2013

AUTOMATIC PERFORMANCE LEVEL ASSESSMENT IN MINIMALLY INVASIVE SURGERY USING COORDINATED SENSORS AND COMPOSITE METRICS

Sami Taha Abu Snaineh
University of Kentucky, staha77@yahoo.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Sami Taha Abu Snaineh, Student

Dr. Brent Seales, Major Professor

Dr. Raphael Finkel, Director of Graduate Studies

AUTOMATIC PERFORMANCE LEVEL ASSESSMENT IN MINIMALLY INVASIVE SURGERY USING COORDINATED SENSORS AND COMPOSITE METRICS

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By
Sami Taha Abu Snaineh

Lexington, Kentucky

Director: Dr. Brent Seales, Professor of Computer Science

Lexington, Kentucky

2013

ABSTRACT OF THESIS

AUTOMATIC PERFORMANCE LEVEL ASSESSMENT IN MINIMALLY INVASIVE SURGERY USING COORDINATED SENSORS AND COMPOSITE METRICS

Skills assessment in Minimally Invasive Surgery (MIS) has been a challenge for training centers for a long time. The emerging maturity of camera-based systems has the potential to transform problems into solutions in many different areas, including MIS. The current evaluation techniques for assessing the performance of surgeons and trainees are direct observation, global assessments, and checklists. These techniques are mostly subjective and can, therefore, involve a margin of bias.

The current automated approaches are all implemented using mechanical or electromagnetic sensors, which suffer limitations and influence the surgeon's motion. Thus, evaluating the skills of the MIS surgeons and trainees objectively has become an increasing concern. In this work, we integrate and coordinate multiple camera sensors to assess the performance of MIS trainees and surgeons.

This study aims at developing an objective data-driven assessment that takes advantage of multiple coordinated sensors. The technical framework for the study is a synchronized network of sensors that captures large sets of measures from the training environment. The measures are then, processed to produce a reliable set of individual and composed metrics, coordinated in time, that suggest patterns of skill development. The sensors are non-invasive, real-time, and coordinated over many cues such as, eye movement, external shots of body and instruments, and internal shots of the operative field. The platform is validated by a case study of 17 subjects and 70 sessions. The results show that the platform output is highly accurate and reliable in detecting patterns of skills development and predicting the skill level of the trainees.

KEYWORDS: Computer Vision, Camera Synchronization, Motion Analysis, Pattern Recognition, Minimally Invasive Surgery Skills Assessment.

Sami Taha Abu Snaineh

Student's Signature

Date

AUTOMATIC PERFORMANCE LEVEL ASSESSMENT IN MINIMALLY INVASIVE SURGERY USING COORDINATED SENSORS AND COMPOSITE METRICS

By

Sami Taha Abu Snaineh

__Brent Seales_____
Director of Dissertation

__Raphael Finkel_____
Director of Graduate Studies

_____

To Humanity

Acknowledgement

I would like to thank my advisor Brent Seales for his guidance and support.

I would like to thank my committee members, Jinze Liu, Judy Goldsmith, and Lawrence Brook.

I would like to thank Dr. Melody Carswel and Michelle Sublette for their help and guidance in the human subject experiment design and IRB issues.

I would like to thank Dr. Ruigang Yang for offering his lab and the Vicon system to use my research.

I would like to thank Yongwook Song for his technical help in setting up and managing the Vicon system.

I would like to thank Jim Hoskins the director of the MIS Training Center in University of Kentucky for his help and guidance in understanding issues related to MIS training and assessment and for his help in setting up and designing the training task.

I would like to thank all the staff of the Computer Science department for their continual support and help.

Most of all, I would like to thank my parents Siham and Suleiman and my lovely wife Amani. Your endless support, love, and encouragement have been my continuous source of inspiration and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES *

* Some of the figures in this proposal were used from other studies under the doctrine of fair use.

Chapter 1

Introduction

This study aims at designing and developing an objective data-driven skills assessment for Minimally Invasive Surgery (MIS). The design employs and coordinates multiple sensors to extract metrics from various objects of the operation scene.

This chapter introduces the concept MIS and the challenges in assessing the performance and the skill levels of surgeons and trainees. We discuss the problem of the assessment along with the limitations of the current approaches. Also in this chapter we discuss the motivations that led to the study and its emergent contributions in improving the assessment. Finally, the chapter outlines the thesis structure.

1.1 Overview

MIS has improved in the last decade and is now popularly used. The typical evaluation techniques for assessing the performance of surgeons and trainees are direct observation, global assessments, and checklists. These techniques are mostly subjective and can, therefore, involve a margin of bias. Therefore, objectively evaluating the skills of the MIS surgeons has caused increasing concern among researchers. This research seeks to improve the MIS objectives of technical skills' assessment using the new technology of computer vision and multiple sensors. These technologies along with kinematic analysis and machine-learning will be used to improve and automate the

assessment process. Also, integrating several assessment techniques in one solution is expected to result in a more accurate and reliable solution.

The objective quantification and assessment of MIS technical skills requires a defined set of metrics. Therefore, techniques to acquire the correct metrics and an analysis model for the data to classify the surgeon's experience are key factors. However, all the previous studies acquired the metrics from either the surgical instruments' placement or the surgeon's hands movements. In addition, the metrics used in previous studies were unrelated, which decreased the reliability of the results. For example, none of the previous studies determined whether the motion may have taken place while the surgeon was looking at the display monitor or not. The typical approach used to read the metrics in the previous work was through electromagnetic sensors attached to either the surgeon's arms or to the surgical instruments. The limitation of what these sensors could measure might have been the reason for the previous studies not analyzing the relationship between the metrics.

Using a multiple-sensor system to study the problem of MIS assessment leads to extracting the relationship between different kinds of motions and developing a better metrics set for the assessment than what other studies used. For example, studying the motion and direction of the surgeon's head might lead to more useful assessment factors because it reveals the surgeon's hand-eye coordination, which is a critical skill in MIS. Also, tracking the surgeon's eye could lead to reliable assessment metrics since the eyes are the main factor in human activities. Eye-tracking and analysis of its metrics has not been studied before to assess MIS technical skills.

The idea is to build multiple non-invasive sensors coordinated in time over many cues of eyes, external shots of body and instruments, and internal shots of operative field. The system combines measurements of the surgical instruments, the surgeon's body movements of arms, head, and eyes in addition to heart rate factors. The coordinated-sensor environment allows us to extract a set of low-level metrics. The low-level measurements (non-fusion measures) are coordinated and combined to allow

higher-level measurements (fusion measures). For example this system provides the ability to analyze and study "blind motion," which is the motion of the surgical instruments or surgeon's hands while the surgeon is looking away from the monitor, or when the instruments are absent from the field of view. This analysis can classify risky or unimportant motion, which is a critical factor in the assessment and can reveal more reliable data than simple observations.

There are several novel ideas in this research that would improve the reliability of MIS assessment. The study utilizes multiple advanced vision systems to improve the assessment accuracy. Vicon, which is an advanced system to track the human body, is used in many researches and industry-areas such as, films, animations, gait analysis, and sports. However, it has never been used in the assessment process. The high accuracy of the Vicon system to track the motion and direction of the arms and the head can lead to higher accuracy in the assessment process than using electromagnetic sensors.

This research also aims to use the head motion and direction change in the assessment. No previous work has investigated how much these or the eye-tracking factors could improve the assessment accuracy. Therefore, the data from multiple vision systems can be used to develop new metrics by analyzing the relationship between each system. Using those ideas along with reading the heartbeat rate for the surgeon during the operation can lead to a robust, reliable, and valid assessment system. This system can be installed in the operating room or the training laboratories and can be helpful and time saving for the master surgeons and trainees. Further, integrating all these factors in order to explore their effects on the accuracy and robustness of the assessment is yet to be studied.

Even with the availability of ideal data, transforming the data into a skill level poses yet another challenge for the research. This challenge is due to the difficulty of quantifying human variability. In this research, mapping quantitative data into skill assessment is required in order to classify the surgeon. Analyzing high dimensionality metrics and quantifying them to reliable assessment measures is challenging. Part of the

thesis aims at discovering how best to analyze the data. There are a number of statistical, machine-learning, and data mining models that showed good reliability in classification, clustering, and finding hidden patterns in high dimensional data. To find the set of metrics that could accurately assess the surgeons and trainees, we used multiple data analysis methods. The methods used are Principal Component Analysis (PCA), a hybrid of partitioning and density-based clustering algorithms, and the neural network algorithm Multi-Layer Perceptron as classifier.

1.2 Motivation and Importance

Soon after the minimally invasive surgery revolution had started, the surgeons' qualifications to perform such operations became a concern. MIS has improved the surgical results for patients [1] and reduced the recovery time. However, it significantly complicated the task and increased stress of the surgeon[2]. Consequently, the study of the MIS ergonomics is being increasingly discussed among researchers and more research is being conducted to reduce stress, improve skills, and evaluate the operation. In addition, evaluating the trainees and increasing the safety of the patient has necessitated the measurement of surgical skills and performance [3]. Several studies and professional organizations like the Royal College of Surgery in England raised the issue and the importance of objectively assessing the surgical performance [4-7]. Thus, evaluating surgeon' skills in MIS has acquired paramount importance in all phases of surgical training and in surgical career in general. In surgical training, the evaluation is important to assess the level of expertise the trainee has gained, and the efficiency of the training process. Also, it gives feedback to the trainees at each step of the training process, which allows them to review and adjust their techniques accordingly. This results in decreased training time. For expert surgeons, the assessment of the surgical skills helps in the ergonomic studies of surgery. This in turn, offers them feedback about

their performance. Also, it contributes to evaluating and improving the training courses and techniques.

The evaluation process is currently subjective, and the most reliable techniques are direct observation, global assessments, and checklists [8]. Those techniques require expert surgeons to observe the trainees while performing surgeries, which demands time, effort, and resources. For example, assessing 20 trainees using the Objective Structured Assessment of Technical Skill (OSATS), a technique which will be discussed in detail later, requires 48 examiners, three hours each [9]. Therefore, the need for objective methods to evaluate MIS trainees is of paramount importance, and has motivated many researchers to look for other approaches.

Even though objective assessment is challenging due to differences in patients, operation setup, working team, and other factors [10], great efforts have been exerted in the past few years to develop objective evaluation techniques [11]. Therefore, MIS researchers in cooperation with researchers from other fields have developed different methods to objectively assess surgeons' skills. However, the literature demonstrates that no general solid and automatic solution has been implemented to assess surgeons as a standalone approach. Some of the methods mentioned in literature used virtual reality systems that assessed the surgeons in the virtual environment, but not in the operating theater where the motor performance could differ significantly [10]. Other methods used external sensors attached to the tools or the surgeon's body. Those sensors could be bulky and require and effort and knowledge to setup, as well as mandate that experts analyze the videos. Other common methods used expert observation, which is manual and subject to bias.

1.3 Thesis Statement

This research aims at improving the assessment of the MIS technical skills by designing and implementing a system that uses multiple non-invasive sensors and computer vision techniques. The proposed approach includes four parts:

1) Tracking the positions and direction of the surgeon's hands and head;

2) Tracking the positions of the surgery tools;

3) Tracking the surgeon's eyes; and

4) Reading the surgeon's heartbeat rate during the operation or the training session.

The results of the first three techniques will be transformed into kinematics parameters and compared to each other to produce other assessment metrics such as velocity, acceleration, deceleration, direction changes, path length, blind motion, blink rate, and fatigue. Then, an analysis model will be used to validate the system, analyze the produced data, and evaluate the surgeon. This research has taken place in the Center for Visualization and Virtual Environment (VIS) laboratories in the University of Kentucky.

The research is based on the hypothesis that integrating tools, arms, head, eyes, and heartbeat factors using advanced vision technology and appropriate data model can lead to great improvements to MIS technical skills assessments. Using multiple non-invasive, real-time, and coordinated sensors over many cues (eyes, external shots of body and instruments, internal shots of body and instruments) can transform the assessment problem to a new domain. Each part evaluates the subject from a different perspective. For example, eye tracking and measuring the fatigue level could assess the ability of the surgeons and their effectiveness of handling tools and performing tasks. Tracking motion and directions of the surgeon's hand in accordance to the display

location could assess hand-eye coordination. Using the kinematics of tools and arms demonstrates the ability of controlling and performing tasks. Heartbeat rate could lead to measuring the physical changes during the operation. Therefore, integrating all of those factors can lead to reliable, valid and applicable solution to MIS technical skills assessment.

1.4 Thesis Contribution Summary

This thesis contributed in:

- Identifying the limitations in the current assessment approaches.
- Proposing a novel design and implementation of a new assessment system.
- Proposing novel assessment metrics that have not been studied before.
- Opening new venues for expansions and more analysis to study other metrics.
- Validating the proposed design system and metrics.
- Building a data model of metrics that can classify the assessment level in three levels' resolution.
- Improving the reliability and accuracy of the objective assessment of MIS skills.

But one of the most valuable contributions made by this thesis is the transformation of the assessment problem by utilizing computer vision technology. This transformation allowed expanding the parameters of the assessment to increase the reliability. This transformation opened the door for more work and contributions to reach a satisfactory level to assess MIS trainees and surgeons. This could lead the computer vision researchers to improve other challenging issues facing minimally invasive surgery.

1.5 Thesis Content

The remaining chapters are organized as follows:

Chapter 2 (Minimally Invasive Surgery Assessment): This chapter reviews MIS assessment to understand the root of the problem. The review includes the types of MIS assessments and a survey of previous work to solve the problem. Then, it discusses the limitations of previous approaches and the challenges these methods face. Finally, the chapter gives a brief summary about the thesis approach to improve the accuracy and reliability of the assessment.

Chapter 3 (System Design and Architecture): This chapter introduces the design of the platform. The platform contains several parts where each part is discussed in detail with description of the requirements to build each part and the theory behind it. The chapter then describes the parts' integration and the time synchronization in order to minimize the capture offset between the subsystems. At the end, it compiles the list of metrics the system can extract and the details of how they are calculated.

Chapter 4 (Experiment Design): In order to test and validate the platform, we developed an experiment to collect data and analyze it. In this chapter, we introduce the experiment design and task description used to collect data, the protocol of recruiting subjects, training them, and assessing their skills level.

Chapter 5 (Data Collection and Processing): This chapter presents a high level of analysis of the metrics on individual basis. The analysis includes studying the correlation coefficient for each metric with the skill level and the variance of each metric within each level of skill.

Chapter 6 (Analysis and Discussion): In this chapter, we present detailed analysis using the Principal Component Analysis (PCA) to understand the features of the data and the interrelationship between subjects and variables, and to detect the skill patterns over time. Visualization of this analysis is provided in the form of plots and interpretation to

clarify the achievements and the interrelationship. The chapter in addition provides analysis to validate the clustering accuracy on different metrics. As a type of classification implementation, we trained a classification algorithm with subset of the data and validated it using the other subset and a 10-fold validation. The result of the classification is presented in this chapter.

Chapter 7 (Conclusion): This chapter summarizes the result achieved by the thesis and the overall contribution of this work toward improving the accuracy and reliability of performance assessment. The chapter includes a summary of the questions the research has answered.

Chapter 8 (Future Work): This thesis achieved answers to several questions but it opened up more questions simultaneously. As a part of this thesis contribution, the chapter describes more ideas, questions, and directions for future research to improve the performance assessment and skills-level recognition.

Chapter 2

Minimally Invasive Surgery Assessment

This chapter gives a detailed background about MIS assessment to enable an understanding of the root of the challenge. Then, it discusses the types of MIS assessments and provides a comprehensive survey of previous work to study the challenges and find solutions. The chapter then manifests the limitations of the previous approaches and the challenges these methods face. A brief discussion about our approach and contributions, and how it contributes towards improving the accuracy and reliability of the assessment follows.

2.1 Assessment Techniques

The literature presented various approaches to assess MIS surgeons and trainees. The common idea among most of the assessment approaches is to acquire metrics for different skill levels while surgeons perform surgical tasks. After the initial step, a statistical analysis is performed to find the correlation between the acquired metrics and the skill level. The common metrics used in the assessment methods are: time, position, motion, kinematics such as, speed and acceleration, force/torque, and others. We discuss each approach in this section along with the metrics used, and how they

have been analyzed. The following is a list of the assessment techniques that have been studied:

- Checklists, direct observation, and video tape observations
- Kinematics and motion analysis
- Virtual reality simulators
- Force/Torque analysis

2.1.1 Assessment Using Checklists, Direct Observation, and Video Tape Observation

The conventional methods of laparoscopic skills evaluation are using checklists, global assessment through direct observation, and/or video tape observation. In direct observation, the expert surgeons observe, assess the trainees and offer feedback about their skills. In video tape observation, the training process or the operation is recorded on a video, and the master surgeons assess the trainee by editing and observing the recorded video. Checklists of subtasks and specific skills are used with direct and video observation [9, 12]. Direct observation gives a better assessment than the latter, because the video fails to give complete information about the surgeon's knowledge of instruments and specific procedures, and efficient use of assistants. Although those methods are proven to be valid and reliable [9, 10, 13, 14], they are time and resource consuming. Further, they are subject to bias since it is an examiner's judgment. For example, evaluating 20 Residents using the Objective Structured Assessment of Technical Skill (OSATS), which is the most common method, took 48 certified surgeons three hours each [9].

OSATS was developed by Martin et al. [9]. It uses direct observation and the assessment is based on a task-specific checklist. The assessors directly observe residents performing surgical tasks on live animals or bench models. They use three types of scoring methods to assess the trainees. These scoring methods are task-specific

checklists for six procedural tasks, seven items of global rating score, and pass/fail judgment. Table 2.1 shows the global rating score and pass/fail judgment for OSATS.

2.1.2 Assessment Using Kinematics and Motion Analysis

The key concept behind the study of kinematics and motion analysis is to track the 3D space positions of objects such as, hands or instruments, then, analyze the data to produce a kinematic signature for each skill level. The main kinematics parameters used are: time, economy of motion, velocity, acceleration, and deceleration. Extensive research on analyzing the relationship between surgical skills and motion analysis, especially hand motion, has taken place recently. These studies in this area show the correlation between the motion and the skills level [15-20]. Therefore, many motion-tracking and analysis tools were developed in the past few years to serve as objective assessment tools for Laparoscopic Surgeons. Further, many tracking-tools and systems were developed to track the motion of laparoscopic instruments. Examples of advanced systems in MIS are the Imperial College Surgical Assessment Device (ICSAD) and the Advanced Dundee Endoscopic Psychomotor Tester (ADEPT).

Table 2.1 The global rating form used to assess technical skill at each of the eight stations in the Objective Structured Assessment of Technical Skill (OSATS). Global rating forms were used in conjunction with task-specific checklists

Global Rating Scale of Operative Performance **Please circle the number corresponding to the candidate's performance in each category, irrespective of training level.**					
	1	2	3	4	5
Respect for tissue	Frequently 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, motion and flow of operation and forward planning	Many unnecessary moves. Frequently stopped operating or needed to discuss next move.		Made reasonable progress but some unnecessary moves. Sound knowledge of operation but slightly disjointed at times.		Economy of movement and maximum efficiency. Obviously planned course of operation with effortless flow from one move to the next.
Knowledge and handling of instruments	Lack of knowledge of instruments.		Competent use of instruments but occasionally awkward or tentative.		Obvious familiarity with instruments.
Suturing and knotting skills appropriate for the procedure	Place the sutures inaccurately and tied knots insecurely and lacked attention to safety.		Knotting and suturing usually reliable but sometimes awkward.		Consistently placed sutures accurately with appropriately and secure knots with proper attention to safety.

Table 2.1 (Continued)

Technical use of assistants. **Relations with patient and the surgical team**	Consistently placed assistants poorly or failed to use assistants. Communicated poorly or frequently showed lack of awareness of the needs of the patient and/or the professional team.		Appropriate use of assistant most of the time. Reasonable communication and awareness of the needs of the patient and/or the professional team.		Strategically used assistants to the best advantage at all times. Consistently communicated and acted with awareness of the needs of the patient and/or of the professional team.
Insight/attitude	Poor understanding of the areas of weakness.		Some understanding of areas of weakness.		Fully understands areas of weakness.
Documentation of procedures	Limited documentation, poorly written.		Adequate documentation but with some omissions or areas that need elaborating.		Comprehensive legible documentation, indicating findings, procedure and postoperative management.
Over all on this task , Should the candidate:				Fail	Pass

ICSAD [8,21] has an electromagnetic tracking system which includes an electromagnetic field generator and two sensors. The sensors are attached to the back of the surgeon's hand. The tracking system is connected to a laptop that has software to analyze the tracked positions of the hands and retrieve the time, motion, velocity, acceleration, and deceleration of the hand movements. Since ICSAD sensors are connected to the surgeon's hands, it can assess real operation in the theater. However, ICSAD cannot measure rotational movements. In addition, the magnetic-field of the ICSAD could disturb the magnetic signals that might be present in the operation theater.

ADEPT [22] is a motion tracking system which uses mechanical methods to capture the motion of the laparoscopic tools. ADEPT is an advanced version of Dundee Endoscopic Psychomotor Tester (DEPT). The system has the standard MIS instruments, an endoscope, and a display. It tracks the three dimensional positions and the rotations of the tools' tips using a dual gimbal mechanism. The sprung top plate, through which the laparoscopic tools pass to perform a task, has access holes to allow tasks with various positioning. Figure 2.1 shows the diagrammatic representation of ADEPT. The gimbals capture the rotation of the tools about its axis using a core that is connected to a potentiometer. The depth of the tools is measured by a slider that passes through the center and is connected to a potentiometer. In addition, there are two more potentiometers to capture the XY values of the tools. In summary, ADEPT uses the following information to assess skills: execution time, instrument error, 3D-coordinates (XYZ), rotation angle, and completion status for all tasks. The reliability and the validity of ADEPT were discussed in many studies such as [23, 24]. In [24], 40 surgeons (20 experienced and 20 junior) performed tasks using ADEPT. The performance parameters used were instrument error, execution time, and task completion. The results show a significant difference in the instrument error between the experienced and the junior surgeons with a lower error rate for the experienced. However, the differences in the other two parameters, execution time, and task completion, were insignificant. The main challenge that faces ADEPT is the mechanical design limitation. The sensors that acquire the data are part of the training system. Therefore, it cannot be used in other training devices and environments or in the real operating theater.

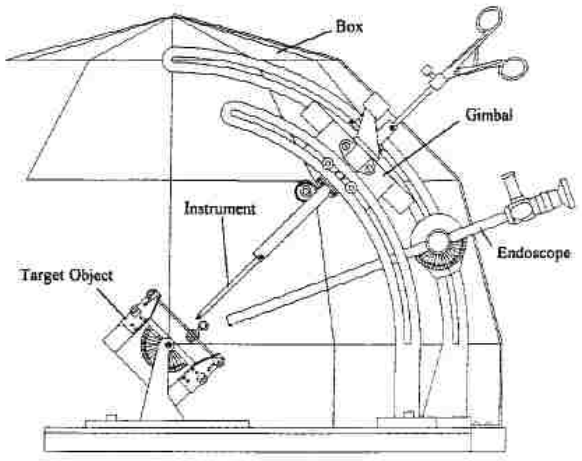


Figure 2.1 Diagrammatic representation of ADEPT [22]

Other researchers used cameras and infrared sensors to track the surgeon's body movement to study the MIS ergonomics. Similar approaches can use cameras to track the positions and collect kinematics data for both, instruments and surgeons for the assessment process. Gillette et al. [25] studied the postural parameter changes that occurred with different operation training tasks in their MIS ergonomics studies. Six cameras were used to track the motion of 37 reflective markers placed on different parts of the operator's body. The parts include the torso, head, upper arms, forearms, wrists, hands, and around each elbow joint. Emam et al. [26] in their laparoscopic suturing ergonomics study used a video-based motion analysis system called (Kinemetrix Model 5.0 3D/3MBM). This system includes three infrared cameras to track the motion of five high contrast markers placed on the surgeon's shoulder and elbow in addition to supination and pronation of the forearm. The parameters used in this study were the angles, the joint of the elbow, the shoulder, and the forearm supination and pronation.

Robotic Video and Motion Analysis Software (ROVIMAS) is software that reads the kinematics data produced by the Da Vinci robotic system and analyzes it to objectively assess the operators. The kinematics parameters that Da Vinci produces are path length,

number of movements, total time, average of path length, average of movements' path length, and velocity parameters for both hands [27]. Different studies have used ROVIMAS as a tool to study the robotic surgery learning curve and surgical skills [27-32]. This software can be used not only to assess Da Vinci robotic procedures, but also in other systems [31]. ICSAD is integrated to capture the kinematics data using electromagnetic sensors [32]. However, methods can be developed to capture kinematics data from non-robotic procedures and use it in similar way to assess the operator.

Cotin et al. in [20] have defined metrics to evaluate laparoscopic trainees in the simulation environment. The defined metrics relied on instrument motion and kinematics analysis of the motion. These metrics are: time to perform a task, path length, motion smoothness which is the change in the acceleration, depth perception which is the total distance an instrument travels along its axis, and response orientation which is the rotation of the instrument about its axis. To validate the proposed metrics, an experiment that includes 20 novice surgeons and a number of expert surgeons was implemented. Their motion was measured using a modified Virtual Laparoscopic Interface (VLI), and specialized software was developed for data processing and visualization of the motion. Each subject had to perform three training tasks multiple times. The results of the experiment demonstrate that the metrics could distinguish between the performances of experts and novices.

2.1.3 Assessment using Virtual Reality Simulators

Virtual reality simulators are used to educate MIS surgeons in the early stages of their training. In the last decade, several computer simulator systems have been developed. Separate sections for the simulators have been added because the simulator can provide more metrics for the assessment, such as, error score. In addition, the method of reading kinematics data in the simulator does not necessarily require tracking sensors or tools. The simulators can record metrics such as, time, positions,

path length, economy of motion, and other parameters. Therefore, researchers used simulators as objective tools to evaluate skills and to study the correlation of skills acquisition between the virtual systems and the actual operating theater. In this section, we describe some virtual reality systems and their usage as assessment tools.

The Minimally Invasive Surgery Trainer-Virtual Reality (MIST-VR) is a simulator that allows the trainee to perform simulated laparoscopic tasks using two standard laparoscopic instruments. The instruments are held in position-sensing gimbals which are connected to a computer [33]. The computer translates and reflects the movement of the instruments into the virtual instruments on the computer display. MIST-VR measures performance by recording and analyzing the completion time, error rate, and economy of movement of each instrument.

LapSim is another simulator that includes eight different tasks. These tasks can be performed through laparoscopic instruments that control the simulation software [34]. LapSim tasks are more realistic than MIST-VR's tasks because they simulate bleeding and structure deformation [35] in addition to providing tasks that are part of a real operation [34]. MIST-VR has been used in many studies and its reliability and validity have been proven overtime [15, 18, 19]. LapSim records various metrics and statistics of both, the left and right instruments to evaluate the performance of the trainee depending on the task being performed. Those metrics are total completion time, instrument navigation time, grasping time, angular path, instrument misses, lifting time, path length, clipping total time, incomplete targets, and blood loss [34]. Other available simulators are Xitact LS500 which incorporates physical objects and virtual abdomen with force feedback, ProMis, Reachin Laparoscopic Trainer, and LapMentor which enables the trainee to perform complete laparoscopic cholecystectomies with force feedback [36].

Kundhal and Grantcharov [37] studied the validity of using virtual reality simulators as an objective measure to evaluate the MIS skills of surgeons. The hypothesis was that the performance in the real operating room correlated with the performance in the

virtual reality simulation environment. A modified OSATS was used to assess the performance in the operating room. Seven tasks of the virtual reality laparoscopic trainer (LapSim) were used to assess the surgeons' performance. For more details about the seven tasks used, see Kundhal et al. [37]. Time, error score, and economy of motion were the primary assessment parameters used to differentiate between the skill levels in the real operating room. The assessment parameters for LapSim are error score, economy of movement, and time. Error score is evaluated by tissue damage, incomplete target areas, badly placed clips, and dropped clips. Economy of movement is evaluated by path length and angular path. Ten surgical residents of different gender and different skill levels participated in the study. Each subject performed three repetitions of seven tasks on LapSim and one laparoscopic cholecystectomy. The cholecystectomy was recorded and assessed by two experts using OSATS. Spearman's test was used for statistical analysis of the data. LapSim tasks and the operating theater tasks were found to be correlated.

The challenge that faces the virtual reality systems is that they can only be used in the virtual environment. The sensors that acquire the data are part of the system. Therefore, these systems cannot be used in other real training and operating environments.

2.1.4 Force/Torque

Force/Torque approach refers to measuring the magnitude and direction of the force and torque the surgeon needs to perform a task. Researchers suggested that the signatures of the force and torque can be used to evaluate technical skills because they are correlated with the experience level of the surgeons [38-42]. Different methods and interfaces measure these magnitudes and directions. Only one group of researchers used the measurement of force/torque for two types of interactions: Tool/Tissue interactions and Tool/Hand interactions.

Rosen et al. [38-41], studied the force/torque and haptic information from the tool/tissue interactions. The goal was to develop an objective laparoscopic surgical skill scale. A three-axis force/torque sensor was installed at the proximal end of a laparoscopic grasper, and a force sensor was installed on the grasper's handle. The sensors were set to measure the force/torque at the tool/tissue interface. Video-recording was used to define different tool/tissue interactions and synchronize each interaction with its corresponding force/torque data measures. The Hidden Markov Model (HMM) was then used on each subject to analyze the data. The results showed differences in force/torque values between subjects at different levels depending on their levels of training and expertise. Also, the results demonstrated similarity in the force/torque signature of subjects at the same level of experience. It is from those differences, that a learning curve of laparoscopic cholecystectomy operation has been developed. The experiment explained skill level differences in: magnitude of force/torque, the types of interaction between the tool and the tissue, and the time of each interaction. The video was analyzed manually by two expert surgeons to define the interactions and synchronize them with the force/torque data. The difference in the median completion time between the novice and the expert surgeons was significant [38]. The novice consumed more time in the idle state (where tools were idle) than the expert [38]. HMM showed differences in the statistical distance between subjects of different levels of experience, and the surgery skill learning curve converged exponentially to the expert level.

Richards et al. [42], measured the force/torque values from the tool/hand interactions. The goal was to study the force/torque values between the tool and the hand interface during each tool/tissue interaction. A three-axis force/torque sensor was installed at the proximal end of the laparoscopic grasper tube and a one-axis force sensor was installed on the grasper's handle. The sensors were calibrated to measure the torque and force at the hand/tool interface. As in [38], Richards et al. [42] used video recordings to define tool/tissue interactions and synchronize the interactions with their corresponding force/torque measures. Two expert surgeons analyzed the video

and defined the interactions. The data gathered from the experiment were analyzed using Vector Quantization analysis and clustered using the K-means algorithm. The experiment showed that the force/torque values in both operations between novices and experts were significantly different. The differences depended on the task being performed. As in [42], the novice surgeons used higher force/torque magnitude in tissue manipulation tasks, whereas in tissue dissection, they used lower force/torque magnitude. As in [38], [42] showed that the novice surgeons took more time to complete the operation than experts by a factor of 1.5-4.8. Richards et al. [42] noticed that the novices spend more time in idle state than the expert do. In the previous two studies, the video was analyzed manually by two expert surgeons to define and synchronize the interactions with the tissue. Thus the approach was not completely objective and automatic. Table **2**.**2** shows a summary of the MIS technical skills studies literature, types of the study, and the metric used.

Table 2.2 Summary of MIS technical skills assessments and the metrics used

Study	Category	Tracking System	No. Subjects	Assessments Parameter Used
Rosen et al. [38-41]	Force/Torque (Tool/Tissue)	NA	8	Time, force and torque magnitude and direction.
Richards et al. [42]	Force/Torque (Tool/Hand)	NA	10	Time, force and torque magnitude and direction.
Cristancho et al. [10]	Kinematics	Electromagnetic sensors	6	Time, position, and movement transitions.
Kundhal and Grantcharov [37]	Virtual Reality	Electromagnetic sensors	10	Time, error score, and economy of movement.
Gallagher and Satava [15] and Gallagher et al. [43]	Virtual Reality	NA	36	Time, error, economy of movement(left and right), and economy of diathermy
Smith et al. [44]	Virtual Reality	NA	10	Time, path length (left and right)
ROVIMAS software [27-30]	Robotic	Electromagnetic sensors and Robot API for capturing data	NA	path length, number of movements, total time, path length average, movements path length average and velocity
Grober et al. [16]	Real Operation	Electromagnetic sensors	2	total number of movements per hand, movement velocity, trajectory, and hand travel distance
Aggarwal et al. [31]	Real Operation	Electromagnetic sensors	19	Time taken, path length, and number of movements for each hand
Dosis et al. [32]	Real Operation	Electromagnetic sensors	5	Number of movements, path length, and time taken in addition to OSATS.
Bann et al. [45]	Lab	Electromagnetic sensors	30	Number of movements, time.
Datta et al. [46]	Lab	Electromagnetic sensors	50	Number of movements, time.

Table 2.2 (Continued)

Hernandez et al. [47]	Lab	Electromagnetic sensors	13	Time, number of movements, and path length in combination with OSATS.
Bodten et al. [4]	Augmented Reality	Cameras	24	Time spent in correct area, strength of the knot.
Chmarra et al. [48]	Lab	Electromagnetic sensors	31	Time, path length, depth perception, motion smoothness, angular area, and volume.
Jayender et al. [49]	Lab	Electromagnetic sensors	13	Time, position, path length, velocity, acceleration, and orientation.
Salgado et al. [50]	Simulator	NA	8	Time, error score, efficiency of movements, instruments misses, and tissue damage
Cotin et al. [20]	Simulator	NA	20+	time , path length, motion smoothness, depth perception , and response orientation
Megali et al. [51]	Simulator	NA	29	Mathematical parameter defined using kinematics parameters.
Allen et al. [52]	Lab	Electromagnetic sensors	30	Time to completion, path length, volume, and control effort
Francis et al. [24]	Lab	Mechanical	40	Instrument error, execution time, and task completion.
Hanna et al. [53]	Lab	Electromagnetic sensors	10	Errors rate, the execution time, and the applied force on the target.
Sokollik et al. [54]	Lab	Ultrasound system	56	Time, position of the instrument in space, number of errors, standardized time, error time, distance efficiency ratio, speed profile, and transit profile.

To summarize the technical methods of assessing the MIS skills, in general, all methods propose extracting information from surgery tasks. We can use this information to develop a model to assess the skills of the surgeons. The differences are in the parameters and the analysis approach used. Most of the studies are based on time and kinematics parameters in the assessment [4, 10, 15-21, 31, 43, 44, 49, 51-55]. The usage of the simulators to evaluate the performance of the trainees in the virtual environment can support the assessment process but cannot serve as a standalone assessment tool. Since the simulation approach does not confirm that the training skills acquired in virtual environment are transferable to the operating theater on an individual basis, the need for an evaluation system in those two environments is still required to confirm whether the trainee has mastered the skills in the area.

Most of the studies used electromagnetic sensors to track and record the positions of the instruments or the surgeon's hands and transform that information into kinematics data [8, 10, 16, 27-32, 45-49, 52, 53]. A few used the Da Vinci robot API system [27, 55] and one study used an ultrasound tracking system [54]. Another approach was to measure the force/torque required to perform a surgical task [38-42]. The studies that investigated the force/torque approach used small sample sizes. The use of such limited samples made it hard to extrapolate and generalize the validity of the approach. Using cameras also made the system bulky, but this bulkiness was isolated from the tools and the surgeon's body and did not interfere with the surgeon's motion. However, the tracked object could be lost if the reference target went out of the camera's field of view while the surgeon was moving. This scenario could cause loss of data. Using the Vicon system could however, minimize the data loss. As the system describes in the next section, the eight cameras can be used so that their field of view can cover all angles of the operating theater. This setup increases the reliability of the tracking system and minimizes the data loss.

2.2 Limitations of the Previous Work

Most of the referenced studies used expert surgeons to manually edit and segment the video into tasks and synchronize it with the kinematics data in order to analyze the skills of the surgeon performing the operation. This approach requires valuable time and resources from the experts. In addition, the trainees will not get continuous feedback on their progress if the expert surgeons are not available.

The systems that use mechanical technology, virtual environment, and robotics surgery instruments are closed to their environment. The sensors that acquire the data are part of the system. These systems cannot be used in all tasks or different operating environments.

Attaching external sensors to the surgical instruments or to the surgeon's body suffers several drawbacks. First, even though the electromagnetic and force/torque sensors are small, because they are attached to the surgeon's body, they might interfere with the surgeon's work. Second, the electromagnetic sensors can be affected by magnetic fields in the surgery and training environment. But the main limitation is studying an isolated type of motion or measure such as, the tools' motion and ignoring the importance of the coordination between the motion of different body and instrument parts. The motion of the tools, hands, head, and eyes are not studied together to find the importance of their interrelationship. The interaction between the head and the eye with objects in the environment, which drives the motion, has never been studied. The reason for this limitation is that the technology used does not provide the capability to extract these measures.

2.3 Importance of the New Approach

None of the studies used the abilities of computer vision. Probably the reason is that the level of the development bar has been too high to use real computer vision in this area. But now that the algorithms are robust enough, computer vision can be a game changer to transform the domain of the assessment challenge. Building and synchronizing a network of camera sensors to extract new metrics from the synchronized motion of different parts and the relationship among each other could be a significant contribution to the improvement of the assessment reliability and accuracy.

The use of the non-invasive camera setup can expand the metrics acquired to assess the skills. For example, the surgeon's head can be tracked and the period of time of looking at the display versus the time looking at the instruments can be studied. This metric can measure the eye/hand coordination which is an important skill to the surgeon. The technology of eye-tracking systems is another option to develop new metrics in the assessment methods. Tracking the surgeons' visual motion may be used to distinguish the levels of experience. Other proposed approaches are to read the physiological changes inside the surgeon's body. Parameters like heartbeat rate and the change of the adrenaline level in the body could be strong qualitative measures of the surgeon's confidence and skills.

The new system design provides several features that are lacking in other systems. The new design encapsulates the bulkiness of the assessment tools from the training or operation environments and minimizes the influence of assessment tools on the surgeon activities. The design also encapsulates the assessment tool from the surgical environment. This encapsulation increases the usage validity of the system in all training and operation environments. The new design provides the capability of fusion motion analysis to capture composite metrics which represent the coordination between different moving parts in the environment. Therefore, the quality of the metrics used would improve since the system distinguishes between the types of motions and

metrics. Finally, studying a wide range of metrics which can only be acquired by this design from different moving parts, including composite metrics, increases the reliability of the assessment and the tolerance to noise and errors.

Chapter 3

System Design

This chapter describes the design of the system. The system includes several parts, which will be discussed in detail. The chapter also presents the description of the used tools, the theory, the challenges faced, and the solutions to overcome them. Finally, the chapter lists the metrics that the system measure, the description of those metrics, and how they have been captured and calculated.

3.1 High Level Architecture

The multi-sensor platform is composed of four subsystems in which each system contains one or more non-invasive sensors. Three of them contain camera sensors and one is a heartbeat rate monitor. Each of the three camera subsystems is responsible for tracking and capturing the 3D positions of one or more moving parts in the surgical or training environment. These subsystems are synchronized and coordinated in order to calculate fusion or combined metrics. The heartbeat rate monitor is responsible on tracking the heartbeat rate during the session activity. Figure 3.1 shows the high level architecture of the following four subsystems.

- Vicon System contains eight MX3+ Cameras
- Eye tracker contains a stereo camera
- Laparoscope contains a stereo camera

- Heartbeat rate monitor

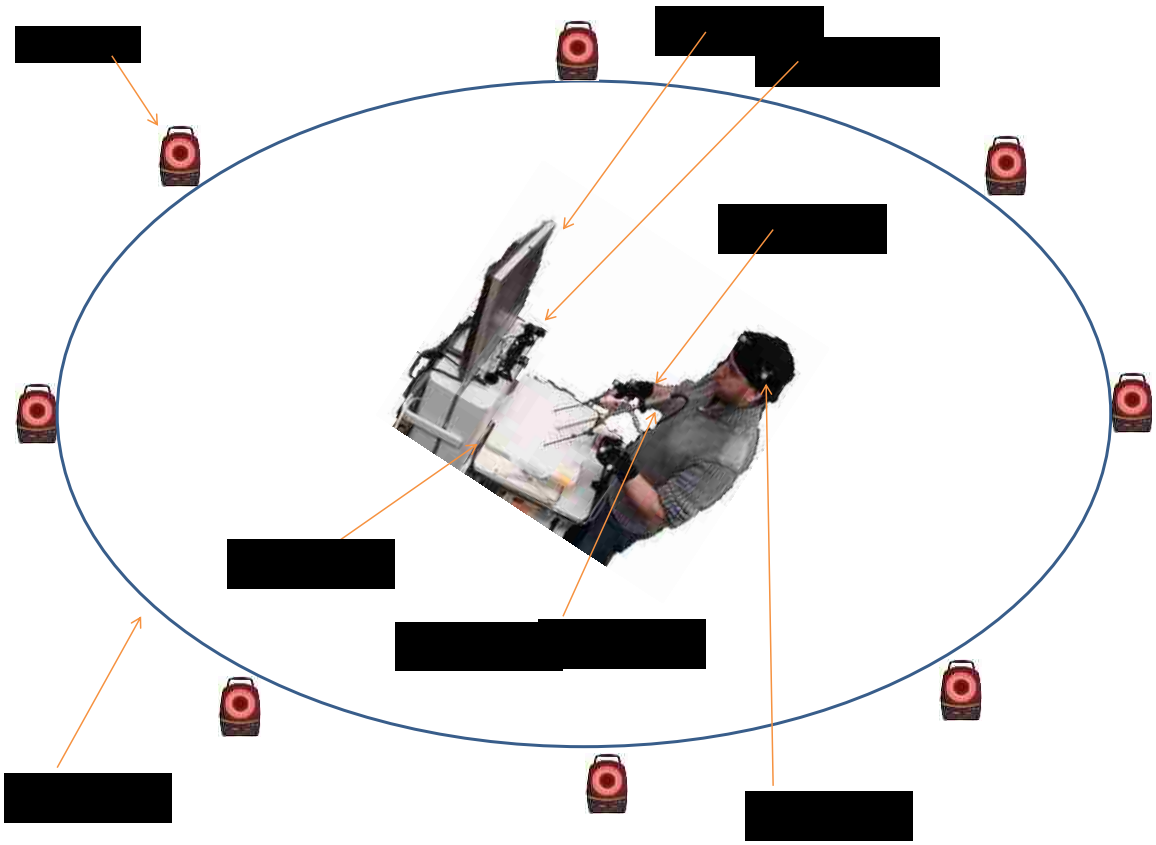


Figure 3.1 High level architecture of the four subsystems

The multi-sensor platform synchronizes and processes the captured data from the subsystems and transforms it into fusion and non-fusion assessment metrics. Those metrics are then analyzed to find their significant and to build a data model to find the hidden patterns of skill levels and to classify the skill level of subjects. The sensors capture data for the surgical instruments, the surgeon's head, hands, eyes, and heartbeat rate. Figure 3.2 shows the block diagram of the data flow in the multi-sensor system. The diagram shows the parts that are monitored by different sensors, the

synchronization step, and the data processing. At the end of the analysis, the fusion and non-fusion metrics are produced to find the skill level.

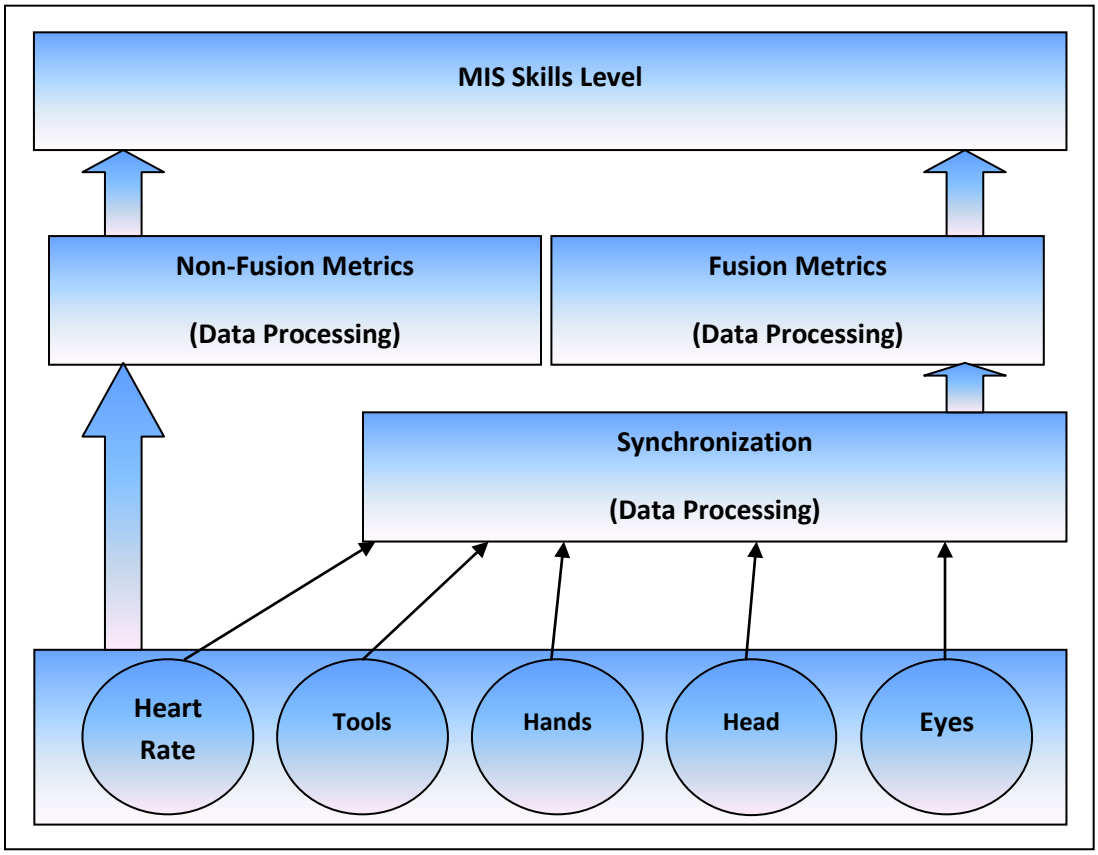


Figure 3.2 Block diagram of the multi-sensor system's data flow

3.2 Tools and Sensors

In this section, we describe each subsystem and how it is used to transform raw data into assessment metrics. In addition, it also presents an overview of each tool, the theory to handle it, and its role in the research.

3.2.1 Endoscope and MIS Tools Detection

MIS uses several types of long and narrow mechanical instruments in addition to one laparoscope. The Laparoscope is a video camera (single or stereo) that transmits a 2D video stream to a screen display. The MIS instruments are hand-controlled tools where the surgeon relies on the video display to move them in order to complete the surgical tasks. These tools are typically inserted through a 0.5-1.5 cm incision [59]. Figure 3.3 shows a sample of one instrument. The MIS operations usually require skills such as grasping, pushing, pulling, cutting, transferring, suturing, knot tying, and needle manipulation. Different tasks require different types of tools [38]. Figure 3.4 shows the MIS operation room setup.

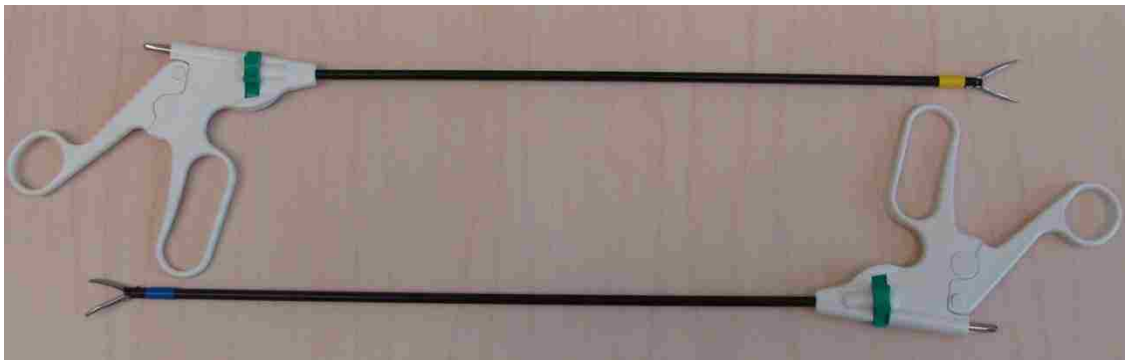


Figure 3.3 MIS mechanical hand-controlled instrument

The laparoscope is a stereo camera inserted in the human body to give the surgeon a field of view for the operation. The available device is Vista Medical Technologies' stereoscope. The lenses used in this device are standard endoscopic lenses with two CCDs positioned slightly apart sharing the same optical path as shown in Figure **3.5**. The cameras capture analog NTSC videos which are routed to a head-mounted display to provide stereoscopic viewing. The disparity between cameras is less than 5mm.

