Surg Endosc (2009) 23:2181–2190
DOI 10.1007/s00464-008-0246-9

Intraoperative monitoring of laparoscopic skill development based on quantitative measures

Sayra M. Cristancho · Antony J. Hodgson · O. N. M. Panton · Adam Meneghetti · Garth Warnock · Karim Qayumi

Received: 29 March 2008 / Accepted: 15 November 2008 / Published online: 31 December 2008

Abstract
Background Methods for evaluating standard skills in the operating room typically are based on direct observation and checklists, but such evaluations are time consuming and can be subject to bias. It often is possible to acquire more objective measurements using surgical simulators. However, motor performance in simulators can differ significantly from that in the operating room. Intraoperative assessment is particularly challenging because of the significant variability between procedures related to differences in the patients, the surgical setup, and the team. This study aimed to evaluate the feasibility of using a new framework for interpreting quantitative measures acquired in the operating room to distinguish between levels of laparoscopic skill development.
Methods Two levels of surgical skill development were observed, namely, those of three fourth-year residents and three attending surgeons performing three laparoscopic cholecystectomies each. Electromagnetic position sensors were attached by the surgeons to a 5-mm curved dissector and a 5-mm atraumatic grasper. From the tools' position histories and video recordings, time, kinematics, and movement transition measures were extracted. Various measures such as the Kolmogorov–Smirnov statistic and the Jensen–Shanon Divergence were used to provide intuitive dimensionless difference measures ranging from 0 to 1. These scores were used to compare residents and expert surgeons executing two surgical tasks: exposure of Calot's triangle and dissection of the cystic duct and artery.
Results The two groups could be clearly differentiated in both tasks during monitoring for the dominant hand (analysis of variance [ANOVA] and Mann–Whitney; $p < 0.05$) but not for the nondominant hand.
Conclusions It is practical to acquire time, kinematic, and movement transition measures intraoperatively using video and electromagnetic position-sensing technologies. Principal component analysis proved to be a useful technique for presenting differences between skill levels based on those measures. The authors conclude that objective assessment of intraoperative surgical motor behavior is feasible and likely practical.

S. M. Cristancho (✉) · A. J. Hodgson
Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada
e-mail: scrista@interchange.ubc.ca

S. M. Cristancho
Facultad de Ingenieria Electronica, Universidad Pontificia Bolivariana Seccional Bucaramanga, Santander, Colombia

O. N. M. Panton · A. Meneghetti · G. Warnock · K. Qayumi
Department of Surgery, University of British Columbia, Vancouver, Canada

O. N. M. Panton · A. Meneghetti · G. Warnock · K. Qayumi
Centre of Excellence for Surgical Education and Innovation, Vancouver General Hospital, Vancouver, Canada

Keywords Laparoscopy · Motor skills · Performance · Quantitative assessment

The minimally invasive surgical (MIS) techniques that have emerged over the past two decades are largely more difficult to master with proficiency than the corresponding open techniques. Inserting tools through keyhole incisions reduces the number of degrees of freedom, virtually eliminates direct tactile feedback, and because of the fulcrum effect, produces motion reversals, all of which pose significant learning challenges. At the same time, it has become increasingly difficult to use operating room time for instruction, so training moves increasingly toward simulated scenarios such as animal and cadaver labs as well as physical and virtual reality simulators [1–3].

A number of simulators have been developed by various groups [4–8]. However, although physical and virtual reality simulators can facilitate the design of graduated training programs by controlling or eliminating some key sources of variability (e.g., differences in patients' anatomy and disease processes) and by presenting task elements to the trainee in increasing order of difficulty, and although these simulators can allow us to acquire objective measurements of the trainee during training, it still is not clear how effective simulators are in developing surgical skills.

It is fairly well established that people with preexisting surgical skill can perform better in simulators than novices and that training in the simulator can improve skills in the simulator [9, 10]. However, there is relatively little understanding of the relationship between performances in simulated environments and those in the operating room [9–11]. Although good correlations of performances between the minimally invasive surgical trainer in virtual reality (MIST-VR) simulator and a pig model of cholecystectomy have been found [12], more complete comparisons between performances in simulators and those in the operating room have been hampered by a lack of universally agreed-on metrics for quantitative skill assessment, by differing skill levels of the participants, and by small sample sizes [9, 12].

Because it is not yet possible to use performance in simulated surgical tasks to measure surgical skill reliably in the live operating room, it remains necessary to perform intraoperative assessments. Current performance evaluation methods used in the operating room include direct observation, global assessments, and checklists [13, 14]. These have proved largely to be reasonably effective and reliable, but they require evaluators to be in the operating room for the entire case to monitor the surgeon's movements and errors. Moreover, they are time-consuming, and some protocols can be subject to bias [15]. Moreover, these types of evaluation typically provide limited information for further focused training because they usually are performed at the "whole procedure" level and normally do not distinguish between elements of the overall surgery judged to be relatively easier or more difficult, nor do they provide detailed commentary and feedback about specific surgical skills. They also are subject to intraoperative variability (due to differences in patients' condition, the operating room staff, equipment, and the like), so reliability can be difficult to establish using these methods [16].

It would therefore be desirable to introduce some elements of objective assessment into the operating room environment, especially if such assessments could be performed with relatively less effort than required for the current manual assessments. Experience with quantitative assessments conducted with surgical simulators has demonstrated that differences in skill levels can be reliably reflected in quantitative measures such as the speed of a surgeon's hand movements or the number of movements needed to perform a task [13, 17–20].

In earlier work, we established that it is feasible to acquire kinematic data in the operating room [21]. To deal with the interprocedure variability that characterizes live operations, we also earlier proposed and validated a surgical modeling tool—the Motor and Cognitive Modeling Diagram (MCMD)—which can be used to represent a procedure hierarchically using a four-level decomposition consisting of phase, task, subtask, and action levels [22]. By attaching particular surgical gestures to the appropriate elements in this decomposition, we can analyze movements in context and aggregate results from similar elements at different points in the same procedure and across different procedures and different surgeons, thereby enabling meaningful comparisons.

