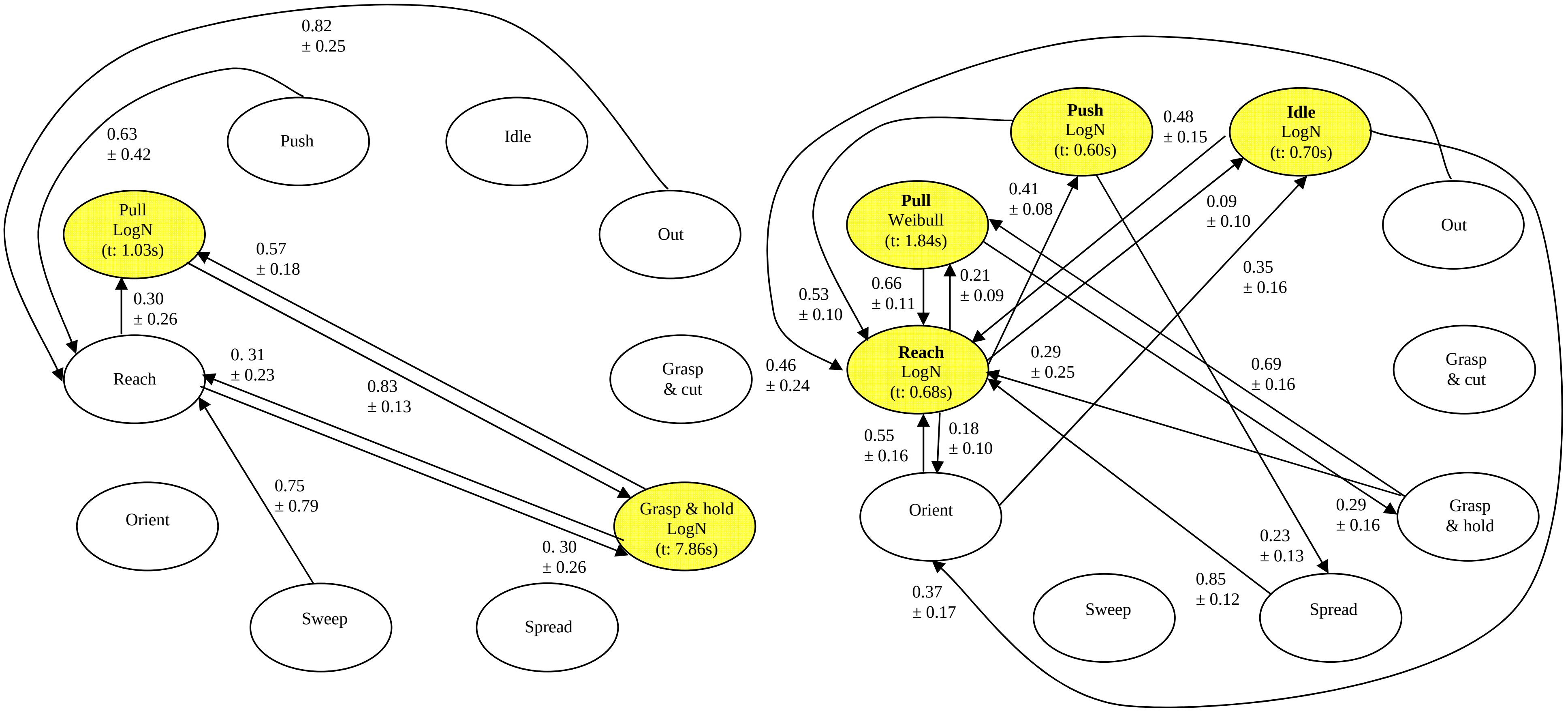
Expert 1: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



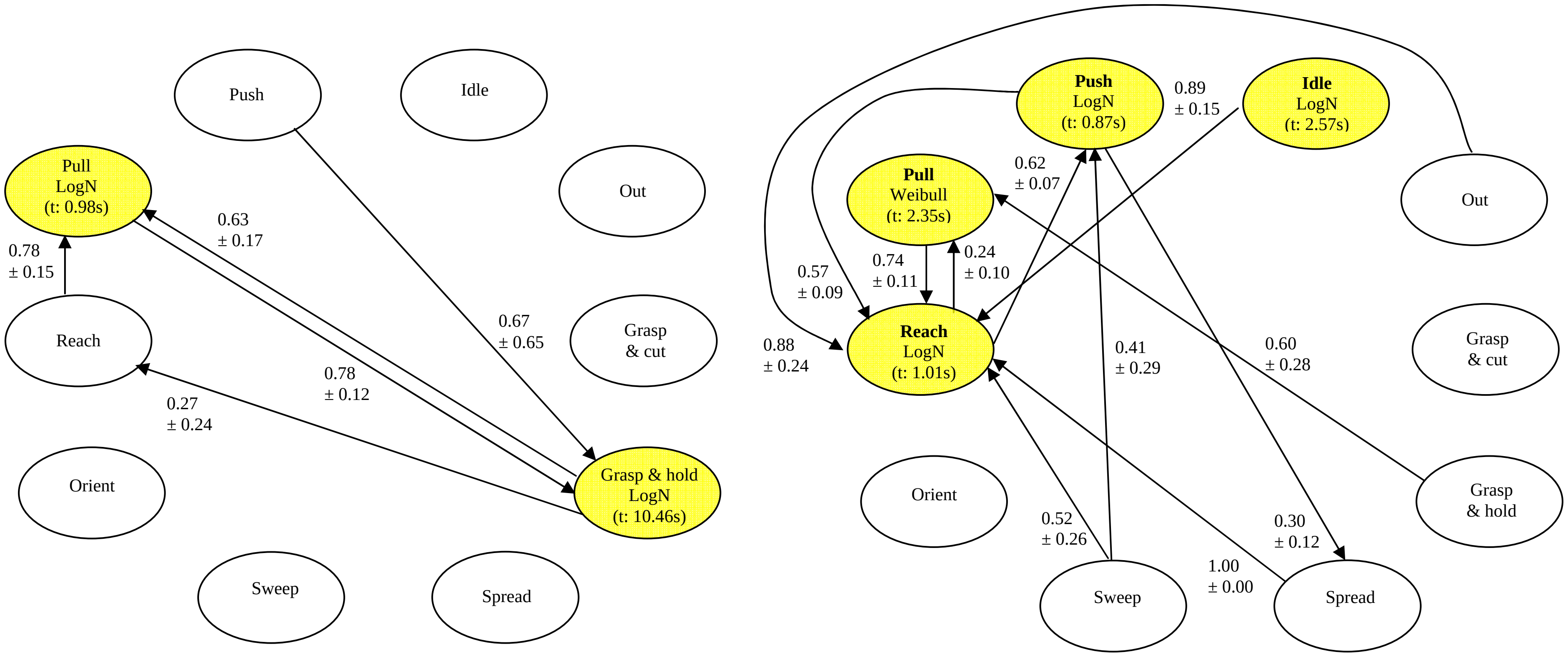
Expert 2: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



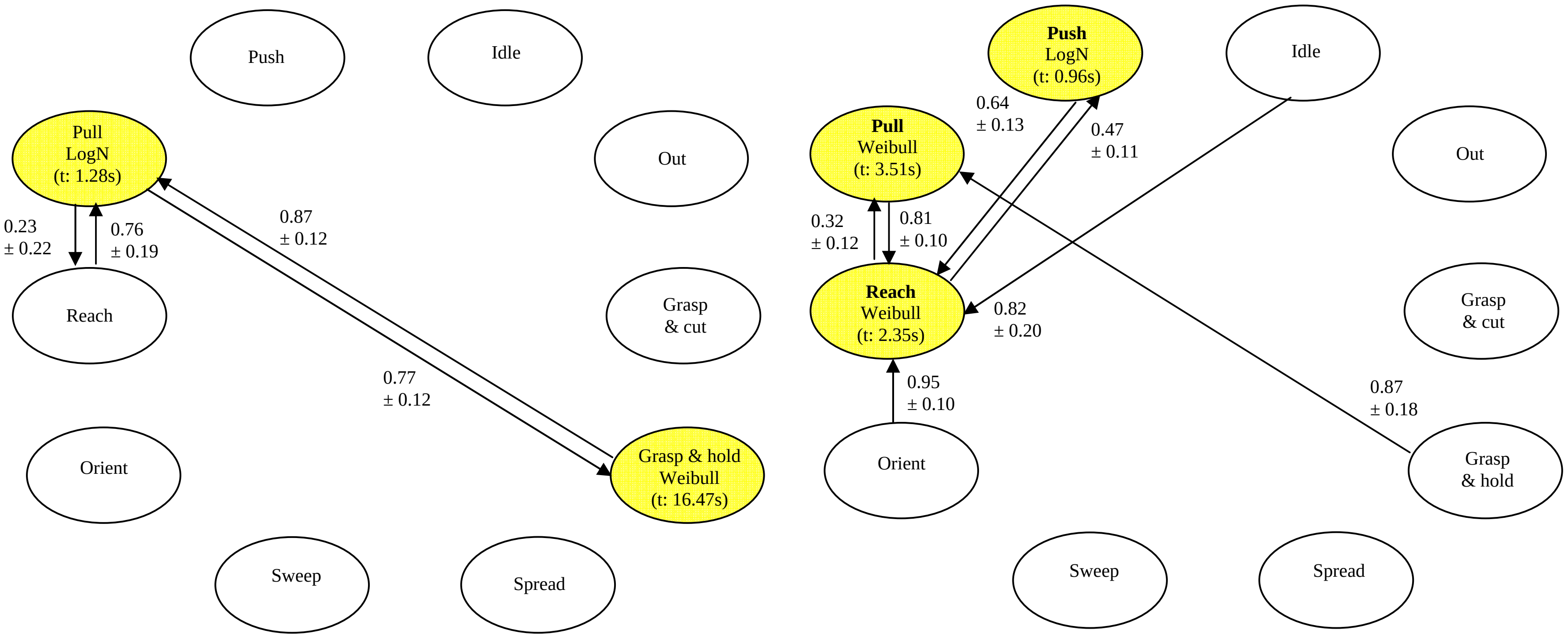
Expert 3: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



Resident 1: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)

