Clin Orthop Relat Res (2008) 466:963–968
DOI 10.1007/s11999-008-0115-9

ORIGINAL ARTICLE

Haptic Feedback Can Provide an Objective Assessment of Arthroscopic Skills

George Chami MD, MRCS, MS(Orth),
James W. Ward, Roger Phillips MSc, PhD, CEng, FBCS, CPIT,
Kevin P. Sherman BMBCh(Oxon), FRCS, PhD

Received: 19 October 2006 / Accepted: 27 December 2007 / Published online: 23 January 2008

Abstract The outcome of arthroscopic procedures is related to the surgeon's skills in arthroscopy. Currently, evaluation of such skills relies on direct observation by a surgeon trainer. This type of assessment, by its nature, is subjective and time-consuming. The aim of our study was to identify whether haptic information generated from arthroscopic tools could distinguish between skilled and less skilled surgeons. A standard arthroscopic probe was fitted with a force/torque sensor. The probe was used by five surgeons with different levels of experience in knee arthroscopy performing 11 different tasks in 10 standard knee arthroscopies. The force/torque data from the hand and tool interface were recorded and synchronized with a video recording of the procedure. The torque magnitude and patterns generated were analyzed and compared. A computerized system was used to analyze the force/torque signature based on general principles for quality of performance using such measures as economy in movement, time efficiency, and consistency in performance. The results showed a considerable correlation between three haptic parameters and the surgeon's experience, which could be used in an automated objective assessment system for arthroscopic surgery.
Level of Evidence: Level II, diagnostic study. See the Guidelines for Authors for a complete description of levels of evidence.

Each author certifies that he has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.
Each author certifies that his institution has approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent was obtained.

G. Chami, J. W. Ward, R. Phillips
Department of Computer Science, University of Hull, Hull, UK

G. Chami, K. P. Sherman
Hull and East Yorkshire Hospitals, NHS Trust, Hull, UK

G. Chami (✉)
Orthopaedic Department, Scunthorpe General Hospital, Scunthorpe, DN15 7BS, UK
e-mail: georgechami@doctors.org.uk

Introduction

Knee arthroscopy is a standard operation in orthopaedic surgery. Mastering the technique is demanding and has a steep learning curve [5]. There currently is no objective assessment system for such a common operation. The current methods of assessment are based on direct observation of the trainee by a trainer, which is time-consuming and subjective by its nature. Furthermore, the current assessment methods do not take into account the forces applied through the instruments, which are important elements in determining the quality of such operations (in which iatrogenic damage to the articular cartilage is easily possible). Training of such skills in a risk-free environment was explored by using a virtual arthroscopy trainer [9, 10] that incorporates an objective scoring system [9]. Such training methods have not been widely used to date because of cost and uncertainty about whether the skills obtained in virtual environments can be transferred to actual surgical experience [1, 4]. Force patterns in real arthroscopy, using an instrumented arthroscopic probe, were analyzed in a previous publication [2]. The authors noted a difference in the torque magnitude signal between a skilled surgeon and a trainee. However, a literature search revealed no published studies that measured, or used, the important element of

force patterns applied on the instruments to assess the quality of performance in real arthroscopic surgery.

The primary aim of this study was to assess whether measurements of the forces that surgeons apply through the arthroscopic instruments can be used to assess performance. In other words, are there any differences in the force patterns that are generated on the arthroscopic instruments by experienced versus inexperienced surgeons that can be used to assess performance?

The secondary aim was to identify whether the main features in the force/torque signature (efficiency of movement, efficiency in time, and consistency in performance) could be used to develop an automated objective system to assess and compare performance between subjects during knee arthroscopy. We also wished to determine which of these features offered the most potential for objective assessment.

Materials and Methods

A force/torque sensor was fitted onto the base of the handle of a standard knee arthroscopic probe. The sensor is capable of measuring the range of forces that are applied through the instruments (50 N force and 500 mNm torque). The sensor obtained 170 measurements per second of force parameters (Fx, Fy, Fz) and torque parameters (Tx, Ty, Tz). We obtained the necessary approvals for the instrument to be used in real surgery for research purposes, including (1) obtaining ethical committee approval, (2) testing the instruments for electrical safety as per HEI 95 (Code of Practice for Acceptance Testing of Medical Electrical Equipment), (3) meeting Medical Devices Agency Guidelines in the UK, and (4) modifying the sensor to meet safety and sterility requirements.