We have shown that the MCMD has sufficient generality to describe different laparoscopic procedures, including cholecystectomy and colectomy procedures [22]. By incorporating a data stream comprising measurements of the position and orientation of selected surgical tools, the MCMD can support assessments of surgical motor performance based on time, transition, and kinematic measures.

Although this evaluation is not yet automated (identification of the phase, task, subtask, and action boundaries still relies on a manual video analysis), it is sufficiently developed for us to test whether analyses of live operating room data are repeatable enough to distinguish reliably between surgeons of different skill levels or not and thereby provide a level of discrimination between groups similar to that found in simulator settings. Therefore, the study reported here aimed to evaluate whether a quantitative analysis of live surgeries based on MCMD will produce relatively low intrasubject variation, moderately low intragroup variation (among a group of surgeons at a similar level of skill development), and significant intergroup variation, thereby demonstrating that quantitative measures can be used to distinguish between surgeons of different skill levels in the live operating room.

Materials and methods

Subjects

We recruited two sets of subjects to represent different levels of skill development: residents (represented by three fourth-year surgical residents, that is, at the earliest stage of training at which they could perform a laparoscopic cholecystectomy under supervision), and experts (represented by three attending surgeons, each of whom had performed hundreds of laparoscopic cholecystectomies).

Both the Vancouver Coastal Health Authority and the University of British Columbia (UBC) research ethics boards granted ethics approval for this study. Residents were protected from coercion by ensuring that their clinical supervisors were not involved in the participation request process and so did not know which particular residents were invited to participate in the study nor the reply of any individual resident. We were not able to protect against supervisors knowing the identity of residents who accepted the invitation to participate because their surgeries were instrumented and monitored, and the presence of the equipment was obvious. However, clinical supervisors were never shown data with any identifying information attached.

Experimental setup

We observed surgeons in the operating room performing three laparoscopic cholecystectomies (LC) per subject using standard surgical tools. Laparoscopic cholecystectomy was the procedure of choice because it is one of the earliest MIS procedures a resident learns to perform and has become widely used as an index operation for ongoing assessment of laparoscopic skills [23].

To measure the positions of surgical tools, we provided two custom-designed clips to which small (~ 1 cm^3) Polhemus electromagnetic tracking sensors (Polhemus Inc., Colchester, VT, USA) were attached. At the beginning of the procedure, the surgeon attached the clips to two surgical tools: a 5-mm curved dissector for the dominant hand and an atraumatic grasper for the nondominant hand. These tools were chosen because they are used for some of the most skill-dependent tasks in LC surgery.

We used the Polhemus electromagnetic tracking system because our earlier work [24] showed that its sampling rate of 120 Hz was high enough to capture all relevant frequencies of the tool movement without running into aliasing (undersampling) problems. Previous studies [25] have shown its accuracy to be sufficient for the type of analyses presented in this report. The Polhemus product manual [26] quotes an accuracy of 2 mm within a 1-m^3 working volume. In the operating room setting, we were able to position the transmitter at an average distance of 20–25 cm from the receiver, which is well within this volume [25].

The sensor's repeatability (ability to detect differential movement) typically is an order of magnitude better than its accuracy [26]. The sensor clips were precalibrated so that we could determine the three-dimensional location and orientation of the tool tip at each sampling instant. Figure 1 presents our experimental setup in the operating room.

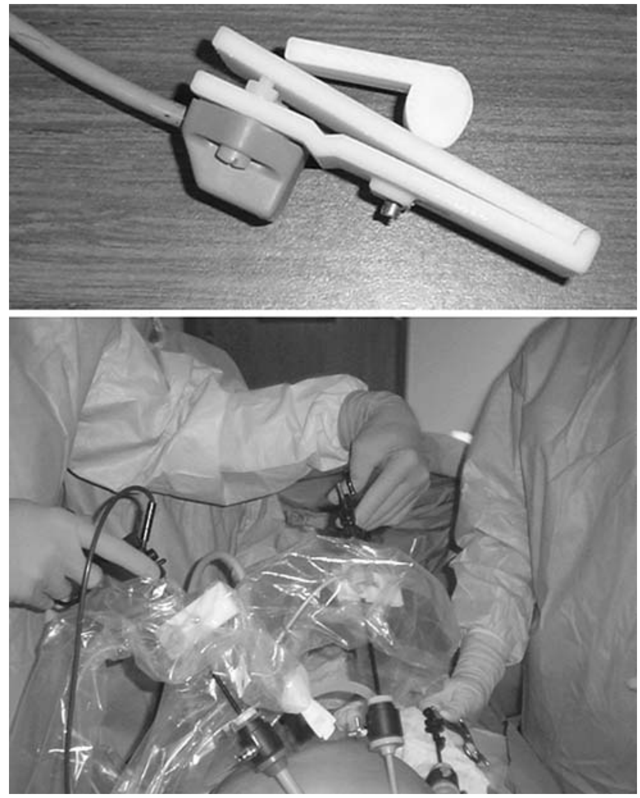

Fig. 1 *Top:* Custom-designed sensor clip with a Polhemus electromagnetic sensor attached. *Bottom:* Electromagnetic sensors attached to surgical tools in use in the operating room during a live procedure

The standard laparoscopic camera (Stryker Inc., Kalamazoo, MI, USA) and tools were used, and the tasks were digitally recorded using video to allow segmentation of the kinematic data stream using the labels in the MCMD. All equipment used was approved by the Biomedical Engineering Department at UBC Hospital, and sterilized using ethylene oxide where appropriate.