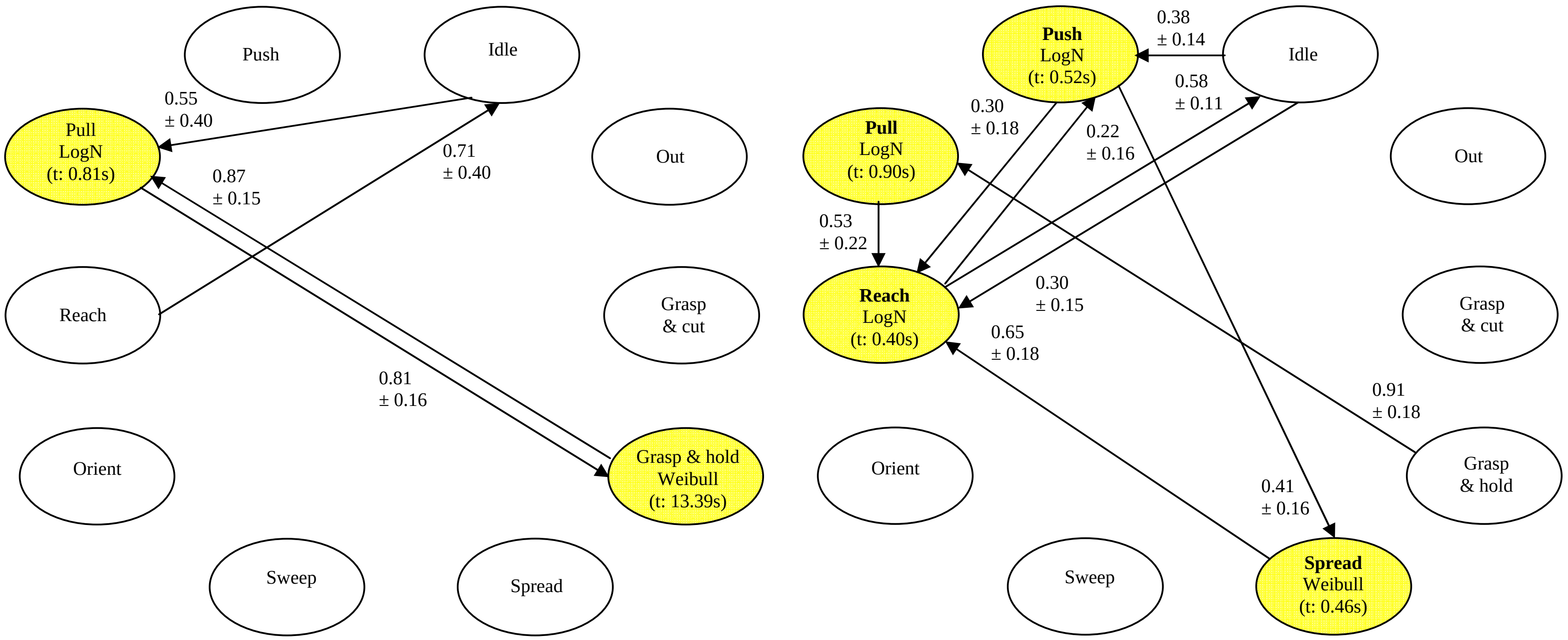


Resident 2: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)

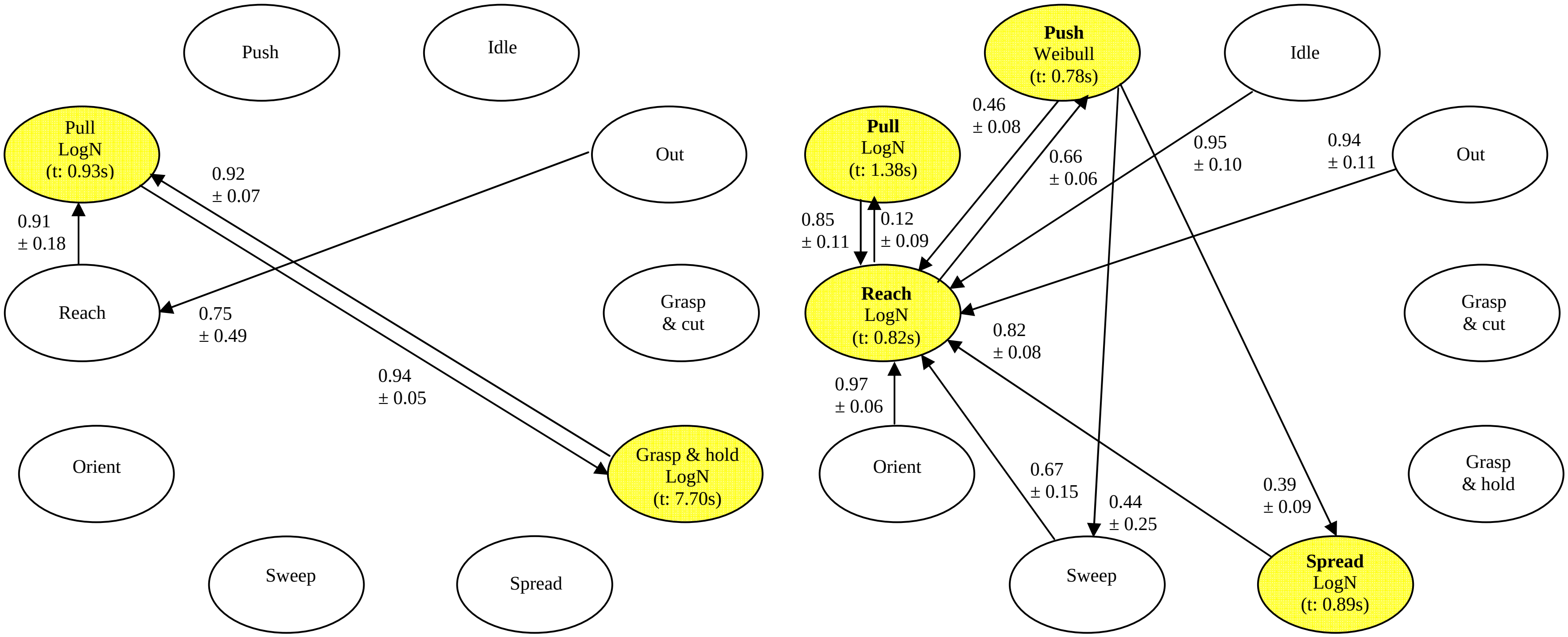


Resident 3: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)

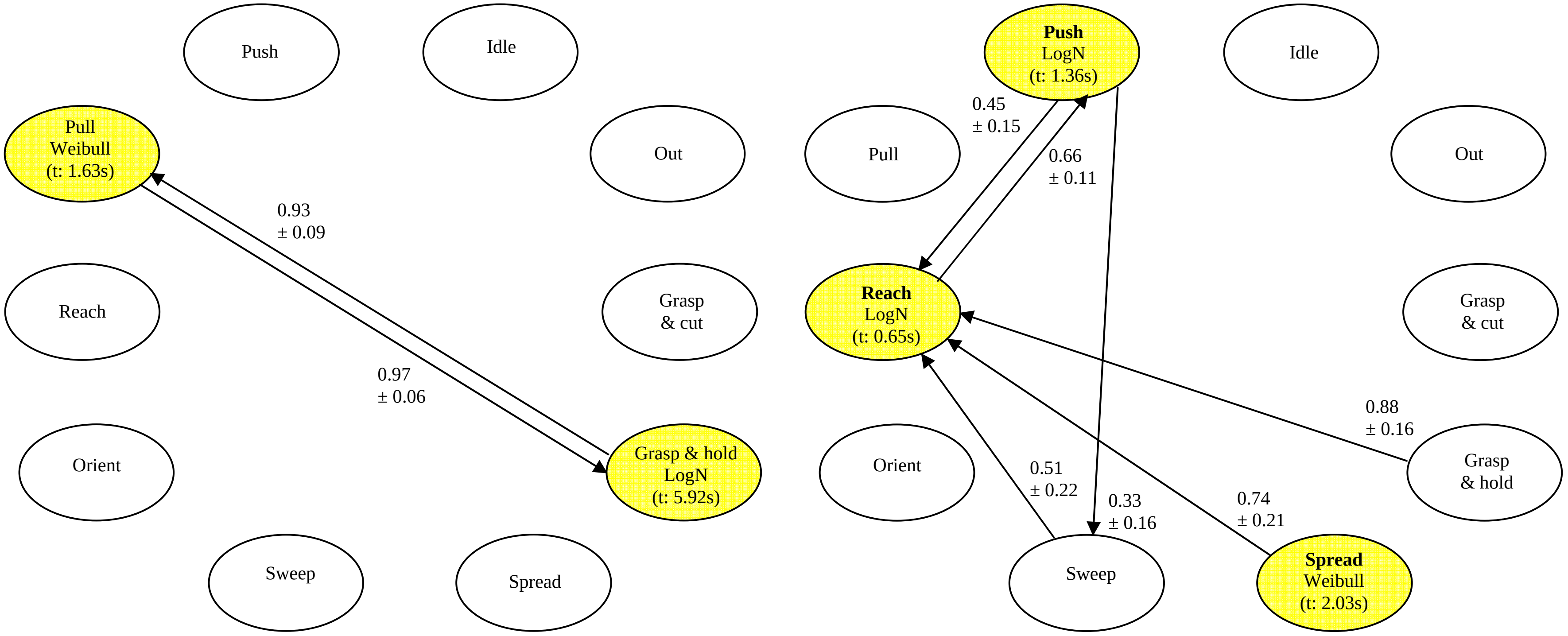
Dissect CD/CA



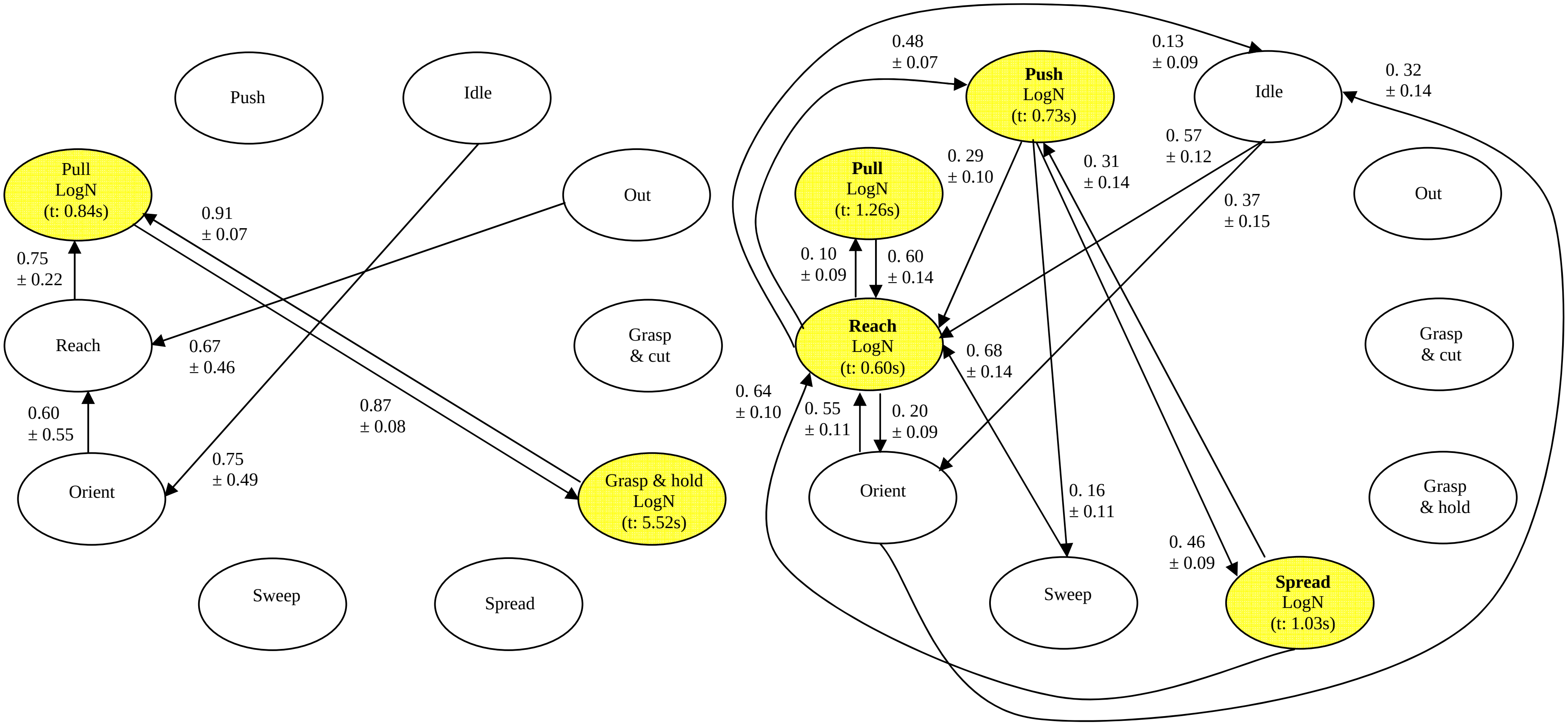
Expert 1: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