The instrument was used in 10 knee arthroscopic procedures by five surgeons (two experts, three trainees) who had different levels of experience in knee arthroscopy; the two experts (orthopaedic consultant and experienced staff grade specialist) had both performed more than 2000 knee arthroscopic procedures, whereas each of the junior trainees had performed fewer than 50 procedures. The diagnostic part of the procedure was standardized by dividing it into 11 separate tasks in the three compartments of the knee (Table 1). The expertise of the surgeons was ranked from 1 to 5 based on the number of knee arthroscopies performed (from five to more than 2000 knee arthroscopies). The instrument's sensor was zeroed before performing each task. The video image obtained through the standard arthroscopic camera was recorded. The data were imported and mathematically analyzed by computer, rather than by an observer, to remove any human bias in the assessment process as described subsequently.

The torque magnitude (TM) was used as the objective measure because it provides the maximum variation in the signal because of the long lever arm of the probe. Torque magnitude was calculated as the length of the vector (Tx, Ty, Tz) for the recorded results and plotted on a graph. This graph then was overlaid on the video recording of each task (Fig. 1).

The signal for each task consisted of two main components, which we termed navigation component and task component. The navigation component consisted of numerous features or peaks that corresponded to the movement of the instrument in the knee to reach the target (articular cartilage, meniscus, or ligament). The task component consisted of features that corresponded to the maneuvers for performing the task, such as probing the cartilage or meniscus.

The TM for both types of features in each part of the knee was measured manually as described previously [2]. Although the results showed the TM for both features varied in the three compartments of the knee, there was a wide difference between the task and navigation TM features; their mean was 140 mNm versus 46 mNm, respectively. It was noted the friction between the probe

Table 1. Summary of tasks performed during a knee arthroscopy session

Number	Task description
T01,T02	Medial compartment: probe femoral condyle and tibial plateau (press twice on each surface in three points)
T03,T04	Medial meniscus: continuous run, then probe three points twice; probe inferior surface by continuous run, then lift margins twice
T05	Anterior cruciate ligament: feel the edges, then pull twice
T06,T7	Lateral compartment: probe femoral condyle, tibial plateau (press twice on each surface in three points)
T08,T10	Lateral meniscus: continuous run, then probe three points twice; probe inferior surface by continues run, then lift margins twice
T09	Popliteal hiatus: pull twice
T11	Patellar femoral joint: probe at three points; press each twice

Reprinted with permission from Chami G, Ward J, Wills D, Phillips R, Sherman K. Smart tool for force measurements during knee arthroscopy: in vivo human study. *Stud Health Technol Inform.* 2006;119:85–89, Copyright (2006), with permission from IOS Press.

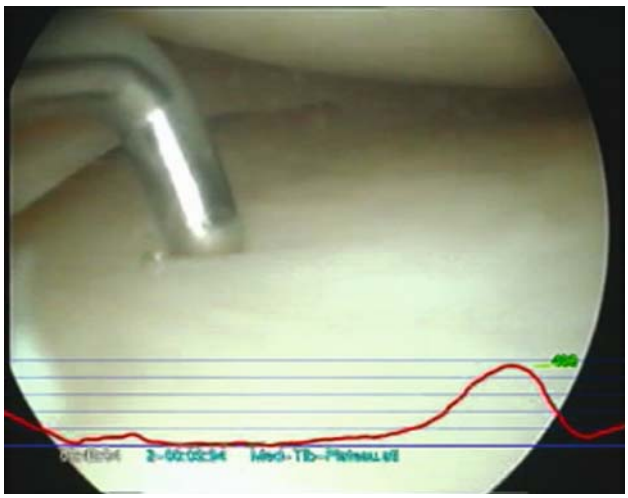

Fig. 1 A force/torque graph is overlaid on a video recording. The graph scrolls right to left in time with the video recording and shows peak values. Reprinted with permission from Chami G, Ward J, Wills D, Phillips R, Sherman K. Smart tool for force measurements during knee arthroscopy: in vivo human study. *Stud Health Technol Inform.* 2006;119:85–89. Copyright (2006), with permission from IOS Press.

and the soft tissue generated peaks with a maximum TM of 15 mNm. The previous observations suggested TM features could be used to automatically identify the origin of the feature, which could be originating from task performance (task features), from an irregular navigation pattern (human-induced noise), or from friction with the soft tissue mainly at the portal entry site (mechanically induced noise).