Figure 3.4 the MIS setup in the operation room [65]

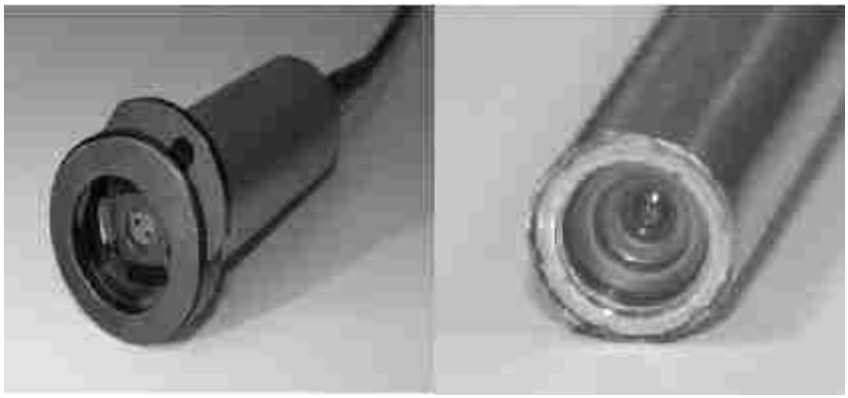


Figure 3.5 Vista stereoscope with single channel endoscopic lens

The initial intention of using the stereoscope in this research was to track the MIS tools, find matching points between left and right stereo images, and reconstruct the 3D positions of the tools using triangulation geometry. In order to reconstruct 3D positions of the tools, we need to calibrate the stereo camera in the endoscope. In addition, to

find matching points in the inner body environment, high specular markers are placed on the tools.

The attempt to use the available stereoscope has failed for several reasons. However, we can get better results using endoscopes that have better stereo-cameras than the Vista. This stereo-camera has larger lens disparity than the Vista stereoscope. We describe below, the approach that we have tried, the reason why it does not work, and where it can be useful.

Camera Calibration

Cameras in computer vision can be modeled as ideal pinhole cameras. This model is important to extract the properties of the world from images and to describe the mapping between the three dimensional world's coordinate and the two dimensional image plane. To be able to describe the mapping, the intrinsic and extrinsic parameters of the cameras must be approximated through the calibration process. The intrinsic parameters include the focal length and center of projection. The extrinsic parameters include the rotation and transformation matrix of the camera in relationship to the other camera. Using those parameters with the triangulation of epipolar geometry, we can construct the 3D coordinates. The intrinsic and extrinsic parameters and the relationship to image coordinates can be expressed in the following equations.

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = M_{\text{int}} M_{ext} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3.1}$$

$$M_{\text{int}} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$M_{ext} = [R \mid t] \tag{3.3}$$

Where (x,y) is the image coordinate for the (X,Y,Z) three dimensional coordinate, M_{int} is the intrinsic camera matrix which includes the focal length (f_x, f_y) and the center of projection (c_x, c_y), and M_{ext} is the extrinsic matrix that is composed of 3x3 rotation matrix R and translation vector t [56].

Cameras usually suffer from several types of distortion such as radial and tangential distortion. The amount of distortion increases the further the pixel is from the image center. Therefore, a model to remove the radial distortion is necessary. A second order polynomial describing the distortion is good for moderate level distortion cameras [57].

$$r_d = r(1+\kappa_1 r^2) \tag{3.4}$$

Using the un-distortion model proposed by Heikkila and Silven [58], two coefficients of radial distortion and two for tangential distortion are computed. The following set of equations describes the model where (u_n, v_n) represents the normalized undistorted image coordinate, and (u_d, v_d) represents the normalized distorted image coordinate. (k_1, k_2) are the second- and fourth-order coefficients of radial distortion. (p_1, p_2) are the de-centering coefficient tangential distortion.

$$\begin{bmatrix} u_d \\ v_d \end{bmatrix} = \begin{bmatrix} u_n \delta u_d^{(r)} + \delta u_d^{(t)} \\ v_n \delta v_d^{(r)} + \delta v_d^{(t)} \end{bmatrix} \tag{3.5}$$

$$\begin{bmatrix} \delta u_d^{(r)} \\ \delta v_d^{(r)} \end{bmatrix} = \begin{bmatrix} 1+\kappa_1 r_n^2 + \kappa_2 r_n^4 + \cdots \\ 1+\kappa_1 r_n^2 + \kappa_2 r_n^4 + \cdots \end{bmatrix} \tag{3.6}$$

$$\begin{bmatrix} \delta u_d^{(t)} \\ \delta v_d^{(t)} \end{bmatrix} = \begin{bmatrix} 2\rho_1 u_n v_n + \rho_2 (r_n^2 + 2u_n^2) \\ \rho_1 (r_n^2 + 2v_n^2) + 2\rho_2 u_n v_n \end{bmatrix} \tag{3.7}$$

Tools Tracking

The calibration model results can be used to reconstruct 3D points of the MIS tools. Therefore, the tools are tracked and their 2D positions are extracted. In order to reliably track the tools, shiny markers are placed on them. The markers are black/white one inch length rectangles. The corners can be found in the image by computing the second derivative based on Shi and Tomasi's definition [59] which is based on Harris' corner detector. The discovered corners can then be tracked across consecutive frames. Shi and Tomasi compute the following matrix to find the good features:

$$H(x,y)=\begin{bmatrix} \sum_{neighborhood}\left(\frac{dI}{dx}\right)^2 & \sum_{neighborhood}\left(\frac{d^2I}{dxdy}\right) \\ \sum_{neighborhood}\left(\frac{d^2I}{dxdy}\right) & \sum_{neighborhood}\left(\frac{dI}{dy}\right)^2 \end{bmatrix} \tag{3.8}$$

Where

I: the intensity of the pixel.

d_x, d_y: the horizontal and vertical displacements of the neighborhood window center.

The results of this approach showed that stereo reconstruction model is unreliable to be used in the context we need. This approach has failed because of the high accumulative error. Measuring Euclidean distance between two points is used as an experiment to study the accuracy of the model. In this experiment, we developed a virtual ruler to help the surgeons in clinical analysis like measurements and decision making. The virtual ruler is used to analyze the accuracy of the reconstruction and its validity to be used in the assessment context. We have tested the ruler on a Da Vinci and a Vista stereoscope.

Table 3.1 and Figure 3.6 report the measured distance and the true distance between two detected points. The table shows ten trials of distance measured between markers on the instruments pointing to phantom organs or the distance of the white bars of the markers on the instruments. The markers are tracked on various numbers of frames in each trial, then the distance is calculated in each frame and the mean and standard deviation are reported. Figure 3.7 shows the uncertainty curve for Vista's error. The percentage error is 8.55% and the sources of the errors can be identified from calibration estimation error, tracking error, image noise, and distortion error.

Table 3.1 The true distance and the measured distance between two points on phantom organs using the stereo reconstruction model using Vista stereoscope

Index	Type	# of frames	True Distance (mm)	Measured Distance(Mean) (mm)	Standard Deviation stddev	Actual Error
1	Liver	19	46	40.1	1.09	5.9
2	Liver	9	112	113.5	11.7	1.5
3	Lung	16	26	21.5	4.48	4.5
4	Lung	7	36	34.5	4.44	1.5
5	Marker	45	25.4	22.6	0.53	2.8
6	Marker	47	25.4	26.13	4.26	0.73
7	Kidney	14	57.15	56.257	13	0.893
8	Live	12	99	75.3	0.8	23.7
9	Lung	20	114	112.1	0.5	1.9
10	Lung	16	86	81.5	0.2	4.5

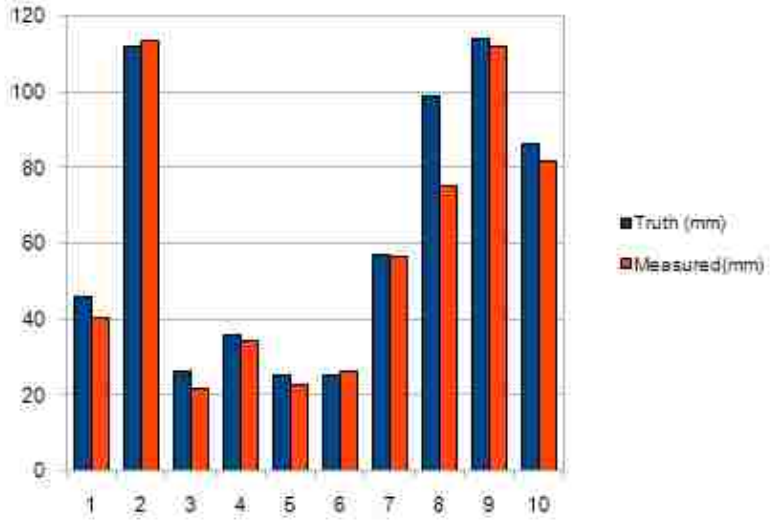


Figure 3.6 Measured vs. truth distance using Vista stereoscope

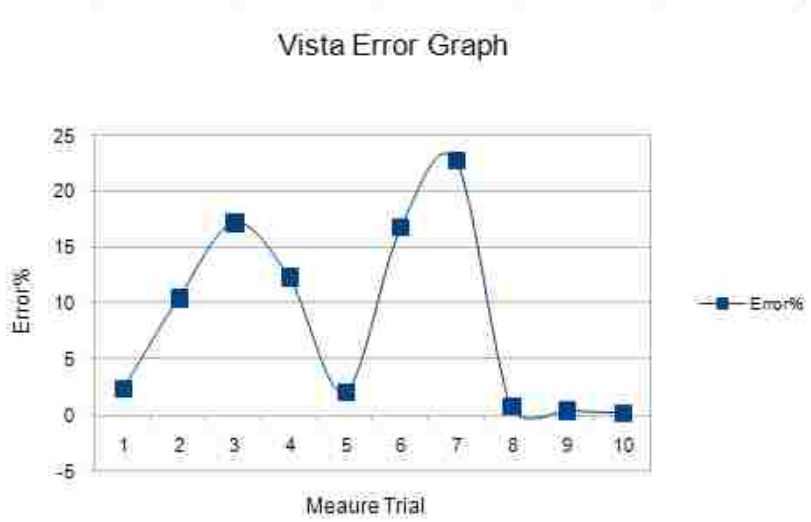


Figure 3.7 The error curve for the measurement using Vista stereoscope

The experiment showed that the error significantly correlates to the distance of the markers from the endoscope. The farther the markers are from the endoscope, the larger is the error. Table 3.2 and Figure **3.8** show that the measured distance decreases by increasing the distance between the object and the cameras. The distance from the

endoscope tip shown in Table 3.2 is measured in inches where the values represent the distance from the tip plus four inches. This area is the approximate working region of the tools in MIS.

Table 3.2 Sample tests of the measurement tools by increasing the distance between the object and the cameras using Vista stereoscope

Distance from the tip of the scope +4 inch	# of frames	True Distance (mm)	Measured Distance(Mean) (mm)	Stddev
1	23	25.4	29.6	0.8
2	45	25.4	22.6	0.53
3	21	25.4	19.51	1.25
4	22	25.4	17.13	0.73
5	15	25.4	16.91	0.49
6	23	25.4	16.53	0.33
7	108	25.4	15.7	0.39
8	54	25.4	15.35	0.83
9	24	25.4	16	1.25
10	71	25.4	16.2	1.17
11	32	25.4	17.07	2.7

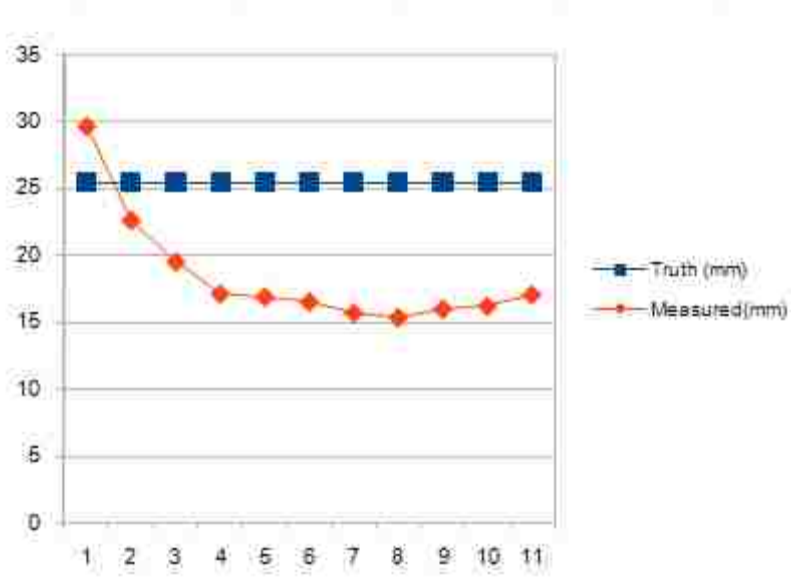


Figure 3.8 The measured vs. truth distance by varying the distance between the cameras and the object using Vista stereoscope

Using a stereo camera with separate optical paths demonstrates a wider baseline than the shared optical path. The maximum working volume increases dramatically at the same time. Separate optical paths can greatly improve the quality of stereo reconstruction. Unfortunately, that camera is unavailable in the laboratory. But we tested the model using a bi-channel stereoscope available at the UK's Albert B. Chandler Hospital. This stereoscope is part of the Da Vinci robot. As Table 3.3 shows, the results are more accurate and reliable than when using the single-channel stereoscope.

Table 3.3 The sample tests of the measurement tools on different phantom organs using Da Vinci cameras. The table shows the true distance and the measured distance

Index	Type	# of frames	True Distance (mm)	Measured Distance(Mean) (mm)	Standard Deviation stddev	Actual Error
1	Liver	98	70	77	3.1	7
2	Liver	39	82	80	2	2
3	Liver	11	112	118.5	3.1	6.5
4	Kidney	17	40	41.5	5.09	1.5
5	Lung	31	35	35	4.3	0
6	Lung	22	114	102	5.8	12
7	Lung	70	94	78	1.7	16
8	Lung	16	82.5	82.4	1.85	0.1
9	Lung	53	76	82.7	2.9	6.7
10	Lung	70	91.5	95.4	5.5	3.9
11	Lung	14	33	29.5	2.6	3.5
12	Lung	14	51	50.7	1.98	0.3
13	Liver	22	64	72.7	3	8.7
14	Liver	12	73	69	3.8	4
15	Liver	13	38	37.1	3.75	0.9
16	Liver	11	56	54.99	3.7	1.01
17	Liver	16	46	50.4	7.9	4.4
18	Liver	11	121	114	3.2	7
19	on liver	50	12	13.5	1.17	1.5
20	on liver	35	19	20.6	1.97	1.6
21	on liver	35	22.2	25.1	1.64	2.9
22	Kidney	24	57	59.5	4.3	2.5
23	Kidney	117	47.6	46.7	1.23	0.9
24	Kidney	379	53	57.2	1.44	4.2
25	Kidney	8	40	35	8.2	5

The uncertainty curve shown in Figure 3.9 demonstrates that the Da Vinci stereoscope's percent error is 6.9% compared to 8.55% using the Vista stereoscope. However, outliers increase the percent error. In addition calibration estimation errors, tracking errors, image noise, and distortion errors can contribute inaccuracy in distance measurements.

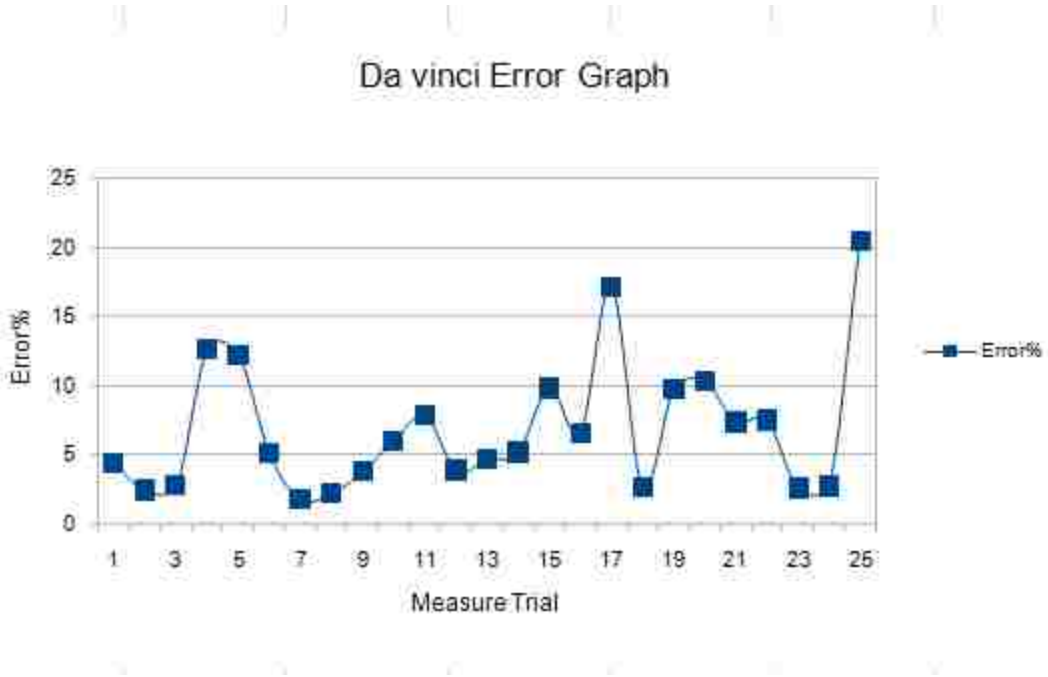
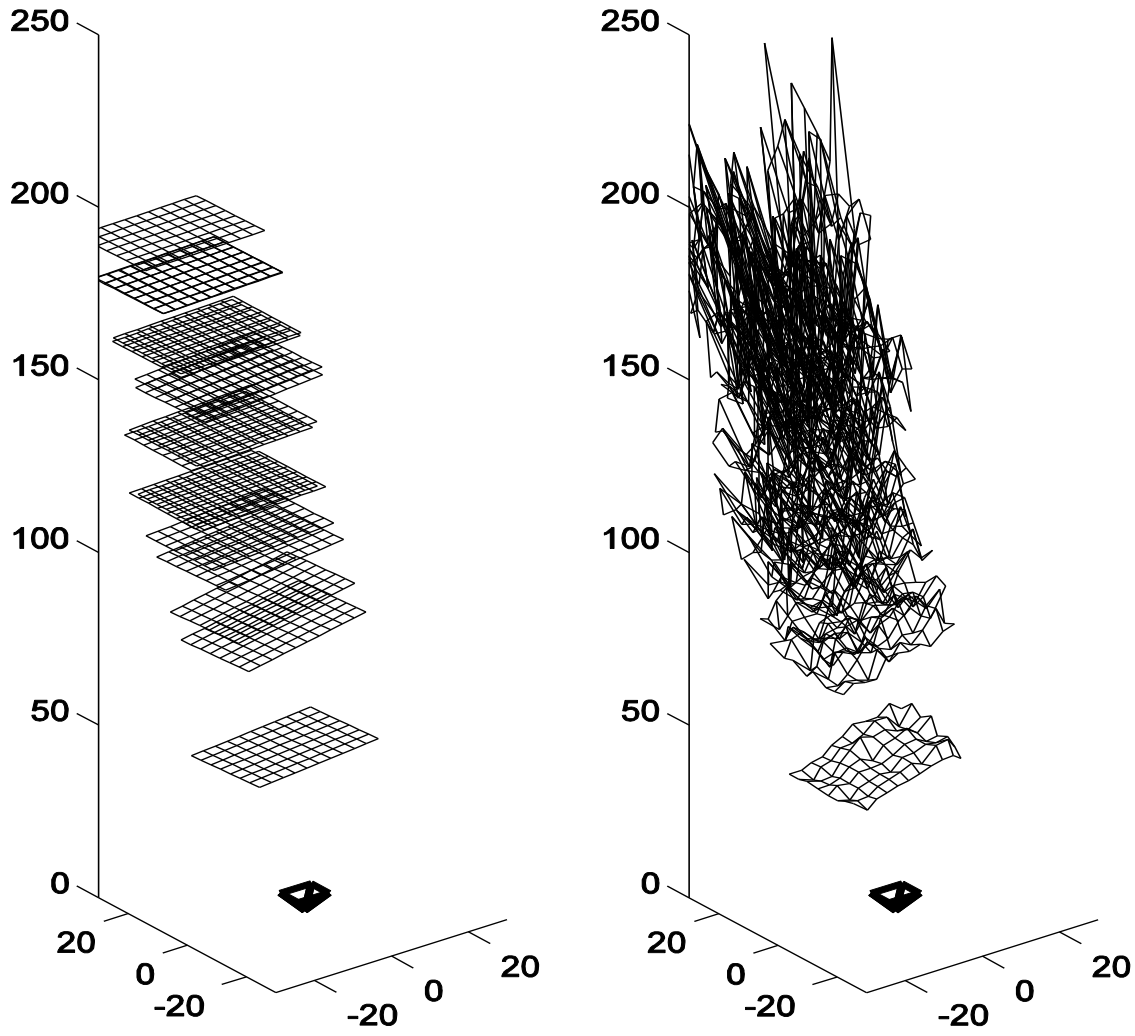


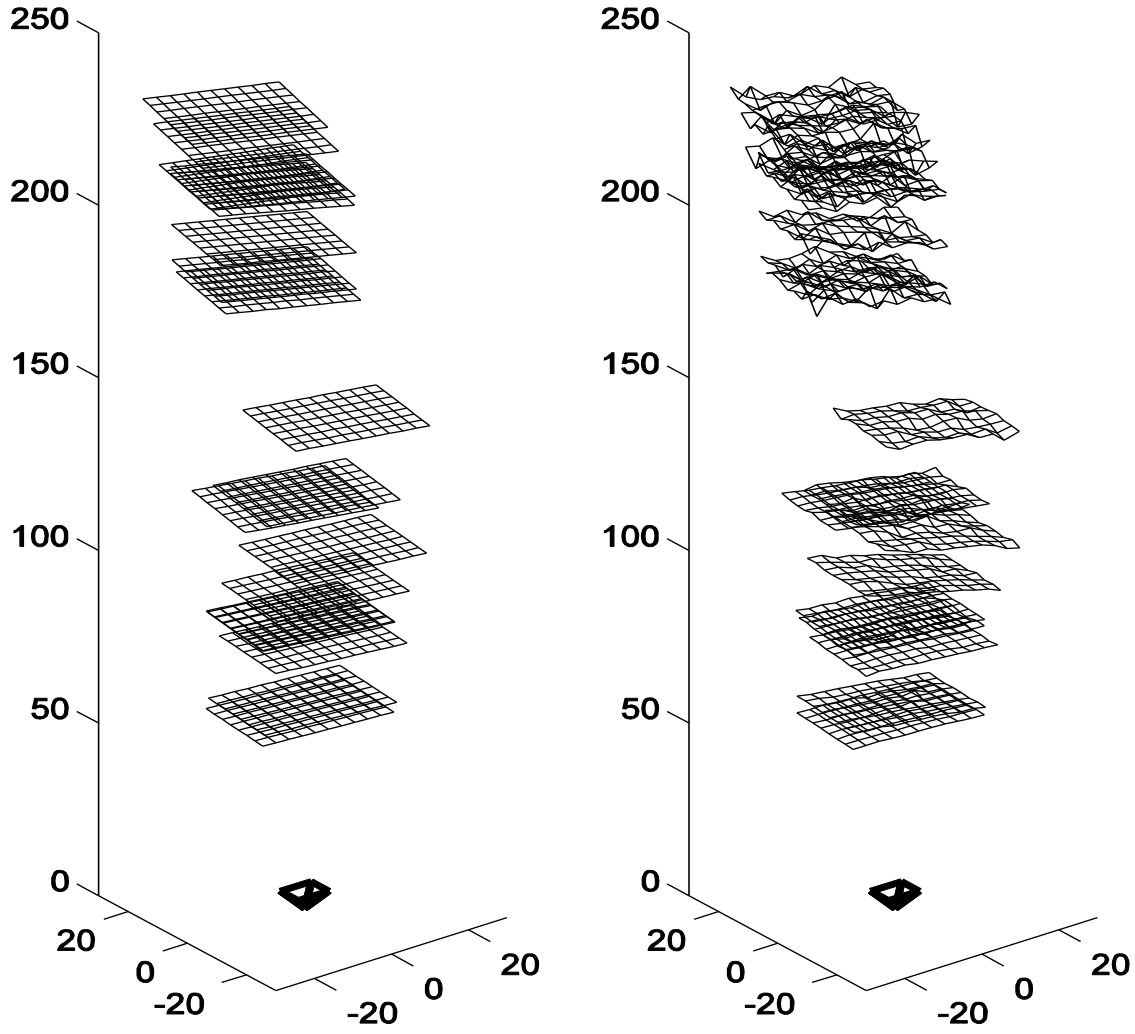
Figure 3.9 The error curve for the measurement using Da Vinci stereoscope

Another experiment we performed in the laboratory in cooperation with another colleague is the reconstruction of 3D images for the checkerboard used in the calibration process using the two types of stereoscope. Eighteen image pairs were captured for the checkerboard other than the images used in the calibration step. Each image contained the checkerboard pattern positioned roughly parallel to the image plane, orthogonal to the Z axis (depth), at increasing distances from the endoscope. Reconstruction by stereo triangulation was performed on each pair of matching feature points. Figure 3.10 and Figure 3.11 show the reconstructed patterns alongside the original points as calculated from the images. The images show that in both cases, reconstruction error increases with distance from the endoscope. In the case of the single-channel scope, however, the error is rather high from the start and rapidly deteriorates beyond about 60mm. The short baseline between its cameras accounts for this short distance, as the useful reconstructed volume increases in direct proportion to the disparity of the stereo pair.



(a) (b)

Figure 3.10 Single-channel reconstruction. The image on the left displays the approximate true positions of the planes, and the image on the right displays the reconstructed views. Reconstruction quality drops rapidly with distance. The camera appears at the origin.



(a) (b)

Figure 3.11 Bi-channel reconstruction. The image on the left displays the approximate true positions of the planes, and the image on the right displays the reconstructed views. The camera appears at the origin.

The reconstruction quality differs greatly based on the design of the stereo camera system. The viewing volumes available for accurate metric reconstruction are directly related to the baseline of the camera pair. Increasing the distance between the cameras also increases the angle of the rays projected from them to world points, reducing the effects of residual error on stereo triangulation. The setup of single-channel endoscope, shared optical path, and camera separation less than 5mm reduces the accuracy of the data, as image noise alone introduces significant errors into reconstruction. The context in this research is that we need to reconstruct 3D positions for points over a stream of images. This error accumulates over time and the result becomes unreliable.

Since the results show the accuracy in the single-channel scope is low, then it cannot be used in a process that accumulates data over a long time. In many of the calculated metrics for the assessment, we need to integrate distances over thousands of frames and these metrics are used to calculate other metrics. The error accumulates and increases as we track the points and calculate the distance over frames. Within a few thousands of frames, this error could accumulate to become more than the actual value.

The accuracy of the assessment and classification using machine learning algorithms relies on the accuracy of the low level features extracted from the system. As a conclusion of these experiments, we propose that the stereo 3D reconstruction is useful in clinical analysis of MIS, like real time measurements and decision making. However, it is computationally intensive and has other accuracy and reliability-related challenges if it is needed to accumulate data over time. Therefore, we decided to forgo it to see how necessary it is for skills assessment. To overcome this limitation, we used the laparoscope camera to only detect whether the tools were moving in or out of the field of view to distinguish the motion in the field of view from the motion outside it. The tracking of the tools' 3D positions was achieved using the Vicon system which is described in the next section.

3.2.2 Tools Detection

The Tools' detection task is achieved through: offline training and 2D tracking. In the training stage, a set of images with markers placed on the tools is collected. The training data is used to model the marker's color intensity. In the tracking stage, markers are automatically detected and traced.

Training Stage

Color-based tracking is vulnerable to lighting variations as the marker may be present differently over frames. However, the main or only source of light is the laparoscope. Further, the surgery and training environment is reddish and has a limited set of colors. So the lighting variation is minimal and mainly based on the orientation and relative angle between the tools and the laparoscope. Therefore, highly distinguished color markers are placed on the tools. To enhance the detection, a training procedure is developed to collect all the possible color values that the marker may appear in, within different frames, and a model is built based on this training data. Here, a 3D Gaussian model is used to imitate the marker's intensity change in Hue Saturation Value (HSV) color space. We assume each pixel has a 3D vector: $p = \{h, s, v\}$. The marker's color distribution can be formulated as:

$$f(p) = \frac{1}{(2\pi)^{\frac{d}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(p-\mu)^T \Sigma^{-1}(p-\mu)\right) \qquad (3.9)$$

Where $\mu = \{\bar{h}, \bar{s}, \bar{v}\}$ is the mean value of all the collected marker's pixels. $\sum$ is the corresponding 3 X 3 covariance matrix. p represents a pixel that is measured by this model, which is a 3D vector. d is the dimension of the data vector (here, it is equal to 3). So function f defines how likely that pixel x is from the marker. The training step

takes place once, but not on every usage. The pixels in the training stage are manually selected.

Tracking Stage

The marker is automatically detected by the color model described above. A threshold t is predefined which decides whether a pixel p is a marker pixel or not (only $f(p) \geq t$). Due to random noise and limited lighting variations, false positive pixels could be wrongly assigned as markers. To handle this problem, we take into account the neighborhood when each pixel $I(x, y)$ is processed. We use an indicator function 1_f to imply whether a pixel $I(x, y)$ belongs to the marker:

$$1_f(I(x,y)) = \begin{cases} 1, & if \Sigma_i \Sigma_j w_{ij} f(I(x+i, y+j)) \geq \lambda t \\ 0, & else \end{cases} \quad (3.10)$$

Where $i, j \in [-k, k]$. k specifies the size of the neighborhood, weight function w_{ij} decides how much contribution of the neighboring pixel $I(x+i, y+j)$ can be modeled as a Gaussian function. The constant value λ depends on the number of neighboring pixels which can be obtained from training.

After applying function 1_f in formula 3.10 to every pixel of the image I, a mask image is obtained with each pixel equal to either 0 or 1. Then, we use a Depth First Search algorithm (DFS) to retrieve all the connecting regions (whose indicator is 1) on the mask image. If no region is detected, then, the tool is not present in the image.

Two different color markers are placed on the surgical instruments. Two instances of the detection algorithm run to detect the left and right surgical instruments based on their color. The data gathered by the detection algorithm combined with the head tracking and eye-tracking enables us to calculate fusion and non-fusion metrics. More

details about data processing and metrics calculation from the surgical tools are in the Tracking Section 3.3.

3.2.3 Vicon System

The hands, head, and tools tracking data are obtained by using a remote, video-based system that uses contrast to identify the 3D position of high contrast markers. The Vicon system contains eight MX3+ cameras installed on the ceiling of the room. The architecture of the Vicon system, as Figure 3.12 shows, contains eight cameras, MX Ultranet unit, and Vicon software. The eight cameras cover and record a stream for the Capture Volume Area. The Ultranet provides power, synchronization, and communication for the eight cameras.

An MX3+ is a high quality camera fitted with a sensitive sensor. It consists of a distinct video camera, a strobe head unit, a suitable lens, and optical filter. Further, it provides high speed and low latency motion capture. Figure 3.13 shows the MX3+ camera and Table 3.4 presents the camera's specifications.

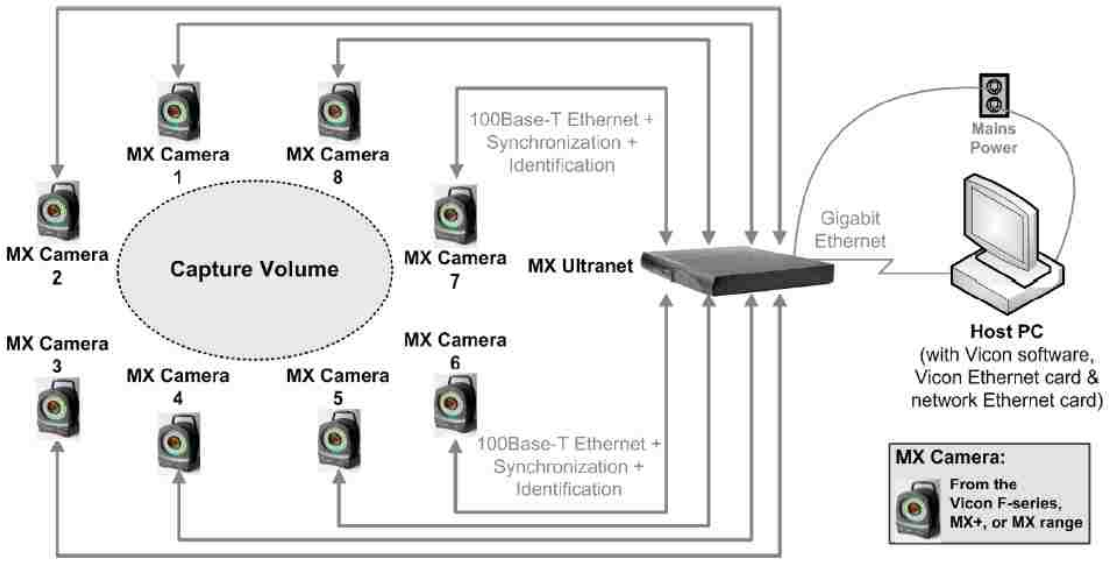


Figure 3.12 Basic Vicon MX architecture [60]

Figure 3.13 Vicon MX3+ Camera [60]

Table 3.4 Technical specification and performance indicators for the MX3+ cameras [60]

Component	**Specification**
Sensor Type	CMOS
Sensor size (Megapixels)	0.3
Sensor size (mm)	5.52 mm (H) x 4.89 mm(V) 8.15 mm(Diagonal
Sensor dynamic range	60 dB
Pixel size	9.9 microns x 9.9 microns
Photosensitive pixels	659 H x 494 V
Shuttered	Yes
Lens format	C-mount options
Size (mm)	215 (H) x 138 mm (W) x 182 kg
Weight	2.1 kg
Resolution (pixels)	659 H x 494 V
Maximum frame rate (fps) for full resolution	242
Aspect ratio	4:3
VGA monitor mode	60 kHz h x 100 Hz v
Threshold grid size	66x50
Threshold grid tile dimensions (pixels)	10x10

The Vicon is used to capture 3D positions of a template of markers. The template design we developed contains five parts. Two pairs of markers are placed on the upper tip of the surgical tools with separation of two inches. The markers' positions are extrapolated to find the 3D positions of the tips inside the body or the training box. On each of the two surgical gloves, we placed three markers in a triangle shape, used to

track the 3D positions of the subject's hands in order to calculate the kinematics and rotation metrics. We also placed three markers in a triangle shape on the surgical hat to calculate the kinematics and rotation of the subject's head. Figure 3.14 shows the template while a subject is performing a task.

Because the Vicon cameras are stationary, they do not need frequent calibration. The calibration procedure takes less than two minutes and the calibration data can be used as long as the cameras have not moved. Three steps are needed to calibrate the Vicon system. First, we apply auto-threshold to detect all infra-red reflectors from the scene, other than the designed template. Second, we capture a stream of images for the Vicon calibration 3-Marker wand to calibrate the cameras. Third, we set up the origin for the capture volume by placing the L-Frame in the required position in the room and capture a few images for that setup.

To validate the accuracy of the Vicon tracking and compare it with the 3D reconstruction using the Vista stereoscope discussed above, we calculated the distances between each pair of markers placed on the left and right tools. The tools are solid and the distance between each pair is fixed. The approximate distance between each pair is two inches. However, the exact distances measured manually by a ruler in mm are 47mm on the left tool and 46mm on the right tool with a potential of slight human error in the measurement. These numbers are used as ground truth. The distance is measured between the centers of both markers. Table **3.5** summarizes the tracking results over 26281 frames and shows that the standard deviation of the measurement over this number of frames is about 0.2mm for the left and the right tool. The percentage error is significantly lower than the error we measured using the Visa stereoscope experiment. The percent error in the left tool measurement is 0.25% where the error using the Vista is 8.55%. From this result, we conclude that using the Vicon to track the tools is more reliable than the Vista stereo reconstruction. The true distance between the tools is used to validate the positions of the points. In the data processing, if the distance between each pair is larger than a threshold of 5mm, we consider it as an outlier and

estimate the points of the markers using the previous and the next frame and adjust the positions accordingly.

The triangle markers are not put on a solid object to minimize the influence on the motion. Therefore, the distance between the points might slightly change from subject to subject based on the size of the hands and head. In addition, the distance between the hands triangle markers might slightly change during the experiment of one subject based on the status of the palm, whether it is closed or open. But this possible change is considered part of the motion.

Table 3.5 Vicon tracking validation and accuracy

	Left Tool	**Right Tool**
Number of Frames	26281	26281
True Value	47	46
Mean	47.11665	46.029235
Standard Deviation	0.267064	0.2230697
Percent Error	0.25	0.06

Figure 3.14 Template of markers used to track 3D positions and rotations of the subject's head, hands, and the surgery tools. The top two pictures show the real markers and the bottom one shows the markers' resolution from top view.

3.2.4 Eye Tracker

The eye-tracking data is obtained using a remote, video-based system that uses contrast to identify pupil location and size and the reflection of near-infrared and infrared non-collimated light to identify the cornea. The vector between pupil and cornea is used as an index of gaze direction. The device is FaceLAB system produced by

SeeingMachines which is not intrusive and is widely used in psychological, advertising, and human-computer interaction research. We used FaceLAB 4.3 integrated remote eye/head tracking system (www.seeingmachines.com). The eye-tracking data will be used to estimate fatigue and other assessment metrics. The measures are taken from the continuous stream of coordinates stored by the software to record fixation location and duration, as well as estimate the size of the pupils (and estimate size changes across time) in addition to garnering information about blinking. FaceLAB's architecture comprises of a software application and stereo camera which enables 3D tracking. Since this device non-intrusive and can detect several features of the eye, it can be a useful tool to improve the assessment of the MIS technical skills. The following are the features that FaceLAB can detect:

- Head-Pose
  The orientation of the subject's head in 3D coordinates. It has six parameters: three for describing the 3D position and three for describing 3D orientation. The Head-Pose is measured in world coordinate frame by transforming the head coordinate, X_w, using the following equation:

$$X_w = R.X_H + T \tag{3.11}$$

  Where x_H is in head coordinate, R is the rotation matrix, and T is the translation vector.

- Gaze
  The gaze is two rays, one for each eye. The ray is represented by an origin point which is the center of the eyeball and a unit vector which is the direction from the origin point towards the object being viewed.

- Saccades
  The saccade is defined according to FaceLAB as a fast motion of the eye to change the gaze point between fixation points. FaceLAB can accurately measure the saccades even inside short eye blinks.

- Viewed Object
  This feature is the interaction of the subject with the surrounding environment using the gaze direction and the head pose. This can be achieved by defining the surrounding environment and building the world model using simple shapes.

- Eye Closure
  The eye closure is the percentage of the coverage of the iris for each eye.

- PERCLOS
  This feature is an indicator for the fatigue which is based on the percentage of extended period of the eyes' closure time excluding the regular eye blinks.

- Eye Blinks
  The blink event is defined as a rapid eye closure followed by an eye opening.

- Pupillometry
  This factor is the measure of the diameter of each pupil.

- Facial Features
  FaceLAB can measure the face features and determine the facial gestures and changes by tracking points on the lips and eyebrows in 3D.

This system requires two types of calibration, camera calibration and head-monitor calibration. The cameras can be stationary by attaching them to the stationary surgical monitor. Thus, the calibration data can be stored without the need to recalibrate on every use. The calibration process is simple and can be undertaken by the laboratory technician. The process of calibration consists of placing the calibration kit shown in Figure 3.15 in different angles while the system is capturing pictures. Twenty snapshots are needed for the calibration.

The monitor calibration is needed to find the information matrix about the border of the monitor in the real world compared to the head and gaze direction of the subject. This calibration is a subject-specific so it should be done for every subject. A profile can be saved for the users of the system with their calibration data. However, re-calibration needs to be done more often than the camera calibration. The process of re-calibration

is simple and takes about 20 seconds. The process requires the subject to track a dot on the monitor that moves from left to right and top to bottom.

Figure 3.15 FaceLAB eye tracker cameras with calibration kit

3.2.5 Heartbeat Monitor

As an internal physiological variable, we decided to study the change in the heartbeat rate to find whether it correlates with the skill level. The heart beat rate was obtained by a heart rate monitor device. The heart monitor is a tool that measures the heart's electrical activities with each heartbeat over time. The device we used was a Polar RS800CX as shown in Figure 3.16. This device functions like a watch that can record the heart beat rate of the subject and transfer it to a computer using an infrared unit. The heartbeat sensors that read the activities are skin electrodes embedded in a rubber belt. The rubber belt includes an infrared unit to transmit the sensors read to the watch. The belt can be put on the subject's chest. This heart monitor is widely used in sports and by athletes.

Figure 3.16 RS800CX heart beat monitor

3.3 Metrics Extraction

This section presents the flow of the data being processed in order to extract the assessment metrics but it does not present the metrics list and their details. The list of extracted metrics and their details are presented in Section 3.6. The system extracts two types of metrics: fusion and non-fusion. Those metrics are extracted from the surgical tools, the surgeon's head, hands, eyes, and the surgeon's heartbeat. Many of the metrics are transformations from the 3D positions of the markers' template into economy of motion, kinematics, and rotational data.

3.3.1 Extracting Metrics from Surgical Tools

Extracting the tools' metrics requires synchronization and coordination of three systems: the laparoscope, the Vicon, and the eye tracker. The markers' template as explained in Section 3.2.3 includes two pairs of markers placed on the upper tip of the surgical tools with separation of two inches. The 3D positions of the markers in space and time are tracked using the Vicon system. The markers' positions are extrapolated to find the 3D positions of the tips inside the body or the training box given that the length of the tool is known. The tools' detection algorithm is used to find whether the left and right tools are present in the monitor at given time. The eye tracker is used to detect whether the subject's gaze intersects with the monitor at a given time. The non-fusion metrics are extracted from the Vicon data. The fusion metrics are extracted using the data of the three systems together. Figure 3.17 shows the block diagram and the flow of data to extract the metrics.

3.3.2 Extracting Metrics from the Surgeons' Head and Hands

Extracting the head and hands metrics requires synchronization and coordination of three systems: the laparoscope, the Vicon, and the eye tracker. The triangle markers placed on the head hat and hands' gloves are used to calculate the kinematics and rotational motion of the head and hands. The 3D positions of the triangle vertices are tracked in space and time by the Vicon. The tools detection algorithm is used to find whether the left and right tools are present in the monitor at a given time. The eye tracker is used to detect whether the subject's gaze intersects with the monitor at a given time. The non-fusion metrics are extracted from the Vicon data. The fusion metrics are extracted using the data of the three systems together. Figure **3.18** shows the block diagram and the flow of data to extract the metrics.

The training or operation stage is stationary in the field of view. Therefore, the eight cameras are set to cover the stage from all angles to reduce the chance of occlusion to

any marker in the template. If occlusion of the template's markers occurs, it is handled by extrapolating the points in one frame and over multiple frames by estimating the distance between points. Because the capture rate is set to 120 frames/second if an occlusion occurs to a point in a triangle or the lined markers on the tools, its position is estimated by the position of the marker in the previous and the next frames.

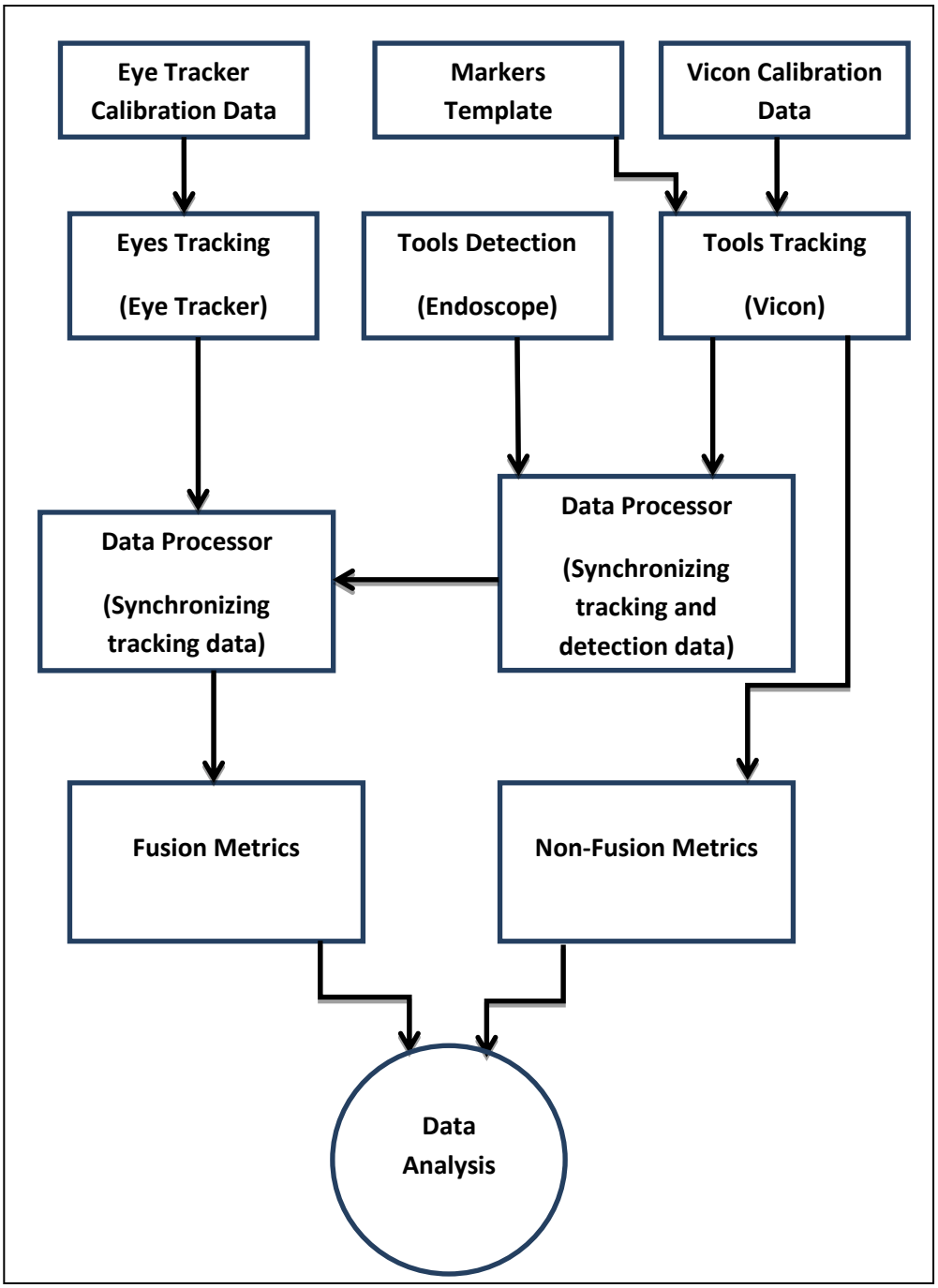


Figure 3.17 MIS tool-tracking and data transforming subsystem block diagram

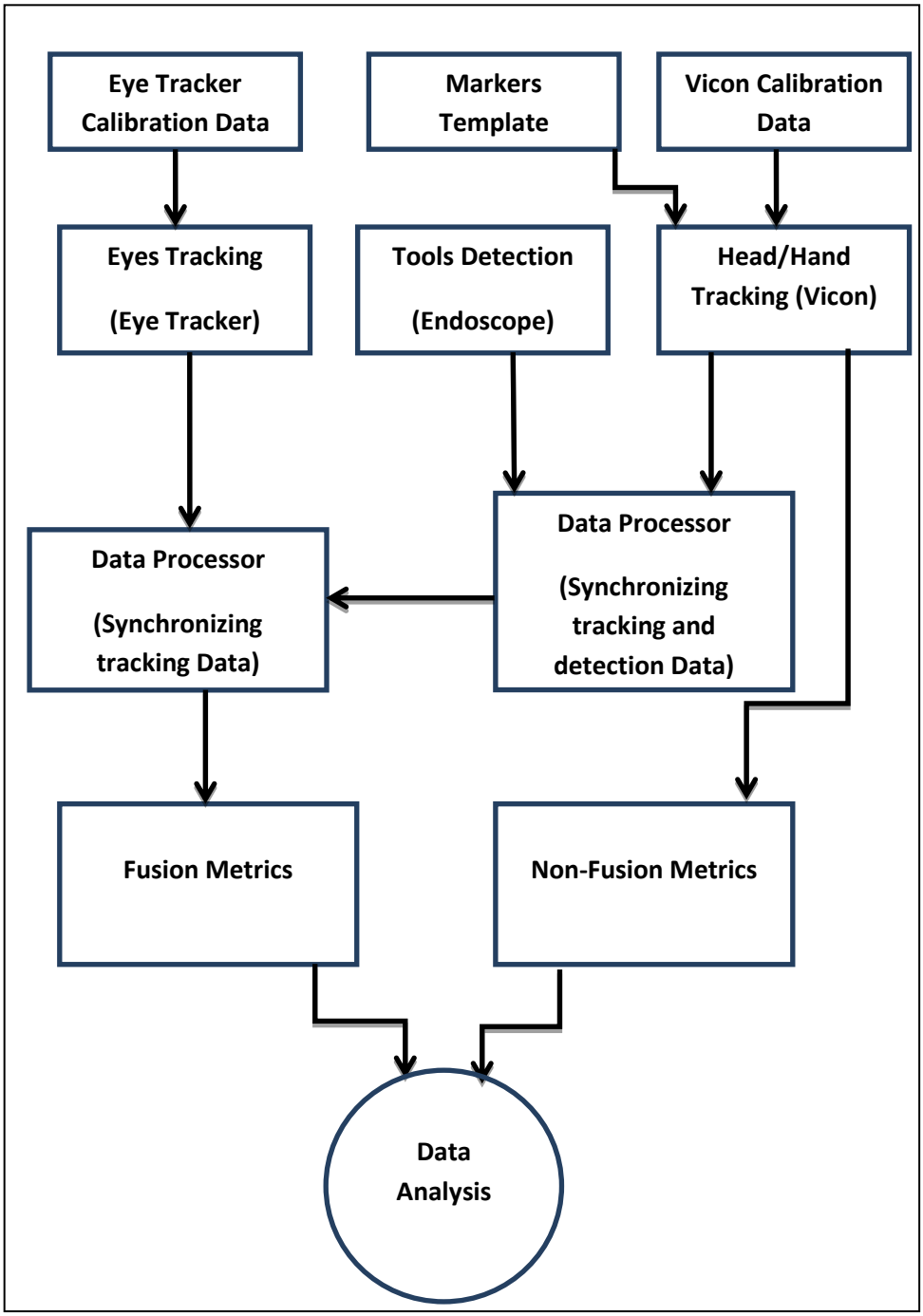


Figure 3.18 Block diagram for tracking hands and head to extract metrics

3.3.3 Extracting Metrics from Eyes

One of the important contributions of this research is the study of the correlation of the eyes' features with the MIS skills level. Metrics extracted from eye features might add some quality to the assessment process. The eye-tracker system is used to extract the metrics related to the features of the eyes and face. The list of metrics and their details are described in Section 3.6.4. Figure **3**.**19** shows the block diagram and the flow of the data to extract the eye metrics.