Analytical methods

This study focused on analyzing the movements of the dominant hand for two subtasks in MCMD for LC identified by the expert surgeons as the most demanding steps of the procedure in terms of the surgical dexterity required. These two subtasks, part of the "isolate cystic duct [CD] and cystic artery [CA]" surgical task, are called "expose triangle" and "dissect cystic duct," respectively. In addition, each subtask was further decomposed into a set of 10 fundamental surgical actions (or gestures) defined by Cristancho et al. [22] as pull, push, reach, sweep, spread, orient, grasp and hold, grasp and cut, idle, and out. We performed analyses of intrasubject, intragroup, and intergroup variability at both the subtask and action levels.

Variability analysis at the subtask level

The position data from the two main subtasks, "expose triangle" and "dissect CD/CA," were separated from the electromagnetic sensor's data stream by manual identification of the start and end times to the nearest 1/30 of a second using video analysis. The "expose triangle" subtask consists of retracting the gallbladder and dissecting some of the surrounding tissue so as to open the cystic pedicle space and identify where the cystic duct and cystic artery lie. It begins when the gallbladder is first stretched out and ends when the cystic pedicle is identified. The "dissect CD/CA" subtask consists of identifying and isolating the cystic duct from the cystic artery by dissecting the surrounding tissue. It begins when the tip of the tool is first inserted between the two anatomic structures and ends when both structures have been completely freed from one another.

The video clips were further decomposed at the action level by identifying the beginnings and endings of the fundamental surgical motions, and the time records were used to segment the corresponding sensor data stream. The three-dimensional tip velocities were derived from the position records through numeric differentiation using a generalized cross-validation algorithm with an approximate cutoff frequency of 20 Hz. The smoothing parameter (B) was set to a constant value of 10^{-11} for all data processing after a wide variety of data streams from the various subjects and procedures had been analyzed and the resulting smoothing parameter values had been averaged [27].

For each subject (i) and procedure (j) (S_{ij}) and for each subtask, we created a cumulative distribution function (CDF) for the absolute value of each velocity component in the three directions relative to the tool tip (lateral, axial, and vertical). These CDFs represent the proportion of samples with an absolute velocity component in the given direction that is less than the value of the corresponding abscissa. These give a more complete picture of the velocity distributions than a simple measure such as the mean or median absolute velocity.

We next constructed three matrices for each subtask in which each row corresponded to a particular combination of subject and procedure. Each matrix represented a velocity in one of the three cardinal directions: lateral, axial, and vertical. Each row contained 11 velocity components. These 11 velocity components, derived from the corresponding cumulative velocity distributions, represented the velocity values at 5%tile increments between the 25th and 75th percentiles, as shown in Table 1. This range was selected because it broadly represents the range of the velocities used by each surgeon without becoming overly subject to extreme values. We independently checked a subset of our data to test the sensitivity of our results to the selection of the end points of this range and found no significant impact.

Table 1 Symbolic representation of a data matrix for the axial velocity components (indicated by "a") of one of the surgical subtasks*

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

* Each row consists of 11 boxes. Each box contains the corresponding percentile value for the axial velocity of the dominant hand during the "expose Calot's triangle" surgical subtask

Once the data from each category were arranged into the corresponding matrix form, we normalized it by dividing each element by the column standard deviation (defined across all procedures and subjects), and then used the principal components analysis (PCA) dimension reduction technique to extract the dominant contributors to overall variability [28].

A PCA analysis replaces a large number of physical measurements with a small number of more abstract weighting terms. The result of this analysis is a new matrix with the same number of rows but a much reduced number of columns, which allows a simpler presentation of the processed data to the surgical trainer.

Applying this approach to each movement direction separately, we found that for all the cases computed, we typically could reduce the number of columns to two while still explaining 90% to 99% of the variability (e.g., see the data for the axial movement direction shown in Fig. 2).

To test the hypothesis that level of skill development is apparent in intraoperatively acquired quantitative measurements, we computed the contributions to variability in the PCA weight space (defined as the mean squared distance of points in the weight space relative to the global mean position across all subjects and groups) due to intrasubject, intragroup, and intergroup variability. We evaluated the null hypothesis that residents and experts all have the same mean positions in this weight space using a nested analysis of variance (ANOVA) test with four degrees of freedom at the subgroup level and one degree of freedom at the group level ($\alpha = 0.05$). We also computed distances from each subject's center to the experts' center (as the reference group) and applied a Mann–Whitney test to evaluate whether residents were, on the average, located further away from the experts' center than the experts themselves, which would demonstrate significant group separation.

Variability analysis at the action level

At the more detailed action level, we compared performances of individual subtasks by decomposing them into

