Int J CARS (2015) 10:981–991
DOI 10.1007/s11548-015-1194-1

ORIGINAL ARTICLE

# Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty

**Narges Ahmidi[1] · Piyush Poddar[3] · Jonathan D. Jones[4] · S. Swaroop Vedula[1] · Lisa Ishii[2] · Gregory D. Hager[1] · Masaru Ishii[2]**

Received: 9 March 2015 / Accepted: 25 March 2015 / Published online: 17 April 2015


**Abstract**

*Purpose* Previous work on surgical skill assessment using intraoperative tool motion has focused on highly structured surgical tasks such as cholecystectomy and used generic motion metrics such as time and number of movements. Other statistical methods such as hidden Markov models (HMM) and descriptive curve coding (DCC) have been successfully used to assess skill in structured activities on bench-top tasks. Methods to assess skill and provide effective feedback to trainees for unstructured surgical tasks in the operating room, such as tissue dissection in septoplasty, have yet to be developed.

*Methods* We proposed a method that provides a descriptive structure for septoplasty by automatically segmenting it into higher-level meaningful activities called strokes. These activities characterize the surgeon's tool motion pattern. We constructed a spatial graph from the sequence of strokes in each procedure and used its properties to train a classifier to distinguish between expert and novice surgeons. We compared the results from our method with those from HMM, DCC, and generic metric-based approaches.

*Results* We showed that our method—with an average accuracy of 91 %—performs better or equal than these state-of-the-art methods, while simultaneously providing surgeons with an intuitive understanding of the procedure.

*Conclusions* In this study, we developed and evaluated an automated approach to objectively assess surgical skill during unstructured task of tissue dissection in nasal septoplasty.

✉ Narges Ahmidi
nahmidi1@jhu.edu

Piyush Poddar
ppoddar1@jhu.edu

Jonathan D. Jones
jdjones@jhu.edu

S. Swaroop Vedula
vedula@jhu.edu

Lisa Ishii
learnes2@jhmi.edu

Gregory D. Hager
hager@jhu.edu

Masaru Ishii
mishii3@jhmi.edu

1 Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

2 Department of Head and Neck Surgery-Otolaryngology, Johns Hopkins Medical Institutes, Baltimore, MD, USA

3 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

4 Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

**Keywords** Unstructured activities · Partially observed time series · Surgical skill assessment · Feature extraction · Septoplasty · Feedback

## Introduction

Traditional methods of surgical skill evaluation have been subjective, with supervising surgeons subjectively evaluating trainee surgeon competence and skill in relatively unstandardized methods [1,2]. Recent changes to the accreditation process for surgery training programs require that these programs now measure a trainee's competence and surgical skill objectively. However, valid, quantitative methods for assessing technical surgical skill are rare, and the need for the development of these methods is great.

Septoplasty is a commonly performed surgery designed to relieve nasal obstruction. It achieves this goal by correcting

deviations from midline of the nasal septum—the structure that separates the nasal passage and nostrils into left and right sides. Deviations are a common anatomic cause of narrowing of the nasal airway aperture that leads to increased airway resistance and a sensation of nasal obstruction. One of the most important steps in septoplasty involves separating the skin of the septum from the underlying bone and cartilage. This step is typically performed with a Cottle elevator—a thin long tool with a dull blade at the tip that can be used to cut adhesions, regions where the skin is tightly bound to the underlying structure, thus separating the skin from the bone and cartilage. After this step, the bone and cartilage can be appropriately assessed and irregularities corrected [3,4].

Septoplasty is a high volume index procedure (260,000 cases in 2006, USA) performed by head and neck surgeons [5]. Training programs must ensure that residents are competent to perform this surgery before they graduate. However, this procedure still lacks an objective, standardized, data-driven metric for evaluating technical skill. Current methods for assessment rely upon the number of procedures performed and other subjective approaches that are not standardized across training programs [6].

Surgical procedures such as septoplasty pose unique challenges to traditional subjective skill assessment approaches. The first and foremost issue is that the evaluating surgeon and the trainee surgeon cannot both look at the surgical field at the same time, because the surgical site is literally within the nose, with access and visualization of the surgical site occurring through the nostril. Because the nostril is a small opening, the operating surgeon's head looking down into the nasal passage blocks an observer's view, i.e., only one surgeon can view the surgical site at a time. This makes both teaching and evaluating septoplasty difficult. Second, the most important step in the procedure, mucosal flap elevation, involves unstructured surgical tool motion that cannot be decomposed into a sequence of predefined segments. Similar unstructured tool motion is seen in many other surgical procedures requiring blunt and sharp tissue dissection. Other challenges for objective skill assessment during septoplasty include patient-specific variations in anatomy and constant changes in the reference coordinate frame due to patients' head movements [1].

Existing techniques for objective surgical skill assessment [7–11] are not applicable to unstructured tool motion in septoplasty. The state of the art for such assessment metrics in laparoscopy procedures is summarized in [12]. Generic aggregate metrics of time and motion efficiency [10,11] may not be computed in the septoplasty context because the amount of dissection that is required varies across patients [13]. Other statistical approaches such as hidden Markov models (HMMs) [14,15] and descriptive curve coding (DCC) [16,17] have previously been applied to objectively assess skill for structured tasks performed on inanimate bench-top models. These statistical methods have yet to be applied to objectively assess surgical skill in the operating room for procedures such as septoplasty, which involve unstructured surgical tool motion. However, describing tool motion using generic metrics or statistical models would still not help trainees learn how to perform the procedure.

Our primary goal was to elucidate the structure of tool motion observed in septoplasty such that faculty surgeons can effectively teach the procedure, assess performance, and provide meaningful feedback to trainees, and such that trainees can efficiently understand the intention and nature of the tool motions, guiding strategies, and abstract planning during septoplasty.