3.3.4 Extracting Heart Metric

The change in the heartbeat rate is the only metric extracted from the heart monitor. The metric is measured by calculating the first derivative of the heartbeat rate over time. The data of the heart rate monitor is not synchronized with other systems.

3.4 Sensors Synchronization

We used the Network Time Protocol (NTP) to synchronize the capture of the sensors together and reduce the time offset between the systems to an approximate of 16.6 milliseconds, which is the time to capture one frame using the FaceLAB. Then, the time in milliseconds is recorded for each frame by each system. Finally, the time of all the frames is coordinated using the time offsets calculated in the first step and mapped to the corresponding frames among the sensors. This mapping helps in classifying the motion and other measures into the blind and non-blind in order to find the metrics based on this classification. Blind motion is the analysis of the motion while the trainee is looking away from the field of view or the tools are undetected in the internal snapshots. Non-blind motion is the analysis of the motion while the trainee is looking at the display and the tools are detected in the internal snapshots. The overlap between the time of looking away and the time when the tools are undetected is measured in

order to not use redundant data. The NTP runs silently and continuously on the three systems to keep the offset at the level required.

To verify that all the systems start logging data simultaneously, we connected them together such that they commence logging information after one mouse click and stop logging after another mouse click. When a subject starts the session, the data collector clicks the mouse. When the session ends, the data collector clicks the mouse again. The clicking process can be done by the subjects themselves to start and end the session. But since the motion analysis represents a primary factor in the data analysis, and clicking the mouse requires motion from the subject, it should be done by all of the subjects or none of them to avoid bias in data collection. However, this step can be automated by replacing the mouse click by a button clickable by the subjects' foot.

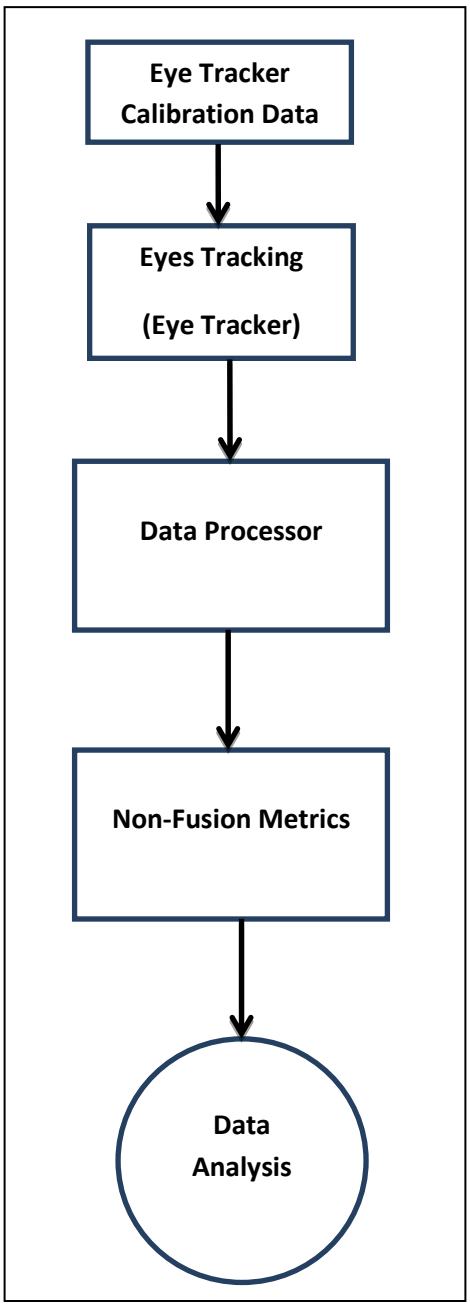


Figure 3.19 Block diagram for tracking eyes to extract metrics

3.5 Data Analysis

Does having read all the data and metrics, ideally imply that the problem has been solved? Quantifying human behavior is a challenging problem. MIS objective technical skills assessment using human measures needs this type of quantification. So the answer to the previous question is absolutely not. Analyzing these high dimensionality metrics and factors and quantifying them to reliable assessment measures is challenging. However, there are a number of statistical, machine learning, and data mining models that showed good reliability in classification, clustering, and finding hidden patterns in high dimensional data. Therefore, the use of a robust analysis model is important to achieve a reliable assessment.

We used four types of analysis on the data:

1. Metrics individual analysis to find the correlation between each metric and the skill level.
2. A multivariate data analysis to reduce the dimensionality of the metrics in order to find hidden patterns in the data. The multivariate method we used is Principal Component Analysis (PCA).
3. Clustering analysis to study the reliability of different sets of metrics to cluster the data into three clusters: novice, intermediate, and expert. The clustering algorithm used is a hybrid of partitioning and density-based algorithms.
4. Classification analysis to study the reliability of different sets of metrics to build a classification model that can find the class of the subjects among novice, intermediate, and expert classes. The classification algorithm used is Multi-Layer Perceptron.

In order to validate the models and analysis, we used real test data in addition to cross-validation analysis. However, different data analysis approaches can be used. Chapters Five and Six present more details about data analysis. Figure 3.20 shows the block diagram of the metrics flow to the data analysis to detect the subject's skill level.

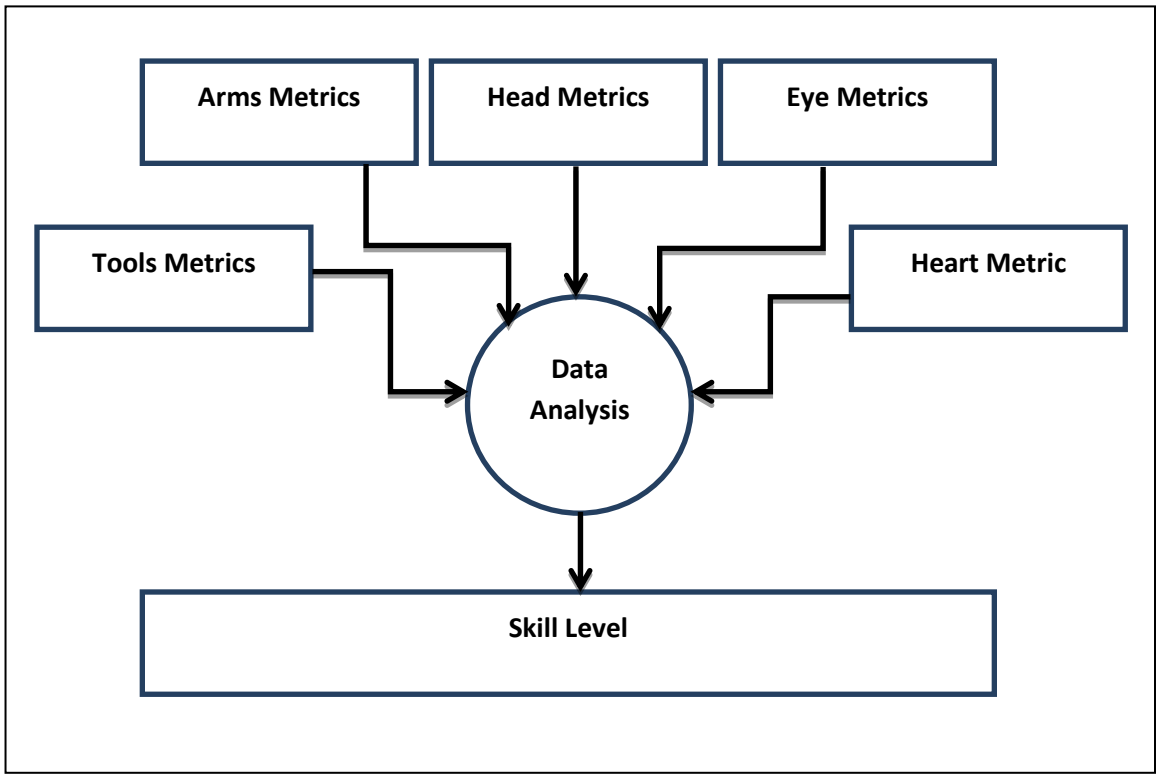


Figure 3.20 Block diagram for the data flow of metrics to detect the skill level

3.6 Extracted Metrics

We setup the system to extract 55 metrics and the option is open for more analysis to extract other metrics. The extracted metrics can be categorized into three types:

- Time metrics
- Economy of motion, Kinematics, rotational metrics
- Stress and fatigue metrics

Each of these categories contains fusion and non-fusion metrics. In the calculation and extraction of metrics, if time or frame rate are needed, we used the time and frame rate of the Vicon since it has higher resolution than other systems. The Vicon frame rate is 120 frames/second.

3.6.1 Time Metrics

Time metrics include time values and ratios throughout the training session. Table **3.6** presents a list of the time metrics and their description.

Table 3.6 List of time metrics

Metric Name	Description
Completion_time	Time taken to complete the task
display_looking_time	Time spent looking at the display
looking_away_time	Time spent looking away from the display or the instruments are absent from the field of view
display_looking_ratio	The ratio of time looking at the display to the total time
display_away_ratio	The ratio of time looking away from the display to the total time

3.6.2 Time Metrics Extraction

The system records the time of each frame captured by each subsystem in milliseconds. Each frame of each subsystem is marked as a blind or non-blind frame. The completion time is calculated by taking the difference between the first frame and last frame. *display_looking_time* is calculated by integrating the time of the non-blind frames subsets. *looking_away_time* is calculated by integrating the time of the blind frames subsets. The completion time should equal the sum of the *display_looking_time* and the *looking_away_time*. The ratios are the percentage of time of looking at or away compared to the total completion time.

3.6.3 Economy of Motion, Kinematics, and Rotational Metrics

This category includes metrics related to economy of motion, speed, acceleration, and jerk for the surgery tools and the trainees' head and arms. In addition, the rotation of the head and the hands is measured. Table 3.7 presents a list of the metrics in this category and their description.

Table 3.7 List of Economy of Motion, Kinematics, and Rotational metrics

Metric Name	Description
head_path_length	The path length of the head over the task
head_speed_mean	The average speed of the head motion over the task
head_speed_var	The variance in the speed of the head motion over the task
head_acceleration_mean	The average acceleration of the head motion over the task
head_acceleration_var	The variance in the acceleration of the head motion over the task
head_direction_change	The accumulation of the head direction change over the task
direction_change_frequency	The frequency of changing the head direction over the task
l_path_length	The path length of the left hand over the task
l_speed_mean	The average speed of the left hand motion over the task
l_speed_var	The variance in the speed of the left hand motion over the task
l_acceleration_mean	The average acceleration of the left hand motion over the task

Table 3.7 (Continued)

l_acceleration_var	The variance in the acceleration of the left hand motion over the task
l_direction_change	The accumulation of the left hand direction change over the task
r_path_length	The path length of the right over the task
r_speed_mean	The average speed of the right hand motion over the task
r_speed_var	The variance in the speed of the right hand motion over the task
r_acceleration_mean	The average acceleration of the right hand motion over the task
r_acceleration_var	The variance in the acceleration of the right hand motion over the task
r_direction_change	The accumulation of the right hand direction change over the task
l_path_length_looking_away	The path length of the left hand over the task while the subject is looking away from the display
l_speed_looking_away	The average of the speed of the left hand over the task while the subject is looking away from the display
l_acceleration_looking_away	The average of the acceleration of the left hand over the task while the subject is looking away from the display
r_path_length_looking_away	The path length of the right hand over the task while the subject is looking away from the display
r_speed_looking_away	The average of the speed of the right hand over the task while the subject is looking away from the display

Table 3.7 (Continued)

r_acceleration_looking_away	The average of the acceleration of the right hand over the task while the subject is looking away from the display
l_p_path_length	The path length of the left instrument over the task
l_p_speed_mean	The average speed of the left instrument motion over the task
l_p_speed_var	The variance in the speed of the left instrument motion over the task
l_p_acceleration_mean	The average acceleration of the left instrument motion over the task
l_p_acceleration_var	The variance in the acceleration of the left instrument motion over the task
r_p_path_length	The path length of the right instrument over the task
r_p_speed_mean	The average speed of the right instrument motion over the task
r_p_speed_var	The variance in speed of the right instrument motion over the task
r_p_acceleration_mean	The average acceleration of the right instrument motion over the task
r_p_acceleration_var	The variance in the acceleration of the right instrument motion over the task

Metrics Calculations

The rotation in the hands and head is measured by integrating the change in the triangle surface normal for the markers placed on the hands and head. The surface normal of a triangle as shown in Figure **3.21** can be calculated as follows assuming the triangle vertices are vectors of three:

$$E_1 = P_3 - P_2$$

$$E_2 = P_2 - P_1$$

$$n = E_1 \, X \, E_2 = [x_n \;\; y_n \;\; z_n]$$

$$\hat{n} = \frac{1}{|n|} n$$

$$|n| = \sqrt{x_n + y_n + z_n}$$

$$\alpha = \arccos \frac{n(T_2).n(T_1)}{|n(T_2)|.|n(T_1)|}$$

$$\alpha_{total} = \sum_{i=0}^{n} \alpha_i \qquad (3.12)$$

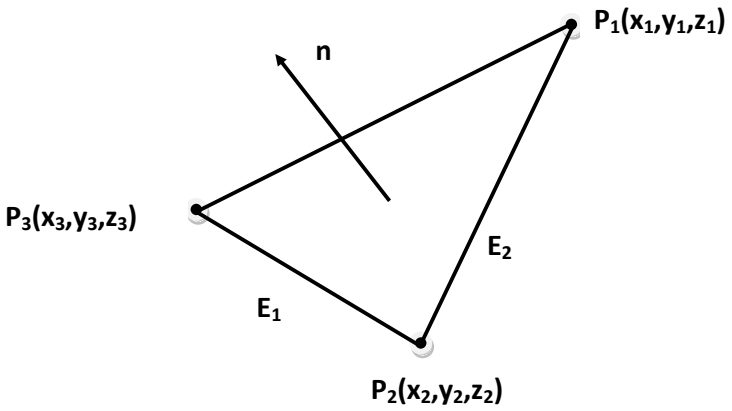


Figure 3.21 The triangle template to calculate the direction change in the hands and head

The path length for the instruments, hands, and head is calculated by integrating the displacement distance of one of the triangle vertices between every two frames. The 3D position of the marker is detected in each frame. Then, the Euclidean distance of the marker between each pair of consequence frames is calculated. This displacement is integrated to get the total path length over the period of the task.

$$r(p_1, p_2) = \sqrt{(x_2 - x_1) + (y_2 - y_1) + (z_2 - z_1)}$$

$$r_{total} = \sum_{i=0}^{n} | r(T_i) - r(T_{i-1}) | \qquad (3.13)$$

Where “p1”, “p2” are the 3D positions of a marker in to consecutive frames and “r” is the displacement.

The instantaneous speed, acceleration, and jerk (smoothness) are the first, second, and third derivative of the displacement over time.

$$v = \frac{dr}{dt} \tag{3.14}$$

$$a = \frac{dv}{dt} \tag{3.15}$$

$$j = \frac{da}{dt} \tag{3.16}$$

Where “v” is the instantaneous speed, “a” is the instantaneous acceleration and “j” is the jerk.

The mean in the speed and acceleration is measured by calculating the average instantaneous speed and acceleration. The variance is calculated over windows of frames by dividing the frames into subsets. We calculate the mean for each subset and then, calculate the variance over the subsets.

The blind path length, speed, and acceleration are calculated in the same way but while taking into consideration whether the head direction intersects with the monitor or not, and whether the corresponding tool to the metric (like left or right) is present in the field of view or not.

3.6.4 Stress and Fatigue Metrics

Stress and fatigue metrics are measures that might represent the level of stress and fatigue the subjects face in the operating session. Examples of this category are blinking frequency, change in blinking duration, change in blinking frequency, change in motion smoothness, fatigue, and heart rate change. Table 3.8 compiles a list of stress and fatigue metrics. The completion status is included in this category.

Table 3.8 List of stress and fatigue metrics

Metric Name	**Description**
Completed	Whether the subject successfully completed the task or not
l_smoothness	The smoothness of the left hand
r_smoothness	The smoothness of the right hand
Fatigue	
l_smoothness_fatigue_ratio	The change in the smoothness relation to the change of the fatigue (left hand)
r_smoothness_fatigue_ratio	The change in the smoothness relation to the change of the fatigue (right hand)
fatigue_perclosVairance	
Saccade	
blinking_frequency_mean	The mean of blinking frequency
blinking_frequency_change	The differential change in blinking frequency over time
blinking_duration_change	The change in blinking duration
blinking_duration_change_mean	The mean of the change in blinking duration
head_interaction_percentage	
gaze_interaction_percentage	
differential_heart_rate	The differential change in the heart rate over time

3.7 Data Normalization

The metrics values have different measurement units and different natures. In order to analyze them together and build a composite data model, these values have to be normalized. There are several ways to normalize data. The unity-based algorithm we followed subtracts the minimum value of a vector from each value and divides it by the

difference between the minimum and maximum as the equation 3.17 below shows. The range of the result values are in the unit interval [0,1]

$$X_{i,0to1} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{3.17}$$

Where:

X_i = Each data point i

X_{min} = The minimum among all points in one metric

X_{max} = The maximum among all points in one metric

$X_{i,\ 0\ to\ 1}$ = The data point i normalized between 0 and 1

3.8 Metrics Novelty

The novelty of this research is not only in the transformation of the assessment problem into a new domain and the new design of the assessment tool, but also in the metrics studied to reach a reliable set that can accurately assess the trainees and surgeons. We proposed and studied many new composite and non-composite metrics coordinated in time that are not used in previous studies. The reason is that the technology used does not provide the capability to extract them. Those metrics can only be extracted using computer vision technology to coordinate cues of eyes, external shots of body and instruments, and internal shots of the operative field. Table **3.9** contains the list of metrics that are novel and have never been studied before.

Table 3.9 List of proposed novel metrics

Metric Name
display_looking_time
looking_away_time
display_looking_ratio
display_away_ratio
head_path_length
head_speed_mean
head_speed_var
head_acceleration_mean
head_acceleration_var
head_direction_change
direction_change_frequency
l_direction_change
r_direction_change
l_path_length_looking_away
l_speed_looking_away
l_acceleration_looking_away
r_path_length_looking_away
r_speed_looking_away
r_acceleration_looking_away
l_smoothness
r_smoothness
Fatigue
l_smoothness_fatigue_ratio
r_smoothness_fatigue_ratio
fatigue_perclosVairance
Saccade
blinking_frequency_mean

Table 3.9 (Continued)

blinking_frequency_change
blinking_duration_change
blinking_duration_change_mean
head_interaction_percentage
gaze_interaction_percentage
differential_heart_rate

As this chapter has shown, the proposed system is capable of observing the environment, and captures metrics and their relationships with each other. Unlike previous studies, this capability is not limited to capturing the metrics from the motion of a single part, but it looks at the problem as a whole and every part is important in the assessment process.

Chapter 4

Case Study

Now as the system is built and set up to capture a large number of fusion and non-fusion metrics, we need to test and validate it to prove its reliability. To validate the system we need a task that mimics the surgical skills. This chapter introduces the design and description of the experiment and the protocol followed to recruit human subjects. The description includes the task for the experiment, data collection protocol, subjects recruiting protocol, training protocol, and other details. Chapters Five and Six introduce the implementation of the experiment, the data analysis, and results.

In collaboration with the Center for Advanced Training and Simulation at the University of Kentucky (www.mc.uky.edu), we adopted one of the training tasks that all MIS trainees have to pass in their first semester of training. The protocol of the experiment was designed to meet the requirements of the Institutional Review Board (IRB) at the University of Kentucky.

4.1 Study Design

Human subjects were recruited to participate in the study to explore the relationship between the metrics, the system measures, and the skill level of performing laparoscopic surgery training tasks. The subjects were divided into three groups with three levels of training to perform one simple task. The expert group was trained for five

hours. The intermediate group was trained for 2.5 hours. The novice group was only introduced to the tasks. Seventeen participants were recruited and data was acquired in multiple sessions for each subject.

4.2 Study Population

We recruited 17 research participants from among the graduate and professional students. Medical students were not specifically targeted for this study, but they were not excluded if they chose to participate. The participants were aged between 22 and 40 years. We targeted students who reported frequently performing tasks requiring a high level of eye-hand coordination and fine motor skills such as playing certain video games. Normal or corrected to normal visual acuity and normal color vision were required. We did not exclude any participants based on race or ethnicity.

4.3 Subject Recruitment Methods and Privacy

We recruited graduate and professional students by sending emails and a mass email to these types of students. The information was sent as a text.

4.4 Informed Consent Process

Subjects were contacted through email or by phone to provide a reminder and were offered the chance to ask questions of the research team.

When participants arrived at the lab, they had the opportunity to see the equipment they would be using to perform the experiment. Participants were given two copies of

the informed consent form, one for them to read and discuss with the researcher, and one for them to keep for their records. The subjects were given time to read the form thoroughly and then, the researcher paraphrased the form and asked the participants if they had questions. After all questions were answered to the satisfaction of the participants, they were asked to sign the consent form if they still wanted to participate. These consent forms are maintained in a locked filing cabinet in Room 304E1 of the UK Center for Visualization and Virtual Environments, 329 Rose Street • Lexington, KY. The forms are kept separate from data. Appendix A contains a copy of the collected form of informed consent.

4.5 Research Procedures

Since it is difficult to obtain approval to set up all the needed tools in a real operating theater and perform the experiments on real procedure, this research used the training box that is available in the laboratory. Participants learned to perform a simple task similar to those taught to medical students just starting surgical training. This task was performed on a surgical simulator located in the Center for Visualization and Virtual Environments at the University of Kentucky. A picture of the simulator is shown in Figure **4.1**. The simulator is an endoscopy training box consisting of a digital camera controller, a light source, a fiber optic cable, and a zero-degree 10-mm camera mounted above the simulator cover at a 90-degree angle to the right of the participant. It also includes a curved canvas screen representing the torso of a patient with several small incisions through which a laparoscope and surgical instruments (“graspers”) are inserted. Underneath the canvas surface, there is a platform on which to-be-manipulated objects are placed. Images of the movement of the instruments and objects are presented on the monitor display.

Figure 4.1 The training box setup

Participants were taught to perform the Pegboard Ring Transfer task which is a standard task in the MIS training curriculum. As shown in Figure 4.2, the task aims to pick up ten 1-cm rings from the rings carrier using the left probe, transferring them to the right probe, and placing them on ten pegs using a grasper. The task is then repeated by grabbing the rings using the right grasper, transferring them to the left grasper, and placing them on the ten pegs again. This task covers the grasping, pushing, and transferring skills. Possible errors in this task are inappropriate hand use and dropping rings off of the pegs. A successful candidate should complete the task in 240 seconds and a medium-skilled one in 480 seconds. Candidates taking more time with errors should be considered novice.

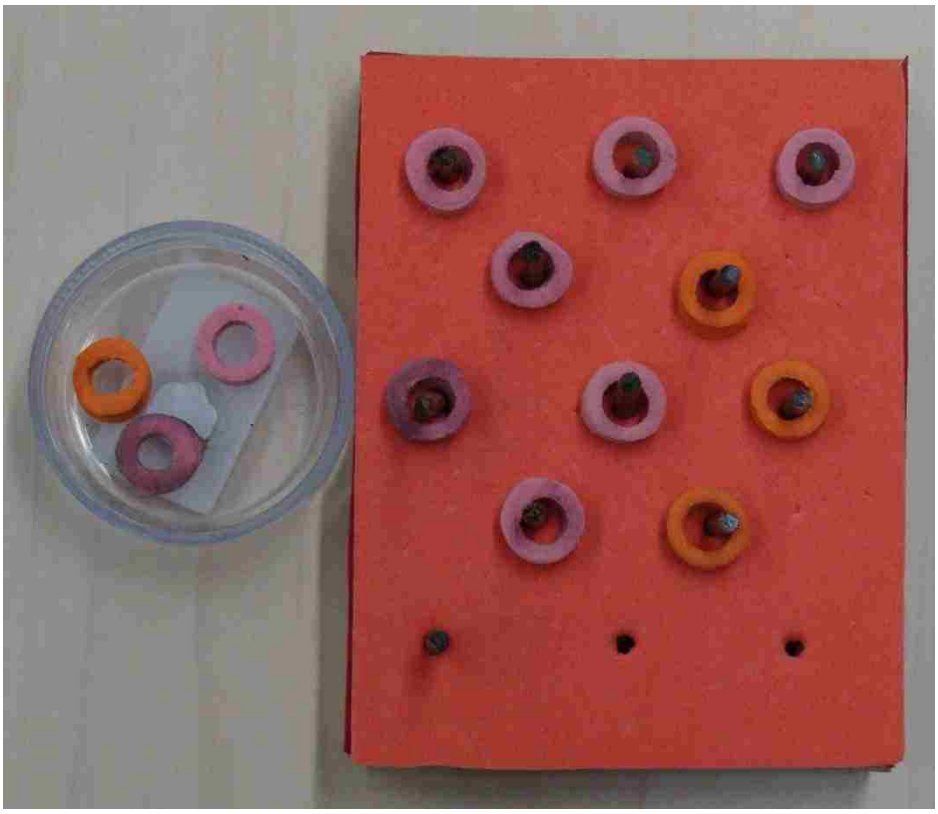


Figure 4.2 Pegboard ring transfer task

4.6 Resources

All phases of the research were conducted at UK's Center for Visualization and Virtual Environments (CVVE). Students were reminded of the location of the experiment two days prior to the scheduled time. The CVVE provided technical support, housed all equipment, and provided areas in which participants could take breaks and relax.

4.7 Potential Risks

The risks participants were likely to experience or did experience were no greater than those experienced when playing video games. These risks include some dizziness and fatigue to the arms, hands, and shoulders. We encouraged rest breaks in the middle of the session to prevent these potential discomforts.

4.8 Safety Precautions

In order to minimize risks that may be associated with a breach of confidentiality, we identified data only with code numbers rather than any personal identifiers. The video recordings were kept on file for a maximum of five years to be destroyed later. It should be noted that the video recordings were not of the participants themselves; rather they were of their performance as seen through the laparoscope and the motion of the markers on the arms and head. The video recording did not reveal parts of the subject's body. All raw and recoded data were kept in locked filing cabinets in the laboratory suites at the Center for Visualization and Virtual Environments.

4.9 Benefit versus Risks

There were no direct benefits to the participants, other than an increased understanding of procedures used in research on human performance.

4.10 Research Material, Records, and Privacy

All research materials were collected as part of this project. No existing data was used. The sources of data have been described in detail in the section on Research

Procedures above, which include 1) eye-tracking data, stored numerically as eye positions relative to the video monitor/display, 2) videotapes of the movement of the instruments controlled by the participants (no portion of the participant's body is videotaped), 3) arms- and head-tracking data, stored numerically as markers positions (no portion of the participant's body is videotaped), and 4) heart beat rate. Methods of maintaining the data are described in detail the section on Safety Precautions above.

4.11 Confidentiality

Code numbers were assigned to all participants, and the association between code number and name were destroyed after the participants completed their role in the research. Thus, all data records include the code number and there is no remaining link to the participants' names. We used the videotapes of the positions of the surgical instruments manipulated by the participants, positions of markers represent movement of the arms and head of the subject, positions of the eyes, and heart beat rate. These tapes do not show any portion of the participant's body. The tapes will be destroyed after at most five years of the study start date.

4.12 Payment

Graduate and professional participants in the study were volunteers. Those participants had the option to be compensated $20 in the form of cash for their participation.

4.13 Subject Complaints

Contact information for both, the researcher and the Office of Research Integrity (ORI) were provided on the informed consent form. We also made contact information pertinent to both, the ORI and the Principal Investigator (PI) in charge of data collection, available to participants through the Center for Visualization and Virtual Environment's receptionist and operator.

4.14 Discussion

Initially the plan was to design a case study that included three MIS training tasks. In this scenario, the data would have been distributed over three different tasks. Thus the effort and analysis would be redundant over three small sets of data. To increase the significance of the size of the dataset and the analysis result, we decided to focus on one training task and perform several experiments to test the validation and robustness of the platform. We selected the pegboard ring transfer task and dropped the cannulation and robe pass tasks from the study. Several reasons led to this choice. First, the complexity of the pegboard ring transfer is higher than the complexity of the other two. Thus, the subject needs more time and effort to master, which gives a window for analyzing three skill levels and can reveal more discriminant features to the metrics measured. Second, the pegboard ring transfer task covers more MIS skills than the other two. The pegboard ring transfer task covers grasping, pushing, and transferring skills while cannulation covers pushing and pulling, and the robe pass covers transferring skills. Finally, the pegboard ring transfer task is more popular in studies and training centers than the other two. The tasks used in the studies discussed in Chapter Two vary based on the nature of the experiment and the environment where the study is taking place. The virtual environment studies use computer-based tasks. The studies in the operation theater use real surgeries on humans but they are performed by surgeons.

Using a real surgery task as a case study is more significant but performing our study in a real operation room is a challenge at this point. Other laboratory studies used the pegboard ring transfer or different tasks.

As briefly mentioned, studies in the literature have taken place in three different environments: laboratories, virtual reality, and operation theaters. The primary investigators of the studies that were performed in the operation theaters were surgeons. The studies which took place in the laboratories or virtual environments were undertaken by surgeons or researchers from other fields such as, engineers. Many of the studies that were performed in the operation theaters started initially in the laboratories. Our study commenced in the VIS Center laboratory and our endeavor is to work on conducting a study that will validate the system in a real operating theater.

The number of subjects used in the studies reviewed in Chapter Two varies between two and 56 subjects. In our study, we have used 17 subjects in 70 sessions. The larger the dataset is the more significant and reliable the result is. Many of the previous researchers studied and performed the analysis on individual metrics. Few used approaches that are based on multivariate analysis like Support Vector Machines (SVM) and Hidden Markov Model. We composed metrics and studied data models that include multiple metrics therefore we needed large set of data.

Two ideas that were considered have affected the system design significantly. The system had to be designed to enable it to acquire fusion metrics through fusion motion analysis and beget a wide range of metrics that might correlate to the skill level. In our experiment, we oversampled the metrics in data collection to study their significance and find a reliable combination of metrics for the assessment. Some of the metrics were used in previous studies and many of them were new metrics. The fusion metrics were all new since this idea was the first time being investigated. Also, the head- and eye-tracking metrics including the fatigue and stress metrics were all new and had never been studied before. All the researchers studied a limited number of metrics as a result of their system limitations. Few studies used 78 metrics in the analysis. But these studies

were performed using the Da Vinci robot. The robot provided those metrics from its system. Therefore, they could not be generalized to environments other than the robotic and could only evaluate Da Vinci tasks.

The system we built was designed to be non-intrusive. The markers of the template were placed on the gloves and hat which minimized the influence on the subject's motion. The only intrusive part in the experiment was the heart rate monitor belt. Subjects had to wear this rubber belt across their chest. The belt was lightly intrusive but did not impact the subject's motion.

As part of the IRB requirements, everything managed by the subjects had to be sterilized. To facilitate the process, we kept aside one pair of gloves and one hat with markers on them and second set of new gloves and hat without markers. We asked the subjects to wear a new hat and gloves before wearing the hat and gloves that include markers. If this approach was used in a real environment, the markers could become a standard on the gloves and hats. The heart rate monitor belt was sterilized at the beginning and end of each session. The belt sterilizing method was resorted to after consulting a nurse.

There were protocols governing the management and handling of a subject's visit and a sample of the steps followed by one subject is provided here as an instance:

- Initially the subject shows interest in participating.
- Some information is communicated to the subject on what he/she is expected to do and how long it will take.
- If the subject decides to participate, an appointment (with a date and time) is scheduled for a session.
- The subject is reminded of the appointment two days before the date.
- If the session is the subject's first, then, an overview of the lab and the experiment is provided.

- The informed consent form in Appendix A is handed to the subject and explained in detail with regard to the experiment, risk, privacy, expectations, and compensation.
- The subject takes time to read the informed consent form before signing it.
- We explain in detail the task the subject is expected to perform. A demo run is performed for him/her to get an idea about the task and then, he or she is given the chance to perform it once to experience the complexity of the task.
- The heart rate monitor belt is sterilized and given to the subject to wear in a private area.
- The gloves and hat are given to the subject to wear before the gloves and hat that include the marker.
- The subject's gaze is then calibrated relative to the display monitor in a process that takes less than 30 seconds.
- The subject is requested to prompt the researcher when ready to start logging data.
- The subject starts performing the task.
- During the process the investigator collects information about the subject's performance to assign a skill level as reference. As described in Chapter Three, the subjects are evaluated based on training time, completion time, completion status, and errors such as, dropping the rings at the transfer stage.
- When the task is completed, the investigator stops the data logging.
- The belt, hats, and gloves are collected from the subject.
- The subject gets a receipt to collect the substitution, schedule another appointment for a training or data collection session, and then released.
- The subject's raw data is assigned a code as described in the confidentiality Section 4.11.
- The informed consent form is archived in a locked cabinet.
- The data is archived for processing.

- Any connection between the data and the subject is removed.

In the training sessions, the subjects are only required to show up in the lab and perform the task for a specific number of minutes as described in the experiment design. The information of how long each subject had to undertake training is retained, but any other connection between the data and the subject is deleted. If the session was not the first session, all the steps above are followed but with less details especially related to the task and the informed consent form, because it is assumed that the subject is already familiar with them.

This chapter describes the procedure and protocols used with the human subjects involved in the experiment. These protocols are approved by the IRB at the University of Kentucky. More forms are needed in order for the experiment to be approved by the IRB. The forms are not reported in this thesis. This chapter at the end presents a discussion on the experiment and throws up a comparison with other studies. In the next chapter, we discuss the results of the case study and the values of the measured metrics as individuals.

Chapter 5

Dataset Collection and Results

This chapter presents a high level of analysis of the metrics including the individual correlation significance between each metric and the skill level. Bar graphs that represent the change of the metric value compared to the skill level are also presented for some of the metrics that have a high Pearson's correlation coefficient ($r>0.5$). At the end of the chapter we show a brief comparison between the captured metrics and previously studied metrics including the novelty of the large number of metrics we were able to measure

5.1 Metrics Analysis

As mentioned in Chapter Three, there are many new metrics that can be measured by the multi-sensor system which have never been studied before. The transformation of the problem and the utilization of computer vision technology enable the measurements of those new metrics. The aim is not to acquire as many metrics as possible. The system is designed and built in order to provide the capability to capture metrics with potential correlation to the skill level. We need to find the correct set of metrics that can classify the performance and skills in high accuracy and robustness. Therefore, we oversampled the collected metrics with the intention of finding out which were correlated with the skill level and which good combination could produce a reliable data model. However, to validate the system and the new metrics, we need to

evaluate them in a significant case study. The data was captured for the subject while performing the peg board transfer task as described in Chapter Four. A subset of the subjects was trained to reach the intermediate and the expert levels. Fifty-five metrics were extracted for each subject in each session in order to analyze the correlation between the metrics and the experience level.

The analysis and discussion in this chapter include 58 records out of 70 and the former includes 19 novices, 14 intermediates, and 25 experts. There are also three sets of data for two more subjects that were captured from a distribution of four novices, four intermediates, and four experts out of a total of 12. The second set of data is captured after the model for data analysis was built and the data was used for validation. The reason for capturing more data at the expert level was to study the effect on the metrics with more training at that level. Figure 5.1 shows the Pearson correlation coefficient (r) for the 55 extracted metrics sorted on the absolute correlation coefficients.

Table 5.1 List of 55 metrics with their correlation with skill level and the P-value sorted on their absolute correlation coefficient

Variables	Pearson Correlation Coefficient (r value)	Absolute Correlation Coefficient |r|	Pearson Correlation Coefficient (P Value) <
Completion_time	-0.95	0.95	1.31793E-30
r_direction_change	-0.91	0.91	1.76593E-23
r_path_length	-0.86	0.86	2.89883E-18
display_looking_time	-0.85	0.85	1.82231E-17
r_p_path_length	-0.85	0.85	1.82231E-17
l_path_length	-0.84	0.84	1.00622E-16
head_direction_change	-0.84	0.84	1.00622E-16
l_direction_change	-0.82	0.82	2.20758E-15
Completed	0.74	0.74	2.49378E-11
l_p_path_length	-0.7	0.7	8.01388E-10
head_path_length	-0.68	0.68	3.69317E-09
l_path_length_looking_away	-0.67	0.67	7.584E-09
r_path_length_looking_away	-0.67	0.67	7.584E-09
looking_away_time	-0.66	0.66	1.51545E-08
direction_change_frequency	-0.61	0.61	3.36857E-07
display_looking_ratio	0.55	0.55	7.28292E-06
display_away_ratio	-0.55	0.55	7.28292E-06
gaze_interaction_percentage	0.54	0.54	1.14838E-05
head_speed_mean	-0.43	0.43	0.000738471
Saccade	-0.37	0.37	0.00420177
blinking_duration_change	-0.35	0.35	0.007005563
blinking_duration_change_mean	-0.34	0.34	0.008939785
head_acceleration_mean	-0.29	0.29	0.02709483
head_speed_var	-0.28	0.28	0.033125102
l_acceleration_mean	-0.26	0.26	0.048552283
l_speed_looking_away	0.26	0.26	0.048552283
l_smoothness	-0.25	0.25	0.058226043
l_acceleration_var	-0.24	0.24	0.069399968
r_acceleration_var	-0.23	0.23	0.082222951
Fatigue	-0.21	0.21	0.113411232

Table 5.1 (Continued)

head_acceleration_var	-0.2	0.2	0.13206597
r_speed_var	-0.18	0.18	0.176165015
l_speed_mean	-0.17	0.17	0.201841432
head_interaction_percentage	0.14	0.14	0.294403823
r_speed_mean	-0.13	0.13	0.330590422
r_p_acceleration_mean	0.13	0.13	0.330590422
r_p_speed_mean	0.13	0.13	0.330590422
r_smoothness	-0.12	0.12	0.369454249
r_speed_looking_away	0.12	0.12	0.369454249
l_p_speed_mean	0.12	0.12	0.369454249
r_acceleration_mean	-0.11	0.11	0.410957667
l_p_acceleration_mean	0.1	0.1	0.455031512
r_p_speed_var	-0.09	0.09	0.501574418
r_acceleration_looking_away	0.07	0.07	0.601501345
r_p_acceleration_var	0.07	0.07	0.601501345
l_speed_var	0.06	0.06	0.654524616
blinking_frequency_mean	-0.06	0.06	0.654524616
fatigue_perclosVairance	-0.05	0.05	0.709298766
blinking_frequency_change	-0.04	0.04	0.765574328
r_smoothness_fatigue_ratio	0.03	0.03	0.823079436
differential_heart_rate	0.03	0.03	0.823079436
l_acceleration_looking_away	-0.02	0.02	0.881523615
l_smoothness_fatigue_ratio	0.01	0.01	0.940602014
l_p_acceleration_var	-0.01	0.01	0.940602014
l_p_speed_var	0	0	1

Some of the measured metrics showed insignificant correlation with the skill level. One of the study's goals was to find a good set of metrics that could classify the performance and skill levels with high accuracy and reliability, thus, the metrics were oversampled from the beginning. Further, the task we used in this case study may not have been hard enough to show the correlation between some metrics. For example, the fatigue and stress metrics may have needed a harder and a more stressful task in

order to study their correlation and examine whether they may actually contribute to improving the assessment. What we are trying to highlight is that the metrics that showed insignificant correlation in this study may need more analysis to find out their significance in the evaluation process and their correlation with the complexity of the performed task. Therefore, more analysis is needed to find out which of the metrics were functions of experience, which were functions of task complexity, and which were functions of both. In Chapter Eight which is about future work we discuss this idea and propose a case study with a more complex task.

In this chapter and the next, we shall focus on the metrics that show a high correlation with the skill level. The case study showed 18 metrics had a significant correlation ($|r|>0.5$) to the skill level. Many of these metrics were new to the skills assessment analysis. In addition, the result showed that the metrics related to speed and acceleration, which were widely used in previous studies, had a low correlation and were thus, not the best metrics to use in the assessment. Table 5.2 compiles the list of metrics with correlation $r>0.5$. The shaded rows in the Table are new metrics proposed by this study.

Table 5.2 List of metrics with r >0.5. The shaded rows are new assessment metrics

Variables	Absolute Correlation |r|	P-Value
Completion_time	0.95	1.31793E-30
r_direction_change	0.91	1.76593E-23
r_path_length	0.86	2.89883E-18
r_p_path_length	0.85	
display_looking_time	0.85	1.82231E-17
l_path_length	0.84	1.00622E-16
head_direction_change	0.84	1.00622E-16
l_direction_change	0.82	2.20758E-15
Completed	0.74	2.49378E-11
l_p_path_length	0.7	8.01388E-10
head_path_length	0.68	3.69317E-09
l_path_length_looking_away	0.67	7.584E-09
r_path_length_looking_away	0.67	7.584E-09
looking_away_time	0.66	1.51545E-08
direction_change_frequency	0.61	3.36857E-07
display_looking_ratio	0.55	7.28292E-06
display_away_ratio	0.55	7.28292E-06
gaze_interaction_percentage	0.54	1.14838E-05

5.2 Variance Analysis

Table 5.3 and Figure 5.1 show the variance of the metrics among each category of subjects. In all metrics reported in Table 5.3, the variance at the novice level is higher than the variance at the intermediate and expert levels. The variance in the intermediate level is higher than the variance at the expert level. The only exception is in *l_p_path_length* metric which is highlighted in the Table. The variance of the intermediate level for that metric is less than that of the expert. These values of variance among the subjects' levels show the similarity of the performances in the expert category compared to the novice category. The similarity among the subjects increases by moving from novice to expert. Further, the difference in the variance

between the novice and the intermediate compared to the difference between the intermediate and expert shows that the first is larger than the second except for the *completion_time*. This difference shows the intermediate level performance is closer to the expert than it is to the novice.

5.3 Pearson's Correlation Analysis

Figures 5.3–5.19 show the bar graphs of each metric for the 58 subjects. In those graphs N represents the novice level, M represents the intermediate level, and E represents the experienced level. As we see in this set of Figures, the norm of the values is close among one level but differs between levels. This similarity among subjects at the same level and the difference among levels are less because the correlation coefficient is smaller. Metrics that have |r|<0.5 are not reported.

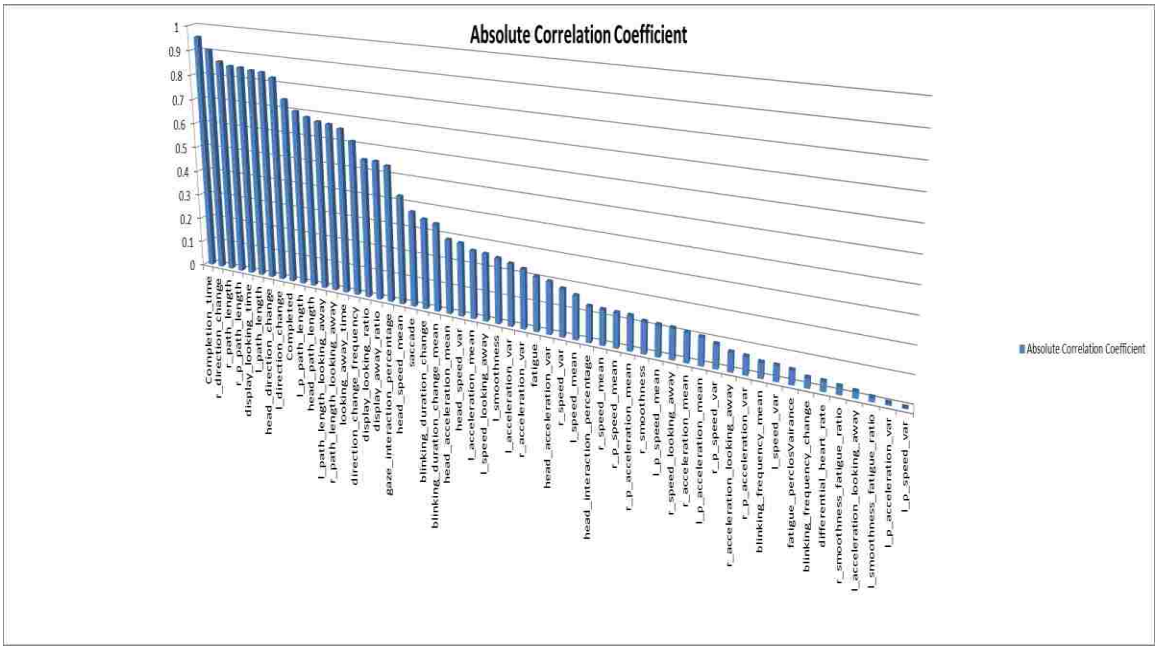


Figure 5.1 The absolute correlation between the measured metrics and the skill level

Table 5.3 The variance of the metrics values for the novice, intermediate, and expert subjects

Variables	Novice Variance	Intermediate Variance	Expert Variance	All Subjects Variance
Completion_time	0.013351	0.010323	0.000776	0.082647
r_direction_change	0.021797	0.005144	0.001650	0.079841
r_path_length	0.040926	0.002027	0.001216	0.085527
display_looking_time	0.054518	0.020253	0.001232	0.083693
r_p_path_length	0.008952	0.000554	0.000322	0.014855
l_path_length	0.034149	0.001220	0.000521	0.070553
head_direction_change	0.035022	0.003422	0.000378	0.046226
l_direction_change	0.045129	0.001538	0.000396	0.066154
l_p_path_length	0.020502	0.000905	0.001639	0.016128
head_path_length	0.057846	0.005056	0.000111	0.042123
l_path_length_looking_away	0.071944	0.003274	0.000749	0.056546
r_path_length_looking_away	0.074890	0.006556	0.000291	0.052134
looking_away_time	0.029474	0.006553	0.000052	0.020101
direction_change_frequency	0.072599	0.007239	0.002163	0.048859
display_looking_ratio	0.053406	0.029619	0.001078	0.035908
display_away_ratio	0.053406	0.029619	0.001078	0.035908
gaze_interaction_percentage	0.018236	0.012148	0.000783	0.013347

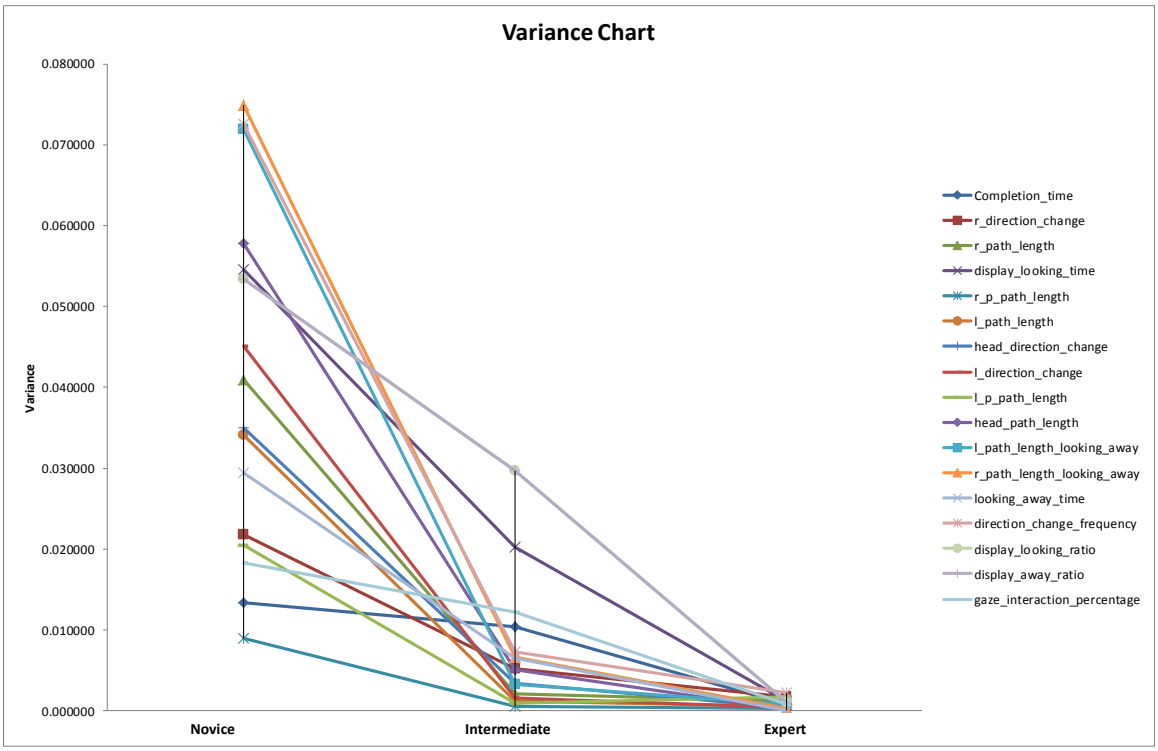


Figure 5.2 The variance of the metrics values for the novice, intermediate, and expert subjects

The metric that has the most significant correlation with the skill level is the completion time. Completion time represents the time taken by each subject to finish the required task. The Pearson correlation value for completion time is $|r| = 0.95$ and p-value = 1.31793E-30. Figure 5.3 shows a bar graph of the completion time and the subjects' skill level. The graph shows a clear trend of time drop from the novice to expert level. There is a level of variance within each experience level. There is also a clear level of variance between each pair of experience categories, especially between the intermediate and expert levels for this metric. The correlation coefficient shows that 5% of the records do not tightly follow this trend. As we can see in Figure 5.3 some of the intermediate subjects consume more time than some subjects in the novice level to complete the task. However, this metric only represents the completion time but not the completion status: whether it was successfully completed or if there were errors in performance.

