Journal of Biomedical Informatics 59 (2016) 102–114

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Affordable, web-based surgical skill training and evaluation tool ☆

Gazi Islam [a,*], Kanav Kahol [b], Baoxin Li [c], Marshall Smith [d], Vimla L. Patel [a,e]

[a] Department of Biomedical Informatics, Arizona State University, 13212 E Shea Blvd, Scottsdale, AZ 85259, United States
[b] Public Health Foundation of India, New Delhi, India
[c] Department of Computer Science, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, United States
[d] Banner Good Samaritan Medical Center, 1111 E. McDowell Road, Phoenix, AZ 85006, United States
[e] The New York Academy of Medicine, 1216 Fifth Ave, New York, NY, United States

ARTICLE INFO

Article history:
Received 2 December 2014
Revised 1 November 2015
Accepted 3 November 2015
Available online 10 November 2015

Keywords:
Skill assessment
Feature detection
Computer vision
Web-based evaluation tool
Surgery

ABSTRACT

Advances in the medical field have increased the need to incorporate modern techniques into surgical resident training and surgical skills learning. To facilitate this integration, one approach that has gained credibility is the incorporation of simulator based training to supplement traditional training programs. However, existing implementations of these training methods still require the constant presence of a competent surgeon to assess the surgical dexterity of the trainee, which limits the evaluation methods and relies on subjective evaluation. This research proposes an efficient, effective, and economic video-based skill assessment technique for minimally invasive surgery (MIS). It analyzes a surgeon's hand and surgical tool movements and detects features like smoothness, efficiency, and preciseness. The system is capable of providing both real time on-screen feedback and a performance score at the end of the surgery. Finally, we present a web-based tool where surgeons can securely upload MIS training videos and receive evaluation scores and an analysis of trainees' performance trends over time.

1. Introduction

Surgical skill is dependent on both psychomotor and cognitive proficiency [1]. Surgeons have to learn the art of executing fine motor movements while maintaining tissue integrity within the human body. This is a challenging task, one that requires a significant time investment both from senior surgeons mentoring the residents and the residents themselves. However, this traditional one-to-one apprenticeship model has significant limitations. First, there is a dearth of trained surgeons, especially in the developing world, who qualify as evaluators. Secondly it requires dedicated time of the surgeon. Hence, it is almost impossible to immediately evaluate every surgical training exercise. The trainees are only assessed before moving to the next stage of training. As a result feedback becomes summative and this lack of granularity and proximity in the assessment fails to provide meaningful guidance on how to improve the subtle aspects of the individual's clinical skills [2].

Due to the demand for greater accountability and patient safety in health care delivery today; effective surgical performance measurement is becoming an absolute necessity [3]. To ensure the best surgical performance, systematic simulator training programs are being developed and integrated into traditional training in hospitals [4]. It is a new and progressive way to improve surgical resident training and surgical skills learning. The simulation systems allow multiple practice sessions, objective measurement of skills and the ability to deliver remote training. While simulation training is emerging as an effective means of practicing in a consequent free environment, the evaluation of surgical proficiency in simulation also requires the constant presence of a competent surgeon. This means that evaluating surgical dexterity still remains highly subjective and does not yield quantitative data [5]. High-end virtual reality (VR) simulators use sensors that map movements into virtual space where they are analyzed by algorithms [6–8]. While this method does offer a degree of objectivity it still has practical limitations. One such confine is that the introduction of sensor mechanisms both interfere with the surgeons' movements [9] and add adaptations that are not seen in real surgical environments. There are also human constraints, such as thinking of a surgical simulator as a videogame, which has no didactic value [10]. Prior experience with videogames has also been shown to be a handicap when using simulators [11]. And yet another concern is that such systems will create a false sense of security, built on

☆ We acknowledge the support of NSF (Grant No. 0904778). We thank Akshay Vankipuram for his help in editing the manuscript.

* Corresponding author. Tel.: +1 480 239 9709.
E-mail addresses: gislam@asu.edu (G. Islam), kanav.kahol@phfi.org (K. Kahol), Baoxin.li@asu.edu (B. Li), mark.smith@bannerhealth.com (M. Smith), vimla.patel@asu.edu, vpatel@nyam.org (V.L. Patel).

http://dx.doi.org/10.1016/j.jbi.2015.11.002
1532-0464/© 2015 Elsevier Inc. All rights reserved.

the development of incorrect habits while getting used to a virtual environment [10]. Thus the solution lies in developing an objective proficiency measurement system for surgery that will be (a) reliable, (b) replicable in real surgery, and (c) affordable.

Use of a video camera in surgical skill training has become very valuable as this video data can be used for future assessment. The video data can be further analyzed with computer vision analysis without the use of any external sensors. Computer vision analysis (herein referred to simply as "computer vision") has produced many technological breakthroughs in the last few decades. It has been successfully used for object detection, tracking, motion detection and analysis. A variety of computer vision applications that have been invented in the past few years can be applied in clinical as well as other domains of biomedical informatics to solve many problems [12,13]. In this research, open source computer vision [12] was used to track physicians' hand and surgical tool motion from captured video and with the goal of assessing surgical dexterity from the analysis. The tracking of a surgeon's hand and surgical tool movements can be one of the most important features in assessing surgical performance. This project addresses the issues of cost, replicability in a real environment, and the measurement of skills without interfering in surgeons' psychomotor movements by adding bulky sensors.

The system was used in minimally invasive surgical (MIS) techniques. Unlike open surgery, MIS uses tiny incisions to operate in constrained environments [14] which requires high levels of psychomotor proficiency and significant training. The specific psychomotor skills and hand-eye coordination needed for this type of surgery were reinforced through box-trainers and computer-enhanced simulation trainers, and the movements of instruments were captured by cameras by which the videos could later be analyzed by a computer vision system.

The proposed system used an optical flow algorithm to analyze the surgical field video in real time and provide dynamic feedback scores in order to assess surgical training performance. We assessed surgical proficiency as a multidimensional vector composed of motion smoothness, surgical gesture proficiency, and number of errors. Surgical proficiency is multidimensional by its nature and it was important to develop a system that captured each of the dimensions. This required a different algorithm for each of the proficiency measures. We then tested the hypothesis that this system could accurately capture surgical proficiency by differentiating between the performances of experts, intermediates, and novices.

Our system is capable of providing trainees both real time and summative feedback. Real time feedback was employed to provide measures to surgeons during the training which helped them dynamically control the learning experience. Summative feedback helped summarize their performance to gain knowledge about trends. We also developed an Internet based tool where users could upload pre-recorded surgical exercise videos and receive an immediate proficiency score.

2. Background

Laparoscopic surgery requires many hours of systematic practice on a simulator to acquire psychomotor skills. One simulator box that has commonly been utilized is the Fundamentals of Laparoscopic Surgery (FLS), and it has become one of the most widely used simulators for training. It is endorsed by the American College of Surgeons (ACS) and establishes a standard set of didactic information and manual skills serving as a basic curriculum to guide surgical residents, fellows, and practicing surgeons in the performance of basic laparoscopic surgery [15]. The training is composed of a laparoscopic trainer box which consists of a number of non-procedure specific simulation exercises incorporating most of the psychomotor skills necessary for basic laparoscopic surgery. We employed exercises from the FLS simulator to develop our system.

The FLS box offers a number of exercises. However for this study, three exercises were considered focusing on hand-eye coordination, ambidexterity, and depth perception (Fig. 1) [8]:

- **Peg transfer:** The peg transfer exercise requires the trainee to lift six objects with a grasper/dissector with the non-dominant hand, transfer the object midair to the dominant hand and then place each object on a peg on the opposite side of the board. There is no importance placed on the color of the objects or the order in which they are placed or where. Once all six objects have been transferred, the process is reversed. The exercise is timed and a penalty is imposed for any peg dropped out of the reach of the tool.
- **Intracorporeal suture:** This suturing task involves the placement of a suture precisely through two marks on a Penrose drain that has been slit along its long axis, and then tying an intracorporeal knot in the suture. The knot must have one double throw and 2 single throws. A penalty is imposed if the drain is avulsed from the block to which it is secured by double-sided adhesive tape.
- **Shape cutting:** This cutting exercise requires cutting out a circle from a square piece of gauze. One hand should be used to provide traction on the gauze using the grasper and to place the gauze at the best possible angle to the cutting hand. A penalty is imposed for any deviation from the line demarcating the circle.

In iterative practice sessions on the FLS, a user receives only summative feedback after a certain number of repetitions. Currently there are two feedback systems available. One requires the presence of an observer to monitor exercise completion time and committed errors which are sent to the FLS to score. The FLS typically returns the result (pass or fail) in 4–6 weeks. The second system requires post-hoc analysis on a recorded surgery of a competent surgeon or trainer to subjectively assess the surgical dexterity of the trainee by providing a composite score, which lacks inter-rater reliability and is an expensive process. However, both these methods remain primarily subjective in nature.