In this paper, we propose a method that provides a descriptive structure for septoplasty by automatically segmenting it into higher-level activities. These activities characterize surgeons' tool motion patterns, translate into clinically meaningful information, and encode surgical skill. We evaluate our method for its ability to discriminate between levels of surgical skill and compare our proposed method with three state-of-the-art methods: HMMs, DCC, and generic metrics computed from the raw signal (Fig. 1). HMMs and DCC have been used successfully in surgical skill assessment for structured tasks, such as suturing, which are performed in a sequence of actions. The generic metrics serve as a baseline comparison and provide a first-hand insight to the low-level signal structure. Finally, we discuss the advantages and drawbacks of our proposed method versus the three other methods in terms of their ability to deal with the live-patient surgical signal, unstructured multi-surgeon tool motions, and whether their output can be translated into actionable feedback.

## Experimental setup

We used data from a cohort study of five faculty surgeons (experts) and nine trainee resident and fellow surgeons (novices), conducted over 2 years and across five sites. We captured data at all sites using a common protocol and equipment. Our data collection procedure did not interfere with routine patient care in any manner, and all pertinent data collection equipment was sterilized according to the standard operating room procedures before each case. The institutional review board at Johns Hopkins approved our study.

The signal for our analyses comprised of kinematic data describing motion of the Cottle elevator during septoplasty. As shown in Fig. 2, we affixed an electro-magnetic (EM) sensor (6° of freedom) to the Cottle elevator ("Cottle sensor"). We performed a pivot calibration for both tips of the Cottle elevator before each procedure. The patient's head moves during surgery as a result of dissection with the Cottle elevator. For about 70 % of the procedures in our study, we measured patients' head motion using a second sterilized sen-

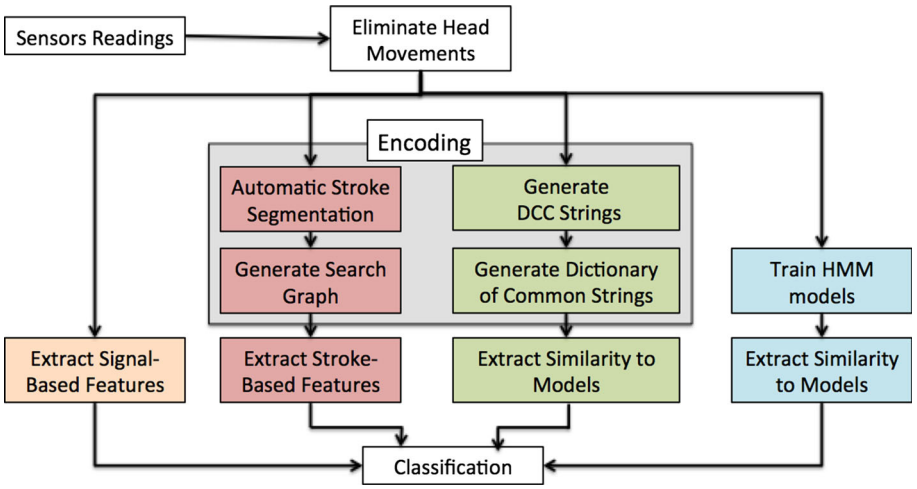



**Fig. 1** Overview of the proposed approach (*red*), in comparison with three state-of-the-art approaches, DCC (*green*), HMM (*blue*), and baseline generic metrics

sor placed within the folds of a sterile towel tightly wrapped around the patient's head ("head sensor"). We used an EM field generator (Aurora®, Northern Digital, Inc., Ontario, Canada) to track the Cottle and head sensors.

We captured video recordings of the procedure to annotate for segments when the Cottle was in use and when it was idle, which of the two ends of the Cottle elevator were being used, and for the operating surgeon when more than one surgeon performed the procedure. We processed data only from segments of the procedure when the Cottle elevator was in use. We recorded video of the procedure using two Kinect®devices (Microsoft, Inc., Redmond, WA, USA) rigidly fixed to a tripod. We developed custom data collection software to capture synchronized video and kinematic data.

Our data collection process involved minimal input from the operating surgeon, who circled the perimeter of the nose with one tip of the Cottle elevator at the beginning of each surgery. We used the data from this segment to register the tool tip with the location of the patient's nose (Fig. 4, left).

In this study, we used data from 86 procedures. An expert surgeon operated in 60, and a novice operated in 26 procedures. Both an expert and a novice surgeon operated in 14 procedures. These multi-surgeon cases add to the complexity of the analysis since we partially observed the surgeon, with the possibility that her performance is influenced by that of the previous surgeon.

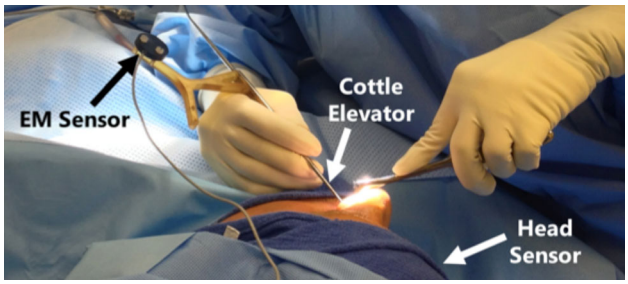



**Fig. 2** Cottle elevator with electromagnetic sensor used by surgeons in this study to elevate the mucosal flap in a septoplasty procedure. A second EM sensor (head sensor; not visible in picture) was used to measure patients' head motion during surgery

## Methodology

In this section, we describe methodology for the four different approaches we took to build a binary classifier for surgical skill assessment in septoplasty: stroke-based, signal-based, DCC [16], and HMM [14].

As the first step, we automatically subtract the patients' head movements from the tool motion signal (Fig. 1). In our primary approach, we automatically extracted higher-level structures called "strokes" from the tool signal and built a search graph based on the strokes. We represent each graph with a set of features and use them to classify trials as expert or novice.

In our second approach, we investigated generic features that can be extracted from the raw tool motion signal (before stroke segmentation), such as profiles of velocity, acceleration, and dominant frequencies. In two other model-based approaches, we used HMMs and DCC-based string dictionaries to compute measures of similarity to expert and novice models. The similarity measures then served as features for classification purposes.