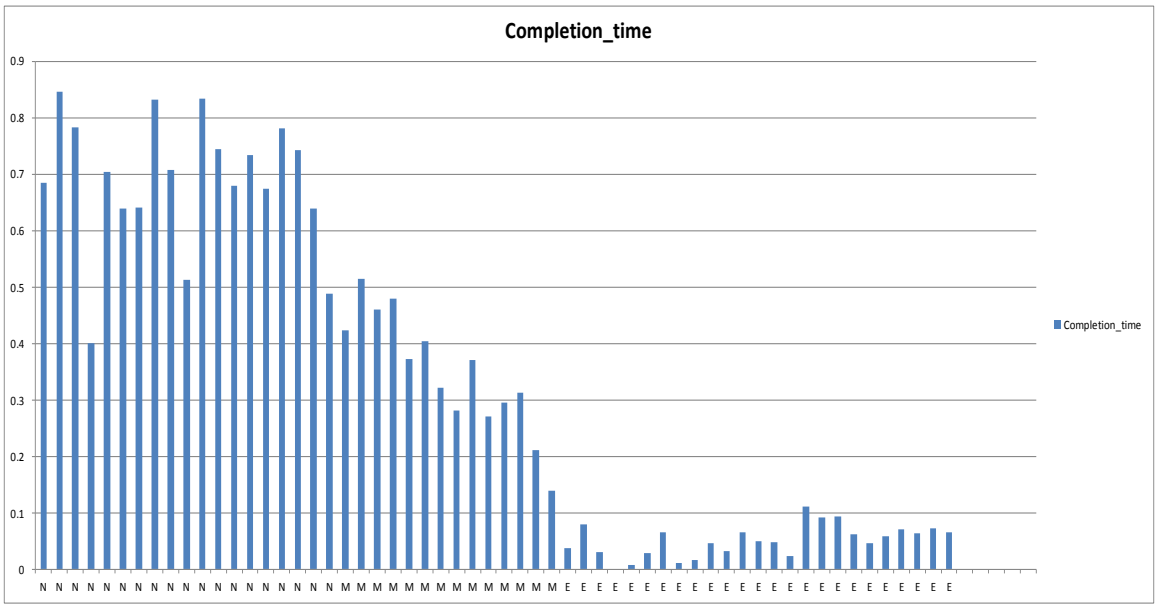


Figure 5.3 Completion time with |r| = 0.95

The metric that has the second highest correlation is r_direction_change which is the direction change in the right hand. The Pearson correlation for this metric is |r|= 0.91. Figure 5.4 shows a trend of magnitude of this metric with the skill level. The more experienced the subject is, the less the magnitude of this metric. The variance in the metric magnitude between the novice and the intermediate level is large but it is smaller between the intermediate and expert levels. We can see from Figure 5.4 that the magnitude of some of the intermediate subjects is less than it is at the expert level. None of the metrics' magnitudes at the novice level is less than any of those at the intermediate or expert levels. We also can see that the variance among novice levels is larger than it is among the intermediate and expert subjects.

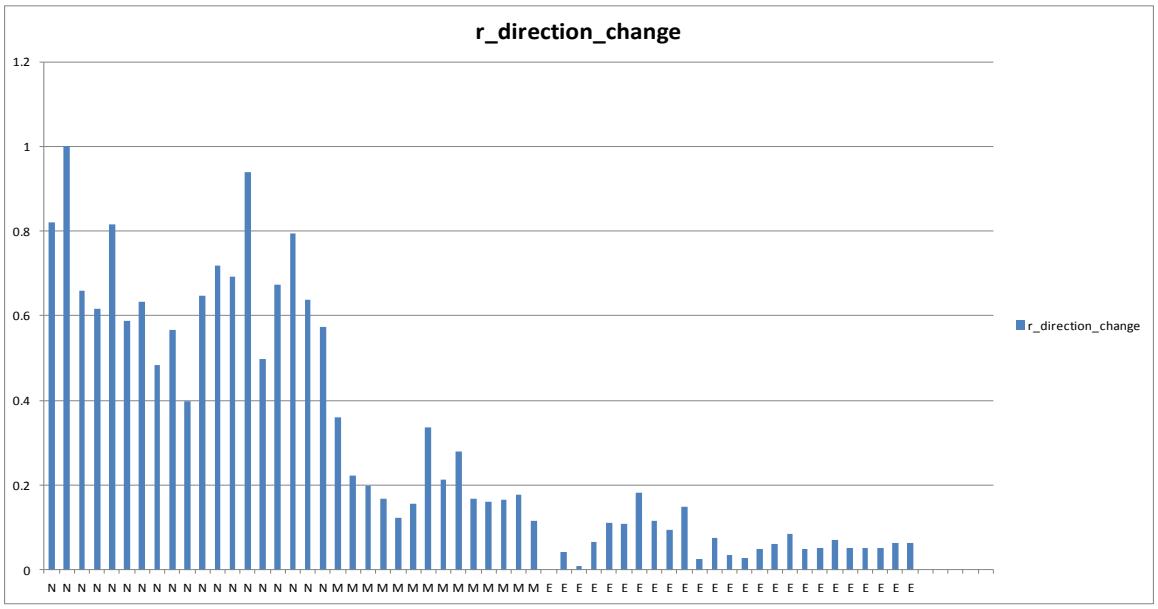


Figure 5.4 Right hand direction change with |r| = 0.91

The path length of the right hand (r_path_length) is the third significant metric because the Pearson correlation value shows |r| = 0.86. Figure 5.5 shows the differences in the magnitude of the right hand path length. Similar to the previous two metrics, there is a change in the magnitude based on the level of experience and a clear drop in the magnitudes between the novice and the intermediate levels, along with a slight drop between the intermediate and the expert levels. As we can see in the graph there are a number of expert subjects that have a higher magnitude of r_path_length than the intermediate subjects. This number is more than it is in the r_direction_change because the correlation coefficient for r_path_length is less. But none of the novice subjects has a magnitude lower than that of any of the expert subjects. These observations show that the performance of the intermediates is closer to the performance of the experts than it is to the novices'.

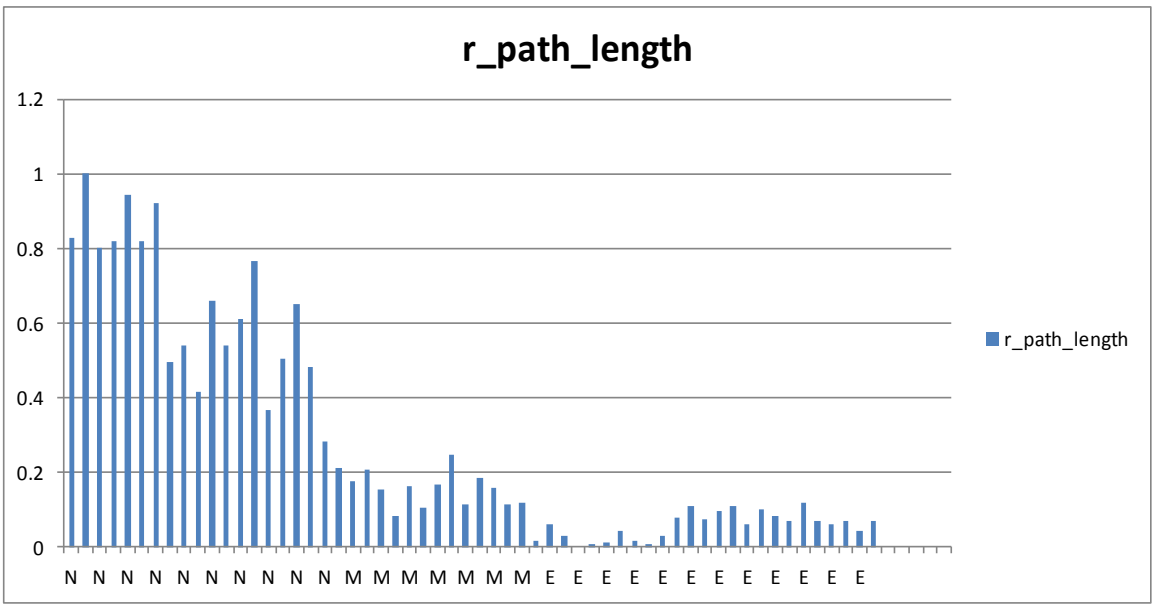


Figure 5.5 Right hand path length with |r| = 0.86

The fourth metric is display_looking_time |r|=0.85. This metric represents the magnitude of time of looking at the display from the completion time of the task. This metric is a fusion one thus; to measure it, we need coordination between the Vicon, the eye-tracker, and the laparoscope systems. Even though there is a similar trend within the magnitudes of this metric like the metrics discussed above, there is an interesting result for four records at the novice level. Figure 5.6 shows the four records at the novice level where the display_looking_time is low compared to the rest of the novice subjects. On tracking those records, we found that three of them belonged to one subject in three sessions and the fourth to a different subject. This subject based on this result did not change the direction of his head and gaze while performing the task. We could interpret from the graph that the magnitudes of this metric at the intermediate level is closer to that of the novice than it is to the expert. It was seen that as the Pearson correlation decreased in the metric, the level of similarity in the metric's magnitude among experience levels increased.

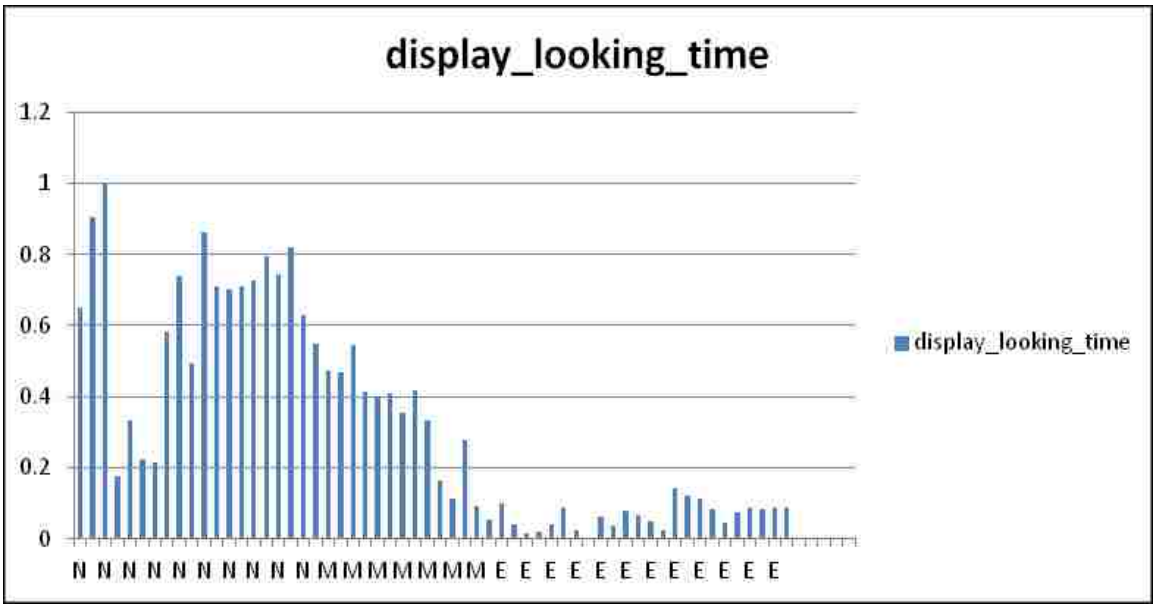


Figure 5.6 Time looking at the display with |r|=0.85

The right probe path length r_p_path_length represents the magnitude of the path length of the instrument controlled by the right hand. The Pearson correlation for this metric is 0.85. As Figure 5.7 shows, the magnitudes of the intermediate and expert subjects are close to each other. There is a clear difference in the magnitudes of the novice over the intermediate level. Two records from the novice level showed magnitudes higher than all the other novices at a significant rate. After tracking those two instances, we found that they belonged to the same subject in two different sessions.

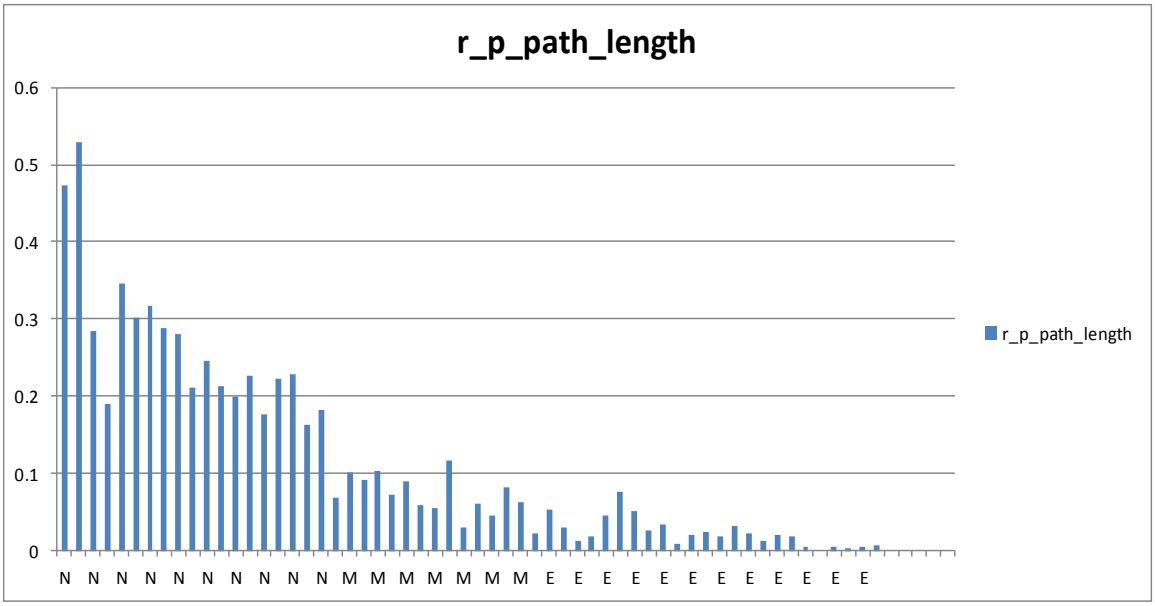


Figure 5.7 Right probe path length with |r|=0.85

The features of the rest of the metrics with correlation higher than 0.5 are reported in Figures 5.8–5.19 in Appendix B. Each graph represents the magnitudes of one metric at the three different levels sorted on the absolute value of the Pearson's correlation coefficient. All Figures show that as the level of significance decreases, the clarity of the trend of magnitude reduction decreases at all levels. This reduction is more obvious between the intermediate and expert levels than it is between the novice and the intermediate. In many cases, there is also a clear increase in the magnitude variability among subjects of the same level as in Figure 5.15 looking_away_time (|r|=0.66) and Figure 5.16 direction_change_frequency (|r|=0.61). In addition, the trend of the variability within the magnitude of metrics among the novice subjects continues to be higher than it is at the other two levels in all the graphs but at different levels. For example, the variability in l_path_length_looking_away (|r|=0.67) in Figure 5.13 and direction_change_frequency (|r|=0.61) in Figure 5.16 is higher than it is in l_p_path_length (|r|=0.70) in Figure 5.11.

Gallagher and Satava [15] and Gallagher et al. [43] have used MIST-VR system to compare and assess laparoscopic psychomotor skills. Thirty-six subjects participated in the study: 12 experienced, 12 intermediate (inexperienced), and 12 novices. Each

subject performed ten sessions on the MIST-VR. Each session included six MIST-VR tasks. The metrics collected for the evaluation were: completion time, error, economy of movement (left and right), and economy of diathermy. The studies reported that all the metrics showed differences among the three groups which reflected the differences in levels of experience among the groups. In trial one, the experienced group was the fastest and the novice group was slowest. Less experienced and novice groups showed significant improvements in completion time through the trials, whereas the experienced group showed less improvement. The novice and intermediate groups showed higher error rates with the highest evident for the novice. Both groups showed a significant drop in the error rate up to trial four. Economy of motion for the left and right instruments also showed that the experienced group had higher and better economy than the other groups, while all the groups showed significant improvement in the ten trials. Similarly, the economy of diathermy which was used in tasks five and six showed differences among the three groups. All groups reached the performance variability plateau by trial five.

The graphs and discussions presented above for the metrics with $|r|>0.5$ match the results achieved by Gallagher and Satava and Gallagher et al. The novice subjects converged to the expert levels through the intermediates' by training. As the experience level increased, the differences among them dropped gradually until they reached the experienced level. At that stage, the change rate at the novice level was higher than it was at the intermediate and expert levels. Further, the magnitude variance and differences between the intermediate and expert became smaller compared to the difference between the novice and other levels. Our results showed that metrics related to speed and acceleration, which were used in Gallagher and Satava and Gallagher et al. studies and widely used in other studies, were not the best metrics to use for the assessment. The significance of the correlation of these metrics was low. We introduced many new metrics that showed high correlation to the skill level. Many of the metrics used were fusion metrics from fusion motion analysis compounded by coordinating multiple sensors. These metrics could only be extracted using computer

vision technology to coordinate cues of the eyes, external shots of the body and instruments, and internal shots of the operative field.

This chapter presented the results and analysis of the measured metrics on an individual basis. It showed the correlation coefficient of each metric with the skill level and the effect of training on the magnitudes of the metrics. The data represented 55 metrics which were taken for 17 subjects in 70 sessions. The system design and one of the study's goals aimed at gaining the capability to collect a wide range of metrics with potential of correlation to the skill level. These metrics were fusion and non-fusion metrics. The collated metrics were oversampled. A large set of metrics showed significant correlation but the others showed low correlation. Many of the significant metrics were new metrics which were the result of transforming the assessment problem into a computer vision problem.

We think some of the metrics, especially the ones related to stress and fatigue could have shown more significant correlation using more stressful tasks as they might be functions of experience and complexity together. The chapter also discussed the metrics variance and difference in the variance levels among groups and within each group. The individual metrics did not produce reliable or highly accurate data for assessment. But studying the metrics as a whole presents the complete picture and gives high accuracy and creates a precise model for the assessment. The next chapter will analyze the relationship between the subjects and among the metrics to find the hidden patterns in the metrics to present a reliable assessment model.

Chapter 6

Analysis and Discussion

This chapter presents a detailed analysis using Principal Component Analysis (PCA) to understand the features of the data and the interrelationship between subjects and variables, and detect the skill patterns over time. Visualization and interpretation of this analysis are presented to clarify the achievements and the interrelationship. The analysis includes a detailed study about the reliability and the effect of noise in the data on the features of the acquired metrics. The chapter in addition provides supervised and unsupervised data mining analysis in order to emphasize more the features of the data and achieve a reliable model to classify subjects. The chapter proposes a classification model that can evaluate subjects with high accuracy. Finally, the chapter presents comparison of the results with previous studies' results and compiles a list of outcomes.

6.1 Introduction

The goal of this analysis is to find a set of metrics that can accurately and reliably assess the surgeons and trainees. As we have seen in Chapter Five, the collected metrics are oversampled and have various correlation significances. The assessment of the subject's skills cannot rely on one or two metrics only to a level of tolerance for measurement errors in order to increase the reliability and robustness. The dimensionality of the data is high and a pattern of skill level cannot be seen within it. To

reduce the dimensionality of the data in order to find the patterns and the hidden information, we used Principal Component Analysis (PCA).

PCA is a statistical technique to decompose and reduce the dimensionality of the data in order to detect features of variance in it. These features enable us to understand the relationship between variables and how the data is structured if the dimensionality is high and not humanly readable. By calculating the eigenvectors and eigenvalues for the data covariance matrix, we can find components that represent the variance directions of this data. The set of components that explain the variance of most of the variables can be used to study the structure of the data. This data composition also allows visualizing the high dimensionality on papers in two or three dimensions by using the significant components. The principal components that can be extracted from data are less than or equal to the data dimension. The central axis of the direction of maximum variance is the best component. The axis with the second maximum variance is the second-best component and so on. These axes are the eigenvector with the largest eigenvalue, second-largest value, and so on. In this study, each acquired metric represents a dimension/variable in the model and each subject's record represents a vector. For more details about PCA review [69]. In our analysis, we used the Unscrambler® X from CAMO software to implement the PCA analysis [72].

The PCA is used to study the structure of the data we collected and understand the hidden features in the variance of the variables. We also used the highest principal components to visualize the data variance and the clusters within it. Finally, the PCA is used to find a reliable number of metrics to build a robust assessment model for a classifier. To find the reliable and robust model, we performed various experiments through adding Gaussian White Noise to the data. The robustness experiments we performed have been explained later in this chapter.

6.1.1 PCA Using Various Number of Metrics

We initially sorted the normalized data in descending order based on the significance of their correlation with the skill level. We built three models from the data in order to perform analysis and find the effect of the variables and the best model to provide the assessment.

The first model includes metrics that have a Pearson correlation coefficient ($|r|>0.5$ or P value < 1.14838E-05). The list of metrics and their correlation coefficients and P values can be found in Table 6.1. The number of metrics that have the required significance is 18. We decided to drop the completion time and completion status for reasons explained in the next section 6.2.1. Thus the number of metrics left in this model was 16. We will refer to this model from now on as the 16-metric model. These metrics represent the dimensions of this model in the PCA analysis. The dimensions of the data matrix are either 58x16 or 70x16 based on the number of data sessions used in each specific experiment. Fifty-eight is the data captured in the sessions before the analysis model was built. Seventy were the total data sessions including the 12 sessions captured after the data model was built for validation.

The second model includes all the 55 measured metrics. We built this model to study the features of all the metrics combined and the effect of lower significant-metrics on the structure of the data while comparing it with the 16-metric model. We can see which model performed better by clustering the data. The PCA vectors' dimensions for this model were 55 and the data matrix dimension was 58x55. We will refer to this model from now on as the All Measured Metrics model.

The third model includes the three metrics with the highest correlation coefficients but excludes the completion time and completion status. These three metrics are r_direction_change, r_path_length, and display_looking_time. We built this model to study the features of the most significant metrics and compare it with the 16-metric model. The main result we were looking for by building this model was the effect of

reducing the number of metrics on the accuracy, reliability, and robustness of the assessment. We needed to find the accuracy versus robustness using this model compared to the 16-metric model. The PCA vector dimension for this model was three and the matrix dimension was 58x3. We will refer to this model from now on as the 3-metric model.

In all PCA models, the matrix built looks like the matrix in Equation 6.1 where each row represents a record of a subject and each column represents the magnitude of one metric for all subjects. M is the value of a variable for a subject, m is the subject's number, and n is the number of metrics (variables).

$$\begin{bmatrix} M_{11} & M_{12}.....M_{1n} \\ M_{21} & M_{22}.....M_{2n} \\ ..... & ..... \quad ..... \\ M_{m1} & M_{m2}....M_{mn} \end{bmatrix} \tag{6.1}$$

6.1.2 Validation

To analyze the performance and accuracy of these models, we used both, cross-validation and a test set validation in addition to a perturbation study to find the effect of noise on them.

6.1.3 Leave-one-out validation (LOOCV)

LOOCV is a type of cross-validation. In this technique, a single data record (data for one subject in one session) from the original dataset is used for validation and the rest of the dataset for training. The method is then repeated such that every record in the original dataset is used as a validation sample. We used this level of validation because the dataset we have is not large.

6.1.4 Perturbation

To study and compare the robustness and reliability of the 16-metric and 3-metric models, we perturbed the data. We added various levels of noise to the data of both models and plotted graphs to show the effect of the noise on the values of each experience level in the first two principal components. This analysis shows the mutual support that metrics could provide each other if noise or corruption affected some other metrics. The noise we added to the data was Gaussian white noise and the level was controlled by the magnitude of the variance. The magnitude of the variance was added as a percentage of each variable data span.

Two different kinds of perturbation were applied separately on the data. The first one was noise that was gradually increased on all the variables of the data to find out the tolerance of both models to noise. The two models were compared with each level of noise. The second experiment was that a large level of noise was applied to one variable in the dataset of both models. The variable we picked was the one whose correlation to the skill level was most significant. The most significant metric was the r_direction_change (r =0.91). The magnitude of the noise was large and that dominated the value of the variable while enlarging the error within it. This experiment was useful to study the robustness of each model and determine the degree to which each one could tolerate corruption in capturing one or more variables.

6.1.5 Score Graph

The score plot which can be generated from the PCA result is an important tool to understand the relationship among the studied subjects. The graph can be generated using any eigenvectors against each other. The plot maps summarize the relationship between the subjects in the principal components' subspace. Each principal component covers a ratio of the data variance and thus, the higher the variance covered by a component, the more the plot manifests the relationship. Usually the most important components are the first three. Here, we show the score plot either between the first

and the second principal components or the first and the third principal components. Also, the variance coverage of each component is added to the graph.

6.1.6 Loading Plot

The loading plot is another important tool to understand the relationship and the correlation among the variables. The loading plot can be viewed as a summary or a map of the variables and shows the size of the contribution of each variable to the principal components.

The PCA scores and loading vectors can be calculated using several methods in which the Singular Value Decomposition (SVD) is a popular one. According to SVD, any matrix X_{nxm} can be decomposed into three matrices $X = T_0 . S . P^T$ where T_0 is an mxn normalized PCA scores matrix. S is an mxm matrix which contains the singular values in the diagonal. P^T is the transposed mxm loading matrix. The PCA scores can then be calculated by $T = T_0 . S$ [73].

6.2 Principal Component Analysis

This section presents the results of PCA on the three data models.

6.2.1 PCA Analysis on the 16-metric Model

To find a set of reliable metrics to evaluate the subjects and validate the system, we have chosen from the 55 measured metrics, a subset that has a correlation coefficient r> 0.5. Table 6.1 shows the list of metrics that has significant correlations sorted based on the absolute correlation coefficient. The details about this list of metrics and how they are calculated can be found in Chapter Three.

Table 6.1 List of metrics with r >0.5

Variables	Absolute Correlation |r|	Description
Completion_time	0.95	Time take to complete the task
r_direction_change	0.91	right hand accumulation direction change
r_path_length	0.86	right hand path length
display_looking_time	0.85	The time spent looking at the display
r_p_path_length	0.85	right probe path length
l_path_length	0.84	left hand path length
head_direction_change	0.84	the head accumulation direction change
l_direction_change	0.82	left hand accumulation direction change
Completed	0.74	task successfully completed or not
l_p_path_length	0.7	left probe path length
head_path_length	0.68	head path length
l_path_length_looking_away	0.67	left hand path length while looking away from the display
r_path_length_looking_away	0.67	right hand path length while looking away from the display
looking_away_time	0.66	total time of looking away from the display
direction_change_frequency	0.61	the frequency of head direction change
display_looking_ratio	0.55	the ratio of time looking at the display compared to the total completion time
display_away_ratio	0.55	the ratio of time looking away from the display compared to the total completion time
gaze_interaction_percentage	0.54	the percentage the gaze intersect with the display(the previous attributes measured using the face direction, but this one is the eye gaze)

We decided to eliminate the completion time (r=0.95) and completion status (r=0.74) from the analysis since completion status was manually recorded and completion time was the main measure in the manual evaluation. We then applied the PCA analysis to find a relationship between the subjects' skills. The analysis produced components where the first three components could describe 94% of the original data variance. The first component (PC-1) described 74.8%, the second component (PC-2), 16.7%, and the third component (PC-3) 2.5%. Table 6.2 shows the first three principal components.

Table 6.2 The contribution of the first three principal components

	PC-1	**PC-2**	**PC-3**
Contribution	74.8%	16.7%	2.5%
Accumulation of Contribution	74.8%	91.5%	94%

The score graph in Figure **6.1** shows how the performance of the 58 subjects related to each other. The keys in the graph are as follows: N=novice; M=intermediate; and E=Expert. A letter combined with the word "validation" represents the leave-one-out cross-validation result for each subject. Subjects close to each other have similar properties, whereas the properties of the subjects far from each other are dissimilar. As the graph shows, the novice subjects in the large red ellipse are scattered, do not have a consistent pattern or behavior among themselves, and are far from other subjects' properties in the other two categories represented by other ellipses. This result means that the novices perform differently even amongst themselves, which indicates that they do not have the right technique to perform the task, and they use their background which differs from person to person. However, the diamond dots (novice) converge toward the middle green ellipse (intermediate) properties area because they build their skills and learn the proper technique by practicing. The triangle dots (intermediate) in the green ellipse are less scattered and the distance among subjects becomes smaller.

This behavior explains that the subjects' properties become similar in nature because their techniques become more refined. Finally, the small blue ellipse, which represents the expert level, shows the properties of all subjects are dense and converged in a tight cluster, indicating that the properties of the performance of those subjects are more uniform, because they use the proper technique. We asked a subset of the experts' group to practice more in order for us to learn whether their values on the x-axis (PC-1) could go farther towards the negative side. The result showed that giving the expert subjects more training did not change their properties and the results continued to be in the same cluster area. Since this study is based on detecting human experience development, there is no sharp threshold between different levels. Therefore, there might be an overlap between different levels, especially between the intermediate and expert levels.

The loading graph in Figure 6.2 shows how metrics correlate with each other. The dots close to each other have high positive correlation whereas dots on opposite sides have negative correlation. For example, the graph shows positive correlation between the ratio of time spent looking at the display measured by the forehead direction and the ratio of the gaze intersection with the display. There is negative correlation between the time ratio spent looking at the display and the path length while looking away from the display. From our observation, the novice subjects tend to frequently redirect either their head or their gaze or both to the incision place instead of the display. This behavior indicates the difficulty they faced in trying to coordinate their visual perceptions and the psychomotor. Further, the experts tended to freeze the tools while they were looking away from the display, whereas the novices continued the trend of motion.

If we look at both graphs in Figure **6.1** and Figure **6.2** we find that the metrics (display_looking_ratio, gaze_interaction_percentage) contribute more in the expert and intermediate clusters. The metrics on the right side of the loading graphs contribute more in the novice and intermediate clusters. We can use these properties to give automated objective feedback to the trainees about what to improve in order to reach a

higher experience level. As a result, this system could decrease the training time by giving this continual feedback. In addition, through the analysis, we could estimate the time the trainees required to reach the level of experience based on the performance in previous sessions. To test the precision and validity of the model, the score graph shows the result of a leave-one-out cross validation performed among the 58 subjects. The validation result is presented in the graph in shapes: square for novice, star for intermediate, and plus for expert. The difference between the real and predicted data is small and none were predicted at a different level.

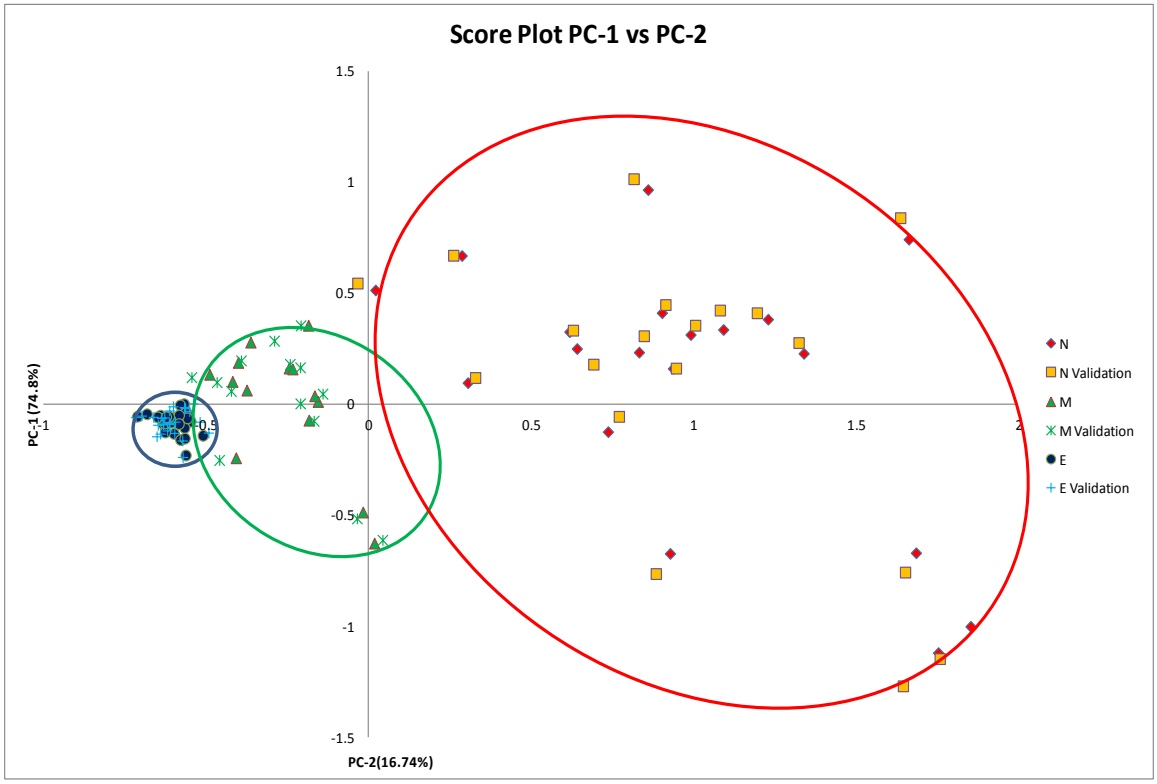


Figure 6.1 PCA score plot. PC-1 (74.8%) vs. PC-2 (16.7%)

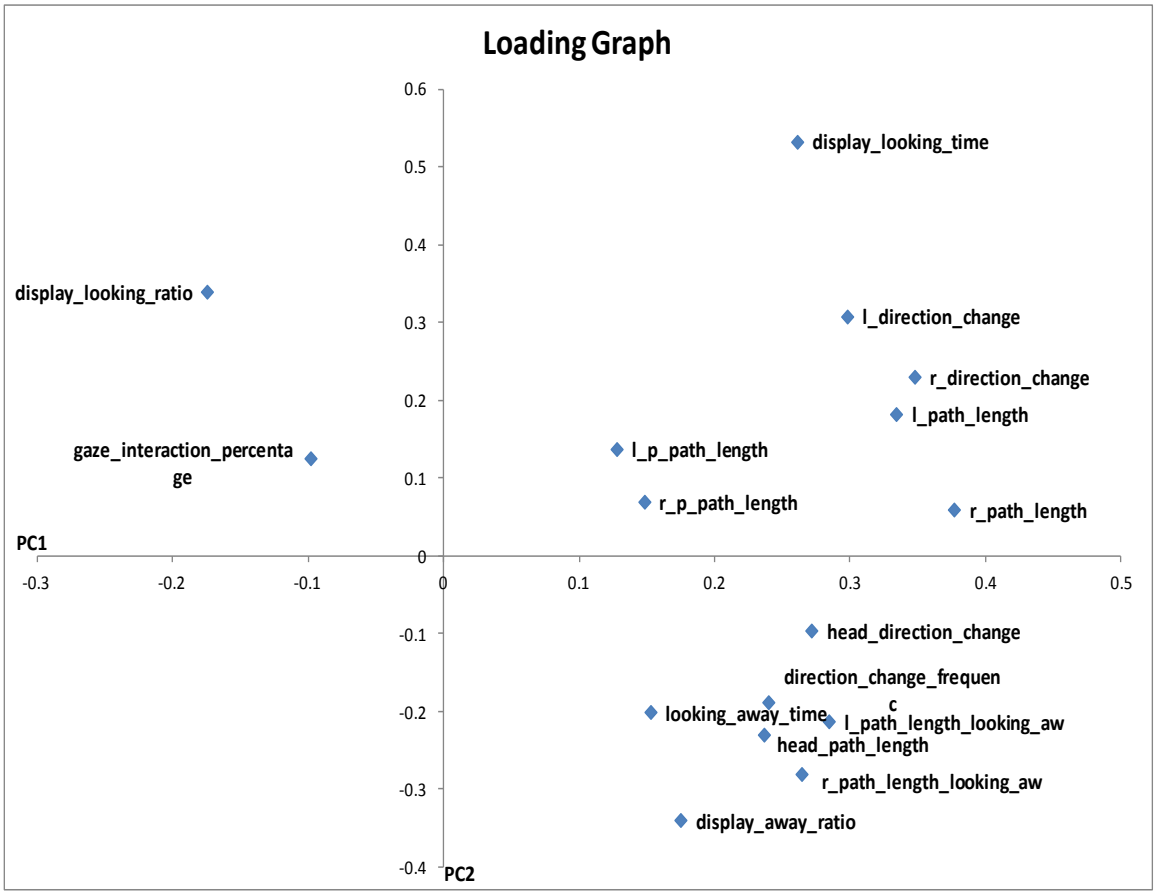


Figure 6.2 PCA loading plot. PC-1 vs. PC-2

Figure 6.3 shows the score plot for the first principal component (PC-1 47.8%) vs. third component (PC-3 2.5%). Similar to Figure **6.1**, the plot shows the relationship among subjects and how the data is clustered based on the skill level. Table 6.3 shows the calibration and validation contribution of the first four principal components using the metrics in Table 6.1. The accumulation contribution in the fourth component is 96%. The contribution of the other components is small as Figure 6.4 shows. The graph also shows that the validation curve of the explained variance grew to more than 90% at the third component.

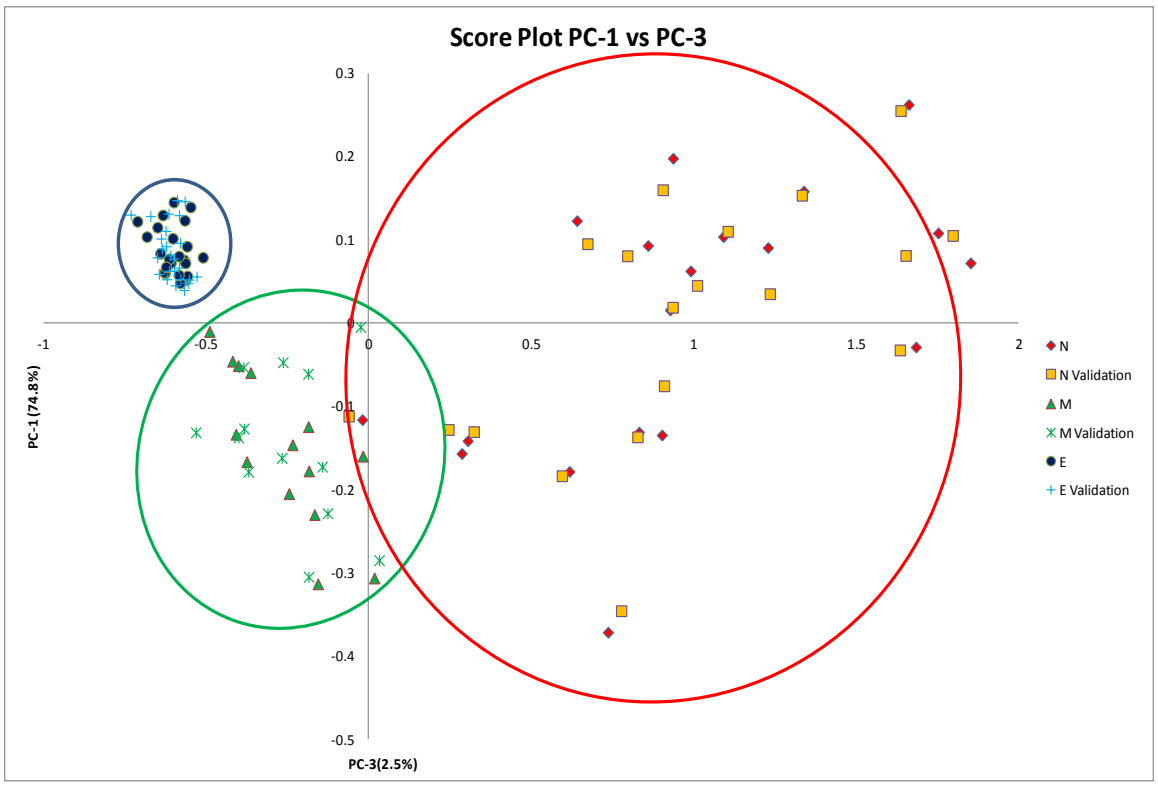


Figure 6.3 PCA score plot. PC-1 (74.8%) vs. PC-3 (2.5%)

Table 6.3 The contribution of the first 4 principal components using the metrics in Table 6.1

	PC-1	PC-2	PC-3	PC-4
Calibration	74.84521	91.58828	94.05439	96.00191
Validation	69.50325	88.73438	90.55489	92.63554

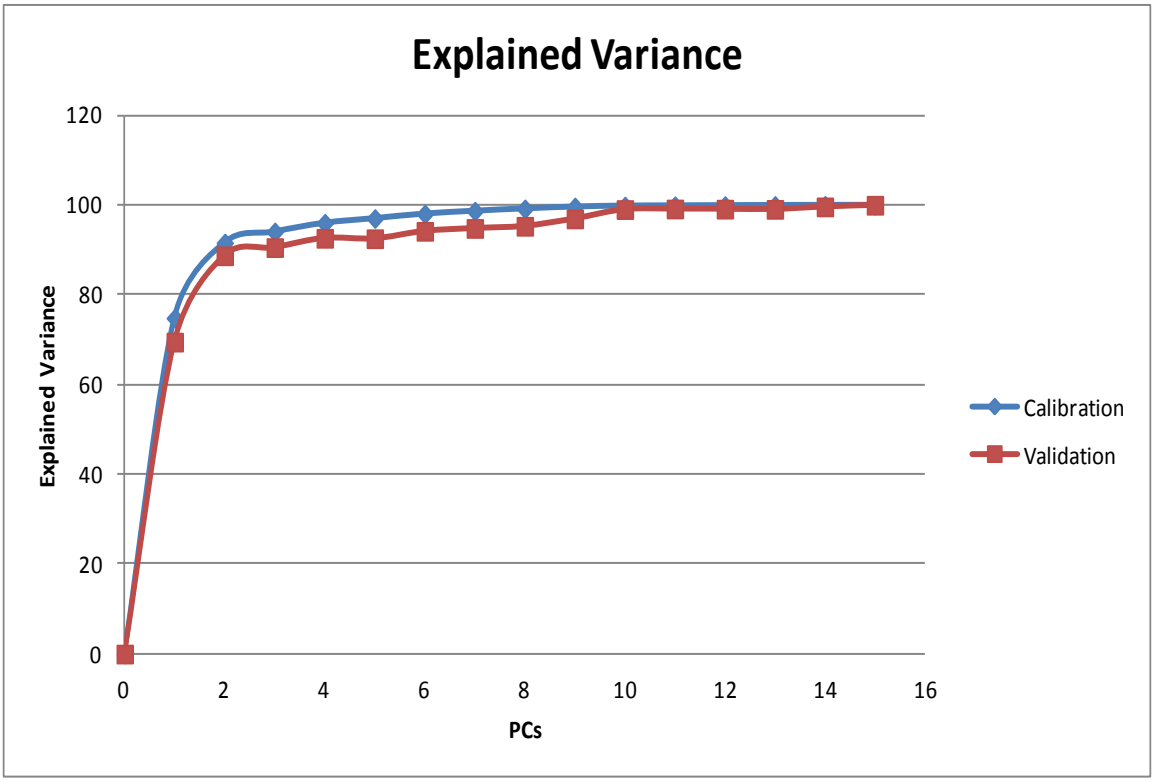


Figure 6.4 The variance contribution of the first 15 principal components using the metrics in Table 6.1

6.2.2 PCA Analysis on All Measured Metrics Model

We used all the metrics described in Chapters Three and Five in the analysis model to compare its accuracy with the one that used the metrics with r>0.5. As Table 6.4 and Figure **6.5** show, the first three principal components can explain 74.89% of the variance compared to 94% using the 16-metric model. At the eighth principal component only 89.8% of the variance is explained. In addition, the difference between the calibration and validation contribution in this model is large. At the 8th principal component, the accumulation of the validation contribution is 77.3% compared to 90.5% in the third principal component of the model built using the selected 16.

Table 6.4 The first 8 principal components' contribution using all measured metrics

	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8
Calibration	47.46	66.92	74.89	79.07	82.42	85.26	87.70	89.80
Validation	42.36	60.90	69.40	70.17	72.36	74.11	75.93	77.31

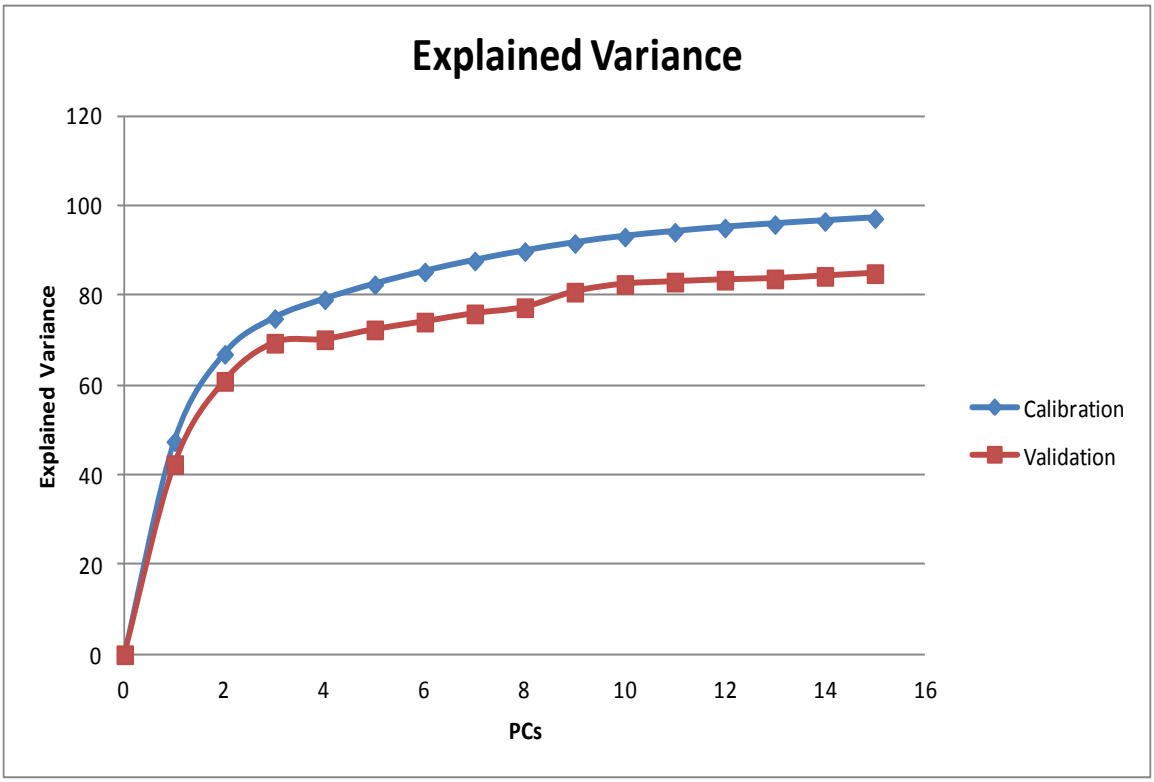


Figure 6.5 The variance contribution of the first 15 principal components using all metrics

Figure 6.6 presents the score plot for PC-1 vs. PC-2 for the model built using all metrics. The graph shows trends in the data similar to the trends in Figure **6.1** but the overlap between the clusters has increased. The experts in the blue ellipse are more scattered and closer to the intermediate and novice subjects. This result means the possible error in classifying subjects using this model could be higher than the previous model. In addition, the mean of the distance between the calibration and the validation

value is 0.083 compared to 0.036 in the previous model. Thus, the 16-metricmodel is more reliable and more accurate.