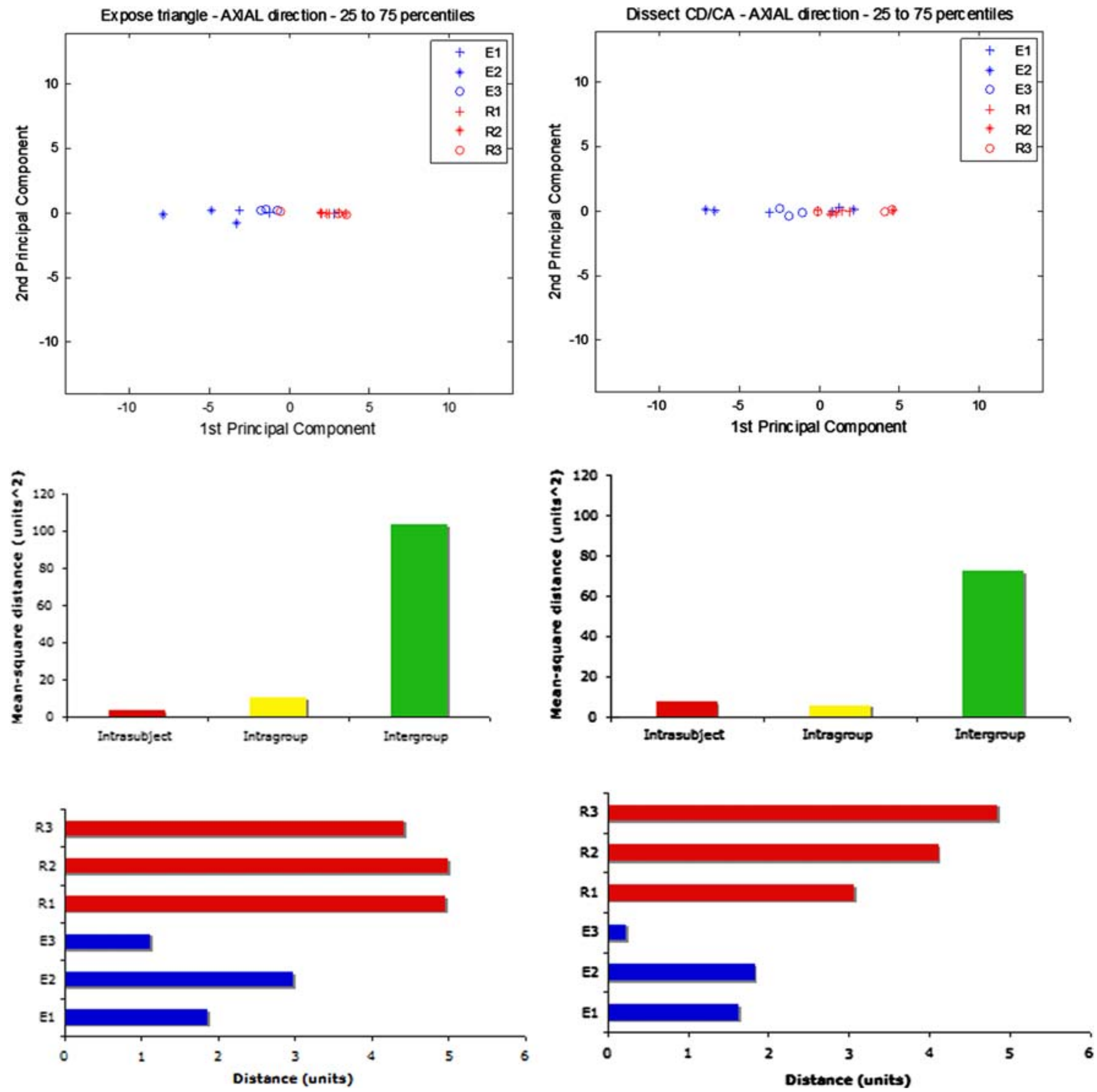

Fig. 2 *Top row*: Subtask-level principal components analysis (PCA) for the dominant hand for two key subtasks: "expose Calot's triangle" and "dissect cystic artery/cystic duct." Each point corresponds to a separate performance of the subtask by the indicated surgeon. *Middle row*: Contributions to variability in the PCA weight space due to intrasubject, intragroup, and intergroup variability. Variability is measured as the mean sum of squared distances from the global mean position. *Bottom row*: Distances from each subject's center to the experts' center used to compute the Mann–Whitney test ($p = 0.049$ for both subtasks)

the 10 elemental tool tip motions—push, pull, reach, orient, sweep, spread, grasp and hold, grasp and cut, idle, out—described in our earlier study [22]. We used manual video analysis to identify the start and end times of each action type and the resulting time stamps to segment the corresponding tool tip position data. Velocities, accelerations, and jerks were computed analytically using the spline approximants determined when the generalized cross validation procedure described earlier was applied.

Because each action typically was performed many times in executing a subtask, we modeled transitions between actions (states) using a semi-Markov model [29] and constructed empirical distributions of holding times, velocities, accelerations, and jerks in each state, together with a matrix representing the transitions between states. We refer to velocity, acceleration, and jerk profiles collectively as kinematic measures.

To compare performances between two subjects—resident (R) and expert (E)—performing a particular subtask, we defined V_R and V_E as the group of kinematics distributions for all executed actions, h_R and h_E as the group of holding time distributions for all executed actions, and T_R and T_E as the state transition probability matrices. We then computed difference measures for the kinematics

(V_R vs. V_E) and holding time (h_R vs. h_E) distributions using the Kolgomorov–Smirnov statistic (K-S) [30–32] and difference measures for state transitions using the Jensen–Shannon divergence statistic (JSD) [33–35]. These statistics are useful because they produce values ranging from 0 (no difference) to 1 (maximum possible difference) and thus can be compared directly without further normalization procedures. Across all subjects and executions, we obtained a $k \times w$ matrix of difference measures, where k corresponds to the number of subject comparisons (R_I vs. E_J) and w to the number of extracted performance measures.

$$\begin{bmatrix} R1 \text{ vs. } E1 \\ R1 \text{ vs. } E2 \\ \vdots \\ R3 \text{ vs. } E3 \end{bmatrix} = \left[\begin{array}{c|cccc|ccc} JSD_{R1\text{ vs. }E1} & D_{push}(h)_{R1\text{ vs. }E1} & \cdots & D_{out}(h)_{R1\text{ vs. }E1} & D_{push}(V)_{R1\text{ vs. }E1} & \cdots & D_{out}(V)_{R1\text{ vs. }E1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ JSD_{R3\text{ vs. }E3} & D_{push}(h)_{R3\text{ vs. }E3} & & D_{out}(h)_{R3\text{ vs. }E3} & D_{push}(V)_{R3\text{ vs. }E3} & & D_{out}(V)_{R3\text{ vs. }E3} \end{array}\right]$$

In our analysis, k equals 15 (9 comparisons of residents to experts and 2 sets of 3 intragroup comparisons, for example, R1 vs. R2, R1 vs. R3, and R2 vs. R3 for the residents and a similar set for the experts) and w equals 71 (1 JSD value for the transition matrix and 7 sets of 10 K-S values; each set of 10 values corresponds to the 10 different kinds of surgical actions; the first group of 4 sets represents summary K-S values comparing the average holding times, velocities, accelerations, and jerks across all visits to each action state, whereas the second group of 3 sets represent the K-S values comparing the detailed velocity, acceleration, and jerk distributions concatenated across all visits to each action state).

