Wireless Body Sensor for Objective Assessment of Surgical Performance on a Standardised FLS Task

Georgina S.J. Kirby*
The Hamlyn Centre & McLaren Applied Technologies Ltd,
United Kingdom
georgina.kirby@mclaren.com

Richard M. Kwasnicki
The Hamlyn Centre,
Imperial College London,
United Kingdom
rmk107@ic.ac.uk

Caroline Hargrove
McLaren Applied Technologies Ltd,
United Kingdom
caroline.hargrove@mclaren.com

Jonathan L. Rees
University of Oxford,
United Kingdom
jonathan.rees@ndorms.ox.ac.uk

Mikael H. Sodergren
The Hamlyn Centre,
Imperial College London,
United Kingdom
m.sodergren@ic.ac.uk

Guang-Zhong Yang
The Hamlyn Centre,
Imperial College London,
United Kingdom
g.z.yang@ic.ac.uk

Benny P.L. Lo
The Hamlyn Centre,
Imperial College London,
United Kingdom
benny.lo@ic.ac.uk

ABSTRACT

Advances in Body Sensor Networks have prompted increasing numbers of low cost, miniaturised sensors being used in many different applications with one being the capture of hand movement data for surgical skills assessment. Despite these advances, existing assessment techniques are still predominantly subjective and resource demanding. Combining surgical training with a reliable objective assessment technique would ensure that trainees are correctly evaluated and credentialed as they progress through their training hence, ensuring competence and reducing critical medical errors.

This paper proposes the use of wearable, wireless inertial sensors for capturing motion data and enabling objective assessment of trainee surgeons' performance in carrying out one of the FLS (Fundamentals of Laparoscopic surgery) tasks; the peg transfer. A novel approach has been developed for the segmenting of specific peg movements enabling performance to be measured entirely objectively.

The features derived from the whole task as well as features for each of the segmented movements were analysed through unsupervised machine learning algorithms to look for useful measures of performance as well as patterns to identify differences between expert and trainee performance. Encouraging results in the peg transfer task, where a successful classification of expertise was obtained for all participants against gold standard assessment, prompt further investigation into the development of advanced performance metrics for a wider range of surgical training tasks.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition; I.2.3 [**Artificial Intelligence**]: Support Vector Machine; K.3.2 [**Computer and Information Science Education**]: Objective assessment tool

BODYNETS 2014, September 29-October 01, London, Great Britain

DOI 10.4108/icst.bodynets.2014.257019

General Terms

Algorithms, Experimentation, Measurement, Performance, Human Factors

Keywords

Surgical dexterity, Machine learning, Body Sensor Networks, Inertial Sensors

1. INTRODUCTION

Competence in surgery is a combination of many factors, more specifically: surgical knowledge, patient care, communication, decision-making and manual dexterity [2]. Minimally Invasive Surgery (MIS) requires a high level of technical skill and complex control of instruments where vision and mobility is restricted, making hand eye coordination difficult, along with the introduction of the fulcrum effect [7]. Surgeons are likely to underperform if they lack manual dexterity.

The assessment of technical skills is required for ongoing training, selection to specialist surgical programmes, reaccreditation and further research such as measuring the effect of training interventions [8]. For those especially gifted or skilled trainees, they can accelerate to more advanced training levels [1], such that the programme is more competency-based rather than time-based. Current options for assessment range from timing the duration of a task, to peer review of performance using the OSATS (Objective Structured Assessment of Technical Skill), to more objective, measurable outcomes such as the use of the ICSAD (Imperial College Surgical Assessment Device) or ADEPT (Advanced Dundee Endoscopic Psychomotor Tester).

Limitations exist for all methods with timing often being too simple and not reflective of quality and expert review being highly resource demanding and subjective. Although ICSAD and ADEPT provide measures of dexterity, the equipment is bulky and expensive making it prone to errors, restrictive and unnatural to the user, limiting their practical use and transfer to the operating room [3], [10]. Wireless sensor gloves are capable of measuring kinematics of finger joints and range of motion, which can also be used to assess hand movements. The accuracy of this system is affected by the fit of the glove to the surgeon's hands and sensors still need to be miniaturised to advance the sensor glove technology [5].

Virtual reality systems have more recently become available for laparoscopic training. These systems, however, are a costly form of skill acquisition compared to bench models and are still not easily accessible for all trainees. Apart from the large cost involved with these systems, another limitation is that the evidence for the transfer of skills from VR simulators to the operating room remains weak. This is partly due to the fact that systems to assess skill in the operating room lack operative technique standardisation and also objectivity [1].