Several video and sensor-based systems have been developed to capture a user's motion and other important features which can be later analyzed objectively and quantitatively and correlated with the skill level [16]. However, most of the quantitative skill assessment systems available are sensor-based, i.e., sensors are integrated into surgical tools or surgeon's gloves to track different movement features. Although these sensory systems capture motion features quite well, there has not been enough work done in combining tool and hand movement data together to assess the surgical proficiency. Other researchers compelled surgeons to wear sensors to monitor certain features or body's center of pressure value and observed the correlation with the skill execution, but unfortunately these kinds of data are not sufficiently comprehensive to successfully assess the skill level. Moreover, skill assessments are usually performed only in simulated training environment since the integration of wearable sensors in live surgery interferes with proper surgical skill execution [16]. And even then the sensors have to be sterilized to be used in real surgery which increases the cost of the entire surgical procedure [17]. Unfortunately, none of these techniques described accurately assess actual surgical competence.

These drawbacks of sensor integration caused a shift in focus toward video based systems. Some video-based systems use only external cameras to capture hand movement while performing sur-

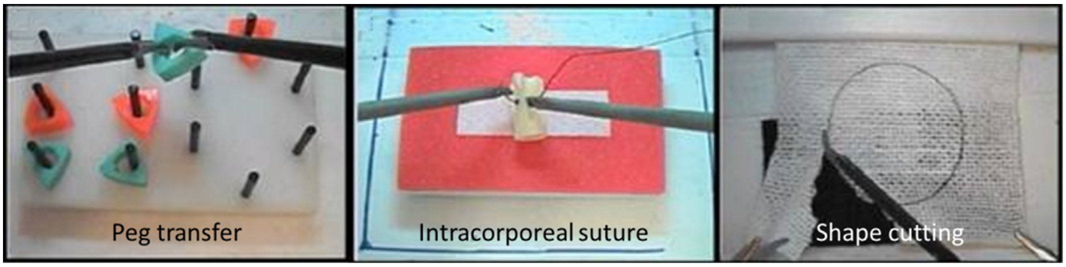

Fig. 1. Exercises for fundamentals of laparoscopic surgery.

gical exercise. This approach requires the camera to have an unobstructed view of the hands and is heavily reliant on the consistency of the ambient lighting, a lack of background noise, and the resolution of the video [16]. And due to the use of long instruments in MIS, it is almost impossible to derive tool movement data from a hand movement video. Other video-based approaches track surgical tool-tip; however, this analysis is solely dependent on the travel-length value of the tool-tip. At this point no work has been done that combines analyses from both hand and surgical tool movement. This is critical in surgery where surgeons' hand and tool movements are important components of ensuring patient safety.

Thus there exists a need for a complete performance evaluation system with automatic tracking. In the present study, we aimed to develop a video-based assessment tool for laparoscopic surgery training to assess the generic and specific technical aspects of surgical features. This method of assessment has previously been proven highly effective in improving technical skills acquisition and self-assessment [18]. The recorded videos can be shared securely via online for quicker distribution to experts for assessment [19]. Thus an Internet based video sharing tool was developed where recorded videos could be uploaded and assessed automatically by the developed algorithm.

3. Methodology

The proposed system in this research was designed to work with the Fundamentals of Laparoscopic Surgery trainer box. Video recordings rendered from inside the box and two external cameras were analyzed to provide formative feedback scores on the screen upon the completion of each exercise (Fig. 2). The idea was to assess both hand and tool data and calculate a score which is fed back to the user.

3.1. Object segmentation

The exercises in the FLS require surgeons to move certain objects. Analysis of the movements, jerkiness and placement of the objects can help provide assessments of the surgeons' level of proficiency. For example, a piece of gauze that is cut by the surgeon provides valuable information about the surgeons' proficiency, since the shape of the cut, its size and duration to execute that cut are all important cues. The system therefore required an efficient object segmentation algorithm. We tested a number of computer vision algorithms to analyze motion.

The color image was converted from a Red–Green–Blue (RGB) color space to Hue–Saturation–Value (HSV) color space. In the RGB model, images are represented by three components, one for each of the primary colors – red, green and blue. However, the HSV color space can capture the distinct image features better than RGB color space [20,21]. The HSV image contains 3 channels: hue (H), saturation (S) and value (V) where Hue is a color attribute and represents a dominant color [22]:

$$H \leftarrow \begin{cases} 60(G-B)/S & \text{if } V=R \\ 120+60(B-R)/S & \text{if } V=G \\ 240+60(R-G)/S & \text{if } V=B \end{cases}$$

If $H < 0$ then $H \leftarrow H + 360$.

The HSV image of the video was split into a single channel Hue image. To isolate marker of the tool and hands, a histogram of the Hue channel was calculated. Color of the tool-marker was red and the gloves were purple whose corresponding histogram values were found and binary thresholding applied to detect the tool and the hand respectively. In this case, if the pixel value in the hue channel matched the histogram color value, it was changed to 255 i.e., white. The rest of the pixels were set to 0 i.e., black.

$$\mathrm{dst}(x,y) = \begin{cases} \mathrm{maxValue} & \text{if } \mathrm{src}(x,y) > \mathrm{threshold} \\ 0 & \text{otherwise} \end{cases}$$

Finally to maximize the elimination of noise, advanced morphological transformation was used. Advanced morphological transformation performed as a noise filter by using erosion and dilation as basic operations. Erode and dilate functions use the specified structuring element that determines the shape of a pixel neighborhood over which the minimum and maximum (respectively) is taken [12].

$$\mathrm{dst} = \mathrm{open}(\mathrm{src}, \mathrm{element}) = \mathrm{dilate}(\mathrm{erode}(\mathrm{src}, \mathrm{element}), \mathrm{element})$$

where

$$\mathrm{erode}(x,y) = \min_{(x',y') \in \mathrm{kernel}} \mathrm{src}(x+x', y+y')$$

$$\mathrm{dilate}(x,y) = \max_{(x',y') \in \mathrm{kernel}} \mathrm{src}(x+x', y+y')$$

Figs. 3 and 4 show the noise-free detection tools and hands respectively.

3.2. Motion detection

Measuring motion and its features as mentioned before is critical to assessing surgical proficiency. In the video, the sequence of images are taken at a fixed time interval. The motion of objects in 3-D induces 2-D motion in the image plane. The motion is called *optical flow*. There are several methods for calculating optical flow. By applying a frame-to-frame differencing technique to find the object silhouette, and motion history images (MHI), the dynamic part of the scene was captured. Frame-to-frame differentiating function calculates absolute difference between two arrays of consecutive frames [12].

$$\mathrm{dst} = |\mathrm{src1} - \mathrm{src2}|$$

The function extracts templates by thresholding frame differences and then updates the motion history image by passing the resulting silhouette.

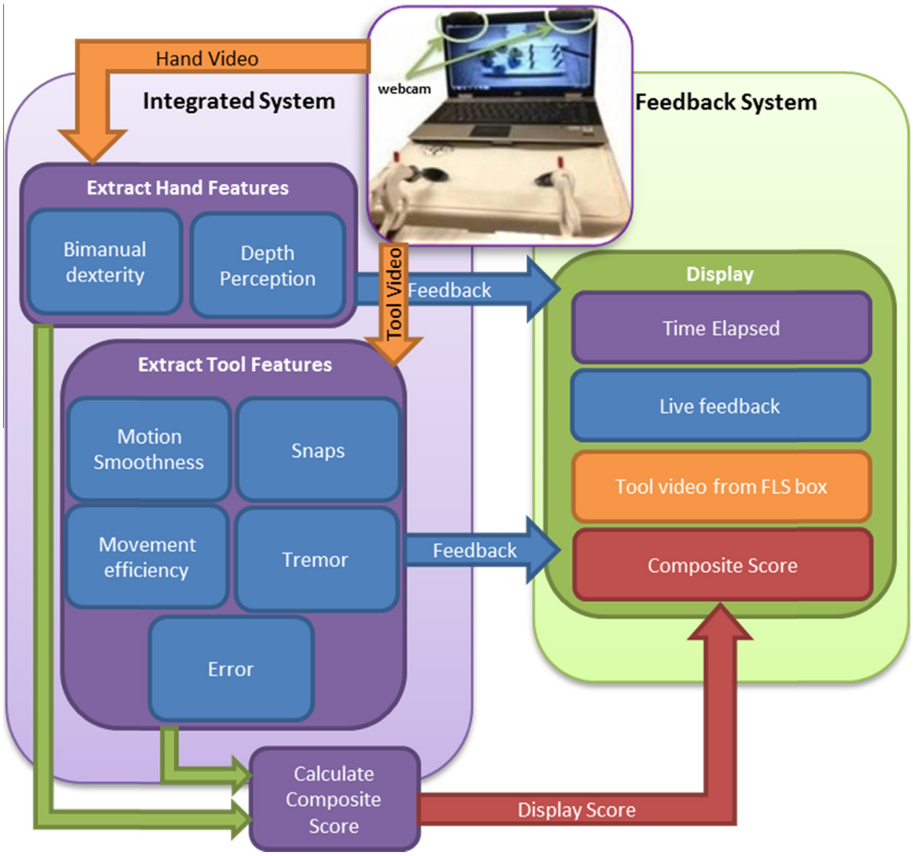

Fig. 2. Proposed system with integrated hand/tool video and a feedback system.