Given the large number of elements in each row, we again used principal components analysis to reduce the dimensionality of the difference matrix and a Mann–Whitney test to test the null hypothesis that there is no distinction between groups when subjects from the same skill level (R vs. R or E vs. E) are compared with subjects from different skill levels (E vs. R).

Results

In this section, the two variability analyses at the subtask and the action levels are reported to determine whether detectable differences in kinematic and time patterns exist across skill levels and which measures are most useful in separating surgeons along this spectrum when the motor performances of the dominant and nondominant hands are analyzed.

Variability analysis at the subtask level

The results of the PCA decomposition for the 25th to 75th percentile axial velocity data for the two subtasks analyzed are shown in Figs. 2 and 3. More than 99% of the variability in both plots is explained by the first principal component, which suggests that a relatively simple summary measure such as the average velocity may be sufficient to represent the variation at the subtask level and that more complex representations may be unnecessary. The results for the lateral and vertical directions are similar to those presented in Figs. 2 and 3 for the axial direction.

The ANOVA for both the "expose triangle" and "dissect CD/CA" subtasks in monitoring of the dominant hand showed a significant intergroup contribution to variability ($p = 0.03$ and 0.02, respectively), which indicates that the velocity measure is able to distinguish between residents and experts (Fig. 2). This result is consistent with the data points shown in the upper plot, in which the intrasubject repeatability is generally high and there is an apparent separation between the residents at the right side of the plot and the experts spread over a wider range on the left side of the plot. In contrast, the test failed to find such a distinction between the residents and experts in monitoring of the nondominant hand in both subtasks ($p = 0.09$ and 0.10, respectively) (Fig. 3).

Similarly, the Mann–Whitney test indicated that dominant hand performance provided group separation for both subtasks ($p = 0.049$) but that the nondominant hand did not ($p = 0.28$ for "expose triangle" and 0.52 for "dissect CD/CA").

Variability analysis at the action level

At the action level, the PCA analyses included 71 difference measures based on kinematics, holding times, and transition probabilities. We found that five principal components were required to represent 80% of the variability in