To assist processing of the measurements, the signal was first passed through a 0- to 4-Hz band pass filter (Hanning window) using SIGVIEW v 1.95 (SignalLab, Pforzheim, Germany). The range of the band was selected because human movement is unlikely to generate a complete task feature of less than 0.25 seconds in duration. Filtering of the signal had a minimal effect on the TM while automatically removing part of the mechanically induced noise (Fig. 2). Second, the filtered signal was analyzed using Excel® 2003 (Microsoft Corp, Redmond, WA). The peaks and troughs in the signal were identified (Table 2, Columns C, D) in which a feature in the signal was identified by one peak between two troughs. The TM of each feature then was calculated by subtracting from the magnitude of the peak the average of the minima of the surrounding two trough magnitudes (Table 2, Column F). The duration for each feature was calculated by the interval between the two minima of the surrounding troughs of that feature (Table 2, Column J). Manual analysis of the results showed approximately 95% of procedure features had a TM greater than 90 mNm, whereas approximately 94% of the features present in the navigation component were less than 90 mNm. The features were categorized based on their magnitude into one of three types, namely task features (greater than 90 mNm), human-induced noise features (15–90 mNm), and mechanically induced noise features (0 to less than 15 mNm) using equations in an Excel® spreadsheet (Table 2, Columns G–I). The identified task features were correlated with the video recording, and in comparison with manual classification, the automatic process was able to correctly identify approximately 96.2% (484 hits identified from a total of 503) of the task features in all groups of users. The percentage of task features versus human-induced noise features was calculated to indicate the efficiency in movements. Efficiency of movement was analyzed by comparing the navigation signal across the users. Efficiency of movement was calculated by analyzing the TM signal for each user using the signal-to-noise ratio. The calculation was done by dividing the number of task features available in the signal by the number of human-induced noise features after removing the mechanically induced noise features and then presented as a percentage. The number of task features versus total time was calculated to indicate the efficiency in time. Time efficiency was calculated as the time needed to produce one task feature by dividing the number of task features by the total time of the procedure. This method was used because two trainees did not complete some of the more difficult tasks, such as T08 or T09 (Table 1); thus, using total procedure time as the metric would have introduced bias. Two additional calculations to indicate the consistency of performance were performed by calculating the standard deviation of the task feature durations and standard deviation of the TM for the task features, respectively. Additional description of the rationale for using these features is included in the Results section.

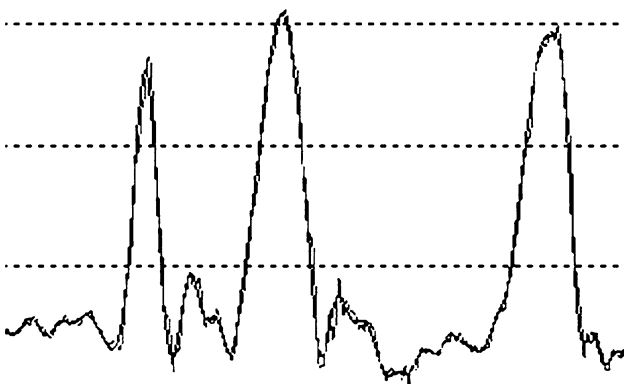

Fig. 2 The image shows an overlay of the filtered signal (band pass, 0- to 4-Hz) (dark line) and the unfiltered signal.

The Spearman's rho coefficient was used to test the correlation between the results and the surgeons' experience using SPSS 13.0 for Windows (SPSS Inc, Chicago, IL).

Table 2. Sequence of equations used in Excel® spreadsheet to automate the analysis process

Row numbers in Excel®	Time in seconds	Torque magnitude	Troughs	Peaks	Peaks and troughs	Peak magnitude	Task feature	Human-induced noise	Mechanically induced noise	Duration of task features
1	A	B	C	D	E	F	G	H	I	J
2	Time	$=\sqrt{(Tx^2 + Ty^2 + Tz^2)}$	=IF(AND (B2 < B1, B2 < B3), B2,"")	=IF(AND (B2 > B1, B2 > B3), B2,"")	"Copy Peaks column D and Troughs Column C"	=E2−((E1 + E3)/2)	=IF(OR (F2 > 90, F2 = 90), F2,"")	=IF(AND (F2 < 90, F2 > 15), F2,"")	=IF(OR (F2 < 15, F2 = 15), F2,"")	=IF(G2 = "","", A3-A1)

Description of the function of each Excel® spreadsheet cell formula: formula in column B: length of the vector (Tx, Ty, Tz), which is recorded from the force torque sensor; C: search for minima of troughs; D: search for peaks; E: copy and paste results from Columns D and C to here; F: calculate the magnitude of each feature by subtracting the magnitude of the peak form the average of the two surrounding minima of the troughs; G: categorize as task feature if the peak magnitude is more than 90 mNm; H: categorize as human-induced noise feature if the peak magnitude is between 15 and 90 mNm; I: categorize as mechanically induced noise feature if the peak magnitude is less than 15 mNm, J: calculate the duration of each task feature.