Despite all of the technological advances discussed above, existing assessment techniques are still relatively subjective and ensuring assessment becomes more of an objective performance measure would be a fairer and more effective way of progressing trainees and accrediting practicing surgeons [6].

In order to address the challenges outlined above, the aim of this work is to assess the feasibility of using a minimal number of small, lightweight, wireless, body worn sensors to measure key hand movements and hence, assess the performance of the user. The improvements from existing systems include the reduction in size and cost, improved usability of the whole system (hardware and software), new technical approaches to segregate hand movements from the set tasks and measure task parameters and an entirely objective assessment of a participant's performance on set surgical tasks.

2. METHODS

2.1 Study Setup and System Specifications

To capture hand movement, wireless IMU (Inertial Measurement Unit) sensors, see Figure 1, are used. They are sized 35x15x15mm and consist of a 3 axis accelerometer, a 3 axis gyroscope, and a 3 axis magnetometer, but the magnetometer was not used due to the high likelihood of magnetic field distortions in a clinical setting. Sensor specifications are detailed in Table 1.

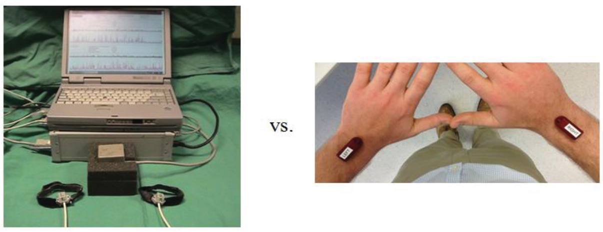


Figure 1. ICSAD device compared to new wireless device

The 12 data streams, 6 from each hand, were collected live over Bluetooth whilst 31 participants (a number of trainees and surgeons with varying experience) performed 4 FLS tasks on an FLS box trainer (a simulated training tool commonly used in laparoscopic training); see Figure 2.

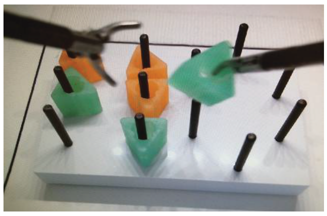
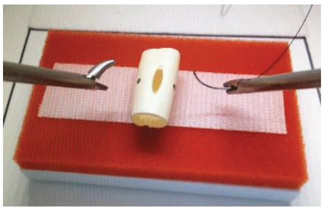
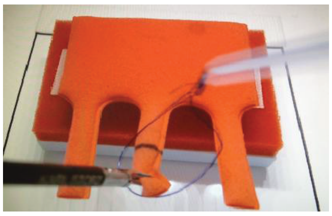
Table 1. Sensor specifications

	Rate (Hz)	Range
Accelerometer	12/50/100/200/400 /800	+/- 2g, 4g, 8g
Gyroscope	100/200/400/800	250/500/2000dps

The FLS programme is a method developed for the teaching and assessment of laparoscopic skills and is considered the gold standard for surgical assessment in the USA [9]. Each of the FLS tasks were addressed separately with data collected from 31 participants on the peg task, 25 participants on the precision cutting task, 24 participants on one of the suturing tasks (intracorporeal knot tying) and 26 participants for the ligating loop task. The FLS tasks are detailed further in Table 2. To assess construct validity, i.e., the ability to distinguish between surgeons of different experience, participants were scored on their timing and accuracy of the tasks (as outlined in standard FLS documentation) alongside the collection of motion data from the sensors and a video of the task.

Figure 2. FLS box trainer containing a peg transfer task

Table 2. Description of FLS tasks performed for this study

FLS task	Brief description of task
Peg transfer 	• Two graspers • Transfer 6 pegs to right side of board (pick up with left hand, pass to right) • Transfer 6 pegs back to left side of board (pick up with right hand, pass to left)
Precision cutting 	• Grasper and scissors • Make initial cut at bottom of gauze • Cut around circle and keep within 2mm of line
Intracorporeal knot 	• Two needle drivers • Drive needle through black dots • Pull suture through • Perform a double throw to tie first knot • Perform single throws to tie last two knots and cut
Ligating loop 	• Grasper and endoloop • Place grasper through endoloop and grasp foam finger • Tighten loop around foam within 1mm of black line

This study was split into 4 main areas: derivation of whole task performance parameters, feasibility studies and initial clustering for performance, derivation of a novel parameter by segmenting peg movements and finally, the clustering of data using all performance parameters for an overall competency-based score.