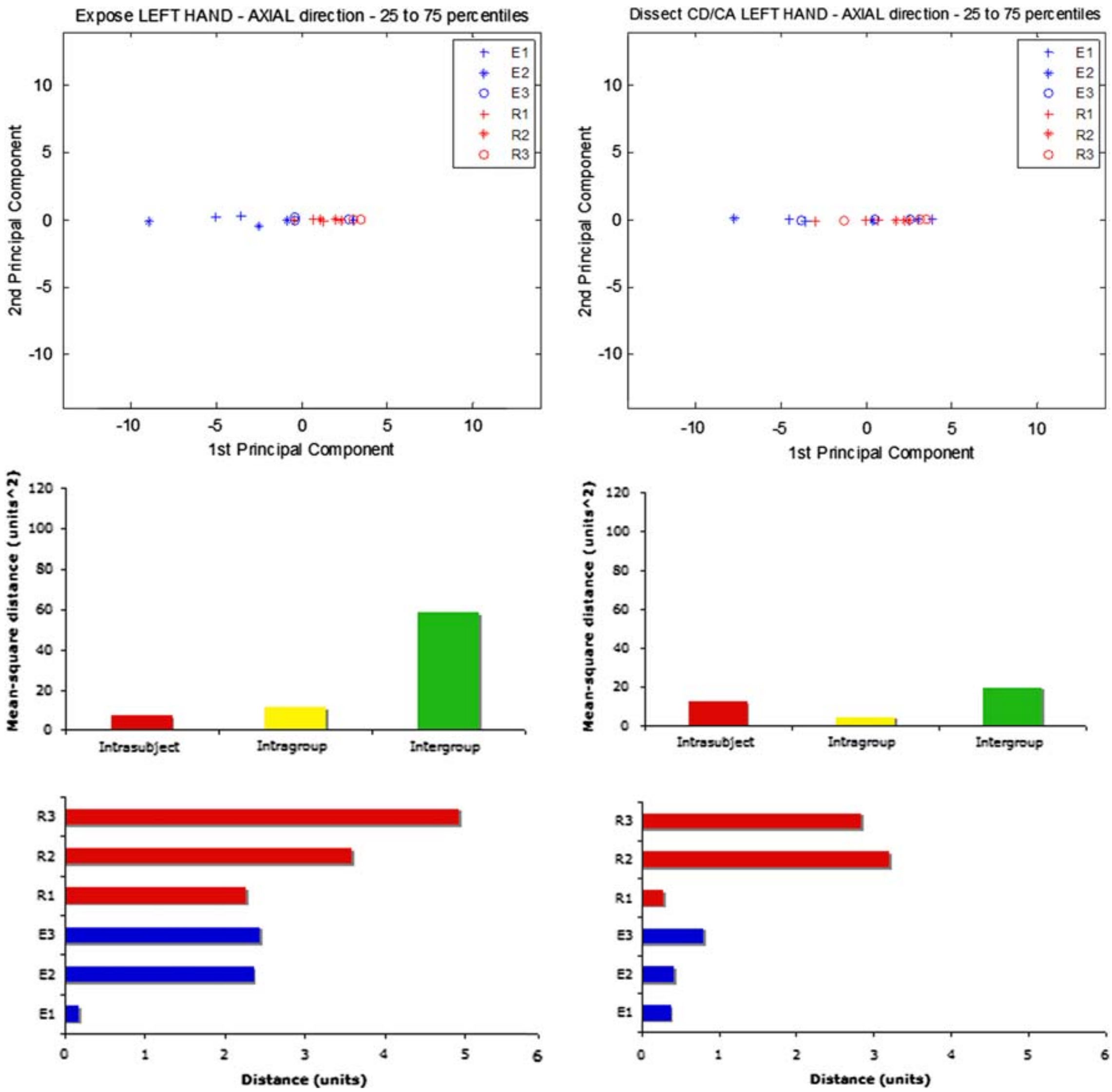

Fig. 3 *Top row*: Subtask-level principal components analysis (PCA) for the nondominant hand for two key subtasks: "expose Calot's triangle" and "dissect cystic artery/cystic duct." Each point corresponds to a separate performance of the subtask by the indicated surgeon. *Bottom row left*: Contributions to variability in the PCA weight space due to intrasubject, intragroup, and intergroup variability. Variability is measured as the mean sum of the squared distances from the global mean position. *Bottom row right*: Distances from each subject's center to the experts' center used to compute the Mann–Whitney test ($p = 0.28$ for "expose"; $p = 0.52$ for "dissect")

the original data. Figure 4 plots the first two principal components to facilitate graphic interpretation of the data.

The Mann–Whitney test indicated that performance during the "expose triangle" subtask ($p = 0.001$) clearly differentiates the set of comparisons among peers (i.e., expert with expert or resident with resident) from comparisons among subjects belonging to different groups (i.e., expert with resident), but not during the "dissect CD/CA" subtask ($p = 0.35$). This is somewhat different from the PCA analysis at the subtask level, in which the test is significant for separating between groups in both subtasks. However, there appears to be some intermingling of subjects from the two groups in the "dissect CD/CA" subtask, which could be related to the fact that no group differentiation was found at the action level.

Discussion

This study aimed primarily to evaluate the feasibility of using quantitative measurements obtained during live surgeries to assess motor aspects of surgical skill. The most important early test of a proposed assessment technique is whether it can reliably distinguish between subjects at the extreme ends of the skill development spectrum.

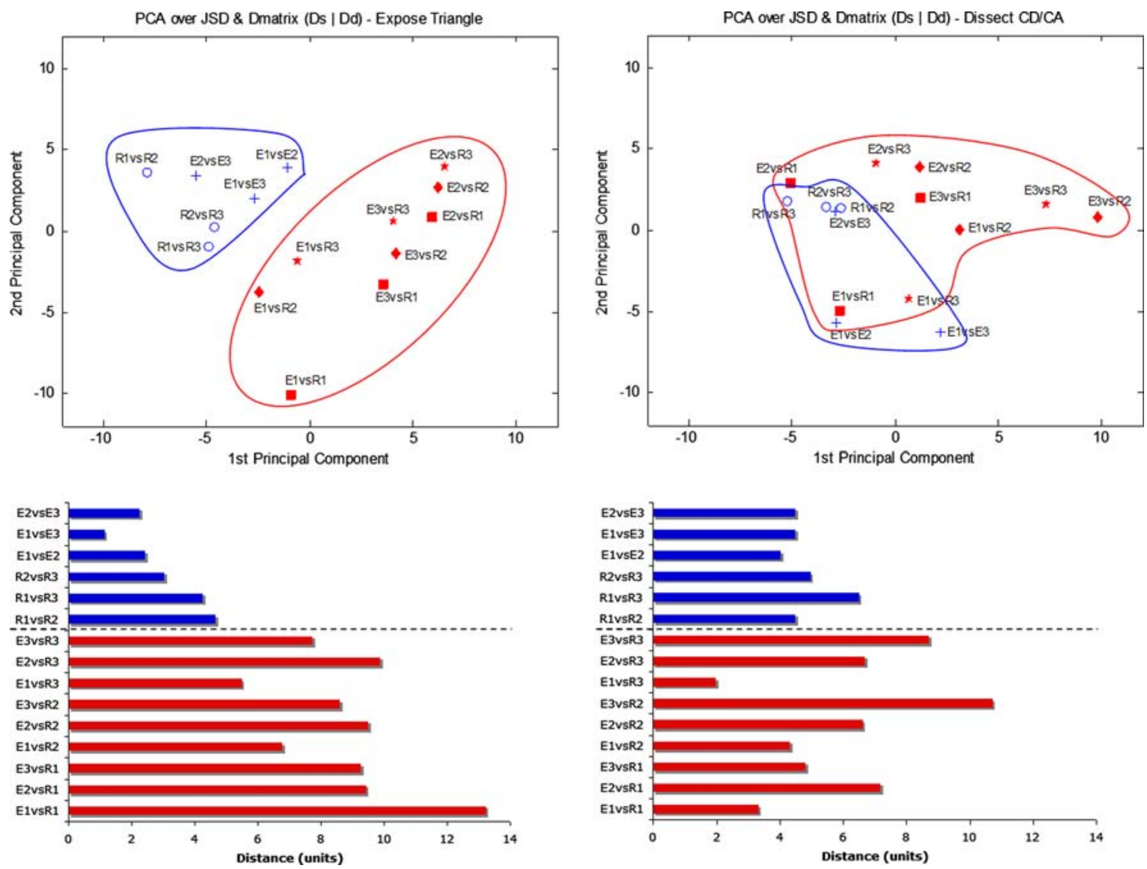

Fig. 4 *Top row*: Action level principal components analysis (PCA) for the dominant hand for the "expose Calot's triangle" and "dissect cystic artery/cystic duct" subtasks. Each point corresponds to a selected comparison of the action-level data between pairs of surgeons either at the same or different levels of skill development. *Bottom row*: Distances from each subjects' pair to the experts' center used to compute the Mann–Whitney test ($p = 0.001$ for "expose"; $p = 0.35$ for "dissect")

In this study, we demonstrated that it is feasible to acquire intraoperative position data using electromagnetic sensors and to process the data using manual video segmentation in the context of MCMD, which serves as a template for the procedure. The interprocedure variability seemed low enough to enable distinctions at both the subtask and action levels between resident and expert surgeons during monitoring of two surgical subtasks (i.e., "expose triangle" and "dissect CD/CA") judged by the expert surgeons to be among the tasks in the procedure most dependent on surgical motor skills. Although the nondominant hand behavior did not exhibit significant group separation (possibly because this hand plays a more "passive" role in simple retraction of the gallbladder), we conclude that a quantitative assessment technique based on tool position measurements is feasible, although considerable work must yet be done to make it a practical technique.