### Subtracting patient's head motions

Let $O_{1:N}^{(w)}$ be the Cottle tip positions during a recorded surgical trial. Each data point $O_f^{(w)}$ is a Cartesian position of the tool tip in the world coordinate system. The origin of the world coordinate system is associated with the EM field generator. The world coordinate system changes from one trial to the next, and the working field (the patient's head) is also not

stationary over the duration of a trial. Consequently, similar surgical actions are represented differently in the world coordinate frame. We therefore transformed tooltip data from the world coordinate frame to the coordinate system of the septal plane, which provides a more consistent reference frame for data across all procedures in our dataset.

We defined an orthonormal coordinate system $\langle \mathbf{u}_x^f, \mathbf{u}_y^f, \mathbf{u}_z^f \rangle$ originating at the center of the septal plane with $u_z$ being the normal of the plane (Fig. 3). Because the septal coordinate system moves with respect to the world coordinate system, we estimated its location at each time frame $f$. We estimated the initial septal plane $\langle \mathbf{u}_x^0, \mathbf{u}_y^0, \mathbf{u}_z^0 \rangle$ as the plane formed by the first and third principal components of the active tool tip trajectory during nose circling registration. We then updated the location of the septal plane at each time frame by subtracting head movements measured from the reference sensor.

Having obtained the septal coordinate system and its center location $\mathbf{c}_f$ at each time frame $f$, we transferred the raw observation $O_f^{(w)}$ from the world frame to the septal frame $O_f^{(s)}$ (for simplicity $O_f$) where $O_f = [x_f, y_f, z_f]^T$.

$$O_f^{(s)} := [\mathbf{u}_x^f, \mathbf{u}_y^f, \mathbf{u}_z^f]^T (O_f^{(w)} - \mathbf{c}_f)$$

In the case of procedures for which we did not use a head sensor, we estimated the location of the septal plane as a proxy measure for head motion. We assumed that the patients' head motion during surgery was constrained to one degree of freedom (side-to-side rotation). The estimate for the septal plane at a point in time was the plane that best fit the most recent tooltip data. We validated our method to estimate the septal plane and head motion using data from procedures where we used a head sensor. On average, the septal plane we estimated using our approach differed from the true septal plane by 3.4° and 7 mm.

### Automated segmentation of strokes

We identified that surgeons use stroking motions of the Cottle to elevate the mucosal flap off the underlying cartilage, based on insights about the septoplasty procedure from expert surgeons and exploratory analyses (Fig. 3, left). We developed a method to automatically extract strokes using tool tip kinematic data. We defined a stroke $S_i$ to start ($s_i$) when the Euclidean distance from the active tool tip to the septal plane is at a local minimum, and end ($e_i$) at the nearest frame following $s_i$ when the distance from the active tool tip to the septal plane is at a local maximum.

$$S_i = O_{s_i:e_i}$$

To avoid detection of extraneous strokes, we applied a moving average filter to smooth the tool tip position data, and constraints to the stroke duration, length, and distance from the beginning or end of strokes to the center of the nose (Fig. 3).

### Search graphs

Strokes made with the Cottle elevator during septoplasty encode clinically meaningful information about the procedure. Each stroking motion during septoplasty serves to achieve elevation of the mucosal flap away from the septal cartilage and bone, eventually covering the area of the septum while searching for adhesions between the mucosa and the cartilage. These objectives are clinically meaningful. Excessive force applied to elevate the mucosa may cause tearing and results in septal perforation, which leads to undesirable postoperative outcomes. The extent and rate of septal plane coverage reflects the surgeon's efficiency in elevating the mucosal flap. Our goal is to define and extract features that can characterize these activities and explain the differences between expert and novice movements.

We extracted the semantic information encoded by strokes during septoplasty by representing each procedure as nodes on a graph. We used the stroke segments to form a two-dimensional directed graph $G = (V, A)$ on the septal plane, where $V$ is a set of vertices and $A$ is a set of arcs. Each vertex $v_i$ represents one stroke that we computed by projecting the starting position of the stroke onto the septal plane:

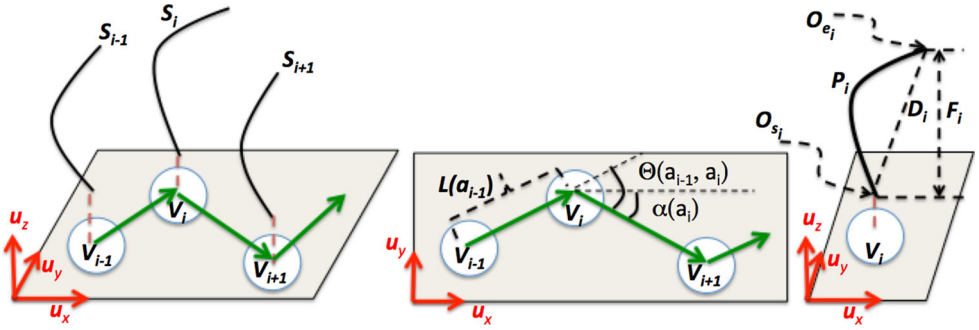
$$v_i := [O_{s_i} \cdot \mathbf{u}_x^{s_i},\ O_{s_i} \cdot \mathbf{u}_y^{s_i}]$$

Each arc $a_i$ is a directed edge from $v_i$ to $v_{i+1}$. Because these graphs reveal the surgeon's search pattern on the septal plane, we call it a "search graph". Figure 4 shows an example search graph for septoplasty procedures performed by expert and novice surgeons.

### First approach: stroke-based features

We defined a set of functions on the graph arcs and vertices to extract features from the graph (Fig. 3). The off-plane functions measure the trajectory length of the stroke $P_i$, distance travelled by the stroke $D_i$, height of the stroke $F_i$, and the time of completion $T_i$. The on-plane functions compute the length of the arcs $L(a_i)$, the absolute angle of the arcs $\alpha(v_i)$, and the relative angles $\theta(v_i)$ between two adjacent arcs.