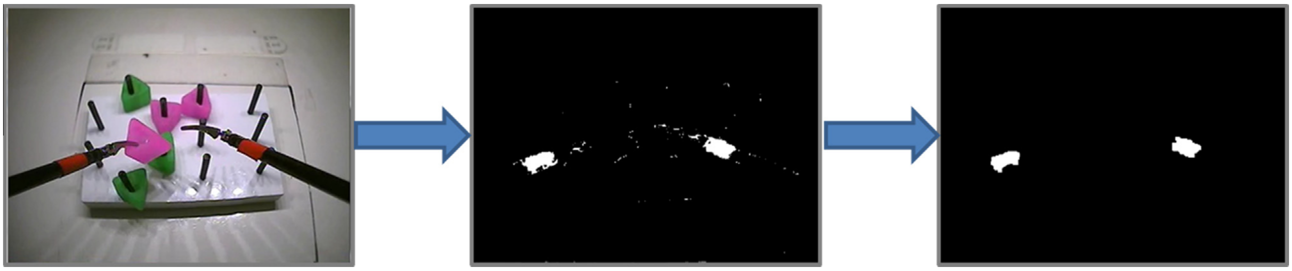

Fig. 3. Noise-free tool detection applying advanced morphological transformation.

$$\mathrm{mhi}(x,y)=\begin{cases}\text{timestamp} & \text{if silhouette}(x,y)\neq 0\\ 0 & \text{if silhouette}(x,y)=0 \text{ and mhi} < (\text{timestamp}-\text{duration})\\ \mathrm{mhi}(x,y) & \text{otherwise}\end{cases}$$

That is, MHI pixels where motion occurs are set to the current timestamp, while the pixels where motion happened before specified duration are cleared (Fig. 5) [23].

The gradients of the resulting motion history image were taken to produce a mask of valid gradients. First the derivatives *Dx* and *Dy* of MHI are calculated to find the gradient orientation as:

$$\mathrm{orientation}(x,y)=\mathrm{TAN}^{-1}\frac{Dy(x,y)}{Dx(x,y)}$$

After that, the mask was filled to indicate where the orientation was valid. For the local motion segments, small segmentation areas were first rejected and then the orientation was calculated using regions of interest (ROIs) that bound the local motions; then the areas of valid motion within the local ROIs were calculated. Any such motion area that was too small was rejected. Fig. 6 shows the detected motion in both hand and tool videos. The algorithm recorded the coordinates of the center point of motion in pixel and the gradient of movement in degrees.

3.3. Feature extraction

Two arrays of data were extracted from both left and right hand and tool positions (pixel) and gradients (degree). Coordinates of the tool were analyzed to extract smoothness, extra movement, and

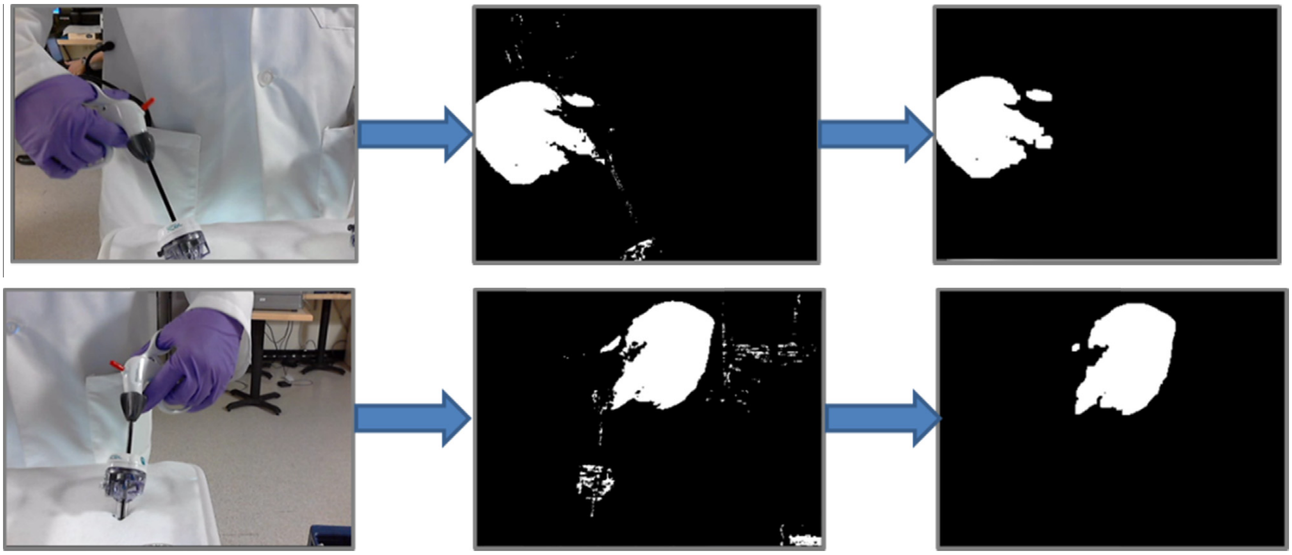

Fig. 4. Noise-free hand detection applying advanced morphological transformation.

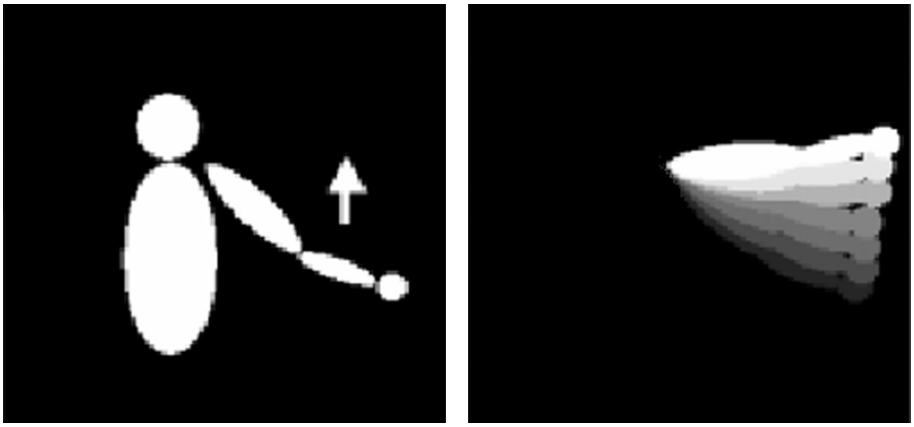

Fig. 5. Example of motion history image (MHI) of hand movement.

movement efficiency. Tremor was calculated by analyzing both the position and direction of movement. Angular values of the direction of hand movement were interpreted into 4 gestures – left, right, up, down and stationary for further extraction of features like perception of depth and hand movement efficiency.

3.3.1. Motion smoothness (tool video)

Smooth steady motion is one of the most important features in assessing surgical skill [24]. From the coordinates of the tool position, Euclidean distance between every frame was calculated. If the hand movement is x, then hand motion, acceleration and jerkiness could be found from the following equations:

$$\text{Hand motion} = \frac{dx}{dt}$$

$$\text{Hand motion acceleration} = \frac{d^2x}{dt^2}$$

$$\text{hand motion jerkiness} = \frac{d^3x}{dt^3}$$

Tool movement for 5 consecutive frames is summarized in Fig. 7, which shows velocity, acceleration, jerkiness and snap of the tool movement which are the 1st, 2nd, 3rd and 4th derivatives for movement of 5 frames respectively. The figure also shows that deceleration, negative jerk, and snaps features were derived from it, but deceleration and its higher derivatives were ignored since they have no association with smooth and steady motion. Positive jerk value for the entire duration of exercise was calculated to find the counter feature i.e., motion smoothness [25].