Some groups have examined the potential of using forces and torques exerted on the tools or between the tools and the tissues subjected to surgery to discriminate between skill levels [36]. Rosen et al. [37, 38] have done extensive work using force/torque signatures applied at the hand–tool interface to evaluate performance in a porcine model. These authors used Markov modeling to show the feasibility for correctly classifying surgeons into two experience levels based on the similarity of the models representing their low-level tool–tissue interactions to models derived from reference groups representing the two experience levels. They demonstrated that the forces and torques applied by experts and novices differed, as did the time to complete the procedures.

In the future, it likely would be valuable to combine both force and position measurements into a single assessment process. Our group has developed a curved dissector tool with an integrated force sensor and position sensor [39], but the instrumentation required involves more significant modifications to the surgical tools, so there are some practical obstacles to making these measurements intraoperatively on a routine basis.

Our current results suggest that objective tool position measurements may ultimately prove useful in assessing a resident's progress in developing surgical motor skills. To reach that goal, further work must be done to assess how many surgeries must be monitored to ensure that the contributions from interprocedure variability can be "averaged out." In addition, we need a better understanding of the relation between intrasubject variability and levels of difference that signify significant steps in skill development.

In this study, we evaluated the end points of the skill development spectrum for residents and experts. In the future, we aim to draw finer distinctions such as whether a given resident is keeping up with his or her year level or not. It also would be instructive to assess differences in performance between expert surgeons for a better understanding of different expert approaches to the same surgical goal. Such assessments could potentially lead to more nuanced discussions of what constitutes "best practice" for particular surgical tasks.

In addition, now that we have the ability to acquire quantitative intraoperative data on a surgeon's motor performance, we can revisit the question of the relationship between these intraoperative measurements and the same surgeon's performance on surgical simulators. By using the MCMD to isolate selected surgical tasks rather than by looking at an undifferentiated stream of data from a live surgical procedure, we may be able to establish stronger correspondences between selected surgical tasks and the corresponding simulations of these tasks. This may ultimately enable us to do validated assessments in a simulated setting. We also could use these comparison metrics to assess the degree to which specific skills acquired in a simulator setting can be transferred to the operative setting by comparing pre- and posttraining patterns in the simulator and the corresponding pre- and posttraining patterns in the live operating room. However, we would first need to ensure that the differences we investigate are larger relative to the intertrial or interprocedure variability.

Because this is a newly proposed assessment method, a considerable amount of further work is needed to make it clinically practical. Future work likely will focus on the following issues:

- Incorporating measurements from l-hook dissectors because they are a common alternative to the curved dissectors. Because surgeons apply cautery more frequently when using l-hook dissectors than when using curved dissectors, and because cautery interferes with electromagnetic tracking systems, it will be necessary to adapt or develop an optical tracking system or a fiberoptic-based system to replace the Polhemus sensor we have used in this study to avoid this cautery-related interference.
- Expanding the set of physical measures assessed by incorporating force sensing. The attending surgeons specifically recommended using force sensing to monitor whether the nondominant hand is providing appropriate traction to the anatomic structure being dissected or not, but force sensing also will likely contribute to improving the characterization of the dominant hand–tissue interactions. This will be a particularly challenging task for three main reasons: (1) force sensors usually are highly sensitive to electromagnetic interference such as that generated during cautery; (2) force sensors need to be placed in the load path, which typically requires modification of the surgical tools; and (3) external force sensors cannot differentiate between forces applied at the tool tip and trocar interaction forces, which may require the development of new force sensors.
- Automating the movement segmentation process to decrease the postprocessing burden of our current manual segmentation process. An interesting approach was proposed recently by Murphy [40] at Johns Hopkins University, who developed automatic techniques based on hidden Markov models to detect raw motion data from a surgical task and segment these into a labeled sequence of surgical gestures. Although this approach to date has been tested only in simulated environments, it would be worth exploring its capability to segment motion data from live surgeries as well.
- Using the method presented in this report to create a database of residents and expert surgeons executing multiple procedures on a variety of patients, coupled with data on the characteristics of each patient, to enable control or compensation for the innate difficulty of each procedure. Such a database would enable us to store sufficient data to test the method's value in drawing finer distinctions along a surgeon's training arc, such as in monitoring year-to-year development.
- Expanding our method to other types of surgeries to form the basis of a comprehensive skill assessment process.

We do not believe that objective assessment methods ever could or should be used to replace the more subjective and nuanced feedback from experienced surgical instructors during the training process. However, we do believe such methods can serve a useful complementary role as an unbiased starting point for evaluation based on quantitative metrics. If such methods can discriminate between skill levels and detect subtle differences in surgical technique, they may someday be used to provide specific, helpful, and relevant feedback to trainees concerning areas in which improvement is needed.