We employed this set of functions to compute features describing each stroke. We specified seven features (described below) based on a hypothesis that experts repeat strokes more consistently and regularly, and with greater efficiency relative to novice surgeons. The off-plane features include Stroke Curvature Consistency, Stroke Duration Consistency, and Stroke Height Distribution, while on-plane features consist of Arc Length Distribution, Absolute Arc



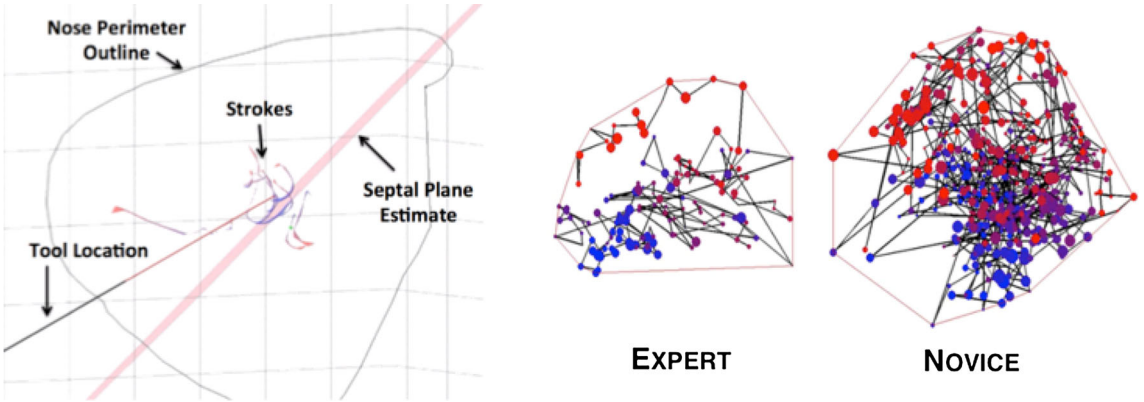




**Fig. 3** (*Left*) representation of the segmented strokes in the form of a graph that connects the starting positions of consecutive strokes after they have been projected onto the septal plane; (*center*) on-plane graph features; and (*right*) off-plane features per stroke

**Fig. 4** (*Left*) three-dimensional visualization of a detected stroke brushing away from septal plane, the PCA-estimated septal plane (*pink*), and manually drawn outline of the nose (*black*) by surgeon. (*Right*) expert and novice two-dimensional search graphs on the septal plane. Color (*blue* to *red*) indicates progression of time. The vertex size is proportional to the height of a stroke. The red outline marks the convex hull of the covered area on the septum

Angle Distribution, Relative Arc Angle Distribution, and Area Coverage Rate.

*Stroke curvature consistency* We hypothesized that surgeons need to perform different types of strokes to adapt to anatomy within the nose. Consequently, we computed measures of local variance instead of the standard global variance measures. Stroke Curvature Consistency (SCC) measures the local consistency of stroke curvatures across strokes. We anticipated that expert surgeons will have more consistent wrist motion that yield strokes with similar curvatures, and therefore a lower SCC, than novice surgeons. We computed the curvature of the $i$th stroke, $C(S_i)$, as the ratio of the stroke trajectory length $P_i$ to the Euclidean distance $D_i$ between the start and end points of the stroke:

$$C(S_i) := P_i / D_i$$

where

$$P_i := P(S_i) := \sum_{f=s_i}^{e_i-1} \|O_{f+1} - O_f\| \qquad D_i := D(S_i) := \|O_{e_i} - O_{s_i}\|$$

We then measured the local consistency of curvatures $\bar{C}(S_i)$ by computing the squared distance between $C(S_i)$ and its local smoothed representation $\mu_i$:

$$\bar{C}(S_i) := (C(S_i) - \mu_i^C)^2 \qquad \mu_i^C = median([C(S_{i-2}) \ldots C(S_{i+2})])$$

We defined the SCC for a trial to be the median of the $\bar{C}$ values of that trial.

*Stroke duration consistency* Stroke Duration Consistency (SDC) measures the local consistency in stroke duration:

$$\bar{T}(S_i) := (T(S_i) - \mu_i^T)^2 \qquad \mu_i^T = median([T(S_{i-2}) \ldots T(S_{i+2})])$$

where $T(S_i) := e_i - s_i$ is the time to finish one stroke. We hypothesized that strokes by expert surgeons will have higher local consistency and therefore lower SDC, when compared with strokes by novice surgeons.

*Stroke height distribution* We computed the stroke height as a proxy for the force applied to elevate the mucosa by measuring the Euclidean distance between the start and end positions of the stroke.

$$D(v_i) := \|O_{e_i} - O_{s_i}\|$$

For a given graph, we computed a histogram from the height of the strokes. We hypothesized that expert surgeons more finely modulated the force they apply to elevate the mucosa than novice surgeons. We anticipated that expert surgeons apply greater force in areas where the flap can be easily elevated and lesser force in areas of greater adhesions (around bone–bone and bone–cartilage junctions) to avoid tearing the mucosal flap. In contrast, we expected novice surgeons to make tentative motions and apply small but uniform amounts of force throughout the procedure.

*Arc length distribution* For a given graph (Fig. 3, center image), we computed the length of all the arcs and then constructed a normalized histogram. The location of each vertex $v_i$ on the septal plane is computed by projecting the starting position of the stroke into the septal plane:

$$X(v_i) := [O_{s_i} \cdot \mathbf{u}_x^{s_i}, O_{s_i} \cdot \mathbf{u}_y^{s_i}]$$

Then, the length of an arc is defined as the distance between its head and tail vertices:

$$L(a_i) := \|X(v_{i+1}) - X(v_i)\|$$

We hypothesized that experts' graphs contain longer arcs, because at each stroke, they manage to elevate larger areas of the mucosal flap. Therefore, they can access farther points on the septum for the subsequent strokes.

*Absolute arc angle distribution* We computed the distribution of absolute angles of arcs as a measure of the probability of choosing a certain direction for the next stroke. We hypothesized that expert surgeons followed a more consistent pattern (a more uniform distribution of absolute arc angles) when compared with novice surgeons. We computed the normalized histogram of the absolute angle of the arcs $\alpha(a_i)$. The absolute angle of each arc is defined as the angle between the arc and the basis of the septal plane coordinate system $\mathbf{u}_x$:

$$\alpha(a_i) := \arccos\left((X(v_{i+1}) - X(v_i)) \cdot \mathbf{u}_x / L(a_i)\right)$$

The absolute arc angles explain the search pattern with respect to the septal plane basis and thus sensitive to anatomical variations in the septum across patients.