2.2 Derivation of Whole Task Performance Parameters

Parameters such as the time taken to complete a task and the number of hand movements have been validated in other surgical skill assessment studies to be able to distinguish experience levels. We derived the number of hand movements (H) from the raw accelerometer data (Acc_x, Acc_y, Acc_z) by capturing the number of zero crossings in the resultant signal (Mag_t):

$$H = \sum_t H(t),$$

$$where\ H(t) = \begin{cases} 1\ if\ (Mag_t > 0\ and\ Mag_{t-1} < 0) \\ 0 \quad\quad otherwise \end{cases}$$

$$Mag_t = \sqrt{Acc(t)_x{}^2 + Acc(t)_y{}^2 + Acc(t)_z{}^2}$$

Features also extracted from the data were: the time taken to complete each task, the variance and range of the separate signals and amount of intensity in certain frequency bands. Following discussions with surgeons, a new parameter was derived called the 'smoothness of movement' (*SM*) and this was done using the change in overall acceleration for each hand:

$$SM = \sum_{t} abs(Mag(t) - Mag(t-1))$$

These parameters were also normalised over time in order to compare values between participants.

2.3 Feasibility Studies and Early Validation

Feasibility studies were undertaken for the purpose of identifying optimal sensor placement, determining the optimal frequency of data collection, validating the parameters derived above and testing the ergonomics of the system.

One sensor was worn on the hand and one on the wrist for a number of tasks to test the optimal location and to find the optimal frequency of data collection, we looked at the error of the sensor in terms of the number of packets lost over wireless transmission and also at how the parameters derived from the data change with frequency. To validate some of the parameters mentioned above, a dataset was initially collected with one participant performing a peg transfer task 3 times: once normally, once with very smooth hand movements and once simulating bad performance (jerky movements). The algorithm to calculate the number of hand movements was also validated and optimised against a virtual reality (VR) simulator which itself gives a validated output of hand movements [11].

In order to assess performance on the task, the features extracted and derived from the motion data, as discussed above, were then the inputs to a k-means unsupervised clustering technique. The aim here was to then give meaning to each cluster in terms of a performance level. Dimensionality reduction had to be addressed as we had 39 features and over 20 data points for each task; features were selected empirically and are listed in Table 3. The performance parameters calculated from the raw data over the whole task (such as the number of hand movements) were inputs to this clustering technique and a competency-based score was determined.

Table 3. Parameters for clustering

1) Time taken	2) Smoothness of movement
3) Normalised number of hand movements	4) Overall maximum hand acceleration
5/6) Range of accelerometer values (each hand)	7/8) Range of gyroscope values (each hand)
9/10) Variance of resultant acceleration (each hand)	11/12) Variance of resultant gyroscope values (each hand)
13/14) Intensity of movement in 0-5Hz frequency band (each hand)	15/16) Intensity of movement in 5-20Hz frequency band (each hand)

2.4 Segmentation of Hand Movements

The segmentation of hand movements was investigated with the purpose of deriving a more advanced parameter for one of the FLS tasks. It was thought that the accuracy of the competency-based score obtained from the k-means clustering described above could be further improved with more detailed analysis of the movements required for each task.

Following the preliminary work detailed above, we focussed on one FLS task: the peg transfer. This task required participants to pick up a peg from one side of the board, transfer it to the instrument in their other hand and then place it down on the other side of the board. This is done for 6 pegs and then repeated to move all 6 pegs back to their original position.

Figure 3 shows the differential signals of the accelerometers placed on both left and right hands, and the ground truth (by manually marking the movements). Pilot data was collected from 31 participants with varying skill level. Differentiating the individual axes of the accelerometer (Figure 3), it was noticed that the 'put-down' movements of the pegs were clear in the data across participants as this was generally done with large, swift movements (circled in blue). This was seen in the right hand data for the first 6 peg movements and the left hand data for the final 6. The figure also shows the ground truth in red (1 denoting a left hand movement and 2 denoting a right hand movement), a peg drop (orange circle) and the attempts to pick the peg up.

The segmentation of these separate hand movements is an important feature to be able to extract from the data for the purpose of objectively assessing the performance on the task, and potentially could be extended to assess similar movements in real operations. From segmenting the peg pick-ups and peg put-downs, it is then possible to obtain the sequence of movements and determine the number of pegs transferred. This will indicate whether a participant has transferred all 12 pegs correctly or whether any pegs were dropped and attempts made to pick them up.