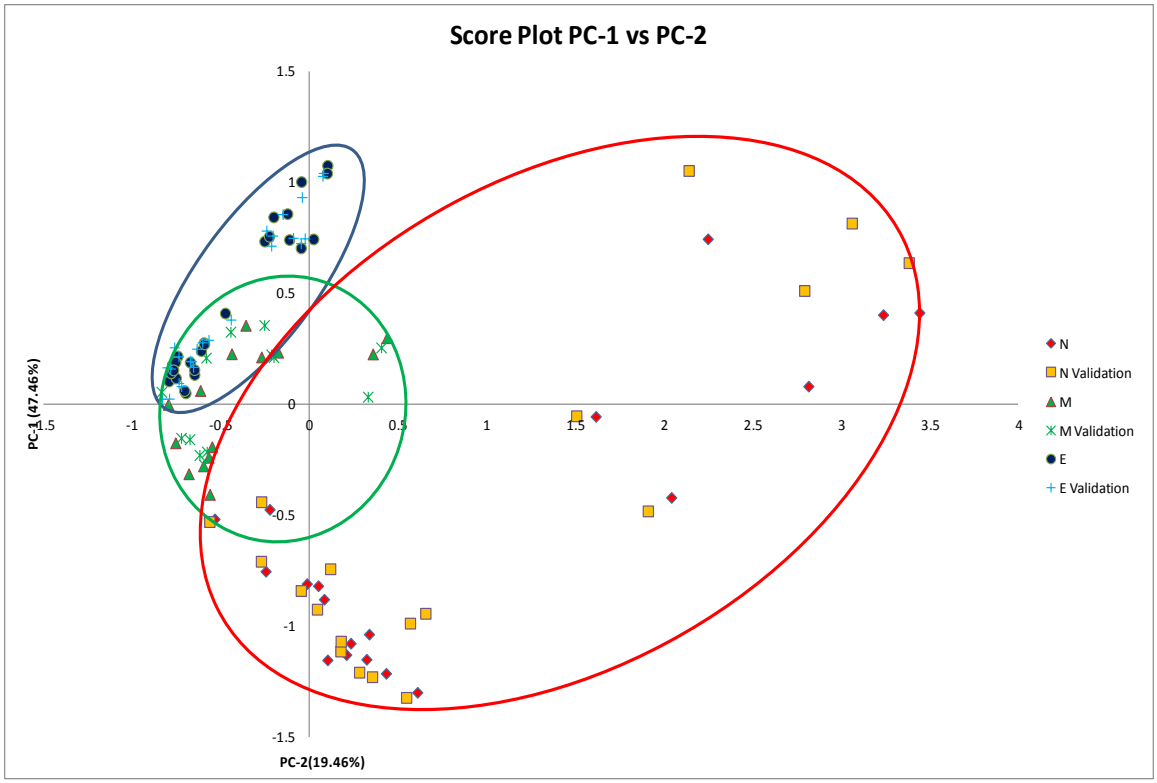


Figure 6.6 PCA score plot. PC-1 (47.46%) vs. PC-2 (19.46%) using all measured metrics model

6.2.3 PCA Analysis on the 3-metric Model

To study the effect of using a large number of metrics on the robustness of the assessment, we performed the PCA analysis on a model built using the three metrics that showed the significance of the highest correlation. Since it was a low dimension model, the PCA analysis only transforms the data into a different space. Since we chose the best three metrics, we expected this model to perform similar to or better than the other models. But we also expected its tolerance to noise and error to be lower, which reduces its robustness. Table 6.5 shows the contribution of the first two principal components which sum up to 98.65% of the variance whereas it is 94% in the first three

components in the 16-metric model. The score plot in Figure 6.7 visualizes the data clusters using PC-1 vs. PC-2 for the 3-metric model. The Euclidean distance between the calibration and validation in the PC-1, PC-2 domain is calculated. The mean distance between the calibration and the validation values using the 3-metric is 0.019 compared to 0.036 in the 16-metric model. The 3-metric model shows better accuracy in the validation model than the 16-metric model. But how robust is this model and how much can it tolerate errors and noise that may affect the data?

Table 6.5 The two principal components' contribution using three metrics

	PC-1	PC-2
Calibration	87.61604	98.65379
Validation	80.27165	95.41236

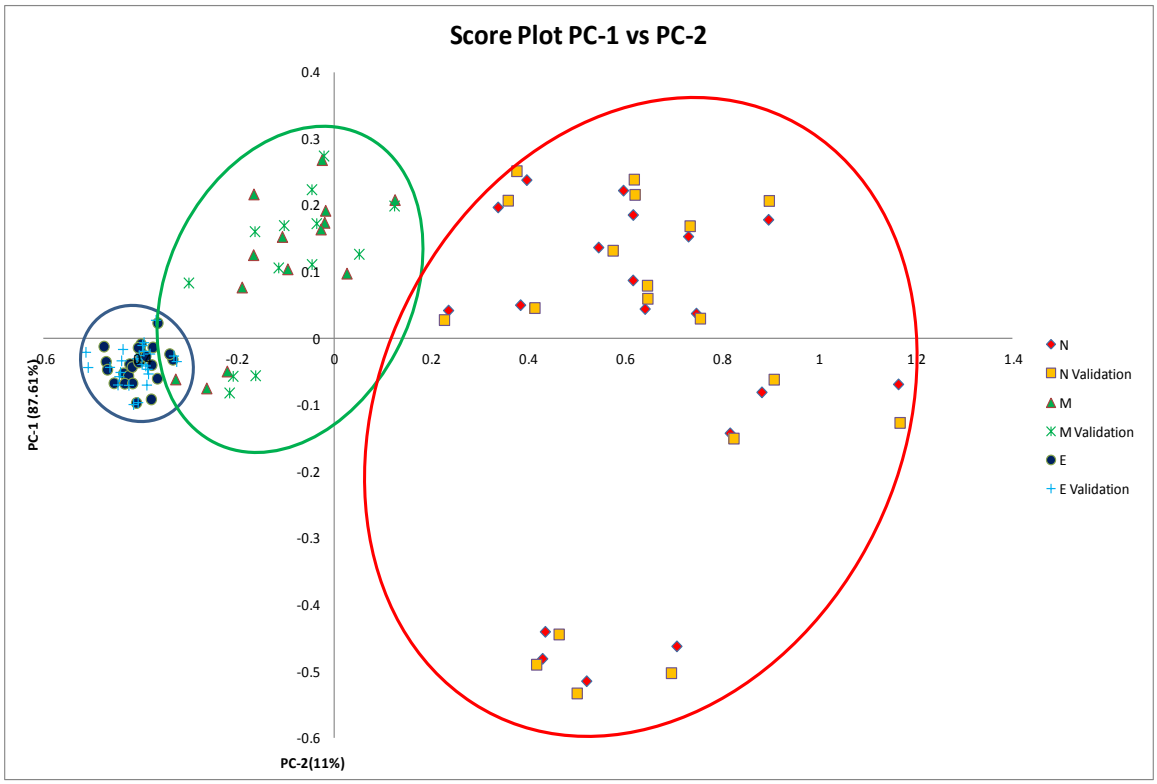


Figure 6.7 PCA score plot. PC-1 (87.61%) vs. PC-2 (11%) using three metrics model

6.3 Robustness

To show the robustness and reliability of the built system and the proposed data models, we performed the experiments described in the introduction of Section 6.1. Two experiments were performed by perturbing the data through applying the Gaussian White Noise on the data models. One experiment entailed applying the noise on all data variables and the second applied the noise on a specified variable.

In the first experiment, we applied various magnitudes of noise on all variables of the data models. The levels of the added noise were 1%, 5%, 10%, 11%, 12%, 13%, 14%, 15%, 17%, and 20% of each variable span. Then, performed a PCA analysis on the perturbed data of both models to study the effects of the perturbation level on the model and the built system's robustness. The score plots in Figure 6.8–Figure 6.19 show how the subjects' features were affected within each experience level. Each of these Figures contains two parts (a) and (b). Figure (a) represents the 16-metric model and Figure (b) represents the 3-metric model.

This experiment showed that both the 16-metric model and 3-metric model are affected by the noise but not in the same way. The effect on the 3-metric model is higher than it is on the 16-metric model. Also the effect on the experienced and intermediate levels is higher than it is on the novice levels in both models and at all the noise levels. To put it simply, this observation means that the experienced and intermediate subjects follow a pattern in order to complete the surgical task, which means the data for the subjects is correlated. The novice subjects do not perform the task based on a pattern and the features of their data are different and less correlated. Adding noise to the data affects that pattern of the experienced and intermediate subjects which reduces the correlation among their subjects. But for the novice subjects the correlation between data is low and thus, the effect of the noise is less.

As the noise level increases, the overlap among the three experience levels increases. The overlap between the experienced and intermediate levels increases at a

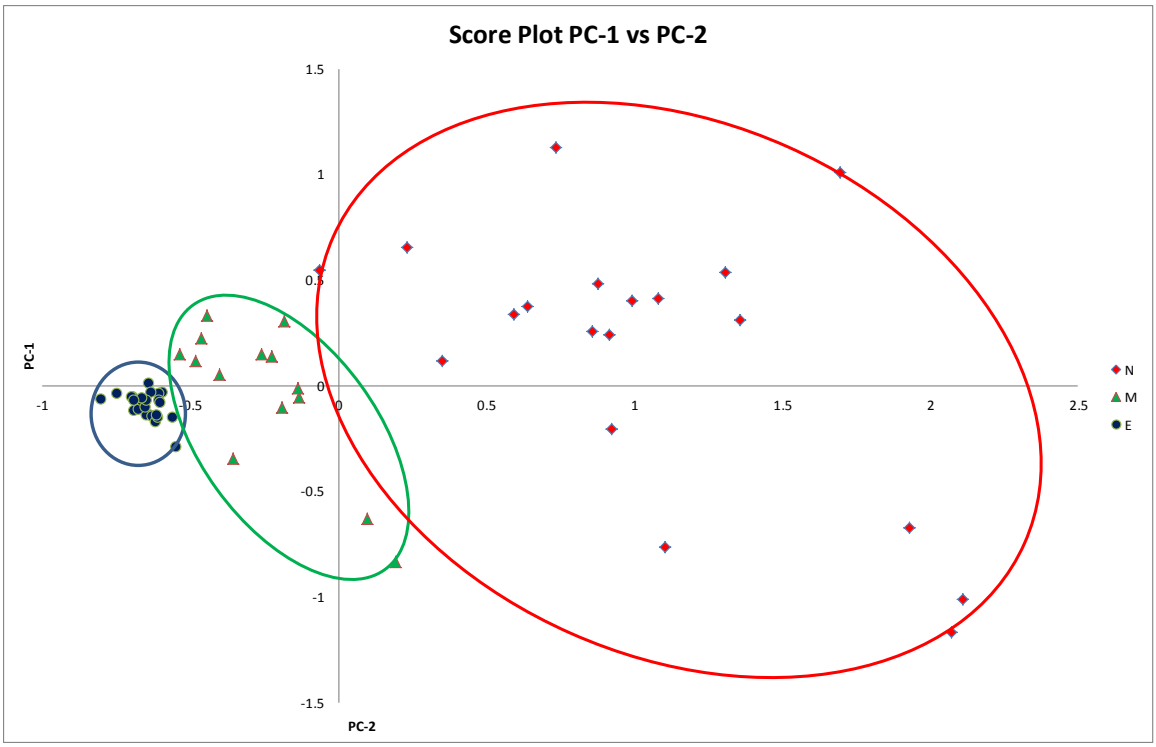
higher pace than the overlap between the intermediate and novice levels. Moreover, the disparity of the subjects within each cluster increases especially within the experience and the intermediate levels. The disparity increases at a higher rate in the 3-metric model than the 16-metric model.

Figure 6.1 shows the score plot of PC-1 and PC-2 of the 16-metric model and Figure 6.7 shows the score plot for the 3-metric model without noise. As described earlier, both models show a tight cluster for the experienced in the blue ellipse, less tightness for the intermediate level in the green ellipse, and a sparse cluster for the novice in the red ellipse. The three clusters are well separated from each other in both models. After applying 1% noise to both models, we can see some effect on the distribution of the points that represent the subjects from Figure 6.8 (a) and (b). But that effect is marginal and the three ellipses show the well-defined clusters.

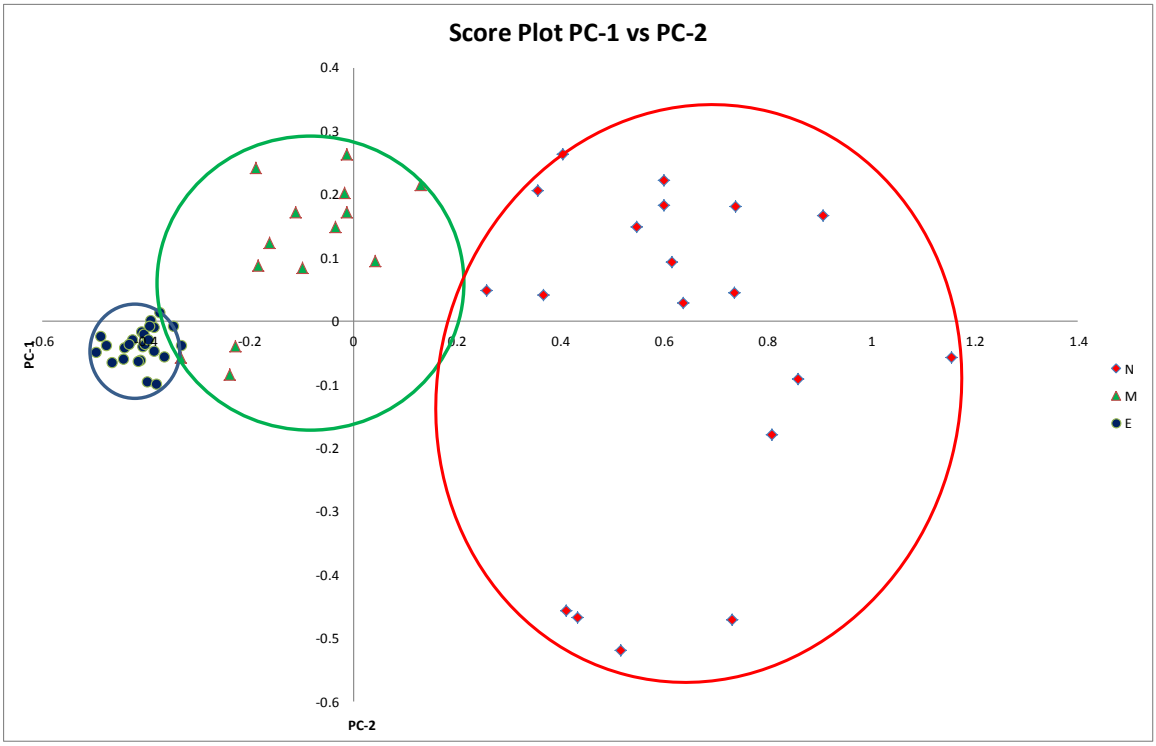
Applying 5% noise increases the disparity among the expert cluster in blue and intermediate cluster in green as Figure 6.9(b) shows. In addition, the Figure shows an overlap between the expert and the intermediate clusters. However, the effect of this level of noise on the 16-metric model as Figure 6.8(a) shows is significantly less. The disparity between the experts and the intermediates is less than it is in the 3-metric model. The overlap between the expert and intermediate subjects in this model is also less overlapped than in the 3-metric model. Both Figures show the novice subjects are minimally affected and both models show significant disparity between the intermediate and novice levels.

The 16-metric model as Figure 6.10(a) shows, tolerates the 10% noise applied on the data. The expert subjects are clustered in the tight blue ellipse. The sparse level among the intermediates is similar to it in Figure 6.1 prior to the introduction of noise. The disparity among the three levels is also clear. However, the 3-metric shows that the expert in the blue ellipse is scattered and comes close to the level of the intermediate in the green ellipse. The overlap between the intermediate and expert levels increases but

the model tolerates the noise and remains capable of distinguishing between the intermediate and the expert on the one hand and the novice on the other.

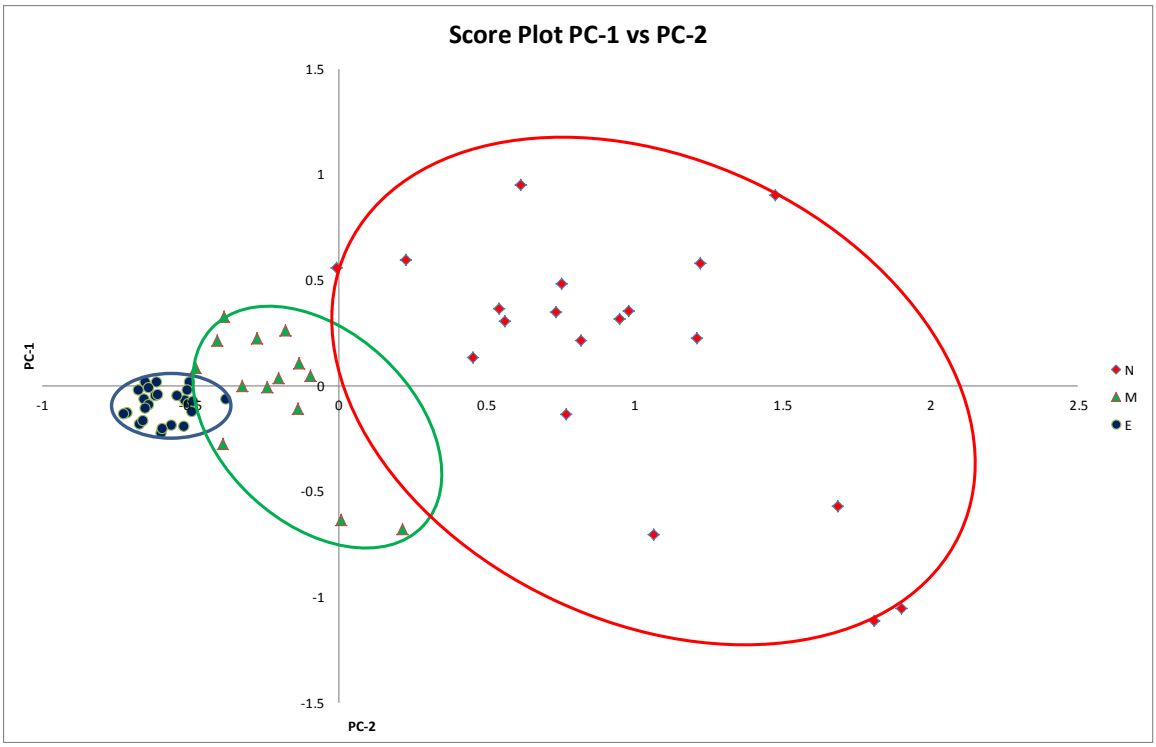


a. 16-metric model

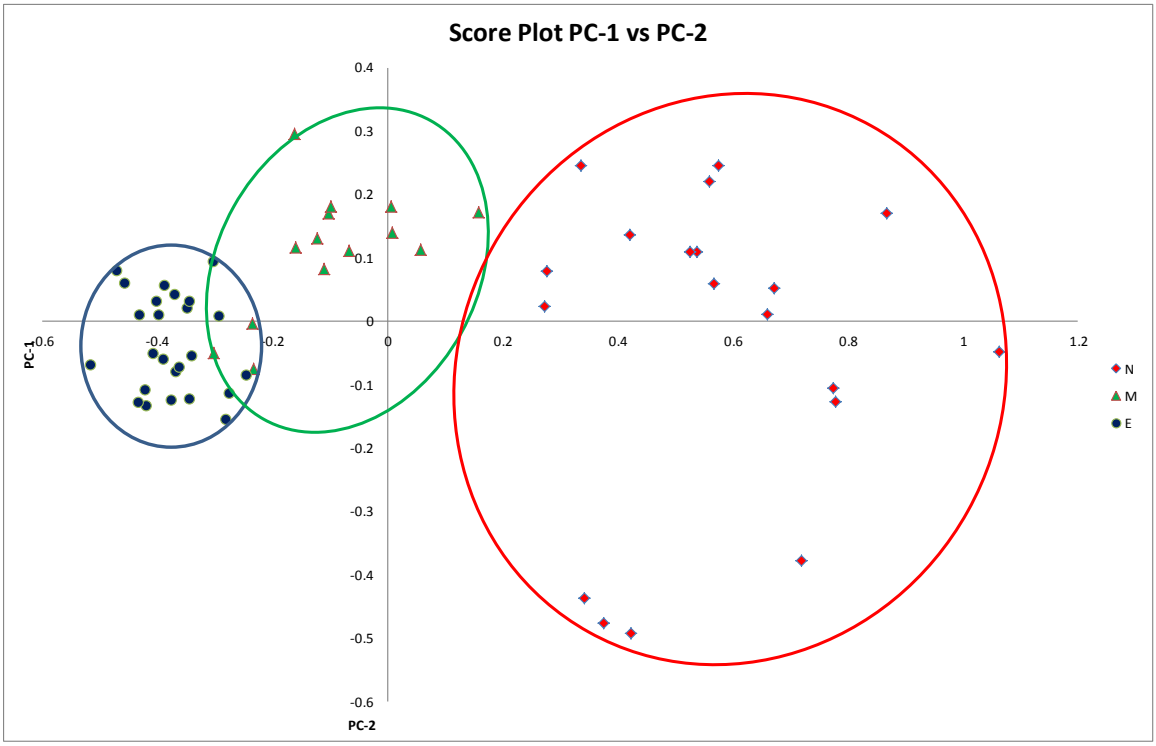


b. 3-metric model

Figure 6.8 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 1% noise applied to all metrics

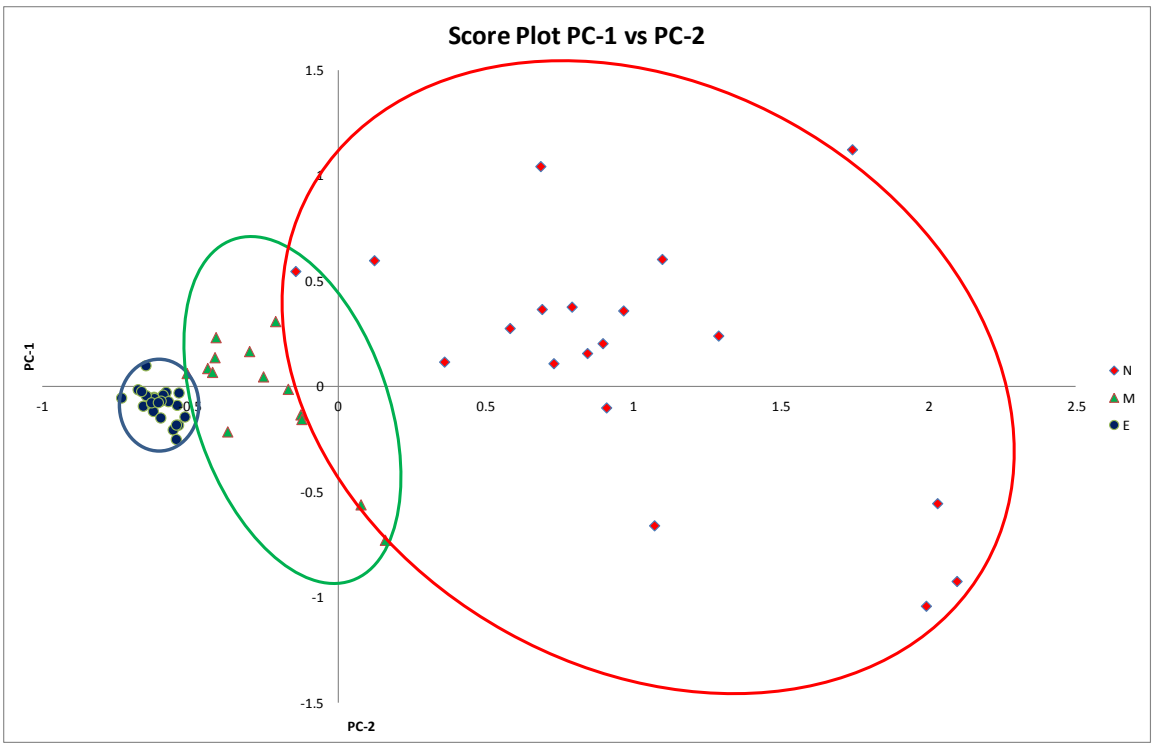


a. 16-metric model

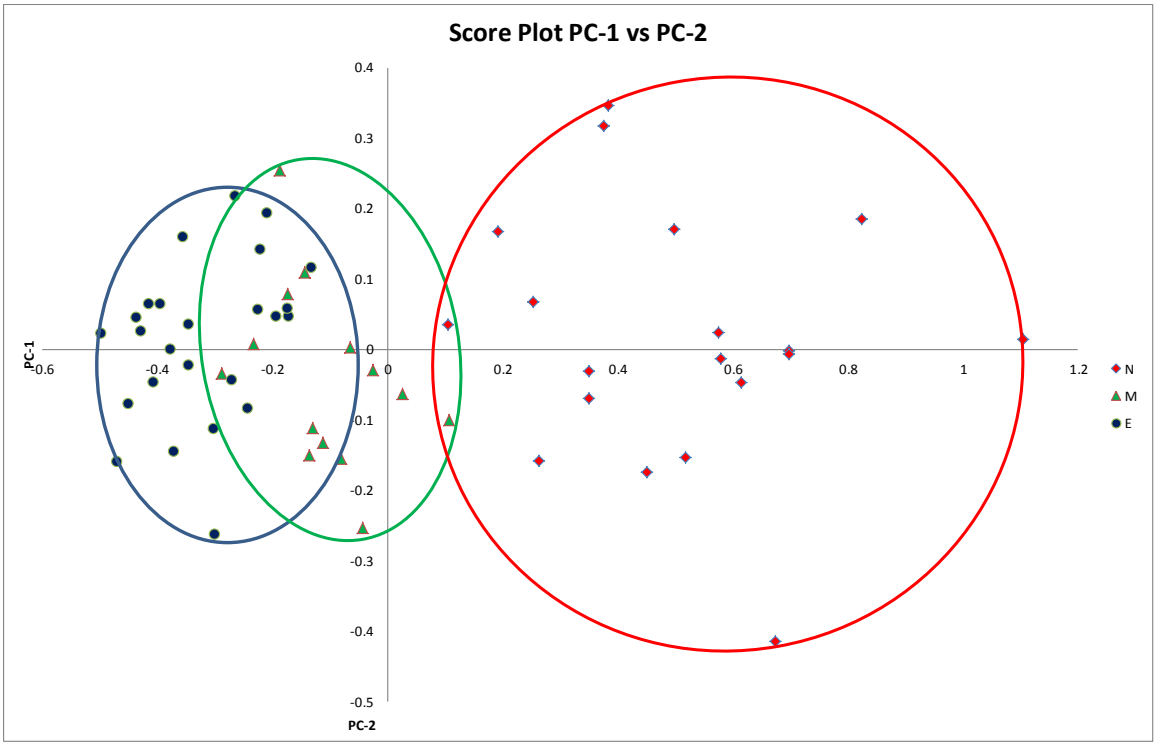


b. 3-metric model

Figure 6.9 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 5% noise applied to all metrics



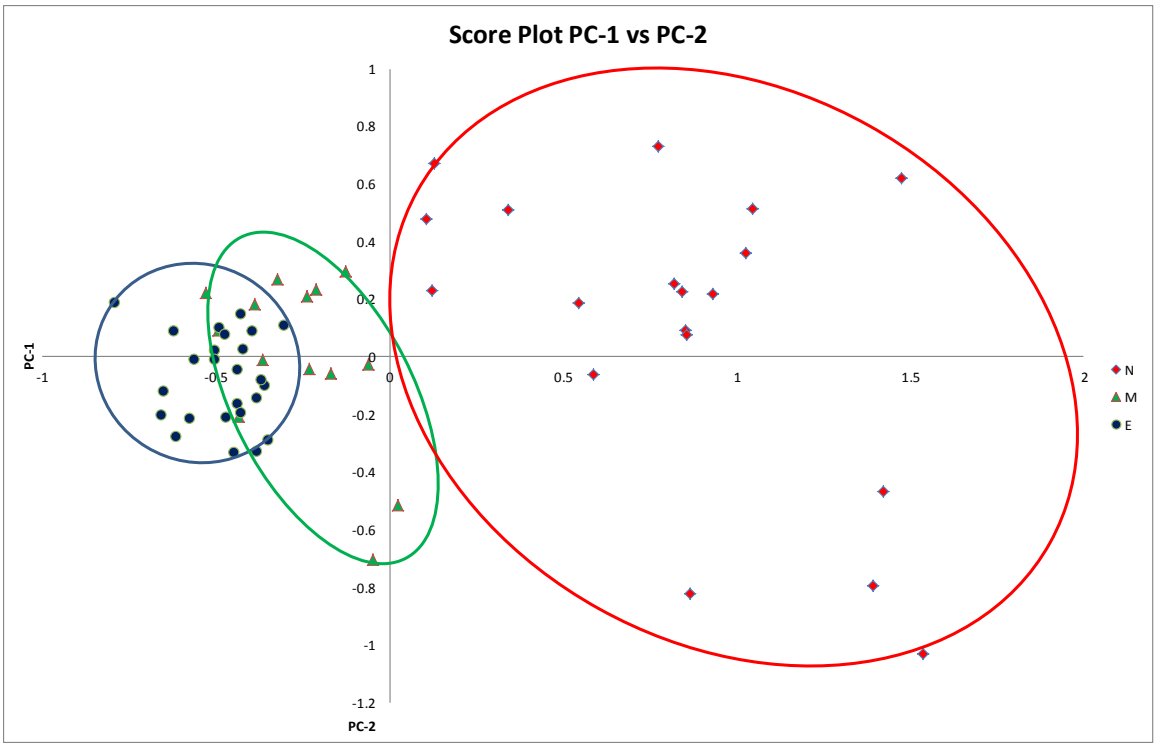
a. 16-metric model



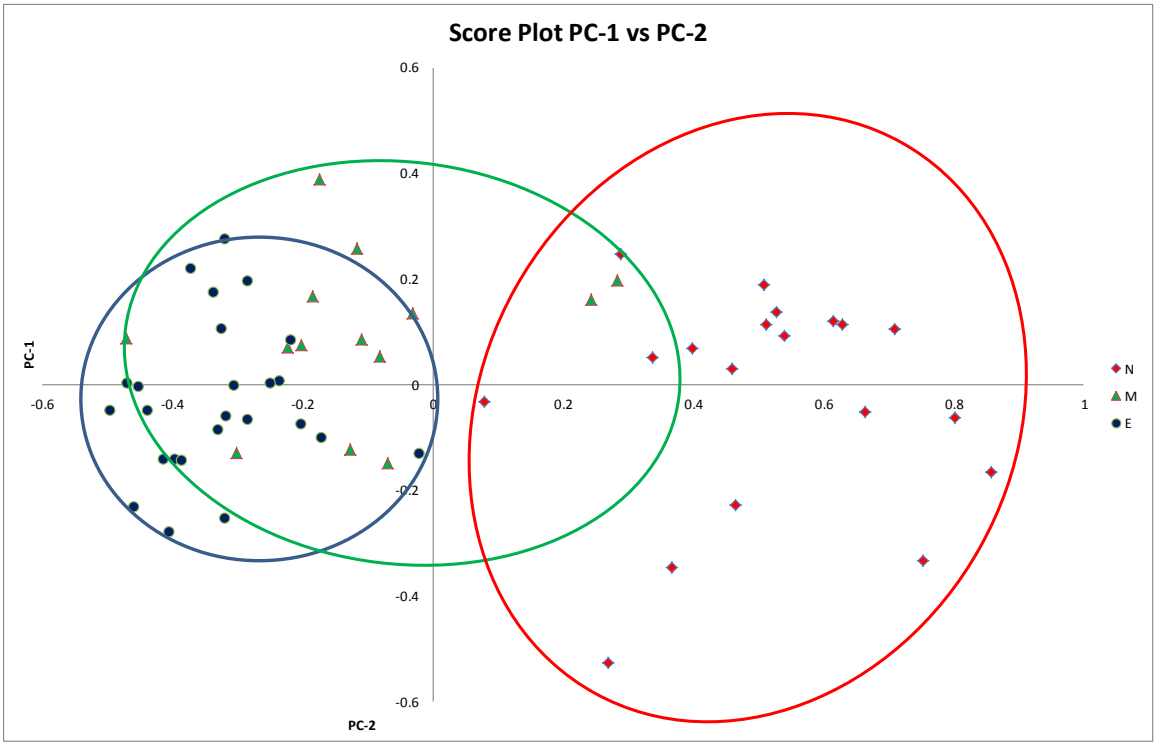
b. 3-metric model

Figure 6.10 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 10% noise applied to all metrics

Applying noise from 11%–15% shows gradual increase in the effect levels of both models. But the 16-metric model shows better tolerance to the noise than the 3-metric model. The increase in the level of sparse in the 16-metric model is less than it is in the 3-metric model as Figure 6.11–Figure 6.15(a) and (b) show. Also, the increase in the overlap rate between the intermediate and expert levels is less in the 16-metric model than it is in the 3-metric model. Both models tolerate the noise and the effect was minimal on the disparity between the intermediate level and the novice level. However, it is clear that the 3-metric model in Figure 6.14(b) where 14% noise is applied and Figure 6.15 (b) where 15% noise is applied, mix the intermediate and expert levels but the 16-metric model shows a kind of disparity between both levels.

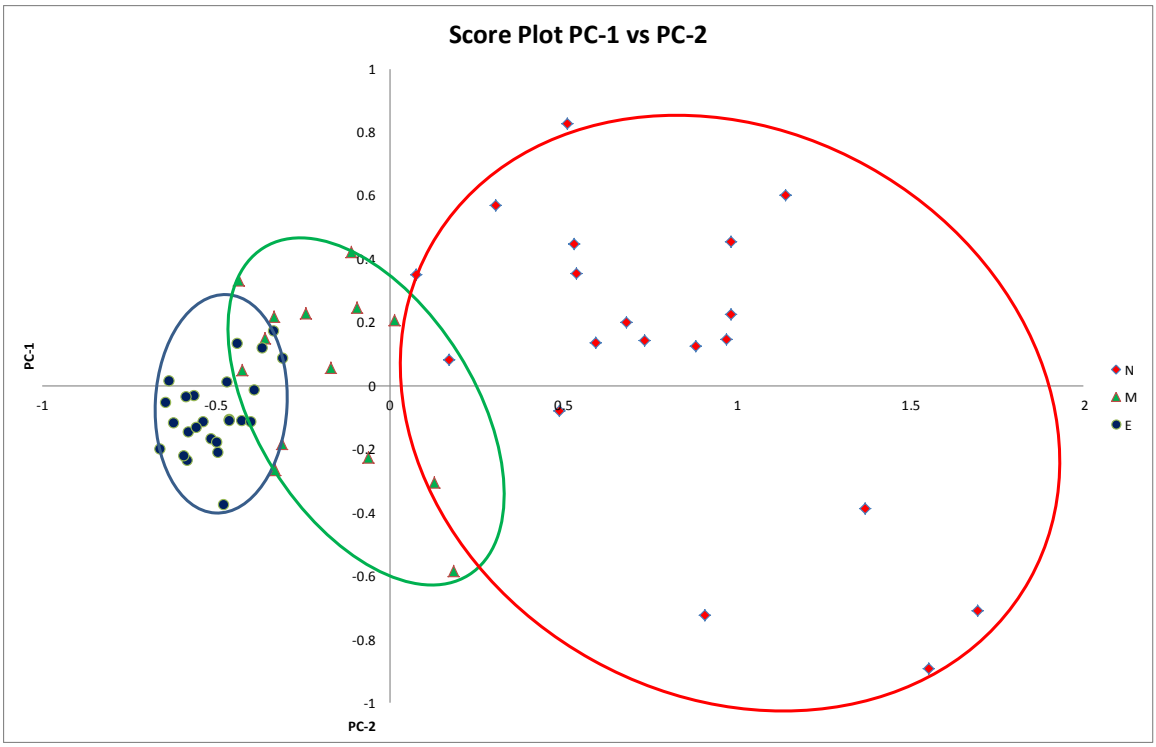


a. 16-metric model

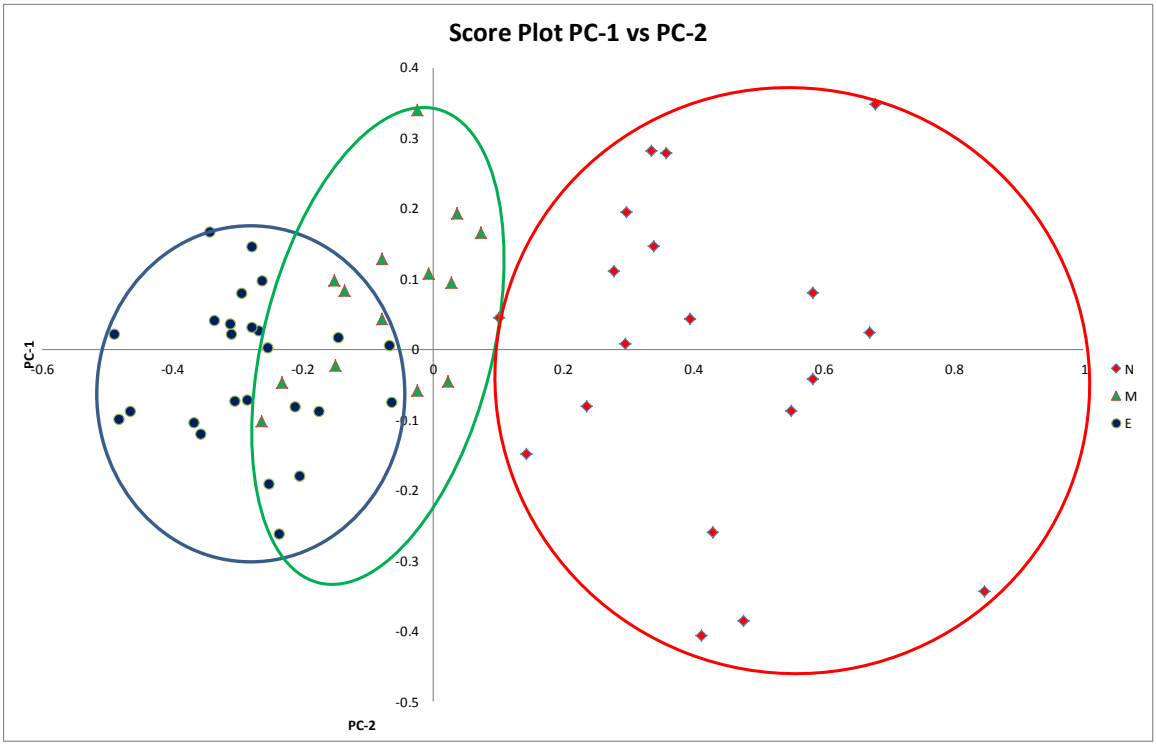


b. 3-metric model

Figure 6.11 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 11% noise applied to all metrics

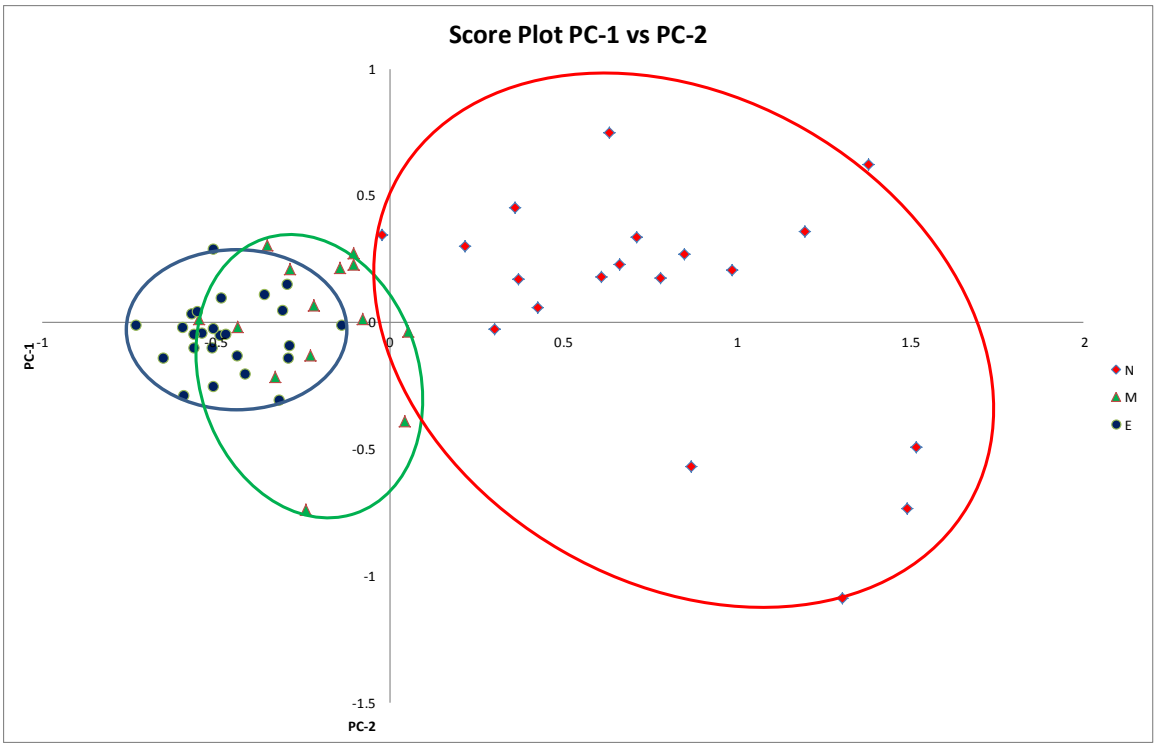


a. 16-metric model

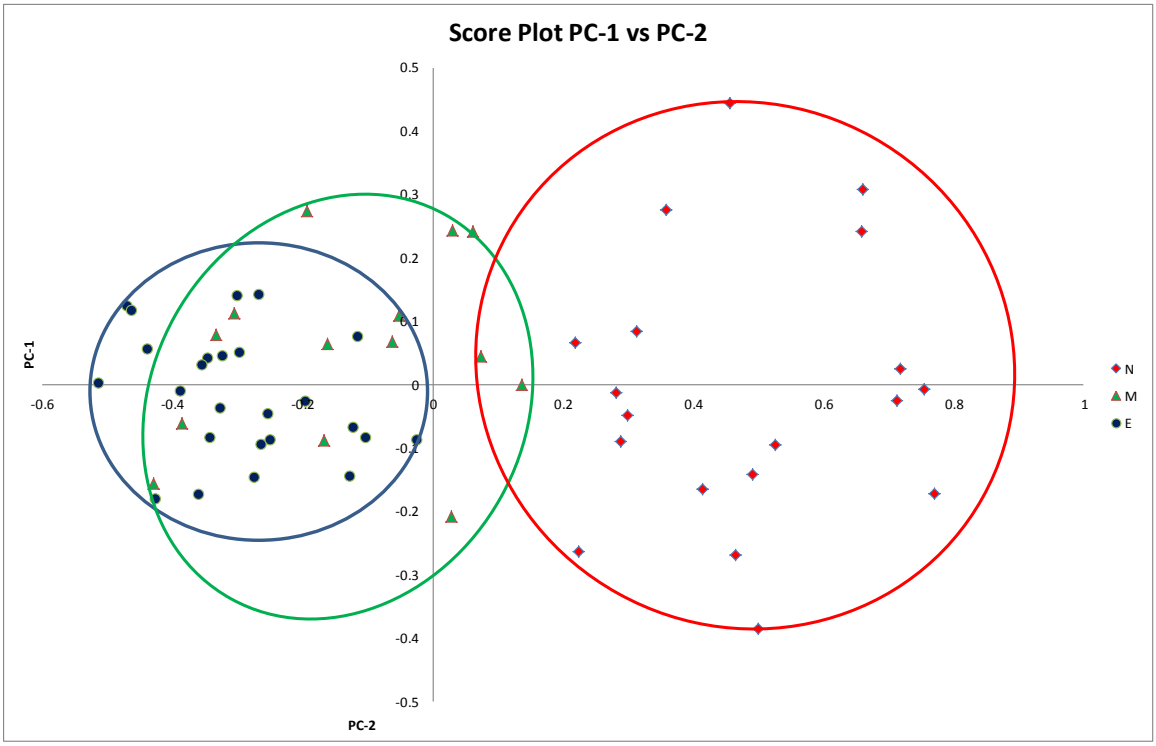


b. 3-metric model

Figure 6.12 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 12% noise applied to all metrics

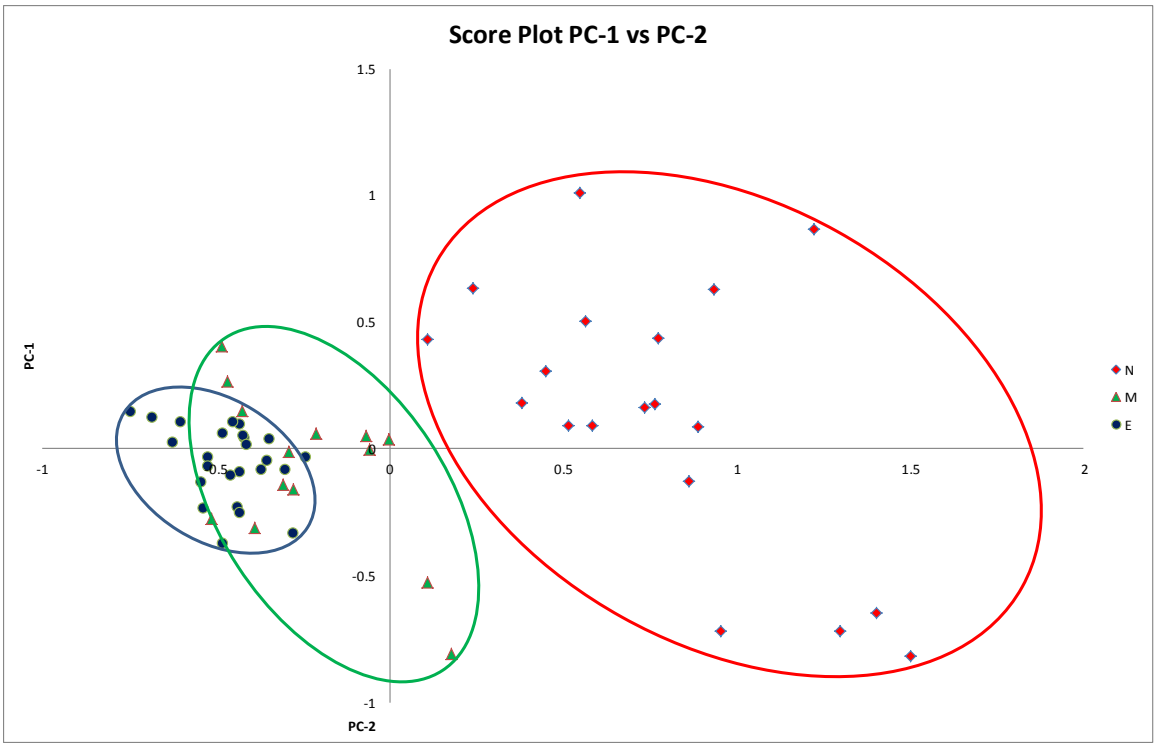


a. 16-metric model

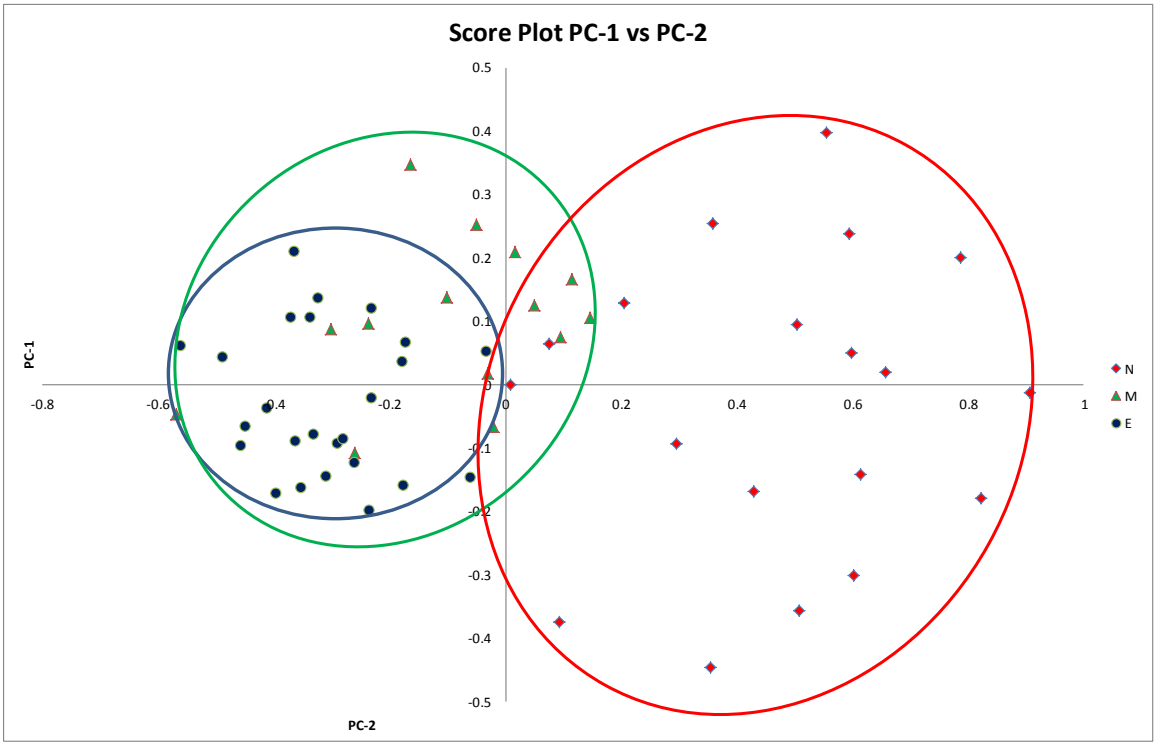


b. 3-metric model

Figure 6.13 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 13% noise applied to all metrics