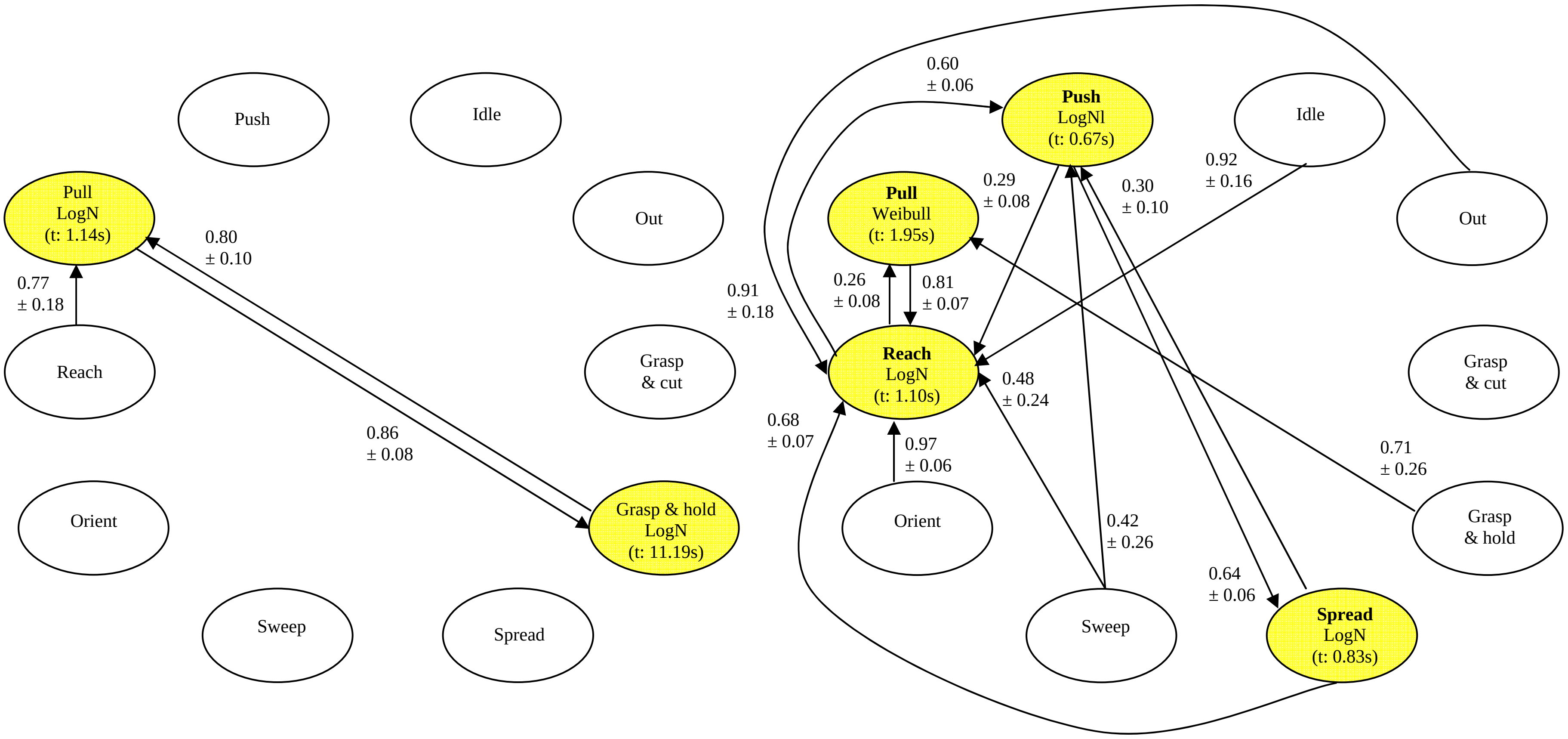
Expert 2: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



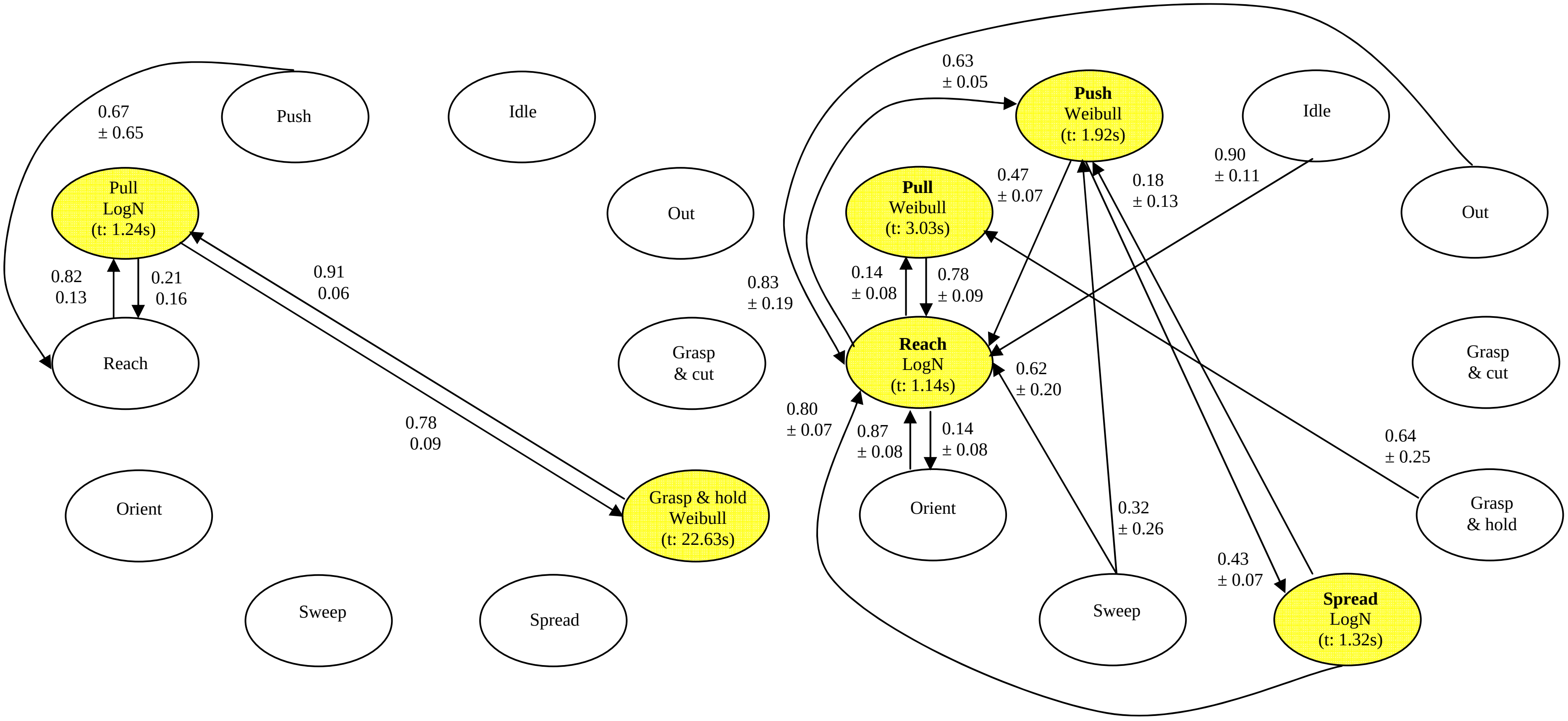
Expert 3: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



Resident 1: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



Resident 2: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)



Resident 3: Non-dominant hand (left) and Dominant hand (right);
Includes significant transitions ($CI_{tran.} < p_{ij}$) and average holding times (t)

APPENDIX E

OR Protocol and Ethics Documents

C.1 OR and tool cleaning protocol

Required equipment: Fastrak Polhemus system (interface unit, magnetic transmitter and receivers), Laptop, and testing surgical tool (disposable Maryland dissector)