Results

Our analysis showed numerous differences between experienced and less experienced surgeons could be detected by the apparatus. Experts showed less noise per task feature and took less time to complete tasks than trainees ($p < 0.01$) (Fig. 3, Table 3). The navigation signal generated by a skilled surgeon had less noise (defined by peaks in the graph during navigation) compared with the trainee, and the magnitude of these human-induced noise features was less than the task features; the skilled surgeons produced task features that were executed on a regular interval with minimal navigation time between the features (Fig. 3).

Experts produced 215% to 310% of noise features for every task feature (approximately two to three noise features are generated for each task feature), whereas trainees produced 399% to 539% (Table 3). Spearman's rho correlation coefficient showed a significant correlation of 1.000 with significance set at 0.01 between the two variables. The result of time efficiency analysis showed experts spent 2 and 3.2 seconds (for Experts 1 and 2, respectively) in navigation and task time to produce one task feature, whereas trainees required 3.35 to 4.66 seconds (Table 3). Again, Spearman's rho correlation coefficient showed a significant correlation of 1.000 with significance set to 0.01. To test consistency in performance; other calculations, such as standard deviation in the TM and the duration for each task feature, were studied and tested against the surgeon's experience. Spearman's rho correlation coefficient showed poor correlation (−0.2) between the expertise of the surgeon and standard deviation in the duration of task feature; however, there was good correlation (−0.974) of the standard deviation in the task feature magnitude (Table 3).

Discussion

We investigated whether the forces applied through the surgical instruments could be used to assess the quality and skill of arthroscopic surgeons. We also sought to identify whether the main features in the force torque signature (efficiency of movement, efficiency in time, and consistency in performance) could be used to develop an automated objective system to assess and compare performance between subjects during knee arthroscopy. Our results indicate it is possible to assess experience objectively using the surgeon's dexterity focusing on parameters extracted from instrument force data. Each of the identified parameters lends itself to automation.

There are some pitfalls associated with the system developed in this study. The number of subjects in the study is small but probably is sufficient to show the

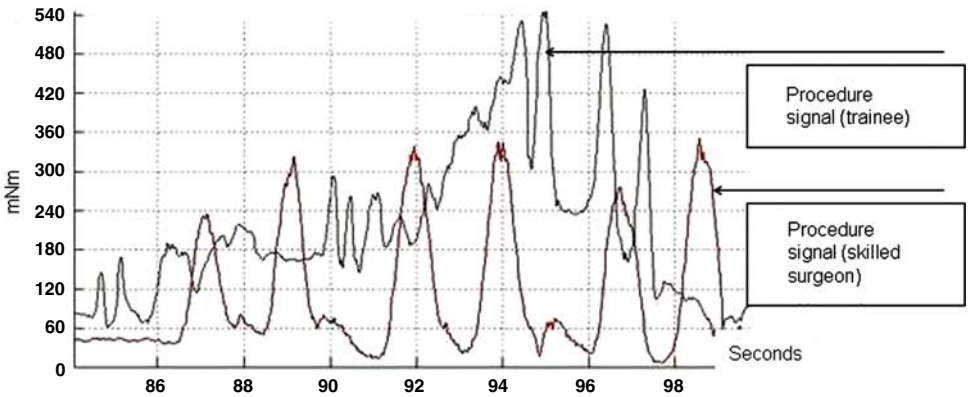

Fig. 3 Procedure signals generated by a skilled surgeon and trainee for the same task (T06) are shown. The skilled surgeon generates more consistent task features and less noise between features. Reprinted with permission from Chami G, Ward J, Wills D, Phillips R, Sherman K. Smart tool for force measurements during knee arthroscopy: in vivo human study. *Stud Health Technol Inform.* 2006;119:85–89. Copyright (2006), with permission from IOS Press.

Table 3. Results of the variable used to assess performance and Spearman's rho correlation test for each set of results against the surgeon's experience

Surgeon	Noise/Task features (%)	Task features versus total time (seconds)	Standard deviation of torque magnitude for task features (mNm)	Standard deviation of task feature duration (seconds)
1 (expert)	215.53	2.01	86.32	0.37
2 (expert)	310.33	3.22	81.06	0.37
3 (trainee)	375.03	3.35	100.09	0.33
4 (trainee)	399.76	4.17	96.29	0.28
5 (trainee)	539.13	4.66	57.16	0.26
Spearman's rho correlation coefficient	1.000*	1.000*	−0.200	−0.974*

*Correlation is significant at the 0.01 level (two-tailed).

potential for the technique. The system lacks assessment of some factors in knee arthroscopy, such as observation of all parts of the knee. However, this system currently could be useful as a supplement for full assessment. Additional work remains to be done to validate the system by assessing the force patterns of a greater number of users performing more knee arthroscopies and to study the learning curve of trainees.