*Relative arc angle distribution* To mitigate the previous feature's anatomical sensitivity, we computed the relative angle between two adjacent arcs. Instead of representing the search pattern in the basis of the septal plane, this new feature encodes the pattern of changes in decision making. We calculated the normalized histogram of the relative arc angles. These relative arc angles (called $\theta(a_{i-1}, a_i)$, as shown in Fig. 3, center image), were computed as:

$$\theta(a_{i-1}, a_i) := \alpha(a_i) - \alpha(a_{i-1})$$

We hypothesized that expert surgeons maintain a structured search pattern, which leads to graphs that contain many instances of smaller $\theta$ values. On the other hand, novice surgeons display random search patterns that result in graphs that are uniformly distributed in $\theta$.

*Area coverage rate (ACR)* We defined the area covered at node $v_i$ as the area inside a convex hull of the search graph after completion of the $i$th stroke (denote as $AC(v_i)$). The convex hull covers the finite set of points, $v_{1:i}$, consisting of the vertex $v_1$ to $v_i$. We defined the ACR at node $v_i$, $ACR(v_i)$, to be the increase in $AC$ with each stroke:

$$ACR(v_i) := AC(v_i) - AC(v_{i-1}) := Convex\ Area(v_{1:i}) - Convex\ Area(v_{1:i-1})$$

We defined the ACR for a given trial to be the median of all the $ACR(v_i)$ values. ACR is an index of the efficiency with which surgeons search for adhesions between the mucosa and underlying cartilage. We hypothesized that expert surgeons will have larger ACR than novice surgeons because experts elevate large areas of the mucosal flap with each stroke. We anticipated that novice surgeons make tentative strokes that do not fully elevate the mucosal flap and therefore return to dissect previously elevated parts of the septum, leading to small ACR values.

**Second approach: signal-based features**

As the second approach for skill classification, we defined three generic features derived from the raw, non-idle kinematic signal in the septal coordinate frame: velocity profile, acceleration profile, and frequency spectrum profile. These low-level features may encode patterns in motor behavior of surgeons. We hypothesized, for example, that movements by expert surgeons have a higher velocity and frequency than those by novice surgeons.

*Velocity distribution* Given that the time difference between two consecutive frames in our signal is constant, we computed the velocity of the Cottle elevator at frame $f$ as the backward differences such that $V_f = O_f - O_{f-1}$. We then computed a normalized histogram from the velocity magnitudes of each trial.

*Acceleration distribution* We estimated acceleration by calculating backward differences between successive values of velocity, expressed as $A_f = V_f - V_{f-1}$. We represented

**Table 1** Accuracy of skill classification for stroke-based features

| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | All |
|---|---|---|---|---|---|---|---|---|
| LOTO | [69.70] | [69.70] | [75.76] | [80.30] | [77.27] | [77.27] | [65.15] | [90.91] |
| | (61.90) | (74.40) | (69.35) | (79.17) | (72.32) | (71.43) | (69.94) | (87.50) |
| LOUO | [66.67] | [68.18] | [59.09] | [72.73] | [59.09] | [66.67] | [59.09] | [74.24] |
| | (59.52) | (72.32) | (53.57) | (71.43) | (56.25) | (63.10) | (65.18) | (73.51) |

Values are %[Micro] (Macro) averages. Features are Stroke Curvature Consistency (F1), Stroke Duration Consistency (F2), Stroke Height Distribution (F3), Arc Length Distribution (F4), Absolute Arc Angle Distribution (F5), Relative Arc Angle Distribution (F6), Area Coverage Rate (F7), and the combination of all features through majority voting (ALL)

**Table 2** Accuracy of skill classification for signal-based features

| | SF1 | SF2 | SF3 | All |
|---|---|---|---|---|
| LOTO | [63.64] | [60.61] | [75.76] | [71.21] |
| | (66.07) | (61.90) | (72.02) | (72.02) |
| LOUO | [45.45] | [53.03] | [43.94] | [46.97] |
| | (50.00) | (52.38) | (45.24) | (51.19) |

Values are %[Micro] (Macro) averages. Features are profile of Velocity (SF1), Acceleration (SF2), Frequency (SF3), and their combination through majority voting (ALL)

each trail with a normalized histogram of the magnitude of acceleration values.

*Frequency spectrum profile* To compute the spectral profile of the kinematic signal, we first calculated the magnitude of the signal in the septal coordinate frame and removed the mean from the signal. We then applied a 256-length fast Fourier transform. We then down-sampled the 256 dimension FFT magnitudes and use it as the feature vector. We hypothesized that, due to increased dexterity in the wrist, experts will perform more higher-frequency movements than novices.

### Third approach: hidden Markov model

As the third approach for skill classification, we modeled each skill class (expert and novice) as an HMM and predict class labels using maximum-likelihood classification [14]. We investigated a variety of HMM configurations, including different states (1–20) and Gaussian mixture components per state (3–20). We trained the HMMs using different combinations of features from the kinematic signal including position and orientation, velocity and orientation, with and without normalization, in both the septal and world frames.

### Fourth approach: descriptive curve coding

We implemented the DCC [16] approach as the fourth method for skill classification. In this method, we encoded the tool tip trajectory as a string of symbols chosen from a predefined alphabet. In the training phase, we built a dictionary of common strings per skill class. In the testing phase, we computed the likelihood of a given test sample under the expert and novice models by aggregating similarity of all the dictionary entries.

We investigated the DCC approach using two alphabets: $A1$, which consists of seven words and is sensitive to changes of direction larger than 45°, and $A2$, with 19 words and sensitivity of 22.5°. In addition, we encoded both the raw signal in the world frame as well as the signal in the septal plane to determine robustness of our findings.