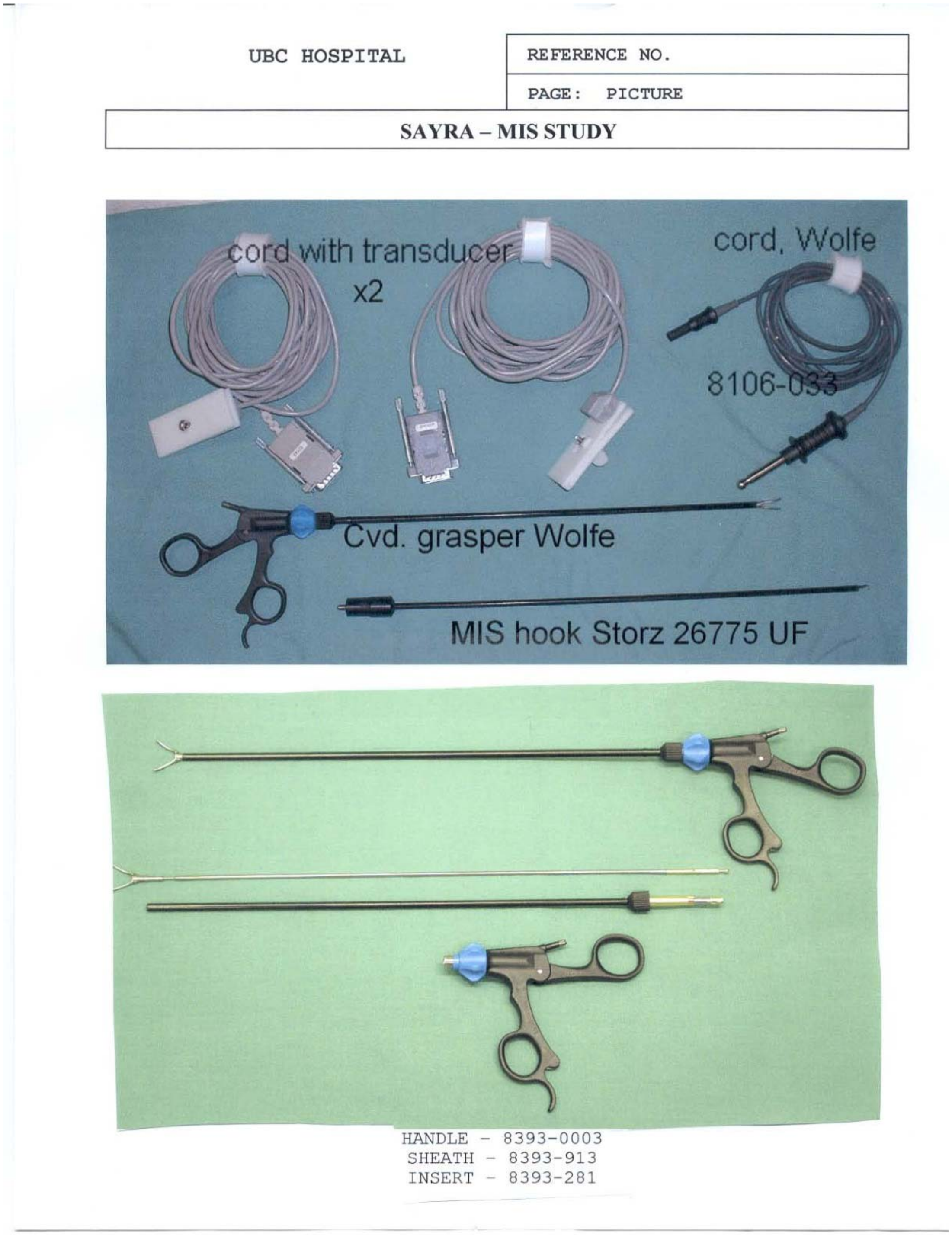
Start up time: 45min before scheduled set up of the operating room

Procedure:

- Pick up experimental tools at Sterile Processing Department
- Using the testing tool, train the surgeon in attaching the clip (it is the device that allows attaching the magnetic sensor to the tool shaft)
- Ask for surgeon’s signature on the consent form
- Initial equipment setup:
  - ✓ Connect laptop and Polhemus
  - ✓ Create two text files: freq_test.txt and case#1.txt
  - ✓ Attach the transmitter to the reference site (box) with duct tape
  - ✓ Use the testing tool to perform the frequency test (i.e., acquire 10 seconds of data and compute the actual sampling frequency) – see How to use GUI below
- Set up video recording:
  - ✓ Turn on STRYKER video equipment

  - ✓ Fill in case information
  - ✓ Select the ‘camera’ icon
  - ✓ Select the ‘film’ icon
  - ✓ It is ready to start recording
- Leave the operating room and wait until patient is anaesthetized
- Procedure begins and surgeon secures the clip to the experimental tools
- When the surgeon indicates that he/she is ready to use the experimental tools, ask him/her to carefully drop the cables on the floor so to not damage the sterile field.
- Plug in the cables to the tool interface units and turn it on
- How to use GUI:
  - ✓ Turn on Polhemus and wait until light stops flashing
  - ✓ Open FTGui
  - ✓ Load text file using the ‘logged’ button
  - ✓ Select ‘continuous’ mode
  - ✓ Select ‘options’, then ‘output’, and select ‘metric’
  - ✓ It is ready to start acquiring data
- Start simultaneously video recording and Polhemus equipments
- After gallbladder extraction, stop both equipments
- Save video on DVD from STRYKER video recording system, pick up tools from nurse station and leave OR

Cleaning protocol designed by Sterile Processing Department at UBC Hospital



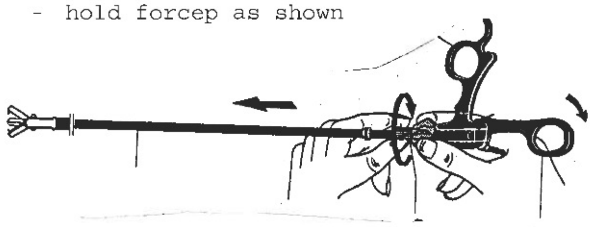
UBC HOSPITAL	REFERENCE NO.
	PAGE: PICTURE

SAYRA – MIS STUDY

HANDLE – 8393-0003
SHEATH – 8393-913
INSERT – 8393-281

UBC HOSPITAL	REFERENCE NO.
	PAGE: 1 OF 2
SAYRA - MIS STUDY	
DEPARTMENT: STERILE PROCESSING	LAST REVISION: JAN 2007
DISTRIBUTION: ASSEMBLY/DECONTAM	REVIEWED BY:
MASTER	APPROVED BY:

DISASSEMBLE:
- hold forcep as shown

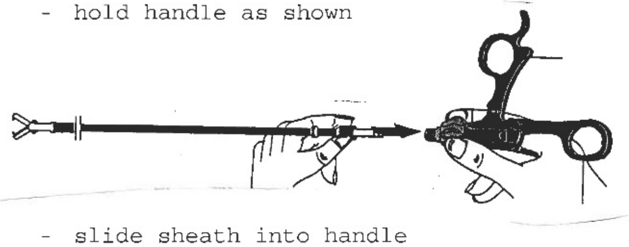


- turn knurled locking collar in direction of arrow and pull off sheath
- holding sheath firmly, turn the jaw insert and remove insert from sheath

CLEAN:
- brush lumen of sheath
- process all parts through washer disinfector

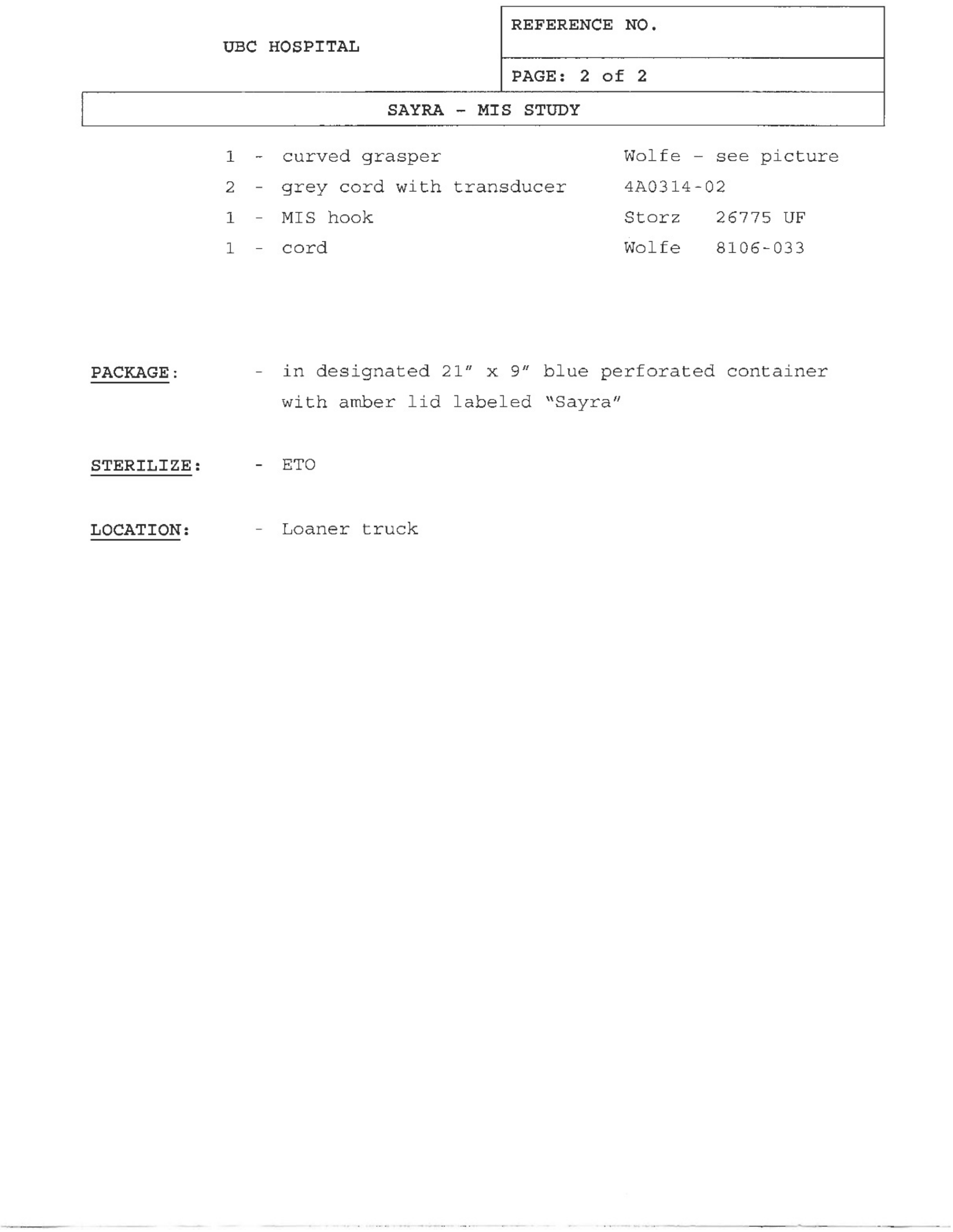
TEST:
- sheath must be tested with insulation tester

ASSEMBLE:
- insert jaw insert into sheath and turn to lock
- hold handle as shown