In order to find the ground truth for the training sets, the video of each task was recorded and synchronised with the data and the individual peg movements were segregated into windows; these were used as inputs to the training of the model.

A Support Vector Machine (SVM) approach was found to be successful in detecting the patterns in the accelerometer and gyroscope. A number of different supervised machine learning approaches were also tried, namely k-means and markov chains,

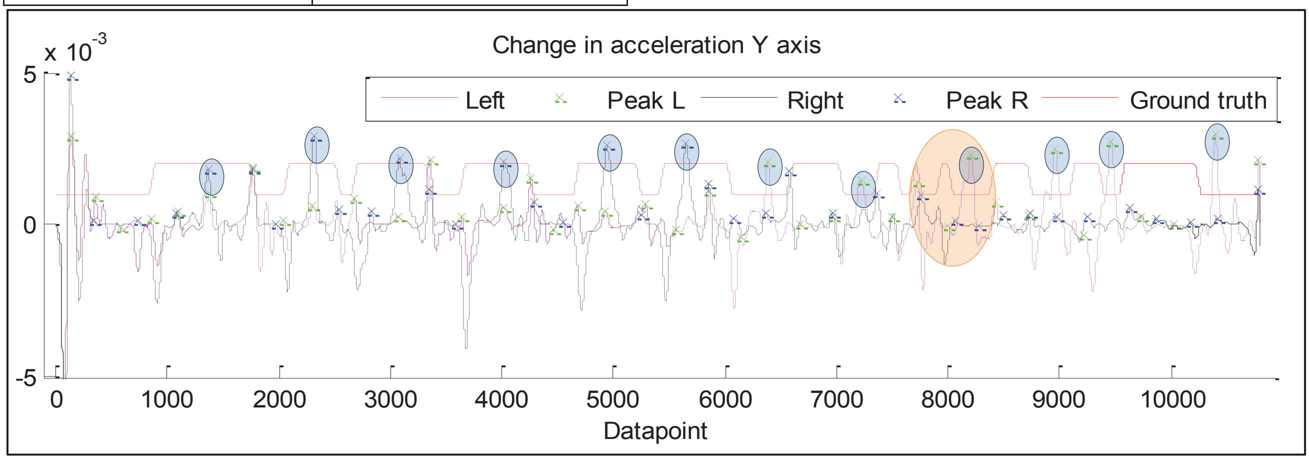


Figure 3. Change in acceleration to find peg put-down movements

in order to recognise the patterns in this data. Mostly these were unsuccessful as the data was largely varied between participants due to experience and different techniques in performing the task. We pre-processed the 12 raw data streams and extracted 10 features for each window of data and these were inputs to the SVM training. We were then able to train the model on 2 datasets leaving 29 to test the model. An SVM approach classifies data points to one of two classes that have been defined in the training phase. Given some training data T, a set of n windows of the form

$$T = \{(x_i, y_i) |\ x_i \in \mathbb{R}^{10}, y_i \in \{1, 2\}\}_{i=1}^{n}$$

where the y_i is either 1 or 2, indicating the class to which the point x_i belongs. Each x_i is a 10-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = 2$ [4].

From the segmentation of these peg movements, we determined the number of pegs moved in total for each participant and the sequence of movements and then used these as inputs to the k-means clustering algorithm. A perfect sequence of movements should read:

'LRLRLRLRLRLR**R**LRLRLRLRLRL'

where L and R represent the Left and Right hand movement respectively, and the **R** in the middle is the two right hand movements (RR) that occur as the first 6 pegs have all been transferred to the right hand side and the first peg is picked up for transfer back.

Four more parameters were also derived to compliment the output of the segementation of hand movements, more specifically, the percentage of time spent with: a dominant left hand, a dominant right hand, a stationary left hand and a stationary right hand.

2.5 Further Clustering for Performance on Peg Task

The same k-means unsupervised clustering technique was used as described in Section 2.3 however, the performance parameters calculated from the raw data over the whole task (such as the number of hand movements) were this time combined with the advanced parameters derived from the SVM algorithms and all were inputs to the clustering algorithm for the peg task. Figure 4 shows the design of the approach used to cluster the data into different performance levels.