**Table 3** Accuracy of Skill classification for the HMM [14] method using position, orientation, and velocity of the tracked tool

| | pos+orn (s) | pos+orn (w) | vel+orn (s) | vel+orn (w) |
|---|---|---|---|---|
| LOTO | [70.93] | [63.95] | [61.62] | [58.13] |
| | (62.82) | (58.91) | (63.78) | (53.65) |
| LOUO | [51.55] | [29.06] | [45.34] | [53.48] |
| | (55.19) | (35.00) | (52.11) | (55.76) |

Values are %[Micro] (Macro) averages for kinematics represented in the world (w) or septal (s) frame

**Table 4** Accuracy of skill classification for the DCC [16] method using two alphabets A1 and A2 (sensitive to 45 and 22.5° change of direction, respectively)

| | A1 (s) | A1 (w) | A2 (s) | A2 (w) |
|---|---|---|---|---|
| LOTO | [84.28] | [91.66] | [81.03] | [86.28] |
| | (79.84) | (88.45) | (73.80) | (81.98) |
| LOUO | [89.24] | [90.21] | [90.07] | [83.86] |
| | (86.10) | (88.92) | (89.55) | (81.58) |

Values are %[Micro] (Macro) averages for kinematics in the world (w) or septal (s) frame

## Training and evaluation

For ground truth, we considered faculty surgeons as experts and resident/fellow surgeons as novices. We used a kernel support vector machine (SVM) for classifying surgical skill (expert vs. novice) with different features as the input: stroke-

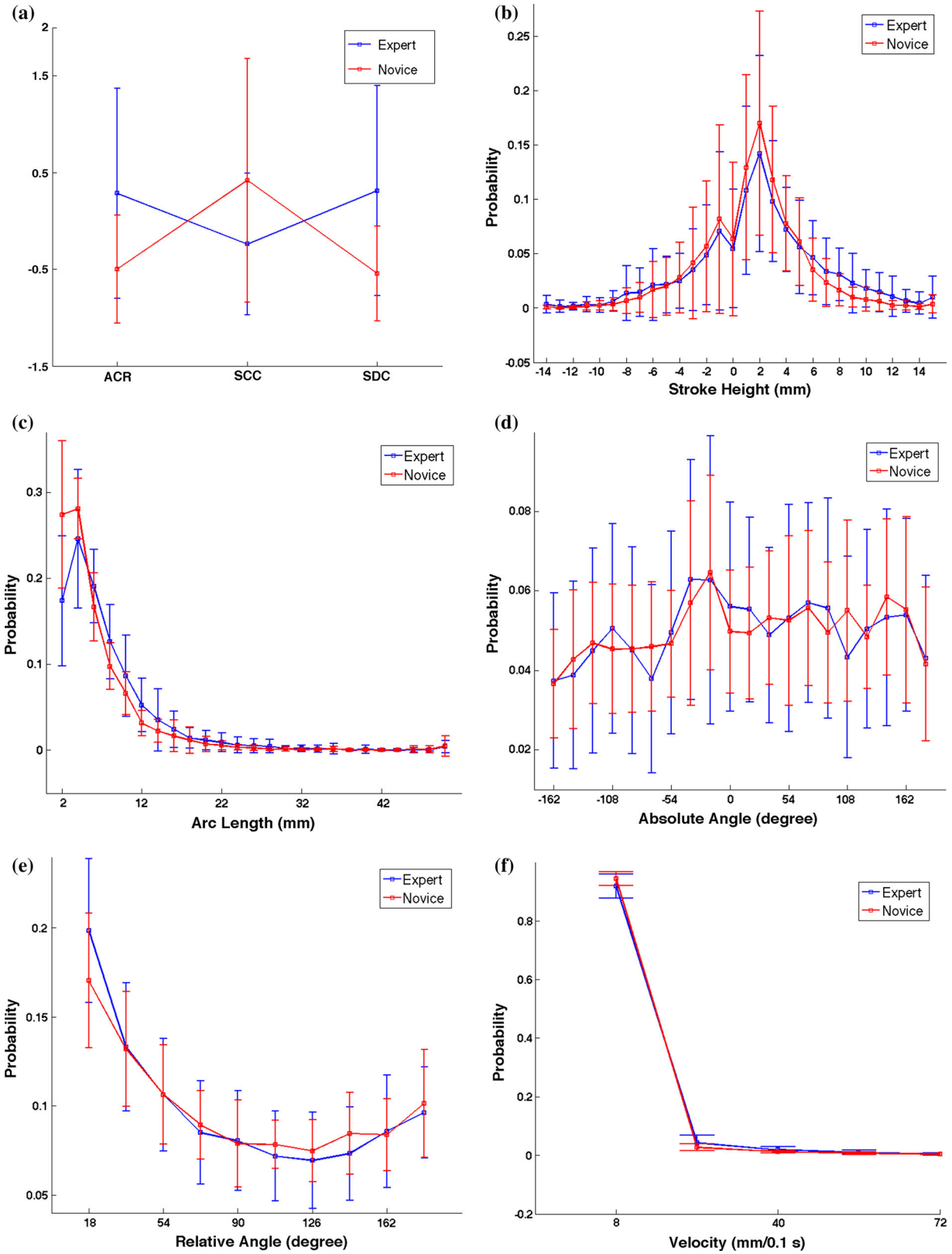



**Fig. 5** Feature representation of the average expert and novice. Stroke-based features are Area Coverage Rate (ACR), Stroke Curvature Consistency (SCC), Stroke Duration Consistency (SDC), Stroke Heights, Arc Lengths, Absolute Angles, and Relative Angles. Signal-based features are velocity, acceleration, and FFT magnitudes