- slide sheath into handle
- turn sheath until it clicks into place
- screw on knurled collar
- test the function of the forcep, then **loosen the knurled collar one full turn**

cont…



UBC HOSPITAL	REFERENCE NO.
	PAGE: 2 of 2

SAYRA - MIS STUDY

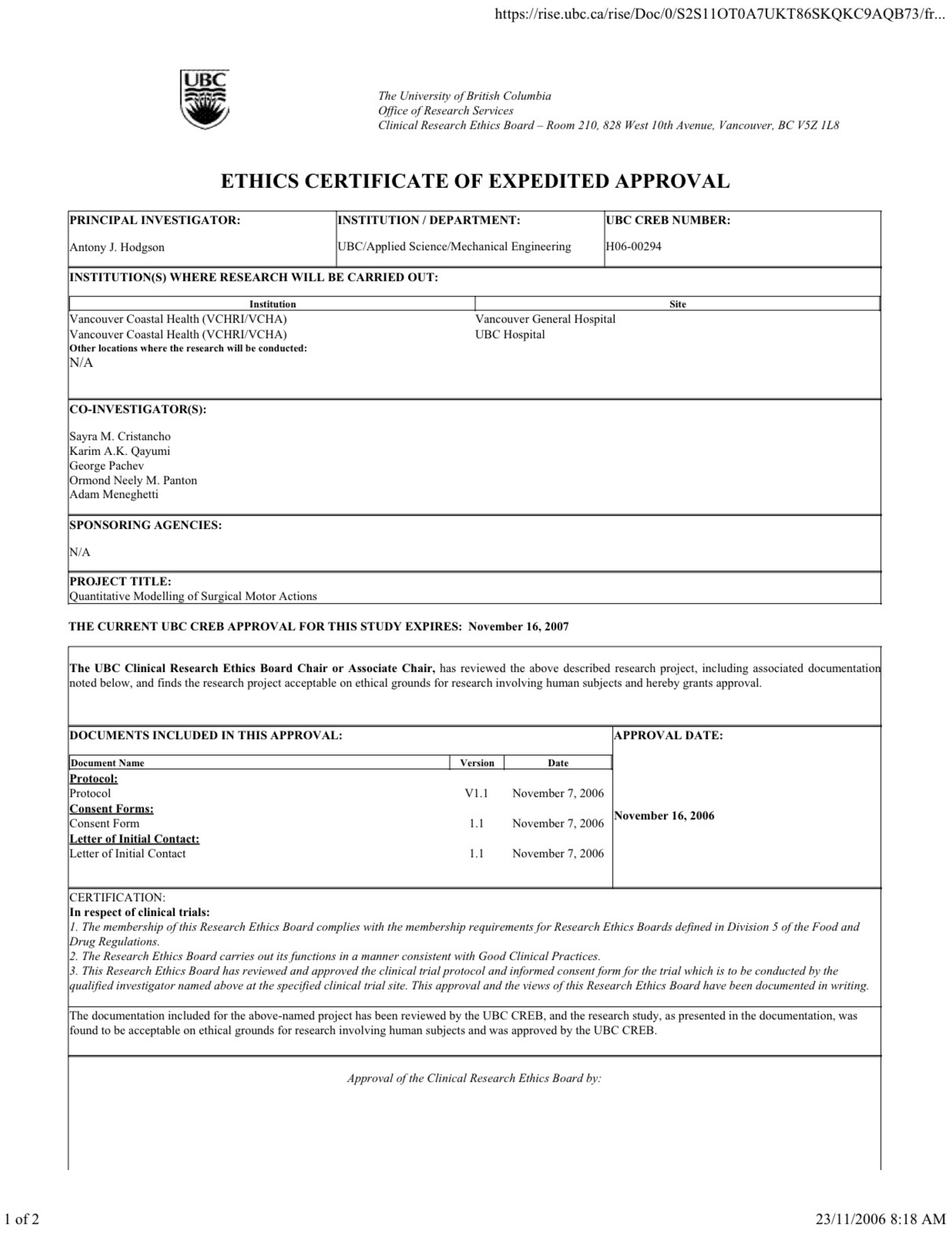
1 - curved grasper	Wolfe - see picture
2 - grey cord with transducer	4A0314-02
1 - MIS hook	Storz 26775 UF
1 - cord	Wolfe 8106-033

PACKAGE: - in designated 21″ x 9″ blue perforated container with amber lid labeled "Sayra"

STERILIZE: - ETO

LOCATION: - Loaner truck

C.2 Ethics Certificates of Approval



https://rise.ubc.ca/rise/Doc/0/S2S11OT0A7UKT86SKQKC9AQB73/fr...

The University of British Columbia
Office of Research Services
Clinical Research Ethics Board – Room 210, 828 West 10th Avenue, Vancouver, BC V5Z 1L8

ETHICS CERTIFICATE OF EXPEDITED APPROVAL

PRINCIPAL INVESTIGATOR:	INSTITUTION / DEPARTMENT:	UBC CREB NUMBER:
Antony J. Hodgson	UBC/Applied Science/Mechanical Engineering	H06-00294

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
Vancouver Coastal Health (VCHRI/VCHA)	Vancouver General Hospital
Vancouver Coastal Health (VCHRI/VCHA)	UBC Hospital

Other locations where the research will be conducted:
N/A

CO-INVESTIGATOR(S):

Sayra M. Cristancho
Karim A.K. Qayumi
George Pachev
Ormond Neely M. Panton
Adam Meneghetti

SPONSORING AGENCIES:

N/A

PROJECT TITLE:
Quantitative Modelling of Surgical Motor Actions

THE CURRENT UBC CREB APPROVAL FOR THIS STUDY EXPIRES: November 16, 2007

The UBC Clinical Research Ethics Board Chair or Associate Chair, has reviewed the above described research project, including associated documentation noted below, and finds the research project acceptable on ethical grounds for research involving human subjects and hereby grants approval.

DOCUMENTS INCLUDED IN THIS APPROVAL:

Document Name	Version	Date
Protocol:		
Protocol	V1.1	November 7, 2006
Consent Forms:		
Consent Form	1.1	November 7, 2006
Letter of Initial Contact:		
Letter of Initial Contact	1.1	November 7, 2006

APPROVAL DATE:

November 16, 2006

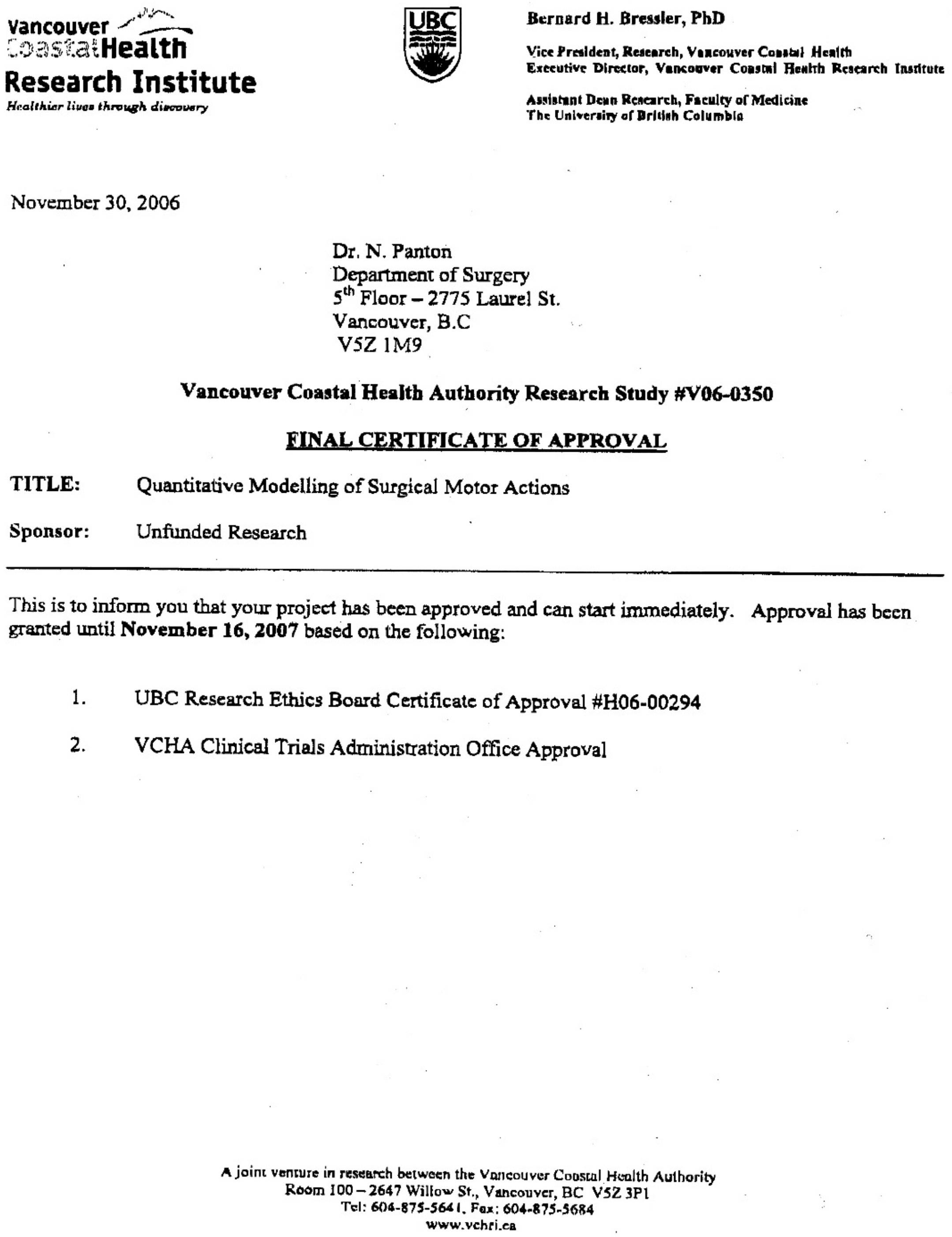
CERTIFICATION:
In respect of clinical trials:
1. The membership of this Research Ethics Board complies with the membership requirements for Research Ethics Boards defined in Division 5 of the Food and Drug Regulations.
2. The Research Ethics Board carries out its functions in a manner consistent with Good Clinical Practices.
3. This Research Ethics Board has reviewed and approved the clinical trial protocol and informed consent form for the trial which is to be conducted by the qualified investigator named above at the specified clinical trial site. This approval and the views of this Research Ethics Board have been documented in writing.

The documentation included for the above-named project has been reviewed by the UBC CREB, and the research study, as presented in the documentation, was found to be acceptable on ethical grounds for research involving human subjects and was approved by the UBC CREB.

Approval of the Clinical Research Ethics Board by:

1 of 2

23/11/2006 8:18 AM



Vancouver CoastalHealth
Research Institute
Healthier lives through discovery

Bernard H. Bressler, PhD

Vice President, Research, Vancouver Coastal Health
Executive Director, Vancouver Coastal Health Research Institute

Assistant Dean Research, Faculty of Medicine
The University of British Columbia

November 30, 2006

Dr. N. Panton
Department of Surgery
5th Floor – 2775 Laurel St.
Vancouver, B.C
V5Z 1M9

Vancouver Coastal Health Authority Research Study #V06-0350

FINAL CERTIFICATE OF APPROVAL

TITLE: Quantitative Modelling of Surgical Motor Actions

Sponsor: Unfunded Research

This is to inform you that your project has been approved and can start immediately. Approval has been granted until **November 16, 2007** based on the following:

1. UBC Research Ethics Board Certificate of Approval #H06-00294
2. VCHA Clinical Trials Administration Office Approval

A joint venture in research between the Vancouver Coastal Health Authority
Room 100 – 2647 Willow St., Vancouver, BC V5Z 3P1
Tel: 604-875-5641, Fax: 604-875-5684
www.vchri.ca

C.3 Request for subject participation and Consent Form

Quantitative Modelling of Surgical Motor Actions

Request for Participation

Would you like to help to apply and evaluate a new methodology for quantifying and assessing minimally invasive surgical performance so that future residents can keep track of their own progress?