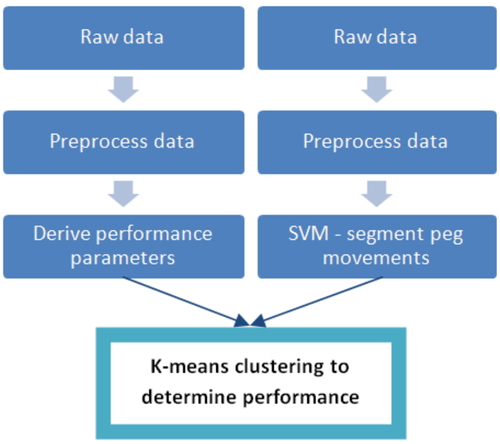


Figure 4. System design

This k-means clustering technique was then used to classify participants with a new competency-based score; Level 1 to Level 4 (novice to expert performance, respectively).

3. RESULTS

3.1 Finding Optimal Sensor Placement and Frequency

Determining the optimal sensor placement was an important part of the study, not only to ensure reliable data but also to find the best ergonomic location. It was shown that the raw accelerometer signals were highly correlated; see Table 4.

Table 4. Correlation of sensor worn on wrist vs. hand

	Before adjustment	After adjustment
Resultant accelerometer	0.79	-
Resultant gyroscope	0.33	0.95

The gyroscope signals appear to have weak correlation but this is due to a time delay between the rotation of the forearm and the rotation of the hand during movements. Adjusting for this delay by finding the same peaks in each signal improves the correlation to 0.95.

In terms of optimal frequency, we found the parameter values fluctuated for sample rates below 20Hz and the sample rate of 50Hz gave a more stable value with the lowest average error score of <3%.

3.2 Validation of Performance Parameters

Data from the 3 simulated peg transfers mentioned in Section 2.3 showed smooth and jerky movements were easily distinguished by the smoothness parameter (see Figure 5) though, expectedly, a similar number of overall movements were made.

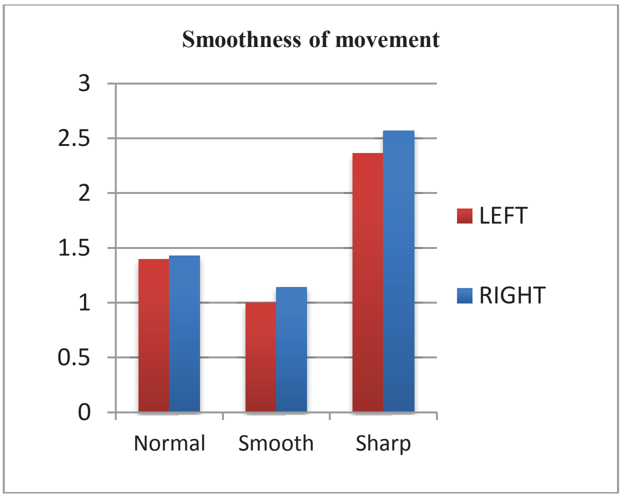


Figure 5. Simulated good and bad performance

The new system demonstrated a very high correlation between the number of hand movements parameter and the same parameter from the VR simulator, 0.82 and 0.99 for the left and right hands, respectively, with both values having a p-value of <0.0001 showing high significance; see Figures 6 and 7.

For 4 of the 16 tasks performed in this sub study, the left hand performed some movements away from the simulator that wouldn't have been detected by the VR simulator but would have been detected by the left hand inertial sensor. This can be seen for the first 4 tasks in Figure 6 and explains the slightly lower overall correlation value for the left hand.

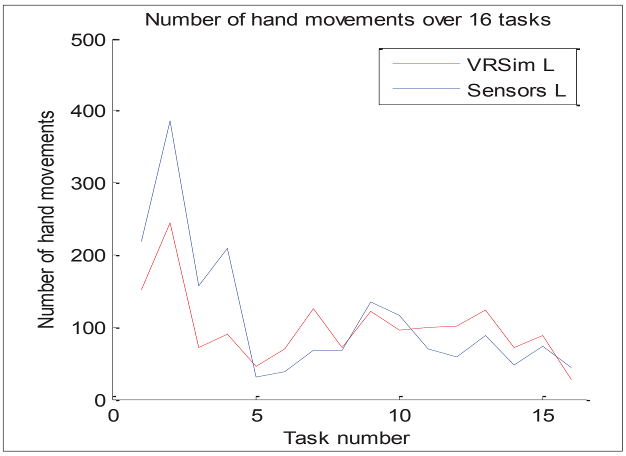


Figure 6. Validation graph for the number of hand movements parameter (left hand)