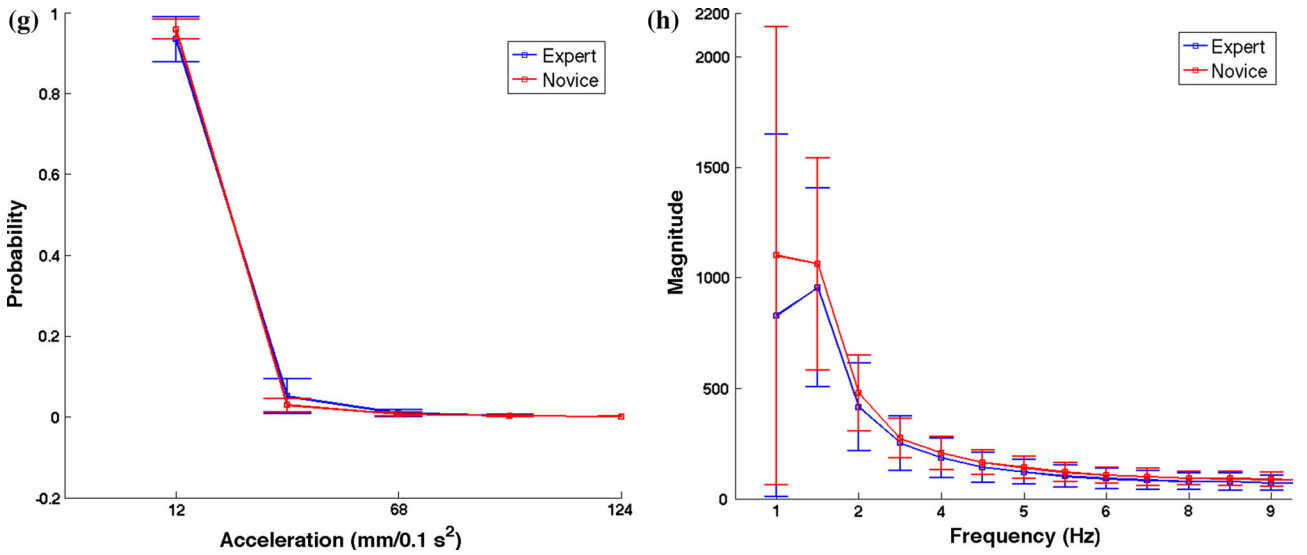



**Fig. 5** continued

based, signal-based, and DCC-based similarity features. For the HMMs, we performed maximum-likelihood classification.

We tested the classifiers under two cross-validation setups: Leave-One-Trial-Out (LOTO) and Leave-One-User-Out (LOUO). In the LOTO setup, we used one procedure (referred to as a trial) as the test data and all the remaining trials for training. In the LOUO setup, we used all trials performed by one surgeon as the test data. When a trial was performed by two surgeons, we concatenated the data from segments performed by each surgeon and considered it as a separate trial.

We computed accuracy of the classifier in two ways—the micro-average (ratio of correctly classified samples to the total number of samples) and macro-average accuracy (average of true positive rates of each class).

All histogram-based features were normalized (removing mean and standard deviation) followed by PCA dimensionality reduction. We tested multiple configurations of bin counts (between 10 and 30 with increments of five) and number of principal components (between one and the number of bins in increments of two). Similarly, for FFT frequencies, we tested for different sampling rates (10–30). We report the best results for single features as well as the combination of features through majority voting over both cross-validation methods (LOUO and LOTO) and both accuracy calculations (micro- and macro-averages).

## Results

Stroke-based features and DCC-based similarity measures were highly predictive of surgical skill. Both these approaches involve encoding the raw signal in a higher-level representation before classification—search graphs for stroke-based features and dictionaries for DCC. In contrast, signal-based features and HMM-based similarity measures were poorly discriminative of surgical skill; both these approaches relied on the raw signal (see Tables 1, 2, 3, 4).

We observed that experts have a lower SCC (Fig. 5a), in accordance with our hypothesis that their stroke curvatures remain locally consistent. Experts tool motion was also at higher frequencies compared with novice tool motions (Fig. 5h). Thus, expert surgeons learn to perform a sequence of similar shaped strokes with higher frequency as they acquire skill with experience.

Our analyses show that expert surgeons have a higher SDC than novices (Fig. 5a), as against our initial hypothesis that experts perform strokes of consistent duration. Our findings may be explained by the observation that expert surgeons better adapt to changes in septal anatomy than novices. Consequently, expert surgeons deploy shorter strokes in areas of high adhesions and longer strokes in areas of low adhesions.

Expert surgeons have a greater ACR and larger stroke heights than novices (Fig. 5a, b). These findings are consistent with our hypothesis that novice surgeons make tentative tool motions, whereas expert surgeons make more definitive and decisive tool motions. As a result, experts elevate larger areas of the mucosal flap with larger forces when compared with novice surgeons.

The distribution of absolute angles illustrates that expert surgeons prefer to first elevate the septum deep into the nose (depth-first approach), resulting in small absolute angles. On the other hand, novice surgeons tend to elevate the width of the outer flap before dissecting deeper into the nose (breadth-first approach), resulting in large absolute angles (Fig. 5d).

The distribution of relative angles demonstrates that experts have smaller values than novice surgeons. This observation is consistent with our hypothesis that experts systematically dissect the tissue, resulting in sequences of strokes along a straight line and small absolute angles. Novice surgeons tend to be more unplanned in their approach with strokes positioned more in a zigzag fashion along the septum and resulting in large absolute angles (Fig. 5e).

Finally, experts search for adhesions more efficiently than novice surgeons (Fig. 5c). The arc length distribution shows that the distance between start location of consecutive strokes is larger for experts (about 10 mm) than for novices (about 2 mm). This observation is consistent with our hypothesis that expert surgeons elevate larger areas of the mucosal flap with each stroke and thereby effectively position their tool further away for the next stroke. Novice surgeons tend to search and elevate smaller areas of the mucosal flap.

For 30 % of cases (with no head sensor), we estimated the septal plane location (at each time frame) and thus introduced a small noise (7° rotation and 7 mm shift) to the adjusted tool tip signal. We expect that the error in estimating the septal plane has a minimal impact on the performance of our feature-based classifier because the features are computed by averaging the signal over time, thereby suppressing the added noise.

## Discussions

With an accuracy of 91 %, our system performs on par with the state of the art metrics for skill assessment of structured surgical activity as summarized in [12]. Current techniques for automatic objective skill assessment falls into two major categories: generic aggregate metrics and statistical models (such as HMM). Generic aggregate metrics have been shown to accurately assess skill in structured tasks in the training laboratory and in box trainers and virtual reality simulators. Such global metrics are typically applicable to compare skill across surgeons only when they perform the same structured task.