We are performing a study to intraoperatively apply a new methodology for quantifying and assessing motor and cognitive aspects of surgical performance. As a participant, you will be asked to perform 2-3 laparoscopic cholecystectomy procedures as you normally do and under the supervision of your attending surgeon, using standard surgical tools. We will provide you with two small plastic clips to each of which will be attached a small magnetic sensor cube approximately 10-15 mm on edge. Whenever feasible during the procedure, you will clip the sensor to the tool you are currently using; the movements of the laparoscopic tool will then be tracked while you perform your normal surgical tasks. The surgery will be videotaped so that the investigator can later correlate the movement patterns with specific phases of the surgery. The acquired data will be processed afterwards to calculate the kinematic features (eg, velocities, accelerations, and jerks) of the tool movements.

If you are interested in participating, please call or email Sayra M. Cristancho (822-8785, scrista@interchange.ubc.ca) for information on enrolling in this study. Thank you for your consideration.

Principal Investigator: Neely Panton, MB, BS, FRCSC, FACS.

Graduate Student: Sayra M. Cristancho, PhD. Candidate, Department of Mechanical Engineering, (604)822-8785

Co-Investigator(s): Antony Hodgson, PhD, UBC Department of Mechanical Engineering, 604-822-3240; Adam Meneghetti, MD, UCSF; Karim Qayumi, MD, PhD; George Pachev, PhD.

This study is for the PhD thesis of Sayra M. Cristancho, UBC Department of Mechanical Engineering.

THE UNIVERSITY OF BRITISH COLUMBIA

Department of Mechanical Engineering
6250 Applied Science Lane
Vancouver, B.C. Canada V6T 1Z4

Tel: (604) 822-3240
Fax: (604) 822-2403

Vancouver, September 2006

To: Surgical Residents (4th and 6th year) at UBC
Subject: Recruiting research participants

Dear Potential Participant:
Would you like to help to apply and evaluate a new methodology for quantifying and assessing minimally invasive surgical performance so that future residents can keep track of their own progress?

We are developing a new methodology for quantifying and assessing motor and cognitive aspects of surgical performance. We will be analyzing performance data acquired during actual surgical tasks (performed in the operating room), and testing whether or not the methodology so developed is able to distinguish between trainees at different levels of development; we would therefore like to invite you to participate.

If you choose to do so, you will be asked to perform 2-3 laparoscopic cholecystectomy procedures as you normally do and under the supervision of your attending surgeon, using standard surgical tools. We will provide you with two small plastic clips to each of which will be attached a small magnetic sensor cube approximately 10-15 mm on edge. Whenever feasible during the procedure, you will clip the sensor to the tool you are currently using; the movements of the laparoscopic tool will then be tracked while you perform your normal surgical tasks. The surgery will be videotaped so that the investigator can later correlate the movement patterns with specific phases of the surgery. The acquired data will be processed afterwards to calculate the kinematic features (eg, velocities, accelerations, and jerks) of the tool movements.

If you are interested in participating, please call or email Sayra M. Cristancho (822-8785, scrista@interchange.ubc.ca) for information on enrolling in this study. Thank you for your consideration.

Participation is purely voluntary and potential subjects are under no obligation to participate.

Principal Investigator: Neely Panton, MB, BS, FRCSC, FACS.

Graduate Student: Sayra M. Cristancho, PhD. Candidate, Department of Mechanical Engineering, (604)822-8785

Co-Investigator(s): Antony Hodgson, PhD, UBC Department of Mechanical Engineering, 604-822-3240; Adam Meneghetti, MD, UCSF; Karim Qayumi, MD, PhD; George Pachev, PhD

THE UNIVERSITY OF BRITISH COLUMBIA

Department of Mechanical Engineering
6250 Applied Science Lane
Vancouver, B.C. Canada V6T 1Z4
Tel: (604) 822-3240
Fax: (604) 822-2403

Consent Form

Quantitative Modelling of Surgical Motor Actions

Principal Investigator: Neely Panton, MB, BS, FRCSC, FACS, UBC Faculty of Medicine, Dept. of Surgery
Graduate Student: Sayra M. Cristancho, PhD. Candidate, Department of Mechanical Engineering, (604)822-8785
Co-Investigator(s): Antony Hodgson, PhD, UBC Department of Mechanical Engineering, 604-822-3240
Karim Qayumi, MD, PhD, UBC Faculty of Medicine, Dept. of Surgery;
Adam Meneghetti, MD, UCSF, UBC Faculty of Medicine, Dept. of Surgery;
George Pachev, PhD, UBC Faculty of Medicine, Division of Educational Support and Development.

You are invited to participate in this study, which is for the graduate thesis of the student named above. The information gathered in this study (tool motions, as well as videotapes of the procedures) will be used to compare how subjects handle surgical tools in actual surgical tasks. The graduate student and the co-investigators will be the only ones with access to the data, and the identity of the participants will not be disclosed in any resulting publications. The results of this study will be used to design a future larger intraoperative study.

Purpose:

The purpose of this pilot study is to apply a new methodology for quantifying and assessing motor and cognitive aspects of surgical performance by analyzing performance data acquired during actual surgical tasks (performed in the operating room), and to test whether or not the methodology so developed is able to distinguish between trainees at different levels of development. You have been asked to participate in this experiment because you are currently in a hospital-based training position.

Exclusion Criteria:

Residents below their 4th year of training will be excluded since they are only allowed to perform selected aspects of the procedure and therefore, continuous data recording would be impossible.

Study Procedures:

If you choose to participate, we will ask you to perform 2-3 laparoscopic cholecystectomy procedures as you normally do (residents will perform under the corresponding attending surgeon's supervision), so your total commitment to this study will be the duration of these procedures (ie, up to ~5h). During the surgery, you will use standard surgical tools. We will provide you with two small plastic clips to each of which will be attached a small magnetic sensor cube approximately 10-15 mm on edge (ie, approximately sugar-cube-sized). Whenever feasible during the procedure, you will clip the sensor to the tool you are currently using; the movements of the laparoscopic tool will then be tracked while you perform your normal surgical tasks. The surgery will be videotaped so that the investigator can later correlate the movement patterns with specific phases of the surgery. The acquired data will be processed afterwards to calculate the kinematic features (eg, velocities, accelerations, and jerks) of the tool movements. We have carefully designed the measuring tool in consultation with attending surgeons so as to ensure that it will not interfere with performing the surgical procedure; however, should you at any time feel that it is interfering with your surgical activity, you may remove it (this process takes less than 1 second) and proceed without it.

Confidentiality:

Your confidentiality will be respected. No information that discloses your identity will be released or published without your specific consent to the disclosure. However, research records identifying you may be inspected in the presence of the Investigator or his or her designate by representatives of the UBC Research Ethics Board for the purpose of monitoring the research. _However, no records which identify you by name or initials will be allowed to leave the Investigators' offices.

The videotapes will be stored in a locked filing cabinet in the Centre of Excellence for Surgical Education and Innovation. The computer data will be stored on a secure computer system in files protected by a password known only to the investigators. The only identifying information that will be associated with any publication of the data will be the year level of your residency. The videotapes and data will be stored indefinitely and may be used again for derivatives or extensions of this research project, but will not be used for any other purposes without your explicit permission.

You will be able to ask either Dr. Hodgson or the responsible graduate student to review your own data at the conclusion of the entire study for purposes of feedback and personal improvement.

Benefits/Remuneration/Compensation:

You will be offered a $10 café gift card for participating in this study. Other than that, there are no explicit benefits you will receive.

Risks

You may experience some anxiety from being observed and measured. Otherwise, there are no known risks associated with this study.

Compensation for Injury:

Although we anticipate no increased risk of injury, signing this consent form in no way limits your legal rights against the sponsor, investigators, or anyone else.

Conflict of Interest:

None of the investigating team has any financial or other material interest in the outcome of this study, nor is it being sponsored by any entity with a financial interest in the outcome.

Contact for information about the study:

If you have any questions or desire further information with respect to this study, you may contact Sayra M. Cristancho at 604-822-8785 or Dr. Antony Hodgson at 604-822-3240.

Contact for concerns about the rights of research subjects:

If you have any concerns about your treatment or rights as a research subject, you may contact the Research Subject Information Line in the UBC Office of Research Services at 604-822-8598.

Right To Withdraw:

Your participation in this study is entirely voluntary and you may refuse to participate or withdraw from the study at any time without jeopardy to your standing in your training program. Data collected up to the point of your withdrawal from the study will be kept for data analysis purposes under the strict provisions of confidentiality described above.

Consent:

Your signature below indicates that you have received a signed and dated copy of this consent form for your own records and that you consent to participate in this study.

Your signature does not imply that you have waived any legal rights in agreeing to participate in this study.

____

Subject Signature Date

____

Printed Name of the Subject signing above.

Witness Signature Date

Printed Name of the Witness signing above.

Principal Investigator / Delegated Representative Date

Printed Name of the Principal Investigator or Delegated Representative signing above.

APPENDIX F

Discussion of 6D vs. 8D in Simulator Experiment

Since we had available time and kinematic data, for the PCA analysis we first defined a 8D data set composed of 6 rms (root-mean-square) velocity components (lateral, axial, and vertical velocities for 'Peel skin' and 'Detach segment' respectively) and the average time spent at each subtask.

The normalized PC coefficients (**Figure F.1**) indicated that times did not provide much information to the analysis since their contributions for the first principal eigenvector PC1 were considerably lower than those provided by velocities. This suggested that kinematics perform better than time in differentiating subjects' performances; therefore, we reduced the dataset and only concentrated on analyzing velocities in a 6D case for describing the main computations of section 4.3.1.2.

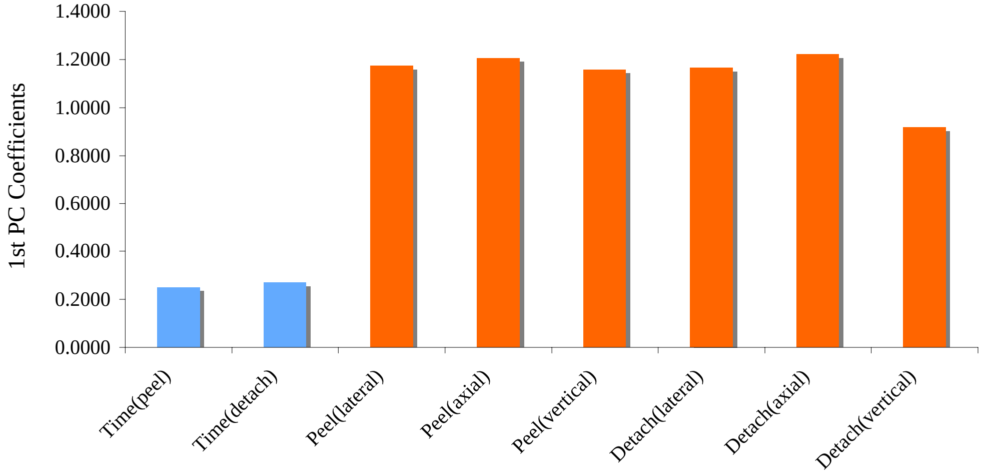


Figure F.1: Normalized (multiplied by $\sqrt{\# of \cdot PCs}$) coefficients for PC1 in 8-D data set. Blue color means time parameter; orange color means kinematic parameter

APPENDIX G
Variability analysis for PCA results – OR study

G.1 One-way anova [McDonald 2008]

When to use it

In a one-way anova (also known as a single-classification anova), there is one measurement variable (e.g., mean velocity) and one nominal variable (e.g., subjects in the residents' group). Multiple observations of the measurement variable are made for each value of the nominal variable.