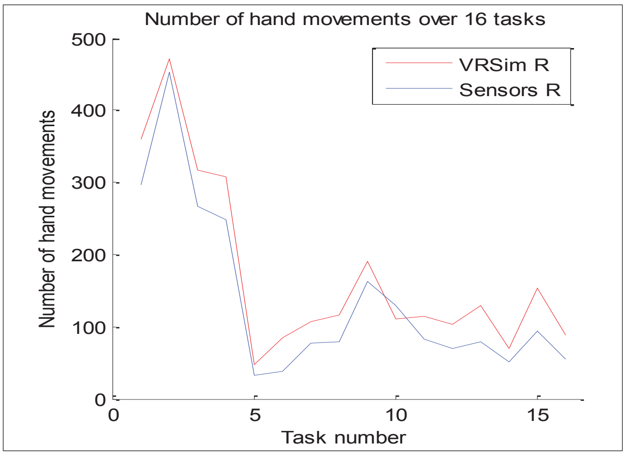


Figure 7. Validation graph for the number of hand movements parameter (right hand)

3.3 Segmentation of Hand Movements using Supervised Machine Learning

By windowing the data using the ground truth and the information highlighted in Figure 3, it was possible to split the peg pick-up and peg put-down movements in order to train the SVM. One whole peg movement and the patterns occurring in the data are shown in Figure 8.

Once trained, the test data was passed through the algorithm and an output was given, for example:

> *sequence: 'LRLRLRLRLRLRLRLRLRLRLRLRL'*
> *extra peg movements (over 12): 1*

This 'extra peg movements' parameter (novel peg parameter) was used as a new input to the k-means clustering technique along with the 4 parameters describing the percentage of time with each hand being dominant over the other and the percentage of time that each hand spends stationary over the whole task; see Table 5.

Table 5. New parameters for clustering

1) Left hand dominant (%)	2) Right hand dominant (%)
3) Left hand stationary (%)	4) Right hand stationary (%)
5) Extra peg movements (over the desired number)	

3.4 Clustering Participants into Performance Levels using Unsupervised Machine Learning

Following the extraction of a number of performance metrics and in particular, the segmentation of the hand movements during a peg task, the k-means based clustering technique was applied. Features were selected empirically as listed in Table 3 and then combined with the more advanced metrics listed in Table 5.

Results of clustering when using only the whole task performance parameters are shown in the top graph in Figure 9; only 3 of the 21 input dimensions are plotted in order to visualise the results but all parameters in Tables 3 and 5 were inputs to the algorithm.

By assessing the performance of participants using the FLS gold standard and the number of hand movements made, we were able to give meaning to the 4 clusters in terms of competency level with level 1 being the lowest (i.e. novice) and level 4 being the highest (i.e. expert). Since no participant dropped a peg outside the field of view, the gold standard score was calculated purely on the time taken to complete the task.