3.3.2. Snaps (tool video)

Extra movement such as snaps is another important factor that may be used to determine surgical proficiency. It was calculated by the 4th derivative of hand movement [25]. Total value of positive snaps was calculated to find this movement feature in every exercise.

$$\text{Snaps} = \frac{d^4x}{dt^4}$$

3.3.3. Movement efficiency (tool video)

On surgical simulators measures of economy of tool movement have been shown to be reliable, valid, and objective measures of technical competence [26]. Once the Euclidean distance between tool positions and the starting frame tool position was calculated for each frame, the values were averaged to find the tool movement efficiency. Average movements of both left and right hands

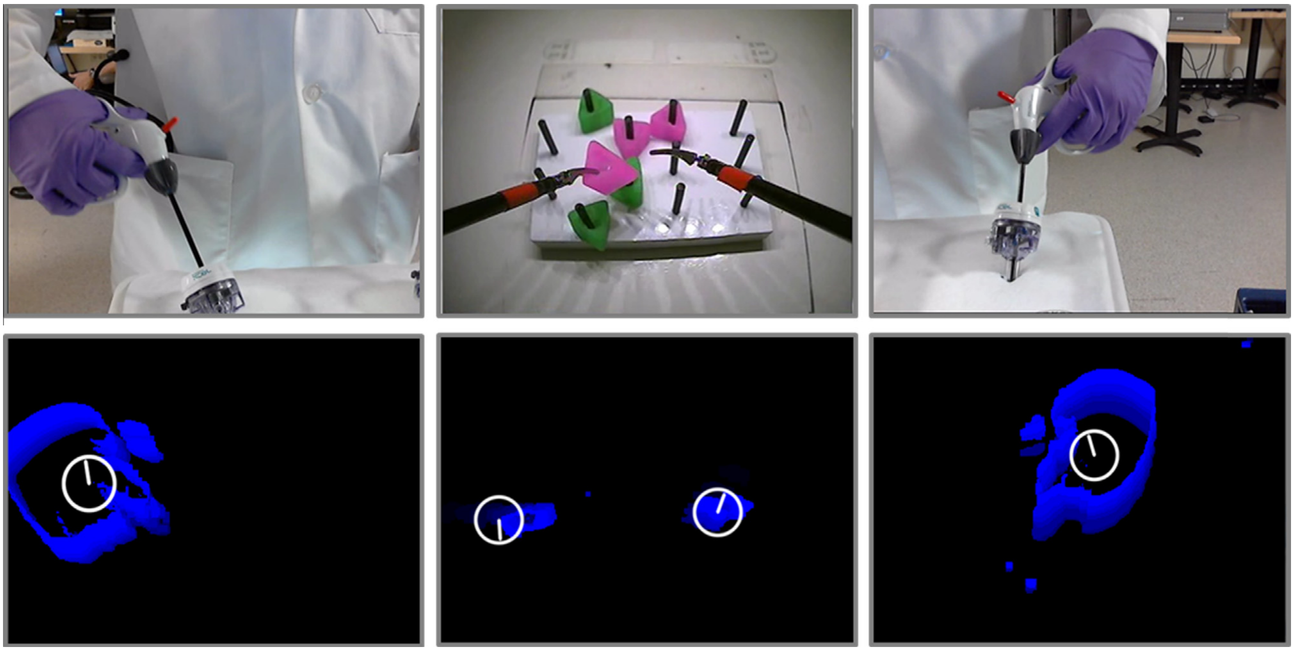

Fig. 6. Detected motion in hand and tool videos.

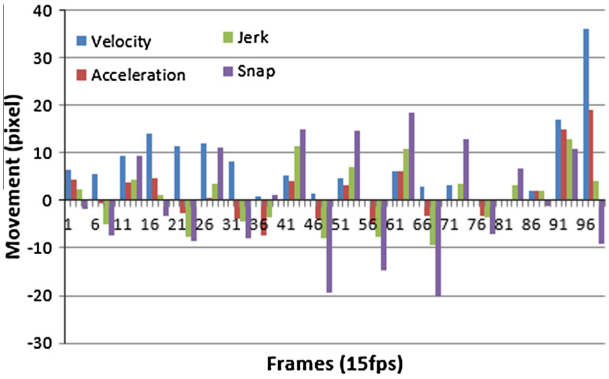

Fig. 7. Different order derivatives of tool movement across five frames.

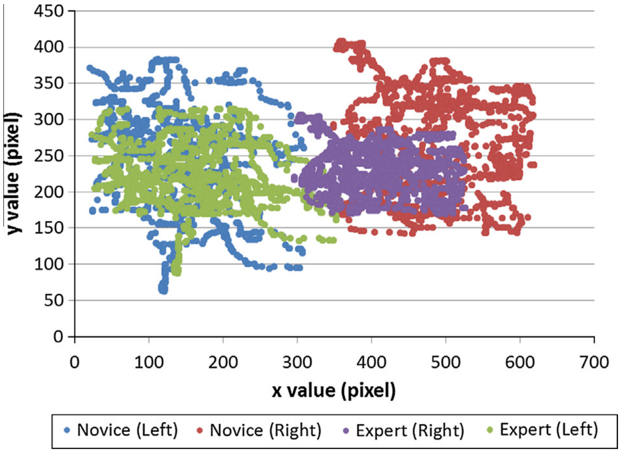

Fig. 8. Trajectory of tool movement of left and right hands of novice and expert surgeons.

of novices and experts are shown in Fig. 8. It is clear that novices expended much more movement than the experts for the same exercises.

3.3.4. Tremor (tool video)

Tremor is an involuntary, roughly sinusoidal component inherent in a normal human hand motion. It has been found to consist of a "mechanical–reflex" component which is thought to originate from the central nervous system and has a frequency range of 8–12 Hz [24]. Imprecision in laparoscopic surgery due to tremor has long been a concern. Surgical tool tremor was calculated by the motions that change directions in 2 frames and had a value of no more than 5 pixels. It was characterized by the following function where numbers in the parentheses refer to the frame numbers:

```
If ((direction[6] != direction[4]) && (movement[6] –
   movement[4] <= 5)) {
   If ((direction[4] != direction[2]) && (movement[4] –
   movement[2] <= 5)) {
      If ((direction[2] != direction[0]) && (movement[2] –
   movement[0] <= 5)) {
         tremor = true;
      }
      Else
         tremor = false;
   }
}
```

3.3.5. Depth perception (hand video)

Repetitive motion toward the direction of the tool was recorded as up-down motion and accounted for the perception of depth. Table 1 categorized up-down movements with a range of angular movements. This repetitive hand movement is observed mostly in the inexperienced residents due to, (1) translation of the 2-dimentional image of the operating field from the video screen into a 3-dimentional mental image [27], (2) learning to operate using long instruments, and (3) mastering ambidexterity and eye-hand coordination. This motion was calculated by the total number of hand motions both in up and down directions.

Table 1
Segmentation of movement from angular values.

Angle	Direction
$45^\circ <, \leqslant 135^\circ$	Up
$135^\circ <, \leqslant 225^\circ$	Left
$225^\circ <, \leqslant 315^\circ$	Down
$45^\circ \geqslant, > 315^\circ$	Right
0	No movement

3.3.6. Bimanual dexterity (hand video)

Bimanual dexterity is a measurement of how well a resident is able to optimize the use of both hands [28,29]. If the resident ignores the non-dominant hand, then he/she probably has not mastered the bimanual dexterity. Bimanual dexterity was calculated by comparing percentage of idle time for each of the hands for the total time to complete the exercise.

3.4. Detecting error

Errors vary from task-to-task. Three different algorithms were developed to detect task-specific errors from the tool videos and record separately in peg transfer, intracorporeal suturing, and shape cutting exercises.

3.4.1. Peg transfer

The developed algorithm detected the colorful triangular objects and computed the total number by counting all of the pixels inside the blue rectangle (Fig. 9). If at the end of the exercise all the objects were not inside the rectangle, then the missing number of objects was recorded as errors.

3.4.2. Intracorporeal suturing

During intracorporeal suturing, if the Penrose drain was pulled excessively from its original location, then it was considered to be tissue damage [30]. The program detected the Penrose drain and automatically calculated the deviation from its original position during the exercise (Fig. 10).

3.4.3. Shape cutting

In the shape cutting exercise, the user excised the required piece of gauze and then placed that piece under the camera, and upon pressing ‘spacebar’, the program automatically detected the shape of both the inside and outside cuts. The program saved an image (snapshot) and converted it from an RGB to grayscale image. Then it employed a Gaussian blur filter to smooth the image and also a Canny’s filter to find edges. Finally the Hough transformation found the actual circle and two binary images were produced for the contents outside and inside the circle respectively (Fig. 11). By counting the number of white pixels, the program automatically found the outside and inside imperfection and provided a combined precision of cut score.