References

1. Bridges M, Diamond D (1999) The financial impact of teaching surgical residents in the operating room. Am J Surg 177:28–32

2. Babineau T, Becker J, Gibbons G, Sentovich S, Hess D, Robertson S, Stone M (2004) The “cost” of operative training for surgical residents. Arch Surg 139:366–370
3. Britt L (2007) Simulation training: what are real questions that must be answered? Am J Surg 194:220
4. Torkington J, Smith SGT, Rees BI, Darzi A (2001) The role of the basic surgical skills course in the acquisition and retention of laparoscopic skill. Surg Endosc 15:1071–1075
5. Gallagher AG, Lederman AB, McGlade K, Satava RM, Smith CD (2004) Discriminative validity of the minimally invasive surgical trainer in virtual reality (MIST-VR) using criteria levels based on expert performance. Surg Endosc 24:660–665
6. Scott D, Rege R, Bergen P, Guo W, Laycock R, Tesfay S, Valentine J, Jones D (2000) Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 10:183–190
7. Derossis AM, Bothwell J, Sigman HH, Fried GM (1998) The effect of practice on performance in a laparoscopic simulator. Surg Endosc 12:1117–1120
8. Rosser JC Jr, Rosser LE, Savalgi RS (1997) Skill acquisition and assessment for laparoscopic surgery. Arch Surg 132:200–204
9. Feldman LS, Sherman V, Fried GM (2004) Using simulators to assess laparoscopic competence: ready for widespread use? Surgery 135:28–42
10. Park AE, Witzke D (2002) The surgical competence conundrum. Surg Endosc 16:555–557
11. Grantcharov TP, Rosenberg J, Pahle E, Funch-Jensen P (2001) Virtual reality computer simulation. Surg Endosc 15:242–244
12. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P (2004) Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 91:146–150
13. Moorthy K, Munz Y, Sarker S, Darzi A (2003) Objective assessment of technical skills in surgery. BMJ 327:1032–1037
14. Wanzel KR, Ward M, Reznick RK (2002) Teaching the surgical craft: from selection to certification. Curr Probl Surg 39:574–659
15. Alleman A (2005) Have you wondered about your colleague’s surgical skills? Am J Med Qual 20:78–82
16. Hammond I (2006) Training, assessment, and competency in gynaecologic surgery. Best Pract Res Clin Obstet Gynaecol 20:173–187
17. Darzi A, Datta V, Mackay S (2001) The challenge of objective assessment of surgical skill. Am J Surg 181:484–486
18. Datta V, Bann S, Mandalia M, Darzi A (2006) The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg 192:372–378
19. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, Bello F, Darzi A (2007) An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245:992–999
20. Smith CD, Farrell TM, McNatt SS, Metreveli RE (2001) Assessing laparoscopic manipulative skills. Am J Surg 181: 547–550
21. Hwang H, Lim J, Kinnaird C, Nagy AG, Panton ON, Hodgson AJ, Qayumi KA (2005) Correlating motor performance with surgical error in laparoscopic cholecystectomy. Surg Endosc 20:651–655
22. Cristancho S, Hodgson A, Panton N, Meneghetti A, Qayumi K (2006) Assessing cognitive and motor performance in MIS for training & tool design. Stud Health Technol Informatics 119:108–113
23. Kologlu M, Tutuncu T, Yuksek YN, Gozalah U, Daglar G, Kama NA (2004) Using a risk score for conversion from laparoscopic to open cholecystectomy in resident training. Surgery 135:282–287
24. Cristancho S, Hodgson A, Panton N, Meneghetti A, Qayumi K (2007) Feasibility of using intraoperatively acquired quantitative kinematic measures to monitor development of laparoscopic skill. Stud Health Technol Informatics 125:85–90
25. Milne AD, Chess DG, Johnson JA, King GJW (1996) Accuracy of an electromagnetic tracking device: a study of the optimal operating range and metal interference. J Biomech 29:791–793
26. Polhemus Incorporated (2002) 3Space Fastrak user’s manual. OPM00PI002. Colchester, VT
27. Dohrmann C, Busby H, Trujillo D (1988) Smoothing noisy data using dynamic programming and generalised cross-validation. J Biomech Eng 110:37–41
28. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
29. Haverkort B (2001) Markovian models for performance and dependability evaluation: lectures on formal methods and performance analysis, 1st edn. Springer, Berlin, pp 38–83
30. Drew J, Glen A, Leemis L (2000) Computing the cumulative distribution function of the Kolmogorov–Smirnov statistic. Comput Stat Data Anal 34:1–15
31. Von Mises R (1964) Mathematical theory of probability and statistics. Academic Press, New York
32. Hodgson AJ, McBeth PB (2002). Comparing motor performance on similar tasks in different settings: statistical characteristics of a nondimensional difference measure. Internal Document, Neuromotor Control Laboratory, Department of Mechanical Engineering, University of British Columbia
33. Dagan I, Lee L, Pereira F (1999) Similarity-based models of word cooccurrence probabilities. Mach Learn 34:43–69
34. Majtey AP, Lamberti PW, Prato DP (2005) Jensen–Shannon divergence as a measure of distinguishability between mixed quantum states. Phys Rev A 72:052310-1–052310-6
35. Endres D, Schindelin J (2003) A new metric for probability distributions. IEEE Trans Inf Theory 49:1858–1860
36. de Visser H, Heijnsdijk EA, Herder JL, Pistecky PV (2002) Forces and displacements in colon surgery. Surg Endosc 16:1426–1430
37. Rosen J, Hannaford B, Richards G, Sinanan M (2001) Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng 48:579–591
38. Rosen J, Brown J, Chang L, Sinanan M, Hannaford B (2006) Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Trans Biomed Eng 53:399–413
39. Kinnaird K (2004) A multifaceted quantitative validity assessment of laparoscopic surgical simulators. Master’s Thesis, Department of Mechanical Engineering, University of British Columbia, Vancouver
40. Murphy T (2004) Towards objective surgical skill evaluation with hidden Markov model-based motion recognition. Master’s Thesis, Johns Hopkins University, Baltimore, MD