There is no published system for assessing training in real knee arthroscopy for comparison with our results. The expert surgeons produced considerably fewer human-induced noise features for each task feature, suggesting experts were able to perform the required task with far fewer collision incidences between the probe and the anatomic structures during navigation. This suggested there were fewer hit-and-miss trials to reach the target, which in turn produced more efficient movements. Furthermore, the experts were more efficient in using the operative time.

It appeared experts used a larger variety of force magnitudes than the trainees to perform the task. This could be the result of the expert's ability to adjust the force on the instrument according to the needs of the procedure, for example, using more force to probe the meniscus to confirm its integrity while using less force to probe the articular cartilage to minimize iatrogenic damage.

A few trials for performance assessment have been published in laparoscopic cholecystectomy using force data obtained from an instrumented laparoscopic grasper equipped with a three-axis force/torque sensor at the hand-tool interface with similar objectives as this study. Markov's statistical model has been used to compare the performance of an expert with that of novices [3, 6–8]. The Markov approach was able to produce one figure that indicated how close the performance of a novice was to an expert's performance, ie, a performance index. It was able to correctly classify 87.5% of the surgical procedures into correct categories (either novices or experts), but scoring of different components of skills, such as efficiency in time or movement, was not obtained. Our approach was able to provide assessment of different skill components, which

could provide useful feedback to the trainee to identify potential areas for improvement and training. The Markov model used to assess performance in laparoscopic surgery required the procedure to be performed in a similar order and method as that used by the expert surgeon (whose movements were modeled and used as the standard for comparison). Thus, the better the match between these two performances was, the higher the score was. The study showed this method was very successful in predicting the user's level of expertise, but it is not yet clear how this technique would affect the outcome of the assessment for another expert surgeon who used different steps or maneuvers in the same task. Contrastingly, our system can compare skills with a perfect performance (no human-induced noise and minimal navigation time) rather than comparing it with a prerecorded procedure having a set order of tasks performed by an expert, making it potentially useful for making an objective assessment of performance for all surgeons without the need to follow a particular order of steps.

We presented a new method and defined parameters or features to assess performance during real knee arthroscopy. The identified parameters correspond to the existing general consensus of good arthroscopic skills. These parameters can be recorded and analyzed completely by a computer rather than by a human observer, supporting the potential for an objective scoring system for assessment of performance in real knee arthroscopy based on the pattern of forces and torques applied to the arthroscopic instruments. This could supplement formal traditional methods of assessment, in real knee arthroscopies or a knee simulator for arthroscopic training, and could provide instant feedback to trainees regarding dexterity and time efficiency without the presence of a surgeon trainer.

Acknowledgments We thank the Department of Computer Science at Hull University and the operating room staff at Castle Hill Hospital in facilitating the study. We thank all who participated in the study and data collection.

References

1. Bliss JP, Hanner-Bailey HS, Scerbo MW. Determining the efficacy of an immersive trainer for arthroscopy skills. *Stud Health Technol Inform*. 2005;111:54–56.
2. Chami G, Ward J, Wills D, Phillips R, Sherman K. Smart tool for force measurements during knee arthroscopy: in vivo human study. *Stud Health Technol Inform*. 2006;119:85–89.
3. Dosis A, Bello F, Gillies D, Undre S, Aggarwal R, Darzi A. Laparoscopic task recognition using Hidden Markov Models. *Stud Health Technol Inform*. 2005;111:115–122.
4. McCarthy A, Harley P, Smallwood R. Virtual arthroscopy training: do the "virtual skills" developed match the real skills required? *Stud Health Technol Inform*. 1999;62:221–227.
5. Miller WE. Learning arthroscopy. *South Med J*. 1985;78:935–937.
6. Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE Trans Biomed Eng*. 2001;48:579–591.
7. Rosen J, Solazzo M, Hannaford B, Sinanan M. Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions. *Stud Health Technol Inform*. 2001;81:417–423.
8. Rosen J, Solazzo M, Hannaford B, Sinanan M. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. *Comput Aided Surg*. 2002;7:49–61.
9. Sherman KP, Ward JW, Wills DP, Mohsen AM. A portable virtual environment knee arthroscopy training system with objective scoring. *Stud Health Technol Inform*. 1999;62:335–336.
10. Sherman KP, Ward JW, Wills DP, Sherman VJ, Mohsen AM. Surgical trainee assessment using a VE knee arthroscopy training system (VE-KATS): experimental results. *Stud Health Technol Inform*. 2001;81:465–470.