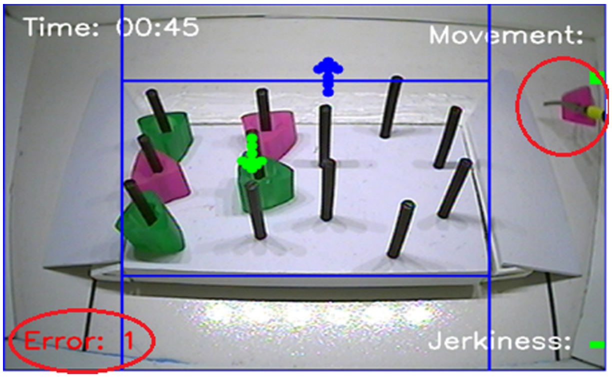

Fig. 9. Error recording in peg transfer exercise by automatic detection of an object outside the test area.

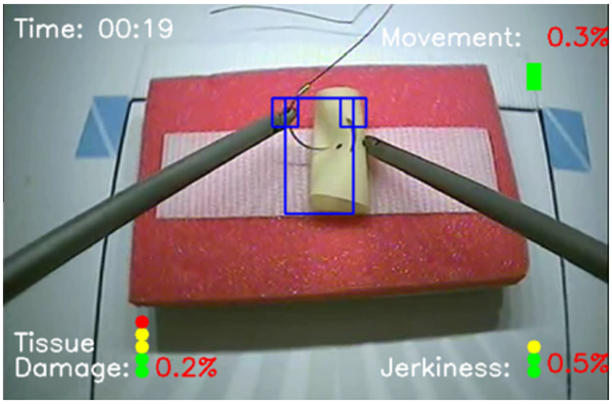

Fig. 10. Detection of tissue damage during exercise.

3.5. Multivariable linear regression

The computer vision algorithm extracted six features from each of the laparoscopic cholecystectomy videos along with their corresponding hand movements. Each of the videos was rated by expert surgeons in three categories: smoothness, efficiency, and precision. Sets of three of the extracted features were found to be associated with each of the rated features by the experts. For example, the smoothness score was associated with the jerkiness, snaps and tremor; and the efficiency score was associated with movement efficiency, depth perception and bimanual dexterity. Three of the extracted features were grouped into a single feature which the experts rated. Multivariate linear regression models were built using the tool “Weka” to find the relationship between extracted features and experts’ scores.

$$h_a(\text{smooth}) = a_0 + a_1(\text{jerkinesss}) + a_2(\text{snaps}) + a_3(\text{tremor})$$

$$\begin{aligned} h_b(\text{efficiency}) = b_0 &+ b_1(\text{movement efficiency}) \\ &+ b_2(\text{depth perception}) + b_3(\text{bimanual dexterity}) \end{aligned}$$

The precision score was solely associated with the number of error committed and the time taken to complete the exercise. After discussions with expert surgeons, the equation for measuring error for each of the surgery types was created and is given below:

$$\begin{aligned} \text{error}_{\text{Peg transfer}} = \frac{1}{5} &\times ((\text{completion time(if} > 48) - 48)\ \text{s} \\ &+ (\text{No. of drops} \times 25)\ \text{s}) \end{aligned}$$

$$\begin{aligned} \text{error}_{\text{Intracorporeal suture}} = \frac{1}{12} &\times ((\text{completion time(if} > 112) - 112)\ \text{s} \\ &+ \text{tissue damage}) \end{aligned}$$

$$\begin{aligned} \text{error}_{\text{Shape cutting}} = \frac{1}{10} &\times ((\text{completion time(if} > 98) - 98)\ \text{s} \\ &+ \text{cut imperfection}) \end{aligned}$$

48 s, 112 s and 98 s are the average time for experts to complete peg transfer, intracorporeal suture and shape cutting exercise respectively.

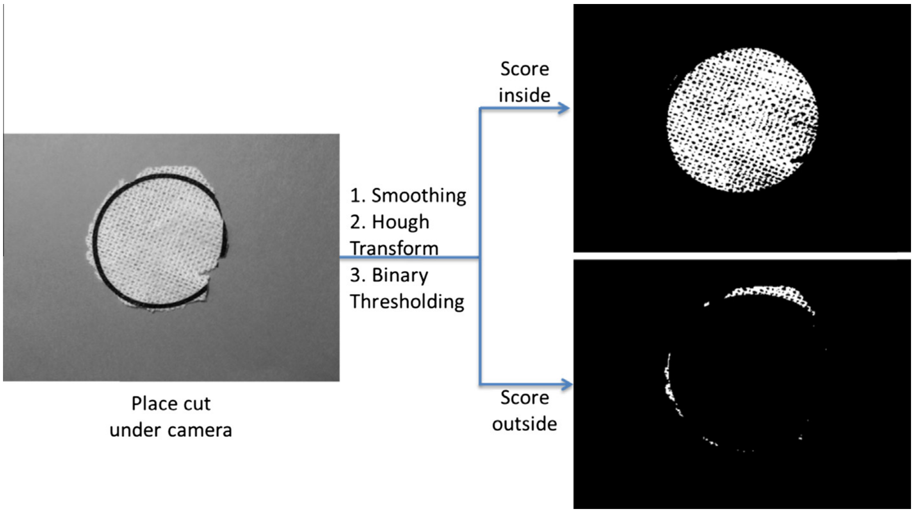

Fig. 11. Automatic measuring of precision in shape cutting task.

3.6. Final scores

Once the exercise was completed, the system would calculate scores for each feature and display it on the screen along with a composite score. Features from hand and tool videos were normalized to 100. Finally, the following scoring scheme was employed for each exercise to calculate the composite score:

$$\text{score}_{\text{Peg transfer}} = \frac{h_a(\text{smooth}) + h_b(\text{efficiency})}{2} - \text{error}_{\text{Peg transfer}}$$

$$\text{score}_{\text{Intracorporeal suturing}} = \frac{h_a(\text{smooth}) + h_b(\text{efficiency})}{2} - \text{error}_{\text{Intracorporeal suture}}$$

$$\text{score}_{\text{Shape cutting}} = \frac{h_a(\text{smooth}) + h_b(\text{efficiency})}{2} - \text{error}_{\text{Shape cutting}}$$

3.7. Training efficacy

A control group and a test group of subjects were used to test the effectiveness of the feedback system. No feedback was provided to the control group, but the test group was provided real-time feedback on motion smoothness, movement efficiency, number of committed errors and elapsed time. Each subject from both of the groups repeated a single surgical exercise after three weeks. For each of the subjects, a performance score for each parameter was calculated and added to form a composite score. A learning curve was created by plotting the composite score over time using the Matlab program. We expected the learning curve for the test group to be steeper than that of the control group (see Fig. 12).

3.8. Webtool integration

We also designed and implemented an Internet based system which could automatically evaluate a FLS exercise video. The website was hosted on a secured server where users could login and upload videos and receive instant evaluation scores on different performance features. Both computer vision and neural network algorithms were run to provide the assessment score. Users could also access their previous assessment scores and observe the progress. All these videos were de-identified and stored in a secured repository. The website also contained the rater's account where expert surgeons could login into the system and rate the de-identified videos in smoothness, efficiency, and precision. These ratings were then used to train the neural network to improve the accuracy of the evaluations.

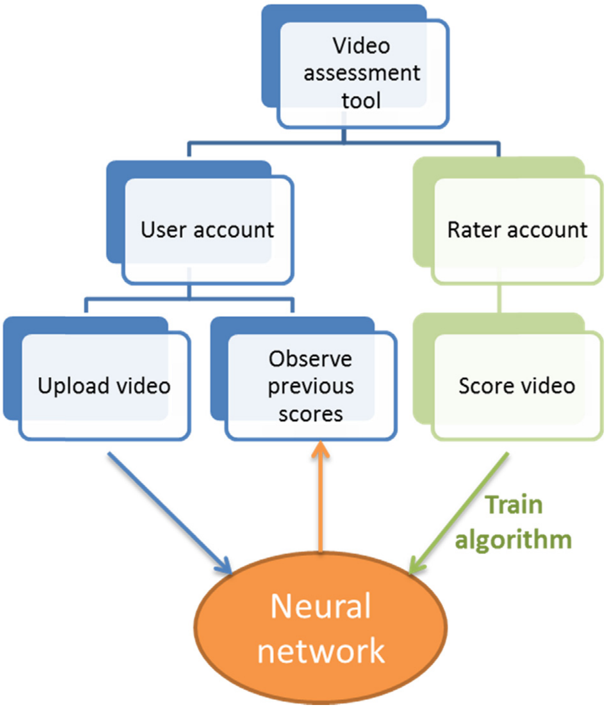

Fig. 12. Internet based system for training of the algorithms.