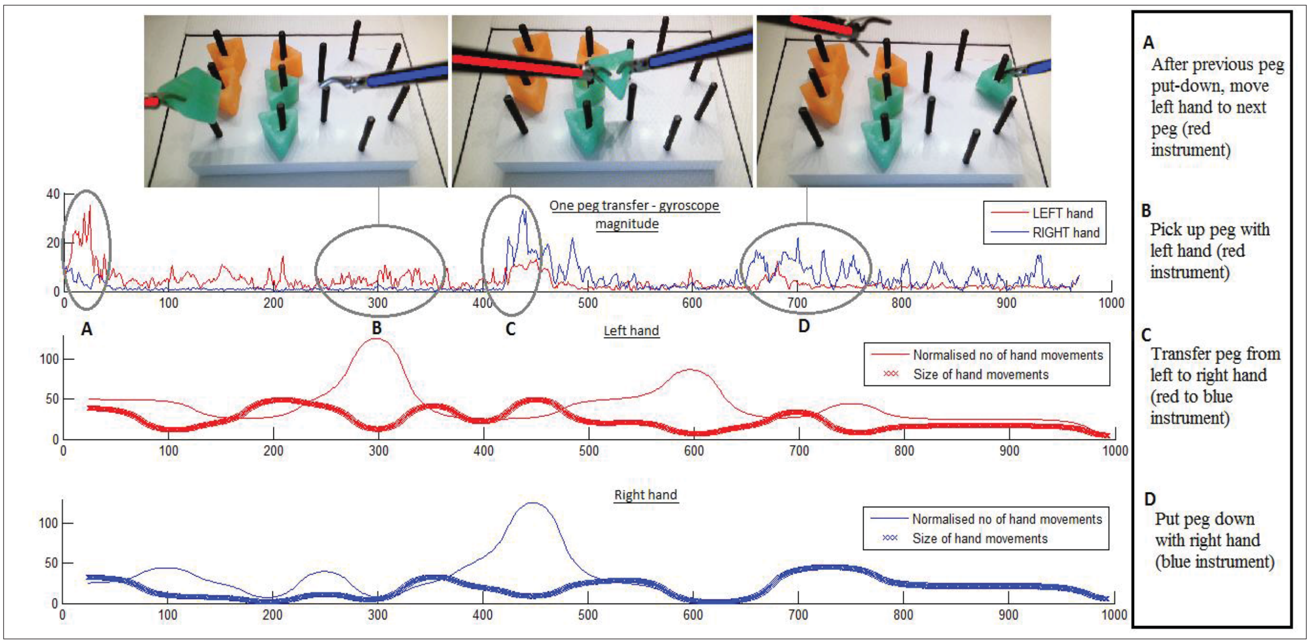


Figure 8. In depth analysis of one complete peg transfer

Table 6. Average time for each cluster

	Average time taken (s)	
	Without novel peg parameter	**With novel peg parameter**
Competency level 1	460.24	453.80
Competency level 2	174.83	162.92
Competency level 3	105.60*	150.46
Competency level 4	117.17*	107.42

*Level 4 had a higher average time taken than Level 3

Comparison of our results against the gold standard score can only be done at a high level to ensure that the fundamentals of the relationship between performance and expertise still hold. Therefore, the average time of clusters should decrease as the competency level increases. This is shown in Table 6 and it can be seen that only the novel clustering technique was correct by the gold standard (where better performance follows a shorter time). Assessment on these FLS tasks needs to move away from purely time-based scoring which encourages participants to perform the task only to achieve a shortened time but not to achieve the best quality of movements or task outcome.

In order to then look at more detailed accuracy of our clusters we looked at those participants who were reassigned to a different cluster with the addition of the novel parameter. It can be seen that by using the novel peg segmentation parameter, a substantial number of the datapoints (28%) were reassigned to new clusters. We found that these datapoints were correctly reassigned to cluster with those of similar performance in terms of time taken and the number of hand movements hence, improving the accuracy of the competency-based score.

As an example, one novice took longer than the minimum threshold of 300 seconds to complete the task. They were initially clustered into 'Comptency Level 2' but with the addition of the novel peg parameter were reassigned to correctly cluster with those also above the time threshold in 'Level 1'.

Table 7 shows the improvement in clustering for assessing task performance by comparing the correlation between the gold standard FLS scores and the competency-based scores obtained from each clustering technique. Using silhouette plots we were also able to determine that the more advanced technique gave a better separation between the 4 clusters.

Table 7. Correlation of clustering techniques with the FLS gold standard

	Correlation	**p-value**
Without novel parameter	0.52	<0.001
With novel parameter	0.80	<0.001

Two areas have been highlighted in Figure 9 and these correspond to one expert performer and a cluster of novice performers. It can be seen that the novice and expert participants have been correctly classified with competency levels 1 and 4, respectively, showing that the sensor location and parameters derived directly from the motion data are sufficient to assess performance on the peg task.

4. DISCUSSION

This work explores options for enhancing surgical training and assessment through wearable technology, providing an in-depth kinematic analysis of a standard surgical task.