Null hypothesis

The statistical null hypothesis is that the means of the measurement variable are the same for the different categories of data; the alternative hypothesis is that they are not all the same.

How the test works

The basic idea is to calculate the mean of the observations within each group, then compare the variance among these means to the average variance within each group. Under the null hypothesis that the observations in the different groups all have the same mean, the weighted among-group variance will be the same as the within-group variance. As the means get further apart, the variance among the means increases. The test statistic is thus the ratio of the variance among means divided by the average variance within groups, or Fs.

This statistic has a known distribution under the null hypothesis, so the probability of obtaining the observed Fs under the null hypothesis can be calculated.

The shape of the F-distribution depends on two degrees of freedom, the degrees of freedom of the numerator (among-group variance) and degrees of freedom of the denominator (within-group variance). The among-group degrees of freedom is the number of groups minus one. The within-groups degrees of freedom is the total number of observations, minus the number of groups. Thus if there are *n* observations in *a* groups, numerator degrees of freedom is *a*-1 and denominator degrees of freedom is *n*-*a*.

Assumptions

Any ANOVA test makes two assumptions: (1) normality, and (2) homoscedasticity which indicates that the within-group variances of the groups are all the same.

It is possible to test the goodness-of-fit of a data set to the normal distribution. McDonald does not suggest doing this, because many data sets that are significantly non-normal would be perfectly appropriate for an anova. Instead, if having a large enough data set, McDonald suggests simply looking at the frequency histogram. If it looks more or less normal, performing an anova is likely feasible.

In terms of homoscedasticity, the usual test for homogeneity of variances is Bartlett's test. This test is used when having one measurement variable, one nominal variable, and one want to test the null hypothesis that the variances of the measurement variable are the same for the different groups.

If the data do not fit the assumptions, it is often possible to find a data transformation that makes them fit. To transform data, a mathematical operation (i.e., log-transformation, square-root transformation) is performed on each observation, and then these transformed numbers are used in the statistical test. If any transformations are used to adjust the data to match the assumptions, then it is necessary to use either the Kruskal–Wallis or Welch's anova instead of one-way anova when there are more than two groups.

G.2 Nested anova [McDonald 2008]

When to use it

One uses a nested anova when having one measurement variable and two or more nominal variables. The nominal variables are nested, meaning that each value of one nominal variable (the subgroups) is found in combination with only one value of the higher-level nominal variable (the groups). Nested analysis of variance is an extension of one-way anova in which each group is divided into subgroups and subgroups into sub-subgroups, etc.

Null hypotheses

A nested anova has one null hypothesis for each level. In a two-level nested anova, one null hypothesis would be that the subgroups within each group have the same means; the second null hypothesis would be that the groups have the same means.

Assumptions

Nested anova tests, like all anovas, assume that the observations within each subgroup are normally distributed and have equal variances.

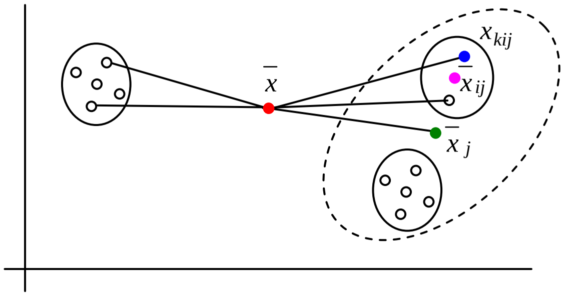
How the test works

In a one-way anova, the test statistic, Fs, is the ratio of two mean squares: the mean square among groups divided by the mean square within groups. If the variation among groups (the group mean square) is high relative to the variation within groups, the test statistic is large and therefore unlikely to occur by chance. In a two-level nested anova, there are two F statistics, one for subgroups (Fsubgroup) and one for groups (Fgroup). The subgroup F-statistic is found by dividing the among-subgroup mean square, MSsubgroup (the average variance of subgroup means within each group) by the within-subgroup mean square, MSwithin (the average variation among individual measurements within each subgroup). The group F-statistic is found by dividing the among-group mean square, MSgroup (the variation among group means) by MSsubgroup. The P-value is then calculated for the F-statistic at each level. For a nested anova with three or more levels, the F-statistic at each level is calculated by dividing the MS at that level by the MS at the level immediately below it.

In the present research, we tested log-transformation for our data; these seemed to produce adjustments which satisfied both normality and homoscedasticity assumptions. We therefore used ANOVA tests to test our null hypotheses.

G.3 Statistical analysis for the OR study

In order to show significance in group separation provided by the principal component analysis, we implemented a nested variability analysis (procedure-subject-group) in terms of distance measures on the PCA space by showing total variability to be the sum of three variability components : intrasubject, intragroup, and intergroup, based on the following analysis for a 2-dimensional PCA space.



k : trial # n_k : total # of trials, all subjects
i : subject # n_i : total # of trials per subject
j : group # n_j : total # of trials per group

By defining:
x_{kij} : position of trial ‘k’ for subject ‘i’ who belongs to group ‘j’
$\overline{x_{ij}}$: position of center of all trials for subject ‘i’ who belongs to group ‘j’
$\overline{x_j}$: position of center of all trials for all subjects belonging to group ‘j’
$\overline{x}$: position of overall center for all trials, all subjects, all groups

we then have :

$$V_T = \sum_j \sum_i \sum_k \left(x_{ki} - \overline{x}\right)^2 = \underset{\text{intrasubject}}{\sum_j \sum_i \sum_k \left(x_{kij} - \overline{x_{ij}}\right)^2} + \underset{\text{intragroup}}{\sum_j \sum_i n_i \left(\overline{x_{ij}} - \overline{x_j}\right)^2} + \underset{\text{intergroup}}{\sum_j n_j \left(\overline{x_j} - \overline{x}\right)^2}$$

As an example we present the variability computation and significance test for one of our operating room analyses.

Based on the PCA algorithm, we first computed the positions of each subject's trial (x_{kij}: E_{ik} and R_{ik}), each subject's center ($\overline{x_{ij}}$: E_i and R_i), each group's center ($\overline{x_j}$: E and R) and the overall center ($\overline{x}$).

	Point on PC1 axis	**Point on PC2 axis**
E11	-3.6362	0.5454
E12	-0.69	-0.2278
E13	0.8773	0.0472
E21	-1.2903	-0.4023
E22	-2.3522	-0.0706
E23	-3.7353	-0.1246
E31	0.0448	-0.0499
E32	0.434	-0.09
E33	0.1927	-0.3788
R11	1.7107	0.032
R12	0.3482	0.3139
R13	0.4581	0.267
R21	1.4062	0.226
R22	1.1498	0.0122
R23	1.4617	-0.0041
R31	0.5337	-0.3256
R32	1.3706	0.2461
R33	1.7162	-0.016

	Subjects' means	
E1	-1.1496	0.1216
E2	-2.4592	-0.1991
E3	0.2238	-0.1729
R1	0.839	0.2043
R2	1.3392	0.0780
R3	1.2068	-0.0318

	Groups' means	
Center E	-1.1283	-0.0834
Center R	1.1283	0.0835
Overall mean	-6.1679E-17	5.5555E-06

Distances between each subject's trial and subject's center (intrasubject), distances between each subject's center and group's center (intragroup) and distances between each group's center and the overall center were then calculated (intergroup), and the sum of squares (SS) were computed for each component.

$\sum_{j}\sum_{i}\sum_{k}\left(x_{kij} - \overline{x_{ij}}\right)^2$ intrasubject

$\sum_{j}\sum_{i} n_i\left(\overline{x_{ij}} - \overline{x_{j}}\right)^2$ intragroup

$\sum_{j} n_j\left(\overline{x_{j}} - \overline{x}\right)^2$ intergroup

Table1	Distances of each trial w.r.t each subject's mean	Dist. square
E11	2.5224	6.3626
E12	0.5774	0.3333
E13	2.0283	4.1140
E21	1.1865	1.4077
E22	0.1673	0.0280
E23	1.2782	1.6338
E31	0.2172	0.0472
E32	0.2259	0.0510
E33	0.2082	0.0434
R11	0.8886	0.7895
R12	0.5029	0.2529
R13	0.3860	0.1490
R21	0.1624	0.0264
R22	0.2005	0.0402
R23	0.1475	0.0217
R31	0.7344	0.5394
R32	0.3226	0.1041
R33	0.5096	0.2597

SS - Intrasubject 16.2040

Table2	Distances of subject mean w.r.t group's mean	Dist square	D^2 * 3
E1	0.2062	0.0425	0.1275
E2	1.3359	1.7847	5.3541
E3	1.3551	1.8364	5.5092
R1	0.3136	0.0983	0.2950
R2	0.2109	0.0445	0.1335
R3	0.1395	0.0195	0.0584

SS - Intragroup 11.4777

Table3	Distance of group men w.r.t to overall mean	D^2	D^2 * 9
Center E	0.4195	0.1760	11.4587
Center R	2.0432	4.1747	11.4587

SS - Intergroup 22.9174

We evaluated the null hypothesis that residents and experts all have the same means using a nested ANOVA test, and the mean sum of squares (MS) explained by intrasubject, intragroup and intergroups components were then plotted.