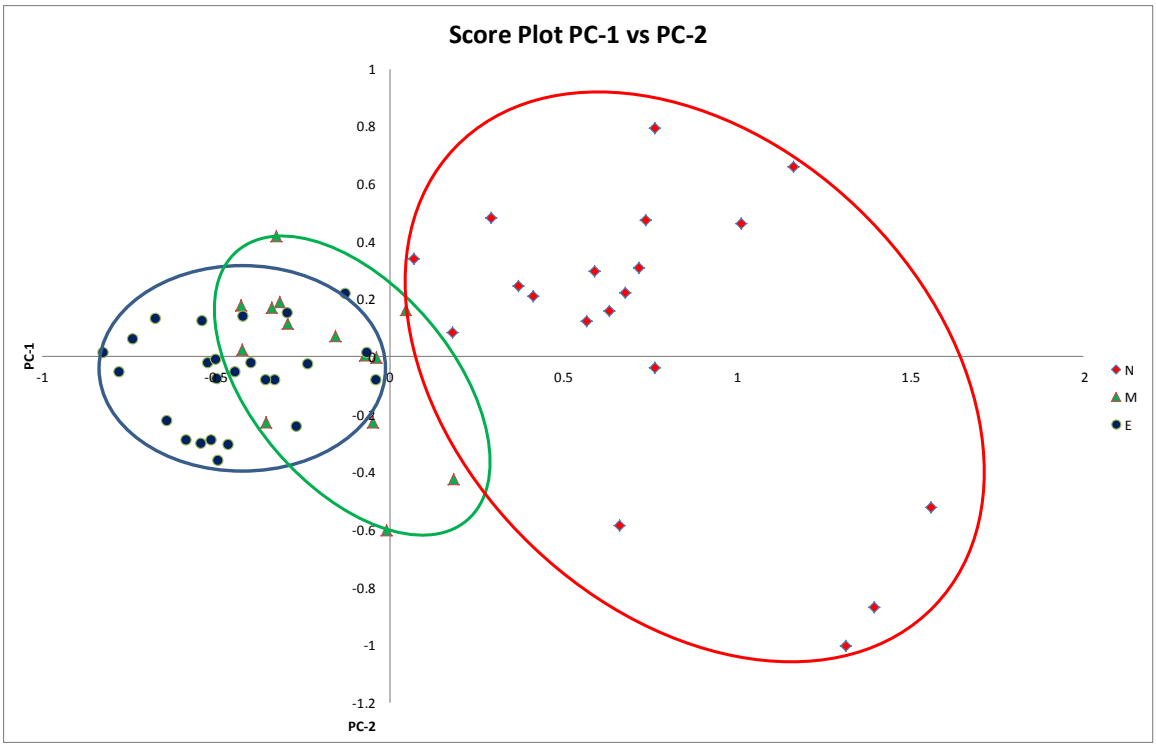


a. 16-metric model

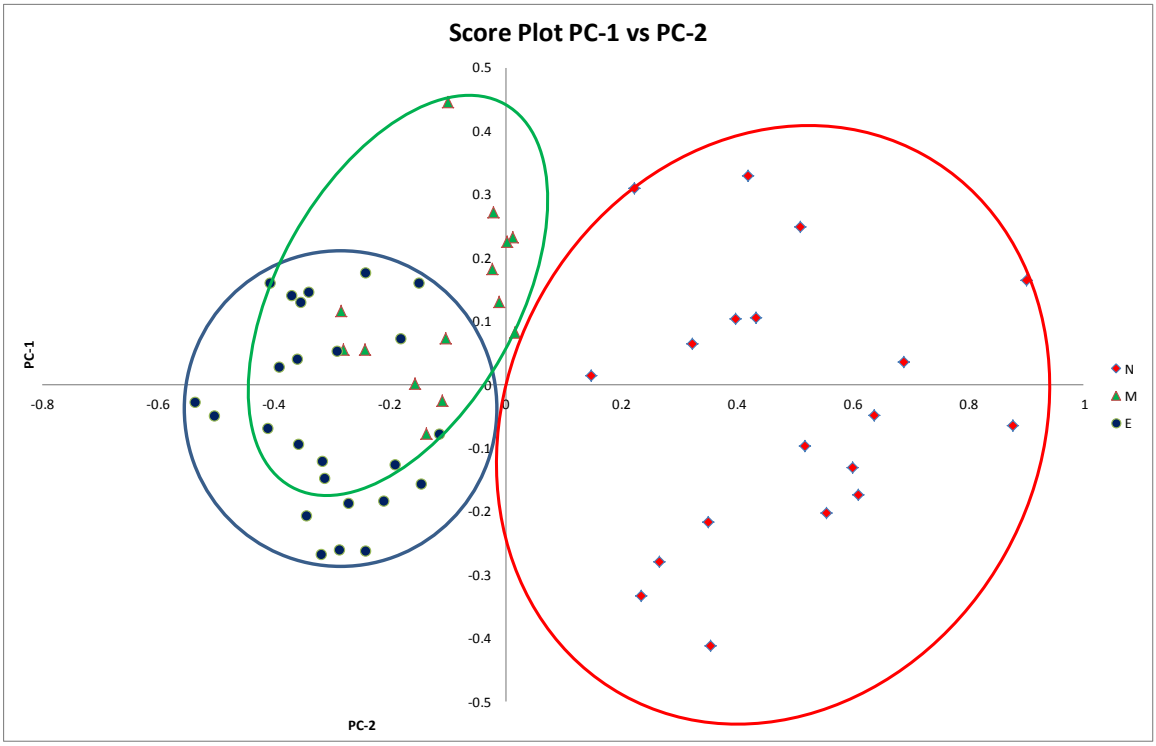


b. 3-metric model

Figure 6.14 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 14% noise applied to all metrics



a. 16-metric model

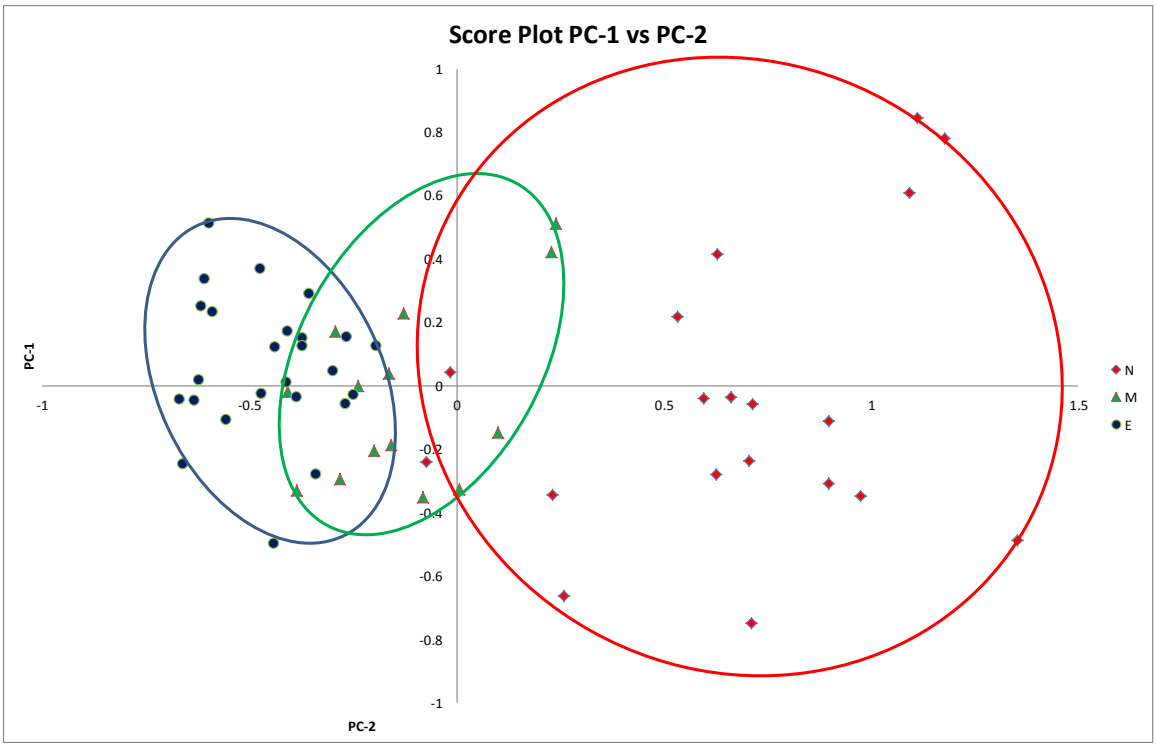


b. 3-metric model

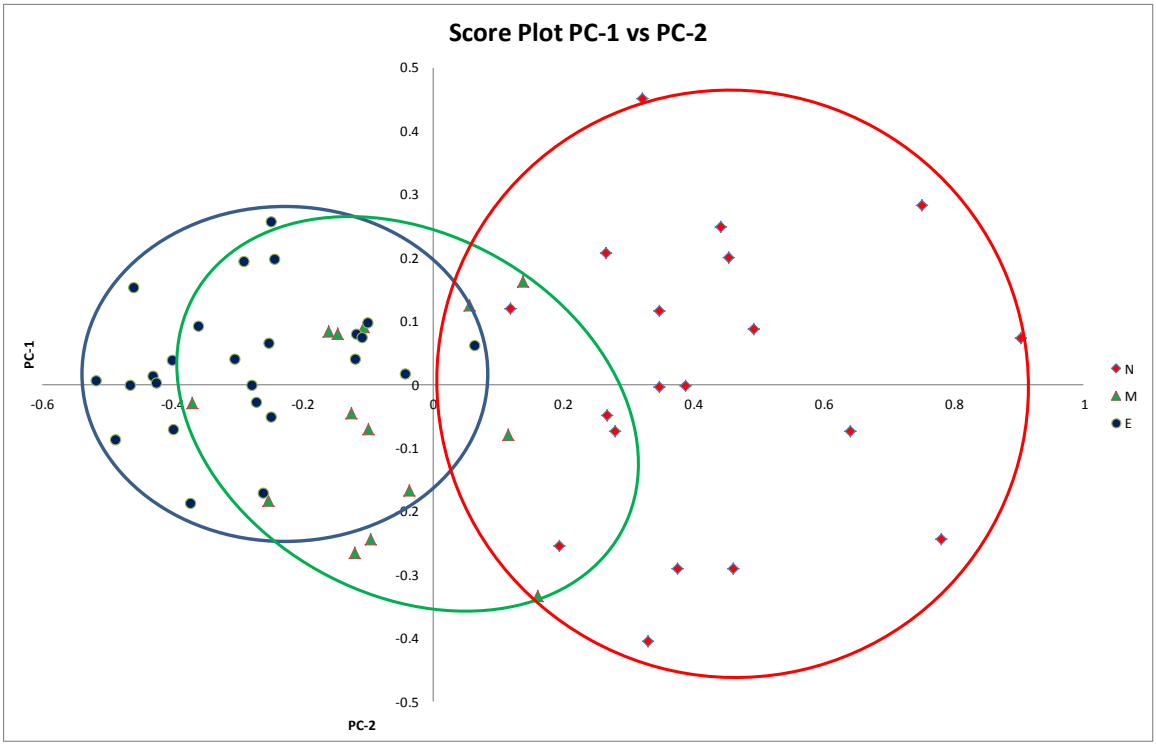
Figure 6.15 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 15% noise applied to all metrics

The trend of the effect of increasing the noise magnitude to 17% and 20% continues and in the 3-metric model, as Figure 6.16(b) 17% and Figure 6.17(b) 20% show, the overlap between the novice level and other experience levels starts occurring. At level 20% the subjects at the three experience levels are mixed and the potential error assessing the subject is high. In the 16-metrtic model, there is still a clear disparity between the novice and the other levels.

This discussion of the noise effect shows the robustness of the 16-metric model over the 3-metric model. This result proves the significance of using more metrics that correlate to the skill level in improving the reliability of the assessment as the metrics mutually support each other.

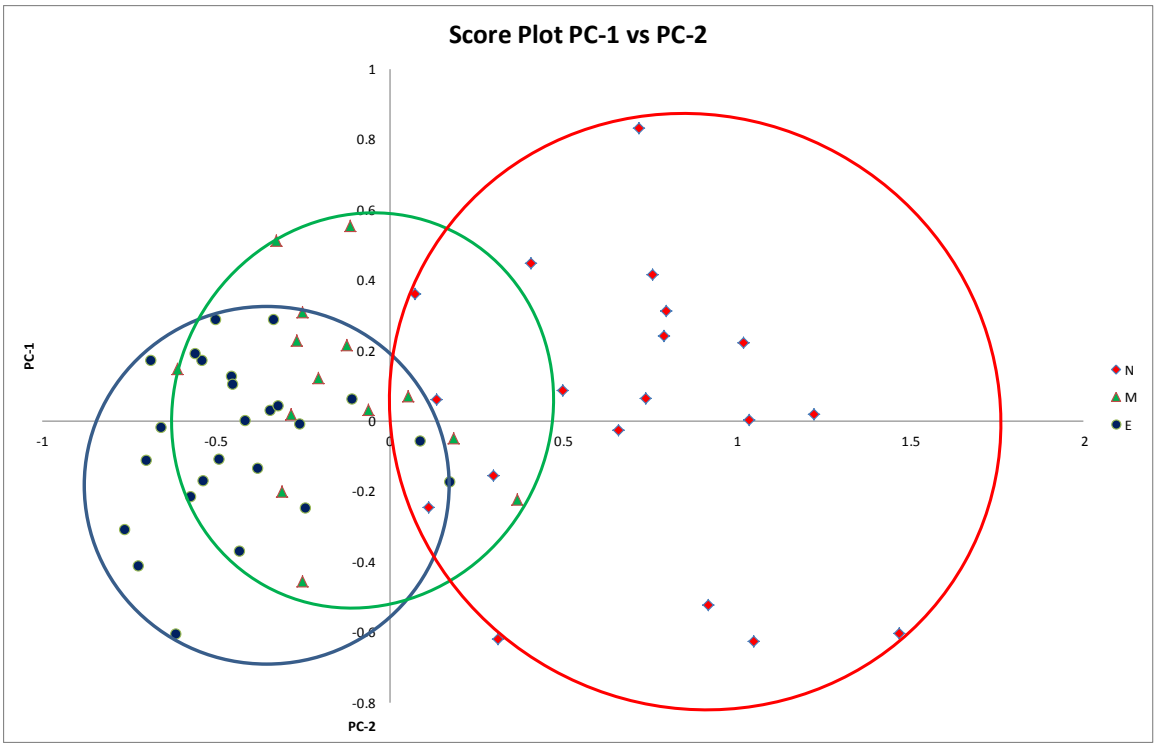


a. 16-metric model

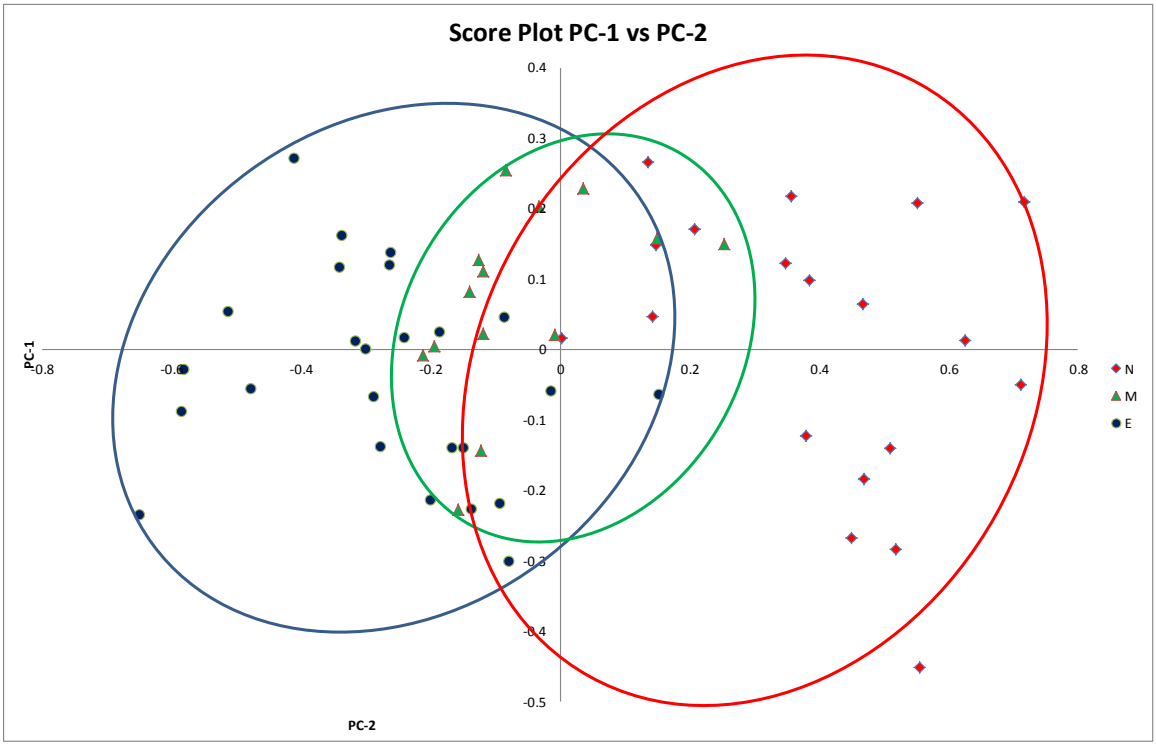


b. 3-metric model

Figure 6.16 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 17% noise applied to all metrics



a. 16-metric model

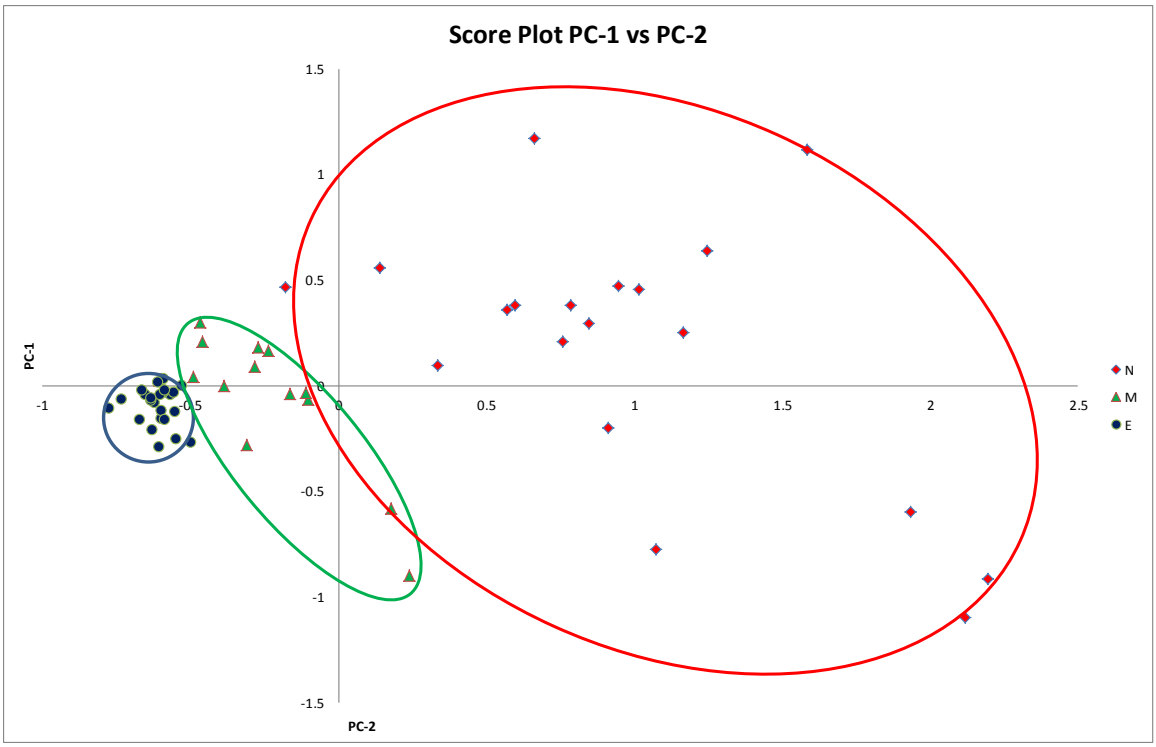


b. 3-metric model

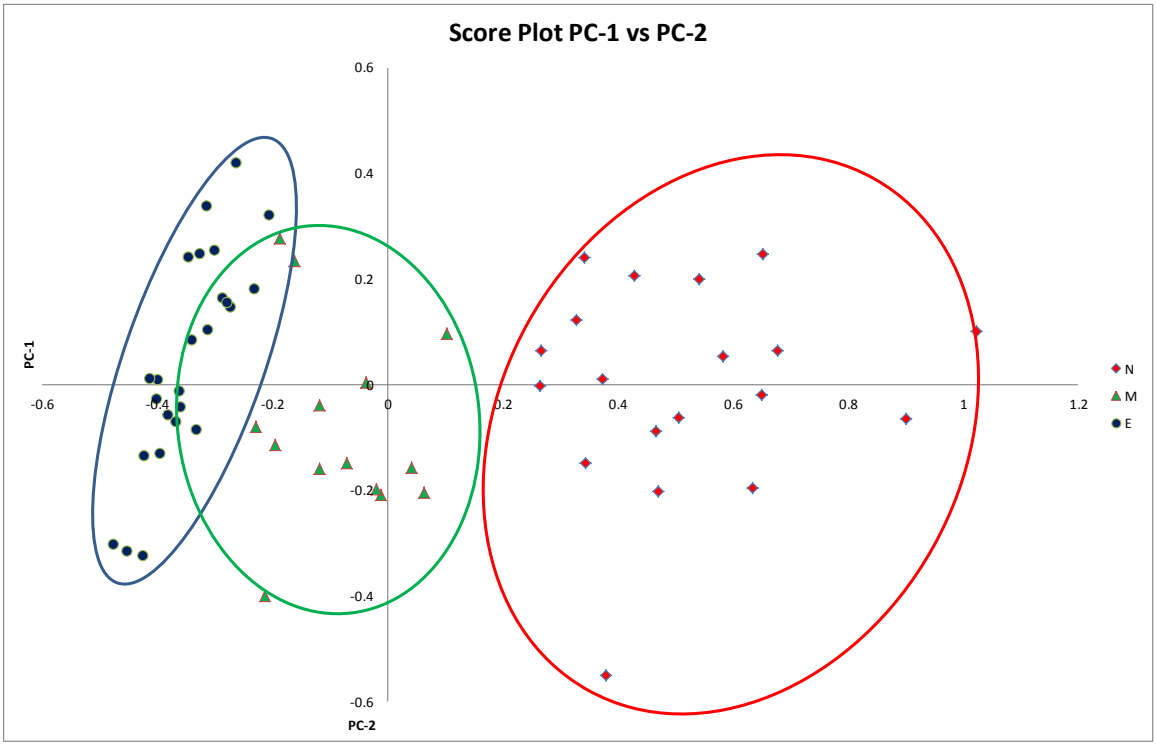
Figure 6.17 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 20% noise applied to all metrics

The second robustness experiment we performed is the application of a large magnitude of noise to one of the variables in both data models. The goal of this experiment is to find the effect of the variables' mutual support on the reliability of the assessment models built. The variable we picked is r_direction_change which has the highest level of correlation to the experience level. Two magnitudes of noise are applied to the 3-metric model and 16-metric model. The magnitudes of noise we added were 50% and 90% which could simulate a total data corruption to one of the variables.

The 16-metric model showed high level of tolerance to both levels of noise while the 3-metric model is largely affected, especially the intermediate and expert levels. As Figure 6.18 (a) 50% and Figure 6.19(a) 90% show the tolerance of the 16-metric model, the 50% and 90% noise effect is marginal on all experience levels which indicates the level of mutual support of the metrics in the model. Figure 6.18(b) and Figure 6.19(b) show how much the 3-metric model is affected by both levels of noise. The expert level is severely affected and the blue points are scattered, which means the similarity among the expert subjects are affected and their performance is dissimilar. That indicates the mutual support among the metrics of this model is less which reduces the level of tolerance to error and noise. An interesting aspect of the 3-metric model is that the effect level is minimal when noise is increased from 50% to 90%. The difference in Figure 6.18(b) and Figure 6.19(b) is marginal. This difference indicates that increasing the noise on the same variable to a higher level may not affect the model, but applying large level of noise to one more variable will lead to failure in the model. Similarly, the difference between Figure 6.18(a) and 6.18 (b) is almost invisible. That means the 16-metric model tolerates any level of noise on one variable and a large level of noise on more than one variable whereas the 3-metric model could fail when a large level of noise is added to one more variable.

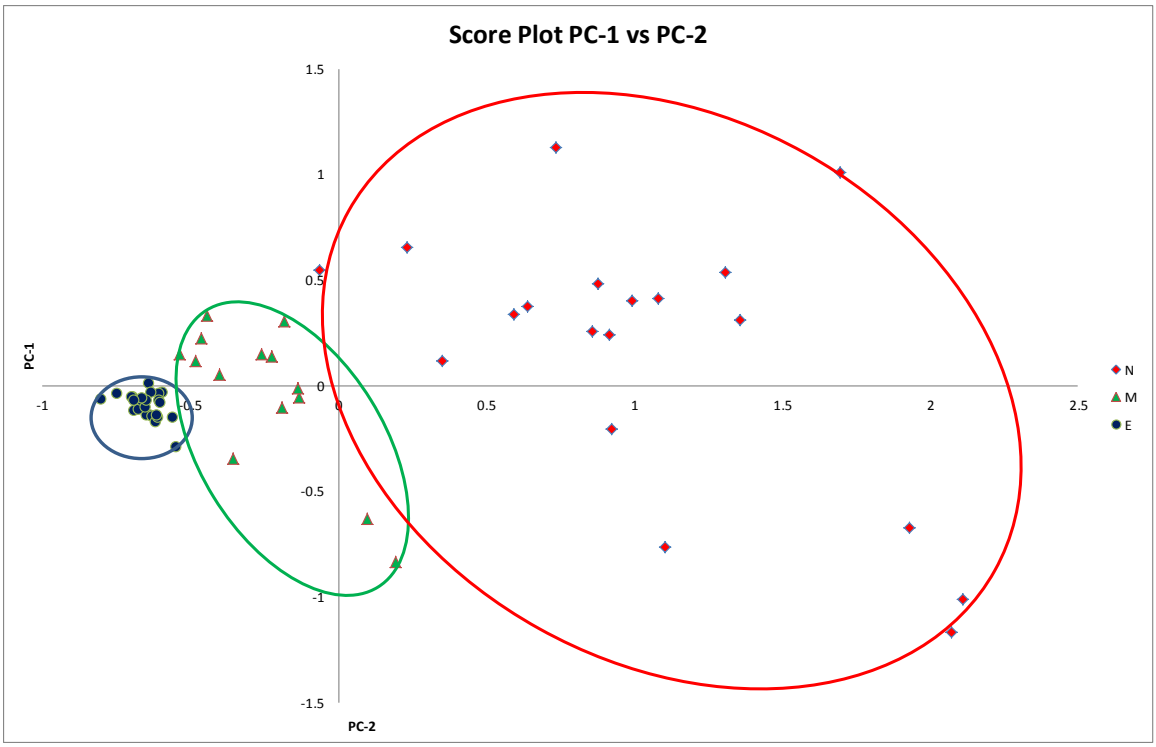


a. 16-metric model

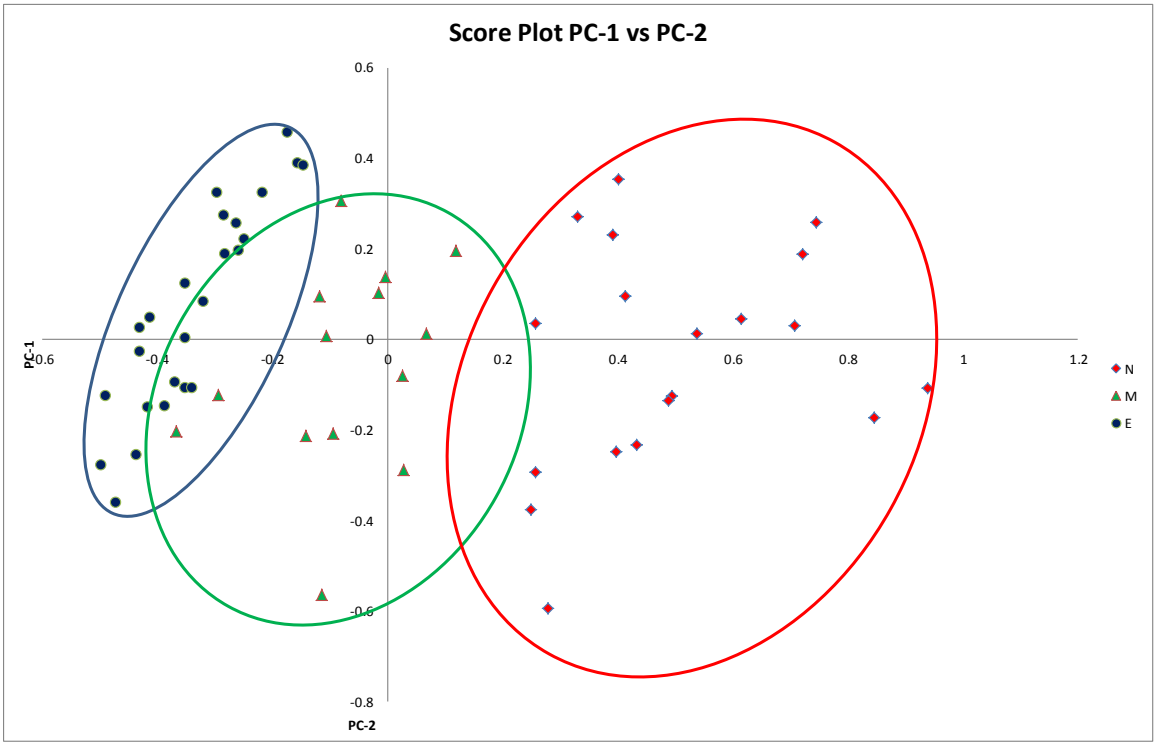


b. 3-metric model

Figure 6.18 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 50% noise applied to one metric



a. 16-metric model



b. 3-metric model

Figure 6.19 Score plot of PC-1 vs. PC-2 for the 16-metric and 3-metric models with 90% noise applied to one metric

6.4 PCA Validation with Real Data

After finishing the model and all the analysis, we decided to capture data for two more subjects at all experience levels to study the validity of the model using real data and compare the newly added data to the previous subjects. Two sets of data were taken for each subject for each skill level for a total of 12 trials. Figure 6.20 shows the score plot for both old and new sets of data. The data for the new subjects are marked differently in the plot. The legends N_1, N_2, M_1, M_2, E_1, and E_2 represent the data for the first and the second subjects at the novice, intermediate, and expert levels. As we see, the new subjects reside in their proper clusters and they converge from novice to expert as their skills increase. In addition, the features of the new data comply with the features of the old data at all levels.

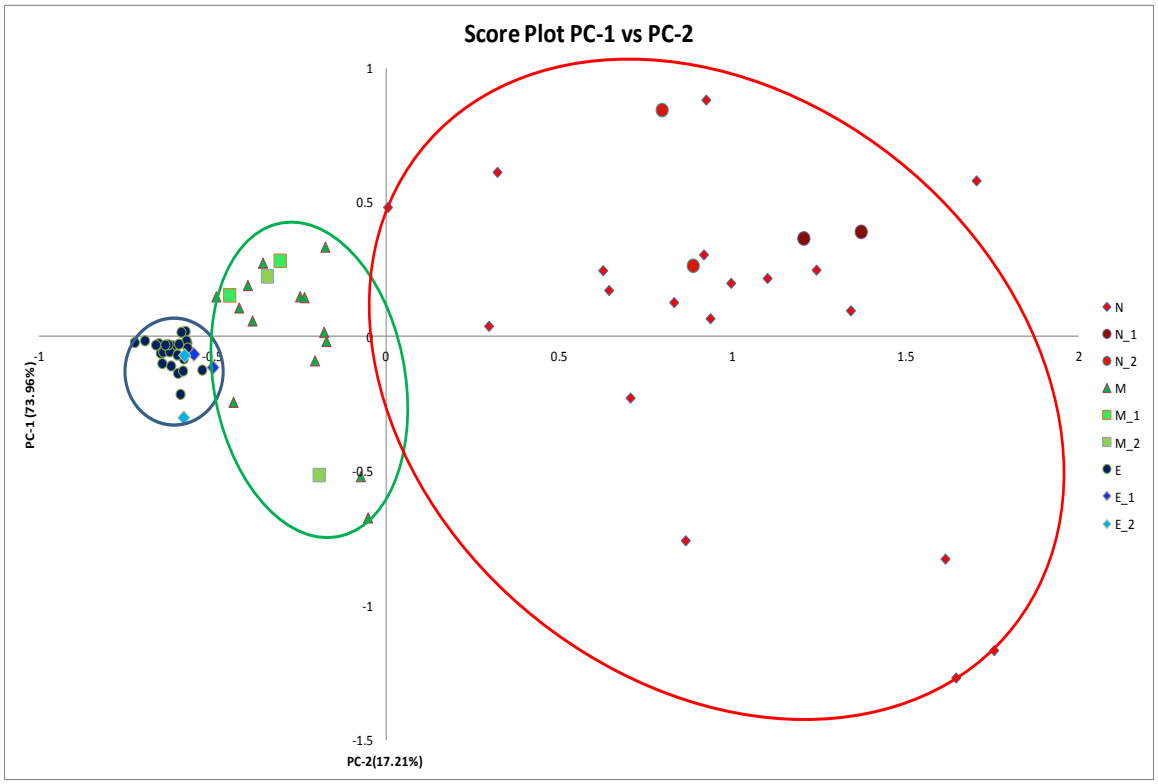


Figure 6.20 PCA score plot. PC-1 vs. PC-2 to validate data captured for two subjects in all experience stages

For more illustration, we found the centroid point of each cluster and plotted it in Figure 6.21 with the data of the new subjects. The figure shows the relationship between the centroid of the cluster and each subject. The Manhattan distance and the Euclidean distance are calculated between the centroid of each cluster and each of the new points. The result of this calculation is reported in Table 6.6 and Table 6.7. As both Tables show, the novice data is closer to the novice centroid, the intermediate data is closer to the intermediate centroid, while the expert data is closer to the expert centroid.

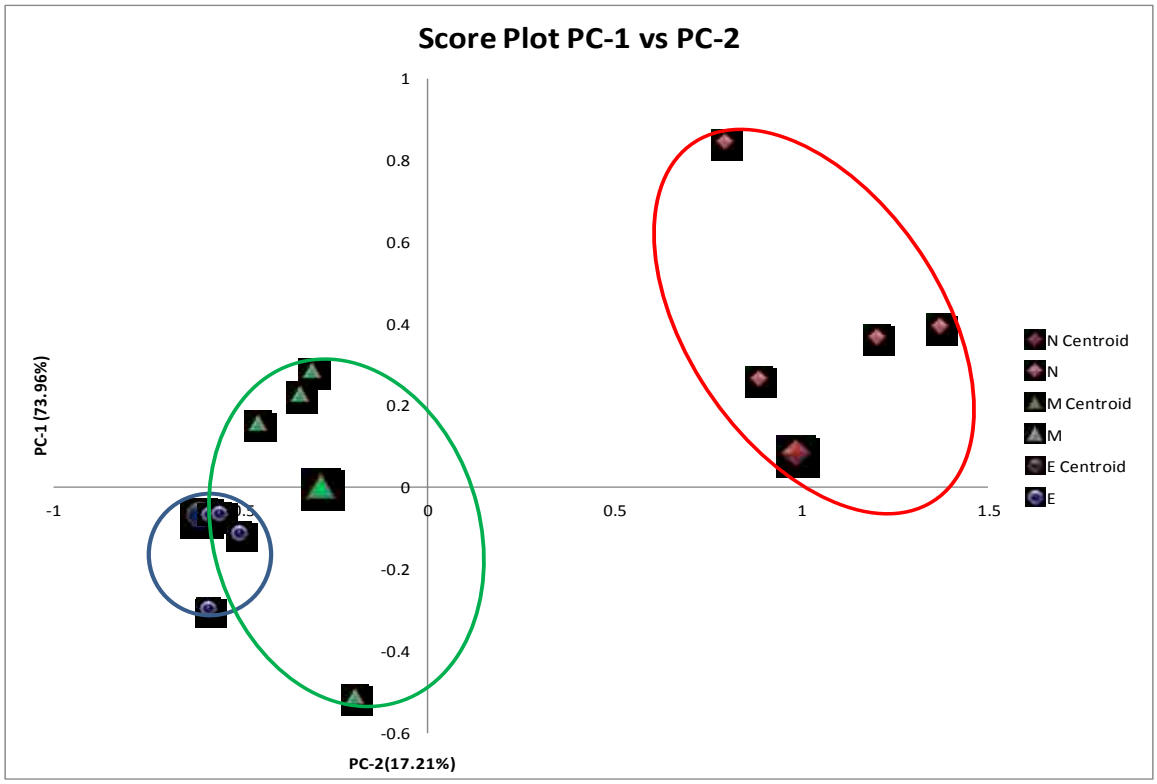


Figure 6.21 PCA score plot for the centroid of each skill level and the new captured data

Table 6.6 Manhattan distance between the centroid of each cluster and individual data for each subject

	Manhattan Distance To N Centroid	Manhattan Distance To M Centroid	Manhattan Distance To E Centroid	Minimum Distance	Centroid
N_2_1	**0.284401**	1.434207	1.826059	**0.284401**	**N**
N_2_2	**0.953724**	1.923943	2.315796	**0.953724**	**N**
N_1_1	**0.691575**	2.045483	2.437335	**0.691575**	**N**
N_1_2	**0.50122**	1.855128	2.246981	**0.50122**	**N**
M_1_1	1.495842	**0.302626**	0.651835	**0.302626**	**M**
M_1_2	1.512537	**0.319321**	0.377511	**0.319321**	**M**
M_2_1	1.474165	**0.280949**	0.557326	**0.280949**	**M**
M_2_2	1.778449	**0.607556**	0.85841	**0.607556**	**M**
E_2_1	1.721613	0.367705	**0.025045**	**0.025045**	**E**
E_2_2	1.955673	0.601765	**0.254141**	**0.254141**	**E**
E_1_1	1.684239	0.330331	**0.152157**	**0.152157**	**E**
E_1_2	1.690555	0.336647	**0.055206**	**0.055206**	**E**

Table 6.7 Euclidean distance between the centroid of each cluster and individual data for each subject

	Euclidean Distance To N	Euclidean Distance To M	Euclidean Distance To E	Minimum Distance	
N_2_1	**0.208964**	1.200603	1.529596	**0.208964**	**N**
N_2_2	**0.785662**	1.370933	1.67382	**0.785662**	**N**
N_1_1	**0.491823**	1.701516	2.030482	**0.491823**	**N**
N_1_2	**0.357559**	1.534618	1.863536	**0.357559**	**N**
M_1_1	1.310385	**0.282111**	0.462368	**0.282111**	**M**
M_1_2	1.442199	**0.226022**	0.271296	**0.226022**	**M**
M_2_1	1.338944	**0.230558**	0.394663	**0.230558**	**M**
M_2_2	1.323986	**0.524099**	0.607428	**0.524099**	**M**
E_2_1	1.577591	0.30512	**0.0246**	**0.0246**	**E**
E_2_2	1.618732	0.425518	**0.233078**	**0.233078**	**E**
E_1_1	1.50095	0.243782	**0.116053**	**0.116053**	**E**
E_1_2	1.550452	0.278028	**0.051594**	**0.051594**	**E**

6.5 Cluster Analysis

The data in this study represents human skills, therefore, getting high accuracy in a clustering analysis is challenging for several reasons. An individual's pace of learning is different from another. There is no specific quantitative threshold to draw a line to divide people based on their skill levels. There is some overlap and a transitioning period between one level and another. In addition, the distribution of the data variance as we have seen in the PCA analysis increases the difficulty of the clustering analysis. Using a partitioning clustering algorithm could cluster the data, but it might falsely cluster subjects in the overlap areas between skill levels, especially between the intermediate and expert. The novice data is largely scattered and has no specific pattern which could be a challenge to the partitioning algorithms. Further, the small distance between the expert and intermediate data and the paucity of the novice data could be a challenge to density-based clustering.

To overcome these challenges we decided to use a hybrid algorithm of partitioning and density-based clustering. The Waikato Environment for Knowledge Analysis (Weka) is an open source data mining tool that provides an interface to a clustering algorithm that wraps kmeans in a density based algorithm [70]. The algorithm is called MakeDensityBasedClusterer which uses kmeans output as a seed to perform density based clustering. The algorithm initially uses kmeans to construct the clusters based on the distance from the centroid. Then MakeDensityBasedClusterer reconstructs the clusters based on the density using normal distribution.

Applying the algorithm on the collected data produced clusters summarized in Table 6.8. The result shows that one subject was falsely put in a different cluster than what the manual assessment undertook. When we looked up the code for that subject we found that it was I_N_3. The code maintains that the data is the third trial as novice for this subject. The third trial is taken in the third session of training, in which the subject is expected to transition from novice to intermediate. The mis-clustered subject is marked

by the blue arrow in Figure **6.22** (PC-1 vs. PC-2) and Figure **6.23** (PC-1 vs. PC-3). The graphs show the score plots presented at the beginning of this chapter. The mis-clustered subject is close to the intermediate subjects and has features similar to them. Figure 6.24 shows the cluster distribution among the data and the mis-clustered subjects. The color legend represents the ground truth and the cluster ID represents the generated clusters.