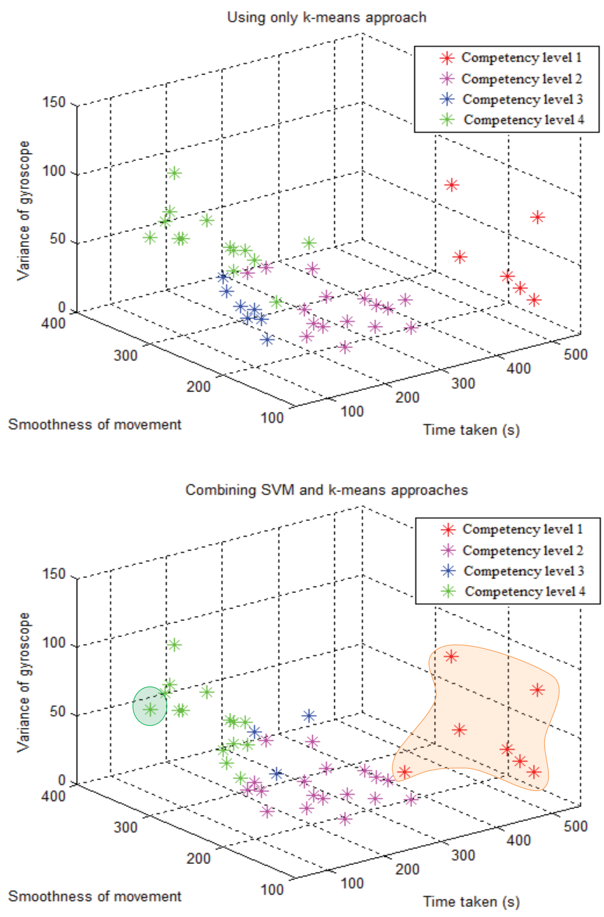


Figure 9. Clustering without (top) and with (bottom) novel peg segmentation parameter

From the feasibility studies, we see that the high correlation of data between hand and wrist locations promises a better position for the sensors than previous systems. This will help to enable a more feasible transfer to the operating theatre. Validations against a VR simulator and other simulated studies showed the reliability of the derived performance parameters. Standard FLS assessment on a peg transfer task involves a peer assessment of the time taken and the number of pegs dropped inside and outside the field of view. By deriving this from the sensor data, the need for subjective and time-consuming peer assessment is removed.

Beyond this preliminary study, further data collection is in progress and more advanced feature selection methods are being developed. All subjects in this study had only one or two attempts at the task and so by collecting data from participants repeating the task further we will look into performance over this longer exposure period, where we would expect the higher competency clusters to be slightly adjusted as participants reach proficiency in the task. Introducing more training sets into the SVM design will also be explored to further test the reliability of the algorithms when collecting more data and to ensure a high accuracy of clustering of expertise.

Overall, the approach of combining two machine learning techniques, one to determine a novel parameter and the other to cluster the data, shows a promising way to entirely objectively and quantitatively assess performance on a peg transfer task. This encourages further study into the data mining of the other FLS tasks and data collection has started for these other tasks. This study shows promising steps towards a wireless, low cost, automated and objective surgical training tool.

5. REFERENCES

[1] Aggarwal, R. et al. 2004. Laparoscopic skills training and assessment. *The British journal of surgery*. 91, 12 (Dec. 2004), 1549–58.

[2] Alvand, A. et al. 2011. Innate arthroscopic skills in medical students and variation in learning curves. *The Journal of bone and joint surgery. American volume*. 93, 19 (Oct. 2011), e115(1–9).

[3] Alvand, A. et al. 2012. Simple visual parameters for objective assessment of arthroscopic skill. *The Journal of bone and joint surgery. American volume*. 94, 13 (Jul. 2012).

[4] Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Machine learning*. 297, (1995), 273–297.

[5] King, R.C. et al. 2009. Development of a wireless sensor glove for surgical skills assessment. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*. 13, 5 (2009), 673–679.

[6] Moorthy, K. et al. 2004. Bimodal assessment of laparoscopic suturing skills: construct and concurrent validity. *Surgical endoscopy*. 18, 11 (Nov. 2004), 1608–12.

[7] Moorthy, K. et al. 2004. Dexterity enhancement with robotic surgery. *Surgical endoscopy*. 18, 5 (May 2004), 790–5.

[8] Moorthy, K. et al. 2003. Objective assessment of technical skills in surgery. *BMJ (Clinical research ed.)*. 327, 7422 (Nov. 2003), 1032–7.

[9] Okrainec, A. et al. 2013. Feasibility of remote administration of the Fundamentals of Laparoscopic Surgery (FLS) skills test. *Surgical endoscopy*. 27, 11 (Nov. 2013), 4033–7.

[10] Schijven, M.P. et al. 2002. The Advanced Dundee Endoscopic Psychomotor Tester (ADEPT) objectifying subjective psychomotor test performance. *Surgical endoscopy*. 16, 6 (Jun. 2002), 943–8.

[11] Watterson, J.D. et al. 2002. A Randomized Prospective Blinded Study Validating Acquistion of Ureteroscopy Skills Using A Computer Based Virtual Reality Endourological Simulator. *The Journal of Urology*. 168, 5 (Nov. 2002), 1928–1932.