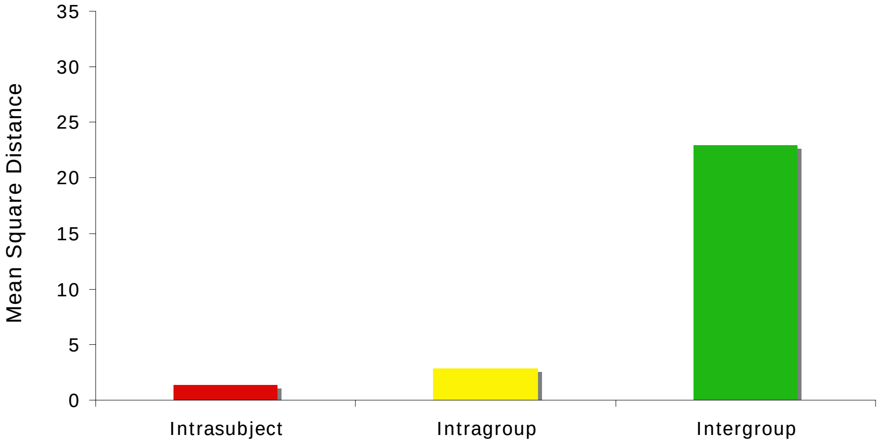
H01: Subjects within each group all have the same means

Source of variation	Mean square (MS)	Degrees of Freedom
Intrasubject	1.3503	12
Intragroup	2.8694	4
Fsubgroup	*P-value*	*F critical at α=0.05*[*]
2.12	**0.140**	**3.26**

H02: Experts and residents have the same means

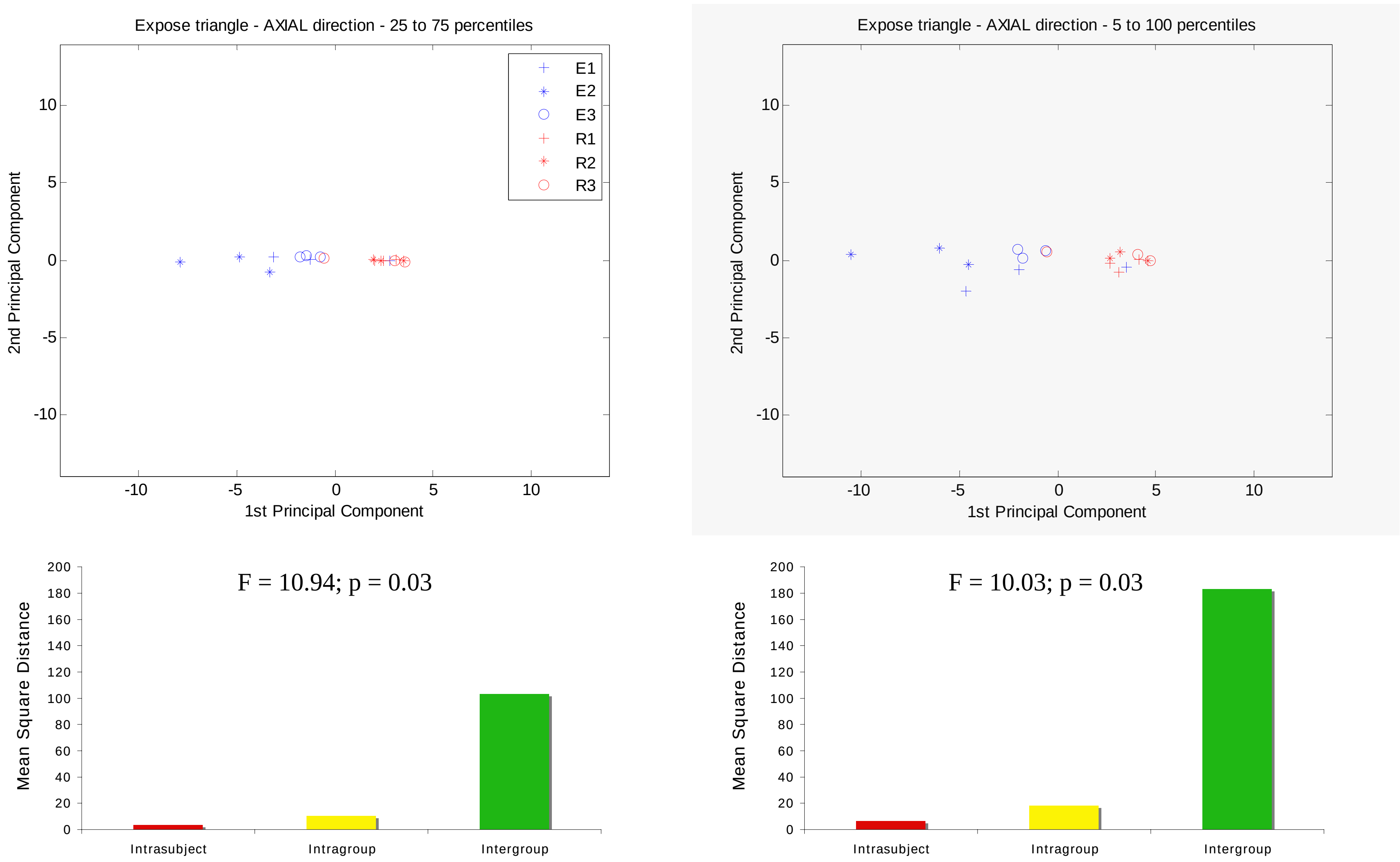
Source of variation	Mean square (MS)	Degrees of Freedom
Intragroup	2.8694	4
Intergroup	22.9173	1
F	*P-value*	*F critical at α=0.05*[*]
7.99	**0.047**	**7.71**

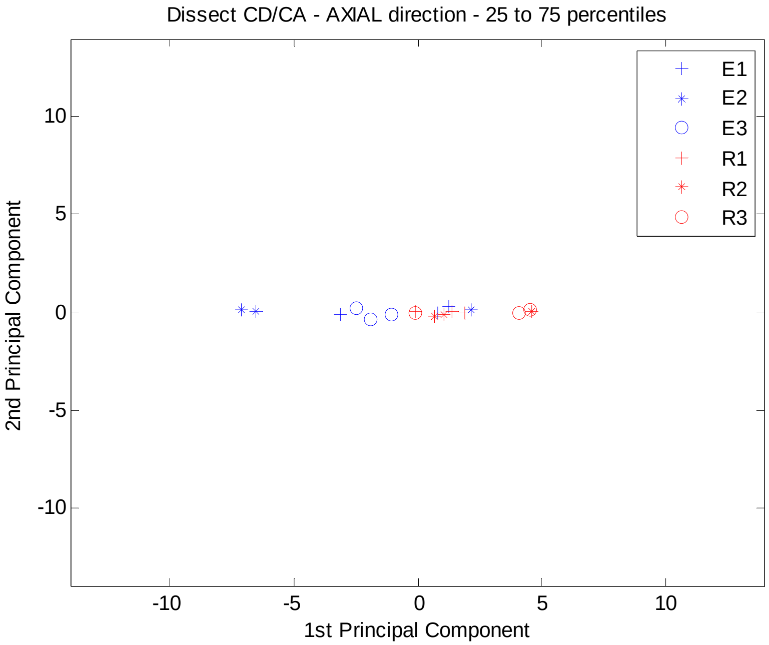
[*] From table of critical values for F distribution

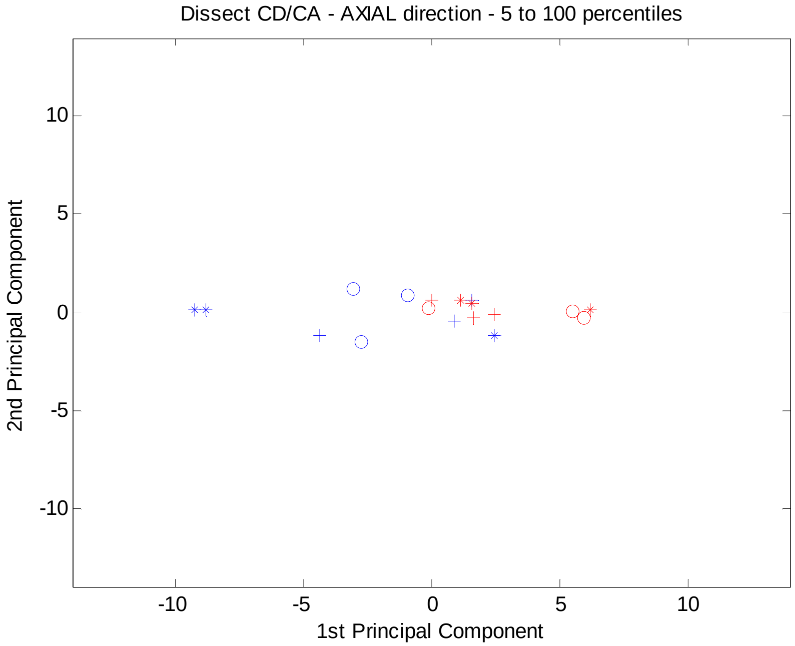


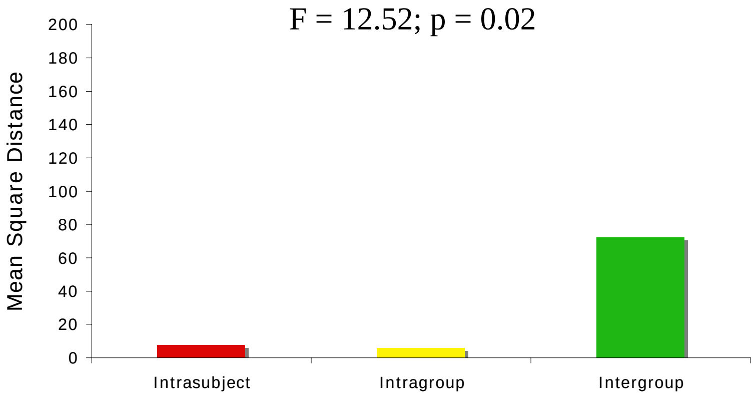
APPENDIX H

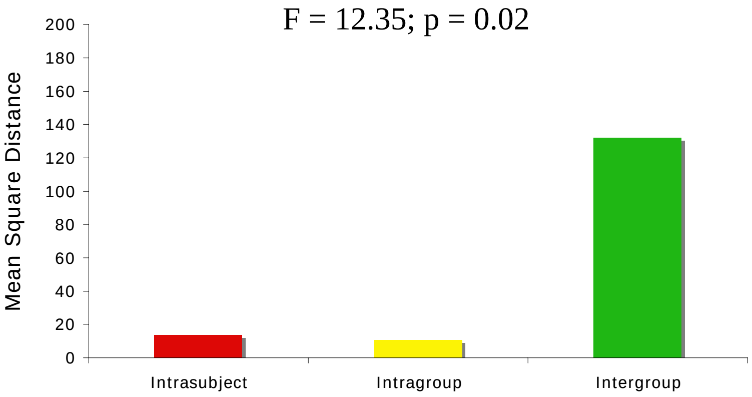
PCA and variability analyses using median (25th to 75th) vs. full percentile ranges (5th to 100th)




Dissect CD/CA - AXIAL direction - 25 to 75 percentiles
E1
E2
E3
R1
R2
R3
10
5
0
-5
-10
2nd Principal Component
-10
-5
0
5
10
1st Principal Component


Dissect CD/CA - AXIAL direction - 5 to 100 percentiles
10
5
0
-5
-10
2nd Principal Component
-10
-5
0
5
10
1st Principal Component


F = 12.52; p = 0.02
200
180
160
140
120
100
80
60
40
20
0
Mean Square Distance
Intrasubject
Intragroup
Intergroup


F = 12.35; p = 0.02
200
180
160
140
120
100
80
60
40
20
0
Mean Square Distance
Intrasubject
Intragroup
Intergroup