Table 6.8 The truth and result clusters based on the 16-metric model

Skill Level	Truth Clusters	Resulted Clusters	Number of mis-clustered subjects	Error Rate
Novice	19 (32.76%)	18 (31.04%)	1	1.72%
Intermediate	14 (24.14%)	15 (25.86%)		
Expert	25 (43.1 %)	25 (43.1)		
Total	58 (100%)	58 (100%)		

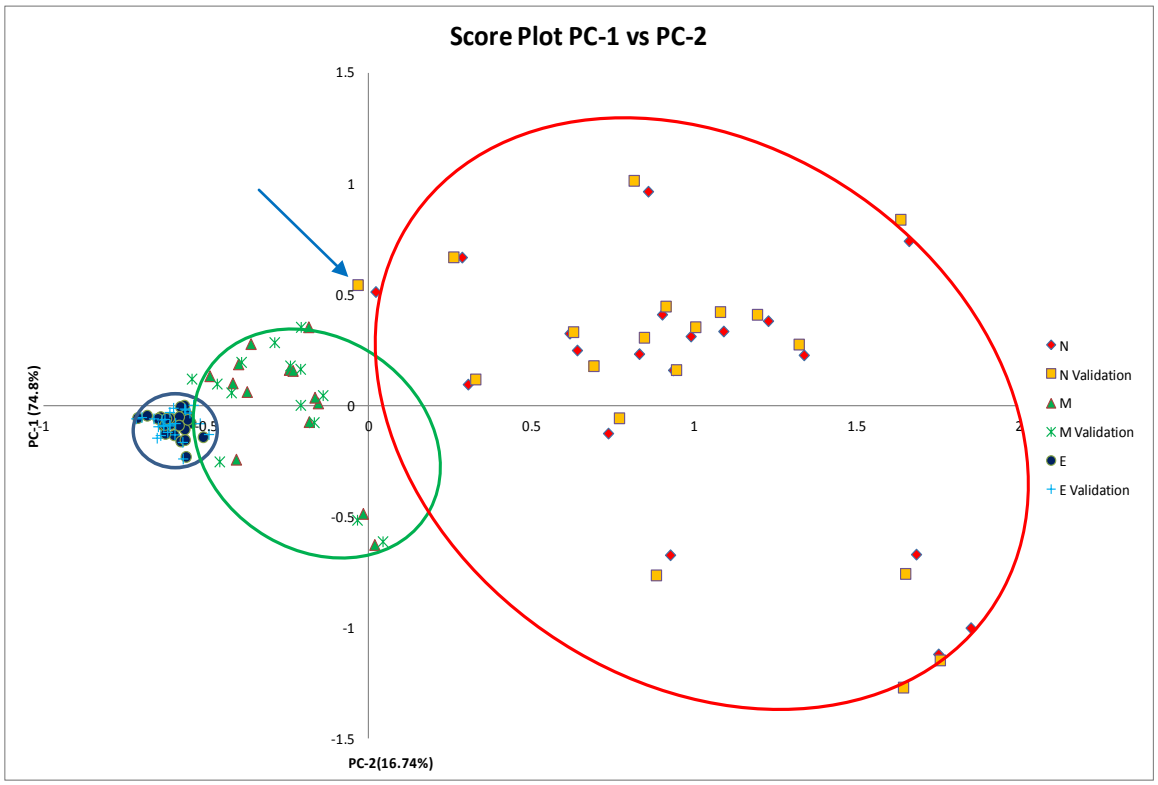


Figure 6.22 PCA score plot of PC-1 vs. PC-2. The mis-clustered subject is marked by the blue arrow

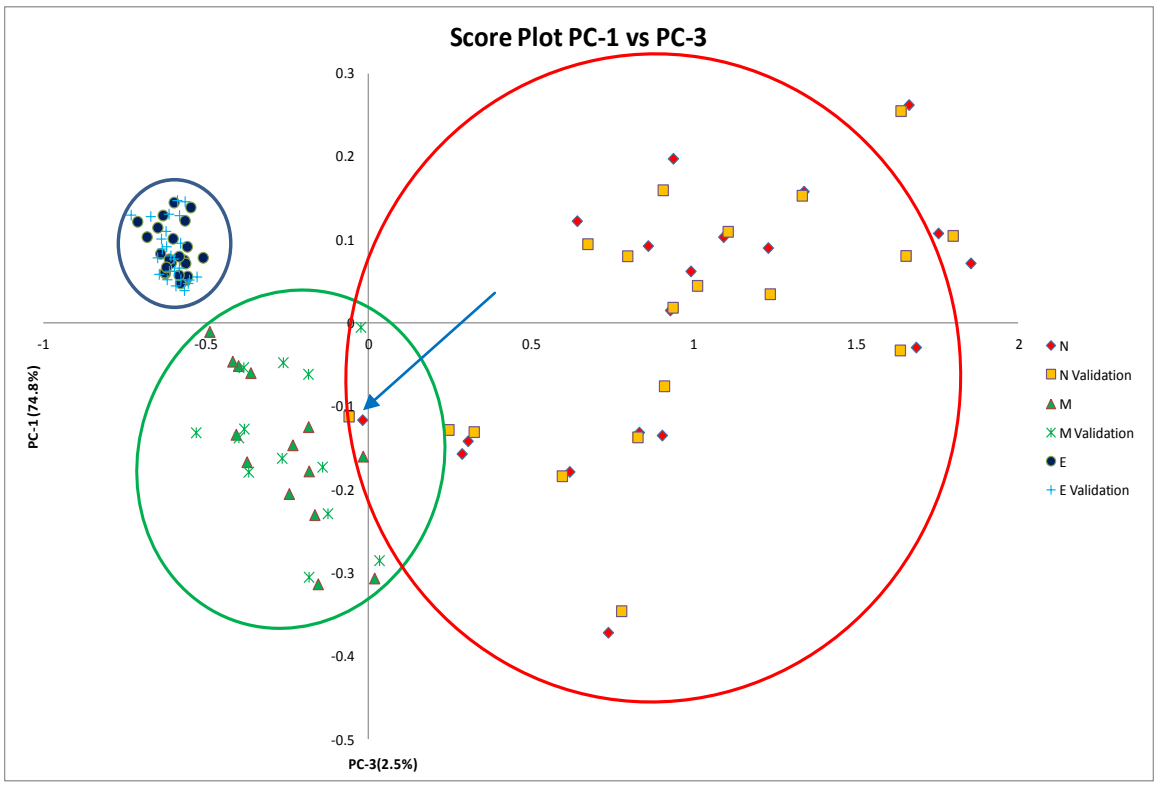


Figure 6.23 PCA score plot of PC-1 vs. PC-3. The mis-clustered subject is marked by the blue arrow

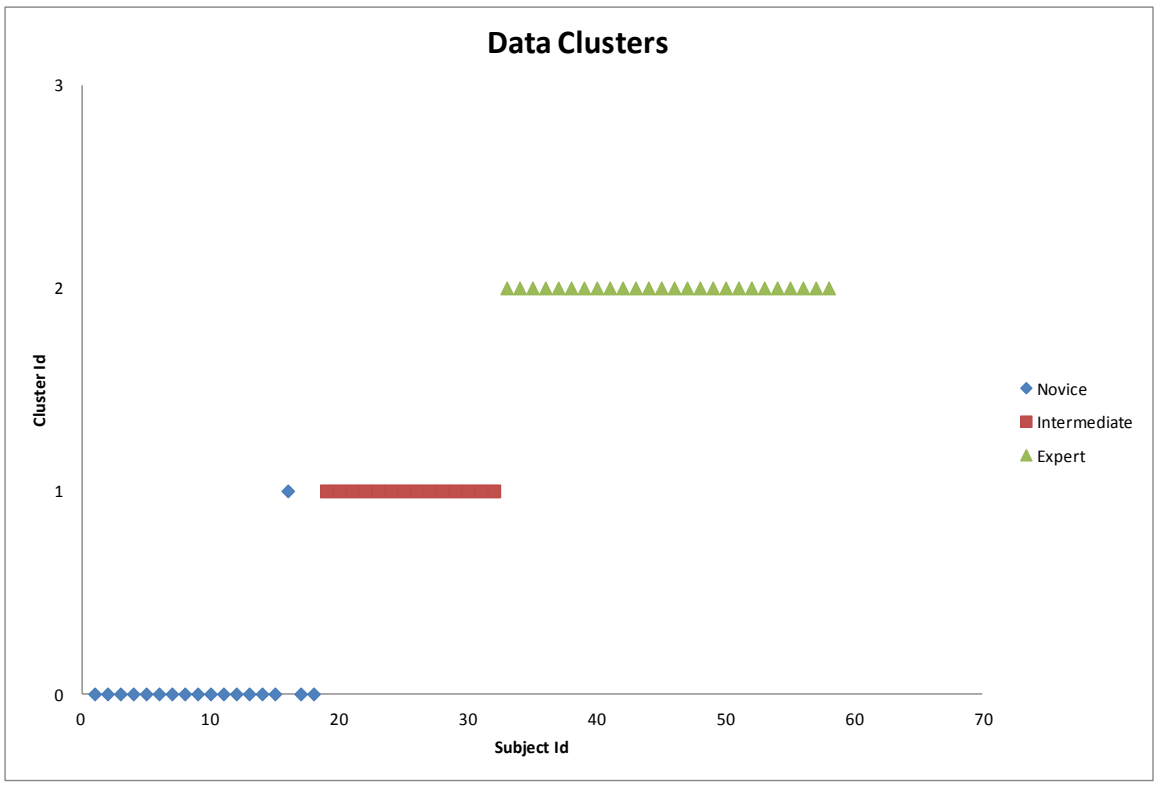


Figure 6.24 Data clusters. The color legend represents the ground truth and the cluster ID represents the generated clusters.

To study the clustering accuracy using various numbers of metrics sorted based on their absolute correlation coefficient, we repeated the clustering analysis by removing metrics from the lowest correlation coefficient to the highest. All measured metrics are included in this experiment. The error rate curve of clustering is shown in Figure 6.25. The graph shows that using the fewest metrics with the highest correlation does not necessarily give the highest accuracy. The error rate is above 13% using the highest two attributes. The rate decreases until it reaches 1.72% between 7-16 metrics which are listed in Table 6.1. The error rate then starts to increase until it reaches 55.17%. This result shows the importance of developing an algorithm that can extract the correct metrics to assess MIS skills objectively and reliably.

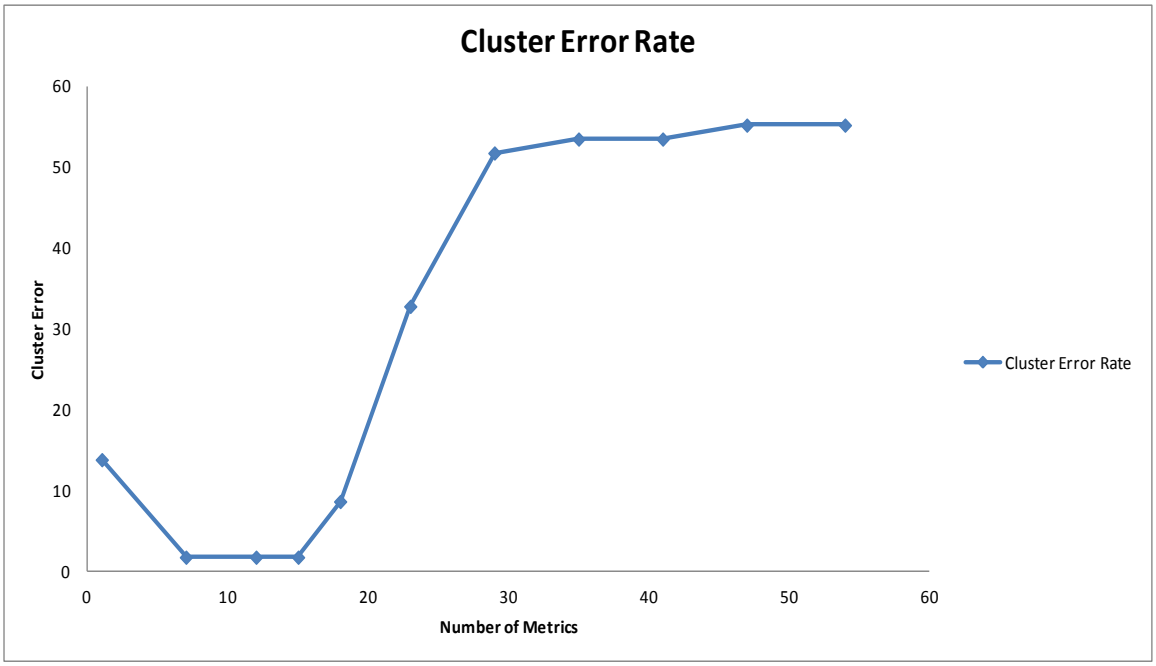


Figure 6.25 Error rate curve of mis-clustering based on the number of attributes used in the experiment.

6.6 Classification

The collected dataset in the case study is probably small to train many kinds of classifiers. If we train a model of tree classification such as C5, the dataset is not large enough to cover all cases. Therefore, many of the metrics will not affect the decision made by the tree. To validate the system as a classifier, we decided to use the Multi-Layer Perceptron (MLP) which is a neural network algorithm. The reason is that all the input data will contribute to building the model; even the dataset is small. Since the size of the dataset is not large enough to reach the recommended level to train an MLP, we built the model using one hidden layer and minimized the number of nodes in the hidden layer to prove the concept of the possibility of building a reliable skill level classifier. To learn more about the MLP algorithm see [71]. Figure 6.26 shows the MLP network built using 16 metrics input and 3 levels output. The MLP model is validated in three stages, first using a test set, second using 10-fold validation, and finally, using the

data set for the two subjects collected after the model is built. Using 10-fold means 90% of the data is used for training which is 52 data records and 10% for testing which are eight subjects.

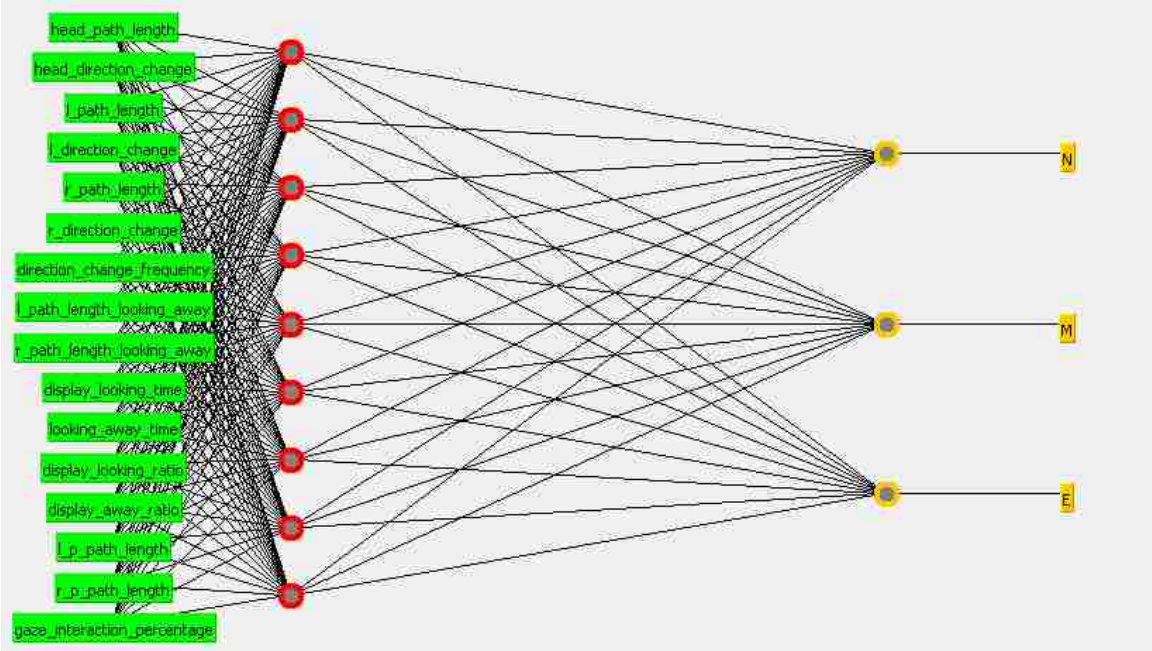


Figure 6.26 The MLP network built using 16 metrics input and 3 levels output

6.6.1 Classification Test Set Validation

The data is divided into two parts, training set and test set. Out of 58 subjects, 16 are allocated for testing.

Training data (42 subjects): N:13 , M:11 , E: 18

Test Data (16 subjects): N:6, M:3, E:7

The model built using the 16 metrics listed in Table 6.1 was able to correctly classify the 16 subjects in the test set with error rate = 0%. None of the subjects was classified incorrectly. The confusion matrix of the classification result is given in Table 6.9. Table 6.10 shows the detailed result including the probability distribution for each class of each subject. As the table shows, all subjects are predicted correctly with high probability. In predicting novice and expert classes, the probability exceeds the 99% except for subject number six which has 78%. When predicting a novice subject, the probability of that subject being an expert is 0%. Similarly when predicting an expert, the probability of the subject being novice is 0%. When predicting the intermediate subjects, the probability is less than the other classes. Further, in the intermediate class, there is a marginal probability of the subject to be in expert or novice classes. This observation explains the features of the intermediate progressing from the novice to the expert level.

Table 6.9 The confusion matrix of MLP classification model built using metrics on the test set

	N	**M**	**E**
N	6	0	0
M	0	3	0
E	0	0	7

Table 6.10 The test set classification results

Subject Number	Actual	Predicted	Error	Distribution		
				N	M	E
1	1:N	1:N		0.998	0.002	0
2	1:N	1:N		0.998	0.002	0
3	1:N	1:N		0.998	0.002	0
4	1:N	1:N		0.997	0.003	0
5	1:N	1:N		0.995	0.005	0
6	1:N	1:N		0.78	0.22	0
7	2:M	2:M		0.279	0.721	0
8	2:M	2:M		0.003	0.986	0.01
9	2:M	2:M		0.002	0.949	0.049
10	3:E	3:E		0	0.005	0.995
11	3:E	3:E		0	0.009	0.99
12	3:E	3:E		0	0.008	0.992
13	3:E	3:E		0	0.007	0.993
14	3:E	3:E		0	0.011	0.989
15	3:E	3:E		0	0.01	0.99
16	3:E	3:E		0	0.01	0.989

6.6.2 Classification 10-Fold validation

The confusion matrix in Table 6.11 shows the classification result of 10-fold classification. One intermediate subject was classified as novice with error rate 1.72% and accuracy rate 98.27%. As in the test set trial, the 10-fold validation result shows the probability of prediction novice and expert classes in most cases as being above 98% except for the mis-classified subject. This subject (marked by red shadow in Table 6.12) shows that the probability of being intermediate is 41% and the probability of being novice is 59%.

Table 6.11 The confusion matrix of MLP classification model built using 16-metric and 10-fold validation

	N	M	E
N	19	0	0
M	1	13	0
E	0	0	25

Table 6.12 10-fold validation classification results

Subject Number	Actual	Predicted	Error	Distribution		
				N	M	E
1	1:N	1:N		0.999	0.001	0
2	1:N	1:N		0.703	0.297	0
3	3:E	3:E		0	0.018	0.982
4	3:E	3:E		0	0.012	0.988
5	3:E	3:E		0	0.014	0.985
6	2:M	2:M		0.006	0.99	0.004
1	1:N	1:N		0.972	0.027	0.001
2	1:N	1:N		0.996	0	0.004
3	3:E	3:E		0	0.007	0.993
4	3:E	3:E		0	0.005	0.995
5	3:E	3:E		0	0.011	0.989
6	2:M	2:M		0.002	0.955	0.043
1	1:N	1:N		0.995	0.005	0
2	1:N	1:N		0.999	0.001	0
3	3:E	3:E		0	0.027	0.972
4	3:E	3:E		0	0.011	0.988
5	3:E	3:E		0	0.009	0.991
6	2:M	2:M		0.015	0.983	0.002
1	1:N	1:N		0.8	0.2	0
2	1:N	1:N		0.999	0.001	0
3	3:E	3:E		0	0.021	0.979
4	3:E	3:E		0	0.012	0.988
5	3:E	3:E		0	0.007	0.992
6	2:M	1:N	+	0.591	0.407	0.002
1	1:N	1:N		0.999	0.001	0
2	1:N	1:N		0.998	0.002	0
3	3:E	3:E		0	0.023	0.977

Table 6.12 (Continued)

4	3:E	3:E		0	0.006	0.994
5	2:M	2:M		0.001	0.758	0.241
6	2:M	2:M		0.001	0.506	0.494
1	1:N	1:N		0.991	0.009	0
2	1:N	1:N		0.998	0.002	0
3	3:E	3:E		0	0.013	0.987
4	3:E	3:E		0	0.01	0.99
5	2:M	2:M		0.038	0.961	0.001
6	2:M	2:M		0.069	0.93	0.001
1	1:N	1:N		0.999	0.001	0
2	1:N	1:N		0.999	0.001	0
3	3:E	3:E		0	0.008	0.992
4	3:E	3:E		0	0.007	0.993
5	2:M	2:M		0.029	0.97	0.001
6	2:M	2:M		0.006	0.985	0.009
1	1:N	1:N		0.998	0.002	0
2	1:N	1:N		0.998	0.002	0
3	3:E	3:E		0	0.009	0.991
4	3:E	3:E		0	0.007	0.993
5	2:M	2:M		0.01	0.987	0.003
6	2:M	2:M		0.023	0.976	0.001
1	1:N	1:N		0.999	0.001	0
2	1:N	1:N		0.996	0.004	0
3	3:E	3:E		0	0.006	0.994
4	3:E	3:E		0	0.036	0.964
5	2:M	2:M		0.005	0.988	0.008
1	1:N	1:N		0.999	0	0.001
2	3:E	3:E		0	0.005	0.995
3	3:E	3:E		0	0.007	0.993
4	3:E	3:E		0	0.106	0.894
5	2:M	2:M		0.002	0.959	0.039

6.6.3 Classifier in Implementation

After the model had been built and validated, we collected data for two subjects in three levels, two trials in each level. That meant that the total trials were 12 of which 4 were novices, 4 intermediates, and 4 experts. The model was then used to classify this set of data. The confusion matrix in Table 6.13 shows one trial of intermediate subjects was incorrectly classified as expert. The classifier shows in Table 6.14 that the probability of that subject being intermediate is 40% and being expert is 60%.

Table 6.13 The confusion matrix of MLP classification model built using 16-metric on the second test set

	N	M	E
N	4	0	0
M	0	3	1
E	0	0	4

Table 6.14 Classification results using the second test set

Subject Number	Actual	Predicted	Error	Distribution		
				N	M	E
1	1:N	1:N		1	0	0
2	1:N	1:N		1	0	0
3	2:M	2:M		0.013	0.985	0.003
4	2:M	3:E	+	0	0.403	0.597
5	3:E	3:E		0	0.005	0.995
6	3:E	3:E		0	0.004	0.996
7	1:N	1:N		0.999	0.001	0
8	1:N	1:N		1	0	0
9	2:M	2:M		0.002	0.935	0.062
10	2:M	2:M		0.002	0.969	0.029
11	3:E	3:E		0	0.18	0.82
12	3:E	3:E		0	0.102	0.898

6.6.4 Classification Robustness

We ran the built classifier on the data after applying various levels of Gaussian noise. The experiment result showed the robustness of the 16-metric model by retaining a higher level of accuracy for most of the noise levels. Ten-fold cross validation was used in this experiment. Table 6.15 shows the accuracy starts at 98.27% of 0% noise on both models. The result at 1% noise retains similar accuracy in both models. At 5% noise, the 16-metric model retains 98.27% accuracy whereas the accuracy of the 3-metric model drops to 94.82%. At 10% noise, the accuracy of the 16-metric model drops to 94.82% and the 3-metrics model drops to 81.03%. At 14% noise, the accuracy of the 3-metric model retains 81.03% whereas the accuracy of the 16-metric model drops to 89.65%. At 20% noise, the 16-metric model retains 86.20 whereas the 3-metric model accuracy drops to 75.86%. There was an exception of this trend at 15% noise. The accuracy of the 16-metric model suddenly dropped to 72.41% compared to the result of 3-metric model which is 77.58%. In this case, the result of the 3-metric model was better and the drop rate in the accuracy of the 16-metric model was significant. However, at a 20% level of noise, the accuracy rose again for the 16-metric model in order to be compatible with the trend of the accuracy rate with the noise increase. Even though we do not have a clear understanding why this kind of behavior is displayed, it is possible there is human error in handling the data. We need more investigation to give clearer analysis on this record.

The accuracy rate for the 16-metric model after applying 90% noise on one metric is 98.27%. Only one subject is incorrectly classified. However, the accuracy in the 3-metric model is 94.82% and 3 subjects are incorrectly classified. This accuracy proves the significance of mutual support among the used metrics to retain the classification accuracy.

Table 6.15 The classification accuracy and error rates at various noise levels for the 16-metric and 3-metric models

Noise Level	16-Metric Model		3-Metric Model	
	Accuracy Rate	Error Rate	Accuracy Rate	Error Rate
0%	98.27	1.72	98.27	1.72
1%	98.27	1.72	98.27	1.72
5%	98.27	1.72	94.82	5.17
10%	94.82	5.17	81.03	18.96
14%	89.65	10.34	81.03	18.96
15%	72.41	27.58	77.58	22.41
20%	86.20	13.7	77.58	22.41

We used the area under the ROC Curve which is known as Area Under Curve (AUC) as a measure to study the effect of each noise level on the classifier result. Receiver operating characteristic (ROC) curve is a plot of true positive rate versus the false positive rate, which is 1-specificity versus sensitivity. The area can be calculated by integrating the ROC curve. Specificity is the percentage of negative instances that were predicted as negative. Sensitivity or Recall is the percentage of positive instances that were predicted as positive. The ROC curve is usually used in binary classifier. It also can be used with multiple-class classifier by calculating the specificity and sensitivity for each class compared to other classes.

Table 6.16 shows the AUC for the three classes in each data model where each class is compared to other classes. The values in the table show that the AUC for the 16-metric model is one for the first three noise levels (0%, 1%, and 5%). In the 3-metric model, the AUC for classes M and E drops to 0.99 in 1% noise and class M drops to 0.98% in 5% noise. The results of AUC for all classes in both models decrease or retain a fixed value over all noise levels. The values for classes M and E drop more than in the N class. Further, the AUC for the M class in most cases drops more than the E class. This result explains the fact that the intermediates and experts are closer to each other. In addition, the discrimination significance of novices is higher than it is in the other two

levels. Also it explains that the rate of error in the intermediate class is higher than the other classes. At 15% noise the AUC value for class M does not follow this trend in the 16-metic model. In this case, the value of AUC at 15% noise is less than it is for M class in 20% noise. The 3-metric model performs better at 15% noise than the 16-metric model.

Applying 90% noise on one variable shows that AUC values for the experience level are (N=1, M=1, and E=1) whereas in the 3-metric model the values are (N=1, M=0.992, and E=0.995). The 16-metric model performed better than the 3-metric model at intermediate and expert levels and both models have the same AUC value for the novice.

Table 6.16 The AUC for the three classes at various noise levels for the 16-metric model and 3-metric model.

Noise Level	AUC 16-Metric Model				AUC 3-Metric Model		
	N	M	E		N	M	E
0%	1	1	1		1	1	1
1%	1	1	1		1	0.99	0.99
5%	1	1	1		1	0.98	0.99
10%	1	0.99	0.99		0.98	0.78	0.9
14%	1	0.93	0.65		0.98	0.78	0.9
15%	0.97	0.76	0.85		1	0.87	0.91
20%	0.98	0.84	0.97		0.98	0.82	0.91

6.7 Discussion

It is complicated to compare the results we achieved with results of every study reviewed in Chapter Two. There are several reasons for this limitation. Many of the previous studies have investigated the correlation of individual metrics to the skill level but not the composite metrics. These studies did not go further in the analysis and stopped at proving that these individual metrics correlated with the skill levels. In

addition, some of them performed the whole experiment to study a single metric. Many of the experiments included manual factors and the expert's inputs in the process. Few studies performed prediction analysis such as, classification performance of their models. However, arguments about robustness and reliability were not reported. Most of the studies are limited to two levels of expertise: novice and expert. Finally, many of the previous studies have taken place in virtual environments or robotic environments which are indirectly comparable to our study environment. However, few studies used multiple metrics to perform analysis and classification models. We review these studies and compare their results with our achievement below.

Allen et al. [52] used Support Vector Machines (SVM) in an attempt to increase the accuracy of laparoscopic performance evaluation. Four expert subjects and 26 novice subjects participated in the study. Each subject performed three training tasks: pegboard transfer, pass rope, and cap needle. In addition to the SVM analysis, the z-score normalization was performed to compare the results of the two types. The instruments' 3D position and orientation were captured by placing two electromagnetic sensors on each tool. Four metrics were used in the assessment analysis: time to completion, path length, motion volume, and a control effort parameter that measures the applied forces on the instruments. The prediction results of SVM for the three tasks were respectively 93.7%, 91.3%, and 90.0%.

In our study, 17 subjects performed the pegboard ring transfer in multiple sessions of total 70 sessions. The data were taken at three levels of experience: expert, intermediate and novice. We performed two types of validation on the classification model that is built on the 16-metric model. The test set included data of 12 sessions that were captured after the system and analysis models were built and 16 records of the initially collected data which totaled to 28 data records. The test set validation accuracy we achieved was 96.43% with an error rate of 3.57%. The 10-fold validation accuracy was 98.27% and the error rate was 1.72%. Both validation methods we used showed higher classification accuracy than the Allen et al. [52] study even though they used two

levels of classification and we used three which is more challenging. Allen et al. did not provide analysis about the reliability and tolerance of their classifier for noise. However, they used four metrics in their model and we presented a detailed argument about the robustness of 3-metric and 16-metric models and showed the importance of using a more correlated metric to tolerate noise. This result proves the significance of the contribution of the system we built by using multiple coordinated sensors to measure composite metrics and perform fusion motion analysis to improve the assessment results.

Varadarajan et al. [55] have used the kinematics data acquired by the Da Vinci robot API to find a data model that can accurately assess the surgeons. HMMs used as data analysis model to recognize specific skill gestures and sub-gestures for tasks in order to automatically assess robotic MIS. The task used in this study was a bench-top suturing task. Two experts, three intermediate, and three novices comprised the eight surgeons who performed the task. The kinematics data were recorded to train the model. Each surgeon performed the task four times in a total of thirty-two sessions. Varadarajan et al. collected 78 motion variables using the Da Vinci API from both patient and surgeon sides. Linear discriminant analysis (LDA) based on HMM is used to reduce the motion variables. The accuracy of gesture recognition varied depending on the number of dimensions used. The maximum accuracy was obtained when the number of dimensions was between nine and 17. The experiment included three different setups and three different analyses of HMMs. The best accuracy they achieved was at 17 dimensions and 3-state HMM where the accuracy was 87%. The authors of that study concluded the importance and the need for more dimensions to differentiate between motion gestures and performance assessment.

Lin et al. [74] used the same 78 metrics from the Da Vinci API to build a binary Bayes classifier. The experiment included fifteen expert trials and 12 intermediate trials of performing the suturing task. The accuracy rate they got was about 92% using six metrics dimension. Reiley and Hager [75] used HMM to build a classifier using fourteen

metrics from the Da Vinci API on two levels of task and subtask. Fifty-seven trials of suturing at three different expertise levels were used: nineteen experts, nineteen intermediate, and nineteen novices. The accuracy level they achieved was 95% on the task level and 100% on the subtask level.

In our study, we showed the importance of measuring the correct metrics to build an assessment model that includes a large number of metrics with high correlation on the robustness and accuracy of the assessment. The analysis we performed to find the best number of metrics from the 55 metrics we captured showed the best performance is between 7 and 16 metrics. Varadarajan et al. [55] achieved similar result with number of metrics between 9 and 17. Even though our results are close, the significance is not in the number of metrics to use. The significance is to use a collection of metrics that highly correlates to the experience level. Varadarajan et al. [55], Lin et al. [74], and Reiley and Hager [75] were able to measure 78 metrics provided by the Da Vinci API to find a reliable subset for the assessment. That environment is a robotic one and therefore the robot provides the kinematics data for the arms. These researchers have performed various types of analysis to improve the accuracy. The results showed great improvement until they reached 95% on a task level and 100% on a subtask level as reported by Reiley and Hager [75].

These results achieved by Varadarajan et al. [55], Lin et al. [74], and Reiley and Hager [75] show the significant need of the system we designed. They achieved this level by using the Da Vinci robot which offers a wide range of metrics. However, the robotic MIS is not as widely used as the manual MIS and these approaches cannot be used as assessment methods in the manual environment. Before we built our system, it was not possible to collect a wide variety of metrics using two electromagnetic sensors. We have created an assessment tool out of new technology that can measure a wide variety of metrics and add to it, the capability of fusion analysis to produce composite metrics. This system is environment-independent and can be used in labs, robotic, virtual, and real operation environments. We showed promising accuracy that competes

with the accuracies achieved in the robotic environment. It is the computer vision piece, coordinated with other metrics that makes it possible.

Chmarra et al. [48] used Linear Discriminant Analysis to build a classification model in order to automate MIS assessment. The classification model they built included three classes: experience, intermediate, and novice. The number of subjects was thirty-one and distributed as: 10 experienced, 10 intermediate, and 11 novice. Each participant performed four tasks: pipe cleaner, rubber band beads, and circles. Six assessment metrics were used in the analysis extracted from the MIS tools motion. The metrics were total time, path length, depth, motion smoothness, angular area, and volume. Leave one out cross validation was used to test and validate the method. The result showed that the classification method was able to classify 23 participants out of 31 with an accuracy rate 74.2% and error rate 25.8%. As Chmarra et al. reported, the experiment showed significant difference in the skills level between the novice group and both, the intermediate and the experienced groups, but showed insignificant difference between the experienced and the intermediate groups. Similar to the Chmarra et al. result, our experiment showed the difference between the novice and intermediate as being more significant than the difference between the intermediate and expert levels. They used three levels of classifications similar to our experiment. The accuracy Chmarra et al. achieved was low compared to the accuracy achieved by the experiments in the robotics environment described above. But that is what can be achieved in the classical MIS environment without the availability of a tool that can provide the correct metrics for the assessment. The accuracy we achieved is a significant boost because of the type of metrics the built system could provide through synchronized sensors. Chmarra et al. did not provide analysis about the effect of noise on the model.

Rosen et al. [38-41] studied the force/torque and haptic information from the tool/tissue interactions and the tool/hand interactions. A video-recording was manually edited to define different tool/tissue and tool/hand interactions and synchronize each

interaction with its corresponding force/torque data measures. A classifier then developed based on Markov modeling (MM) and a subset of hidden Markov modeling (HMM) was used to classify the subjects in two categories, novice and expert. The number of subjects was ten where five were novices and five were experts. The classification model was reported in [39]. The reported result of the classification using two classes is of accuracy rate 87.5% and error rate 12.5%. The incorrectly classified subjects were experts classified as novice. In this experiment the video editing and the tool/tissue interactions were manually defined; thus the method is not fully automated. The classification results were achieved by developing a non-intrusive system with the capability of producing fusion metrics and this shows improvement in the accuracy and robustness. The result of leave one out validation in principal component analysis and the 10-fold cross validation show the level of robustness and reliability of the assessment using the system and data model. The 10-fold cross validation classification accuracy rate is 98.27% and error rate is 1.72%.

In this chapter, we presented a detailed analysis and discussion of the data collected in the case study. We discussed the accuracy and reliability of the system and the data model. The results showed high accuracy and reliability of the designed platform to provide automated assessment. The data is validated using different methods, and all of them showed the robustness of the platform. As we see in the list of metrics, many of them are new and have never been studied before. In addition, the chapter provided a comparison between our results and previous studies' results and showed the significant improvement in the assessment accuracy. The platform is open to add many new metrics to improve the robustness. Some of these ideas are presented in the future work chapter.

Chapter 7

Conclusion

This chapter summarizes the results achieved by the thesis and the overall contribution of this work toward improving the accuracy and reliability of performance assessment. It also includes a summary of the questions the research has answered.

7.1 Assessment

From the discussion in Chapter One and Chapter Two we conclude that the effort to solve the MIS surgeons' assessment challenge is about finding reliable, valid, and measurable metrics. Most of the metrics used are quantitative parameters and need external sensors to be measured. This thesis has studied the current MIS assessment approaches and identified four categories:

- Checklists, direct observation, and video-tape observations where master surgeons directly or indirectly observe the trainees and provide assessment and feedback about their skills.

- Kinematics and motion analysis using electromagnetic or mechanical sensors, where an object or objects such as hands and instruments are tracked in the 3D space positions and transformed to kinematics data in order to find a correlation with motion signature and skill level.

- Force/Torque analysis where the force and torque metrics are measured to find their correlation significance with the skill level.

- Virtual reality simulators where the trainee practices on computerized simulators and gets assessment and feedback.

These approaches suffer from several problems and limitations. The main source of limitations to improve the assessment is the technology used and the method resorted to, to approach the problem. These limitations can be summarized as follows:

- The checklist and direct and indirect observation methods are subjective and time and resource consuming.

- The electromagnetic and force/torque sensors are attached to the surgeon's body, which might influence the surgeon's work.

- The electromagnetic sensors can be affected by magnetic fields in the surgery and training environment.

- The virtual reality simulators can only assess the subjects in the simulation environment but not in the real training and operation environment.

- The main limitation is studying an isolated type of motion or measure such as the tools' motion and ignoring the importance of the coordination between the motion of different body and instrument parts. The motion of the tools, hands, head, and eyes are not studied together to find the importance of their interrelationship. The interaction between the head and the eye with objects in the environment, which drives the motion, has never been studied.

In this thesis, we proposed a platform that integrates and synchronizes multiple non-invasive sensors to observe and extract individual and composite metrics from the surgery and training environment. These metrics can be used to recognize patterns of surgeon skills development based on their fusion motion and interactions with the

environment. Unlike the currently known methods, the technique used in this system can be automated and therefore, can scale. Using the system in a case study of 58 subjects in addition to 12 subjects for validation showed high accuracy and reliability. The results showed that metrics related to speed and acceleration, which are widely used in previous studies, are not the best metrics for the assessment. We introduced many new metrics to reach a high level of accuracy. Many of the metrics used are composite metrics coordinated in time to get the fusion of motion analysis. These metrics can only be extracted using computer vision technology and coordinated cues of eyes, external shots of the body and instruments, and internal shots of the operative field. The results show the ability to classify over a large number of subjects which suggests a shift in the way to approach the problem.

7.2 Findings

From the discussion and comparison with other studies, we can summarize the outcomes of this research as follows:

- Employed computer vision techniques for skill assessment in MIS. We used computer vision as a non-intrusive technology and the fusion of motion analysis between the tools and different body parts of the trainees.

- Designed and developed an environment independent system that can capture a wide range of metrics in which many showed high correlation with the skill level. The system has been developed by new technology and can be used to provide assessment to MIS trainees and surgeons. The goal is not to find more metrics but to find the correct metrics for performance evaluation.

- The system is able to capture composed and fusion metrics based on the synchronization of the multiple sensors which adds quality to the measured metrics.

- The case study proved the significant importance of acquiring coordinated data from various objects in the surgery environment on the evaluation process.

- In the study, we found a list of metrics that significantly correlates to the MIS skill level that had not been studied before.

- The research provided an extensive study and analysis to prove the accuracy and reliability of the system and the proposed data model for the assessment.

- The study showed the importance of finding the proper number and type of metrics to build a reliable assessment model.

- The identified 16-metrics model can classify performance with less than 3.57% error on a test set and 1.7% error in the 10-fold validation with high robustness and tolerance to noise and error in data.

- As a result of building this system, the case study showed significant improvement to the assessment accuracy and reliability.

- The new system is environment independent. It can be used in all training and surgical environments including the robotic and virtual ones. However, it is more useful in the labs and real operation theater.

7.3 Thesis Contribution

The coordinated computer vision and tracking cues we used in this study were crucial in allowing us to find a new, multi-cue solution to a longstanding problem in laparoscopy, that of automatic performance assessment in a way that is valid, sensitive, and fine-grained. The study collected many coordinated vision/tracking/performance metrics on 70 subjects, which is a large and statistically significant data set. The vision cues ended up being a crucially important part of the composite metrics that gave correct classification results. This approach was our intuition but we were not sure exactly which cues and what combinations are more significant. We showed in this study that these cues could be appropriately collected in a surgical training environment, and that computer vision should now become a part of these medical training environments.

From the assessment of our findings, we present the following original contributions:

- The identification of the limitations in the current assessment approaches.
- A novel design and implementation of an assessment system that integrates and synchronizes multiple non-intrusive sensors to extract metrics from the environment. The extraction process uses cues of eyes, external shots of the body and instruments, and internal shots of the operative field. The technique used by this system can be automated and therefore, can scale. This design is stand-alone and can work in the training environment as well as the operating theaters.
- Based on the new system, this thesis found new assessment metrics that showed merit, robustness, and reliability.
- Many other novel metrics are proposed, some showed no correlation with the skills level and some might need more study using more complex case studies to prove their reliability and validity.

- The system is designed to be open for expansions and more analysis to study other metrics.
- The system and the metrics are validated using a case study of 58 subjects in addition to 12 subjects studied after the data analysis model is built.
- The ability to extract a model of metrics that can classify the assessment level in three-class resolution. The result of the classification shows significant improvement of the previous studies which used different systems.
- Overall this thesis contributed in improving the reliability and accuracy of MIS objective assessment.

But, over and above these contributions, one of the most valuable contributions made by this thesis is the transformation of the way the assessment problem has been thought of for a long time by utilizing the new technology of computer vision. This transformation allowed expanding the parameters of the assessment to increase reliability. In addition, this transformation opened the door wide for more work and contributions to reach a satisfactory approach to assess MIS trainees and surgeons. This stud also encourages computer vision researchers to improve other challenging problems facing the field of minimally invasive surgery. The results of our study reveal objective metrics for analysis of surgical task performance. We believe these findings, in the context of surgical monitoring, are markedly better than all other known methods. It is the computer vision piece, coordinated with other metrics that makes it possible.

Chapter 8

Future Work

This thesis provided answers to several questions but at the same time it opened more questions to answer. As part of this thesis contribution, this chapter describes more ideas, questions, and directions for future research to improve the performance assessment and skills level recognition as well as to improve the designed system. In addition, the chapter reports preliminary results to one section of the future work to motivate ourselves and others to carry on this research to advanced stages. The following is a collection of ideas to improve the system and the case studies to validate it.

8.1 Use More Complex Case Study

The task used in the case study is a pegboard transfer task. The time to complete this task is short and the training time to master it is relatively short. The case study showed significant correlation coefficients for some metrics and low correlation for others. Many of those metrics are categorized in the study as stress and fatigue metrics. Pegboard transfer is not complex enough to show stress and fatigue in a few minutes. A suggestion is to design more complex case studies using a task that takes a longer time to complete, a longer time to master, and causes more stress and fatigue on the trainee than the pegboard transfer task. A suggested task is suturing which takes more time and requires higher control and experience of psychomotor skills as Figure 8.1 shows.

Further, studies need to be performed to investigate individual features of the metrics and find which is a function of expertise, function of task complexity, or function of both, and how we can utilize each type in the assessment process.

Figure 8.1 Suturing training task

8.2 Larger Number of Subjects

The case study used data for 70 sessions for 17 subjects. A larger set of data boosts the significance of the results. In addition, using a larger data set helps in using analysis methods that require large set of data in order to give reliable results.

8.3 Reduce Cost and Increase Mobility

New computer vision technology can be studied to reduce the cost and increase the mobility of the system. An example is replacing the expensive Vicon system by four or six Kinect cameras. If those cameras can give similar accuracy and reliability as Vicon,

the cost will be significantly reduced. In addition, Kinect cameras are smaller, lighter in weight, and can reduce the setup cost and increase the mobility of the system.

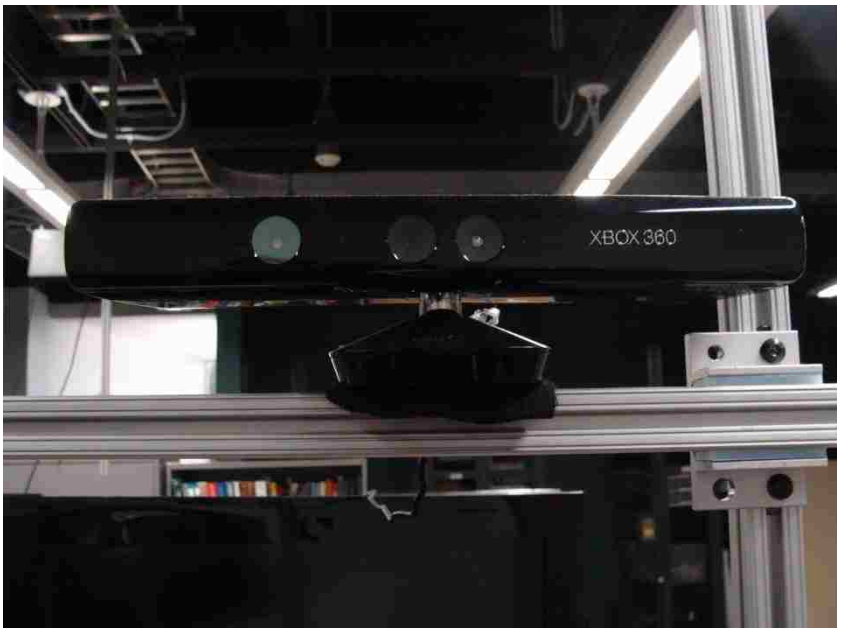


Figure 8.2 Kinect camera

8.4 Find New Metrics

The 3D positions, time, and system synchronization raw data may be retained to extract more metrics. For example we can extract the volume of motion for head, hands, and instruments in the body while looking at or away from the field of view. Many other ideas can be studied to produce metrics which might correlate to the skill level.

8.5 Set up the System in a Training Center

Set up the system in the Center for Advanced Training and Simulation at the University of Kentucky to continually capture data for trainees and validate the system in a real training environment. Capturing a large set of data for real trainees enables us to build an open library of the raw data and metrics. More researchers can use the library to perform more analysis that could improve the assessment and introduce new ideas to improve the system.

8.6 Detect Progress Pace and Custom Feedback

Conduct study to detect the progress pace of a trainee. We can use algorithms to compare the progress of trainees with the reference data we have to detect how long they need to reach the experience level. Based on the principal component analysis discussion in Chapter Six, we can provide custom feedback to the trainee based on the values of metrics and find in which specific area they need to improve. This custom feedback could help speed up the learning and decrease the training time.

8.7 Segment Tasks and Detect Errors

Conduct research to develop computer vision algorithms that can segment the tasks into subtasks. The assessment metrics can be extracted based on the subtask instead of the whole task in general. For example, in pegboard transfer, we can study developing an algorithm that can segment the right hand stage of work and left hand. We can also try to segment the task into subtasks of picking rings from the ring holder, transferring and placing them on the pegs. The metrics for different subjects can be associated with these subtasks. Since we compare specific subtasks, this segmentation can increase the

quality of the assessment. Achieving reliable results in this study could improve the feedback given to the trainee to be more specific and determine the specific subtasks which the trainee is not performing well. In addition, we can develop a computer vision algorithm to detect whether the instruments are moving to achieve a subtask, still in an idle state, or jerking. Further, we can detect errors such as dropping a ring or placing a ring in the wrong place.

8.8 Increase Skill Level Resolution

For complex tasks, we can increase the resolution of the skill level. Instead of using novice, intermediate, and expert, we can use a scale of ten to assess the level of the trainees. This idea however, may not be useful for a simple task like pegboard transfer, but it is important for complex tasks to improve detecting the progress level and customize feedback.

As a preliminary result, we performed a PCA on the data we captured to find out if it was possible to develop a scale instead of three levels. A scale of seven was developed based on the session of training which lasted 30 minutes. The assignment of levels to the subjects was imprecise. The reason is because the idea of including this part of analysis had been added after disconnecting the captured data from the subjects' information. In the estimation process, we used the time and date of the captured data using the computer file system. Figure 8.3 shows the score graph of this experiment. The legends in the graph (L_1-L_7) represent the assumed level of experience where L_1 is the lowest level of experience and L_7 the highest level. The graph shows that the overlap in the new levels is higher than the overlap in the three levels analysis, especially in the early stages of training and in the experience level. After reaching a level that is considered the experienced level, the extra training does not change the behavior and signature of the metrics. Here, L_5, L_6, and L_7 are almost completely overlapped. By using more complex tasks in the experiment, the overlap could be less,

and the result of the relationship among subjects in each level could give better information to understand features specific to a task.

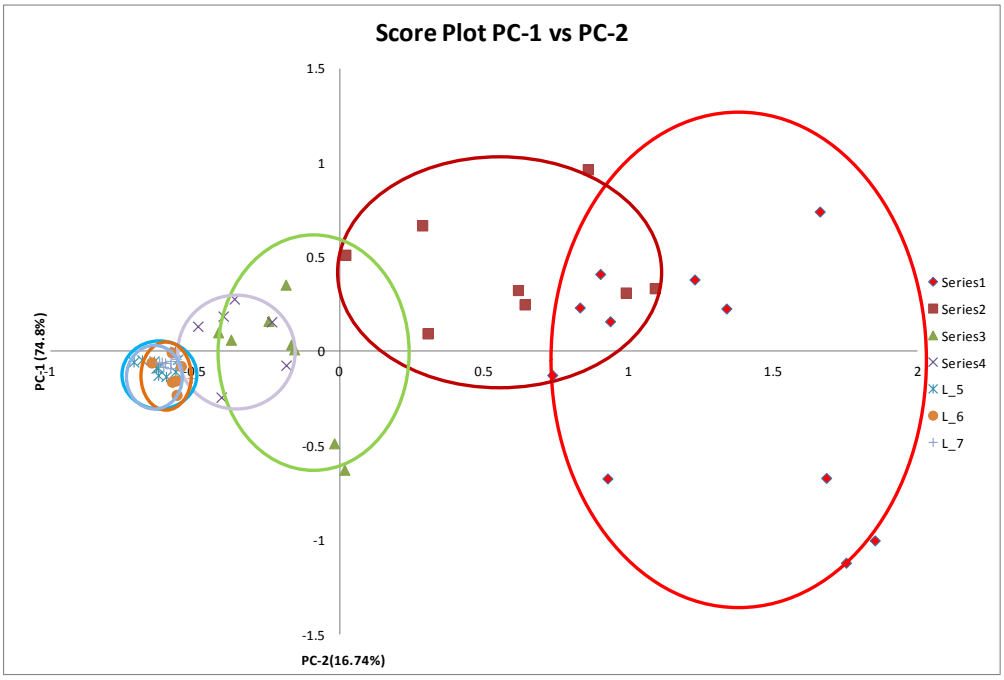


Figure 8.3 PCA score plot of PC-1 vs. PC-2 for a scale of resolution seven

8.9 Assessment Report

Improve the system to produce a detailed assessment report: The idea is to add a feature to produce a report like the checklist used in the OSATS method described in Chapter Two.

8.10 Real Time Feedback

One of the ideas worth studying is the real time assessment and feedback while the trainee is performing the task. The idea is to study the possibility of performing real time and continuous analysis as data being read by the system. This approach enables us to give the trainees continuous assessment and feedback that helps to improve their techniques while they are training. As a possible result of this feature in combination with the task segmentation, the system could identify the part of the task in which the trainee is facing difficulty.

8.11 Assessing New Tools and Environments

Develop a validation study of the multi-sensor system as a tool to validate and assess new surgery tools and environments. The base of the study is using the system to collect data for experts performing tasks using traditional surgical instruments and newly developed instruments or experts practicing in two different environments. This experiment could give comparisons and feedback on which environments or instruments are better to use.

8.12 Plug-N-Play System

Improve the system design to allow the user to include or exclude features. For example, after the improvement, the user can select which metrics to include or exclude in the analysis. If the user does not have all sensors such as, no heart rate monitor or no eye tracker, the user can exclude those sensors. The system should be able to extract metrics which are independent from those sensors and perform analysis based on the available metrics.

8.13 End Note

This thesis contributed toward helping the minimally invasive surgery discipline to automatically assess skills performance and opened more venues to advance the work and develop skill level recognition. Similar ideas can be used in other areas that require psychomotor skills. If the reader of this thesis is interested in more ideas, knowing more about new results, or keen in extending cooperation in similar research, he or she can contact the author.

Appendix A

Appendix A contains the IRB approved consent form used in the data collection process.

Consent to Participate in a Research Study

Technical Skills Assessment of Minimally Invasive Surgeons Using Computer Vision

WHY ARE YOU BEING INVITED TO TAKE PART IN THIS RESEARCH?

You are being invited to take part in a research study about the motion during performing training tasks used in minimally invasive surgery and how this motion relate to the experience of performing the tasks. If you take part in this study, you will be one of about 12 people to do so at UK.

WHO IS DOING THIS STUDY?

This research project is to fulfill a PhD thesis in computer science department. The researcher in charge of this study is Sami Taha Abu Snaineh, PhD Candidate in the department of Computer Science at UK. The researcher is being guided by Brent Seales, PhD, Faculty in the department of computer science.

WHAT IS THE PURPOSE OF THIS STUDY?