4. Results

Video data of medical students and surgical residents (system users) from various post graduate years (PGY) were collected while practicing peg transfer, intra corporeal suturing and shape cutting exercises using the FLS trainer. Each subject conducted the exercise at two consecutive times, and videos of the second trial were analyzed. Each exercise data was captured with three synchronous cameras; two cameras for capturing each of the hand movements

Table 2
Number and expertise levels of subjects.

Group	Experience level	No. of subjects
Expert	PGY5–PGY4	8
Intermediate	PGY1–PGY3	12
Novice	Medical students	32

and one camera for capturing the tool movement. A total of 52 sample videos were collected and grouped as shown in Table 2.

4.1. Inter-rater reliability

Fifty-two de-identified FLS exercise videos were sent to three expert surgeons for evaluation. Each video was analyzed and rated in terms of smoothness, efficiency and precision. To find the inter-rater reliability, Fleiss' Kappa coefficient between experts' score was calculated. Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed numbers of raters when assigning categorical ratings to a number of items or classifying items. Table 3 shows fair to moderate agreement between the raters with significant *p*-value.

4.2. Stratification of expertise

All of the frames of the surgical tool and hand videos were analyzed. Arrays of position and movement gradients were captured from each frame where each sample consisted of both left and right tools/hands. The angular value of the movement gradient was converted to four motions: left, up, right and down. Several algorithms analyzed the displacement and direction of movements and prepared an occurrence matrix for each of the gestures.

Fig. 13 shows the jerkiness, snaps and tremor scores for both the left and right surgical tools of participants from three levels of expertise. The histogram clearly shows that experts had lower jerkiness scores, hence their movements were smoother than those of intermediates and novices. The total numbers of snaps were also calculated and average snaps among the different expertise groups were displayed in Fig. 13. Tremors were difficult to assess subjectively, however the motion detection algorithm was able to detect very subtle movement and the system calculated tremors for both hands and tools. The figure shows a decrease in tremor features with the increasing experience level of the examiners. Due to the use of longer instruments, very subtle hand movement could cause greater tool movement and thus a greater number of tremors were noticeable in tool analysis.

Motion redundancy for both hand movements was analyzed. More than 65% of the redundant motion was observed in up-down direction, which results from extra motion in perception of depth. Fig. 14 shows motion redundancy in up-down direction. It appears that as the level of expertise increases the perception of depth also increases, producing a reduction in vertical redundant motion. In addition, the average traveling distances for the tools were also calculated and it showed that experts required on average 40% less movement than novices to complete the same exercise.

Table 3
Inter-rater reliability: Fleiss' Kappa coefficient between raters.

	Fleiss' kappa	*P*	Agreement
Smoothness	0.4349	<0.0001	Moderate agreement
Efficiency	0.3581	0.0001	Fair agreement
Precision	0.4393	<0.0001	Moderate agreement

Percentage of time the non-dominant hand moved compared to the dominant hand was calculated to assess bimanual dexterity. Both intermediates and novices showed almost 40% less activity in the non-dominant hand as compared to the experts (Fig. 15).

Table 4 presents the results of ANOVA among the different skill levels which was found to be very significant for all six features. $p < 0.05$ was taken as statistically significant.

Random Forest classifier was used for the classification of each gesture data. Sixty-six percent of the data for both hand and tool gestures was used as training data. The remainder 34% of the data was unlabeled and used for testing of each gesture. Each gesture shows a significantly high true positive rate of detected features (Table 5).

Motion smoothness, motion redundancy and tremor value for both hand and tool gestures were normalized and Linear Discriminant Analysis (LDA) was performed. Fig. 16 shows the result of the LDA analysis for both hand and tool gestures where the data roughly conform to 3 distinct regions in a 2-dimensional projection space. These initial experiments validated the hypothesis that LDA could be used to simplify the original data into a simpler, low-dimensional data set. In addition the features from the tool movements were more accurate at different levels of expertise than those of the hand movements.

4.3. Smoothness and efficiency score

Training dataset was used to find the regression parameters. Once all the parameters were determined, the test data set was run to find the correlation coefficient between raters' average score and predicted score. Both equations showed more than 90% correlation with significant *p*-value (Table 6).

4.4. Error detection

For error detection, the automatically generated error score was compared to the observers score for each of the videos. The system was able to detect drops of triangular objects in peg transfer exercise very accurately. The overall sensitivity of the system was found to be 87% (Table 7).

4.5. Efficient skill learning

The proposed video-based surgical skill assessment technique could provide immediate feedback, hence it was also tested as a tool for the efficient skill learning technique. Thirty-two medical students (novices) were used as the control group where no-screen feedback or assessment score was provided. Twenty-two medical students were used as the experimental group where they were provided real-time on-screen feedback and assessment score at the end of every trial. All participants in each group performed the peg transfer and shape cutting exercise at least 16 times. The average score for each of the two groups as a function of the number of trails is shown in Fig. 17, where an improved learning curve was observed for the experimental group. Paired t-test was performed showing a significant difference in performance between the two groups ($p < 0.0001$).

5. Discussion

In the present study we developed and evaluated a video-based assessment tool for laparoscopic surgery training to assess the generic and specific technical aspects of surgical skills. Fifty-four FLS training exercise videos performed by medical students and residents from various postgraduate years of training were analyzed. Several functions were developed that combined a series of com-

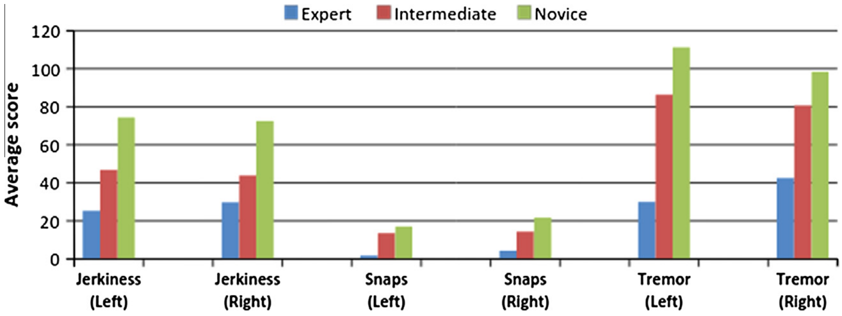

Fig. 13. Jerkiness, snaps and tremor scores as a function of expertise.

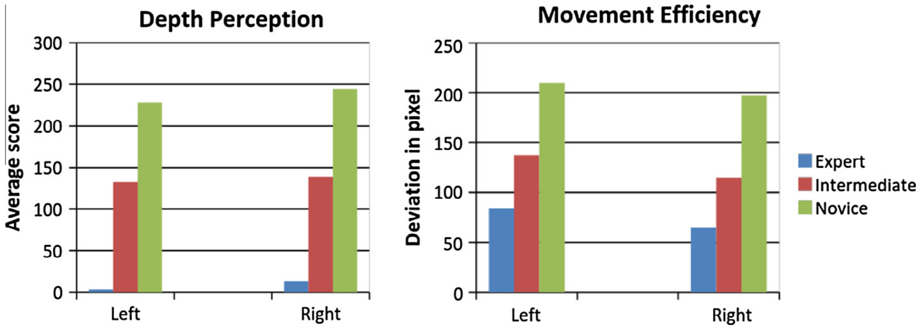

Fig. 14. Perception of depth and movement efficiency as a function of expertise.

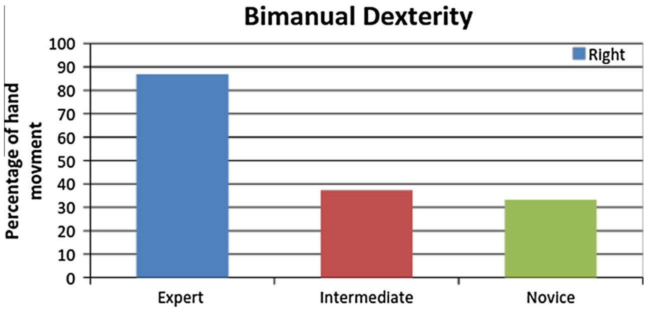

Fig. 15. Comparison of bimanual dexterity as a function of expertise.

Table 4
Comparison of skill levels by expertise, using the tool and hand gestures (ANOVA analysis).

	Gestures	Expert vs intermediate *p*-value	Intermediate vs novice *p*-value	Expert vs novice *p*-value
Tool	Smoothness	0.0032	<0.0001	<0.0001
	Snaps	0.0425	0.0325	0.00345
	Movement efficiency	0.0295	0.0002	<0.0001
	Tremor	0.0002	0.0045	<0.0001
Hand	Depth perception	<0.0001	<0.0001	0.0001
	Bimanual dexterity	<0.0001	<0.0001	<0.0001

Table 5
Random forest analysis of training and testing data sets by tool and hand gestures.