Most generic aggregate metrics computed using tool motion data may not be immediately translated into measuring skill in the OR because of ambiguity in applying them with the highly variable kinematics observed in the OR. For example, generic aggregate metrics of time, path length, speed, and motion efficiency may not be computed in the septoplasty context because the amount of dissection that is required varies across patients. Other metrics such as motion smoothness, force, dominant frequencies, and profile of velocity and acceleration, as they are currently defined for the overall task/procedure, are not applicable directly to septoplasty, because they are dependent on what segment of the surgery is being performed. In our method, we investigated the idea of local smoothness (stroke duration and curvature) and force applied for tissue dissection (stroke height) during septoplasty, and showed that these metrics using the adaptive definitions we used contain discriminatory information on surgical technical skill.

Conventional statistical methods such as HMMs are ineffective for skill classification in the septoplasty context, as our data show, owing to the high variability in unstructured tool motion observed during the surgery. In addition, these methods do not explain how to perform the septoplasty or provide feedback on how to better perform the surgery. Our method extracts the inherent structure in the surgery by transforming the raw unstructured motion data into a sequence of strokes. The properties of strokes that we computed as metrics allows for individualized skill assessment when multiple surgeons operate in a given procedure with relatively short duration of signals from each surgeon. Furthermore, the metrics we computed provide feedback in a format that surgeons can easily understand and efficiently apply in real-life teaching situations in the operating room.

## Conclusion

Our in-depth study on the structure of septoplasty and objective skill assessment demonstrated that extracting higher-level structures to encode tool motion yields valid objective assessments for surgical skill. Our study sheds light on the actual structure of tool motion during septoplasty, which is otherwise considered an unstructured and hard to teach procedure. Thus, our findings provide faculty surgeons with tools to effectively teach the procedure to trainees and objectively evaluate surgical skill and competence. Furthermore, our analyses provide trainees with specific insights into how to perform the procedure like an expert, thereby facilitating more efficient skill acquisition. Whether using stroke-based or DCC-based skill assessment and feedback translates into faster and better acquisition, and longer retention, of skill among trainees needs to be determined through subsequent clinical studies.

One possible technical improvement for the proposed features is to structure them in a way to capture temporal changes throughout a single trial. This information is already contained in the search graph. Discovering these changes would allow to mark specific anomalies in the time series as targeted feedback for trainee surgeons.

**Acknowledgments** We gratefully acknowledge support from NIH 5R21DE022656-02. We would also like to thank participating attending surgeons—Drs. Kofi Boahene, Patrick Byrne, Ira Papel, and Theda Kontis, and trainee surgeons—Drs. Sun Ahn, Amit Kocchar, Linda Lee, Ryan Li, Myriam Loyo, Sofia Lyford-Pike, Peter Revenaugh, David Smith, Babar Sultan, David Mener, Samuel Oyer, Daniel Sun, at the Johns Hopkins Medical Institutions.

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical standard** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

**Informed consent** For this study (approved by Institutional Review Board at Johns Hopkins), no written consent was necessary.

## References


1. Sugden C, Aggarwal R (2010) Assessment and feedback in the skills laboratory and operating room. Surg Clin North Am 90:519–533
2. Champagne BJ (2013) Effective teaching and feedback strategies in the or and beyond. Clin Colon Rectal Surg 26:244–249
3. Fettman N, Sanford T, Sindwani R (2009) Surgical management of the deviated septum: techniques in septoplasty. Otolaryngol Clin North Am 42:241–253
4. Hwang PH, McLaughlin RB, Lanza DC, Kennedy DW (1999) Endoscopic septoplasty: indications, technique, and results. Otolaryngol Head Neck Surg 120:78–82
5. Bhattacharyya N (2006) Ambulatory sinus and nasal surgery in the United States: demographics and perioperative outcomes. Laryngoscope 120:635–638
6. Stewart MG, Witsell DL, Smith DL, Weaver EM, Yueh B, Hannley MT (2014) Development and validation of the Nasal Obstruction Symptom Evaluation (NOSE) scale. Otolaryngol Head Neck Surg 130:157–163
7. Mason JD, Ansell J, Warren N, Torkington J (2013) Is motion analysis a valid tool for assessing laparoscopic skill? Surg Endosc 27:1468–1477
8. Moorthy K, Munz Y, Sarkar SK, Darzi A (2003) Objective assessment of technical skills in surgery. BMJ 327:1032–1037
9. Van Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP, Dankelman J (2010) Objective assessment of technical surgical skills. Br J Surg 97:972–987
10. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, Bello F, Darzi A (2007) An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245:992–999
11. Dosis A, Aggarwal R, Bello F, Moorthy K, Munz Y, Gillies D, Darzi A (2005) Synchronized video and motion analysis for the assessment of procedures in the operating theater. Arch Surg 140:293–299
12. Oropesa I, Sanchez-Gonzalez P, Lamat P, Chmarra MK, Pagador JB, Sanchez-Margallo JA, Sanchez-Margallo FM, Gmez EJ (2011) Methods and tools for objective assessment of psychomotor skills in laparoscopic surgery. J Surg Res 171:e81–e95
13. Reiley CE, Lin HC, Yuh DD, Hager GD (2011) Review of methods for objective surgical skill evaluation. Surg Endosc 25:356–366
14. Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager GD (2009) Data-derived models for segmentation with application to surgical assessment and training. Med Image Comput Comput Assist Interv 12:426–434
15. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation. IPCAI 2012. LNCS, vol 7330, Springer, Heidelberg, pp 167–177
16. Ahmidi N, Gao Y, Bejar B, Vedula SS, Khudanpur S, Vidal R, Hager GD (2013) String motif-based description of tool motion for detecting skill and gestures in robotic surgery. MICCAI 16:26–33
17. Ahmidi N, Hager GD, Ishii L, Gallia GL, Ishii M (2012) Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery. Springer, Berlin