The goal of this study is to develop an approach to objectively assess laparoscopic surgery trainees using cameras and computer vision. This approach can be used for the assessment of new trainees, tools, and training environments for laparoscopic surgery

ARE THERE REASONS WHY YOU SHOULD NOT TAKE PART IN THIS STUDY?

You should not participate in the study if you have poor vision that is not corrected by using contacts or eye glasses.

WHERE IS THE STUDY GOING TO TAKE PLACE AND HOW LONG WILL IT LAST?

This study will be conducted during multiple 30 minute visits to UK's Center for Visualization and Virtual Environments.

The number of required visit depends on the group you are participating in. There are three groups represents three level of experience in performing training tasks. Experienced Level: ten visits for training and data capture. In the last visit, the data will be captured and recorded.

Intermediate level: five visits for training and data capture. In the last visit, the data will be captured and recorded.

Novice level: One visit for one hour. You will be introduced and trained for half an hour and the data will be captured in half an hour.

If you decided to be in the experienced group, you can change any time to be in a different group as long as the time of your training falls within the time limit of the other group.

WHAT WILL I BE ASKED TO DO?

1) You will be asked to perform visual-motor tasks similar to the tasks that new surgeons must learn. You will use grasping instruments similar to long-handled tongs to pick up small objects and move them to new location. While you are performing these tasks, you will not be allowed to directly see what you are doing. Instead you must watch your own movements on a large display.

2) You will receive instruction about how to perform the tasks described above and you will be allowed to practice them. We will collect information about how quickly and accurately you will perform the tasks. A video recording will be made of the images on the display during the session. A continuous record will also be made of where you are looking on the display. The last measurement will be made using cameras for tracking eye and head positions. Information about the position of your eyes with

respect to the screen across time will be automatically translated into a series of coordinate values. There will be no videotape of your actual face or eyes.

3) You will be asked to put colored markers on your arms and head (using hat) during the session. Those markers will be tracked to capture the motion of the arms and the head. Only the coordinates of those markers will be recorded and no video will be recorded for the arms or the head.

4) You will be asked to wear a heart rate watch and belt to monitor the heart beat rate during the session.

WHAT ARE THE POSSIBLE RISKS AND DISCOMFORTS?

The things you will be doing should pose no more risk than those you experience when playing a computer or video game. These risks include the potential for mild dizziness, and possible fatigue to your shoulders, arms and hands. You will be asked to take a rest break every 15 minutes or when you feel tired in order to minimize any such symptoms. The eye-tracking and arms/head tracking procedure involves the use of cameras mounted near the display and the ceiling of the room; they do not involve placing sensors in or near the eyes or the body and we are not aware of any danger associated with its use. The markers will be attached to the arms and the hat is shiny colors that can be available on clothes in stores. The heart rate monitor will be placed on the arm like a watch as it is used training exercises.

DO YOU HAVE TO TAKE PART IN THE STUDY?

If you decide to take part in the study, it should be because you really want to volunteer. You will not lose any benefits or rights you normally have if you choose not to volunteer. You can stop at any time droning the study and still keep the benefits and rights you had before volunteering.

IF YOU DON'T WANT TO TAKE PART IN THE STUDY, ARE THERE OTHER CHOICES?

If you do not want to be in the study, there are no other choices except not to take part in the study.

WHAT WILL IT COST YOU TO PARTICIPATE?

There are no costs associated with taking part in this study.

WILL I RECEIVE ANY PAYMENT OR REWARDS FOR TAKING PART IN THE STUDY?

After completing the study you will receive $20. If you decide during the course of the experiment to stop and discontinue, you will receive $10 for your time.

WHO WILL SEE THE INFORMATION I GIVE?

We will make every effort to prevent people who are not on the research team from knowing how you performed. We will assign your data a code number rather than use your name, and these data will be combined with similar data from approximately 12 people taking part in the study. This combined information will be used when we write up the study to share it with other researchers. However, you will not be personally identified in these written materials. Videotapes of your performance (that is, the coordinates of the instruments under the camera, the coordinates of markers on the head and arms, the positions of the eyes, and the heart beat rate) will be kept for a maximum of five years before begin destroyed.

CAN MY TAKING PART IN THIS STUDY END EARLY?

Yes. You have the right to decide to stop participating at any time. Your decision to stop taking part in this study will not jeopardize your right to participate in other studies. You will not be treated negatively if you decide to stop participating before the study is over. The amount money that you will receive will be determined by how much of the study you completed.

WHAT IF I HAVE QUESTIONS?

Feel free to ask any questions that might come to mind now. Later, if you have questions about the study, you can call Sami Taha Abu Snaineh, 859-536-1881, sstaha2@uky.edu. If you have any questions about your rights as a research volunteer, contact the staff of the Office of Research Integrity at the University of Kentucky at 859-

257-9428 or toll free at 1-866-400-9428. We will give you a copy of this consent form to take with you.

_____
_____

Signature of person agreeing to take part in the study Date

_____

Printed name of person taking part in the study

_____
_____

Signature of person obtaining informed consent Date

Appendix B

Appendix B presents figures 5.8-5.19 from Chapter Five. These figures show the magnitude of the metrics that have Pearson's correlation higher than 0.5.

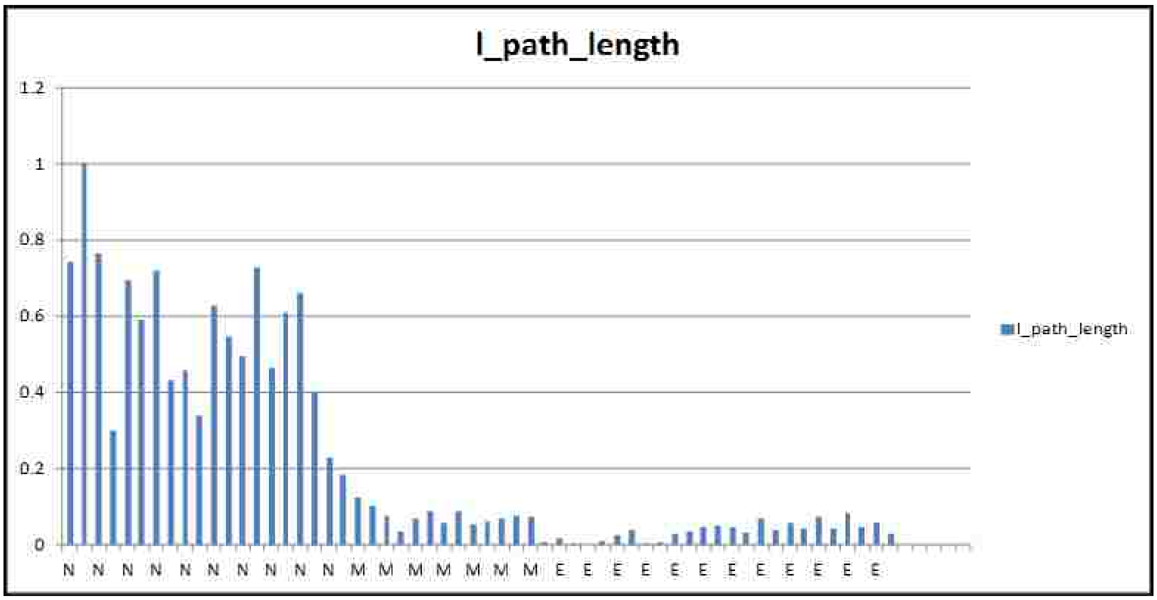


Figure 5.8 Left hand path length with |r|=0.84

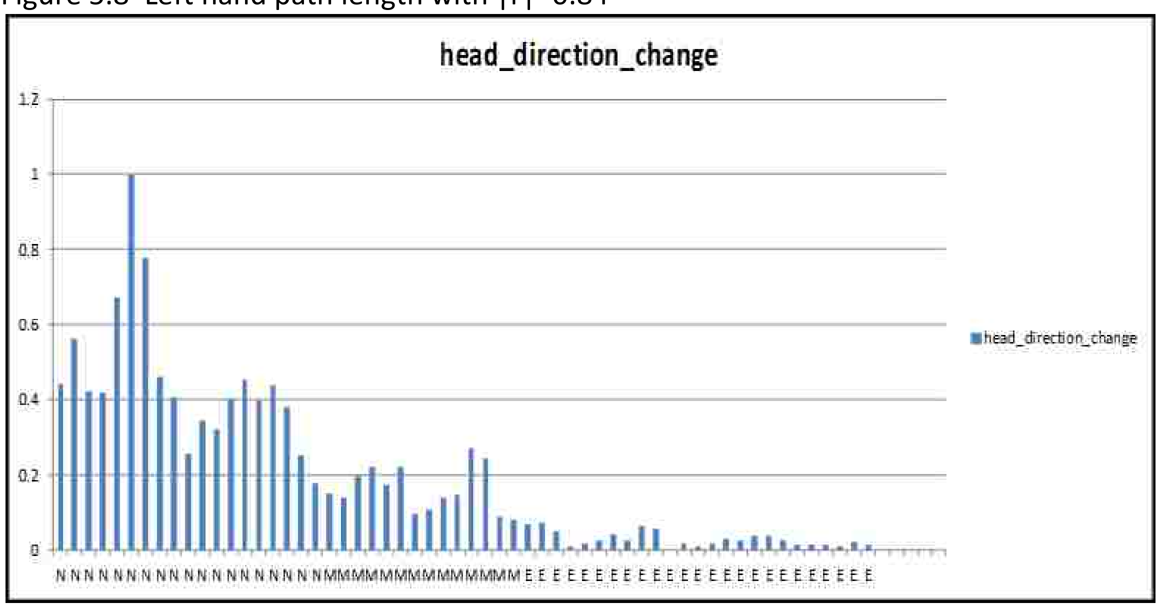


Figure 5.9 The change in head direction with |r|=0.84

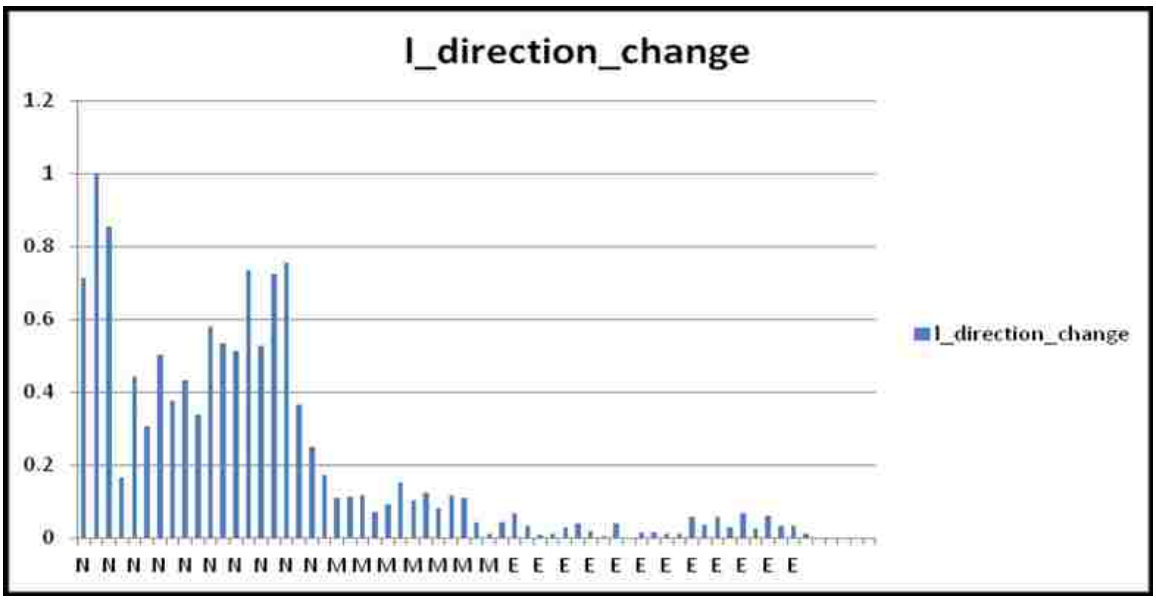


Figure 5.10 The change in the left hand direction with |r|=0.82

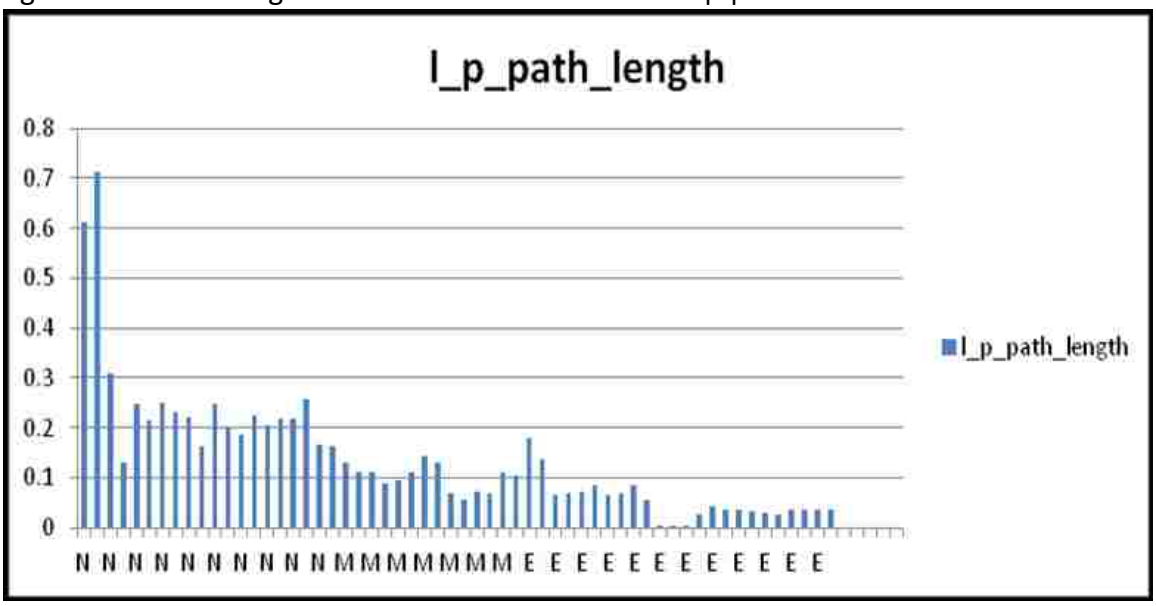


Figure 5.11 Left probe path length with |r|=0.70

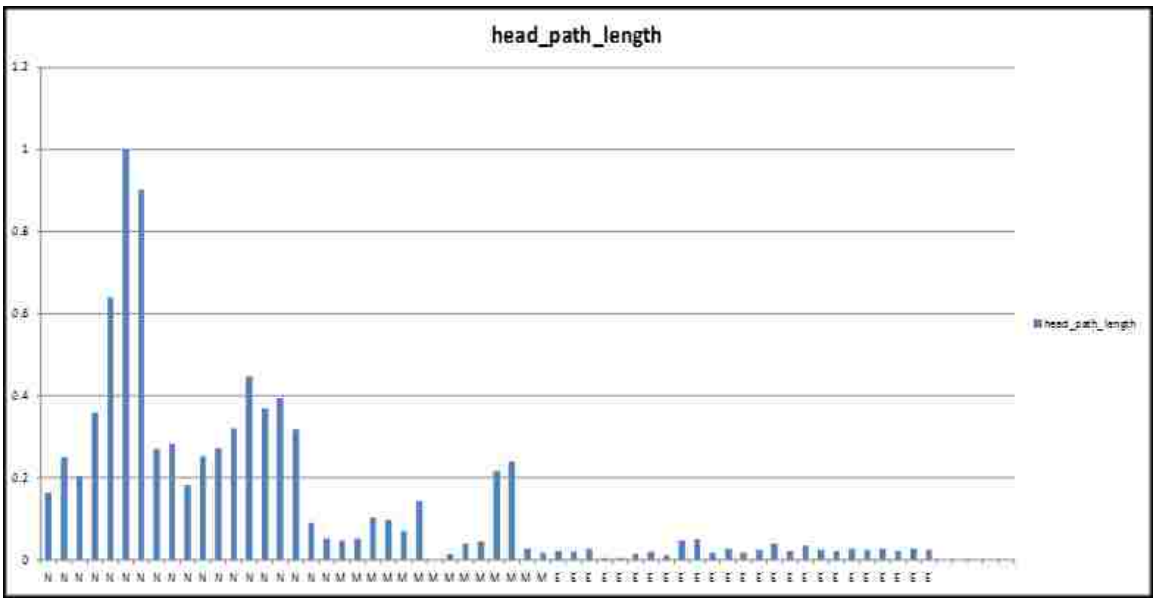


Figure 5.12 Head path length with |r|=0.68

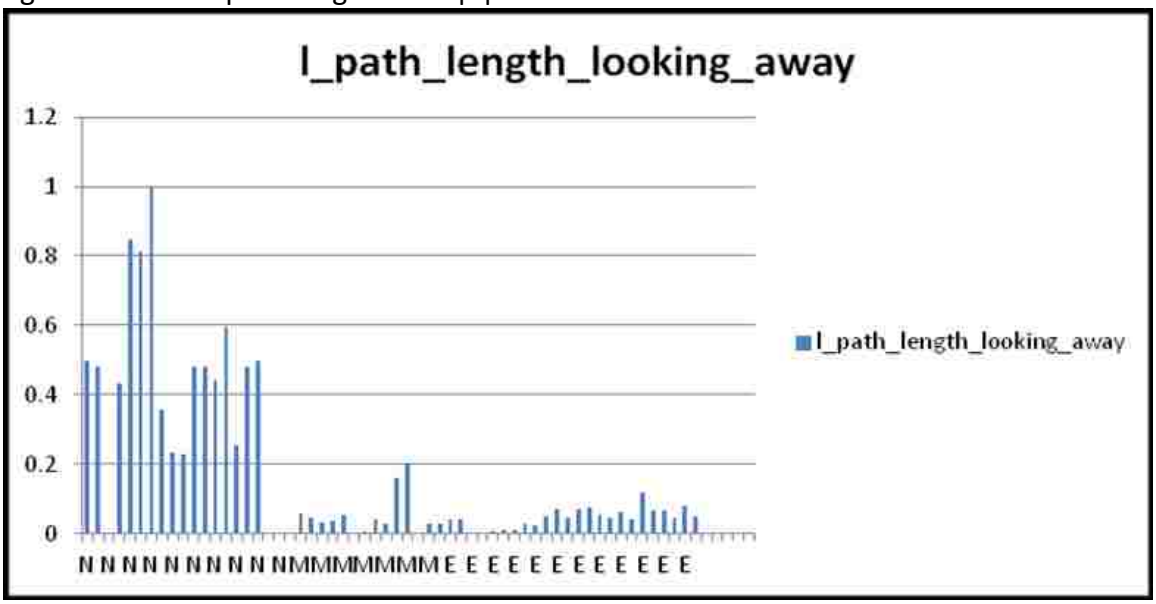


Figure 5.13 Left hand path length while looking away from the display with |r|=0.67

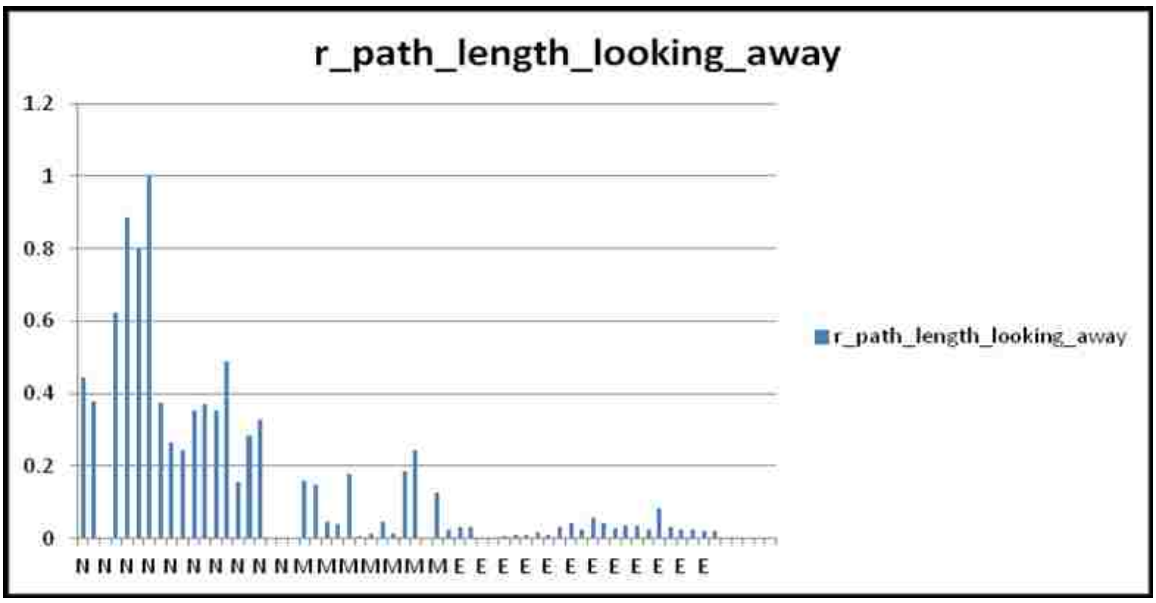


Figure 5.14 Right hand path length while looking away from the display with |r|=0.67

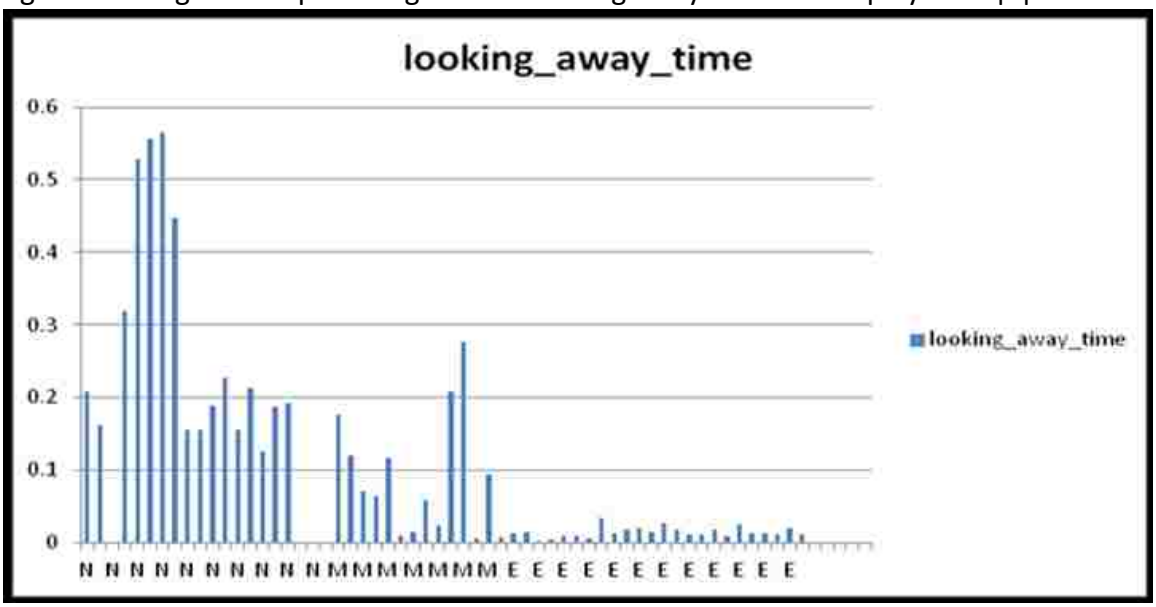


Figure 5.15 The time spent looking away from the display with |r|=0.66

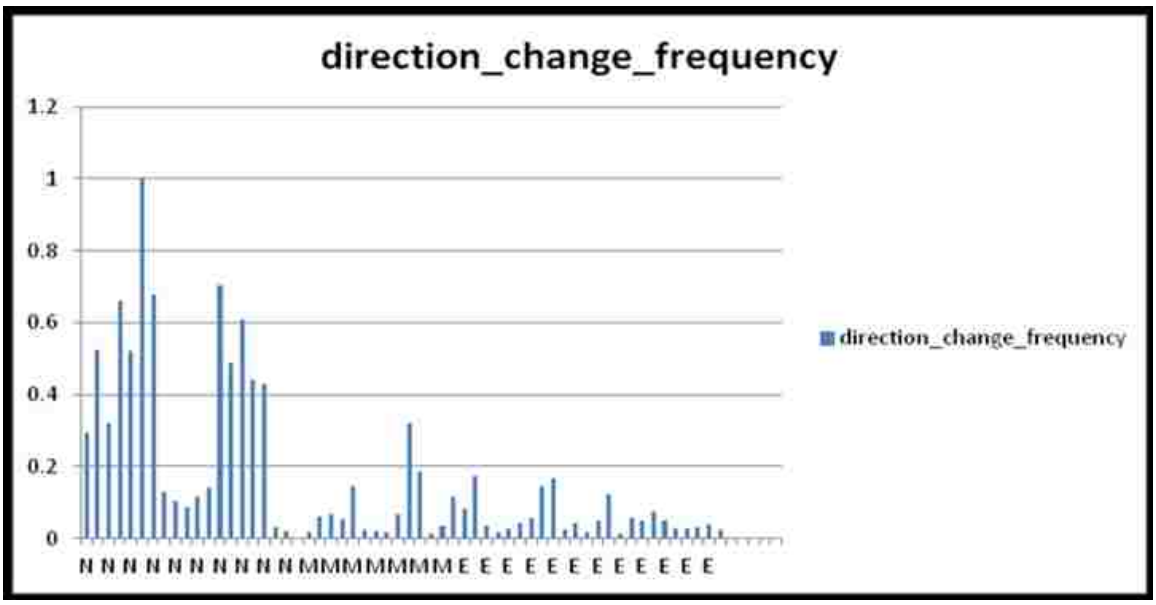


Figure 5.16 the frequency of changing the head direction with |r|=0.61

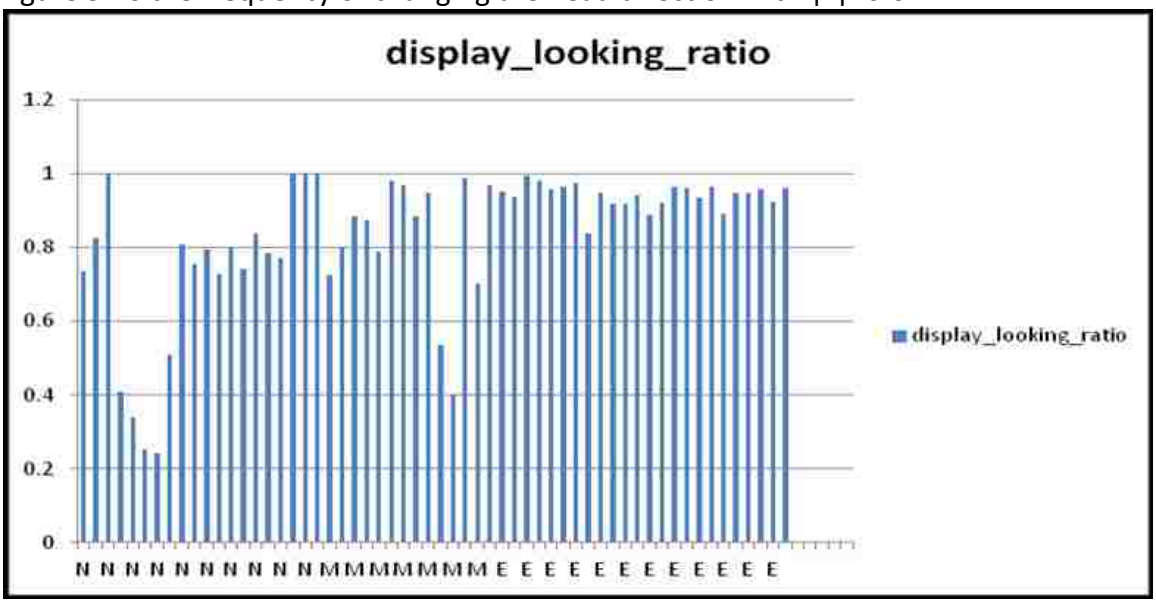


Figure 5.17 The ratio of the time looking away from the display with |r|=0.55

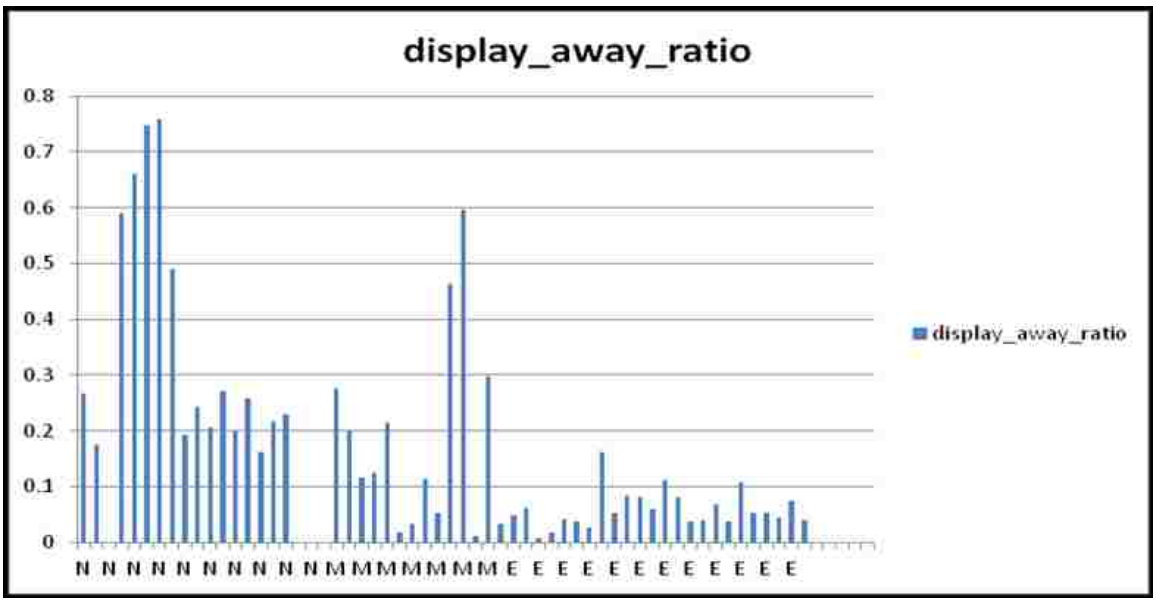


Figure 5.18 The ratio of time looking at the display with |r|=0.55

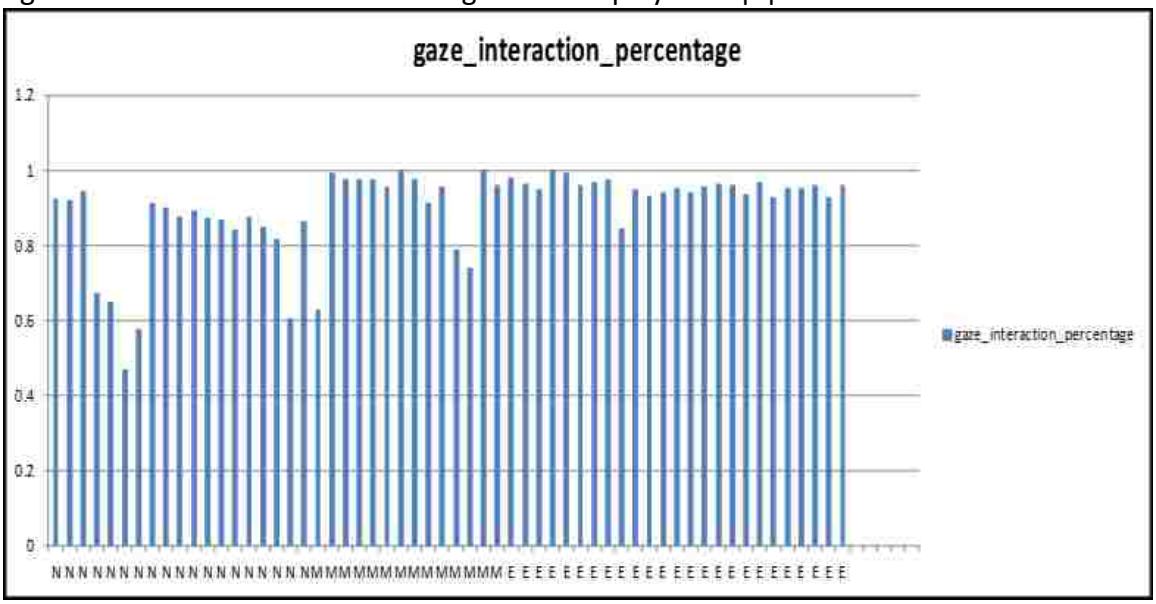


Figure 5.19 The ratio of the gaze interaction with the display with |r|=0.54

References

[1] M. Field, D. Clarke, S. Strup, and W. B. Seales, "Stereo endoscopy as a 3-D measurement tool," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2009, pp. 5748-51, 2009.

[2] R. Berguer, W. D. Smith, and Y. H. Chung, "Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery," *Surgical Endoscopy-Ultrasound and Interventional Techniques,* vol. 15, pp. 1204-1207, Oct 2001.

[3] C. M. Carswell, D. Clarke, and W. B. Seales, "Assessing mental workload during laparoscopic surgery," *Surg Innov,* vol. 12, pp. 80-90, Mar 2005.

[4] S. M. B. I. Botden, I. H. J. T. de Hingh, and J. J. Jakimowicz, "Meaningful assessment method for laparoscopic suturing training in augmented reality," *Surgical Endoscopy and Other Interventional Techniques,* vol. 23, pp. 2221-2228, Oct 2009.

[5] J. C. Rosser, Jr., L. E. Rosser, and R. S. Savalgi, "Objective evaluation of a laparoscopic surgical skill program for residents and senior surgeons," *Arch Surg,* vol. 133, pp. 657-61, Jun 1998.

[6] J. C. Rosser, L. E. Rosser, and R. S. Savalgi, "Skill acquisition and assessment for laparoscopic surgery," *Arch Surg,* vol. 132, pp. 200-4, Feb 1997.

[7] T. R. C. o. S. o. E. a. t. S. a. N. Foundation. (1999). *Surgical Competence - Challenges of assessment in training and practice*.

[8] K. Moorthy, Y. Munz, S. K. Sarker, and A. Darzi, "Objective assessment of technical skills in surgery," *BMJ,* vol. 327, pp. 1032-7, Nov 1 2003.

[9] J. A. Martin, G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents," *Br J Surg,* vol. 84, pp. 273-8, Feb 1997.

[10] S. M. Cristancho, A. J. Hodgson, O. N. Panton, A. Meneghetti, G. Warnock, and K. Qayumi, "Intraoperative monitoring of laparoscopic skill development based on quantitative measures," *Surg Endosc,* vol. 23, pp. 2181-90, Oct 2009.

[11] Moorthy, Munz, Dosis, Bello, and Darzi, "Motion analysis in the training and assessment of minimally invasive surgery," *Minim Invasive Ther Allied Technol,* vol. 12, pp. 137-42, Jul 2003.

[12] C. J. Sultana, "The objective structured assessment of technical skills and the ACGME competencies," *Obstet Gynecol Clin North Am,* vol. 33, pp. 259-65, viii, Jun 2006.

[13] R. Reznick, G. Regehr, H. MacRae, J. Martin, and W. McCulloch, "Testing technical skill via an innovative "bench station" examination," *Am J Surg,* vol. 173, pp. 226-30, Mar 1997.

[14] T. R. Eubanks, R. H. Clements, D. Pohl, N. Williams, D. C. Schaad, S. Horgan, and C. Pellegrini, "An objective scoring system for laparoscopic cholecystectomy," *J Am Coll Surg,* vol. 189, pp. 566-74, Dec 1999.

[15] A. G. Gallagher and R. M. Satava, "Virtual reality as a metric for the assessment of laparoscopic psychomotor skills - Learning curves and reliability measures," *Surgical Endoscopy and Other Interventional Techniques,* vol. 16, pp. 1746-1752, Dec 2002.

[16] E. D. Grober, M. Roberts, E. J. Shin, M. Mahdi, and V. Bacal, "Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence," *American Journal of Surgery,* vol. 199, pp. 81-85, Jan 2010.

[17] V. Datta, S. Mackay, M. Mandalia, and A. Darzi, "The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model," *Journal of the American College of Surgeons,* vol. 193, pp. 479-485, Nov 2001.

[18] C. Sutton, R. McCloy, A. Middlebrook, P. Chater, M. Wilson, and R. Stone, "MIST VR. A laparoscopic surgery procedures trainer and evaluator," *Stud Health Technol Inform,* vol. 39, pp. 598-607, 1997.

[19] N. Taffinder, C. Sutton, R. J. Fishwick, I. C. McManus, and A. Darzi, "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,* vol. 50, pp. 124-30, 1998.

[20] S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, D. Rattner, and S. Dawson, "Metrics for laparoscopic skills trainers: The weakest link!," *Medical Image Computing and Computer-Assisted Intervention-Miccai 2002, Pt 1,* vol. 2488, pp. 35-43, 2002.

[21] A. Dosis, F. Bello, K. Moorthy, Y. Munz, D. Gillies, and A. Darzi, "Real-time synchronization of kinematic and video data for the comprehensive assessment of surgical skills," *Stud Health Technol Inform,* vol. 98, pp. 82-8, 2004.

[22] G. B. Hanna, T. Drew, P. Clinch, B. Hunter, and A. Cuschieri, "Computer-controlled endoscopic performance assessment system," *Surg Endosc,* vol. 12, pp. 997-1000, Jul 1998.

[23] N. K. Francis, G. B. Hanna, and A. Cuschieri, "Reliability of the advanced dundee endoscopic psychomotor tester for bimanual tasks," *Archives of Surgery,* vol. 136, pp. 40-43, Jan 2001.

[24] N. K. Francis, G. B. Hanna, and A. Cuschieri, "The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester - Contrast validity study," *Archives of Surgery,* vol. 137, pp. 841-844, Jul 2002.

[25] J. C. Gillette, N. E. Quick, G. L. Adrales, R. Shapiro, and A. E. Park, "Changes in postural mechanics associated with different types of minimally invasive surgical training exercises," *Surg Endosc,* vol. 17, pp. 259-63, Feb 2003.

[26] T. A. Emam, G. Hanna, and A. Cuschieri, "Ergonomic principles of task alignment, visual display, and direction of execution of laparoscopic bowel suturing," *Surg Endosc,* vol. 16, pp. 267-71, Feb 2002.

[27] A. Dosis, F. Bello, T. Rockall, Y. Munz, K. Moorthy, S. Martin, and A. Darzi, "ROVIMAS: a software package for assessing surgical skills using the da Vinci telemanipulator system," in *Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on*, 2003, pp. 326-329.

[28] Y. Munz, J. Hemandez, S. Bann, F. Bello, A. Dosis, S. Martin, K. Moorthy, T. Rockall, and A. Darzi, "The advantages of 3D visualization in surgical performance with the da Vinci telemanipulation robotic system," in *11th International Congress and Endo Expo/SLS Annual Meeting*, 2002.

[29] Y. Munz, J. Hemandez, S. Bann, F. Bello, A. Dosis, S. Martin, K. Moorthy, T. Rockall, and A. Darzi, "The use of motion analysis in determining the advantages of 3D vision in surgical performance with the da Vinci telemanipulation robotic system," in *Proc. 10th International Congress of the EAES*, 2002.

[30] J. D. Hernandez, S. D. Bann, Y. Munz, K. Moorthy, S. Martin, A. Dosis, F. Bello, V. Datta, T. Rockall, and A. Darzi, "The learning curve of a simulated surgical task using the Da Vinci telemanipulator system," *British Journal of Surgery,* vol. 89, pp. 17-18, Jun 2002.

[31] R. Aggarwal, T. Grantcharov, K. Moorthy, T. Milland, P. Papasavas, A. Dosis, F. Bello, and A. Darzi, "An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room," *Annals of Surgery,* vol. 245, pp. 992-999, Jun 2007.

[32] A. Dosis, R. Aggarwal, F. Bello, K. Moorthy, Y. Munz, and D. Gillies, "Synchronized video and motion analysis for the assessment of procedures in the operating theater," *Archives of Surgery,* vol. 140, pp. 293-299, Mar 2005.

[33] M. S. Wilson, A. Middlebrook, C. Sutton, R. Stone, and R. F. McCloy, "MIST VR: a virtual reality trainer for laparoscopic surgery assesses performance," *Ann R Coll Surg Engl,* vol. 79, pp. 403-4, Nov 1997.

[34] D. T. Woodrum, P. B. Andreatta, R. K. Yellamanchilli, L. Feryus, P. G. Gauger, and R. M. Minter, "Construct validity of the LapSim laparoscopic surgical simulator," *Am J Surg,* vol. 191, pp. 28-32, Jan 2006.

[35] A. Larsson, "An open and flexible framework for computer aided surgical training," *Stud Health Technol Inform,* vol. 81, pp. 263-5, 2001.

[36] R. Aggarwal, K. Moorthy, and A. Darzi, "Laparoscopic skills training and assessment," *Br J Surg,* vol. 91, pp. 1549-58, Dec 2004.

[37] P. S. Kundhal and T. P. Grantcharov, "Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room," *Surg Endosc,* vol. 23, pp. 645-9, Mar 2009.

[38] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan, "Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions," *Stud Health Technol Inform,* vol. 81, pp. 417-23, 2001.

[39] J. Rosen, B. Hannaford, C. G. Richards, and M. N. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE Trans Biomed Eng,* vol. 48, pp. 579-91, May 2001.

[40] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan, "Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model," *Comput Aided Surg,* vol. 7, pp. 49-61, 2002.

[41] J. Rosen, M. MacFarlane, C. Richards, B. Hannaford, and M. Sinanan, "Surgeon-tool force/torque signatures--evaluation of surgical skills in minimally invasive surgery," *Stud Health Technol Inform,* vol. 62, pp. 290-6, 1999.

[42] C. Richards, J. Rosen, B. Hannaford, C. Pellegrini, and M. Sinanan, "Skills evaluation in minimally invasive surgery using force/torque signatures," *Surg Endosc,* vol. 14, pp. 791-8, Sep 2000.

[43] A. G. Gallagher, K. Richie, N. McClure, and J. McGuigan, "Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality," *World Journal of Surgery,* vol. 25, pp. 1478-83, Nov 2001.

[44] C. D. Smith, T. M. Farrell, S. S. McNatt, and R. E. Metreveli, "Assessing laparoscopic manipulative skills," *Am J Surg,* vol. 181, pp. 547-50, Jun 2001.

[45] S. D. Bann, M. S. Khan, and A. W. Darzi, "Measurement of surgical dexterity using motion analysis of simple bench tasks," *World Journal of Surgery,* vol. 27, pp. 390-394, Apr 2003.

[46] V. Datta, A. Chang, S. Mackay, and A. Darzi, "The relationship between motion analysis and surgical technical assessments," *American Journal of Surgery,* vol. 184, pp. 70-73, Jul 2002.

[47] J. D. Hernandez, S. D. Bann, Y. Munz, K. Moorthy, V. Datta, S. Martin, A. Dosis, F. Bello, A. Darzi, and T. Rockall, "Qualitative and quantitative analysis of the learning curve of a simulated surgical task on the da Vinci system," *Surgical Endoscopy and Other Interventional Techniques,* vol. 18, pp. 372-378, Mar 2004.

[48] M. K. Chmarra, S. Klein, J. C. F. de Winter, F. W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surgical Endoscopy and Other Interventional Techniques,* vol. 24, pp. 1031-1039, May 2010.

[49] J. Jayender, R. Estépar, and K. G. Vosburgh, "New kinematic metric for quantifying surgical skill for flexible instrument manipulation," presented at the Proceedings of the First international conference on Information processing in computer-assisted interventions, Geneva, Switzerland, 2010.

[50] J. Salgado, T. P. Grantcharov, P. K. Papasavas, D. J. Gagne, and P. F. Caushaj, "Technical skills assessment as part of the selection process for a fellowship in minimally invasive surgery," *Surgical Endoscopy and Other Interventional Techniques,* vol. 23, pp. 641-644, Mar 2009.

[51] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario, "Modelling and evaluation of surgical performance using hidden Markov models," *IEEE Transactions on Biomedical Engineering,* vol. 53, pp. 1911-1919, Oct 2006.

[52] B. Allen, V. Nistor, E. Dutson, G. Carman, C. Lewis, and P. Faloutsos, "Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks," *Surgical Endoscopy and Other Interventional Techniques,* vol. 24, pp. 170-178, Jan 2010.

[53] G. B. Hanna and A. Cuschieri, "Influence of the optical axis-to-target view angle on endoscopic task performance," *Surg Endosc,* vol. 13, pp. 371-5, Apr 1999.

[54] C. Sokollik, J. Gross, and G. Buess, "New model for skills assessment and training progress in minimally invasive surgery," *Surg Endosc,* vol. 18, pp. 495-500, Mar 2004.

[55] B. Varadarajan, C. Reiley, H. Lin, S. Khudanpur, and G. Hager, "Data-derived models for segmentation with application to surgical assessment and training," *Med Image Comput Comput Assist Interv,* vol. 12, pp. 426-34, 2009.

[56] R. Hartley and A. Zisserman, *Multiple View Geometry (second edition)*: Cambridge University Press, 2003.

[57] R. Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3d Machine Vision Metrology Using Off-the-Shelf Tv Cameras and Lenses," *IEEE Journal of Robotics and Automation,* vol. 3, pp. 323-344, Aug 1987.

[58] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 1106-1112.

[59] G. Bradski and A. Kaebler, *Learning OpenCV Computer Vision with the OpenCV Library*: O'Reilly, 2008.

[60] V. M. S. Limited. (2006). *Vicon MX Hardware System Reference*.

[61] *What Is an Electrocardiogram?* Available: http://www.nhlbi.nih.gov/health/dci/Diseases/ekg/ekg_what.html

[62] Z. D. Martínez, J. F. Menéndez, and M. J. S. Vargas, "See5 Algorithm versus Discriminant Analysis. An Application to the Prediction of Insolvency in Spanish Non-life Insurance Companies," *Investment Management and Financial Innovations,* 2004.

[63] J. De Andrés, "Statistical Techniques vs. See5 Algorithm. An Application to a Small Business Environment," *The International Journal of Digital Accounting Research,* vol. 1, pp. 153-179, 2001.

[64] Available: http://rulequest.com/see5-comparison.html

[65] C. Lio, C. Carswell, S. Strup, J. Roth, and R. Grant, "THE OPERATING ROOM AS CLASSROOM: UNDERSTANDING COGNITIVE CHALLENGES FACING SURGICAL TRAINEES," in *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 2010, pp. 1571-1575.

[66] *Center for Advanced Training and Simulation*. Available: http://www.mc.uky.edu/mis/skill.asp

[67] I. Corporation. *Open source computer vision library*. Available: http://www.intel.com/technology/computing/opencv/

[68] Z. Y. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 1330-1334, Nov 2000.

[69] Esbensen Kim, "Multivariate Data Analysis in practice", Oslo , CAMO Software, 2010

[70] Tammimmagadda S. Knka P., Yaramala V.B., "Implementation of Clustering through Machine Learning Tool", IJCSI Internation journal of Computer Science Issues, Vol. 8, Issu1, January 2011, pp: 395-401

[71] Marsland, Stephen, "Machine Learning: An Algorithmic perspective", Boca Raton, Chap & Hall/CRC, 2009

[72] CAMO software. Available: http://www.camo.com

[73] Kurt Varmuza, Peter Filzmoser, "Introduction to Multivariate Statistical Analysis in Chemometrics", Boca Raton, Chap & Hall/CRC, 2009

[74] H.C. Lin, I. Shafran, T.E. Murphy, A.M. Okamura, D.D. Yuh, G.D. Hager, "Automatic detection and segmentation of robot-assisted surgical motions", in: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 802–810. Springer, Heidelberg (2005)

[75] C. Reiley , G. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery." Computing and Computer Assisted Intervention, vol. 5761, pp. 435–442, 2009.

Vita

Sami Suleiman Taha Abu Snaineh was born in Jerusalem, Palestine. He attended Palestine Polytechnic University (PPU) from 1996 to 2001 to earn a B.Sc. in Computer Systems Engineering. From 2001 to 2003 he worked at PPU as Teaching Assistant. In the summer of 2003 he joined Maharishi University of Management (MUM) in professional master program in Computer Science. He earned M.Sc. in Computer Science from MUM in spring 2007. From 2005 to 2008 he worked at IBM as senior Software Engineer. He joined the University of Kentucky (UK) in 2009 to start graduate studies in Computer Science. While in UK Sami was awarded Kentucky Graduate Scholarship from 2009 to spring 2013. From 2009-2011 Sami worked as Teaching Assistant and Research Assistant for the Computer Science department at UK. He earned the award of the Teacher of the year in 2012. Sami also received a certificate in college teaching and learning from the Graduate School of UK. In 2011 he joined Lexmark International as a senior Software Engineer for embedded systems. In August 2013, he will assume the duties of an Assistant Professor at Palestine Polytechnic University in Palestine.