	Gestures	True positive (%)	False positive (%)
Tool	Smoothness	72.2	17.7
	Snaps	66.7	34.8
	Movement efficiency	60.0	29.7
	Tremor	80.0	4.5
Hand	Depth perception	72.2	16.0
	Bimanual dexterity	94.4	1.6

puter vision algorithms to accurately track the surgical tool-tips, the surgeon's hands and the objects in the surgical scene. The motion detection function tracked the position and direction of the surgical tools and the surgeon's hand movements. It then used this information to extract a number of psychomotor skill assessment features which included jerkiness, snaps, tremor, movement efficiency, perception of depth and bimanual dexterity. For each of the features, ANOVA analysis showed a statistically significant difference in variance between the consecutive expertise groups, i.e. novices-vs-intermediates and intermediates-vs-experts. Also when classified into 3 groups for each of these features (using Random Forest classifier), it showed an average of 74% correctly classified groups according to the expertise level. Each of these videos were de-identified and sent to three experts to be rated in three categories: smoothness, efficiency, and precision. Only scores with high inter-rater reliability were utilized in the development of the scoring algorithms. Sets of three of the extracted features were found to be associated with one of the rated features by the experts. For example, the smoothness score was associated with the jerkiness, snaps and tremor; the efficiency score is associated with movement efficiency, depth perception, and bimanual dexterity; and precision was associated with the number of error committed and the time taken to complete the exercise. A multivariate linear regression model was built for each of the categories and regression parameters were found using the tool "Weka". The entire dataset was split into a 60% training set, 10% cross-validation set and 30% testing set to find the parameters and test the regression models. Analysis results showed a very high correlation (91% aver-

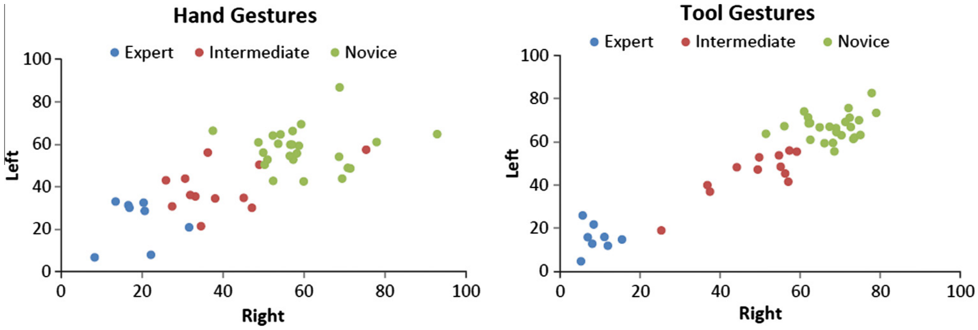

Fig. 16. Linear discriminant analysis of hand and tool gestures at different levels of expertise.

Table 6
Correlation between raters' average and predicted scores.

Features		Correlation coefficient	*p*-value
Smoothness	4.86 − 0.74 × (jerk) − 0.29 × (snap) − 0.92 × (tremor)	0.9241	<0.0001
Efficiency	1.67 − 0.41 × (movement efficiency) + 0.64 × (bimanual dexterity) − 0.94 × (depth perception)	0.9028	<0.0001

Table 7
Correlation between detected and actual errors on three tasks.

		Correlation coefficient	*p*-value
Peg transfer	Drop count	0.98	<0.0001
Intracorporeal suturing	Tissue damage	0.77	0.0008
Shape cutting	Precision of cut	0.86	<0.0001

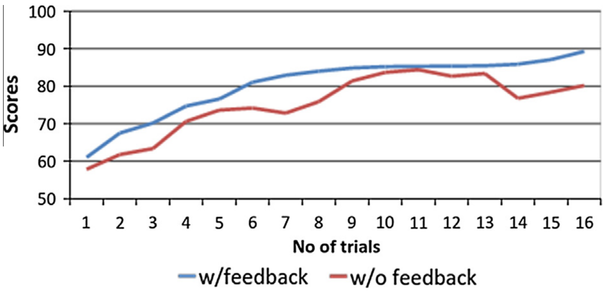

Fig. 17. Learning curve over the number of trials with and without feedback.

age) of experts' ratings with significant *p*-values. The system automatically detected the number of errors committed during each exercise. Errors in this situation were very task specific and identified by continuous observations by the expert observers. Overall the system successfully detected 87% of the errors automatically, thus supporting the hypothesis that the skill assessment provided by the proposed automatic quantitative scoring system is equivalent to that of the gold standard.

Although our current findings contrasted an expert group with a novice group, expertise is best viewed as a continuum with a number of levels that result in unique performance characteristics. The development of expertise is marked by specific transitions corresponding to reorganizations of knowledge and non-monotonic (not linear) increases in the learning curve [31]. To observe the effect of the feedback system on the process of development of expertise, we conducted an experiment with 54 medical students where each performed 2 specific FLS box exercises 16 times. None of the subjects had any prior experience with the task. Thirty-two of the novices were used as the control group where no performance feedback was provided during the learning period. They simply followed the provided guidelines for the tasks. For the other 22 novices (experimental group), on-screen real-time feedback and performance score were provided subsequent to each exercise.

The performance curve for the control group of novices who were not provided any feedback during the course of the 16 exercise session showed an "intermediate effect" in their learning curve. Intermediate effect is defined by Patel et al. [32,33] as "a temporary decline in performance as knowledge is acquired and organized, when a linear increase in performance with the length of training or time on task would be expected". On the other hand, the performance curve of the experimental group exhibited a more steady performance throughout the learning session. Although the rate of learning became saturated after a certain number of trials, the learning curve did not show the intermediate effect in the experimental group. Paired t-test showed significant improvement in the performance of the experimental group over the control group. The average group score showed an 11% better performance for the experimental group over the control group, thus supporting the second hypothesis that the immediate feedback system would increase training effectiveness by reducing the time it takes to attain a desired level of proficiency.

Finally, we developed an Internet based tool where a user could upload FLS exercise videos and receive an immediate assessment score. Users could also track their progress by observing their past scores. All these videos were de-identified and added to an online surgical video repository. Expert raters could observe these videos and provide performance score for smoothness, precision and efficiency features. Once a video rating by a minimum of three expert was obtained, the score was used to retrain the neural network (see Fig. 18).

The main limitation of our study was the dearth of surgical videos. Although videos of the surgical procedures were readily available; capturing hand video required placing an additional camera on the system. Moreover, each tool and hand video was rated by three expert surgeons in three different categories, and only videos with high inter-rater reliability were included in the study. This constraint reduced the total number of usable FLS videos. Despite these shortcomings, there was significant statistical significance in the reliability and validity of the assessment tool. For future studies, we aim to increase this number by capturing more surgical videos. The web-based assessment application is expected be a useful tool for acquiring a large number of video ratings from experts.

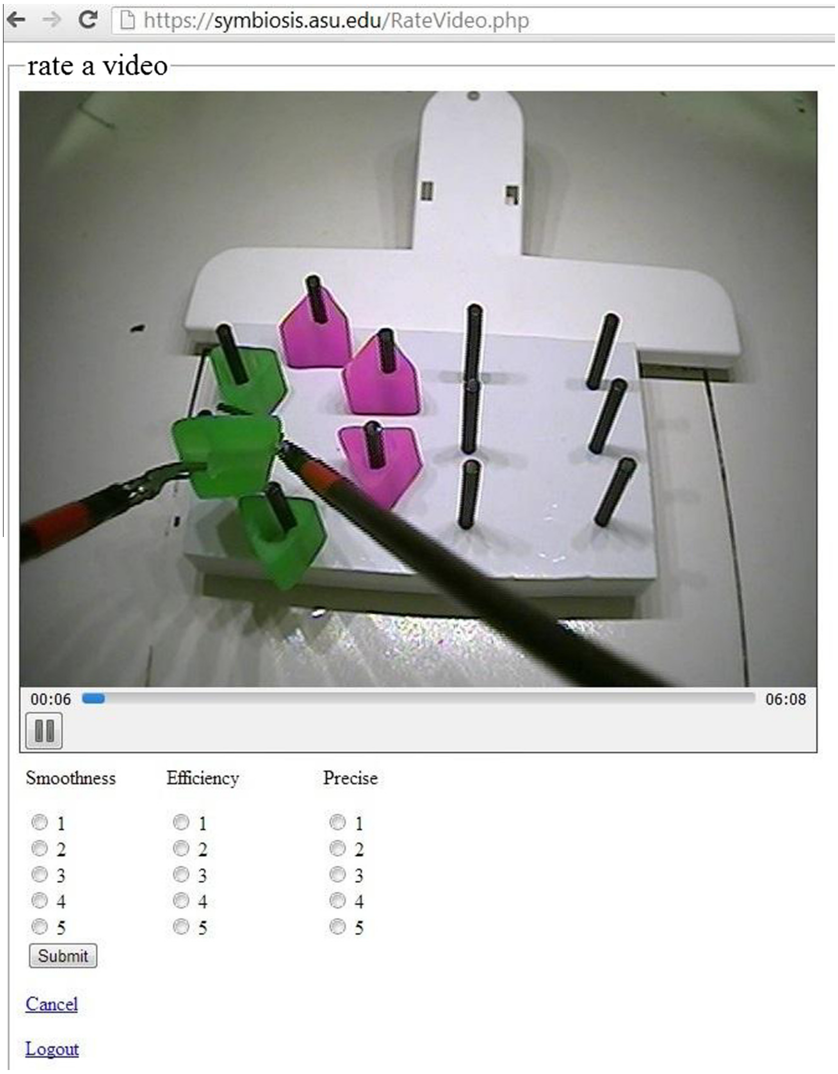

Fig. 18. Internet based video assessment tool for efficiency, smoothness and precision.

6. Conclusion

This research presents a video-based surgical skill assessment technique. We used a computer vision algorithm to analyze the video of a surgeon's hand and surgical tool movement, and extract features like surgical tool-movement smoothness, movement efficiency, individual gesture proficiency, and task specific errors. Data from different surgical residents at various level of training and expert surgeons were collected to train and test the algorithm. Since the research analyzes video data of the surgery rather than any wearable sensors, it is cost effective and also overcomes the drawbacks of most of surgical skill assessment techniques presently available.

The proposed video-based surgical skill assessment technique can provide real-time on-screen feedback, so it is also being tested as a tool for an efficient skill learning technique. After analyzing data from the experiment with an on-screen feedback system, the results showed steeper learning curve than the system without the feedback. More data is being collected for analysis to help further strengthen this hypothesis.

Objective evaluation remains the holy grail of this line of experimentation, and to that effect the ultimate achievement would be an acceptance of this tool by the various Boards of Surgical Specialties as a validated tool. While there are several challenges that need to be addressed to achieve this, such as large scale multisite trials and repeatability, the work presented here lays the foundation for such experimentation. A huge opportunity lies in addressing the need for objective evaluation of both cognitive and psychomotor surgical skills concurrently. We also need to understand how surgical errors evolve [34], and if such a system can help predict an error before it occurs. For example, could increased snap values potentially predict an impending mistake? While these remain important questions, the work here takes the first step at developing a comprehensive, affordable and scalable approach to surgical proficiency determination. In addition to the developed informatics tool presenting a practical solution, it hopefully will also encourage research and investigation in this area.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.11.002.

References

[1] F. Treves, A Manual of Operative Surgery, Cassell and Company, London, 1891.
[2] J. Ende, Feedback in clinical medical education, JAMA 250 (6) (1983) 777–781.
[3] L.T. Kohn, J.M. Corrigan, M.S. Donaldson, To. Err, To Err is Human: Building a Safer Health System, National Academy Press, Washington, DC, 2000.

[4] L.M. Sutherland, P.F. Middleton, A. Anthony, J. Hamdorf, P. Cregan, D. Scott, G.J. Maddern, Surgical simulation: a systematic review, Ann. Surg. 243 (3) (2006) 291–300.

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

[6] A. Dosis, R. Aggarwal, F. Bello, Synchronized video and motion analysis for the assessment of procedures in the operating theater, Arch. Surg. 140 (3) (2005) 293–299.

[7] R. Aggarwal, T. Grantcharov, K. Moorthy, T. Milland, A. Darzi, Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room, Ann. Surg. 247 (2) (2008) 372–379.

[8] R. Aggarwal, T. Grantcharov, K. Moorthy, T. Milland, P. Papasavas, A. Dosis, F. Bello, A. Darzi, An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room, Ann. Surg. 245 (6) (2007) 992–999.

[9] K. Kahol, M. Vankipuram, Hand motion expertise analysis using dynamic hierarchical activity modeling and isomap, in: 19th International Conference on Pattern Recognition, 2008, ICPR 2008, 2008.

[10] I. Oropesa, P. Sánchez-González, P. Lamata, M.K. Chmarra, J.B. Pagador, J.A. Sánchez-Margallo, F.M. Sánchez-Margallo, E.J. Gómez, Methods and tools for objective assessment of psychomotor skills in laparoscopic surgery, J. Surg. Res. 171 (1) (2011) e81–e95.

[11] J. Lynch, P. Aughwane, T.M. Hammond, Video games and surgical ability: a literature review, J. Surg. Educ. 67 (3) (2010) 184–189.

[12] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, O"Reilly Media Inc., Sebastopol, CA, 2008.

[13] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice-Hall Inc., USA, 2001.

[14] R. Aggarwal, T. Grantcharov, K. Moorthy, J. Hance, A. Darzi, A competency-based virtual reality training curriculum for the acquisition of laparoscopic psychomotor skill, Am. J. Surg. 191 (1) (2006) 128–133.

[15] Fundamentals of Laparoscopic Surgery... the definitive laparoscopic skills enhancement and assessment module. <http://www.flsprogram.org/>.

[16] J. Chen, M. Yeasin, R. Sharma, Visual modelling and evaluation of surgical skill, Pattern Anal. Appl. 6 (2003) 1–11.

[17] M. Aizuddin, N. Oshima, R. Midorikawa, A. Takanishi, Development of sensor system for effective evaluation of surgical skill, in: The First IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, 2006, BioRob 2006, 2006.

[18] P.J. Driscoll, A.M. Paisley, S. Paterson-Brown, Video assessment of basic surgical trainees' operative skills, Am. J. Surg. 196 (2) (2008) 265–272.

[19] R. Gonzalez, J. Martinez, E. Menzo, A. Iglesias, C. Ro, A. Madan, Consumer-based technology for distribution of surgical videos for objective evaluation, Surg. Endosc. 26 (8) (2012) 2179–2182.

[20] C. Wen, Y.Q. Shi, X. Guorong, Identifying computer graphics using HSV color model and statistical moments of characteristic functions, in: 2007 IEEE International Conference on Multimedia and Expo, 2007.

[21] A.R. Smith, Color gamut transform pairs, SIGGRAPH Comput. Graph. 12 (3) (1978) 12–19.

[22] S. Sural, Q. Gang, S. Pramanik, Segmentation and histogram generation using the HSV color space for image retrieval, in: Proceedings of the 2002 International Conference on Image Processing, 2002.

[23] M. Shah, Fundamentals of Computer Vision, University of Central Florida, Orlando, FL, 1997.

[24] C.N. Riviere, R.S. Rader, N.V. Thakor, Adaptive canceling of physiological tremor for improved precision in microsurgery, Biomed. Eng. (1998).

[25] M.J. Richardson, T. Flash, Comparing smooth arm movements with the two-thirds power law and the related segmented-control hypothesis, J. Neurosci. 22 (18) (2002) 8201–8211.

[26] E.D. Grober, M. Roberts, E. Shin, M. Mahdi, V. Bacal, Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence, Am. J. Surg. 199 (1) (2010) 81–85.

[27] B. Pamela, E. Andreatta, M. Derek, T. Woodrum, M. Rebecca, M. Minter, Laparoscopic skills are improved with lapmentor™ training, Ann. Surg. 243 (6) (2006) 854–863.

[28] A.A. Gumbs, N.J. Hogle, D.L. Fowler, Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills, J. Am. Coll. Surg. 204 (2) (2007) 308–313.

[29] M.C. Vassiliou, L.S. Feldman, C.G. Andrew, S. Bergman, K. Leffondré, D. Stanbridge, G.M. Fried, A global assessment tool for evaluation of intraoperative laparoscopic skills, Am. J. Surg. 190 (1) (2005) 107–113.

[30] S.P. Rodrigues, T. Horeman, J. Dankelman, J.J. van den Dobbelsteen, F.W. Jansen, Suturing intraabdominal organs: When do we cause tissue damage?, Surg Endosc. 26 (4) (2012) 1005–1009.

[31] V.L. Patel, G.J. Groen, Developmental accounts of the transition from medical student to doctor: some problems and suggestions, Med. Educ. 25 (6) (1991) 527–535.

[32] V.L. Patel, J.F. Arocha, D.R. Kaufman, A primer on aspects of cognition for medical informatics, J. Am. Med. Inform. Assoc. 8 (4) (2001) 324–343.

[33] V.L. Patel, D.R. Kaufman, J.F. Arocha, Emerging paradigms of cognition in medical decision-making, J. Biomed. Inform. 35 (1) (2002) 52–75.

[34] V.L. Patel, T. Cohen, New perspectives on error in critical care, Curr. Opin. Crit. Care 14 (4) (2008) 456–459.